Revised time estimation of the ancestral human chromosome 2 fusion

AUTHORS' RESPONSES TO THE REVIEWERS

Barbara Poszewiecka, Krzysztof Gogolewski, Paweł Stankiewicz and Anna Gambin

The following document contains responses to all remarks, comments, doubts addressed by the reviewers of our manuscript *Revised time estimation of the ancestral human chromosome 2 fusion*. To make it clear and readable for the reader we have provided the original comments given by each reviewer (italic font) and where it was required we have left our clarifications, comments or other forms of response (bold font).

Reviewer 1

The chromosome count of humans is 46, while that of the great apes is 48. This difference is due to a fusion of two chromosomes in human lineage which was previously estimated to have occurred ~0.74 million years ago by Dreszer et al.

In that work an estimation method was proposed that was based on the analysis of the fixed substitutions in the human and chimpanzee genomes since their divergence from the common ancestor. Two observations were exploited: (1) that the substitutions that are densely clustered on the chromosomes show a remarkable excess of AT to GC (biased) substitutions, and (2) that the signal of these Unexpected Biased Clustered Substitutions (UBCS) is strong and stable near telomeres while also peaking at the fusion site of the two ancestral chromosomes.
Dreszer et al. proposed a method to compute the expected number of BCSs, and used it to date the fusion event.

In the paper under review it is stated that the computations in Dreszer et al. were not precise:

"First, we present a novel algorithm for re-calculation of the UBCS statistic proposed by Dreszer et al. [5]. We have corrected their procedure for estimation of the expected number of the so-called clustered substitutions by introducing a inclusion-exclusion principle. <...>
In their approach authors did not discussed the problem of overlapping clusters of substitutions. Here, it means that one substitution may belong to many clusters, which influences the calculation of the expected number of BCS."

Though in the Supplementary Materials of Dreszer et al. one finds:

"However, our definition of clusters allows for a CS to lie in more than one cluster, any one of which can be biased (and hence make the CS a BCS). <...> If a substitution lies in two or more clusters, then the calculation <...> involves a combinatorial computation."

These statements in Dreszer et al. are followed by an example of a computation of an expected number of BCS using the inclusion-exclusion principle.
This computation seems to be equivalent to the one presented in the paper under review, and might substantially diminish its novelty, however it is hard to compare the two methods due to the differences in notations.

We have now unified the notation by introducing symbols p' and $\hat{p}$ consistently with those defined in the Supplementary Materials of Dreszer et al. We have also changed the denotation of the number of substitutions in a bin from s_k to b_k not to be confused with the s_k from the Supplementary Materials of Dreszer et al. denoting the *k*-th substitution in the considered genomic region.

In their example, Dreszer et al. consider a substitution that is contained in two clusters and their method cannot be applicable in the case of substitution contained in 3 or more clustered. We have devised an algorithm for computing the probability that a substitution is BC in such cases.

As stated in the Dreszer et al., the calculation of the probability that a given substitution is Biased Clustered (BC) requires combinatorial computation if it belongs to more than one cluster. However, algorithms that use the inclusion-exclusion principle in a straightforward way may lead to an intractable combinatorial explosion in time and memory consumption during computations. This is due to the fact that they may require generation of all possible assignments of values specifying whether a certain substitution is biased or not to all substitutions in considered clusters.

To address this issue, we propose a completely new algorithm that uses the dynamic programming techniques to apply the formulas for the calculation of the probability that a given substitution is BC derived in the section *"Efficient algorithm for the calculation of the expected number of BCSs"* of the manuscript. This approach reduces the number of combinatorial simulations required for the analyzed data to the extent when it becomes a time and space tractable computation. Actually, to compute the probability that a certain substitution is BC the required memory is proportional to 2^c (where c is the maximum number of substitutions within a cluster among all clusters in which the considered substitution is contained), and the time complexity is proportional to $n \cdot 2^c$ (where n is the total number of substitutions in all clusters that contain the considered substitution).

