Evolutionary analysis of chromosome end extension

Haojing Shao[1], Chenxi Zhou[1], Minh Duc Cao[1], and Lachlan J.M. Coin[1]

[1]**Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD 4072 Australia**

Corresponding author:
Lachlan J.M.Coin[1]

Email address: l.coin@imb.uq.edu.au

ABSTRACT

There are substantial subtelomeric interstitial telomeric sequence (ITS) in the human genome, however the origin of these sequences is not well understood. We investigate the possibility that these ITS have arisen via a process of chromosome end extension to the telomere sequence. By analysing the relationship between subtelomeric duplication and ITS, we identify multiple ITS which were ancestral chromosome telomeric capping sequence. Comparison of chromosome terminal sequence between 15 species reveals an ongoing evolutionary process of chromosome extension, with an average extension rate of 0.0020 bp per year per chromosome. Analysis of SNP data from 1000 genomes demonstrates reduced SNP diversity in subtelomeric regions, indicating that many terminal regions are younger than the remaining autosomal sequence.

INTRODUCTION

Chromosome ends contain telomere sequences and subtelomeric regions. Most human chromosome subtelomeric regions are duplications of other chromosome subtelomeric regions arranged in different combinations, referred to as subtelomeric duplications (STD). STD are highly divergent between species or even different populations of the same species (Mefford and Trask, 2002; Linardopoulou et al., 2005) and have experienced rapid adaptive selection (Mefford and Trask, 2002). The majority of subtelomeric duplications have the same orientation towards the chromosome end (Mefford and Trask, 2002; Linardopoulou et al., 2005). Based on this it has been suggested that they originated from reciprocal translocation of chromosome tips and unbalanced selection (Linardopoulou et al., 2005).

Telomere repeat sequences ($[TAAGGG]_n$) - the capping sequences of chromosome ends - are breakable, acquirable and fusible in the genome. In somatic cells, telomeres are observed to progressively shorten (Blackburn, 2000; Shay and Wright, 2006). If the telomere sequence is lost, the broken chromosome will become unstable (McClintock, 1941; Tanaka et al., 2012; Flint et al., 1994), and multiple types of rearrangements can occur, including chromosome fusion (McClintock, 1941), tips translocation (Sabatier et al., 2005), or direct extension of telomere repeats (Flint et al., 1994). The manual insertion of telomere sequence in the interstitial region results in enhanced chromosome breakages and induces high rates of chromosome rearrangements around the insertion (Kilburn et al., 2001). Interstitial telomeric sequences (ITS) are widespread in the genome (Bolzán and Bianchi, 2006; Lin and Yan, 2008). In subtelomeric regions, they are almost always oriented towards the terminal end of the chromosome, like the STD (Linardopoulou et al., 2005). Human chromosome 2 is a fusion of two ancient chromosomes 2A and 2B which remain distinct in other species, including Chimpanzee(IJdo et al., 1991). At the fusion site of a 2Aq and 2Bp a pair of proper reverse orientated interstitial telomere sequence can be found.

The initial studies of chromosome terminal evolution focused on comparison of the terminal regions of the X and Y chromosomes. Recombination between these regions has been documented(Helena Mangs and Morris, 2007), thus the terminal of sex chromosome are referred to as pseudoautosomal regions(PAR). Prior to the availability of high throughput sequencing, studies focused on variation in PAR gene content between eutherian species. For instance, the human p arm terminal(PAR1) contains 24 genes. Only 2 of these genes have homologous copies in the mouse, but they are located at mouse autosomes. An extension

and attrition model has been proposed for sex chromosome terminal evolution(Graves, 1995). From the whole chromosome arm scale, genome comparison demonstrated the majority of X chromosome p-arm is an extension to original X chromosome(Ross et al., 2005; Helena Mangs and Morris, 2007). The terminal of X chromosome underwent dramatic extension and loss in eutheria. This study investigates a model of chromosome extension in autosomes, using reference genomes from multiple eutherian speciees.

METHODS

Relationship of Interstitial telomere and subtelomeric duplications

Telomere sequences were annotated from GRCh37 repeatmasker database(Smit et al., 1996). We extracted the non-capping telomere sequences inside subtelomeric regions as subtelomeric ITS sequence. We also included in this analysis the ancient subtelomeric region in the chromosome 2 fusion sites(IJdo et al., 1991). We then searched for all subtelomeric duplications (STD) which are either overlapping or adjacent to these subtelomeric ITS using a database of STD(Bailey and Eichler, 2006). We classified the relationship between ITS and STD we divided into 6 types. Type a is a subtelomeric duplication spanning the entire ITS. Type b and type d are duplications overlap with the ITS from the distal and proximal site, respectively (distal and proximal site refers to relationship to the centromere, so that distal site is the edge of the ITS furthest from the centromere). Type c and type e are duplications which are adjacent to (<50 bps) ITS from the distal and proximal sites, respectively. Type f means there is no duplication adjacent to ITS. We counted the number of duplications for each type for each ITS. We also calculated the average number of ITS in each category at figure1. In order to calculate whether observed distribution of these categories are significantly different from what would be expected by chance, we performed 1000 permutations on ITS by randomly sampling the same number of ITS with the same size distribution in the subtelomeric duplication regions.

Chromosome end population genetics analysis

SNP frequencies were extracted from 1000 Genome Project vcf files (v5.20130502). Other mutations are excluded. Chromosome 13, 14, 15, 21, 22 p arm (unknown terminal sequence) and sex chromosomes (different average mutation rate) were excluded from analysis. In order to estimate SNP diversity (unadjusted) , we uniformly divided the subtelomeric duplication regions(definition see above) into 50 windows. Then we uniformly divided sequence 500 kb adjacent to these regions into 50 windows. Considering the difficulty in detecting SNP in duplicated sequences, we downloaded the unmask regions(2.67 Gb in total) from 1000 Genome Project website(Abecasis et al., 2012). We further overlapped these regions with callable divergence region(see below) into merged regions(2.58 Gb). To calculate the adjused SNP diversity, SNP must reside inside these merged regions. If a window contained zero merged region, this window was unable to provide a diversity estimate. Otherwise, the diversity in these merged regions will represent the diversity of this window.

