An inversion is a genomic rearrangement where a piece of DNA is replaced by its reverse complement and re-inserted into the genome. When the reversed DNA piece is re-inserted at its original site, it is called an inversion. Inversions are discovered by sequence analyses and are called scale inversions between human and chimpanzee. Submicroscopic inversions that are too small to be detected by a genomic aligner are referred to as micro-inversions. We call the inversions that are too small to be detected by a classic aligner micro-inversions. There have been reports of inversion polymorphism in human genomes (Bansal et al., 2007; Feuk et al., 2005; Sindi and Raphael, 2010; Szamalek et al., 2006), indicating that inversion is an important feature of intraspecies genomic structure. Breakpoints defined by inversions have been used extensively in genome comparison and ancestral genome reconstruction (Bourque et al., 2002; Ma et al., 2006; Peng et al., 2006; Sankoff, 2006). Chaisson et al. (2006) used inversions to reconstruct a phylogenetic species tree as a new approach to supplement the traditional phylogenetic analysis based on substitutions alone. We observed that some spurious alignments in multispecies alignments are caused by undetected inversions. Discovering such inversions and correcting their alignments can improve alignment quality, which in turn can improve the accuracy of downstream data analysis based on alignments.

Although human and chimpanzee are the most closely related species, the studies on micro-inversions between them (Chaisson et al., 2006; Feuk et al., 2005; Lee et al., 2008; Ma et al., 2006) showed large discrepancies. The earliest result reported 1576 putative inversions (Feuk et al., 2005), and later, it was found that the majority of these were artifacts (Chaisson et al., 2006; Kolb et al., 2009). Chaisson et al. (2006) reported 426 inversions, and the refined method of Ma et al. (2006) identified a similar number of inversions; however, only ~59% of inversions were consistent in these studies (Chaisson et al., 2006). Lee et al. (2008) reported 323 inversions among which 252 could be clearly characterized. These discrepancies indicate that identifying inversions is difficult, even for closely related species. The problem is even more difficult for more distantly related species because of greater sequence divergence, higher frequency of inversions and higher likelihood that the inversions are nested.

All these studies rely on alignments produced by BLASTZ (Schwartz et al., 2003) [now updated to LASTZ (Harris, 2007), though the methodology remains largely the same], which is used by major comparative genomic resources such as the UCSC Genome Browser (Kent et al., 2005) (hereafter referred to as the browser). Below we illustrate some limitations of BLASTZ in aligning inversions, as these observations will guide us in looking for missing micro-inversions in the whole genome alignment. By examining this typical genomic aligner, we argue that there exist pico-inversions that are not detectable by the existing alignment tools.

Briefly, BLASTZ aligns two sequences that are soft masked where repetitive sequences are distinguished from non-repetitive ones. BLASTZ starts by collecting hits between two non-repetitive sequences using a seed of a certain pattern, then extends them that inversions are related to some diseases (Gimelli et al., 2003; Osborne et al., 2001; Visser et al., 2005) and may suppress recombination (Stefansson et al., 2005). There have been reports of inversion polymorphism in human genomes (Bansal et al., 2007; Feuk et al., 2005; Sindi and Raphael, 2010; Szamalek et al., 2006), indicating that inversion is an important feature of intraspecies genomic structure. Breakpoints defined by inversions have been used extensively in genome comparison and ancestral genome reconstruction (Bourque et al., 2002; Ma et al., 2006; Peng et al., 2006; Sankoff, 2006). Chaisson et al. (2006) used inversions to reconstruct a phylogenetic species tree as a new approach to supplement the traditional phylogenetic analysis based on substitutions alone. We observed that some spurious alignments in multispecies alignments are caused by undetected inversions. Discovering such inversions and correcting their alignments can improve alignment quality, which in turn can improve the accuracy of downstream data analysis based on alignments.

Although human and chimpanzee are the most closely related species, the studies on micro-inversions between them (Chaisson et al., 2006; Feuk et al., 2005; Lee et al., 2008; Ma et al., 2006) showed large discrepancies. The earliest result reported 1576 putative inversions (Feuk et al., 2005), and later, it was found that the majority of these were artifacts (Chaisson et al., 2006; Kolb et al., 2009). Chaisson et al. (2006) reported 426 inversions, and the refined method of Ma et al. (2006) identified a similar number of inversions; however, only ~59% of inversions were consistent in these studies (Chaisson et al., 2006). Lee et al. (2008) reported 323 inversions among which 252 could be clearly characterized. These discrepancies indicate that identifying inversions is difficult, even for closely related species. The problem is even more difficult for more distantly related species because of greater sequence divergence, higher frequency of inversions and higher likelihood that the inversions are nested.

All these studies rely on alignments produced by BLASTZ (Schwartz et al., 2003) [now updated to LASTZ (Harris, 2007), though the methodology remains largely the same], which is used by major comparative genomic resources such as the UCSC Genome Browser (Kent et al., 2005) (hereafter referred to as the browser). Below we illustrate some limitations of BLASTZ in aligning inversions, as these observations will guide us in looking for missing micro-inversions in the whole genome alignment. By examining this typical genomic aligner, we argue that there exist pico-inversions that are not detectable by the existing alignment tools.

Briefly, BLASTZ aligns two sequences that are soft masked where repetitive sequences are distinguished from non-repetitive ones. BLASTZ starts by collecting hits between two non-repetitive sequences using a seed of a certain pattern, then extends them...
Again in Figure 1, the dotted antidiagonal line shows a missing inversion. The black and grey diagonal lines are alignments from the browser. The gaps are insertion(s) in human and/or deletion(s) in chimpanzee.