The passage from Supplementary Materials of Dreszer et al. dealing with the inclusion-exclusion principle should be addressed, and mathematical notation used in the two papers should be unified. The "imprecise formula" from Dreszer et al. should be formally introduced,
and its flaws (if any) should be clearly stated.

Dreszer et al. do not present the explicit formula for computing the probability that a given substitution is BC if it is contained in two overlapping clusters in a general case. Instead, the authors provide one example for computing the probability for a very specific arrangement of substitutions in two overlapping clusters. However, the derived formulas are not correct. Please note that in the Supplementary Materials we not only clearly list inconsistencies in the derivation of the formulas, but also precisely point out errors that were made in the final formula of the considered example.

The differences in the estimations should also be explained.
For example, why exactly is the confidence interval shorter in the new study?
Are there other major differences in these methodologies, and if so, why were they introduced?

Figure 2 in the Supplementary Materials shows that the increasing number of windows considered in the calculation of the UBCS statistics, results in higher values of this statistic near telomeres. This shows that the extent of the biased gene conversion phenomenon of in these regions the have been the actual have been underestimated, which influenced on the estimations. Moreover, we based the estimates on the recent genome build, which are more accurate, especially in the telomeric regions.

Little examples, similar to those presented in Dreszer et al., for which the two methods provide different estimations would also be useful.

This being said, it's hard for me to judge how much novelty there is in this paper,
when compared to the work of Dreszer et al, and I believe that a major revision is required to address this issue.

Thank you for that comment. We agree that we might have not sufficiently highlighted this in our manuscript. However, in our article we have provided a novel method for estimating the time of the speciation events. The method was applied to different species and provided results that are consistent with the present state of the art. To the best of our knowledge, it is an innovative approach and importantly can be easily applied to the approximation of the speciation times of species other than those considered in the article.

What is more, there are lots of typos throughout the paper,
and grammar should also be carefully checked.
To give just a few examples:

"Our result sshed light"
"Within a SD"
"the reduction of the number of chromosomes does not have to lead to fetal genetic dysfunction" -> should it be "fatal"?
"The Authors"
"Out next step that we plan to undertake"

The manuscript was thoroughly proofread and typos and grammar/lexicography errors were improved.

