ARTICLES

Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication

Elena V. Linardopoulou[1,2], Eleanor M. Williams[1], Yuxin Fan[1]†, Cynthia Friedman[1], Janet M. Young[1] & Barbara J. Trask[1,2,3]

Human subtelomeres are polymorphic patchworks of interchromosomal segmental duplications at the ends of chromosomes. Here we provide evidence that these patchworks arose recently through repeated translocations between chromosome ends. We assess the relative contribution of the principal mechanisms of ectopic DNA repair to the formation of subtelomeric duplications and find that non-homologous end-joining predominates. Once subtelomeric duplications arise, they are prone to homology-based sequence transfers as shown by the incongruent phylogenetic relationships of neighbouring sections. Interchromosomal recombination of subtelomeres is a potent force for recent change. Cytogenetic and sequence analyses reveal that pieces of the subtelomeric patchwork have changed location and copy number with unprecedented frequency during primate evolution. Half of the known subtelomeric sequence has formed recently, through human-specific sequence transfers and duplications. Subtelomeric dynamics result in a gene duplication rate significantly higher than the genome average and could have both advantageous and pathological consequences in human biology. More generally, our analyses suggest an evolutionary cycle between segmental polymorphisms and genome rearrangements.

The human genome contains an abundance of large DNA segments that have duplicated over the last 40 million years[1,2]. These segmental duplications represent ≥5% of the genome[2] and are frequently found near centromeres and telomeres[3]. Segmental duplications are emerging as important factors in chromosomal rearrangements leading to disease[4] and rapid gene innovation[2], but the mechanisms by which they form are not well understood. Here we focus on the unusually dense concentrations of interchromosomal segmental duplications comprising human subtelomeres, which form the transition zones between chromosome-specific sequence and the arrays of telomeric repeats capping each chromosomal end. Previous cytogenetic studies have shown that human subtelomeres are strikingly polymorphic in content—large segments can be present in or absent from normal alleles[5]—and that the copy number of subtelomeric segments can vary among higher primates[6–9]. This natural plasticity, combined with documented expression of several human subtelomeric genes[10,11], suggests that the evolutionary dynamics of subtelomeric regions could contribute to normal phenotypic variation within and between primate species, as is observed in other organisms (reviewed in ref. 5). However, subtle rearrangements of DNA near the ends of chromosomes are observed in association with human disorders, including mental retardation[12]. Although full sequence coverage has not yet been achieved for all chromosome ends, let alone for multiple alleles of each end, much can be learned from available sequence about subtelomere organization, evolution, variation and function, as well as more generally about the origin and consequences of segmental duplications.

Complex interrelated structures

Our ‘paralogy map’ of subtelomeric segmental duplications (Fig. 1 and Supplementary Table S1) uses all finished sequences of genomic clones submitted to GenBank before April 2003. The map comprises ~2.6 Mb of sequence present in two or more of 33 human subtelomeres (including three allelic pairs). The seven completely sequenced subtelomeres in the set are bounded distally by 0.5–2.4 kb of various tandemly repeated units[13] called telomere-associated repeats (TAR1) and a short sample of the native telomeric arrays[14]. Numerous degenerate telomere-like repeats and TAR1 elements are also situated at varying distances from telomeres[15,16] (Fig. 1). Notably, these repeats are almost always oriented $5'$–$3'$ towards the telomere.

The paralogy map reveals the complex patchwork of sequence blocks shared by human subtelomeres. Different subtelomeres can show >100 kb continuous similarity, but a segment shared by a given chromosome set extends only 13 kb on average before being displaced on at least one subtelomere by a segment with a different chromosomal distribution. In the 33 subtelomeric contigs analysed, we identify 41 homology blocks larger than 3 kb (Fig. 1) (blocks 42–44 are special cases and not counted, see Supplementary Table S2). These blocks occur in 2–18 copies (average of 5), with 88–99.9% identity (Supplementary Table S2). Almost all instances of these blocks are in the same orientation and relative order (Fig. 1). Polymerase chain reaction (PCR) analyses of monochromosomal hybrid cell lines confirm block boundaries defined by sequence

[1] Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North C3-168, Seattle, Washington 98109, USA. [2] Department of Bioengineering, University of Washington, Box 357962, Seattle, Washington 98195-7962, USA. [3] Department of Genome Sciences, University of Washington, Box 357730, Seattle, Washington 98195-7730, USA. †Present address: Departments of Laboratory Medicine and Medicine (Division of Medical Genetics), University of Washington, Seattle, Washington 98195, USA.

alignments and identify at least one additional chromosomal copy for 17 out of 29 blocks evaluated (Supplementary Fig. S1 and Table S3).