Gaplessly in both directions, and finally extends them further while allowing gaps to form alignment blocks. Once there is an alignment between non-repetitive sequences, it behaves as an anchor and extends into the flanking regions where repetitive sequences can be aligned. However, some orthologous repeats cannot be aligned due to large indels that block the alignment extension from the anchor. In Figure 1, the alignments in non-repeat regions extend into repeats, as indicated by the grey diagonal lines. But, there is no alignment on the browser between human positions hg19:chr1:35,358,810 ($b_1$) and 35,359,676 ($e_1$). The corresponding locations in chimpanzee are panTro3:chr1:35,081,773 ($b_2$) and 35,082,501 ($e_2$). The dotted diagonal line is a missing alignment due to the big gaps blocking the alignment extension. A long run of unsequenced positions (which is not rare in the chimpanzee assembly) has the same effect as a big gap. To obtain the alignments of inversions, one sequence is aligned to the reverse complement of the other. Since there are fewer orthologous non-repetitive sequences on the reverse strand (assuming the majority of the sequence is not inverted), there are fewer alignment anchors to extend into repeat regions. Therefore, it is more difficult to align inversions in genomic regions with repeats. Again in Figure 1, the dotted antidiagonal line shows a missing inversion discovered in our study. In this example, when aligning human sequence with the reverse complement of chimpanzee sequence, there is no hit (because hits are only collected between non-repetitive sequences), and therefore, there is no alignment. Even after BLASTZ correctly produces an inversion as an alignment block, it may be screened out in a post-processing step such as chain-net (Kent et al., 2003) at the browser since chain-net gives preference to a long alignment block. When an inversion is surrounded by strong alignments, its flanking alignments may be so strong that they compensate for the low score produced by the misalignment in the middle, and the real inversion is discarded. Chaisson et al. (2006) showed an example where an inversion of 290 bases was screened out in chain-net and spurious alignment was presented where the inversion was supposed to be.

Indels in alignments can be as small as just one base, since the gapped extension step in the alignment procedure is done via a dynamic programming model, which explicitly accounts for gaps as short as one base. However, inversions do not enjoy such special treatment in the computational model. It was pointed out that rearrangements may happen at all scales, but small rearrangements are not detected by the alignment, because the aligner is not designed to handle such small rearrangements (Kent et al., 2003). The shortest inversion found in large-scale alignments is determined by the alignment significance score threshold together with other alignment parameters. For example, an alignment match has a score of 91 or 100 by default in BLASTZ, and the default alignment significance score threshold is 3000. Assuming that there is no gap or mismatch in an alignment, the shortest significant alignment is around 3000/100 ~ 30000/91 (=30 ~ 33) bases. This tells us that the shortest inversion that is detectable by BLASTZ (and thereafter chain-net) is also around this size. One cannot just simply decrease the alignment significance score threshold to find the shorter inversions, because it will produce a large number of spurious alignments. Here we present an approach to detect pico-inplace-inversions between human and chimpanzee. Since there are large discrepancies among previous studies on micro-inversions, we conduct our own analysis on micro-inversions and apply several rules to ensure accuracy of our discoveries. Since our goal of identifying micro-inversions is to help study pico-inplace-inversions, we restrict our analysis to micro-inplace-inversions. After we obtain micro-inversions, we have an initial (under)estimate of the inversion rate between human and chimpanzee. We look for pico-inversions starting from this initial rate and update it once we obtain convincing pico-inversions. After several rounds of updating, this rate stabilizes and gives us a more accurate estimate on the number of inplace-inversions between two genomes. The pipeline of detecting pico-inplace-inversions is implemented in PicoInversionMiner. We use out-group information to preliminarily verify the pico-inversions between human and chimpanzee detected by this tool and use simulations to systematically evaluate the tool.

2 METHODS

2.1 Detection of micro-inplace-inversions

Micro-inversions are long enough to form significant alignments. Many of them are recorded in chain-net alignment. Some are missing due to the artifacts of the aligner or post-processing procedures as we described above. We categorize micro-inplace-inversions into several types based on how we look for them:

- **Type I**: inversions that are recorded in chain-net as individual alignment blocks. For this type, we need to carefully identify the place ones and ensure that an inversion is not counted as multiple small ones represented in several alignment blocks.
- **Type II**: inversions that are not aligned at all due to artifacts of the aligner that we discussed above (also shown in Fig. 1). We look for such inversions between two adjacent alignment blocks. For a pair of such blocks, we unmask the sequence segments between the two blocks and run BLASTZ between these sequences. We then use the approach for Type I to look for inversions among the resulted alignments.
- **Type III**: inversions that are aligned by the aligner but screened out by post-processing procedures. These inversions are inside an alignment block of chain-net. For each alignment block, we unmask both sequences and align the first sequence with the reverse complement of the second using BLASTZ. A resulted alignment is then potentially a micro-inversion.

To avoid spurious and non-orthologous inversions, we enforce several rules to the potential micro-inversions:

(1) An inversion must be surrounded by strong alignments on both flanking regions. The distances from an inversion to these alignments must be within a threshold (e.g. 2000 bases).
The results of micro-inversions help us to design an approach to detect
likely to accompany an inversion. Between human and chimpanzee, the results suggest that a sequence gain or loss is
are much less likely considering the close evolutionary relationship between
deletions. Since independent inversion and indel(s) at exactly the same site
proportion (89%) of (a set of selected) inversions immediately flanked by
one gap on their flanking sites. Kolb et al. (2009) reported an even higher proportion (89%) of a set of selected inversions immediately flipped by
deletions. Since independent inversion and indel(s) at exactly the same site
are much less likely considering the close evolutionary relationship between
human and chimpanzee, the results suggest that a sequence gain or loss is
likely to accompany an inversion.

2. Detection of pico-inplace-inversions

The results of micro-inversions help us to design an approach to detect
pico-inversions: the number of micro-inversions gives an initial estimate
of the number of micro-inversions genome-wide, and finally analyze the time complexity of the
pipeline.

