Sequence analysis

Pico-inplace-inversions between human and chimpanzee

Minmei Hou1,∗, Ping Yao2, Angela Antonou3 and Mitrick A. Johns4
1Department of Computer Science, 2School of Nursing and Health Studies, 3Department of Mathematical Sciences
and 4Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, USA

ABSTRACT

Motivation: There have been several studies on the micro-inversions between human and chimpanzee, but there are large discrepancies among their results. Furthermore, all of them rely on alignment procedures or existing alignment results to identify inversions. However, the core alignment procedures do not take very small inversions into consideration. Therefore, their analyses cannot find inversions that are too small to be detected by a classic aligner. We call such inversions pico-inversions.

Results: We re-analyzed human–chimpanzee alignment from the UCSC Genome Browser for micro-inplace-inversions and screened for pico-in-place-inversions using a likelihood ratio test. We report that the quantity of inplace-inversions between human and chimpanzee is substantially greater than what had previously been discovered. We also present the software tool PicoInversionMiner to detect pico-in-place-inversions between closely related species.

Availability: Software tools, scripts and result data are available at http://faculty.cs.niu.edu/~hou/PicoInversion.html.

Contact: mhou@cs.niu.edu

Received on June 6, 2011; revised on September 22, 2011; accepted on October 6, 2011

1 INTRODUCTION

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 inplace-inversion. Very large inversions can be observed under a microscope. Yunis et al. (1980) reported nine such large-scale inversions between human and chimpanzee. Submicroscopic inversions are discovered by sequence analyses and are called micro-inversions; their sizes range from dozens to millions of bases. Most studies focusing on micro-inversions rely on genomic alignments; however, only ∼59% of inversions are consistent in these studies (Chaisson et al., 2006). Lee et al. (2008) reported 323 inversions among which 252 could be clearly characterized. They have not been any explicit study on pico-inversions; their existence and prevalence have been unknown.

Characterization of inversions has been useful in any aspects of biomedical research. Navarro et al. (2003) suggested that very large inversions might have been responsible for a speed up of the speciation of human and chimpanzee. There is also evidence 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 interspecies 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) 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 alignment block. 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 gray diagonal lines. But, there is no alignment on the browser between human positions hg19:chr1:35,358,810 (b1) and 35,359,676 (e1). The corresponding locations in chimpanzee are panTro3:chr1:35,081,773 (b2) and 35,082,501 (e2). 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 ∼ 30009/100 = 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 place-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 in-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).
We then have human and chimpanzee, the results suggest that a sequence gain or loss is are much less likely considering the close evolutionary relationship between 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 flanked 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.

### 2.2 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 inversion rate between two species. Also, the prevalent existence of gaps on flanking sites of inversions suggests that a potential inversion and gap(s) immediately adjacent to it should be considered a single event instead of multiple independent events. Since pico-inversions are too short to form significant alignments, the approaches used in identifying micro-inversions do not apply here. We use probability analysis to detect them.

In the rest of this section, we first describe the probability model to determine a pico-inversion, then present the approach to detect pico-inversions genome-wide, and finally analyze the time complexity of the pipeline.

#### 2.2.1 The probability model to determine a pico-inversion

For an alignment block containing a potential pico-inversion, we use \( P_{\text{orig}} \) and \( P_{\text{inv}} \) to denote the probabilities of evolutionary events given the original alignment and the alignment with the inversion corrected, respectively. We use a likelihood ratio test between \( P_{\text{orig}} \) and \( P_{\text{inv}} \) to draw a conclusion about the inversion.