Subtelomeres contain members of 25 small families of genes (Fig. 1). There is one gene per 30 kb on average. Eighteen families contain at least one subtelomeric member that encodes a potentially functional protein (Supplementary Table S4). Thus, gain, loss or alteration of subtelomeric genes has a potential phenotypic effect. Subtelomeric genes have highly varied functions and include odorant and cytokine receptors, tubulins, transcription factors and genes of unknown function.

Sequences in the paralogy map and duplicates thereof detected in later assemblies and/or by PCR (an added total of 0.97 Mb) account for ≥83% of the estimated subtelomeric terrain in a typical genome[16]. Approximately 90% of the 490 kb of finished sequence added to nine ends in the latest genome assembly (Build 35) is >90% identical to sequence already represented in our data set; only 26 kb is novel. Thus, our data set represents a reasonably comprehensive sample from which mechanistic information can be derived.

Mechanisms of sequence transfer

To investigate the mechanisms resulting in subtelomeric segmental duplications, we considered two phases in their evolutionary history. The first consists of duplication to a new chromosome, creating a new structural boundary, and the second involves possible interactions between existing duplicates. We analysed the patterns and breakpoints of homology in sequenced subtelomeres to infer the mechanism of interchromosomal sequence transfer that would result in the first step. Two primary models were considered that might give rise to subtelomeric segmental duplications, namely chromosome translocations and DNA transposition.

