Peer Review File

Manuscript Title: The complete sequence of a human Y chromosome

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

The difficulties with sequencing and assembling Y chromosomes has led to their general omission from genome projects. This omission has fed into the longstanding paradigm that Y chromosomes are degenerate wastelands of the genome, despite extensive evidence from a variety of organisms that Y chromosomes are key to many male phenotypes (Lemos et al. Science 2008), can show high adaptive potential despite the loss of recombination (Sandkam et al. Nature Ecology and Evolution 2021), and even hoover up many autosomal genes through duplications (Tobler et al. PNAS 2017).

The authors present an epic analysis of the human Y chromosome sequence and have by all accounts done an exceedingly thorough job of assembling a full human Y chromosome and describing its many features. If anything, my main suggestion would be that the main text is perhaps too thorough – the very large number of analyses presented may obscure any key take-home messages about the Y chromosome that the authors might intend. If there are any key points that the authors wish to make, I would encourage them to organize the results around them and perhaps move some of the less salient material to the already extensive supplemental materials.

Other than that largely editorial suggestion, my other comments are largely minor:
Line 115 - Many readers, including those expert in genomics, may not be familiar with the term "hydatidiform mole". It would be helpful to define it here.

Line 145 – Clarify whether 30X PacBio Hifi coverage was the average across the genome (which implies 15X for the Y) or for the Y chromosome (which implies 60X coverage for the autosomes). Same with ONT coverage on line 150 and details on line159.

Table 1 – what is an exclusive gene and an exclusive transcript, and how does this differ from the tally of other Y chromosome genes and transcripts?

Line 214 HORs are not defined in the main text – I had to find the definition in the supplemental materials. Perhaps define either here at first usage, or since this is a Fig legend, define as well at first usage in the results main text. More importantly, I'm unsure of how a HOR differs from other repeats. Perhaps explain at first usage in the main text?

Line 314 "We confirmed that the structure of the newly assembled T2T-Y is consistent with expectations based on previous studies" This is somewhat cryptic. Briefly explain what those expectations are.

Line 435 – I think this sentence should read "Additionally, transductions can occur in the soma, …"

Supplementary Fig 18 – the WestEurasia label is somewhat less descriptive than the others – for example EastAsia, SouthAsia. West Eurasia could encompass everything from Iceland to Kazakstan. Is this actually the right label, or is it supposed to be Western Europe?

Referee #2 (Remarks to the Author):

This manuscript presents the first human Y chromosome completely sequenced to the last base pair to complement the other 23 chromosomes recently published by the T2T consortium, so we now have a single, fully sequenced genome. The difficulty in sequencing the Y chromosome due to repeats, amplicons and palindromes has a long history of targeted approaches, but this is the first time that the heterochromatic q arm of the Y chromosome has been sequenced at all apart from closing other smaller gaps such as the TSPY array and the centromere, this is the main improvement. It allows for characterisation for the first time of the repeat expansion process. It is also shown that the PAR sequencing is improved and that mapping of short reads is improved, even though it is unclear to me to which extent this is important for the new, more complex parts which, as shown by the companion paper, can only be studied by other long-read sequencing data sets.

While the technical aspects of the study are certainly impressive and the complete Y a great resource (which has already been exploited in the companion Ballast paper), the new biological insights into e.g. evolution of the Y are fairly limited, and the analyses often very descriptive in the same way as genome papers have used to be. This leads in my view to a lot of data presented without much motivation and also without much new insights derived from the facts. More could be done in the results section to motivate analyses, state what is not known on a certain topic, and what we add to knowledge and understanding with this new Y. Added to this, there are a few statements in the midst of facts which I think are very speculative, and more care could be made in marking more uncertain ideas/claims as being just that (see some specific examples below, but more generally: which of the differences between this Y and the hg38 are mistakes in hg38 and which are just polymorphisms?, are all these TSPY copies missing in hg38-Y or are there more in the HG002?

The technical work is generally of very high quality, and the analyses are generally well-described and documented. I would definitely not hesitate to shift to using this Y chromosome over hg38 Y immediately, and this might be the strongest recommendation I can give to the study.

I have the following general comments

1. The paper as written is very long without a clear structure which makes it hard to spot the new insights from small improvements. After all, the hg38-Y was well sequenced for most parts except heterochromatin, so many of the improvements are perhaps not so striking quantitatively and I think it is a very interesting study even without emphasising potential errors in hg38-Y or spending a lot of the manuscript on some more rarely studied phenomenon such as non B DNA, human contamination etc. They might be a general interest if the authors could motivate why they study them better and quantify more what has been learned by doing so. For example, it is not enough to state l 461 that "Non-B DNA structures are known to affect a variety of important cellular processes, such as replication, gene expression, and genome stability" as the only motivation. Otherwise, I think most of the sections on non-B DNA, mappability/variant calling, transducer repeats and human contamination can be transferred for most parts to the SI, which would only increase the readership of the main paper (and the subset of people interested in these can study the SI). I am not convinced that the improved mapping over hg38-Y demonstrated is particularly important, most of the new sequence added is heterochromatic and not really useful for mapping short reads anyway, most of the improvements are important for long read mapping as clearly showed in the companion Hallast paper
2. The Figures are well prepared and laid out, but they contain a lot of information that is often not even mentioned in the main text, and it is up to the reader to make interpretations of the apparent patterns but without explanation and statistical tests in many cases this is difficult. For example, in Figure 1, panel A is the basic description we would expect, but panel B contains hard-to-evaluate

information (like which methylation profile should I trust and what is its level of certainty, what can I learn form the sequence identity among the centromere repeats, which differences between the two Ys are certain to be due to real difference and to previous assembly artefacts, is panel c important up and above that the palindromes are differently organised in these two haplogroups and to which degree is this surprising in light of previous studies, and finally, why do we need panel d at all since this is hardly mentioned and interpreted in the main text. Similar comments could be made of the other Figures, but I suggest a majority of their content could be extended Figures, see comment 1. Above
3. I think some statements on "improvements" etc, could be toned down without making the manuscript less interesting. While assembling heterochromatin is impressive, it is unfair to imply that half the Y was not sequenced before as meaning half the information was missed, after all the comparison between the two Y chromosomes can not even quantify this since they are for different haploigroups and, therefore, very different. This could be clearer even in the abstract
4. Likewise, I see no good reason to emphasise errors in hg38-Y unless they are completely certain, refraining from this would not decrease the value of the study. For example, in several places, errors are implied because the assemblies differ, like for the centromere, the number of TSPY repeats and the organisation of the TSPY families.
5. PAR1. The PAR1 analyses are interesting but not easy to understand, how does the PAR1 from the T2T X chromosome differ from this new PAR1 of the Y chromosome, and how does that affect mapping quality? What is it about the PAR1 sequence that makes it harder to assemble than most of the Y, I.e. what is the source of the sequence bias in PACbio coverage and is nanopore similarly affected
6. The authors know how to denovo assemble complex regions better than anyone else, I believe, and I would therefore appreciate more guidelines to doing this going forward based on this Y chromosome, I.e. how much HIFI and nanopore coverage is needed? And which ratio of each?, the sources of biases in coverage etc.
7. Is there any reason to believe that structural changes could have during cell divisions in the cell line used? Any way to check? Y chromosomes can be lost from somatic cells in vivo, but can they also be rearranged?