For each window, diversity is calculated as the average of the base pair diversity, i.e. $\bar{h} = \frac{\sum_{j=1}^{L} h_j}{L}$, where L is the total (callable) size of the window. The base pair diversity is defined as $h = 1 - \sum_{i=1}^{n} {f_i}^2$, where f is allele frequency and n is the number of observed alleles. Finally, we summarize the average diversity from all the chromosomes for each bin as $\bar{\bar{h}} = \frac{\sum_{k=1}^{c} \bar{h}_k}{c} = \frac{\sum_{k=1}^{c} \frac{\sum_{j=1}^{L} (1 - \sum_{i=1}^{n} {f_{ijk}}^2)}{L}}{c}$, where c is total number of available chromosomes.

In order to calculate regional divergence between Human and Chimpanzee genomes at chromosome ends, we downloaded the alignments from human(GRCh37) to chimpanzee(panTro4) at UCSU(Kent et al., 2002). In brief, this alignment was unique for each base pair. The divergence regions were defined as the human-aligned regions(2.74 Gb). The nucleotide divergence was calculated as the percentage of substitution between two sequences. Non-SNP mutations were ignored in the estimation.

We performed a local regression on diversity using R(function geom_smooth in ggplot2(Wickham, 2016) package and parameter is method=loess,span=0.7).

Human extension rate analysis

We downloaded pairwise alignment files from UCSC(Kent et al., 2002). These files contain regional alignments from multiple species to Human reference GRCh37. Initially, 21 genomes (Chicken(galGal3), Chimp(panTro4), Cow(bosTau7), Dog(canFam3), Gibbon(nomLeu1), Gorilla(gorGor3), Horse(equCab2), Marmoset(calJac3), Mouse(mm10), Orangutan(ponAbe2), Rat(rn6), Rhesus(rheMac3), Sheep(oviAri3),

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

Baboon(papHam1), Cat(felCat5), Elephant(loxAfr3), Kangaroo(dipOrd1), Panda(ailMel1), Pig(susScr2), Rabbit(oryCun2) and Zebrafish(danRer10)) were analyzed. For each human autosome as well as ancient chromosome 2A and 2B(IJdo et al., 1991), we sorted the alignments by human chromosome and location. We searched for the most terminal end alignment. Because short alignments could result from common repeat elements and subtelomeric duplications, we only selected the alignments longer than human longest subtelomeric duplication (154k) to represent ancient chromosome sequence. Because sequence divergence and genome assembly quality will significantly affect the alignment length. Baboon, Kangaroo, Panda, Pig, Rabbit and Zebrafish genomes were hard to represent large ancient chromosome sequence and removed from the analysis.

We defined the ancient chromosome end for the last common ancestor of species A (typically human) and species B as the most terminal end of homologous sequence between the two species. Unique terminal sequence in species A which starts subsequent to the ancient chromosome end is referred to as chromosome extension sequence which has occurred in species A since the most recent common ancestor of A and B (referred to as MCRA(A,B)) . In estimating the size of this extension sequence, unknown sequence regions (”N” regions) are discarded. Total autosome extension sequence (s) is the sum of all autosome terminal extension sequences in species A since MCRA(A,B). The autosome expansion rate of species A since MCRA(A,B) is estimated as $p = \frac{s}{t}$, which t is the estimated MRCA time. The average human autosome extension rate since the divergence of human and other primates is estimated as $\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k} = \frac{\sum_{i=1}^{k} \frac{s_i}{t_i}}{k}$, which k is the number of primates and $\bar{t}$ is the mean of estimated MRCA time.

We downloaded all available pairwise alignments for eight species (Human(GRCh37), Cow(bosTau7), Dog(canFam3), Horse(equCab2), Mouse(mm10), Rat(rn6), Sheep(oviAri3), Cat(felCat5)) from UCSC(Kent et al., 2002). We also downloaded the unknown sequence annotation files named ”gap.txt.gz” from UCSC to infer the terminal unknown sequence(gap). If a species terminal is annotated as ”telomere” and there is another gap within 154kb to the terminal, this terminal is regarded as uninformative and removed from analysis like human 13p. Mouse(mm10) chromosome 1 to 19 p arms(3Mb gap with telomere and centromere) and chromosome 4 and 9 q arms(too many gaps) are removed from the analysis. For the remaining terminal, we perform a similar analysis as the human. Sex chromosomes are excluded from this analysis(see discussion).

RESULTS

Some subtelomeric interstitial telomeric sequences represent ancient chromosome ends

There are multiple interstitial telomeric sequences (ITS) in the human genome which are orientated in the same direction at subtelomeric regions (Linardopoulou et al., 2005). For example, 6 telomere sequences are in the first 110kb of 18p (permutation $p = 0.039$, see Methods, Figure 2). We investigated the relationship between all chromosome end ITS and subtelomeric duplications (STD) of 1kb or more (Figure 1, Table S1, S2, Methods). All ITS are either fully contained within a duplication or within 50bp of a duplication (Figure 1). The vast majority overlap or are next to a duplication on the distal side of the ITS (15 sites) rather than proximal (3 sites, of which each site also has a distal-side duplication overlap, see Table S1). The STD duplication on the distal side of ITS suggests that duplication events occurred at the end of the ITS.

As a clear example of this, we could identify two subtelomeric ITS(chr8:170440-170577, chr19:59097932-59098077), which occur on the most proximal end of a subtelomeric duplication, and which moreover have no further subtelomeric duplication which is more proximal. In other words, if we were to remove all sequence distal to these ITS, then they would form the terminating telomeric sequence of a chromosome terminal without any subtelomeric homology. We could find no non-primate mammalian sequence homologous to sequence distal to the chr8:170440-170577 ITS, indicating that it may be the ancestral telomere sequence for the last common ancestor of primates.