2.2 Modeling the relationship between pico-inversions

For the null model, we consider that in an alignment block $M$, matches, substitution blocks and gaps follow a categorical distribution where
$p = (p_0, p_1, \ldots, p_{\text{gap}}, p_{\text{inv}})$. We then have

$$P_{\text{inv}} = P(r_{\text{inv}}, \ldots, r_{\text{gap}}, \ldots, r_{\text{inv}}) = \left( \prod_{i=0}^{20} p_i^{r_i} \right)^{20}$$

where $r_{\text{inv}}$'s ($i > 0$) and $r_{\text{gap}}$ are the counts of micro-inversions, substitution blocks of length $i$ and $M$, respectively. With a maximum likelihood estimate, we get

$$P_{\text{inv}} = \left( \prod_{i=0}^{20} \hat{p}_i^{r_i} \right)^{20}$$

Let $p_{\text{inv}}$ denote the probability of having an inplace-inversion (of any length) at any position in the genome. For the alternative model, we consider that matches, substitution blocks, gaps and inplace-inversion(s) follow a categorical distribution where $p' = (p_{\text{orig}}', p_1', \ldots, p_{\text{gap}}', p_{\text{inv}}')$. After the inversion sequence is reversed with its complement in an
alignment block $M$, the new optimum alignment becomes $M'$. Let $\text{INV}$ denote an inversion event. We now have

$$P_{\text{inv}} = P(r_{\text{inv}}', \ldots, r_{\text{gap}}', \text{INV}(p', M')) = P(r_{\text{inv}}', \ldots, r_{\text{gap}}', \text{INV}; p', M') \Pr(\text{INV}(p', M'))$$

$$= \left( \prod_{i=0}^{20} \hat{p}_i^{r_i} \right)^{20} p_{\text{inv}}$$

We use the count of micro-inversions as an initial estimate of the number of inversions $C_{\text{inv}}$. To compute $p_{\text{inv}}$, we will update it with newly identified pico-inversions. Now $C_{\text{inv}} = \sum C_0 + C_{\text{gap}} + C_{\text{inv}}$. Let $\hat{p}_i = \hat{p}_i/C_{\text{inv}}$ and $\hat{p}_{\text{gap}} = C_{\text{gap}}/C_{\text{inv}}$ be the maximum likelihood estimates of $p_i$ and $p_{\text{gap}}$, respectively. We then have

$$p_{\text{inv}} = 1 - \sum C_0 - \hat{p}_i - \hat{p}_{\text{gap}}$$

Representative values of $\hat{p}_i$'s and $\hat{p}_{\text{gap}}$ are shown in Table 2 under iteration 0. We notice that $\hat{p}_i(1:1) > \hat{p}_i$ (the probability of $i$ contiguous substitutions assuming their independence), which verifies the non-independence of substitutions.
caused by the inversion, and the other one is considered an independent indel by the inversion. (c) Although gaps are immediately next to the inversion, they are located on different sequences. Only one of them is considered to be caused by the inversion, and the other one is considered an independent indel event. However, it does not matter in the computational model which gap is counted as part of the inversion event. (d) The gaps are not immediately next to the inversion, and they are considered independent indel events.

Fig. 2. Consideration of gaps around an inversion. Inside each box, an inversion is already corrected by replacing it with its reverse complement so two sequences in the box produce the perfect alignment. (a) and (b) Gaps are immediately next to the inversion, and they are considered to be caused by the inversion. (c) Although gaps are immediately next to the inversion, they are located on different sequences. Only one of them is considered to be caused by the inversion, and the other one is considered an independent indel event. However, it does not matter in the computational model which gap is counted as part of the inversion event. (d) The gaps are not immediately next to the inversion, and they are considered independent indel events.

Fig. 3. Dot-plot illustration of notations used in PicoInversionMiner. The diagonal lines correspond to the original alignments. The thick antidiagonal line is the reverse hit, and the dotted lines on two sides of it indicate the gapless extension.

For every subsequence \( s_q \) of at least five bases in \( s'_2 \), \( s_q \) in \( s_2 \) is replaced by its reverse complement \( s'_2 \) to form \( s'_q \). \( M' \) denotes the global alignment between \( seq_1 \) and \( seq_2' \) under the restriction that no gap is allowed in \( s'_2 \), because we do not allow a gap inside a pico-inversion. \( n_{max} \) denotes the \( t_a \) whose \( M' \) has the highest \( P_{est} \), and its alignment is \( M'_{max} \). (a) The segment in the first sequence aligned to \( n_{max} \) may be a potential pico-inversion. We then compute \( P_{est} \) using \( M' \) and compare it with \( P_{est} \) of \( M'_{max} \). If \( P_{est} > P_{est}(0.01) \), it means that this potential inversion most likely will not pass the subsequent likelihood ratio tests and is discarded.

The concern of non-orthogonal alignment in detecting pico-inversions is the same as in detecting micro-inversions. With the observation that spurious alignments usually have higher rates of mismatches and gaps, we enforce two criteria to exclude false positive pico-inversions. First, suppose there are \( m \) mismatches in \( M'_{max} \) where there are \( n \) aligned positions, the probability of having at least \( m \) mismatches in \( n \) positions must be less than a threshold \( P_{est} \) (e.g. 1%). For simplicity, we assume mismatches to be independent and use binomial distribution here. Second, suppose there are \( g \) gaps in \( M'_{max} \) (excluding the ones considered part of the inversion event as explained before), the probability of having at least \( g \) gaps in \( n \) positions must also be less than \( P_{est} \). When a gap is allowed to exist, its length must be within \( L_{gap} \) bases (e.g. five bases, since most gaps are within five bases observed from the whole genome alignment). \( P_{est} \) and \( L_{gap} \) are referred to as orthogonal thresholds.