Referee #3 (Remarks to the Author):

28-Dec 2022 Nature T2T human Y chr
Rhie and colleagues continue their effort to build a complete human genome assembly by adding a Y chromosome sequence to their previous haploid T2T genome assembly. This is a big step towards producing a biologically realistic genome. Previously, much of the Y chromosome was represented as gaps and the pseudo-autosomal regions were partially duplicated from the homologous X chromosome sequence. Now, the biological copy number of PARs (diploid) and the remainder of the sex chromosomes (haploid) are accurately represented. I have explored and analyzed both the Hg38 and new T2T-Y sequence and can vouch personally for the massive improvement of this resource. The authors conduct some analyses and make interesting observations about repeat structure of previously unknown DNA sequences as well as the methylation landscape therein. Otherwise, the article itself does read a bit as a resource paper (and rightfully so), mostly comparing the T2T-Y to Hg38. This comparison has limited space for inference of biological/evolutionary processes.

Below, I present some musings about how I perceived the article and propose some alternative approaches, methods or contextualization. Most of my general comments should be viewed as ideas for improvement and not required changes. There are some cases (flagged **) that I would like to see least be more explicitly evaluated or contextualized in the text.

General comments:

################
1. Ploidy of T2T
The now-representative PARs produces an issue that is well known (reviewed by Carey et al. 2022, Cell Genomics) and will become more problematic as the human genome assembly becomes more realistic: meitocally homologous sequences are present (similar but not exactly duplicated sequences), which confound short read mapping when unique positions are required. The authors resolve this issue using the previous (and really outdated) approach of simply masking one copy of the PAR (and the whole Y in XX individuals). This seems like an obvious use case to build and map to a graph. The ad hoc approach of masking and mapping to different genomes depending on the sex of an individual seems problematic, especially if sex metadata is not available. **This complexity and potential for downstream analysis issues should be discussed in the text.

################
2. The rest of the HG002 genome
I was a bit surprised to see the HG002 Y chromosome published itself without a full assembly of the genotype from which the sequence arose. With the added Y chromosome, the T2T becomes even less representative of a biological genotype, but also continues the precedent of human assemblies as mosaics among multiple genotypes. Obviously the entire HG002 genome assembly has now been produced, and the existence of this assembly is referenced throughout but never actually discussed in the methods, results etc. **Since the Y sequence was produced with lots of other genome-wide data, the exclusion of the rest of the genome needs to be explicitly discussed in the methods, and not just in passing (e.g. 149, 358, 361, 412, etc). Throughout, the HG002 Y is treated as its own distinct entity (annotation, polishing, etc.), however, this sequence and the reads/sequences mapped to it do not exist in isolation. I can envision many situations where this would introduce biases.

################
3. Variant detection and interpretation esp. in repeat regions
To me, the potential for better variant calling and annotation is probably the most important contribution of a complete Y assembly. However, the variant detection and analysis left quite a lot up in the air. The fact that most variants can be pushed forward to the new assembly is promising, but also makes the reader think that the main point is that, perhaps we really don't need the new assembly. This of course incorrect - we are going to do a MUCH better job of determining the position and functional affect of variants with a more complete reference genome. For example, the method (liftover from existing databases) does not allow for control of the repeat structure when mapping to the new sequence. These should have a major effect on where variants can be detected de novo. Accurate assemblies of non-exact duplicate repeats will let more reads map uniquely and potentially resolve spurious heterozygous calls. Crucially, much of the new sequence will only be typable with reads much longer than illumina. Simulated illumina reads and liftover seemed like a strange choice. Additionally, these sections [480-652] were difficult to read, far more detailed than the rest of the article (roughly 25% of the total results text) and its impact was difficult to understand. To really illustrate the promise of the new Y sequence, I was expecting de novo variant detection using HiFi reads, especially since these are wildly available and this exact situation was highlighted in this groups' earlier work on variant detection in the T2T genome release. Regardless of how this section is re-written, throughout, **the authors must make it clear exactly how much of the Y chromosome is actually covered by uniquely mapping short reads (or lift overs).

################
4**. Polishing
How was the bulk of the Y chromosome polished (lines 159, ex dat fig 2, etc) with the technology specified? I understand the bioinformatic pipeline (ex dat fig 2) and techniques (Merfin etc), but the data fed into it and the structure of the Y chromosome does not seem like it would be conducive to this approach. 33x illumina reads seems like way too low coverage to do any decent polishing, especially in the heterozygous PARs, and too short to uniquely map to the bulk of the Y

chromosome sequence. At least discuss this and how ONT and HiFi can be used as independent lines of evidence since these have high/non-random error rates or were used to generate the scaffold respectively.

###############
5. Comparisons between T2T-Y and Hg38
In general, I had trouble following the logic and value of these analyses. The results seem confounded between biological differences between the genotypes, assembly/annotation methods and completeness. Since the basic sequencing and bioinformatics methods are so wildly different, I think the reader will assume that all sequence and gene differences are simply methodological, which I think is the point [249, etc] (to show that T2T-Y is better), but how this controls for biological differences needs to be more clearly laid out. **However, there are cases where the comparison between T2T-Y and Hg38 is supposed to infer biological processes (e.g. 262, 461, etc) - in these cases, the effects of different sequencing techniques needs to be explicitly mentioned and controlled for. Lastly, please be more precise when reporting statistics between the two assemblies. For example: **[232, etc]: do percentage stats here and elsewhere include the gapped sequence in Hg38?

**Line specific comments:

[171]: What are the mean and error units here? Is this kb or coverage? What is the error specified (SE, CI, etc.)?

[253, 255, etc]: The methods make it seem like you only mapped the refseq/gencode/etc sequences to the HG002 Y chromosome to do the annotation. I imagine this is not the case and the sequences were mapped to an unpublished entire genome. Please specify this. If this isn't the case, and these sequences were just mapped to the Y scaffold, please discuss what biases might appear by taking an entire genome-wide dataset and mapping to a small subset of the genome (e.g. false best hits, etc.). This is especially true for novel Iso-seq reads. Lastly, there must have been some RNA-seq too, right? Isoseq-only gene models can miss quite a bit.