We propose that some current subtelomeric ITS, including these two specific examples, were the ancient chromosome end telomere sequences. We considered three alternate models for the origin of subtelomeric ITS, but contradictions are found with each. The first is the random model, assuming the distribution ITS was mediated by random duplication or rearrangement the subtelomeric sequence. However, random permutations find equal random sequences with duplications at both proximal and distal breakpoints, which is different from the observed distal-preferential distribution(Figure 1). The second is the reciprocal tips translocation model (Linardopoulou et al., 2005) for subtelomeric duplication.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

However, this model does not involve the chromosome terminus, nor does it create new telomere repeat sequence, so it cannot explain the origin of ITS.

The third model is that subtelomeric ITS are insertion sequences of double-stranded break(DSB) repair(refer to ITS insertion model). This hypothesis is from Nergadze et al(Nergadze et al., 2004), which was originally proposed for (non-subtelomeric) intrachromosomal ITS. Under this model, ITS sequences inserted into a subtelomeric duplication should be detectable as insertion sequences relative to the paralogous copy. However, of the 166 STD pairs with at least one pair containing an ITS (class I, see Methods, Figure 1 Table S3), the vast majority (92%) of STD pairs have an ITS of exactly the same size. 8 duplication pairs(5%) showed a different size of ITS. Only 2 sites within 5(5/166, 3%) duplication pairs show complete deletion of ITS. These two deletion sites(total size 1.2 kb, 4.2 kb) contain many other sequences at both proximal and distal site of ITS(size 185, 192 bp). Thus these two ITS are likely deleted with a larger deletion or trimmed by multiple rounds of subtelomeric rearrangement. Another prediction of the ITS insertion model is that sequence orthologous to sequence containing an ITS with most recent common ancestor prior to ITS insertion event should also be detectable as an ITS insertion. However, in comparison of human ITS to 5 primates, we found no case that proximal sequence and distal sequence were joined without the intervening ITS(see Methods, Table S4). There is no proper ITS deleting example to support the hypothesis of subtelomeric interstitial ITS originating from an insertion of repairing DSB. We also observe an excess of distal-to-ITS only rather than proximal-to-ITS alignment in this primate comparison, supporting the role of end extension in the formation of ITS sequences. Taken together, this evidence suggests that multiple directly duplicated events at the ancient chromosome terminal have played an important role in the formation of present-day ITS.

Mechanism of chromosome end extension

The mean size of subtelomeric ITS (336 bps) is much shorter than capping telomeric sequence. When subtelomeric ITS are used as chromosome end capping telomeres, they create a dysfunctional chromosome end(Sabatier et al., 2005). There are multiple ways to repair the dysfunctional chromosome end (Sabatier et al., 2005; IJdo et al., 1991; Flint et al., 1994), including chromosome end fusion (Figure 3a), telomere extension (Figure 3b), duplication or translocation of another chromosome end (Figure 3c). Relics can be found for all of these events in the human genome(Figure 3d). Chromosome end fusion is found at ancient chromosome 2A and 2B fusion into chromosome 2 (IJdo et al., 1991), which can be seen from a characteristic inverted interstitial telomere sequence. Telomere extension to telomere is indistinguishable from common telomere shortening and lengthening unless the non-telomeric sequence is also involved in the extension. A common observation for ITS or capping telomere is that has TAR1 (telomere associated repeat 1) element inside, and furthermore, the proximal telomere identity is lower than the distal telomere identity(Figure 3d). This suggests that the ancient telomere broke and a new telomere with TAR1 was added. The duplications of other subtelomeric regions to the shortening ITS are the relics of duplication or translocation of other ends to dysfunctional chromosome end. These genome observations are identical to all observations from in-vitro telomere repair models (Sabatier et al., 2005; IJdo et al., 1991; Flint et al., 1994), suggesting that joining sequence to chromosome ends could occur spontaneously as a result of repairing the dysfunctional chromosome ends both in-vitro, as well as in vivo in our ancestors.

Population genetics at chromosome ends

We used the 1000 Genome Project (Abecasis et al., 2012) data to estimate average genetic diversity at chromosome ends (See methods)[Figure 4]. From these data, we found 54% reduction in diversity at subtelomeric duplication regions[Figure 4a,4b]. Because it may be hard to detect SNPs in these regions(which will artificially decrease diversity), we removed the uninformative regions as the 1000 Genome Project suggested (Abecasis et al., 2012) and adjusted the estimation(see methods). The diversity is still 15% lower than the adjacent regions. Next, we investigated the divergence of these regions from chimpanzee and found that the divergence was sharply increased at STD(Figure4c,4d) which was consistent with study(Sequencing and Consortium, 2005). We further found that this increased divergence was mediated by the alignment to paralogous sequence(70%), indicating that chimpanzee missed the homologous sequences(see extension rate section). Combining these observations, the STD regions are special regions in the human genome that have lower diversity and high divergence(Figure 4e,4f). The chromosome end extension hypothesis could fully explain these observations. The new extension sequences at STD didn't have mutations(zero diversity) and took time to accumulate mutation in population(low diversity). And also, these extension sequences didn't exist in other species, such as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