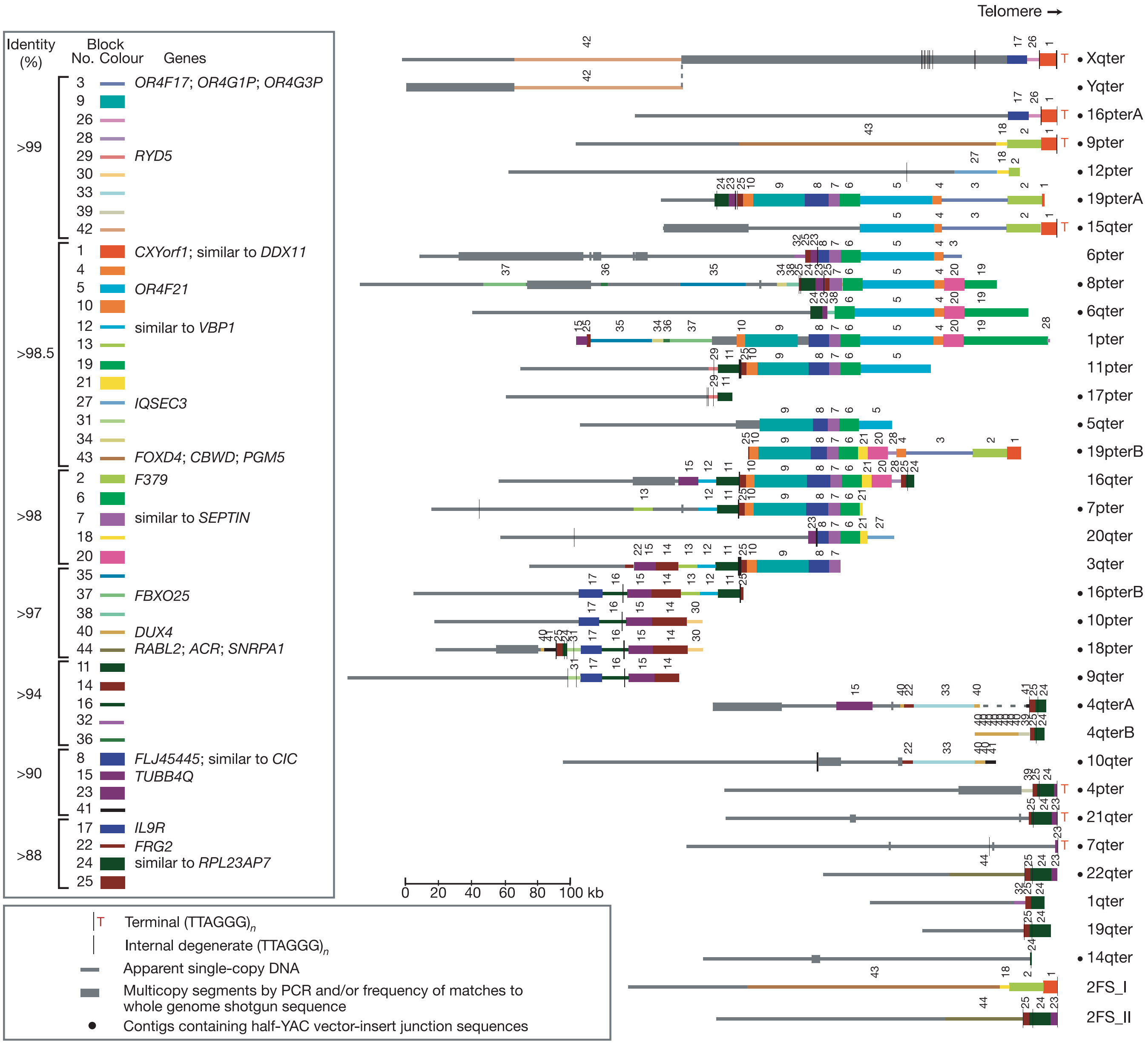


Figure 1 | Subtelomeric paralogy map. Subtelomeric contigs (Supplementary Table S1 lists constituent accession numbers and localization methods) are aligned at telomeres or to maximize alignments of paralogous blocks. Copies of a given block have the same colour, line width and number. Only blocks 15 and 40 on 4q, 22 on 3q, 34–37 on 1p, and 38 on 6q are in inverted orientation relative to other corresponding block copies. 2qFS_I and _II represent ancestral telomeres fused head-to-head at 2q13-14; other internal paralogies are not displayed or analysed here. A and B indicate allelic variants. Yq/Xq pseudoautosomal homology extends distal of dotted line. See Supplementary Table S4 for details about the subtelomeric gene copies.

Several observations argue for the translocation model (Fig. 2) and against transposition. First, subtelomeric blocks do not have characteristic features associated with known transposons or their insertion sites[17]. Proposed targets for insertions of segmental duplications by a more general transpositional model[18] are also not found at subtelomeric homology breakpoints. Second, the preserved centromere–telomere orientation and order of most duplicated blocks and degenerate telomeric repeats, and the embedded patterns of shared blocks (Fig. 1), argue against a transpositional model.

Instead, the block patterns are consistent with patchwork formation by numerous translocations involving the tips of chromosomes and subsequent transmission of unbalanced chromosomal complements to offspring (Fig. 2). In this model, each translocation event has the potential to create a new homology boundary and define a new block. Figure 3 illustrates how two translocations led to the duplication of a subtelomeric segment (block 4 plus 5) and its juxtaposition between different neighbours on chromosomes 15q and 8p. The sequence of events can be inferred from the state of interspersed repeat elements at homology breakpoints. Chromosomes 15q and 16q represent ancestral states, and the intermediate state of 6p reveals temporal separation of the two translocations leading to the block configuration on 8p.

Translocations can result from aberrant repair by either non-homologous end-joining (NHEJ) or homologous recombination; both are major mechanisms of double-strand break (DSB) repair in mammalian cells[19,20]. To deduce the relative contribution of NHEJ and non-allelic homologous recombination (NAHR) to subtelomeric block juxtapositions, we examined all homology breakpoints at single-nucleotide resolution. The presence of repetitive elements of the same class (or paralogous genes) at the homology boundary in both aligned junction sequences, often with a transition from high to lower sequence identity within the repeat, is strongly suggestive of homology-based repair (for example, see Supplementary Fig. S2, where the original state can be recognized by characteristic direct repeats flanking the *Alu* element). In contrast, the absence of aligned repeats or the presence of a truncated repetitive element (or gene) at the homology boundary in one sequence is indicative of non-homologous end-joining (as at breakpoints A and B, Fig. 3).

