Ongoing human chromosome end extension driven by a primate ancestral genomic region revealed by analysis of BioNano genomics data

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
[*]Correspondence: l.coin@imb.uq.edu.au

ABSTRACT

The majority of human chromosome ends remain incomplete due to their highly repetitive structure. In this study, we use BioNano data to anchor and extend chromosome ends from two European trios as well as two unrelated Asian genomes. Two thirds of BioNano assembled chromosome ends are structurally divergent from the reference genome, including both deletions and extensions. The majority of extensions are homologous to sequences on chromosome 1p, 5q and 19p. These extensions are heritable and in some cases divergent between Asian and European samples. We identified two sequence families in these sequences which have undergone substantial duplication in multiple primate lineages, leading to the formation of new fusion genes. We show that these sequence families have arisen from progenitor interstitial sequence on the ancestral primate chromosome 7. Comparison of chromosome end sequences from 15 species revealed that chromosome end divergence matches the corresponding phylogenetic relationship and revealed a rate of chromosome extension since the primate divergence of 80-440 kbp per million years.

Introduction

The genome sequence of chromosome ends in the reference human genome assembly are still incomplete. In the latest draft of the human genome[1] 21 out of 48 chromosome ends were incomplete; amongst which five chromosome ends (13p,14p,15p,21p,22p) are completely unknown and the remaining chromosome ends are capped with 10kb-110kb of unknown sequence. There are many interesting observations in the chromosome end regions which remain unexplained, such as the observed increase in genetic divergence between Chimpanzee and Humans towards the chromosome ends[2].

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 [3,4] and have experienced rapid adaptive selection[3]. Subtelomere length polymorphism is also found in humans[5]. The majority of subtelomeric duplications have the same orientation towards the chromosome end[3,4]. Based on this it has been suggested that they originated from reciprocal translocation of chromosome tips and unbalanced selection[4].

Telomere repeat sequences ($[TAAGGG]_n$) - which are the capping sequences of chromosome ends - are breakable, acquirable and fusible in the genome. In somatic cells, telomeres are observed to progressively shorten [6,7]. If the telomere sequence is lost, the broken chromosome will become unstable[8–10], and multiple types of rearrangements can occur, including chromosome fusion[8], tips translocation[11], or direct addition of telomere repeats[10]. The manual insertion of telomere sequence in interstitial region results in enhanced chromosome breakages and induces high rates of chromosome rearrangements around the insertion[12]. Interstitial telomeric sequences (ITS) are widespread in the genome[13,14]. In subtelomeric regions, their orientations are almost always towards the terminal end of the chromosome, like the STD[4].

The quality of assembly for the chromosome ends largely depends on the sequencing technology. To correctly assemble highly duplicated regions like chromosome ends, sequence reads or read pairs spanning the repeat are required[15]. Recently, the NanoChannel Array (Irys System) from BioNano Genomics[16] was introduced. This technology can generate barcodes on DNA fragments which are hundreds of kilobases long by detecting the distance between specific enzyme recognition sites. Alignment and genome assembly are performed based on numerous distinct site distance fragments. These very long fragments enable construction of individual physical maps, as well as completing the reference in unknown regions[17]. In this manuscript, we report on observed subtelomeric dynamics using this technology for the first time. We conclude that these genome dynamics reflect ongoing chromosome extension and deletion and identify genomic regions which have undergone substantial extension in multiple primate lineages.

Results

Assembling chromosome ends using BioNano data supports updates to GRCh38 reference

We downloaded data from BioNano Genomics(see Methods). The data contains eight samples, including two family trios (Ashkenazi, CEPH) and two Chinese samples. The raw data has been assembled into contigs (with N50 of 3.3MB) and aligned to the GRCh37 reference. We used the most distal unique aligned sequences at the chromosome ends to anchor individual chromosome ends (see Methods). Overall, 19 reference chromosome ends were extended from the GRCh37 reference (Figure 1, Supplementary Figure 1). The extension sequences supported a re-orientation of chromosome 2q ends in the new GRCh38 reference (Supplementary Figure 2a). We observed extension sequence at 9q in one family which matched the GRCh38 extension of this chromosome(Supplementary Figure 2b). We also observed a 260kb extension sequence in one sample at 16p terminal which was originally discovered in 1991[5](Supplementary Figure 1.31).

Discovery of heritable chromosome end polymorphism

We revealed characteristics of chromosome ends by analysis of completed chromosome ends of the eight samples. We focused on autosomes in this study. Five unknown heterochromatic acrocentric chromosome short-arms (13p, 14p, 15p, 21p and 22p) were excluded from analysis. The chromosome 1p end was also excluded because of the discontinuous alignment of the BioNano contigs at a reference gap(Supplementary Figure 1.1). For the remaining 38 chromosome ends, we compared them to the corresponding GRCh37 chromosome ends (Method, Figure 1, Supplementary Figure 1).

Chromosome end polymorphism was frequently observed versus the GRCh37 reference (Fig.1 and Supp. Fig. 1). Amongst the total of 304 per-sample chromosomes ends from the 8 samples, only 36.5% are identical to the reference genome. For the remaining chromosome ends, 49.3% contain deletions (average size=55k, standard deviation, SD=77k, Supplementary Table 1) and 41.8% of chromosome ends are extended (average size=71k, SD=79k, Supplementary Table 2). Chromosome end polymorphism within the 8 samples was observed at 47.4% (18/38) of the chromosome ends. For example, four types of extension sequences (unaligned,1p,19p,16q) are observed in 9q arm amongst 8 samples (Figure 1a). Chromosome end polymorphism are widespread amongst all samples studied (Supp. Fig. 1). We observed that the daughter chromosome end sequences (both deletion and extension) are almost always (93.4%, 71/76) also observed in one of her parents, indicating that the chromosome end data are unlikely to be a result of an assembly artefact.

Inferring the origin of extension sequences

We recognized the origin of the extension sequences by re-aligning the BioNano labels (enzyme recognition site) patterns to GRCh37 reference. Most of the extension sequences were unaligned due to relatively short size and the 10kb density of enzyme recognition sites (average length of unaligned and aligned contigs was 53k and 97k respectively). All the aligned extension sequences are aligned to one of 1p, 3q, 5q, 16q or 19p.