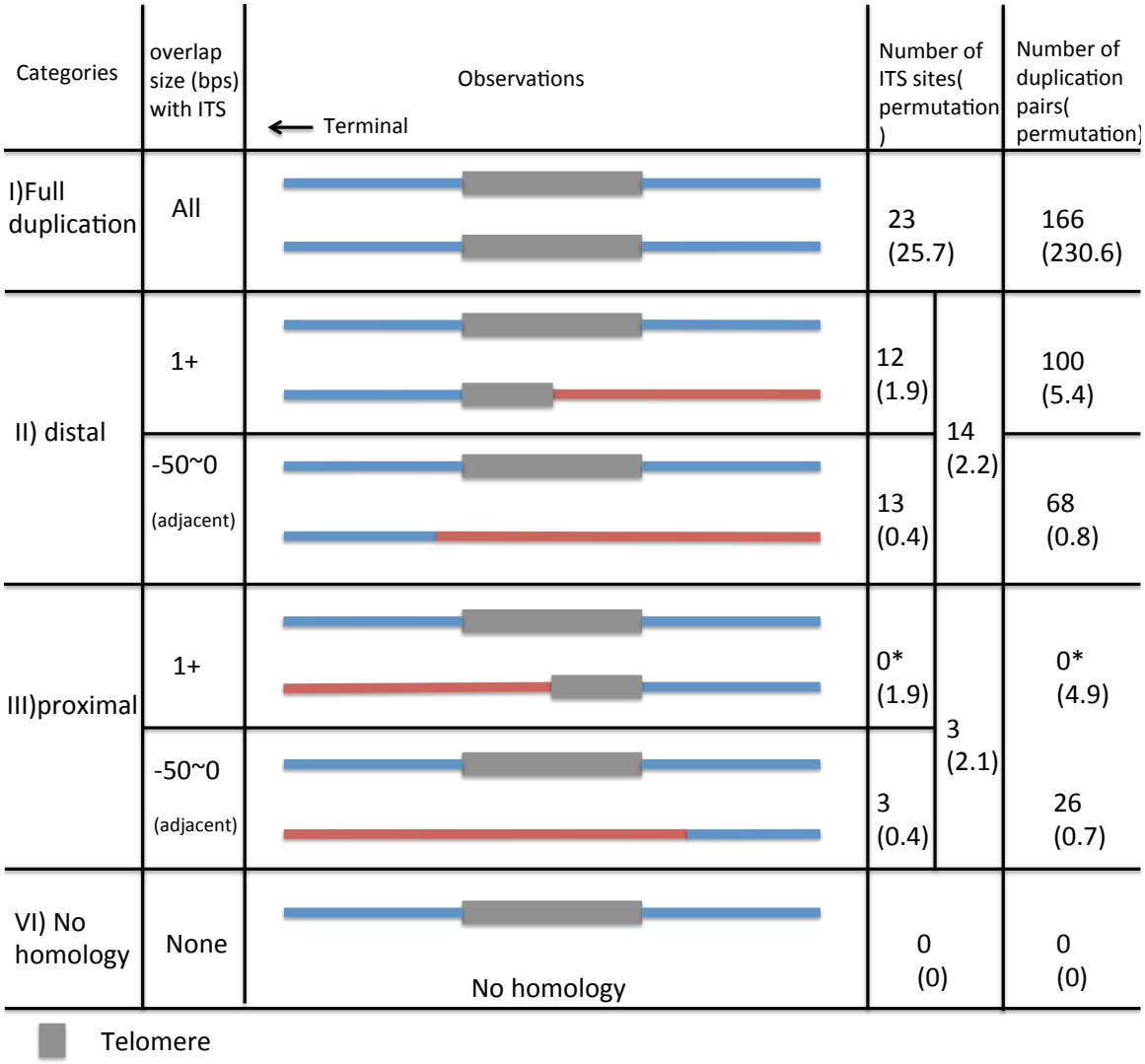

Figure 1. Summary of interstitial Telomeric Sequence(ITS).

*Summary of homology boundary for all subtelomeric ITS. All possible relationships between the sequence containing the ITS (top sequence in each cell), and homologous sequence (bottom sequence in each cell) are shown. Blue indicates homologous sequence and red indicates non-homologous sequence. Grey indicates telomere sequence. I) The homologous sequence spans the entire ITS(the homologous sequence may contain different size of ITS(5%) or no ITS(3%). See Table S3). II) The homologous sequence overlaps the distal breakpoint only. III) The homologous sequence is next to (<50bp) ITS distal breakpoint. IV) The homologous sequence is overlapping the proximal breakpoint only. V) ITS is next to (<50bp) ITS proximal breakpoint VI) No homologous sequence is observed. * means updating the orientation of 2q and 12p ITS as GRCh38. The permutation P-value of distal categories(II) are less than 0.001(Details see Table S2).*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

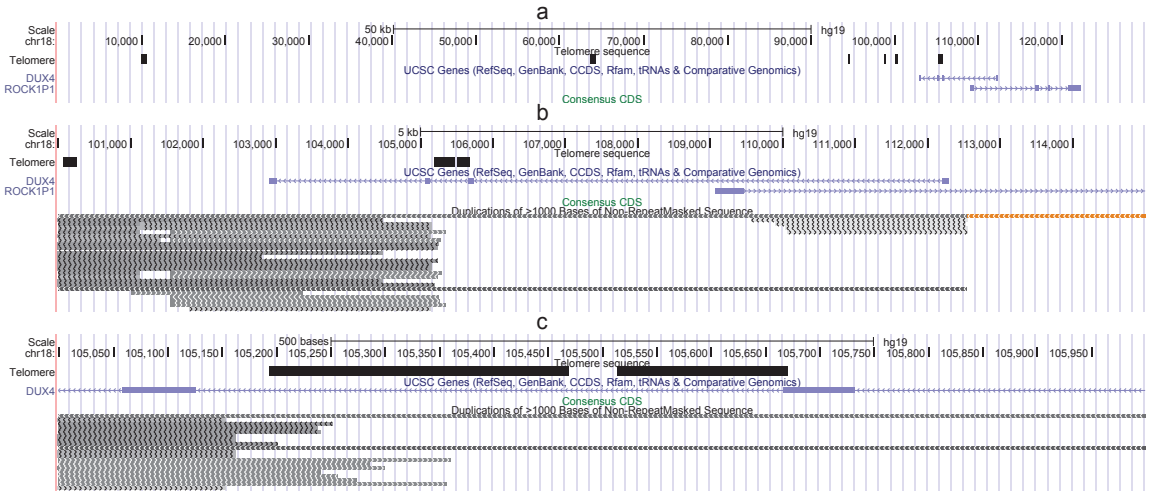

Figure 2. Interstitial Telomeric Sequence(ITS) from UCSC browser.
UCSC browser (Kent et al., 2002) displaying chr18p ITS subtelomeric duplications and genes at three different scales: a) 0-130k; b) 100-114k and c) 105-106k.

chimpanzee. Thus the aligner could only align these sequence to paralogous sequence and result in higher divergence.