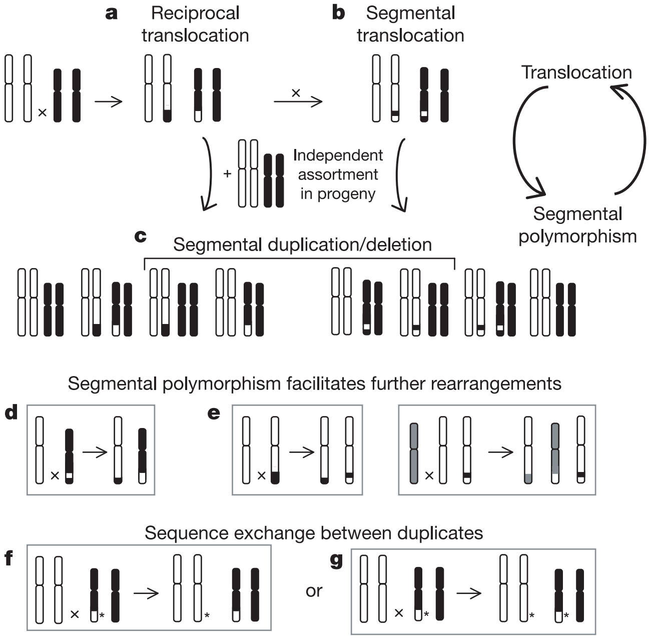

a
Reciprocal translocation
×
b
Segmental translocation
×
+
Independent assortment in progeny
Translocation
Segmental polymorphism
c
Segmental duplication/deletion
Segmental polymorphism facilitates further rearrangements
d
×
e
×
×
Sequence exchange between duplicates
f
×
*
*
or
g
×
*
*
*

Figure 2 | A translocation-based model of segmental duplication and polymorphism. **a**, A terminal duplication/deletion can arise if a translocation product and an intact homologue are passed from parent to offspring, creating a segmental polymorphism (**c**). **b**, A segmental duplication/deletion can arise if a second interchromosomal exchange occurs between the translocated chromosomes. **d**, **e**, Segmental polymorphism can facilitate further rearrangements by promoting translocations through interchromosomal homologies (**d**) or causing translocation or other rearrangement owing to the absence of homology (**e**). **f**, **g**, Both reciprocal and non-reciprocal homology-based sequence transfers are possible between duplicates generated by any of the above steps. Asterisk indicates sequence variant.

We identified a complete, non-redundant set of 56 junction-sequence alignments, each representing a unique translocation event, in the sequenced subtelomeres. We deduced the repair mechanism in 53 of these cases (Fig. 4, Supplementary Fig. S3 and Table S5). The vast majority seem to result from NHEJ (49/53, 92%) (Fig. 4b). We infer repeat-mediated NAHR for only four (8%) of the events (Fig. 4a), three of which involved *Alu* repeats. In the 15 cases of NHEJ for which structures representing both original partners and one translocation derivative were available, we found ≤5 bp of homology between the original sequences at the junction site (for example, Fig. 3b). Small insertions found at eight junction sites are consistent with NHEJ-mediated translocations; eight cases of apparent large deletions could have formed either by translocation or intrachromosomal deletion (Supplementary Table S5 and Fig. S4)

Although duplication borders of segmental duplications were found in genome-wide analyses to be enriched in recently active *Alu* repeats[21,22], interspersed repeats are not enriched at the DSBs leading to subtelomeric segmental duplications. Of a total of 102 independent DSBs (Fig. 4), 45% occurred within a repetitive element (10.8% in *Alu* elements), close to the frequency expected from subtelomeric repeat content (Supplementary Table S6). We do, however, find degenerate telomeric repeats at 4% of these DSBs, whereas they occupy 0.5% of subtelomeric sequence (Supplementary Tables S5 and S6). Subtelomeres are notably enriched in degenerate telomeric repeats relative to adjacent single-copy sequence or other genomic regions (~10- and ~100-fold, respectively) (Supplementary Table S6). These repeats could have been appended during DSB repair, the postulated genesis of other interstitial telomere-like repeats[23]. Breakpoint 22 in Supplementary Table S5 is a clear example of such a process. Although we cannot rule out a functional role for these repeats, they are probably scars of many past DSB repairs.