We identified (see Methods) two sub-telomeric duplication families (family A and B of length 9kb and 8kb respectively) which both have 16 copies in the human genome. Family A is observed in all of the 1p, 3q, 5q, 16q, 19p extension sequence regions; whereas family B is only observed in 1p,3q 16q and 19p. We identified all homologous sequences to these two families in the non-redundant nucleotide database which we used to construct a phylogenetic tree for each family(see Methods, Figure 2a,2b, Supplementary table 3,4). These duplications were only found in close species and were highly duplicated in human (A:16,B:16 copies) and chimpanzee(A:9,B:7 copies). Both trees are clustered into 3 parts: namely i. interstitial duplication across multiple species, ii. human subtelomeric duplication and iii. chimpanzee subtelomeric duplication. The interstitial duplications are clustered between species on a longer branch, suggesting they form an orthologous group. We conclude that the subtelomeric cluster duplicated from an ancient interstitial sequence in the genome millions years ago (see Discussion). The human and chimpanzee subtelomeric copies do not cluster together, suggesting these subtelomeric copies are recently and independently duplicated in each species.

Identification of the ancient shortening chromosome end telomeric sequence

There are multiple interstitial telomeric sequences (ITS) in the human genome which are orientated in the same direction at subtelomeric regions[4]. For example, 6 telomere sequences are in the first 110kb of 18p (Figure 3a). The reciprocal tips translocation model[4] for subtelomeric duplication does not involve the chromosome terminus, nor create new telomere repeat sequence, and thus we are unaware of any mechanism for explaining these sequences. We investigated the relationship between all chromosome end ITS and duplications of 1kb or more (Supplementary Figure 3, Supplementary Table 5, Method). We found that all ITS are either fully contained within a duplication, or within 50bp of a duplication (Figure 3d). The vast majority overlap or are next to a duplication on the distal side of the ITS (16 sites) rather than proximal (3 sites, of which each site also has a distal-side duplication overlap, see Supplementary Table 5). The dominance of distal side ITS duplication suggests that duplication events occurred at the terminal-end of the ITS, and that the current subtelomeric ITS were actually the ancient chromosome end telomere sequence. We also observed multiple homologies to different regions with variable degree of overlap