Please discuss how annotations using sequence reads from different alleles [287, etc] than those in the reference sequence can be problematic when forming gene models and counting transcripts, especially in repetitive sequences.

[392-400] The logic here was confusing to me. Phylogenetic analyses can't confirm anything [394], especially without an outgroup. I think you should either do the analysis correctly (as stated [surprisingly] in line 398) or perhaps just drop this section.

Author Rebuttals to Initial Comments:

Reviewer 1

The difficulties with sequencing and assembling Y chromosomes has led to their general omission from genome projects. This omission has fed into the longstanding paradigm that Y chromosomes are degenerate wastelands of the genome, despite extensive evidence from a variety of organisms that Y chromosomes are key to many male phenotypes (Lemos et al. Science 2008), can show high adaptive potential despite the loss of recombination (Sandkam et al. Nature Ecology and Evolution 2021), and even hoover up many autosomal genes through duplications (Tobler et al. PNAS 2017).

The authors present an epic analysis of the human Y chromosome sequence and have by all accounts done an exceedingly thorough job of assembling a full human Y chromosome and describing its many features. If anything, my main suggestion would be that the main text is perhaps too thorough – the very large number of analyses presented may obscure any key take-home messages about the Y chromosome that the authors might intend. If there are any key points that the authors wish to make, I would encourage them to organize the results around them and perhaps move some of the less salient material to the already extensive supplemental materials.

Thank you for your supportive comments! As suggested, we reorganized the main text and figures around our main findings. We have also trimmed the manuscript by moving some non-essential method descriptions to the Online Methods and Supplementary Information.

Other than that largely editorial suggestion, my other comments are largely minor:
Line 115 - Many readers, including those expert in genomics, may not be familiar with the term "hydatidiform mole". It would be helpful to define it here.

We have briefly clarified this in the main text:

> Having been derived from a complete hydatidiform mole, CHM13 has a 46,XX karyotype but is almost entirely homozygous. This simplified assembly of its genome, but prevented assembly of a Y chromosome.

Line 145 – Clarify whether 30X PacBio Hifi coverage was the average across the genome (which implies 15X for the Y) or for the Y chromosome (which implies 60X coverage for the autosomes). Same with ONT coverage on line 150 and details on line159.

We apologize for our previous description of sequencing coverage, which confused multiple reviewers. We had a note in the main text that "all coverage statistics are given relative to the haploid sex chromosomes", so when we noted 30X, we meant 30X on each of the sex chromosomes and 60X genome coverage. To avoid further confusion, we have reverted to the standard of reporting all coverages relative to the haploid genome size and noting the halved coverage of the sex chromosomes where appropriate.

Table 1 – what is an exclusive gene and an exclusive transcript, and how does this differ from the tally of other Y chromosome genes and transcripts?

The exclusive gene and transcript counts referred to features found only in one assembly and not the other. Because we only found exclusive genes and transcripts in the T2T-Y assembly, we have changed this wording to "additional" rather than "exclusive" to avoid confusion.

Line 214 HORs are not defined in the main text – I had to find the definition in the supplemental materials. Perhaps define either here at first usage, or since this is a Fig legend, define as well at first usage in the results main text. More importantly, I'm unsure of how a HOR differs from other repeats. Perhaps explain at first usage in the main text?

The HOR term comes from the satellite repeat literature, and the referenced Altemose *et al.* paper gives a very thorough overview. We now give a brief HOR definition at the beginning of the "Centromere" section, prior to its first reference in Fig. 3:

> Normal human centromeres are enriched for an AT-rich satellite family (~171 base monomer), known as alpha satellite, typically arranged into higher-order repeat (HOR) structures and surrounded by more diverged alpha and other satellite classes. Each HOR copy is nearly identical and comprises a tandemly arrayed set of monomers.

Line 314 "We confirmed that the structure of the newly assembled T2T-Y is consistent with expectations based on previous studies" This is somewhat cryptic. Briefly explain what those expectations are.

The text now includes:

We annotated sequence classes on the T2T-Y as ampliconic, X-degenerate, X-transposed, pseudoautosomal, heterochromatic, and other, in accordance with Skaletsky *et al*. In addition, we were able to classify a more precise annotation for the satellites (including DYZ17 and DYZ19) and the centromere (**Fig. 1** and **Supplementary Table 21**). The X-degenerate and ampliconic regions were estimated to be 8.67 Mb and 10.08 Mb in length, in concordance with previous findings.

Line 435 – I think this sentence should read "Additionally, transductions can occur in the soma, …"

We have updated this sentence as suggested, but based on reviewer feedback we have shortened the SVA section and moved most of the original version to the supplement, including this sentence.

Supplementary Fig 18 – the WestEurasia label is somewhat less descriptive than the others – for example EastAsia, SouthAsia. West Eurasia could encompass everything from Iceland to Kazakstan. Is this actually the right label, or is it supposed to be Western Europe?

The WestEurasia label comes directly from the Simons Genome Diversity Panel nomenclature and is indeed quite broad, covering 75 sub-populations. Nevertheless, we feel that it is important to use the WestEurasia label to be consistent with other studies of these data.

Reviewer 2

This manuscript presents the first human Y chromosome completely sequenced to the last base pair to complement the other 23 chromosomes recently published by the T2T consortium, so we now have a single, fully sequenced genome. The difficulty in sequencing the Y chromosome due to repeats, amplicons and palindromes has a long history of targeted approaches, but this is the first time that the heterochromatic q arm of the Y chromosome has been sequenced at all apart from closing other smaller gaps such as the TSPY array and the centromere, this is the main improvement. It allows for characterisation for the first time of the repeat expansion process. It is also shown that the PAR sequencing is improved and that mapping of short reads is improved, even though it is unclear to me to which extent this is important for the new, more complex parts which, as shown by the companion paper, can only be studied by other long-read sequencing data sets.

We agree that long reads are necessary to study the most repetitive regions; however, T2T-Y does add over 500 kb of newly accessible sequence for short read mapping as well. In addition, the total number of variants called against the T2T-Y reference is typically reduced due to the improved base quality and completeness. In restructuring the paper, we have updated the variant calling section to better illustrate these points using the variant call sets from the 1KGP and SGDP samples.

While the technical aspects of the study are certainly impressive and the complete Y a great resource (which has already been exploited in the companion Ballast paper), the new biological insights into e.g. evolution of the Y are fairly limited, and the analyses often very descriptive in the same way as genome papers have used to be. This leads in my view to a lot of data presented without much motivation and also without much new insights derived from the facts. More could be done in the results section to motivate analyses, state what is not known on a certain topic, and what we add to knowledge and understanding with this new Y.