Human extension rate estimated from species divergence

The comparison of other species chromosome end to human can not only verify the chromosome-end extension hypothesis but also be used to estimate the extension rate. We downloaded pairwise alignments for GRCh37 to 15 well-assembled species from UCSC (Kent et al., 2002). 39 well-assembled autosome ends as well as two ancient chromosomes 2 fusion ends (IJdo et al., 1991) were analyzed (see Methods). By aligning these species to human chromosome ends, we could identify the ancient chromosome end for each species most recent common ancestor with human (MRCA-human) as the most terminal end of the homology and therefore the missing homologous sequence can be defined as human extension sequence since the MCRA (see Methods, Table S5). Figure 5 indicates this process on two chromosome ends (9q and 15q), showing a core homologous region shared amongst eutherian genomes with different extension sequence.

Figure 6a and 6b shows the estimated length of human end extension on each chromosome end since the MRCA of human and other Eutherian genomes. Human end extension for comparison with species that have the same MRCA with human should be identical. For example, 68% of human end-extension relative to cow and sheep(same MRCA with human) are identical in size(within 1kb) while only 5% are identical relative to chimpanzee and cow (different MRCA with human). In the non-primate mammals group, the ancient chromosome ends are highly clustered together, 23 of them are estimated to be identical(labels highlight in Figure 6a,b, see Methods, Table S5). For example in Figure 5a,5b, the non-primate mammals are almost all inferred to have the same ends at 134 kb and 255 kb away from the human terminal at 9q and 15q respectively. Notably, 50(14%) non-primates mammals autosome ends are still serving as current terminals in human(Table S6, see Methods). These chromosome ends contain not only human extension sequence but also another species-specific extension sequence, for example, cat D4q, dog 9p and horse 25q(Figure 5a). The extension sequences for human and other species, together with the identical ancient chromosome ends confirm the ongoing extension of chromosome ends.

. The length of human-specific extension sequences represents a near linear relationship with MRCA time (Sequencing and Consortium, 2005; Scally et al., 2012; Locke et al., 2011; Murphy et al., 2001; Reisz and Müller, 2004)(Figure 6c,d), and are consistent with the accepted phylogenetic tree. One exception is that we identified 783 kb of human-specific chromosome extension sequence versus gorilla, whereas we identified 1744 kb versus chimpanzee, which invokes incomplete lineage sorting. However, this may be resolved by the observation that 30% of the gorilla genome sequence is closer to human or chimpanzee than the latter are to each other (Scally et al., 2012).

We also estimated the extension sequence for seven non-primate mammals against each other. Their

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

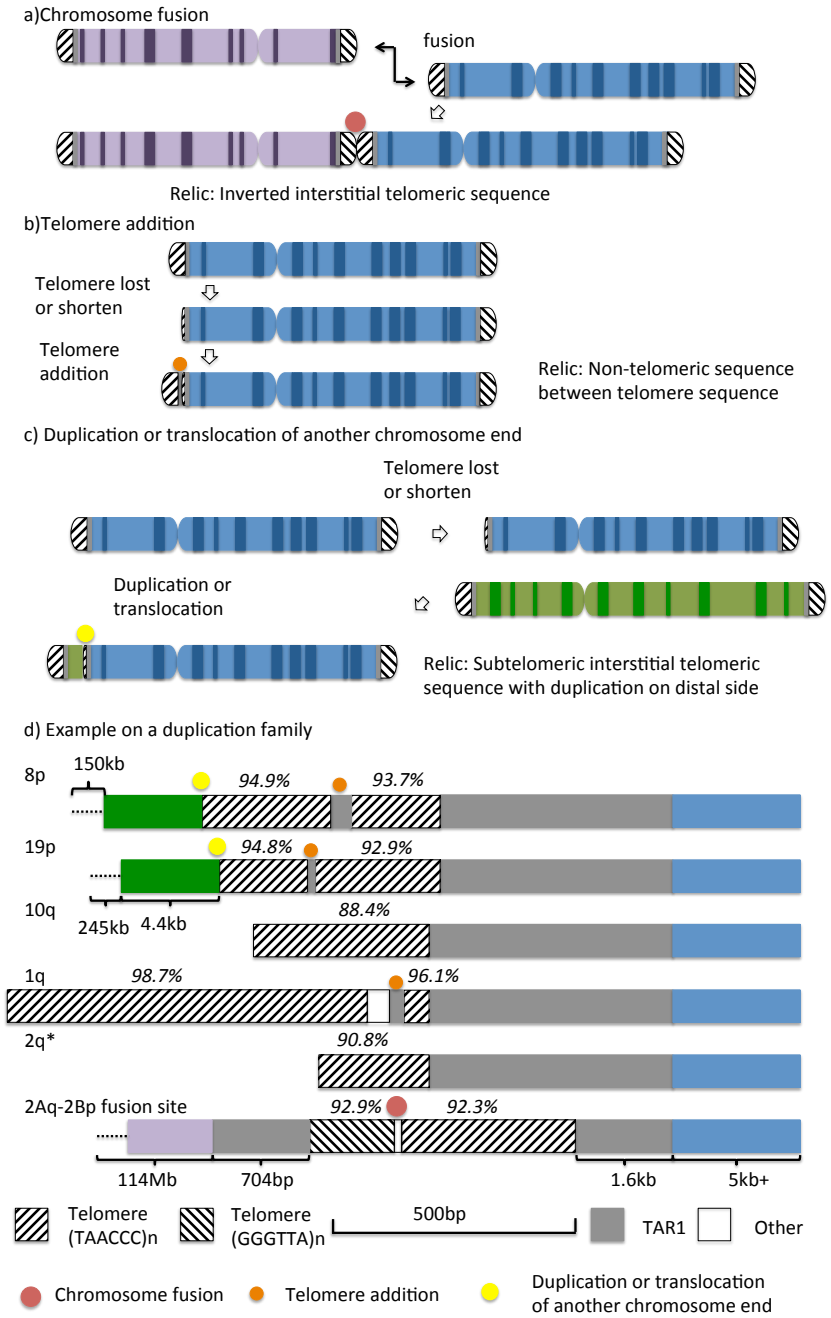

Figure 3. Shortened telomere repair models and example.