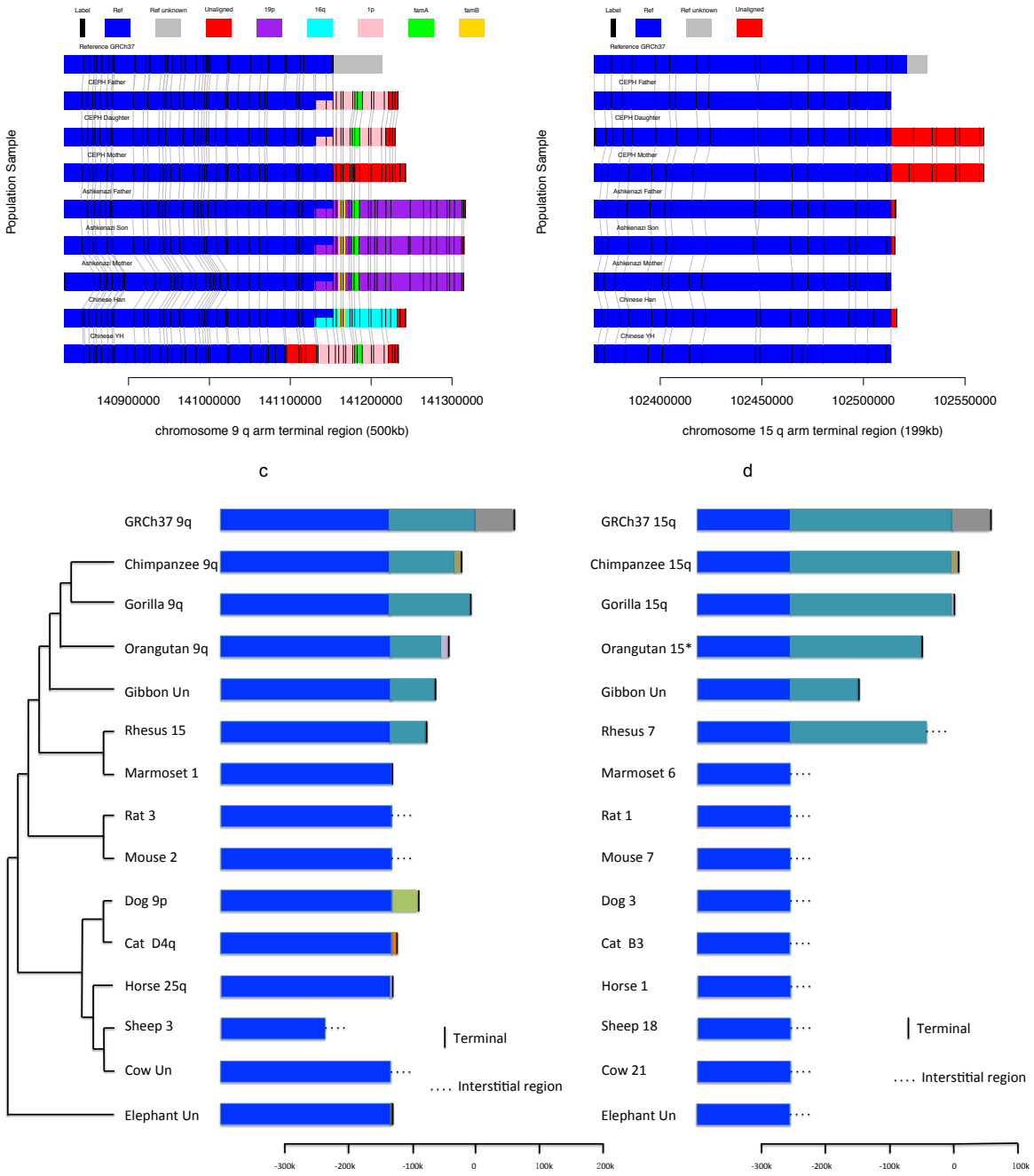

Figure 1. Paralogy map for chromosome tips region. a:) chromosome 9q b:) chromosome 15q for BioNano assembled chromosome ends in eight samples. c:) chromosome 9q and d:) 15q for 14 mammals. The enzyme recognition sites (labels) are marked as black bars and gray connecting lines indicate alignment between samples. Homologous sequence is indicated by color block: blue indicates homology to human reference sequence at the given chromosome; grey indicates unknown reference sequence; purple, cyan and pink indicate homology to 19p, 16q and 1p respectively. The remaining unaligned regions are all colored with red. Overlapping colors indicate a shared homology to multiple sources. Yellow and bright green indicate homology with family A and family B respectively. An overlapping color indicates homology two sequences. In c) and d) phylogenetic trees are drawn for the species[18] . The same color indicates the same homology. * means Orangutan chromosome 15 unlocalized scaffold. The solid black line indicates the location of the chromosome end.

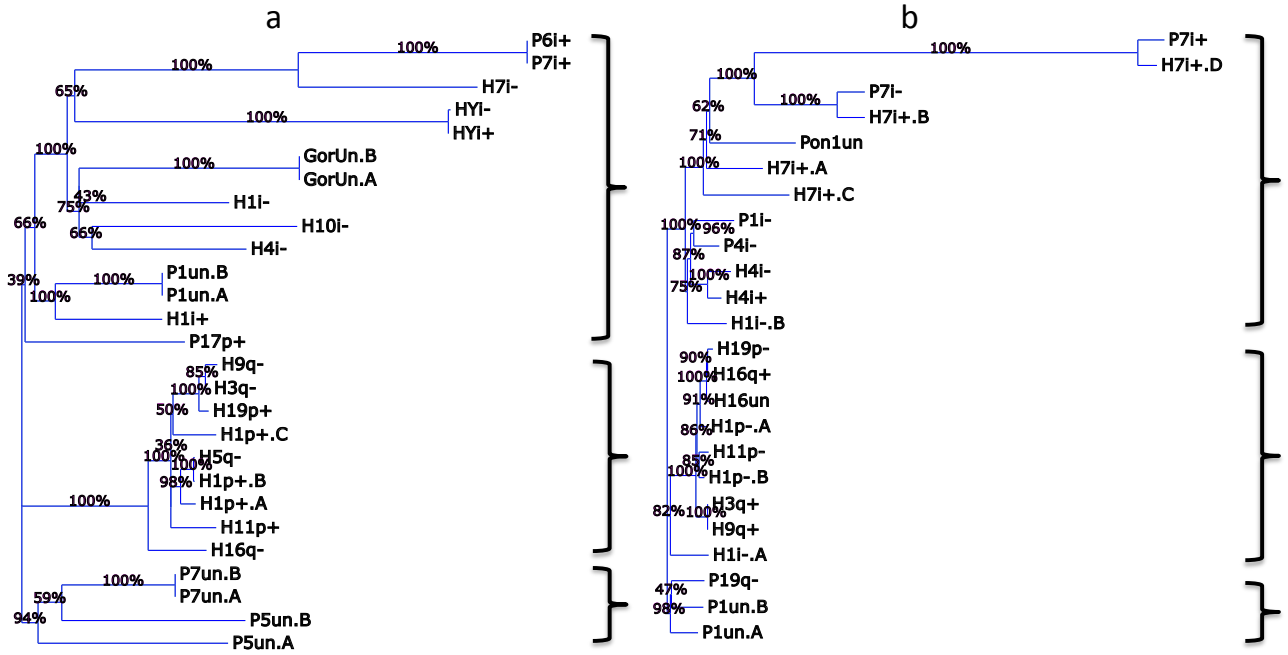

Figure 2. Phylogenetic trees for two sequence familes identified in human extension sequences a) Family A (9kb) b) Family B (8kb). Phylogenetic trees are generated from their homologous sequences. H for Homo sapiens, P for Pan troglodytes, Gor for Gorilla and Pon for Pongo abelii. p for p arm, q for q arm, Un for unplaced contig, un for unlocalised within a chromosome and i for interstitial genome regions. '+' for forward aligned and '-' for reverse aligned. '.A','.B','.C' and '.D' are appended to distinguish different copies in the same region. The brackets indicate clustering into the 3 groups: interstitial; human subtelomeric and chimpanzee subtelomeric. Bootstrap value are showed on each branch.

with the ITS (for example between 10 bps to 168 bps at position 105kb on chr18p, see Figure 3c). This suggests that duplication of chromosome terminal sequence has taken place independently multiple times, and is likely mediated by telomeric homology. We conclude that multiple directly duplicated events at ancient chromosome ends have generated multiple present day ITS sequences at subtelomeric regions.

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[11]. There are multiple ways to repair the dysfunctional chromosome end[10,11,20], including chromosome end fusion (Figure 4a), telomere addition (Figure 4b), duplication or translocation of another chromosome end (Figure 4c). Relics can be found for all of these events in the human genome(Figure 4d). Chromosome end fusion is found at ancient chromosome 2A and 2B fusion into chromosome 2[20], which can be seen from a characteristic inverted interstitial telomere sequence. Telomere addition to telomere is indistinguishable from common telomere shortening and lengthening unless non-telomeric sequence is also involved in the addition. A common observation for ITS or capping telomere is that is has TAR1 (telomere associated repeat 1) element inside, and furthermore the proximal telomere identity is lower than the distal telomere identity(Figure 4d). 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[10,11,20], 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[21] data to estimate average genetic diversity at chromosome ends (See methods) [Figure 5, 5d]. From these data, we found the diversity decreases sharply in the first 500kb of the p-arm and the last 500kb of the q-arm. This is driven by a decline in diversity in subtelomeric duplications in these regions[Figure 5c,5f], but not the remaining non-duplicated chromosome end regions [Figure 5b,5e]. Diversity is typically correlated with divergence. However, the chimpanzee genome project reported that the divergence between human and chimpanzee constantly increases toward the chromosome end, suggesting the chromosome end is the most divergent sequence between human and chimpanzee [2]. Thus chromosome ends are regions with the low diversity and high divergence which is consistent with a model of chromosome