Generation of diverse structures by multiple translocations between chromosome ends (Fig. 2a–c) is just one aspect of subtelomeric dynamics. Once duplicates exist on different chromosomes, they are subject to homology-based reciprocal or non-reciprocal

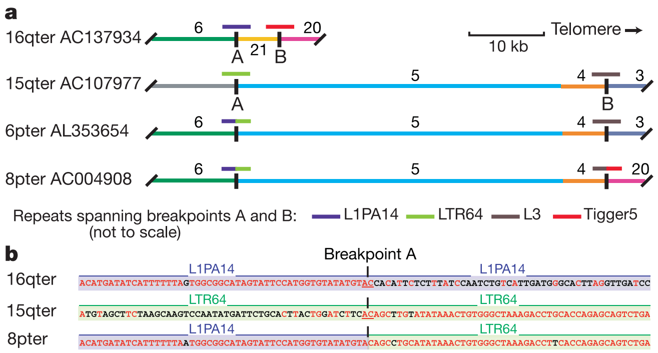

a
16qter AC137934
6
21
20
A
B
10 kb
Telomere →
15qter AC107977
5
4
3
A
B
6pter AL353654
6
5
4
3
8pter AC004908
6
5
4
20
Repeats spanning breakpoints A and B: L1PA14 LTR64 L3 Tigger5
(not to scale)
b
Breakpoint A
16qter
L1PA14
L1PA14
ACATGATATCATTTTTTAGTGGCGGCATAGTATTCCATGGTGTATATGTACCACATTCTCTTTATCCAATCTGTCATTGATGGGCACTTAGGTTGATCC
15qter
LTR64
LTR64
ATGTAGCTTCTAAGCAAGTCCAATATGATTCTGCACTTACTGGATCTTCACAGCTTGTATATAAACTGTGGGCTAAAGACCTGCACCAGAGCAGTCTGA
8pter
L1PA14
LTR64
ACATGATATCATTTTTTAATGGCGGCATAGTATTCCATGGTGTATATGTACAGCCTGCATATAAACTGTGGGCTAAAGACCTTCACCAGAGCAGTCTGA

Figure 3 | Layers of interchromosomal translocations form subtelomeric blocks. **a**, Paralogous blocks have shared colour and number; short coloured lines above indicate different repetitive elements at homology breakpoints A and B, which define two translocations. An intact copy of each repeat is preserved in 16q and 15q sequences spanning the homology breakpoints with 6p and 8p, which contain truncated repeats fused by NHEJ. **b**, Only two identical nucleotides (underlined) are found at the point where the original two sequences were joined at breakpoint A to form a hybrid. Aligned matching bases are red.

sequence transfers (Fig. 2f, g). These events do not generate new block boundaries, but can supplant mutations accrued on one chromosome with those from another copy and spread structures formed by NHEJ to new locations (Fig. 2d). We reasoned that if duplication and subsequent homology-based sequence transfer are separated by sufficient time, the latter could be observed as a significant shift in sequence identity within regions of similarity.

To test for such events, we evaluated fluctuations in sequence identity along a 60-kb region, parts of which are shared with 88–99.5% identity by seven subtelomeres (Fig. 5a). Four computational approaches indicate that homology-based sequence transfers occurred many times between these paralogues. (1) The best-matching pairs (that is, partners in the most recent transfer events) change ≥5 times along the sequences (Fig. 5b and Supplementary Fig. S5). (2) The phylogenetic relationships of neighbouring sections are strikingly incongruent (Fig. 5c). (3) The per cent identity between any two subtelomeres shifts significantly multiple times across their alignment (Figs 5d, e). High similarity is unlikely to result from local selective pressure, because the most similar portions of different sequence pairs do not coincide. (4) Strong statistical support for multiple sequence transfers, ranging from several hundred to several thousand base pairs, is obtained using GeneConv[24] (Fig. 5f and Supplementary Table S7). Thus, subtelomeric blocks on different chromosomes do not evolve independently; instead, continued interchromosomal interactions obfuscate their duplication history. Transfers are also likely to be prevalent among the many subtelomeric blocks that are >98% identical, but only more subtle haplotype analyses might detect these events[25].