*a. Chromosome fusion. b. Telomere extension. c. Duplication or translocation of another end. d. An example in human genome within one duplication family. The main region is GRCh37:chr8:155249-155739. The size of each block is following the legend except the block with bracket. The telomere repeat identity is shown on the top. * means GRCh38 2q. 10q, 1q and 2q are chromosome terminal. The color blocks indicate homology between chromosomes.*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

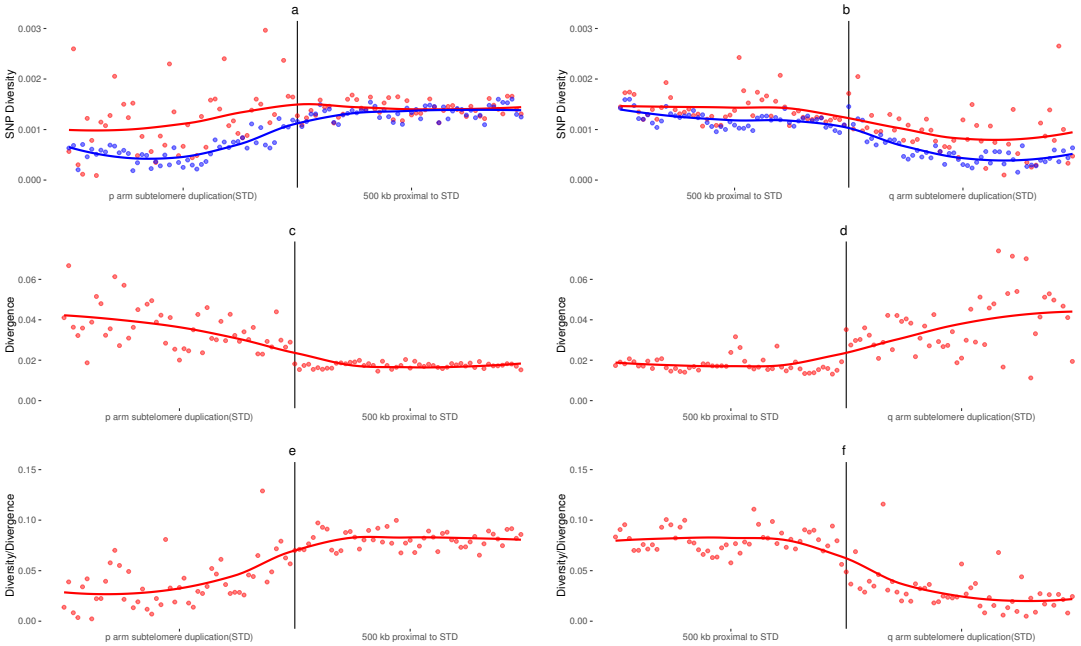

Figure 4. Average population genetic diversity and divergence at chromosome ends. *The chromosome end and adjacent sequence is divided into non-overlapping windows. In each window, we calculated standard and adjusted diversity(**a** and **b**), divergence(**c** and **d**) and adjusted diversity/divergence(**e** and **f**) as dot. Then the line is regression result from the dots. Blue and red are indicated the standard and adjusted estimation of diversity(see methods) in **a** and **b**.*

terminal all contained a sequence which couldn't align by any other species(Figure S2). Their extension size also follows the phylogenetic tree, for example, the rat and mouse have less extension sequence compare to each other than to other non-rodent mammals. It suggests the chromosome extension is widespread in mammals.

We estimated the extension rate by dividing the length of extension sequences by the estimated time since the most recent common ancestor (MRCA)(see Methods). This rate estimated the combined effect of both extension and shortening. If shortening is dominant, there will be zero extension sequence like 3p. Considering chromosome end extinction(see discussion), we only estimate the rate in extant chromosome end. In human, we could estimate this rate from the highly identical chromosome ends(count=23, see Methods) which have clear breakpoints among non-primate mammals(red bold chromosome ID at Figure 6a,6b). The human extension rate per chromosome terminal since the common ancestor of non-primate mammals is ranging from 0 to 0.0099 bp per year with an average rate of 0.0020 bp per year. The Primates, Rodentia, and Eulipotyphla extension rate per chromosome terminal are estimated by comparison to each other. They are estimated to be 0.0021, 0.0036 and 0.0022 bp per terminal per year for Primates, Rodentia and Eulipotyphla, respectively.

DISCUSSION

Our analysis indicated that many subtelomeric duplications have been mediated by subtelomeric interstitial telomeric sequence (ITS) and that the duplications preferentially occur on the distal side of these ITS. This indicates that the interstitial telomeric sequence is the ancient chromosome ends and that duplication occurred via a process of fusion to the capping telomere at the chromosome end. Moreover, the observed extensions in the BioNano sequence data(Shao et al., 2017) appear to be compatible with this hypothesis, although the current resolution of this approach is too large to be conclusive.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