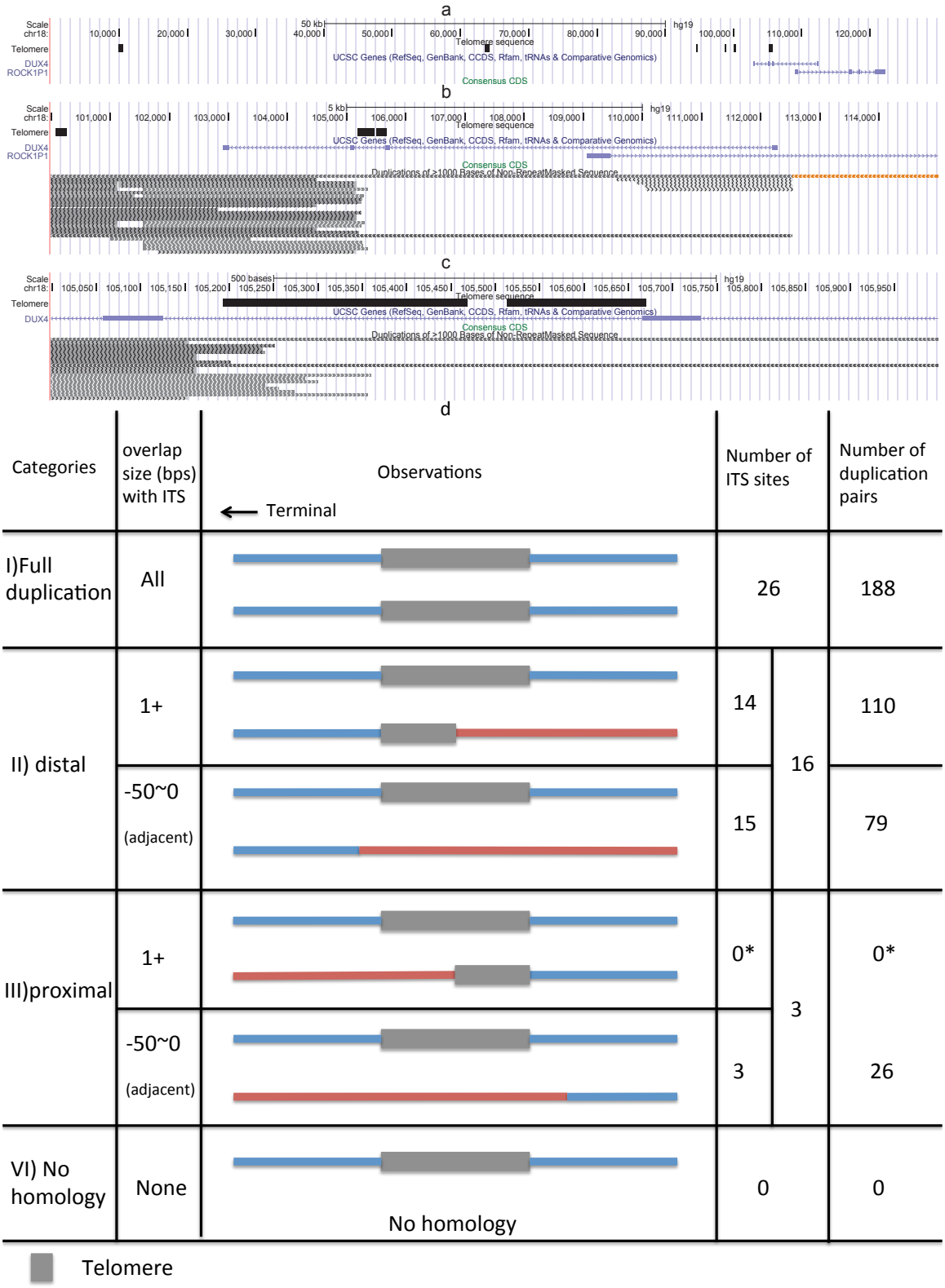

Figure 3. Summary of interstital Telomeric Sequence(ITS). UCSC browser [19] displaying chr18p ITS subtelomeric duplications and genes at three different scales: a) 0-130k; b) 100-114k and c) 105-106k. d) Summary of homology boundary for all subtelomeric ITS. All possible relationships between the sequence 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. II) The homologous sequence overlaps the distal breakpoint. III) The homologous sequence is next to (<50bp) ITS distal breakpoint. IV) The homologous sequence is overlapping the proximal breakpoint. 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.