Dynamics of primate subtelomeres

Recent changes in subtelomeric composition can be detected using fluorescence *in situ* hybridization (FISH) to determine the copy number and location of sequences in chromosomes of different primate species. Previous descriptions of subtelomeric dynamics using this approach were confounded by the use of probes encompassing several subtelomeric blocks, each with a different chromosomal distribution and evolutionary history[5,26]. To refine the analysis of structural changes in subtelomeres, we used four small FISH probes, each of which encompasses a single homology block. This approach reveals an unanticipated degree of recent genomic rearrangement in subtelomeres. Each block varies in copy number and chromosomal location between human individuals (Fig. 6), and FISH detects more chromosomal sites than are evident in the genome assembly or hybrid panel (Supplementary Table S8). We detect content variation at 14 chromosomal ends using just four block probes on three individuals. Further analyses would undoubtedly uncover more variation.

Gross structural polymorphism of human subtelomeres is also evident in the finished sequence of allelic pairs (Fig. 1). The two sequenced alleles of 16p are 99.8% identical in chromosome-specific DNA sequence, transition to much lower identity (~93%) within the adjoining block 17, and have no detectable homology in distal sequence (Supplementary Fig. S6). The 19p alleles also differ grossly in subtelomeric content (Fig. 1). One of the structurally variant 4q alleles (4qA) is found in association with facioscapulohumeral dystrophy[27]. Other cases of gross allelic variation are revealed by PCR analyses of the hybrid panel (Supplementary Fig. S1).

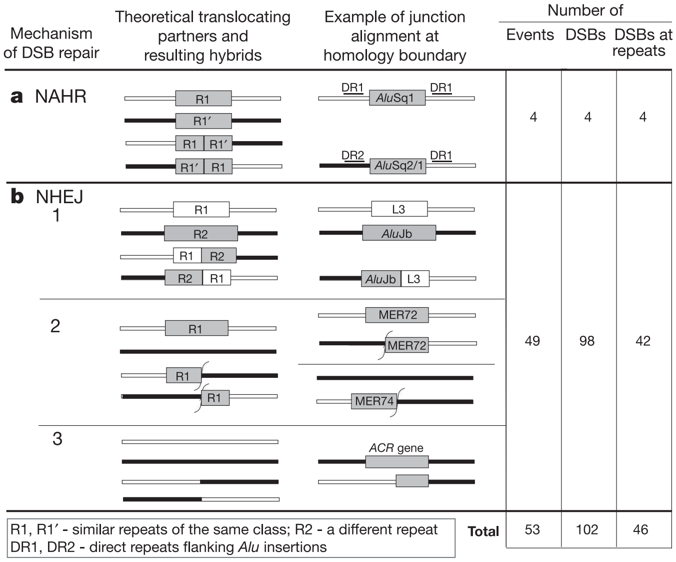

Mechanism of DSB repair
Theoretical translocating partners and resulting hybrids
Example of junction alignment at homology boundary
Number of
Events
DSBs
DSBs at repeats
a NAHR
R1
R1′
R1 R1′
R1′ R1
DR1
AluSq1
DR1
DR2
AluSq2/1
DR1
4
4
4
b NHEJ 1
R1
R2
R1 R2
R2 R1
L3
AluJb
AluJb L3
2
R1
R1
R1
MER72
MER72
MER74
49
98
42
3
ACR gene
R1, R1′ - similar repeats of the same class; R2 - a different repeat
DR1, DR2 - direct repeats flanking Alu insertions
Total
53
102
46

Figure 4 | Most subtelomeric homology breakpoints are consistent with NHEJ. a, b, For each mechanistic scenario, we show both original and derived forms, assuming reciprocal exchange. One derived form would be lacking in non-reciprocal cases. The third column gives a schematic example of each scenario identified in pairwise alignments of subtelomeric homology blocks. Of the complete, non-redundant set of 56 homology breakpoints, 53 were assigned a mechanistic scenario (details in Supplementary Table S5 and Fig. S3). In some cases, two originals and one hybrid were available for comparison (for example, NHEJ group 1). Other predicted states were not among surviving, sequenced alleles.