In our models, the substitution rates affect the detection of pico-inversions greatly. We cannot simply assume the independence of substitutions in this study since it causes significant bias toward more false positive pico-inversions. We call a segment of \( i \) contiguous substitutions (where two flanking positions are matches) a substitution block (of length \( i \)), which is considered a single categorical event outcome, and use \( p_i \) to denote the probability of such an event at any position in the genome. The longest run of substitutions in human-chimpanzee chain-net alignment has 20 bases. Therefore, we consider \( p_i \)'s up to \( p_{20} \). \( p_0 \) corresponds to the probability of no change (e.g. match) at a position. Let \( P_{\text{inv}} \) denote the probability of starting a gap at any position in the genome. Let \( C_i \) and \( C_{\text{gap}} \) be the counts of substitution blocks (of length \( i \)) and gaps (regardless of length) in the whole genome alignment, respectively. \( C_i \) is the count of matches. \( C_{\text{inv}} = \sum C_i + C_{\text{gap}} \). Let \( \hat{P}_i = C_i/C_{\text{inv}} \) and \( \hat{P}_{\text{gap}} = C_{\text{gap}}/C_{\text{inv}} \) be the maximum likelihood estimates of \( p_i \) and \( p_{\text{gap}} \), respectively. When \( C_i = 0 \), \( \hat{P}_i = 1/C_{\text{inv}} \) and \( \hat{P}_{\text{gap}} = 0 \).

We then have \( P_i = 1 - \sum_j \hat{P}_j \hat{P}_{\text{gap}} \) Representative values of \( \hat{P}_i \)'s and \( \hat{P}_{\text{gap}} \) are shown in Table 2 under iteration 0. We notice that \( \hat{P}_i(20) > \hat{P}_{\text{gap}} \) (the probability of i contiguous substitutions assuming their independence), which verifies the non-independence of substitutions.

For the null model, we consider that in an alignment block \( M \), matches, substitution blocks and gaps follow a categorical distribution where \( p = (p_0, p_1, \ldots, p_{20}, p_{\text{gap}}) \). We then have

\[
P_{\text{null}} = P(C_{\text{inv}} = 0, \ldots, C_{\text{inv}} = 20, M) = \left( \prod_{i=0}^{20} \hat{P}_i \right) \hat{P}_{\text{gap}}
\]

where \( x_i, x'_i \)'s \((i > 0)\) and \( x_{\text{gap}} \) are the counts of matches, substitution blocks of length \( i \) and gaps in \( M \), respectively. With a maximum likelihood estimate, we get

\[
P_{\text{null}} = \left( \prod_{i=0}^{20} \hat{P}_i \right) \hat{P}_{\text{gap}}
\]

for an inversion event. We now have

\[
P_{\text{null}}(\text{INV}) = \text{Pr}(x_{\text{gap}}, x'_0, \ldots, x'_{20}, x_{\text{gap}}|M, \text{INV}) = \text{Pr}(x_{\text{gap}}, x'_0, \ldots, x'_{20}, x_{\text{gap}}|M) \text{Pr}(\text{INV}|M)
\]

\[
= \left( \prod_{i=0}^{20} \hat{P}_i \right) \hat{P}_{\text{gap}}
\]

Let \( P_{\text{null}} \) 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 \( p' = (x'_0, x'_1, \ldots, x'_{20}, x'_{\text{gap}}, P_{\text{null}}) \).

After the inversion sequence is replaced with its reverse complement in an alignment block \( M \), the new optimum alignment becomes \( M' \). Let \( \text{INV} \) denote an inversion event. We now have

\[
P_{\text{null}}(\text{INV}) = \text{Pr}(x_{\text{gap}}, x'_0, \ldots, x'_{20}, x_{\text{gap}}|M', \text{INV}) = \text{Pr}(x_{\text{gap}}, x'_0, \ldots, x'_{20}, x_{\text{gap}}|M') \text{Pr}(\text{INV}|M')
\]

\[
= \left( \prod_{i=0}^{20} \hat{P}_i \right) \hat{P}_{\text{gap}}
\]

We use the count of micro-inversions as an initial estimate of the number of inversions \( C_{\text{inv}} \), to compute \( P_{\text{null}} \) and will update it with newly identified pico-inversions. Now \( C_{\text{inv}} = \sum C_i + C_{\text{gap}} + C_{\text{inv}}, x'_i \) \((i > 1)\) and \( x_{\text{gap}} \) are accordingly updated to \( x'_0 \) and \( x_{\text{gap}} \). We then have \( P_{\text{null}} = 1 - \sum_{i=0}^{20} \hat{P}_i \hat{P}_{\text{gap}} - \hat{P}_{\text{gap}} \).