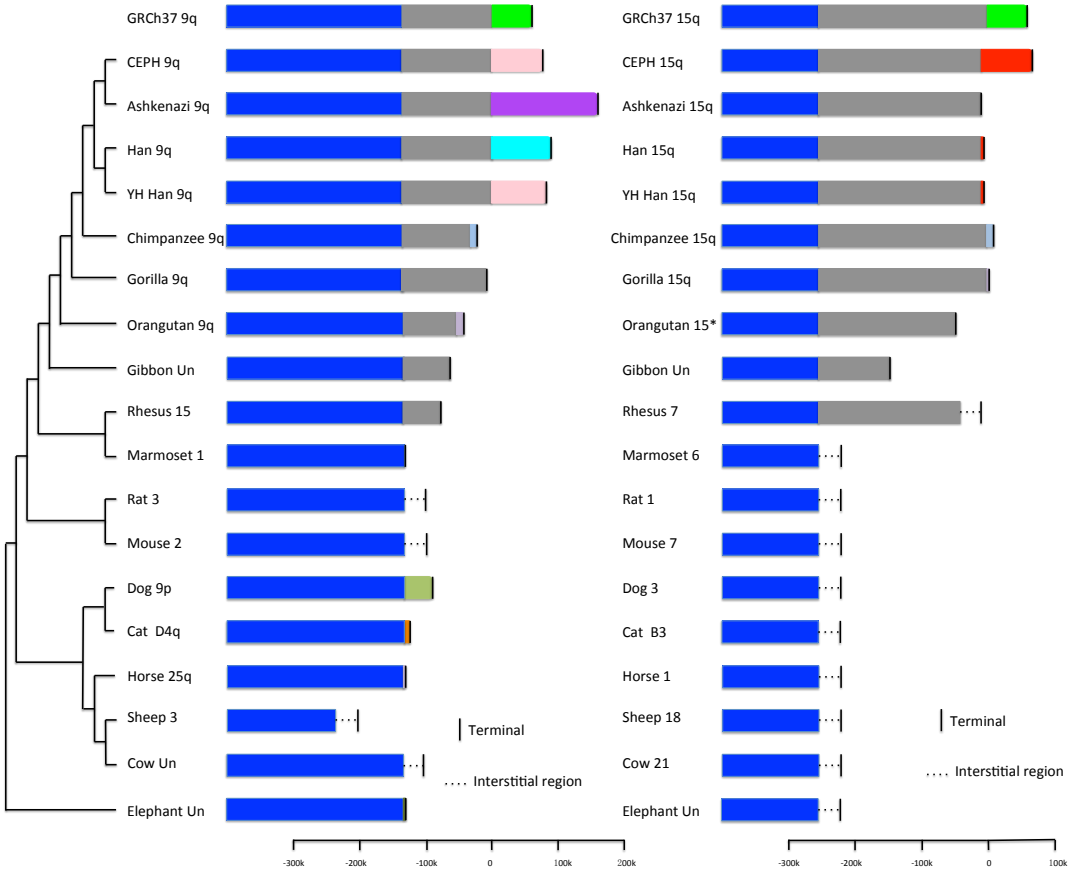

Figure 5. Paralogy map of 9q and 15q between mammals.

Phylogenetics tree are drew for the species(Murphy et al., 2001) and human(Poznik et al., 2016) population on the left. Different colors represent different block of homology sequence. Different species unknown regions are in different color. The non-chromosome-terminal (>1M) homology sequences in other species are not showed. The human population(CEPH, Ashkenazi and Han) terminals are based on BioNano data(Zook et al., 2016) by methods in (Shao et al., 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

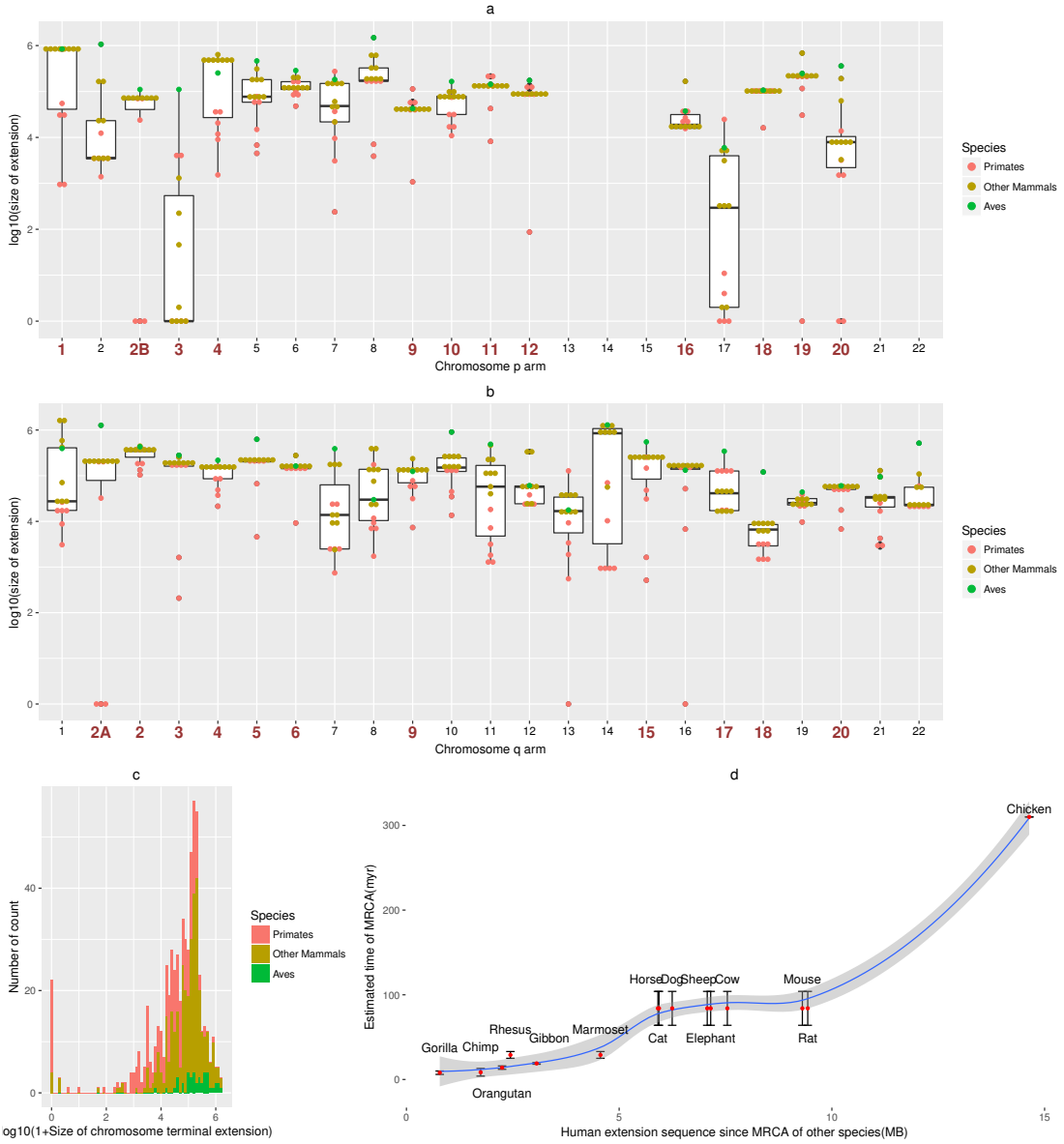

Figure 6. Dotplot, boxplot and size distribution of end extensions.