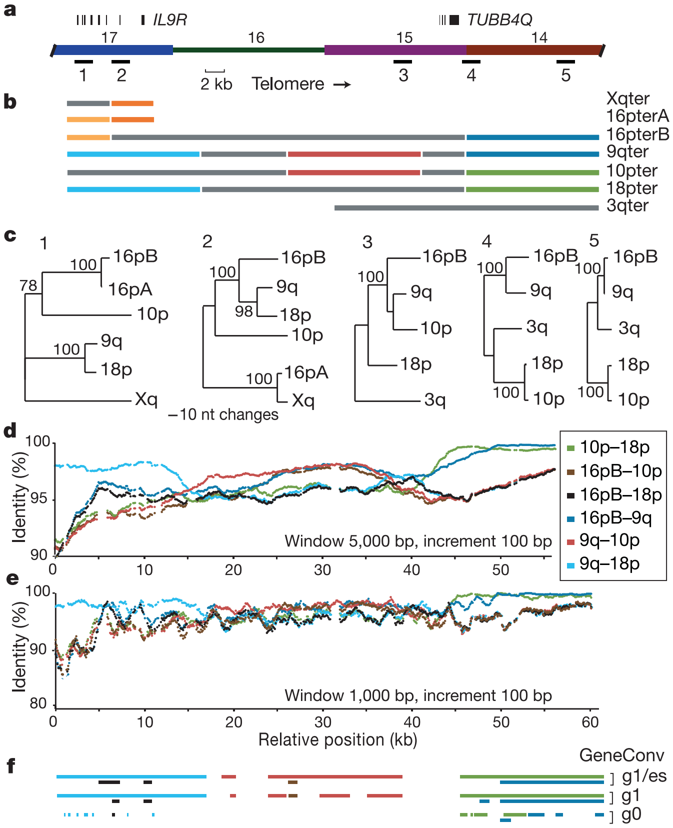
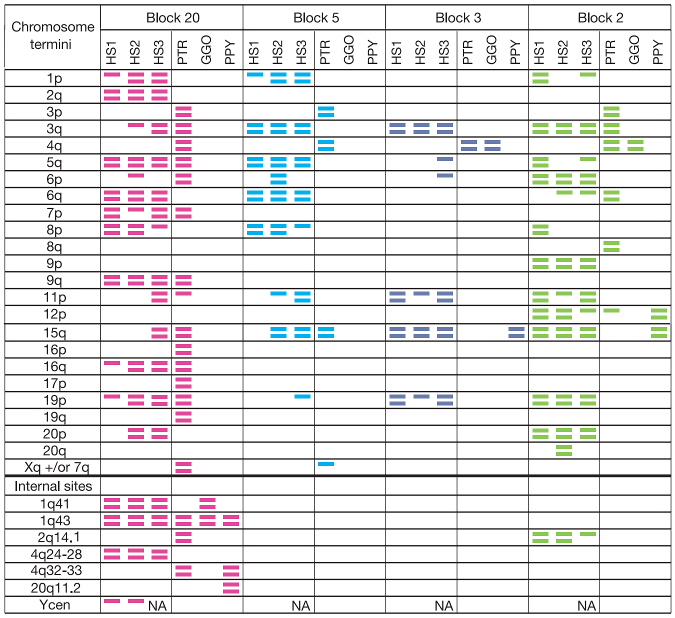


Figure 5 | Homology-based sequence transfers between subtelomeres. a, The region analysed encompasses four numbered blocks, two multi-exon genes, and five sequences sampled for phylogenetic analyses. **b,** Diagram of multiple sequence alignment with colours (excluding grey) indicating the best matching pairs with ≥98% identity in non-overlapping 5-kb windows. **c,** Neighbour-joining trees with bootstrap values (over 1,000 replicates) constructed from 2-kb samples of the alignments. **d, e,** Plot of per cent identity between four subtelomeres in 5-kb (**d**) and 1-kb (**e**) windows. Colours indicate alignments of different pairs. **f,** The same colours indicate transferred segments found statistically significant by GeneConv using different stringency parameters.