We agree that the paper is largely descriptive, but this simply reflects the project's goal to assemble and release this important community resource. However, as suggested, we have added further context to the results to better motivate the analyses presented.

Added to this, there are a few statements in the midst of facts which I think are very speculative, and more care could be made in marking more uncertain ideas/claims as being just that (see some specific examples below, but more generally: which of the differences between this Y and the hg38 are mistakes in hg38 and which are just polymorphisms?, are all these TSPY copies missing in hg38-Y or are there more in the HG002?

We removed all statements about potential mis-assemblies in GRCh38-Y, which, as pointed out, can be difficult to assess. Instead, we refocused on the experience of a user and what they would expect to see when switching to this new reference sequence. We think most readers would be interested to know the differences between the GRCh38-Y and T2T-Y assemblies before planning any new analyses.

The technical work is generally of very high quality, and the analyses are generally well-described and documented. I would definitely not hesitate to shift to using this Y chromosome over hg38 Y immediately, and this might be the strongest recommendation I can give to the study.

Thank you for your vote of confidence!

I have the following general comments

1. The paper as written is very long without a clear structure which makes it hard to spot the new insights from small improvements. After all, the hg38-Y was well sequenced for most parts except heterochromatin, so many of the improvements are perhaps not so striking quantitatively and I think it is a very interesting study even without emphasising potential errors in hg38-Y or spending a lot of the manuscript on some more rarely studied phenomenon such as non B DNA, human contamination etc. They might be a general interest if the authors could motivate why they study them better and quantify more what has been learned by doing so. For example, it is not enough to state l 461 that "Non-B DNA structures are known to affect a variety of important cellular processes, such as replication, gene expression, and genome stability" as the only motivation. Otherwise, I think most of the sections on non-B DNA, mappability/variant calling, transducer repeats and human contamination can be transferred for most parts to the SI, which would only increase the readership of the main paper (and the subset of people interested in these can study the SI). I am not convinced that the improved mapping over hg38-Y demonstrated is particularly important, most of the new sequence added is heterochromatic and not really useful for mapping short reads anyway, most of the improvements are important for long read mapping as clearly showed in the companion Hallast paper

As suggested, we have moved several sections to the Supplementary Information and refocused the main text on what we feel are the most important additions to the reference sequence. We have also revised the mappability and variant calling section to better communicate the improvements, even when analyzing short reads (as noted above).

Specifically concerning non-B DNA, we respectfully disagree that this represents a "rarely studied phenomenon". For example, see the recent review by Guliang Wang and Karen Vasquez "Dynamic alternative structures in biology and disease" which includes 331 references (Nature Reviews Genetics, 2023). However, we agree that the previous draft included too much speculation on this topic, which we have moved to the supplement. We have shortened our description of this topic to 3 sentences in the main text, communicating our primary finding that T2T-Y assembles 6-fold more non-B DNA motifs than GRCh38-Y, with particularly strong enrichment observed within the centromere and heterochromatic Yq12. Non-B DNA motifs are known to be difficult to sequence through (Guiblet *et al.* Genome Res 2018, Weissensteiner *et al.* bioRxiv 2022), and we would like to highlight that our T2T assemblies will enable more studies of these unique motifs and structures.

2. The Figures are well prepared and laid out, but they contain a lot of information that is often not even mentioned in the main text, and it is up to the reader to make interpretations of the apparent patterns but without explanation and statistical tests in many cases this is difficult. For

example, in Figure 1, panel A is the basic description we would expect, but panel B contains hard-to-evaluate information (like which methylation profile should I trust and what is its level of certainty, what can I learn form the sequence identity among the centromere repeats, which differences between the two Ys are certain to be due to real difference and to previous assembly artefacts, is panel c important up and above that the palindromes are differently organised in these two haplogroups and to which degree is this surprising in light of previous studies, and finally, why do we need panel d at all since this is hardly mentioned and interpreted in the main text. Similar comments could be made of the other Figures, but I suggest a majority of their content could be extended Figures, see comment 1. Above

We have split the original Fig. 1 into multiple figures so that they better follow the flow of the results text and can be more thoroughly described both in the main text and captions.

The current hg38 reference contains an arrangement of palindromes and amplicons considered "ancestral" (Teitz *et al.,* Am J Hum Genet 2018) and consists of 8 palindromes with near-identical arms. The rearrangements in AZF/ampliconic regions have been long known, including those affecting fertility (Repping *et al.* Am J Hum Genet 2002), as well as the specific large gr/rg inversion representing the main difference between hg38 and T2T-Y (depicted in the original Fig. 1c, now Fig. 4). Our companion paper (Hallast *et al.*) also found a high incidence of recurrent euchromatic inversions especially among the palindromes. However, there is evidence that natural selection might act toward preserving amplicon copy numbers (Teitz *et al.* Am. J. Hum. Genet. 2018). Importantly, our T2T-Y chromosome contains the same copy number of amplicons as the reference genome, and thus should not pose problems for studies relying on mapping coverage in these regions. The specific amplicon organization is associated with the Y chromosome haplogroups. The T2T-Y organization was found in 7/44 samples analyzed by Hallast and colleagues, including two additional haplogroup J samples. While T2T-Y does not represent the most frequent organization of the human Y chromosome, the same amplicon copy number as the current reference Y chromosome renders it a suitable replacement.

3. I think some statements on "improvements" etc, could be toned down without making the manuscript less interesting. While assembling heterochromatin is impressive, it is unfair to imply that half the Y was not sequenced before as meaning half the information was missed, after all the comparison between the two Y chromosomes can not even quantify this since they are for different haploigroups and, therefore, very different. This could be clearer even in the abstract

We agree that many of the differences observed between the two Y sequences can be due to differences between their source haplogroups, and have toned down statements inferring potential mis-assemblies in GRCh38. However, our statement in the abstract that "more than

half of the Y chromosome is missing from the GRCh38" is simply based on the size of the N-base gap in the GRCh38 reference sequence. Fig. 1 illustrates that both this gap, and the assembled q-arm heterochromatic block in T2T-Y, constitutes over half of the chromosome length. As for sequence quality, the variant calling results presented in Fig. 6 demonstrate the overall higher base quality and completeness of T2T-Y, so we believe some comparison between the two assemblies and their quality is necessary and fair.

4. Likewise, I see no good reason to emphasise errors in hg38-Y unless they are completely certain, refraining from this would not decrease the value of the study. For example, in several places, errors are implied because the assemblies differ, like for the centromere, the number of TSPY repeats and the organisation of the TSPY families.