Note that a gap at an adjacent flanking site of an inversion may not be included in \( x_{\text{gap}} \) since it may be caused by the inversion event and considered in \( P_{\text{null}} \), which we have explained above. Figure 2 shows several cases of gaps around a potential inversion and explains the situations where we consider a gap as part of an inversion event.

We use the likelihood ratio test

\[
D = -2 \log \frac{P_{\text{null}}}{P_{\text{inv}}}
\]

which follows \( p(1) \) to conclude a pico-inversion.

#### 2.2.2 Detection of genome-wide pico-inversions

To look for genome-wide pico-inversions, we scan the chain-net alignment. Every alignment block \( B \) has two sequences, one from human and one from chimpanzee. Note that the inversion could have occurred in either species, but we do not intend to differentiate the two cases at this stage. We simply assume that the inversion occurred in the second sequence in an alignment, which does not affect detecting the locations of inversions between two species. For every five bases of the second sequence (call this segment of bases the query and the segment of the first sequence aligned to the query the query counterpart), we look for a subsequence in the first sequence that is identical to the reverse hit. Since we are looking for inplace-inversions, we restrict this search within a range (e.g. 20 bases) of the original cut. For a thorough survey of the inversion blocks, we scan the chain-net alignment. Every alignment block \( B \) has two sequences, one from human and one from chimpanzee. Note that the inversion could have occurred in either species, but we do not intend to differentiate the two cases at this stage. We simply assume that the inversion occurred in the second sequence in an alignment, which does not affect detecting the locations of inversions between two species. For every five bases of the second sequence (call this segment of bases the query and the segment of the first sequence aligned to the query the query counterpart), we look for a subsequence in the first sequence that is identical to the reverse hit. Since we are looking for inplace-inversions, we restrict this search within a range (e.g. 20 bases) of the original cut.
bases to the left and right sides of the query counterpart). Next, we compute the highest scoring gapless extension of the hit (using the first sequence and the reverse complement of the second sequence), allowing at most two mismatches, within 30 bases is slim. Let \( s_1 \) denote the segment of the first sequence in this extended inversion alignment. Let \( s_2 \) and \( s_3 \) denote the segments of the second sequence aligned to \( s_1 \) in the original alignment and the inversion alignment, respectively. Figure 3 illustrates the notations used here. Note that \( s_2 \) is part of the original sequence, not inverted. \( s_2 \) and \( s_3 \) may be the same, which may indicate an inversion without any gain or loss of bases. In most cases, even when \( s_2 \) and \( s_3 \) are not the same, they overlap. Let \( M \) be the subalignment from \( R \) that covers \( s_2 \) and \( s_3 \) with flanking regions \( f_R \) and \( f_L \) on both sides (e.g. of 20 bases). \( M \) aligns \( seq_1 \) and \( seq_2 \), and it contains the potential pico-inversion.

Note that the inversion alignment found above \( s_1 \) versus the reverse complement of \( s_2 \) may be overestimated from the real inversion. We then search the subsequences of \( s_2 \) to determine the best potential inversion in \( M \).