The above steps produce a set of potential pico-inversions. For each potential pico-inversion, we conduct the likelihood ratio test defined in Formula (3). If the likelihood ratio is greater than a threshold \( T \), the pico-inversion is reported. After the first iteration, a certain number of pico-inversions are determined. For example, when \( T = 6.64 \) (which corresponds to 1% significance level of \( x^2(l) \)), the count is 1986. This count shows that our initial \( p_{est} \) (based on the count of micro-inversions) was an underestimate. We then update \( p_{est} \) using the new number of inversions. At the same time, some substitution blocks and gaps are found to be caused by inversions, and the values of \( p_{est} \) are updated accordingly. Using the updated rates, we rescan the set of potential pico-inversions. For example, when \( T = 6.64 \), the second iteration produces 4129 pico-inversions. We iteratively update the parameters from the newly discovered inversions and rescan the set of potential pico-inversions until the number of determined inversions stabilizes (i.e. \( p_{est} \) converges), which takes several iterations.

2.2.3 Time complexity of PicoInversionMiner The above steps are summarized in Figure 4. There are two major components in PicoInversionMiner. The first component (ScanPotentialPicoInversions) linearly scans the whole genome alignment and obtains the set of potential pico-inversions \( I \). The second component (lines 3-9 in PicoInversionMiner in Fig. 4) iteratively scans \( I \), conducts a likelihood ratio test for each potential pico-inversion and updates parameters until the count of pico-inversions stabilizes. We can analyze the time complexity separately on these two components. Suppose \( x \) is a sequence, \( |x| \) denotes the length of \( x \). Supposing \( X \) is a set, \( |X| \) denotes the number of elements in \( X \). For the first component, the most time-expensive step is at line 16, the global alignment between \( seq_1 \) and \( seq_2' \). Based on this step and assuming the sizes of \( s'_2 \), \( seq_1 \) and \( seq_2' \) are the same for all hits, the time complexity of this component is \( O(|H|) \times |s'_2| \approx |seq_1| \times |seq_2'| \), where \( H \) is the set of reverse hits in an alignment block, and \( |H| \) is the total number of reverse hits in the genome. Since \( |seq_1| \) and \( |seq_2'| \approx |H| \), the time complexity of the first component is therefore \( O(|H|) \times |s'_2| \). The time complexity for the second component is \( O(|K| \times |I|) \) where \( K \) is the number of iterations. Since \( |K| \approx |H| \times |s'_2| \), the overall time complexity of PicoInversionMiner is \( O(|H| \times |s'_2| \times |I|) \). The typical size of \( s'_2 \) is at most dozens of bases. Since we restrict the reverse hits to be close to the alignment diagonal (Fig. 3), the total number of reverse hits is largely linear to the genome size. The running time measured in CPU time is described in Section 3.2.
Fig. 4. Pseudocode of PicoInversionMiner.

PicoInversionMiner(\(B,S,P,C_{\text{init}},T\)) // \(B\): chain-net alignment blocks // \(P\): the initial parameters // \(C_{\text{init}}\): the initial inversion count based on micro-inversions // \(T\): the significance threshold of likelihood ratio test

1. \(I \leftarrow \text{ScanPotentialPicoInversions}(\(B\))\)
   // \(I\) is a set containing \(<M,M_{\text{max}}^{\text{best}},s_{\text{max}}^{\text{best}}\)>

2. \(C_{\text{init}} \leftarrow C_{\text{init}}\)

3. \(V \leftarrow \emptyset, C \leftarrow C_{\text{init}}\)

4. for every entry \(<M,M_{\text{max}}^{\text{best}},s_{\text{max}}^{\text{best}}\> \in I\)

5. \(t \leftarrow \text{likelihood ratio computed from formula (1), (2), and (3)}\)

6. if \(t > T\)

7. \(V = V \cup \{s_{\text{max}}^{\text{best}}\}; C = C + 1\)

8. if \(C > C_{\text{init}}\)

9. \(C_{\text{init}} \leftarrow C\); Update parameters; GOTO line 3;

\(V\) contains the final results of pico-inversions.

4 RESULTS

We used hg19-panTro3 chain-net alignment from the browser to detect both micro- and pico-in-place-inversions between human and chimpanzee. Here we present the results and analyze the basic characteristics of these inversions. For micro-inversions, we compared our result with the result of Chaisson et al. (2006). For pico-inversions, we conducted a preliminary verification based on out-group information.

3.1 Micro-in-place-inversions

Table 1 shows the counts of different types of micro-in-place-inversions between human and chimpanzee. We see that the counts depend on the threshold of the distance from the inversion to its flanking alignment blocks. The shorter the distance threshold, the more significant evidence that the inversion is inplace. Note that because Type III inversions are the ones found inside alignment blocks, they are all surrounded by nearby alignments (from the same alignment block). Therefore, the counts of Type III inversions are the same for different distance thresholds. We take the value of 425 to compute the initial \(\hat{p}_{\text{inv}}\) since this count is the most consistent with what was reported in Chaisson et al. (2006). Figure 5 shows the length distribution of micro-in-place-inversions that are shorter than 400 bases. It is obvious from the plot that shorter inversions have higher frequencies in general, which also implies the existence and prevalence of pico-inversions.

We performed a preliminary comparison of these inversions with the of Chaisson et al. (2006). Of the 426 inversions detected by Chaisson et al. (2006), 424 were converted from hg17 to hg19 by liftOver from the browser. Assume two inversions from two studies are consistent if their human sequences overlap. Among the result of Chaisson et al. (2006), 194 inversions are consistent with the Type I inversions, 17 with the Type II inversions and none with the Type III inversions. Therefore, about 50% of inversions from Chaisson et al. (2006) are found in our study. We took a closer look at the rest that were not found in our study. Among these, 45 inversions’ human sequences are aligned in hg19-panTro3 chain-net with <25% of their lengths, and 147 inversions’ human sequences are aligned to the positive strand of chimpanzee in hg19-panTro3 chain-net; this indicates that the discrepancies on these inversions are largely due to the assembly and alignment differences. The remaining 21 inversions are not reported in our study either because there are nearby rearrangements (and the inversions are determined not to be inplace in our study) or because their PIPs are <95% (which is a criterion of orthologous alignment specified in Section 2.1).