As suggested, we have limited our discussion of these differences to the places we are certain, which is indeed the case for the two regions noted (the centromere and the TSPY array).

The Y centromere in GRCh38 is a model sequence and not an actual sequence assembly (Schneider *et al.* Genome Res 2017). When analyzing the orientation of the repeats in the model sequence, we noted they were in the reverse orientation of all other Y's, indicating an orientation error when placing the model. Later, an updated linear assembly of the Y centromere sequence from the same RP11 clones was completed, but not incorporated into the reference (Jain *et al.* Nat Biotech 2018). However, this sequence did allow us to compare the actual sequence differences between the J1 and R1b haplogroups.

The TSPY array was originally collapsed in hg16 (released 2003) and then flagged as a known assembly issue since hg17 (released 2004). The GRCh38 array contains a ~40 kb N-base gap inside the TSPY array and is ~300 kb shorter than the 700 kb estimate from Skaletsky *et al.* (Nature 2003). Figures from Skaletsky *et al.* show no gap in the TSPY array, but the contigs from that paper as well as assemblies prior to hg16 are not accessible. However, this region was flagged as "issue HG-1532" in all later assemblies, and only recently (2022-02-18), a "patch" sequence was made available in GRCh38.p14 that claims to have corrected the issue (https://www.ncbi.nlm.nih.gov/grc/human/issues/HG-1532). So, it seems fair to note that this region was mis- or partially assembled in the GRCh37 and GRCh38 releases.

We have updated the description of these regions to better explain their history in past references and comparison to the T2T-Y assembly.

5. PAR1. The PAR1 analyses are interesting but not easy to understand, how does the PAR1 from the T2T X chromosome differ from this new PAR1 of the Y chromosome, and how does that affect

mapping quality? What is it about the PAR1 sequence that makes it harder to assemble than most of the Y, I.e. what is the source of the sequence bias in PACbio coverage and is nanopore similarly affected

Mapping quality becomes zero if a read can align to multiple locations with equal mapping scores. Extended Data Fig. 9 shows that the mapping quality for both XX and XY samples are all near zero in PAR1 and PAR2. This is not due to any particular similarities or differences in the T2T X and Y assemblies, but due to the fact that human PAR sequences are too similar to be correctly assigned or phased using short reads alone.

Compared to the rest of ChrY, PAR1 is enriched for GA-microsatellite repeats, which are a known issue for PacBio HiFi sequencing and result in reduced depth of coverage (Nurk *et al.* Genome Res 2020). The ONT reads had better coverage across these regions and were used for patching the HiFi coverage gaps. The exact cause of this particular PacBio sequencing bias is unknown to us (and not fully understood by PacBio either), but non-B DNA structures potentially forming at these repeats have been previously shown to affect polymerization efficiency (Guiblet *et al.* Genome Res 2018).

6. The authors know how to denovo assemble complex regions better than anyone else, I believe, and I would therefore appreciate more guidelines to doing this going forward based on this Y chromosome, I.e. how much HIFI and nanopore coverage is needed? And which ratio of each?, the sources of biases in coverage etc.

The technology landscape is moving incredibly fast, and so the answers to these questions are always changing. As a rough estimate, we generally suggest around 50X coverage of both HiFi and ultra-long ONT for new T2T projects. The particular sequencing biases observed in this study have been previously described (Nurk *et al.* Genome Res 2020, Nurk *et al.* Science 2022, Mc Cartney *et al.* Nat Methods 2022), and some of these biases have now been completely resolved by updated sequencing chemistries. Because of this, we chose to focus this paper primarily on the sequence of the Y rather than the technological aspects.

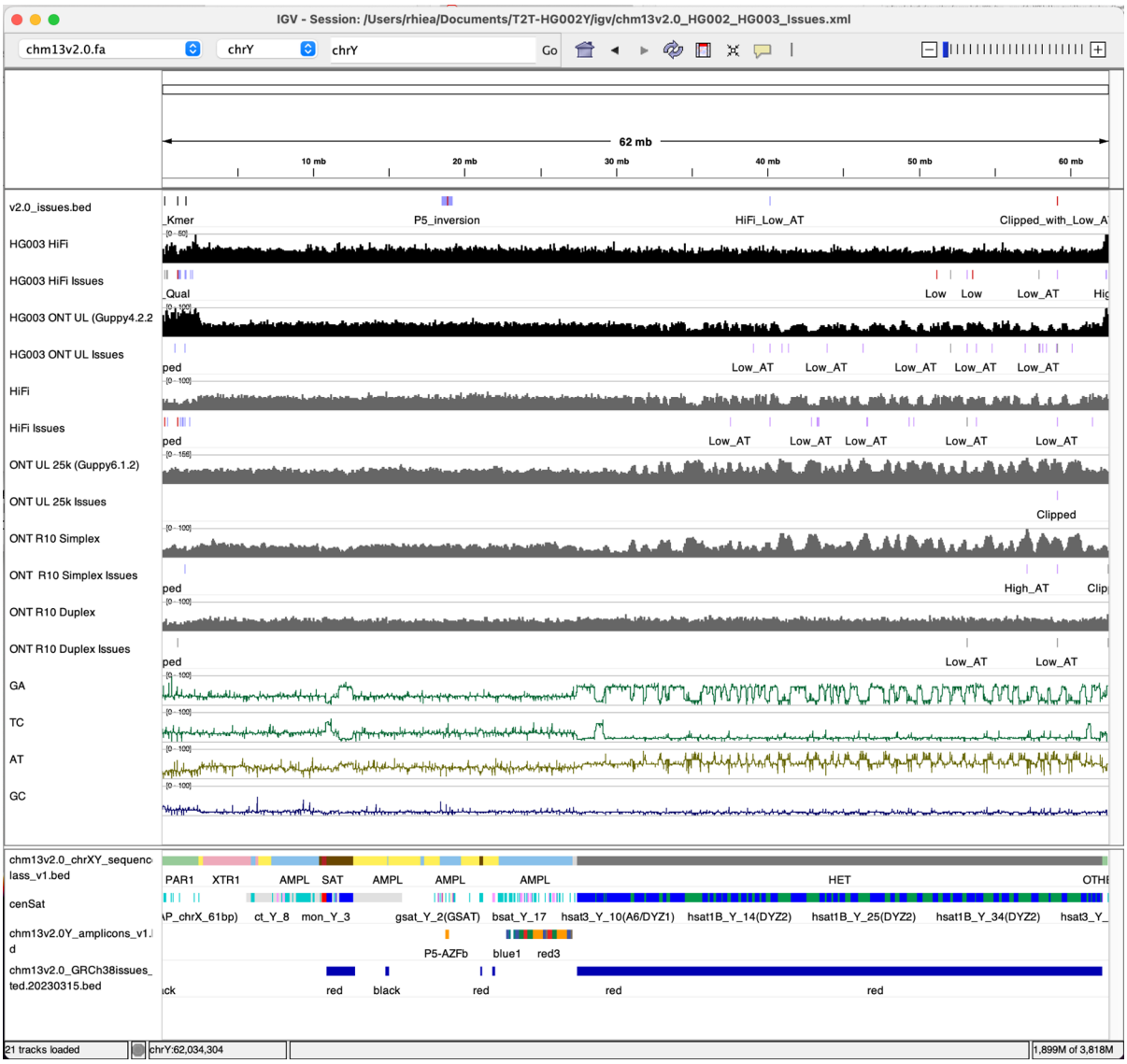
We have taken many of the lessons learned from this project and integrated them into our latest automated assembly process, Verkko. The Verkko paper (Rautiainen *et al.* Nat Biotech 2023) provides more detailed guidance on what is needed for T2T assembly, and is much more replicable than the largely manual methods used here. This is noted in the Discussion and well-demonstrated by the companion Hallast *et al.* study.