For every subsequence \( s_2 \) of at least five bases in \( s_2 \), \( s_2 \) is replaced by its reverse complement \( s_2' \) to form \( seq_2' \). \( M' \) 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 \( n \) whose \( M' \) has the highest \( P_{orth} \), and its alignment is \( M'_{orth} \) (or the segment in the first sequence aligned to \( \hat{P}_{orth} \)). may be a potential pico-inversion. We then compute \( P_{seq} \) using \( M' \) and compare it with \( P_{max} \) of \( M'_{orth} \). If \( P_{orth} > P_{seq} \) (e.g. \( P_{max} \) is less than 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-orthological alignment in detecting pico-inversions is the same as in detecting micro-inversions. With the observation that spuriously aligned segments usually have higher rates of mismatches and gaps, we enforce two criteria to exclude false positive pico-inversions. First, supposing there are \( m \) mismatches in \( M'_{orth} \), where there are \( n \) aligned positions, the probability of having at least \( m \) mismatches in \( n \) positions must be less than a threshold \( \theta \) (e.g. 1%). For simplicity, we assume mismatches in orthology are independent and use binomial distribution here. Second, supposing there are \( g \) gaps in \( M'_{orth} \), the probability of having at least \( g \) gaps in \( n \) positions must also be less than \( \theta \). When a gap is allowed to exist, its length must be within \( \theta \) bases (e.g. five bases, since most gaps are within 5 bases observed from the whole genome alignment). \( P_{orth} \) and \( L_{gap} \) are referred to as orthologous 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 \( \chi^2(1) \), the count is 186. This count shows that our initial \( P_{orth} \) (based on the count of micro-inversions) was an underestimate. We then update \( P_{orth} \) 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 \( \hat{P}_L \), \( \hat{P}_R \), and \( \hat{P}_{seq} \) 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 (e.g. \( P_{orth} \) 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. Supposing \( s \) is a sequence, \( |s| \) denotes the length of \( s \). Supposing \( X \) is a set, \(|X|\) denotes the number of elements in \( X \). For the first component, the most time-consuming 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 \( O(\sum|H_i| \times |s_2'| \times |seq_1| \times |seq_2'|) \), where \( H_i \) is the set of reverse hits in an alignment block, and \( \sum|H_i| \) is the total number of reverse hits in the genome. Since \( |seq_1| \) and \( |seq_2'| \) is bounded by \( |f_R| \) and \( |f_L| \), the time complexity of the first component is therefore \( O(\sum|H_i| \times |s_2'|^2) \). The time complexity for the second component is \( OK \times \Theta \) where \( K \) is the number of iterations. Since \( |M'_{orth} \) \( |seq_1| \) and \( |seq_2'| \) are the numbers of reverse hits in the genome. The time complexity of PicoInversionMiner is \( O(|H| \times |f_R| \times |f_L| \times |s_2'|^2) \). 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.
3 RESULTS

We used hg19-panTro3 chain-net alignment from the browser to detect both micro- and pico-inplace-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-inplace-inversions

Table 1 shows the counts of different types of micro-inplace-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-inplace-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-inplace-inversions

Unless otherwise noted, the results for pico-inversions in this section are obtained by the $T$ threshold that corresponds to...
Among these, 5946 are pico-inversions (i.e. counts of pico-inversions (Table 2). 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. The parameters excluding exons 1.0% 1.1% 1.5% repeats do not change significantly after each iteration. Given the whole genome alignment between human and chimpanzee, the complete sequencing, assembling and aligning these sequences are 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 $f(x) = 1.401.8 \times 0.844^x$ 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

