Author’s response to reviews

Title: Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences

Authors:
Gaik Tamazian (mail@gtamazian.com)
Pavel Dobrynin (pdobrynin@gmail.com)
Ksenia Krasheninnikova (krasheninnikova@gmail.com)
Aleksey Komissarov (ad3002@gmail.com)
Klaus-Peter Koepfli (klauspeter.koepfli527@gmail.com)
Stephen O'Brien (lgdchief@gmail.com)

Version: 1 Date: 20 Apr 2016

Author’s response to reviews:

Dear Scott,

My coauthors and I would like to express our gratitude to the reviewers for their thorough comments. Please find below our response describing the changes made to the paper. The excerpts from the reviews are marked with the ‘>’ sign.

Best regards,
Gaik Tamazian

Reviewer #1 (Thomas Keane)

> I have one major concern though - I think the paper requires a much
>more substantial evaluation of the accuracy of the software (ideally to
>include more than just one other tool - several are cited). The only
>numerical measures of accuracy are identity, length, mismatches, and
>coverage. To me, the goal of the assessment should be to measure the
>structural accuracy of the chromosomes produced - from the local
>structure (few kbp range), longer range accuracy (e.g. tens to hundreds
>of kbp), and the overall chromosome organisation accuracy. Then the
>question of where are the mistakes being made, e.g. are the coding
>regions of the genome correct - for species with splicing, are the
>exons in the correct order+orientation? Are structural variations
>differences between the species being assembled and the reference
>genome represented correctly in the chromosomes? Especially
>transposable element differences, which can be difficult since they
>contain repetitive sequence.

Following the reviewer’s recommendation, we introduced two additional criteria for assembly evaluation. The first criterion is based on contig or scaffold adjacency within the assembled chromosomes; for contigs we also check their adjacency within scaffolds. The second criterion is based on gene integrity: for genes that are located on several contigs or scaffolds we check that their contigs or scaffolds are properly arranged in the assembled chromosomes. The described approach was applied to the chimpanzee genome assembly as described in the revised manuscript (subsection ‘Pan troglodytes assembly’ of section ‘Chromosomer assembly evaluation’). Also, we would like to note that the proposed contig-scaffold-chromosome scheme represents three levels of structural accuracy as the reviewer suggested. Contigs-within-scaffolds, contigs-with-chromosomes and scaffolds-within-chromosomes correspond to the local structure (a few to tens kbp range), longer range accuracy (tens to hundreds of kbp) and the overall chromosome organization (typical scaffold lengths).

Regarding structural variations, we would like to note that Chromosomer does not consider mate-pair reads and thus cannot detect structural differences between the species being assembled and the reference genome except the differences that have been already considered in
the fragments being assembled (e.g., scaffolds). To address the issue of transposable elements (TEs) we propose two approaches. The first one is to mask repetitive elements including TEs before the alignment stage to minimize their impact on the assembly process. The second approach is to use net alignments that combine multiple shorter alignments of non-repetitive regions to a single large-scale alignment spanning through non-aligned transposable elements. Both approaches may be combined as we did to assemble the Tibetan antelope and chimpanzee genomes described in the revised manuscript.

I can think of a few ways to create this sort of evaluation 1) take a species with PE reads, and mate pair libraries of various sizes, create scaffolds with all but the largest mate pair library, run chromosomer and other similar tools, and then use the largest library to see how many of those mate-pairs align in the correct orientation and in the expected size range. 2) Create simulated genomes with structural variations at different size ranges, create PE and mate pair libraries, run chromosomes (and other tools), and then look at the representation of the introduced simulated structural variants. 3) In the real data sets (could use the existing presented datasets), it would be very very useful to know the accuracy of the coding regions, e.g. exon order+orientation, as a way to compare the performance of different tools.

Following the reviewer’s recommendation, we chose the third way: we used Chromosomer to assemble the chimpanzee genome from its contigs and scaffolds and compared the obtained chromosomes to the original GenBank chromosomes by the above-mentioned criteria and described in the revised manuscript.

Minor things:

- the quality of the figures with the dotplots is not good. I can't
>read the legends and most of the dot plot is quite blurry.

The dot plots had been submitted in high resolution. We suppose that the publishing system compressed the original dot plot images in the produced PDF file and put external links to the high-quality files. Please check the PDF file for the links located in the top right corner of its pages.

Reviewer #2 (Dennis Larkin)

> The authors correctly indicate that with a large number of genomes being sequenced with NGS technologies it is possible to use assemblies of closely related species to estimate the order of scaffolds in de novo sequenced genomes. On the other hand, such an approach will likely produce errors in the regions where the de novo sequenced genome is different from the reference genome (lineage-specific structural differences).

We agree with the reviewer that the approach implemented in Chromosomer will fail to determine lineage-specific structural differences. However, this is not the goal of Chromosomer: it sacrifices these differences to produce a putative chromosome assembly that is close to the specified reference genome but also maps as many of the original genome fragments as possible.

>The approach that authors chose is to use BLAST to place the target genome scaffolds on reference chromosomes. From the text, however, it is not clear what will happen when a target genome scaffold maps to more than one region of the reference genome. This likely to indicate two possible scenarios: a) the target genome scaffold is chimeric and the observed structural difference is not real or b) the target
scaffold is correct and the structural difference is real. Because Chromosomer is not taking into account read mappings and does not in any way evaluates quality of target scaffolds and only uses one reference genome, it is unclear how exactly the tool deals with the structural differences between the reference and target genomes present inside the target genome scaffolds.

Chromosomer allows for arrangement of genomic scaffolds of one species in the order corresponding to the genomic structure of the reference species. Consequently, as mentioned above, Chromosomer will not reveal structural variations that happened between different scaffolds in the target genome. On the other hand, Chromosomer is sensitive to structural variations that might occur inside of scaffolds. In the case where different parts of the target scaffold can be equally mapped to different sequences of the reference genome, then, as stated in the paper, the segment is considered either unplaced or unlocalized.

Besides, Chromosomer can use net alignment files to place the target genome scaffolds on reference chromosomes.

There are other mammalian genome-size reference assisted assembly tools., e.g. RACA (Kim et al., 2013) that take into the account read mappings and use more than one genome (reference and outgroup) to minimize the effect of chimeric scaffolds on reference assisted assembly, and to keep those structural differences that are well-supported either by read mappings or comparative data. The authors therefore, should compare the outcome of their tool to the outcomes of the reference assisted assembly tools built specifically for large genomes, like RACA.
Following the reviewer’s request, we compared RACA and Chromosomer using the Tibetan antelope dataset from the RACA paper and described the result in the revised manuscript.

On the other hand, Chromosomer transfers the annotation from reference genome to target predicted chromosomes. Again, there are tools that are built to do similar tasks (e.g., RATT). Therefore, the quality of annotation transfer needs to be evaluated and the results presented.

Since Chromosomer does not modify the original scaffolds, their annotated features also do not change. Thus, the quality of transferred annotated features depends only on the quality of the annotated features at the original scaffolds. Unlike RATT, Chromosomer does not utilize syntenic relations between the scaffolds being assembled and a reference genome and therefore, does not adjust the transferred annotation. Considering these differences, we suggest that Chromosomer and RATT are not comparable to each other because they use different approaches to solve distinct problems.