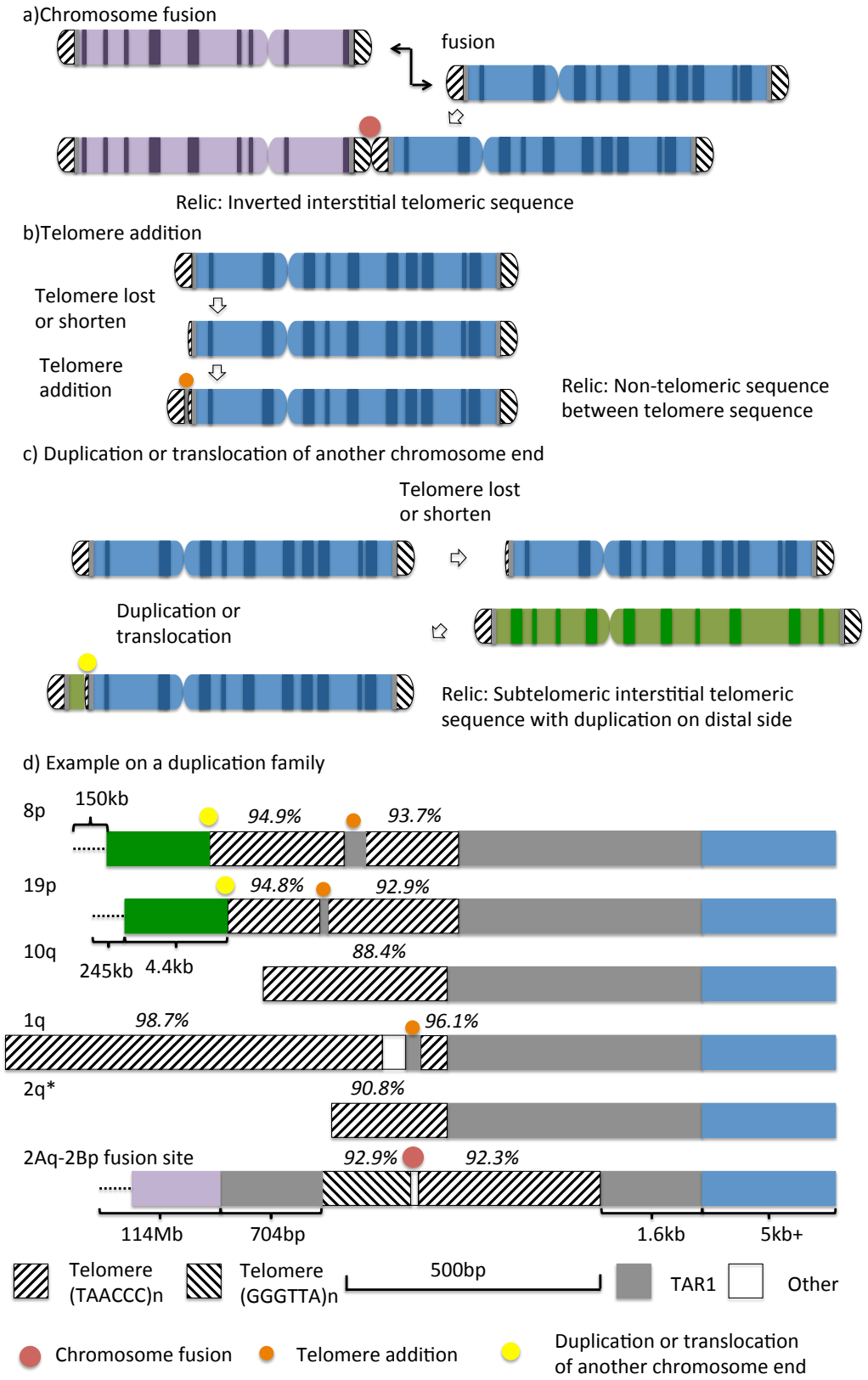

Figure 4. Shortened telomere repair models and example. a. Chromosome fusion. b. Telomere addition. c. Duplication or translocation of other end. d. Example in human genome within one duplication family. The main region is GRCh37:chr8:155249-155739. The size of each block is following the legend except the block with bracket. The telomere repeat identity is showed on the top. * means GRCh38 2q. 10q, 1q and 2q are chromosome terminal. The color blocks indicate homology between chromosomes.

end extension occurring within primates, as newly duplicated end sequence would have a low within population diversity, but would also result in high divergence between species.

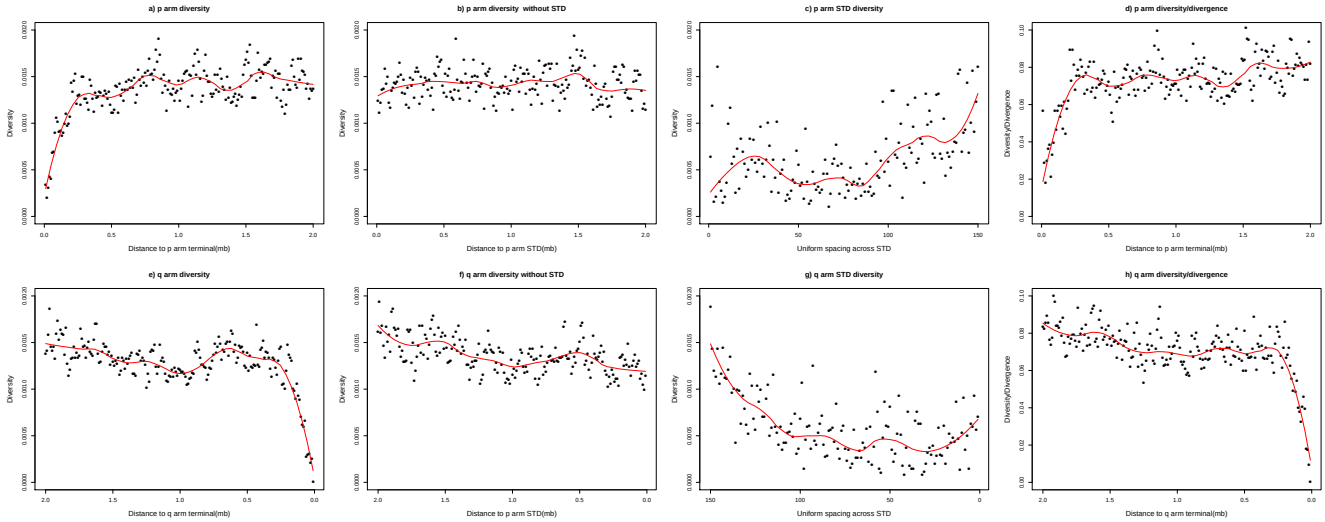

Figure 5. Average population genetic diversity and divergence at chromosome ends. The reference is divided into non-overlapping windows. For 2a and 2e, each bin is a 10kbp window in first 2Mbp sequence in p or q arm. For 2b and 2f, each bin is a 10kbp window in the first 2Mbp sequence next to p or q arm subtelomeric duplication. For 2c and 2g, the regions are uniformly divided into 150 bins for each p or q arm subtelomeric duplication regions. Diversity/divergence is shown in 2d and 2h. Divergence is estimated from human to chimpanzee alignment. Red lines are local regression results.