Reviewer 2

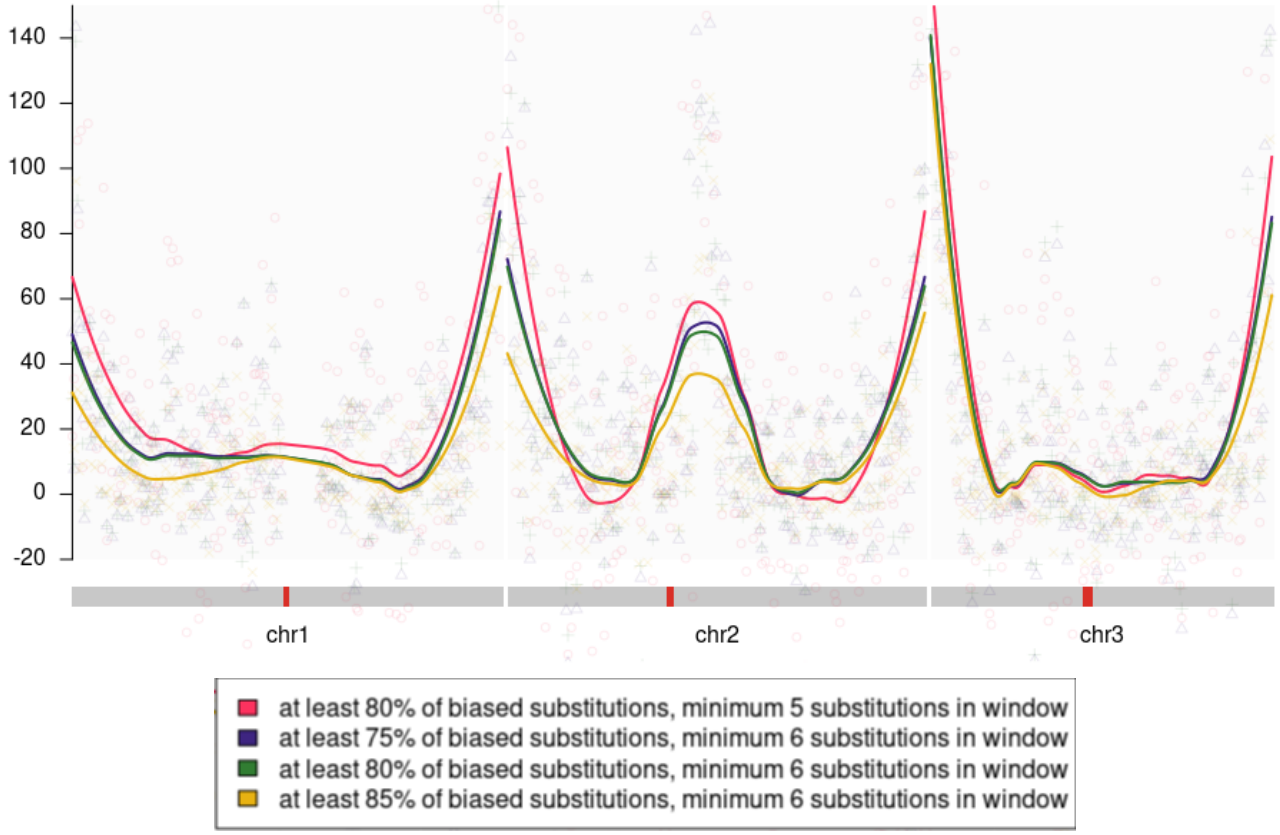
The authors provide a new method for calculation of the UBCS statistics proposed
by Dreszer et al; and applied it to estimate the HSA2 chromosomal fusion time.
While the method is interesting, it is not quite clear how robust it is, and what kind of evidence may support (or disprove) the authors' findings.

One of my concerns about the method is that it is not clear if the method is parameter-free, and if not, how the parameters were chosen and how much the results depend on the particular parameters' values. Namely, the whole UBCS statistics is built around so-called biased clustered substitutions, and the definition of the clustered substitution is that "it belongs to a 300 bp window with at least four other substitutions". I did not find any discussion on why this parameter choice -- 300 bp and 4 substitutions -- is good or optimal in any sense. It is not clear from the provided pseudocode how exactly the results depend on the parameters value. I believe the values we inherited from the original Dreszer's paper, which is OK. But since the main idea of the presented paper is to make Dreszer's method more accurate, the choice of the parameters should be widely discussed.

This question of the reviewer is indeed of high importance, so we want to thank for that.
To assess the robustness of the algorithm with respect to its parameters we have evaluated the data for human and chimpanzee for different values of window sizes (250 and 300), minimal percent of substitutions in window to consider a substitution "biased" (75%, 80%, 85%). and number of substitutions in the window to be considered "clustered" (5 and 6). First of all the trends of the UBCS statistics are conserved for all telomere sites (significant increase of values) as well as around the fusion site of the Chr2. Additionally, we observe a low magnitude standard deviations from the values presented in the main article (see Figure).

These simple tests assure that the method, indeed, is parameter-dependent but the choice of parameter does not influence the final outcome and conclusions.
It is worth mentioning also, that Dreszer et al. have performed their robustness analysis based on their version of the calculations, which provided an insight into the stability of the approach.

Below we present values of UBSC statistics for different choice of parameters (minimal percent of substitutions in the window to consider a substitution "biased" and number of substitutions in the window to be considered "clustered").