7. Is there any reason to believe that structural changes could have during cell divisions in the cell line used? Any way to check? Y chromosomes can be lost from somatic cells in vivo, but can they also be rearranged?

Regarding the stability of the HG002 cell line, which has been sequenced by the Genome in a Bottle consortium using nearly every technology available, we do not see any structural differences when comparing sequencing data from different lots or passages.

To check the integrity of the HG002 cell line versus its original source, we compared against the HG003 cell line, which was derived from the father of the HG002 donor. If the structure of the Y chromosome in HG002 and HG003 is in agreement, then it is very likely that no structural variants arose during cell culture (otherwise, the exact same variants would have to independently arise in both lines). To test this, we collected all available long read data from both HG002 and HG003, which included more recent HiFi (Revio) and ONT (R10) technologies, and mapped it to T2T-CHM13+Y (for HG003 data) and T2T-CHM13 autosomes + HG002XY (for HG002 data).

We observe no evidence for large structural variation between the T2T-Y assembly and any of the additional HG002 or HG003 sequencing data (see the figure below that shows the mapping coverage and issue tracks for various read sets). Also, the HSat1/HSat3 coverage bias observed in our early sequencing data disappears when using the latest chemistries (e.g. see the uniformity of HG003 HiFi and HG002 R10 duplex coverage tracks). The newer HG003 HiFi data shows even better agreement with the HG002 assembly than the older HG002 HiFi data, and the number of sites flagged as potential assembly issues is fewer when using the newer data. Thus, given the agreement of the HG003 sequencing data mapped to the T2T-HG002-Y assembly, we conclude that the T2T-Y assembly is a faithful reconstruction of a human Y chromosome.



Reviewer 3

28-Dec 2022 Nature T2T human Y chr

Rhie and colleagues continue their effort to build a complete human genome assembly by adding a Y chromosome sequence to their previous haploid T2T genome assembly. This is a big step towards producing a biologically realistic genome. Previously, much of the Y chromosome was represented as gaps and the pseudo-autosomal regions were partially duplicated from the homologous X chromosome sequence. Now, the biological copy number of PARs (diploid) and the remainder of the sex chromosomes (haploid) are accurately represented. I have explored and

analyzed both the Hg38 and new T2T-Y sequence and can vouch personally for the massive improvement of this resource.

We are happy to hear this!

The authors conduct some analyses and make interesting observations about repeat structure of previously unknown DNA sequences as well as the methylation landscape therein. Otherwise, the article itself does read a bit as a resource paper (and rightfully so), mostly comparing the T2T-Y to Hg38. This comparison has limited space for inference of biological/evolutionary processes.

Thank you for this feedback. We intended for this to be primarily a resource paper, to be complemented by the companion paper of Hallast *et al.* that included a broader set of Y chromosomes for comparison. To make the paper more readable, we have restructured the main figures and results to better highlight what's new in this complete Y reference.

Below, I present some musings about how I perceived the article and propose some alternative approaches, methods or contextualization. Most of my general comments should be viewed as ideas for improvement and not required changes. There are some cases (flagged **) that I would like to see least be more explicitly evaluated or contextualized in the text.

General comments:

1. Ploidy of T2T

The now-representative PARs produces an issue that is well known (reviewed by Carey et al. 2022, Cell Genomics) and will become more problematic as the human genome assembly becomes more realistic: meitocally homologous sequences are present (similar but not exactly duplicated sequences), which confound short read mapping when unique positions are required. The authors resolve this issue using the previous (and really outdated) approach of simply masking one copy of the PAR (and the whole Y in XX individuals). This seems like an obvious use case to build and map to a graph. The ad hoc approach of masking and mapping to different genomes depending on the sex of an individual seems problematic, especially if sex metadata is not available. **This complexity and potential for downstream analysis issues should be discussed in the text.

We agree that mapping to a pangenome graph is a great idea for dealing with the PARs, and other regions of the genome with high interchromosomal similarity (e.g. the short arms of the acrocentric chromosomes), but this is outside the scope of our study. While there has been some recent progress on variant calling against a graph-based reference, these studies side-step the specific issue of the PARs and masking remains the most popular approach (e.g. the Carey *et al.*

review noted above suggests masking as the most practical solution). For users, we provide both masked and unmasked versions of the assembly. As for sex metadata, it is possible to detect the presence of a Y chromosome from the raw sequencing data, so choice of the correct reference could be automated within a sex-aware variant calling pipeline.

We have added our recommendations therefore to the Discussion:

> We recommend the use of T2T-CHM13 when mapping reads from XX samples and ChrY-PAR-masked T2T-CHM13+Y when mapping XY samples (Supplementary Note 3).

2. The rest of the HG002 genome

I was a bit surprised to see the HG002 Y chromosome published itself without a full assembly of the genotype from which the sequence arose. With the added Y chromosome, the T2T becomes even less representative of a biological genotype, but also continues the precedent of human assemblies as mosaics among multiple genotypes. Obviously the entire HG002 genome assembly has now been produced, and the existence of this assembly is referenced throughout but never actually discussed in the methods, results etc. **Since the Y sequence was produced with lots of other genome-wide data, the exclusion of the rest of the genome needs to be explicitly discussed in the methods, and not just in passing (e.g. 149, 358, 361, 412, etc). Throughout, the HG002 Y is treated as its own distinct entity (annotation, polishing, etc.), however, this sequence and the reads/sequences mapped to it do not exist in isolation. I can envision many situations where this would introduce biases.