### Table 2. A parameter updating process of detecting pico-inversions between human and chimpanzee

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Iterations</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_0$</td>
<td>$9.86 \times 10^{-4}$</td>
<td>$9.86 \times 10^{-5}$</td>
<td>$9.86 \times 10^{-6}$</td>
<td>$9.86 \times 10^{-7}$</td>
<td></td>
</tr>
<tr>
<td>$p_1$</td>
<td>$1.22 \times 10^{-2}$</td>
<td>$1.22 \times 10^{-3}$</td>
<td>$1.22 \times 10^{-4}$</td>
<td>$1.22 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>$p_2$</td>
<td>$3.98 \times 10^{-4}$</td>
<td>$3.98 \times 10^{-5}$</td>
<td>$3.97 \times 10^{-6}$</td>
<td>$3.97 \times 10^{-7}$</td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td>$2.67 \times 10^{-3}$</td>
<td>$2.66 \times 10^{-4}$</td>
<td>$2.66 \times 10^{-5}$</td>
<td>$2.60 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td>$5.42 \times 10^{-7}$</td>
<td>$5.37 \times 10^{-8}$</td>
<td>$5.06 \times 10^{-9}$</td>
<td>$5.06 \times 10^{-10}$</td>
<td></td>
</tr>
<tr>
<td>$p_5$</td>
<td>$1.70 \times 10^{-4}$</td>
<td>$1.57 \times 10^{-5}$</td>
<td>$1.57 \times 10^{-6}$</td>
<td>$1.57 \times 10^{-7}$</td>
<td></td>
</tr>
<tr>
<td>$\hat{p}_{inv}$</td>
<td>$1.51 \times 10^{-3}$</td>
<td>$1.51 \times 10^{-4}$</td>
<td>$1.50 \times 10^{-5}$</td>
<td>$1.50 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>$\hat{p}_{ninv}$</td>
<td>$1.65 \times 10^{-7}$</td>
<td>$9.14 \times 10^{-9}$</td>
<td>$1.73 \times 10^{-10}$</td>
<td>$2.44 \times 10^{-11}$</td>
<td></td>
</tr>
<tr>
<td>$\epsilon = \chi²$</td>
<td>425</td>
<td>2411</td>
<td>4554</td>
<td>6432</td>
<td></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>Orthologous check</th>
<th>Significance level of $\chi²(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>Note</td>
<td>8297</td>
</tr>
<tr>
<td>$P_{Ortho}=0.01$</td>
<td>3292</td>
</tr>
<tr>
<td>$P_{Ortho}=0.05$</td>
<td>2966</td>
</tr>
</tbody>
</table>

$\chi²(1)$ significance level of 0.01 and the orthologous thresholds of $P_{Ortho}=0.01$ and $L_{gap}=5$ (these are the default thresholds in PicoInversionMiner). It takes five iterations of parameter updating for $\hat{p}_{inv}$ to stabilize. An example of the parameter updating process is recorded in Table 2. We see that most parameters excluding $\hat{p}_{inv}$ do not change significantly after each iteration. Given the whole genome alignment between human and chimpanzee, the complete process takes ~2.5 h of CPU time on a regular desktop. The counts of pico-inversions found by PicoInversionMiner are reported in Table 3 with different significance thresholds and orthologous thresholds.

Using the default thresholds, 6007 inversions are found by PicoInversionMiner. The shortest has five bases (because of the query size in PicoInversionMiner), and the longest has 154 bases. Among these, 5946 are pico-inversions (i.e. ≤40 bases). The frequencies of the lengths of these pico-inversions follow an exponential distribution $f(x) = 161/2 \times 0.861^x$ with the goodness of fit $R^2 = 0.96$. We will show by the simulations (in Section 4) that the accuracy of the prediction of pico-inversions is related to the length of pico-inversions. In general, shorter predicted pico-inversions have higher false positive rates. Therefore, we summarize the counts and properties of pico-inversions considering different minimum lengths in Table 4.

### 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 $f(x)$</td>
<td>$161/0.86^x$</td>
<td>$1374/0.87^x$</td>
<td>$1295/0.87^x$</td>
</tr>
<tr>
<td>excluding repeats</td>
<td>0.96</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>Total count excluding repeats</td>
<td>4351</td>
<td>1618</td>
<td>600</td>
</tr>
<tr>
<td>Frequency distribution $f(x)$</td>
<td>$1402/0.84^x$</td>
<td>$1126/0.85^x$</td>
<td>$840/0.86^x$</td>
</tr>
<tr>
<td>excluding repeats</td>
<td>0.94</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>Association with IRs</td>
<td>36.6%</td>
<td>38.4%</td>
<td>39.3%</td>
</tr>
<tr>
<td>excluding repeats</td>
<td>43.4%</td>
<td>36.8%</td>
<td>33.6%</td>
</tr>
<tr>
<td>Association with genes</td>
<td>33.8%</td>
<td>33.8%</td>
<td>33.8%</td>
</tr>
<tr>
<td>excluding repeats</td>
<td>6.7%</td>
<td>18.7%</td>
<td>34.0%</td>
</tr>
<tr>
<td>Association with genes</td>
<td>39.6%</td>
<td>38.4%</td>
<td>39.3%</td>
</tr>
<tr>
<td>excluding repeats</td>
<td>1.0%</td>
<td>1.1%</td>
<td>1.5%</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.