The main method should be described in more detail and possibly avoid terms like "tensor product signs". I also have the following major concerns about the the method description:

page 6:
line9: The upper bound for n is equal to the number of substitutions in the region covered by all windows containing substitution at the coordinate j (starting from the coordinate j − m + 1 and ending with the coordinate j + m − 1) minus 4.

Thank you for this remark! The phrase “minus 4” was introduced by accident. We have corrected it in the revised manuscript.

I believe this is not correct, but maybe the definition should be written in a more detailed way. For example, let's consider the case when there are 8 substitutions, the 4th substitution is on the j-th position. If substitution properly located, then there are 7 representative windows: (1:) the window containing all the substitutions (1-8), (2:) the one containing all except the first (2-8); (3:) (3-8), (4:) (4-8), (5:) (1-7) (all, except the last), (6:) (1-6), (7:) 1-5. In this list there are only windows with at least 5 substitutions. It would be great if the authors clarify how they get such an upper bound.

In the section “Method of the construction of a minimal set of representative windows” in the Supplementary Materials, we have provided an explicit method for choosing representative windows together with a descriptive figure presenting the method and the justification of the upper bound for the number of representing windows.

line 15: "a table of size 2 · n − 1". Is is a table or a vector? If this is a table, its size should be in a form N by M. Moreover, it looks like it should be 2m-1 instead of 2n-1. Actually, it looks like the variable n is used in two different senses in Figure 2 caption, which is very misleading.

Thank you for that comment. The word “table” was changed to “vector” as suggested.

page 7:
line 32:
The variable X_k was never defined.

X_k is a random variable specifying the number of biased substitutions in k-th bin. We have added the definition of this variable in the revised manuscript.

Line 35: it is not clear why two events are independent. Do you actually mean conditionally independent?

Yes, these two events are conditionally independent. We have corrected this mistake in the revised manuscript.

Why does the value of x_k start from 0? Should it start from 5 (since otherwise such a window would not be representative)?

According to the definition mentioned above, X_k is a random variable specifying the number of the biased substitutions in the k-th bin. Therefore, it can have values from 0 up to the number of substitutions in the k-th bin (b_k).

I also have some minor comments:
1) abstract:
... result sshed ... -> ... results shed ...

The typo was corrected.

2) page 1: the sentence "300-500 copies of the 40 kb genomic segments within a SD have been identified near the fusion site in the chimpanzee genome but only 4-5 copies are present in the modern human genome" is not clear. Is this correct that there are at least 12-20 Mb of SDs in the chimpanzee genome? Maybe there is a typo here.

Counterintuitively, this is true. Currently, we work on another manuscript concerning the assembly of the telomeres of the Great Apes. The figure below shows the reads from the chimpanzee genome mapped in the region near the fusion site. The depth of coverage in the region on the left side of the fusion is extremely high (7500x) and accounts for at least 300 copies of this region in the chimpanzee genome. Our findings show that only a handful of these reads do not come from subtelomeric regions of the chimpanzee.