The length and quality of HiFi and ONT reads make it possible to confidently determine which reads originate from the X and Y chromosomes. As shown in Extended Data Fig. 1, when the whole-genome HG002 dataset was assembled, these chromosomes formed an assembly graph component that was completely separate from the rest of the genome. Thus, it was possible to focus our finishing efforts on these chromosomes alone, while leaving the HG002 autosomes in a “draft” state. When mapping HG002 whole-genome data for polishing, validation, etc., we combined the complete HG002 sex chromosomes with the complete CHM13 autosomes as a mapping sink for the autosome reads. This approach worked well and prevented any cross-mapping between the autosomes and sex chromosomes.

We have further clarified this approach in the relevant section of the Methods:

> For polishing, the ChrXY draft assemblies were combined with the T2T-CHM13v1.1 autosomes to prevent mapping biases caused by the incompletely resolved autosomes of HG002 (Hereby T2T-CHM13+XY).

3. Variant detection and interpretation esp. in repeat regions

To me, the potential for better variant calling and annotation is probably the most important contribution of a complete Y assembly. However, the variant detection and analysis left quite a lot up in the air. The fact that most variants can be pushed forward to the new assembly is promising, but also makes the reader think that the main point is that, perhaps we really don't need the new assembly. This of course incorrect - we are going to do a MUCH better job of determining the position and functional affect of variants with a more complete reference genome. For example, the method (liftover from existing databases) does not allow for control of the repeat structure when mapping to the new sequence. These should have a major effect on where variants can be detected de novo. Accurate assemblies of non-exact duplicate repeats will let more reads map uniquely and potentially resolve spurious heterozygous calls. Crucially, much of the new sequence will only be typable with reads much longer than illumina. Simulated illumina reads and liftover seemed like a strange choice. Additionally, these sections [480-652] were difficult to read, far more detailed than the rest of the article (roughly 25% of the total results text) and its impact was difficult to understand. To really illustrate the promise of the new Y sequence, I was expecting de novo variant detection using HiFi reads, especially since these are wildly available and this exact situation was highlighted in this groups' earlier work on variant detection in the T2T genome release. Regardless of how this section is re-written, throughout, **the authors must make it clear exactly how much of the Y chromosome is actually covered by uniquely mapping short reads (or lift overs).

To address this specific question, we replicated the short-read mappability analysis performed in the T2T-CHM13 variants paper (Aganezov *et al.* Science 2022) and added the following to the Results:

> Due to genomic repeats, accuracy of short-read variant calling is heterogeneous across the genome. One approach for improving reliability is to restrict analysis to "accessible" regions based on various alignment metrics. To this end, we followed published protocols to generate a short-read accessibility mask for T2T-Y based on patterns of normalized read depth, mapping quality, and base-calling quality. Our masks reveal that while the heterochromatic long arm (Yq) remains largely inaccessible to short-read analysis, T2T-Y still adds 578 kb of accessible sequence compared to GRCh38-Y (increase of 4.2%, **Table 1**).

In addition to this newly mappable sequence, the T2T-Y also results in a reduced number of variant calls across all Y haplogroups (with the exception of R1b), indicating a higher overall base quality compared to GRCh38-Y. These analyses were performed using real short-read data from

the 1KGP and SGDP datasets mapped directly to the T2T-Y without the use of liftover (the liftover datasets are provided simply for those interested in annotating their dataset). We have left the HiFi-based *de novo* variant calling results for the Hallast *et al.* companion paper.

4**. Polishing

How was the bulk of the Y chromosome polished (lines 159, ex dat fig 2, etc) with the technology specified? I understand the bioinformatic pipeline (ex dat fig 2) and techniques (Merfin etc), but the data fed into it and the structure of the Y chromosome does not seem like it would be conducive to this approach. 33x illumina reads seems like way too low coverage to do any decent polishing, especially in the heterozygous PARs, and too short to uniquely map to the bulk of the Y chromosome sequence. At least discuss this and how ONT and HiFi can be used as independent lines of evidence since these have high/non-random error rates or were used to generate the scaffold respectively.

We apologize for our confusing presentation of sequencing depth. Reviewer #1 was also confused by our reporting of coverage relative to the diploid, rather than haploid, genome size. By the standard metric, HG002 was sequenced to 66X coverage. This is sufficient coverage for variant calling, as DeepVariant is tuned for 50X (Poplin *et al.* Nat Biotech 2018).

We have included an extensive description of the polishing and evaluation process in the Supplementary Methods (pages 12–31). It describes how haplotype-specific markers were utilized to reliably evaluate and polish both ChrX and ChrY, and how all data types were integrated (notable, the use of hybrid DeepVariant which combines evidence from both Illumina and HiFi). We largely followed the prior T2T polishing process (Mc Cartney *et al.* Nat. Methods 2022) and noted all modifications to this process in the supplement.

5. Comparisons between T2T-Y and Hg38

In general, I had trouble following the logic and value of these analyses. The results seem confounded between biological differences between the genotypes, assembly/annotation methods and completeness. Since the basic sequencing and bioinformatics methods are so wildly different, I think the reader will assume that all sequence and gene differences are simply methodological, which I think is the point [249, etc] (to show that T2T-Y is better), but how this controls for biological differences needs to be more clearly laid out. **However, there are cases where the comparison between T2T-Y and Hg38 is supposed to infer biological processes (e.g. 262, 461, etc) - in these cases, the effects of different sequencing techniques needs to be explicitly mentioned and controlled for.

We agree that true biological differences between GRCh38-Y and T2T-Y can easily confound summary statistics. Therefore, we revised the manuscript to remove any statements that suggest mis-assembly in GRCh38 unless there is compelling evidence that the observed differences are not biological in nature. See also our response to Reviewer #2 regarding this point and some known GRCh38 assembly issues.

We have left the summary statistics between assembly size, gene counts, repeats, etc. in Table 1, because we feel it is important for users to know the differences between these two reference assemblies, regardless of the source of those differences.

Regarding the specific cases noted, we revised the statement on annotation differences (262) to read:

> Only seven genes differed in their annotation between GRCh38-Y and T2T-Y, due to presumed Y haplogroup differences (**Supplementary Table 10**).

We found annotation errors in one of the TSPYs and two RBMYs, thereby corrected the number of genes accordingly. The number of gene differences increased from six to seven.