3.2 Pico-in-place-inversions

Unless otherwise noted, the results for pico-inversions in this section are obtained by the \(T\) threshold that corresponds to

Table 1. Counts of micro-in-place-inversions with different thresholds of the distance from the inversion to its flanking alignment blocks

<table>
<thead>
<tr>
<th>Inversion type</th>
<th>Distance thresholds (bases)</th>
<th>Inversion size (bases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>I</td>
<td>85</td>
<td>140</td>
</tr>
<tr>
<td>II</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>III</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>198</td>
<td>276</td>
</tr>
</tbody>
</table>

All inversions have the PIP >95%.

Fig. 5. Histogram of micro-in-place-inversions <400 bases at intervals of 10 bases.
Table 2. A parameter updating process of detecting pico-inversions between human and chimpanzee

<table>
<thead>
<tr>
<th>Parameters</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_0$</td>
<td>9.86 x 10^{-5}</td>
<td>9.86 x 10^{-6}</td>
<td>9.86 x 10^{-7}</td>
<td>9.86 x 10^{-8}</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0.02 x 10^{-2}</td>
<td>0.02 x 10^{-3}</td>
<td>0.02 x 10^{-2}</td>
<td>0.02 x 10^{-6}</td>
</tr>
<tr>
<td>$p_2$</td>
<td>3.98 x 10^{-4}</td>
<td>3.98 x 10^{-5}</td>
<td>3.98 x 10^{-4}</td>
<td>3.97 x 10^{-6}</td>
</tr>
<tr>
<td>$p_3$</td>
<td>2.67 x 10^{-3}</td>
<td>2.66 x 10^{-4}</td>
<td>2.66 x 10^{-3}</td>
<td>2.60 x 10^{-6}</td>
</tr>
<tr>
<td>$p_4$</td>
<td>5.42 x 10^{-6}</td>
<td>5.37 x 10^{-6}</td>
<td>5.06 x 10^{-6}</td>
<td>5.06 x 10^{-6}</td>
</tr>
<tr>
<td>$p_5$</td>
<td>1.70 x 10^{-5}</td>
<td>1.57 x 10^{-5}</td>
<td>1.57 x 10^{-5}</td>
<td>1.57 x 10^{-6}</td>
</tr>
<tr>
<td>$\hat{p}_0p$</td>
<td>1.51 x 10^{-3}</td>
<td>1.51 x 10^{-3}</td>
<td>1.50 x 10^{-3}</td>
<td>1.50 x 10^{-5}</td>
</tr>
<tr>
<td>$\hat{p}_0n$</td>
<td>1.65 x 10^{-6}</td>
<td>1.49 x 10^{-6}</td>
<td>1.73 x 10^{-6}</td>
<td>2.44 x 10^{-6}</td>
</tr>
<tr>
<td>$\hat{c}_{av}$</td>
<td>425</td>
<td>2411</td>
<td>4554</td>
<td>6432</td>
</tr>
</tbody>
</table>

Values for iteration 0 are computed from the chain-net alignment. Values for other iterations are updated using newly found inversions from the last iteration. The count of pico-inversions obtained in the sixth iteration is the same as the one in the fifth iteration.

Table 3. Counts of pico-inversions (≤40 bases) between human and chimpanzee at different significance thresholds and orthologous thresholds.

<table>
<thead>
<tr>
<th>Ortholog check</th>
<th>Significance level of $\chi^2(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>Note</td>
<td>8297</td>
</tr>
<tr>
<td>$P_{h}=0.01$</td>
<td>3292</td>
</tr>
<tr>
<td>$P_{h}=0.05$</td>
<td>2966</td>
</tr>
</tbody>
</table>

Table 4. Analyses of pico-inversions (≤40 bases) produced by PicoInversionMiner

<table>
<thead>
<tr>
<th>Minimum length of pico-inversions</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total count</td>
<td>5946</td>
<td>2428</td>
<td>1106</td>
</tr>
<tr>
<td>Frequency distribution</td>
<td>$f(x)$</td>
<td>$R^2$</td>
<td>$p$</td>
</tr>
<tr>
<td></td>
<td>1617 x 0.86$^a$</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>Total count excluding repeats</td>
<td>4351</td>
<td>1618</td>
<td>600</td>
</tr>
<tr>
<td>Frequency distribution</td>
<td>$f(x)$</td>
<td>$R^2$</td>
<td>$p$</td>
</tr>
<tr>
<td></td>
<td>1402 x 0.84$^b$</td>
<td>0.94</td>
<td>0.91</td>
</tr>
</tbody>
</table>

In this table, 'frequency distribution' refers to the frequency distribution of the lengths of pico-inversions; $R^2$ refers to the goodness of fit of the distribution function; 'repeats' refers to the simple repeats and homopolymers; 'IR' refers to a pair of sequences that are inverted repeats.

Among the 5946 pico-inversions, 4351 are not homopolymers or simple repeats. The frequency distribution of the lengths of these inversions becomes $f(x)=1401.8+0.844^x$ with $R^2=0.94$.

Many of the discovered pico-inversions are simple repeats or homopolymers (a sequence of identical bases). Since the techniques of sequencing, assembling, and aligning these sequences are less reliable, the predication of inversions in these sequences is also less reliable. We examined each pico-inversion whether it is homopolymer (if the inversion’s human or chimpanzee sequence is homopolymer) and whether it is simple repeat (if the inversion’s human sequence overlaps an entry in the simple repeat annotation of hg19). Among the 5946 pico-inversions of a minimum length of five bases, 4351 are not homopolymers or simple repeats. The frequency distribution of the lengths of these inversions becomes $f(x)=1401.8+0.844^x$ with $R^2=0.94$.