Chromosome extension has resulted in formation of new fusion genes

Although duplication of a sequence does not create any new coding sequences, different combinations of duplication could reuse the coding sequences to create new fusion genes. For example, the fusion of chromosome 2A and 2B created a fusion gene (AK123946) [20]. At 18p, the duplication joining to ancient telomere sequence(chr18:105194-105670) created a fusion gene DUX4(Figure 3b). The duplication families A and B both contain coding sequences and pseudogenes. Surprisingly, the majority (7/9 in family A and 4/8 in family B) of duplications have different fusion genes (see Methods, Supplementary Table 3,4). For example, 11p and 16q both contain sequences from families A and B, but the sequence between the two duplications are different which creates two different fusion genes, namely LINC01001 and FAM157C [Supplementary Figure 3]. All the fusion genes are expressed in the Expression Atlas[22], suggesting these recent fusion genes have functional importance. By duplicating regions contain coding sequences andimpact combining them at chromosome ends, new functional fusion genes could be created.

Human extension rate estimated from species divergence

The comparison of other species chromosome end to human can not only verify the extension hypothesis but also estimate the extension rate. We downloaded pairwise alignments for GRCh37 to 15 well-assembled species from UCSC[19]. 39 well-assembled autosome ends as well as two ancient chromosome 2 fusion ends [20] were analyzed (see Methods). By aligning other species to human chromosome ends, we could identify the ancient chromosome end for the most recent common ancestor(MRCA) of human and this species 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, Supplementary Table 6). Some MRCAs are the same between 15 species, thus their ancient chromosome ends are expected to be identical. We observed that 61% (373/615) of the per-species ancient chromosome ends inferred in this way are identical (see Methods, Supplementary Table 6). For example, at 9q and 15q the non-primate mammals are almost all inferred to have the same ends at 134 kb and 255 kb away from the human terminal respectively (Figure 1c,1d). Extension sequences are found not only in human but also in other species when these homologous sequences are also their current chromosome ends, for example, cat D4q, dog 9p and horse 25q(Figure 1c). 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[2,18,23–25](Table 1), and are consistent with the accepted phylogenetic tree. One exception is that we identified 783kbps of human-specific

chromosome extension sequence versus gorilla, whereas we identified 1744kbps versus chimpanzee, which runs counter to the common belief that human is more closely related to chimpanzee than gorilla. However, this may be resolved by the the observation that 30% of the gorilla genome sequence is closer to human or chimpanzee than the latter are to each other[23].

We estimated the human extension rate by dividing the total length of extension sequences by the estimated time since the most recent common ancestor (MRCA) (see Methods, Table 1).The extension rate is more accurately estimated by close species such as primates. Assuming the GRCh37 represents the human full chromosome sequences and the pairwise alignment results represent the ancient chromosome ends, the average human autosome ends (n=41) extension rate since the MRCA with other primates is between 78 and 436 kbps per million years.

Table 1. Human chromosome end extension sequence with estimated MRCA[2,18,23,24] time.

(a) Primates

Species	Chimp	Gorilla	Orangutan	Gibbon	Rhesus	Marmoset
Extension sequence(kb)	1745	783	2249	3059	2444	4560
MRCA(myr)	4-13	6-10	12-16	18-20	25-33	
Rate(kb/myr)	134-436	78-131	141-187	174-192	74-98	138-182

(b) Mammals

Species	Rat	Mouse	Dog	Cat	Horse	Sheep	Cow	Elephant
Extension sequence(kb)	9433	9307	6247	5911	5936	7070	7543	7152
MRCA(myr)	64-104							
Rate(kb/myr)	91-147	89-145	60-98	57-92	57-93	68-110	73-118	69-112

(c) Aves

Species	Chicken
Extension sequence(kb)	14639
MRCA(myr)	310
Rate(kb/myr)	47

Discussion

In this study we provide evidence that chromosome ends are dynamic and have undergone substantial extension since divergence of primates. We show that the process of chromosome-end extension is ongoing and generates significant polymorphism in the population, which is heritable from one generation to the next.

We identified two duplication families which are participating in ongoing chromosome end expansion. These families appear to both originate from the same region on ancient primate chromosome 7(GRCh38, chr7:45793082-45823926), separated only by 13kb; however they have undergone distinct patterns of duplication in subtelomeric regions of at least two primates (Chimpanzee and Human). These families are still participating in subtelomeric duplication in human populations, as can be seen by alignment of these families to observed BioNano chromosome extension sequence. We showed that duplication of these families in sub-telomeric regions has lead to the formation of new fusion genes. This may provide a mechanism for the extension sequence to be adaptively selected and eventually become fixed in the population. We also showed that formation of extension sequence has resulted in genetic diversity within subtelomeric duplications.

Our analysis indicated that many subtelomeric duplications have been mediated by subtelomeric interstitial telomeric sequence (ITS), and that the duplications occur on the distal side of these ITS. This indicates that the interstitial telomeric sequence are 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 appear to be compatible with this

hypothesis, although the current resolution of this approach is too large to be conclusive.

Normally, duplication of a subtelomeric region into another subtelomeric regions will never involve interstitial sequence, thus the interstitial origin of sequences at subtelomeric regions indicate an unknown mechanism. Interestingly, two cases of direct interstitial sequence joining to telomere sequence were described in a human cancer fusion study[9]. One was a 447 bps interstitial chromosome 4 sequence joining to 10q telomere at one side and 4p telomere at another side. The other was a 211 bp interstitial chromosome 2 sequence and a 374 bps chromosome 17 sequence joining to 4q telomere and Xp telomere, respectively. These cases in cancer support the ability of interstitial sequence to join to the telomere. A possible model is that the interstitial sequence is deleted and then directly joined to the telomere, similar to the 'deletion-plus-episome' model for double minutes in cancer[26,27].

There are several potential mechanisms we might consider for annealing of chromosome sequence to the capping telomere sequence[4]: non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), microhomologous-mediated end joining(MMEJ)[28] and single strand annealing(SSA)[28]. The presence of overlapping homology at the breakpoints suggests this mechanism is homology mediated. Moreover, the co-location of the breakpoints with interstitial telomere sequence indicates that the process may be partially mediated by micro-homology of the telomeric repeat sequence. Both MMEJ and SSA are also consistent with the observation that all interstitial telomere repeat sequence are in the same orientation.