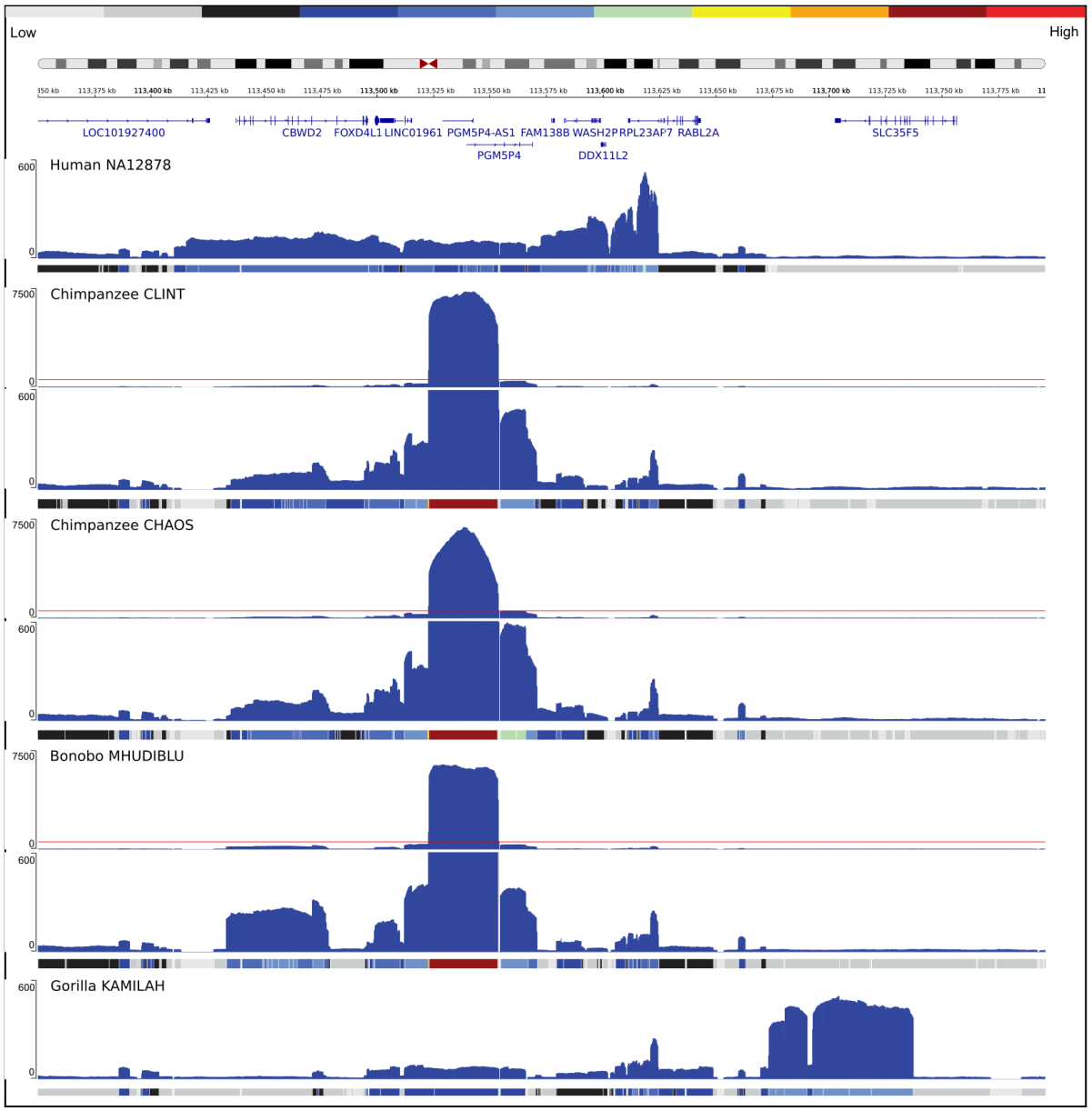


Figure 1. Normalized depth-of-coverage histogram of aligned whole-genome CCS reads of human (NA12787), two chimpanzees (Clint, Chaos), bonobo (Mhudilbu) and gorilla (Kamilah) near the fusion site (chr2:113350000-113800000, NCBI hg38). In the case of the bonobo and two chimpanzees two depth-of-coverage tracks are shown. The Y-axis limit of the top track allows for the presentation of all data. The Y-axis limit of the bottom track allows for the presentation of values apart from the region with extremely high coverage. Red line on the top track marks the Y-axis limit of the bottom track.

3) page 12 (and below):
linage -> lineage
The typo was corrected in all its occurrences.

4) page 13:
Nonetheless, a time point was estimated as 0.67 Mya with 95% confidence interval 1.3 Mya.
It looks like one of the ends of the interval is missed.
The lower bound of the confidence interval was added.