It has been proposed that inverted repeats (IR) [a pair of adjacent or nearby sequences where one is the reverse complement of the other] mediate inversions (Kolb et al., 2009; Small et al., 1997). Many of the discovered pico-inversions are associated with IR: the flanking sites of the inversion are IR, the inversion overlaps IR or the inversion is surrounded by nearby IR (e.g. within a distance of 20 bases from the inversion). Among the above 4351 pico-inversions that are not homopolymers or simple repeats, 46.3% are associated with an IR of at least six bases that are perfect matches, 20.5% are associated with an IR of at least 10 bases that have at most

3271
The inversion starts at chr5:66,492,476, which is just several bases away from the inversion occurred just between this pair.

The reverse complement of an 18-base DNA fragment between this pair. (b) An inversion in 5’ UTR of CD180. The circle in the top sequence indicates a substitution to the inversion. (c) An inversion in 5’ UTR of SPRY1. The inversion starts at chr5:66,492,476, which is just several bases away from the translation start position. The subsequence in shade in the box is lost due to the inversion.

Fig. 6. Representative examples of pico-inversions in human. In each subfigure, the bottom panel is the original alignment of human, chimpanzee and gorilla (in this order) from the browser. Gorilla is used as the out-group in these examples. Small circles indicate substitutions based on the original alignment with the out-group. The sequences above wide arrows are ancestral human sequences after the species split between human and chimpanzee. Rectangular boxes indicate inversions. (a) An inversion in 5’ UTR of C1D180. The inversion starts at chr5:66,492,476, which is just several bases away from the translation start position. The subsequence in shade in the box is lost due to the inversion. (b) An inversion in 5’ UTR of SPRY1. The inversion starts at chr4:124,317,977. The circle in the top sequence indicates a substitution based on the original alignment with the out-group. The sequences above wide arrows are ancestral human sequences after the species split between human and chimpanzee. Rectangular boxes indicate inversions. (c) An inversion in 3’ UTR of C7orf41. The inversion starts at chr7:30,201,616. The reverse complement of an 18-base DNA fragment underlined in the top sequence is inserted in the ancestral human sequence, and leads to the perfect match of a pair of inverted repeats of 23 bases. The inversion occurred just between this pair.

Table 5. Verification of pico-inversions (≥5 and ≤40 bases) using an out-group

<table>
<thead>
<tr>
<th>Category</th>
<th>Human</th>
<th>Chimpanzee</th>
<th>No support</th>
<th>No out-group</th>
<th>Partial support</th>
<th>Total alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repetitive</td>
<td>1246</td>
<td>1914</td>
<td>617</td>
<td>352</td>
<td>222</td>
<td>451</td>
</tr>
<tr>
<td>Non-repetitive</td>
<td>253</td>
<td>376</td>
<td>413</td>
<td>426</td>
<td>127</td>
<td>1595</td>
</tr>
</tbody>
</table>

*Repetitive* refers to inversions that are homopolymers or simple repeats. The columns of ‘human’ and ‘chimpanzee’ record the numbers of inversions within human lineage and chimpanzee lineage, respectively. ‘No support’ indicates that there is no evidence of inversion in either species based on the out-group. ‘No out-group’ indicates that there is no gorilla or orangutan sequence aligned to the human sequence of the inversion in the 46-way alignment. ‘Partial alignment’ indicates that the inversion is not completely contained inside an alignment block from the 46-way alignment.

We conducted a preliminary verification of the pico-inversions found between human and chimpanzee using gorilla (or orangutan, in cases where the gorilla sequence does not exist) as an out-group. The assumption is that if an inversion is real and has occurred in a certain lineage, the alignment between the sequence of this lineage and its out-group must be worse than the alignment between the sequence with the inversion corrected and its out-group. Let $A_h$, $A_c$, $A_g$ and $A'_g$ denote the scores of the global alignments between the out-group and the original human sequence, the corrected human sequence, the original chimpanzee sequence and the corrected chimpanzee sequence, respectively. If $A_h < A'_c$ and $A_g > A'_g$, we conclude that the inversion occurred in the human lineage. If $A_h > A'_c$ and $A_g < A'_g$, we conclude that the inversion occurred in the chimpanzee lineage. For other cases, we conclude that there is no evidence of an inversion based on the out-group information, and the reported inversion is a false positive. For the above global alignments, flanking positions (e.g. up to 20 bases) of the inversion are also included in the alignment computation to ensure the alignment accuracy. The gorilla and orangutan sequences are taken from the 46-way Multiz alignment from the browser. The analysis results are recorded in Table 5. Among the 4351 inversions that are not homopolymers or simple repeats, 352 are not aligned to gorilla or orangutan (and they are considered to have no out-group), and 222 are not completely contained inside alignment blocks that possess an out-group; these inversions are excluded from the analysis. Out of the remaining 3777 inversions, 617 are determined to be false positives. Therefore, the false positive rate of detecting inversions that are not repetitive sequences is $617/(1246+1914+617) = 16.3\%$. The analysis results of pico-inversions that are repetitive sequences are also shown in Table 5, and the false positive rate is much higher at $413/(253+376+413) = 39.6\%$. Note that in this analysis, the number of inversions in chimpanzee is much higher than the number in human. This might be related to the fact that the sequencing quality of chimpanzee genome is not as good as the quality of human genome, and PicoinversionMiner is sensitive to substitutions and short indels in the alignment that may be caused by sequencing errors.

### 4 Evaluation of PicoinversionMiner by Simulation

To systematically evaluate the accuracy of PicoinversionMiner, we apply it on simulated genomic sequences and compute its sensitivity (the percentage of true inversions that are detected) and specificity (the percentage of detected inversions that are true
et al.