Methods

Bionano data analysis

We downloaded publicly available BioNano data for eight humans (a CEPH trio, an Ashkenazi trio and two Han Chinese) from BioNano Genomics website http://bionanogenomics.com/science/public-datasets/. The initial downloaded data contained raw Bionano data, assembled contigs as well as the unique alignments from contigs to GRCh37 generated using RefAlign from BioNano Genomics IrysSolve. This alignment mapped each Bionano contig to the best matching position on GRCh37, and did not require a full-length alignment of contig to reference, thus allowing us to identify chromosome end polymorphism. We assumed that contigs which are aligned to the most distal sequence with highest matching score represent the individual sample chromosome ends. As the technology is currently unable to resolve different chromosome end haplotypes, the chromosome ends can be viewed as a dominant marker for the longest allele.

The reference chromosome end sequences which could not be aligned by are defined as deletion sequences. The reference unknown sequences ("N" region) are removed for deletion size estimation. The unaligned sequence at the distal part of the individual chromosome end is defined as extension sequence. Deletion and extension sequences are all more than 8.9 kb(average labels distance), otherwise, we defined as the reference type. Chromosomes 13p, 14p, 15p, 21p, 22p were removed because of missing reference sequence. Sex chromosomes are removed because they are unequal between sample. 1p is removed when all the sample's contigs are discontinuous at the reference unknown region (chr1:471k-521k) and the remaining of 1p could only be aligned as secondly alignment.

In order to recognize the origin of the extension sequences, we realigned the contigs to GRCh37(373,590 labels) allowing multiple matches by RefAlign. A pre-alignment process in RefAlign was used to merge labels which were close to each other. This process resulted in identification of 346,991 labels for subsequent analysis. The mean distance and standard deviation between adjacent labels was 8.9 kb and 100 kb, respectively. We then used RefAlign to re-align the contigs to the GRCh37 reference enzyme recognition sites (RefAlign parameter: -output-veto-filter _intervals.txt$ -res 2.9 -FP 0.6 -FN 0.06 -sf 0.20 -sd 0.0 -sr 0.01 -extend 1 -outlier 0.0001 -endoutlier 0.001 -PVendoutlier -deltaX 12 -deltaY 12 -hashgen 5 7 2.4 1.5 0.05 5.0 1 1 1 -hash -hashdelta 50 -mres 1e-3 -hashMultiMatch 100 -insertThreads 4 -nosplit 2 -biaswt 0 -T 1e-12 -S -1000 -indel -PVres 2 -rres 0.9 -MaxSE 0.5 -HSDrange 1.0 -outlierBC -AlignRes 2. -outlierExtend 12 24 -Kmax 12 -f -maxmem 128 -BestRef 0 -MultiMatches 5). This re-alignment allowed us to more accurately identify chromosome ends. In particular, for 6p in sample NA12878, the multiple alignments supported a more distal alignment (chr6:73k-364k) than the unique alignment (chr6:245k-1870k). The break of these two contigs was mediated by a connection from 6p to a chromosome 16 interstitial region, which could be an artefact. Similarly, for 16p in NA24385, the multiple alignments supported a more distal alignment (chr16:67k-204k) than the unique alignment (chr16:204k-3728k). The more distal contig could align to both 19p and 16p with no overlap, but the alignment at 19p(chr19: 61k-244k) is longer than 16p making it primarily align to 19p in unique alignment. Notably, the contig with highest matching score for 19p in this sample is another contig(chr19:61k-2341k).

We also realigned individual contigs to other contigs with the same parameter as above. When a label in one sample was aligned to another sample label, we regarded these two labels were connected. For duplication family A and B, we extracted the label IDs from GRCh37 label map. Family A contained 3 to 4 labels. Family B contained 2 labels. For the alignment regions containing family A or B, we only drew these regions when at least 3 or 2 labels were aligned to GRCh37, respectively.

For the CEPH and Ashkenazi trios, we checked the alignments from the child to their parent for every chromosome end(Supplementary Figure 1). If the child chromosome end was aligned to one of its parent with no more than one label difference at the terminal, we considered that the child chromosome end was inherited from the parent.

Subtelomeric duplication analysis

We downloaded all pairs of regions in the human genome (GRCh37) with high sequence identity from the segmental duplication database[29]. We assign the copy number of a base-pair position to be the number of entries in this file which overlap this position. We take all positions with copy number at least 22 to be high copy number duplication families. We order these families by their base-pair length, defined as the longest contiguous stretch of DNA with copy number greater than 22. Only duplication families which have at least one member within 1MB of a chromosome end, were kept (Supplementary Table 7). This resulted in an identification of two duplication families with high copy number, which was selected for evolutionary analysis.

We used NCBI blast to align these two sequences (GRCh37,A:chr1:652579-662291,B:chr16:90220392-90228245) to the human genome, chimpanzee genome and nucleotide collection (NT) database, filtering out sequences with less than 95%(human and chimpanzee) or 85%(other) sequence identity to the full query sequence. Next, we extracted the duplications, aligned them by mafft[30][key option: strategy G-INS-i] and built the neighbor-joining tree by T-rex[31][option:default].

Subtelomeric duplication regions are defined as the longest contiguous duplication sequence starting from the chromosome terminal. From the duplication database[29], both pairs of duplications regions inside subtelomeric duplication regions are defined as subtelomeric duplications. The longest autosome subtelomeric duplication pairs are chr1:317720-471368 and chr5:180744335-180899425. Sex chromosomes are excluded from this analysis because homology between the X and Y chromosomes in p arm terminal is maintained by an obligatory recombination in male meiosis[32].

Subtelomeric Interstitial telomere sequence analysis

Telomere sequences were annotated from GRCh37 repeatmasker database[33]. We extracted the non-capping telomere sequences inside subtelomeric regions as subtelomeric ITS sequence. We also included the ancient subtelomeric region in the chromosome 2 fusion sites[20]. Then we searched for all the subtelomeric ITS and their adjacent duplications in duplication database[29]. Bases on all the possible overlapping, we divided into 6 types. Type a is duplication spanning the whole ITS. Type b and type d are duplications overlap with the ITS from the distal and proximal site, respectively. Type c and type e are duplications 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.