The first number refers to the minimum length of each sequence of the IR; the second number (in the parentheses) refers to the maximum number of mismatches between the pair of sequences of the IR.
and leads to the perfect match of a pair of inverted repeats of 23 bases. The
at chr7:30,201,616. The reverse complement of an 18-base DNA fragment
between this pair. (a) An inversion in 5’ UTR of SPRY1. The inversion starts
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.
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 \( A_h \) and \( A'_h \) 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 \( A_h < A'_h \)
and \( A_h > A'_c \), we conclude that the inversion occurred in the human lineage. If \( A_h > A'_h \) and \( A_h = A'_c \), 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.
Table 6. Verification of pico-inversions (≥5 and ≤40 bases) using an out-
group alignment.

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

‘Repetitive’ refers to inversions that are homopolymers or simple repeats. The columns
of ‘human’ and ‘chimpanzee’ record the number 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.

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
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>Length distribution function</th>
<th>Inversions</th>
<th>Length distribution function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real data</td>
<td>Alignment</td>
<td>0.986</td>
<td>0.0122</td>
<td>0.000397</td>
<td>0.0055</td>
<td>( f(x) = 1 - x^{-1.23} )</td>
<td>2.44 ( \times 10^{-6} )</td>
<td>( f(x) = 1 - 0.861^{-5} )</td>
</tr>
<tr>
<td>Simulation</td>
<td>Sequences</td>
<td>0.987*</td>
<td>0.0122*</td>
<td>0.000398*</td>
<td>0.0055*</td>
<td>( f(x) = 1 - x^{-1.154} )</td>
<td>2.45 ( \times 10^{-6} )</td>
<td>( f(x) = 1 - 0.888^{-5} )</td>
</tr>
<tr>
<td>Simulation</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>
<td>–</td>
<td>–</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 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.

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 inversions). We use the shortest human chromosome, #21, as the starting sequence (so that repeats and duplications already exist) to simulate substitution blocks, indels and pico-inversions. The models of substitution blocks and indels are obtained from the whole genome alignment between human and chimpanzee. The model of pico-inversions is obtained from the pico-inversions discovered by PicoInversionMiner based on 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, \( \hat{p}_{gap} \) for indels and \( \hat{p}_{inv} \) for pico-inversions (Table 2), where \( p_i (i > 0) \sim N\left( \frac{p_{pico}}{1-p_{pico}}, \frac{p_{pico}(1-p_{pico})}{n} \right) \) (where \( n \) is the genome size). For each simulation, we sample \( p_i \)'s (\( i > 0 \)), \( \hat{p}_{gap} \) and \( \hat{p}_{inv} \) from these normal distributions.

Simulation of various divergences. \( \epsilon \) controls the divergence of simulated sequences. See details in the text. Only inversions no shorter than 10 bases are considered in this evaluation (the sensitivity considering all inversions is significantly lower).

\[ f(x) = 1 - x^{-1.23} \]

\[ f(x) = 1 - x^{-1.154} \]

\[ f(x) = 1 - x^{-1.128} \]

\[ f(x) = 1 - 0.861^{-5} \]

\[ f(x) = 1 - 0.888^{-5} \]
values. In all, 100 pairs of sequences are produced for each the same length distributions of indels and pico-inversions from the is 98.7%; when the divergence is between human and chimpanzee, where the PIP

...algorithm of PicoInversionMiner. The algorithm starts from an initial estimate of the inversion rate, 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 , 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 . The initial estimate of 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 significantly, and there are more iterations, which leads to more false positives. Note that for the simulations, the value of in Figure 4 is calculated based on the initial estimate of and the length of the simulated sequence: \(1.63 \times 10^{-7} \times 48M = 8\).

Through 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 \((>) 1\), \(p_{14}\) and \(p_{54}\) 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 \(c = 7.5\), 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—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 ‘CATGT’, and their alignment only cons 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 \(2P_{\text{inv}} \geq 2\ln(2.44 \times 10^{-6} \times 0.9865^5/(1.57 \times 10^{-6})) = 0.74\), 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 unique 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