(a) is for p and (b) is for q arm. Each dot is an estimated size of human extension sequence against a species on each arm. Y axis is the normalized size of extension. Bond red numbers indicate these chromosome extension size are clustered together in non-primates group. (c) Length distribution of extension sequence in histogram. (d) Comparison of total extension sequence with MRCA time in 15 species.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

CONCLUSIONS

The dynamic nature of human chromosome ends has recently been examined using long-fragment optical mapping and sequencing techniques(Young et al., 2017; Shao et al., 2017). In this study, we provide further evidence to support a model of chromosome end extension model(Shao et al., 2017) at subtelomeric regions. By examining the pattern of overlap between interstitial telomeric sequence(ITS) and subtelomeric duplications, we have shown that a number of ITSs represent the ancient chromosome capping telomeres and provided evidence for chromosome extension at these ancient telomeres. In particular we have identified 2 ITSs for which there is strong evidence to support their role as ancient capping telomeric sequences. Other potential explanations for the distribution of subtelomeric ITS, including an insertion modelKilburn et al. (2001), were found to be incompatible with the observed patterns of homology. By examining nucleotide diversity and divergence in subtelomeric regions, we could show that chromosome ends appear to be younger than remaining chromosome from a population genetics perspective. Finally, comparison of chromosome ends amongst 15 species confirms that chromosome extension has taken place on multiple chromosomes in multiple mammalian lineages.

ACKNOWLEDGMENTS

L.C. was supported by an Australian Research Council Future Fellow (FT110100972). The research is supported by funding from the Australian Research Council (DP140103164). H.S. is funded by a University of Queensland scholarship.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Bailey, J. a. and Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, 7(7):552–64.

Blackburn, E. H. (2000). Telomere states and cell fates. *Nature*, 408(6808):53–56.

Bolzán, A. D. and Bianchi, M. S. (2006). Telomeres, interstitial telomeric repeat sequences, and chromosomal aberrations. *Mutation Research/Reviews in Mutation Research*, 612(3):189–214.

Flint, J., Craddock, C. F., Villegas, A., Bentley, D. P., Williams, H. J., Galanello, R., Cao, A., Wood, W. G., Ayyub, H., and Higgs, D. R. (1994). Healing of broken human chromosomes by the addition of telomeric repeats. *Am. J. Hum. Genet.*, 55(3):505–12.

Graves, J. A. M. (1995). The origin and function of the mammalian y chromosome and y-borne genes–an evolving understanding. *Bioessays*, 17(4):311–320.

Helena Mangs, A. and Morris, B. J. (2007). The human pseudoautosomal region (par): origin, function and future. *Current genomics*, 8(2):129–136.

IJdo, J. W., Baldini, a., Ward, D. C., Reeders, S. T., and Wells, R. a. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U. S. A.*, 88(20):9051–9055.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Res.*, 12(6):996–1006.

Kilburn, A. E., Shea, M. J., Sargent, R. G., and Wilson, J. H. (2001). Insertion of a telomere repeat sequence into a mammalian gene causes chromosome instability. *Molecular and cellular biology*, 21(1):126–135.

Lin, K. W. and Yan, J. (2008). Endings in the middle: current knowledge of interstitial telomeric sequences. *Mutation Research/Reviews in Mutation Research*, 658(1):95–110.

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., and Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, 437(7055):94–100.

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S.-P., Wang, Z., Chinwalla, A. T., Minx, P., et al. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–533.

McClintock, B. (1941). The stability of broken ends of chromosomes in zea mays. *Genetics*, 26(2):234–282.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018

Mefford, H. C. and Trask, B. J. (2002). The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.*, 3(2):91–102.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, 409(6820):614–618.

Nergadze, S. G., Rocchi, M., Azzalin, C. M., Mondello, C., and Giulotto, E. (2004). Insertion of telomeric repeats at intrachromosomal break sites during primate evolution. *Genome research*, 14(9):1704–1710.

Poznik, G. D., Xue, Y., Mendez, F. L., Willems, T. F., Massaia, A., Sayres, M. A. W., Ayub, Q., McCarthy, S. A., Narechania, A., Kashin, S., et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide y-chromosome sequences. *Nature genetics*, 48(6):593–599.

Reisz, R. R. and Müller, J. (2004). Molecular timescales and the fossil record: a paleontological perspective. *TRENDS in Genetics*, 20(5):237–241.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., et al. (2005). The dna sequence of the human x chromosome. *Nature*, 434(7031):325–337.

Sabatier, L., Ricoul, M., Pottier, G., and Murnane, J. P. (2005). The loss of a single telomere can result in instability of multiple chromosomes in a human tumor cell line. *Molecular Cancer Research*, 3(3):139–150.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., et al. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175.

Sequencing, T. C. and Consortium, A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

Shao, H., Zhou, C., Cao, M. D., and Coin, L. (2017). Ongoing human chromosome end extension revealed by analysis of bionano and nanopore data. *bioRxiv*, page 108365.

Shay, J. W. and Wright, W. E. (2006). Telomerase therapeutics for cancer: challenges and new directions. *Nature reviews Drug discovery*, 5(7):577–584.

Smit, A. F., Hubley, R., and Green, P. (1996). Repeatmasker open-3.0.

Tanaka, H., Abe, S., Huda, N., Tu, L., Beam, M. J., Grimes, B., and Gilley, D. (2012). Telomere fusions in early human breast carcinoma. *Proc. Natl. Acad. Sci. U. S. A.*, 109(35):14098–103.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Young, E., Pastor, S., Rajagopalan, R., McCaffrey, J., Sibert, J., Mak, A. C., Kwok, P.-Y., Riethman, H., and Xiao, M. (2017). High-throughput single-molecule mapping links subtelomeric variants and long-range haplotypes with specific telomeres. *Nucleic Acids Research*, 45(9):e73–e73.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26624v1 | CC BY 4.0 Open Access | rec: 6 Mar 2018, publ: 6 Mar 2018