Fusion gene analysis

Gene information was obtained by zooming into specific genome regions from UCSC browser[19]. Genes which contain coding sequence which lies within the duplication as well as coding sequence which lies outside the duplication it is considered as a fusion gene. We extracted gene expression information from the Expression Atlas[22] recording the highest expression level in Supplementary Table 3 and 4.

Chromosome end population genetics analysis

Mutation frequencies were extracted from 1000 genome vcf files (v5.20130502). Chromosome 13, 14, 15, 21, 22 p arm (unknown terminal sequence) and sex chromosomes were excluded from analysis as above. We divided the genome into L = 10kbp non-overlapping windows for calculating diversity and divergence. Within a window, diversity is calculated as the average of the base pair diversity, i.e. $\bar{h} = \frac{\sum_{j=1}^{L} h_j}{L}$. 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. We performed a local regression on diversity using the R function(loess.smooth parameter: degree=10,span=0.3). Finally, we summarize the average diversity from all the chromosome 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 is total number of chromosomes. For analysis of subtelomeric duplications only, the window size could not be pre-specified because the size of the regions are variable. As a result, for this analysis we specified the total number of bins to be 150 for each chromosome subtelomere region. In order to calculate divergence between Human and Chimpanzee we first aligned the human and Chimpanzee genomes using bwa[34] (default parameters). Divergence was estimated as the percentage of unaligned base pairs within each 100bp window. The entirely unaligned regions are excluded for this estimation.

Human extension rate analysis

We downloaded pairwise alignment files from UCSC3. These files contain region alignments from species to GRCh37. Initially, 21 genomes (Chicken, Chimp, Cow, Dog, Gibbon, Gorilla, Horse, Marmoset, Mouse, Orangutan, Rat, Rhesus, Sheep, Baboon, Cat, Elephant, Kangaroo, Panda, Pig, Rabbit and Zebrafish) were analyzed. For each human autosome as well as ancient chromosome 2A and 2B[20], 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.

The human most terminal end of homology are defined as the ancient chromosome ends for the last common ancestor of human and this species. For each ancient chromosome end, we calculated the number of species ends, which has one or more similar ends(within 1kb) with other species end.

The sequence starting from the ancient chromosome end for a last common ancestor to the current human chromosome ends are defined as human chromosome extension sequence since the divergence of human with a given species. Human chromosome end unknown sequences("N" regions) are removed from size estimation. Total autosome extension sequence (s) are the sum of all autosome terminal extension sequences. The human autosome expansion rate since the divergence of human and this species 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.

Authors' contributions

H.S. and L.C. conceived the study. H.S. and C.Z. performed the analysis. H.S. wrote the manuscript, which was revised and approved by all authors.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

L.C. is 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

1. Consortium, I. H. G. S. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Sequencing, T. C. & Consortium, A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16136131. DOI 10.1038/nature04072.
3. Mefford, H. C. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.* **3**, 91–102 (2002). DOI 10.1038/nrg727.
4. Linardopoulou, E. V. *et al.* Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94–100 (2005). URL http://www.nature.com/nature/journal/v437/n7055/pdf/nature04029.pdf. DOI 10.1038/nature04029.
5. Wilkie, A. O. *et al.* Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* **64**, 595–606 (1991).
6. Blackburn, E. H. Telomere states and cell fates. *Nature* **408**, 53–56 (2000).
7. Shay, J. W. & Wright, W. E. Telomerase therapeutics for cancer: challenges and new directions. *Nature reviews Drug discovery* **5**, 577–584 (2006).
8. McClintock, B. The stability of broken ends of chromosomes in zea mays. *Genetics* **26**, 234–282 (1941).
9. Tanaka, H. *et al.* Telomere fusions in early human breast carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14098–103 (2012). URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84865542465{&}partnerID=tZOtx3y1. DOI 10.1073/pnas.1120062109.
10. Flint, J. *et al.* Healing of broken human chromosomes by the addition of telomeric repeats. *Am. J. Hum. Genet.* **55**, 505–12 (1994). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1918405{&}tool=pmcentrez{&}rendertype=abstract.
11. Sabatier, L., Ricoul, M., Pottier, G. & Murnane, J. P. The loss of a single telomere can result in instability of multiple chromosomes in a human tumor cell line. *Molecular Cancer Research* **3**, 139–150 (2005).

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

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

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

15. Chaisson, M. J., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics* (2015).

16. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3** (2016).

17. Shi, L. *et al.* Long-read sequencing and de novo assembly of a chinese genome. *Nature Communications* **7** (2016).

18. Murphy, W. J. *et al.* Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).

19. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/186604$\delimiter"026E30F$nhttp://www.genome.org/cgi/doi/10.1101/gr.229102. DOI 10.1101/gr.229102.

20. IJdo, J. W., Baldini, a., Ward, D. C., Reeders, S. T. & Wells, R. a. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 9051–9055 (1991). DOI 10.1073/pnas.88.20.9051.

21. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3498066{&}tool=pmcentrez{&}rendertype=abstract. DOI 10.1038/nature11632. /www.pubmedcentral.nih.gov/articlerender.fc.

22. Petryszak, R. *et al.* Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic acids research* gkv1045 (2015).

23. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).

24. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).

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

26. Storlazzi, C. T. *et al.* Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome research* **20**, 1198–1206 (2010).

27. Carroll, S. *et al.* Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Molecular and cellular biology* **8**, 1525–1533 (1988).

28. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings (2008).

29. Bailey, J. a. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–64 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16770338. DOI 10.1038/nrg1895.

30. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). DOI 10.1093/molbev/mst010.

31. Alix, B., Boubacar, D. A. & Vladimir, M. T-REX: A web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40** (2012). DOI 10.1093/nar/gks485.

32. Ross, M. T. *et al.* The dna sequence of the human x chromosome. *Nature* **434**, 325–337 (2005).

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

34. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).