Most of the Non-B DNA section (461) has been moved to the supplementary information and it is now noted that these differences are due to sequence missing from GRCh38-Y:

> In the T2T-Y, we identified a total of 825,526 repetitive sequence motifs capable of forming alternative DNA structures (non-B DNA), compared to only 138,640 in GRCh38-Y (**Supplementary Table 17**, **Supplementary Note 2**). This nearly 6-fold increase is largely attributed to our use of novel and improved experimental and computational methodology, as non-B DNA motifs, which might form structures during sequencing, are notoriously difficult to sequence through. We found a particular enrichment of these motifs at the newly sequenced centromeric region and heterochromatic region on the Yq arm (**Fig. 1**).

Lastly, please be more precise when reporting statistics between the two assemblies. For example: **[232, etc]: do percentage stats here and elsewhere include the gapped sequence in Hg38?

The initial statistics included the N-bases of the GRCh38-Y (i.e the gaps), which could be misleading. We clarified the noted sentence to highlight that the percentage of characterized

repeats has been increased due to the newly added sequence and have updated the main text and Table 1 accordingly:

> The newly added sequences increased the percentage of identifiable repeats on the Y chromosome from 66.3% to 84.9%, or 17.5 Mb of non-N bases in GRCh38-Y compared to 53 Mb of bases in T2T-Y (**Table 1**, **Supplementary Table 12-13**).

**Line specific comments:

[171]: What are the mean and error units here? Is this kb or coverage? What is the error specified (SE, CI, etc.)?

The mean is the average sequencing coverage and the error unit is the standard deviation. We added units to remove confusion.

[253, 255, etc]: The methods make it seem like you only mapped the refseq/gencode/etc sequences to the HG002 Y chromosome to do the annotation. I imagine this is not the case and the sequences were mapped to an unpublished entire genome. Please specify this. If this isn't the case, and these sequences were just mapped to the Y scaffold, please discuss what biases might appear by taking an entire genome-wide dataset and mapping to a small subset of the genome (e.g. false best hits, etc.). This is especially true for novel Iso-seq reads. Lastly, there must have been some RNA-seq too, right? Isoseq-only gene models can miss quite a bit.

"T2T-CHM13" was missing in the annotation section. We revised the manuscript to properly identify it as "T2T-CHM13+Y", which includes all CHM13 autosomes, CHM13 ChrX, and HG002 ChrY. This set of sequences is versioned in GenBank as "T2T-CHM13v2.0", and we performed all annotations on this assembly (NCBI GenBank accession GCA_009914755.4).

Iso-Seq reads were available and used for both the CAT+Liftoff annotation and RefSeq annotations. We did not generate additional RNA-Seq data for this particular study, but the RefSeq annotation included all available RNA-Seq data from human genomes, including some previously generated HG002 datasets. As described in the "*De novo* RefSeqv110 and GENCODEv38 annotation" section of the Online Methods (with more details in Supplementary Methods):

> A *de novo* RefSeq annotation was performed on both GRCh38 and T2T-CHM13+Y and released (v110) as previously described for other vertebrate genomes. A total of 82,862

> curated RefSeq transcripts, 345,700 cDNAs, 8.65 million ESTs, 9.7 billion RNA-Seq reads, and 83 million PacBio IsoSeq and Oxford Nanopore reads from over 30 distinct tissues were retrieved from SRA and tentatively aligned to the assembly using Splign or minimap2.

The CAT+Liftoff annotation used all GENCODE v38 transcripts as well as the *de novo* assembled Iso-Seq transcripts.

Please discuss how annotations using sequence reads from different alleles [287, etc] than those in the reference sequence can be problematic when forming gene models and counting transcripts, especially in repetitive sequences.

Accurate annotation of repetitive genes is of course a challenge. In fact, during revision, Hallast *et al.* notified us of three protein-coding genes that appeared to be incorrectly labeled due to the high similarity between each copy. Specifically, after a thorough comparison of the exon sequences, RBMY1B and RBMY1D were corrected as two copies of RBMY1A1. In addition, a premature stop codon was found in one TSPY copy (3rd to the last protein-coding copy in the array), which Hallast *et al.* confirmed in other Y haplogroups and suggested should be labeled as a pseudogene (now TSPY10P). Unfortunately, there is no standard for how to refer to ampliconic genes that differ from the copies that happen to be represented in GRCh38. We have added a brief note to the Discussion:

> The completion of ampliconic and otherwise highly repetitive regions of ChrY will also require updates to existing gene annotations that are based on the incomplete GRCh38-Y assembly. How to label and refer to genes within variable-size ampliconic arrays, like TSPY, is an open question. Moreover, the highly repetitive sequences pose new challenges to computational tools developed on GRCh38.

However, we are confident for the rest of the ampliconic gene annotation here because of the highly conserved nature of the genes. We are making no claims of transcript abundance or expression of individual units in this paper. This would require a separate study, with testis-specific expression data paired with sample-specific assemblies of the individual arrays.

[392-400] The logic here was confusing to me. Phylogenetic analyses can't confirm anything [394], especially without an outgroup. I think you should either do the analysis correctly (as stated [surprisingly] in line 398) or perhaps just drop this section.

As the reviewer suggested, we redid the phylogenetic analysis using two outgroups: *Hylobates moloch* (NCBI RefSeq accession NW_022611649.1) and *Pongo abelii* (NCBI GenBank accession KP141780.1). The results are nearly identical. We updated the main text accordingly and moved the phylogenetic tree figure to the Extended Data due to space constraints. The sentence in question now reads:

> Phylogenetic analysis using protein-coding TSPYs from a Sumatran orangutan (Pongo abelii) and a Silvery gibbon (Hylobates moloch) as outgroups confirmed that all protein-coding TSPY copies (including TSPY2) originated from the same branch, which is separated from the majority (all but one) of TSPY pseudogenes (**Extended Data Fig. 6**).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have dealt with all my suggestions, and I have no further comments. This is an excellent analysis of a complete Y chromosome.

Referee #2 (Remarks to the Author):

In their revision the authors simplify the interpretation of their very impressive sequencing of a complete Y chromosome and the result in my opinion is a much easier to read paper. All reviewer comments by me and the other reviewers have been addressed in the rebuttal as well as for most part in the main text. It is acknowledged that this is more of a resource paper than a novel findings paper (except for the heterochromatin characterisation) and together with the companion I find this set of papers very important for future understanding of the human Y chromosome. I still find the short read mapping analysis and human contamination analysis less scientifically interesting but with their shortening and further quantification I agree that they constitute good arguments for adopting this Y chromosome (together with the Y chromosomes of the companion) as a much better reference than the Y of HG38. Thus, I have no further requests for analyses but think that the authors could run over their main display items again and make sure to simplify as much as possible without losing details discussed directly in the text. This could improve the readability of the paper further

Referee #3 (Remarks to the Author):

Rhie et al have made substantial improvements to their manuscript and have addressed all of my previous comments. I have no further suggestions or comments.