Table 6. Comparison between simulation data and real data of the alignment between human and chimpanzee

<table>
<thead>
<tr>
<th>Category</th>
<th>Source</th>
<th>( p_0 )</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>Indels</th>
<th>Inversions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rate</td>
<td>Length distribution function</td>
</tr>
<tr>
<td>Real data</td>
<td>Alignment</td>
<td>0.986</td>
<td>0.0122</td>
<td>0.000397</td>
<td>0.00150</td>
<td>( f(x) = 1 - x^{-1.235} )</td>
</tr>
<tr>
<td>Simulation</td>
<td>Sequences</td>
<td>0.986*</td>
<td>0.0122*</td>
<td>0.000398*</td>
<td>0.00151*</td>
<td>( f(x) = 1 - x^{-1.154} )</td>
</tr>
<tr>
<td></td>
<td>Alignment</td>
<td>0.981*</td>
<td>0.0160*</td>
<td>0.000802*</td>
<td>0.00205*</td>
<td>( f(x) = 1 - x^{-1.128} )</td>
</tr>
</tbody>
</table>

The sequences are generated by the simulator according to the models trained from the whole genome alignment. The simulated sequences are then aligned by BLASTZ.

* These data are the average of 100 simulations.

The locations of mutational events are assumed to be uniformly distributed on the sequences. For each nucleotide substitution, the chances of transition and transversion are set to be 66.7 and 33.3%, respectively, since the ratio between the two is 2:1 as observed from the whole genome alignment between human and chimpanzee. We then use BLASTZ to align the simulated sequences.

We have computed \( \hat{p}_i \) for substitution blocks, \( \hat{p}_{ins} \) for indels and \( \hat{p}_{inv} \) for pico-inversions (Table 2), where \( p_i (i>0) \sim N(\hat{p}_i, \hat{p}_i / \sqrt{n}) \), \( p_{ins} \sim N(\hat{p}_{ins}, \hat{p}_{ins} / \sqrt{n}) \) and \( p_{inv} \sim N(\hat{p}_{inv}, \hat{p}_{inv} / \sqrt{n}) \) (where \( n \) is the genome size). For each simulation, we sample \( p_i 's (i>0), p_{ins} \) and \( p_{inv} \) from these normal distributions. The frequencies of indel lengths follow the power-law distribution \( f(x) = 6182215x^{-1.235} \) with the goodness of fit \( R^2 = 0.98 \) obtained from the whole genome alignment between human and chimpanzee. We then sample indel lengths from their distribution function \( f(x) = 1 - x^{-1.235} \) (and then round down to the nearest integer). Note that there is no difference between insertion and deletion when only two species are considered without an out-group. Therefore, we simulated deletions for all indels for simplicity. The frequencies of inversion lengths follow the exponential distribution \( f(x) = 16172.7 \times 0.861^x \) that we obtained earlier, and we sample inversion lengths from their distribution function \( f(x) = 1 - 0.861^{-x} \) (and then round down to the nearest integer) assuming the shortest inversion has five bases. In each sequence of \( n \) bases, we simulate 0.5\( \hat{p}_{ins} \) indels and 0.5\( \hat{p}_{inv} \) inversions assuming each species evolves at the same rate.

We then align two sequences using BLASTZ (with default parameters). Since there are many duplications originated from chromosome \( #21 \), we use single_cov2 program from the TBA/Multiz package (Blanchette et al., 2004) to post-process BLASTZ output to make sure any human position is aligned to chimpanzee at most once (i.e. a sequence segment in human is aligned to only one copy in chimpanzee, instead of multiple homologous copies in chimpanzee). single_cov2 is used in the place of chain-net, because it eliminates non-orthologous homologous alignments caused by duplications similarly to chain-net, and we did not find a stand-alone public application that created chain-net alignments.

A total of 100 pairs of sequences are produced. The average length of each sequence is \(~48.0\text{M} \) bases. The details of the comparison between simulation data and real data are shown in Table 6. Note that the sequences are generated by the simulator according to the models trained from the real data of the whole genome alignment. Therefore, the frequencies of the simulated events in the simulated sequences are very consistent with the real data, but the values computed from the alignments of these sequences show some differences from the real data. The alignment between the simulated sequences actually shows slightly higher species divergence than the divergence between human and chimpanzee observed from the whole genome alignment.

Figure 7a shows sensitivity versus 1−specificity of PicoInversionMiner applied to these simulated sequences. When considering all inversions (shown in Fig. 7b with a minimum length of five bases), the sensitivity is moderate, and the specificity is very high.

![Fig. 7](image-url)
PicoInversionMiner applied to simulated sequences of different above human–chimpanzee simulation for simplicity. The same length distributions of indels and pico-inversions from the sequences with default thresholds and applied PicoInversionMiner to the alignments between these same starting sequence, but without simulating pico-inversions, sequences using the same substitution and indel models from the pico-inversions in Section 5.

For the control experiment, we also simulated 100 pairs of sequences using the same substitution and indel models from the starting sequence, but without simulating pico-inversions. For the 100 pairs of sequences with inversions simulated, the number of simulated inversions ranges from 110 to 117 with an average of 1.1 inversions per pair of sequences. This number of false positives is actually less than the number of false positives discovered between sequences with pico-inversions simulated. For the 100 pairs of sequences with inversions simulated, the number of simulated inversions ranges from 110 to 117 with an average of 112.9, the number of discovered (pico-)inversions ranges from 44 to 77 with an average of 62.8 (with default thresholds) and the number of false positives ranges from 0 to 6 with an average of 1.8 (which leads to the specificity of 97.5%). This fact is related to the algorithm of PicoInversionMiner. The algorithm starts from an initial estimate of the inversion rate \( P_{inv} \), which is based on the number of micro-inversions between human and chimpanzee, and iteratively updates it with newly discovered (pico-)inversions. The higher the value of \( P_{inv} \), the more potential pico-inversions pass the likelihood ratio test and are determined to be pico-inversions; therefore, there is a higher chance of producing false positives. The iteration stops when there is no increase of \( P_{inv} \). The initial estimate of \( P_{inv} \) is very low. When there are no inversions simulated, the number of pico-inversions discovered by this rate is also very low, and the iteration stops after one or two cycles. However, for the sequences with inversions simulated, the number of pico-inversions discovered in the first iteration raises the value of \( P_{inv} \) significantly, and there are further iterations, which leads to more false positives. Note that for the simulations, the value of \( C_{INV} \) in Figure 4 is calculated based on the initial estimate of \( P_{inv} \) and the length of the simulated sequence: \( 1.63 \times 10^{-7} \times 48 = 8 \).

Though PicoInversionMiner is designed to detect pico-inversions between human and chimpanzee, we would like to test its effectiveness on more diverged sequences. To simulate sequences of different divergences, we assume that the rates of substitution blocks, indels and inversions are constant. For example, if two sequences' \( p_1 \) is 0.06, which is around five times greater than the \( p_1 \) between human and chimpanzee (call this value coefficient \( c \)), their other \( p_1 \)'s (\( p_1 \) is more than 1), \( p_{indel} \) and \( P_{inv} \) are also five times greater than the rates between human and chimpanzee. We then use different \( c \) values to simulate sequences at different divergences. When \( c = 1 \), the divergence is between human and chimpanzee, where the PIP is 98.7%; when \( c = 5.7 \), the PIP becomes ~90%. However, we use the same length distributions of indels and pico-inversions from the above human-chimpanzee simulation for simplicity.

Figure 7b shows sensitivity versus \( 1 - \text{specificity} \) of PicoInversionMiner applied to simulated sequences of different \( c \) values. In all, 100 pairs of sequences are produced for each \( c \) value. When sequences are diverged, the sensitivity of detecting very small pico-inversions is very low. Therefore, only inversions no shorter than 10 bases are considered here. For nearly all cases, the specificity is very high (~97%). The sensitivity is acceptable (e.g. >50%) when \( c \leq 4.5 \), which corresponds to a PIP of 94%. We can also observe that after \( T \) reaches the value corresponding to \( \chi^2(1) \) significance level of 0.1, a less strict \( T \) value does not improve sensitivity much. Therefore, PicoInversionMiner is only effective for very similar sequences.

### 5 DISCUSSION AND CONCLUSION

We see that the sensitivity of PicoInversionMiner in detecting very small pico-inversions is low. Actually, many small inversions are simply not detectable by any means. For example, the reverse complement of “CAATG” is “CATG”, and their alignment only contains one mismatch. It is impossible to distinguish the inversion event from the substitution event in this case.

There are also some cases where the inversion is not detectable due to the limitations of the model used by PicoInversionMiner. For example, suppose that there is a five-base inversion in human and suppose that its alignment with chimpanzee shows a substitution block of five bases. Using values from Table 2 iteration 5, we have \( -2 \ln \text{PIP} = 2ln(2.44 \times 10^{-6} \times 0.9865^5)/(1.57 \times 10^{-6}) = 0.74 \), which is not significant enough to conclude an inversion. Note that the rates of substitution blocks are computed from the whole genome alignment, which also includes spurious alignments (Prakash and Tompa, 2007) and non-orthologous homologous alignments. Therefore, the rates of substitution blocks (especially the large ones) are very likely elevated. When a better quality alignment is available, the substitution block rates can be corrected (and most likely be reduced), and some potential pico-inversions, whose likelihood ratio tests were not significant enough before, may be rediscovered.

We have explained that the shortest significant alignment between human and chimpanzee is around 30–33 bases assuming there are no mismatches or gaps. When there are mismatches or gaps, which is more common between more diverged species, the shortest significant alignment is longer. We defined pico-inversions as the ones too small to be detected by the aligner. Therefore, there is no clear distinction between the shortest micro-inversion and the longest pico-inversion. We arbitrarily chose 40 bases as the largest size of pico-inversions in this article.

All inversions discovered in this article are in-place ones. It may be noted that there are micro-inversions that are transposed to different genomic regions. It can be conjectured that there are also pico-inversions that are transposed. However, there lack studies on micro-inversions that are transposed, partly due to the assembly and alignment challenges. It is even more difficult to detect pico-inversions that are transposed. This can be a future work.

Although we tried to simulate genomic sequences as similar to the real sequences as possible based on the properties of substitution blocks, indels and inversions obtained from the whole genome alignment, the simulation cannot perfectly present the real situation. For example, the simulation assumed uniform distribution of the evolutionary events, which is too simplified and may cause bias in the evaluation results. We presented a preliminary approach (by using an out-group) to verify pico-inversions between human and
chimpanzee. The false positive rate based on out-group information indicates that the specificities computed from the simulations may be elevated. It is a future work to develop more advanced methods to verify the pico-inversions and evaluate the tool.

In summary, inversions are important genomic mutations. However, very small inversions have been ignored for a long time partly due to the technical limitation in sequence alignment methodologies. This study verified the existence of inversions as short as several bases and estimated that there are at least thousands of very small inversions between human and chimpanzee. Detection of such events not only provides a more complete picture of genome evolution, but also helps improve alignment quality (by correcting wrong alignments caused by inversions) and facilitates any downstream data analyses based on alignments. We also presented the software tool PicoInversionMiner, which is effective in detecting pico-inversions between very similar sequences. To find very small inversions between more diverged sequences, we need to explore more sophisticated methods.

ACKNOWLEDGEMENT
We thank the Northern Illinois Center for Accelerator and Detector Development (NICADD) at Northern Illinois University for the free access to its computer cluster to perform the large-scale simulations in this project.

Funding: National Institutes of Health grant (R15 HG005913 to M.H.).

Conflict of Interest: none declared.

REFERENCES

Pico-inversions