Subtelomeric dynamics are not confined to the human lineage. Blocks moved, and copies were lost and gained during primate evolution (Fig. 6). For example, block 20 is present at ≥9 subtelomeric locations in chimpanzee and human, whereas it occurs at only a few internal sites in gorilla and orangutan. The odorant-receptor-gene-containing block 5 was completely lost from the orangutan and gorilla genomes, yet is duplicated in chimpanzee and human genomes in four or more sites. The high similarity between sequenced human copies of these blocks (Fig. 1 and Supplementary Table S2), together with the fact that humans have more copies of three of the four blocks than other primates, argues that the diversity of these particular block distributions arose primarily by recent duplications, rather than by loss of different subsets of ancestral copies. We estimate that an average of 25 independent events, involving relocation or copy-number change of a total of ~1.2 Mb, can account for the observed differences between chimpanzee and human in the subtelomeric distribution of these blocks. FISH analyses also suggest that chimpanzees, like humans, show gross variation in subtelomeric content. Future, less anthropocentric analyses will probably reveal subtelomeric blocks that have been lost in humans but retained and perhaps duplicated in other primates.

Timing and rates of subtelomeric transfers

The very recent nature of the interchromosomal events shaping subtelomeres is apparent from the high similarity of paralogous blocks on different chromosomes and from our cytogenetic analyses. For 28 of the 41 blocks, even the most dissimilar copies exceed 97% identity (Fig. 1 and Supplementary Table S2). Assuming a mutation rate of 10^{-3} substitutions per site per million years[28] (Myr), all the copies of these 28 blocks must have formed by duplication (except the original one) or participated in homology-based sequence transfer in the past 15 Myr during the divergence of humans and great apes.



Chromosome termini	Block 20						Block 5						Block 3						Block 2					
	HS1	HS2	HS3	PTR	GGO	PPY	HS1	HS2	HS3	PTR	GGO	PPY	HS1	HS2	HS3	PTR	GGO	PPY	HS1	HS2	HS3	PTR	GGO	PPY
1p	–	=	=				–	=	=										=		–			
2q	=	=	=																					
3p				=						=												=		
3q		–	=	=			=	=	=				=	=	=				=	=	=	=		
4q				=						=						=	=					=	=	
5q	=	=	=	=			=	=	=						–				=		–			
6p		–		=				=							–				=	=	=			
6q	=	=	=				=	=	=											–	–	=		
7p	=	–	=	=																				
8p	=	=	–				=	=	–										=					
8q																						=		
9p																			=	=	=			
9q	=	=	=	=																				
11p			=	–				–	=				=	–	=				=	–	=			
12p																			=	=	–	–		=
15q			=	=				=	=	=			=	=	=			=	=	=	=			=
16p				=																				
16q	–	=	=	=																				
17p				=																				
19p	–	=	=	=					–				=	–	=				=	=	=			
19q				=																				
20p		=	=																=	=	=			
20q																				=				
Xq +/or 7q				=						–														
Internal sites																								
1q41	=	=	=		=																			
1q43	=	=	=	=	=	=																		
2q14.1				=															=	=	–			
4q24-28	=	=	=																					
4q32-33				=		=																		
20q11.2						=																		
Ycen	–	–	NA						NA						NA						NA			

Figure 6 | Chromosomal distribution of four subtelomeric blocks. FISH was conducted on three unrelated humans (HS1–3), chimpanzee (PTR), gorilla (GGO) and orangutan (PPY) (see Supplementary Methods). Coloured bars indicate sites at which FISH signals were consistently observed on both homologues (two bars) or only one homologue (one bar). Colours correspond to Fig. 1. Chromosome locations are given according to the human karyotype. No signal was observed for block 5 in gorilla and orangutan; its presence was also not detected by PCR (Supplementary Table S3). NA, not applicable.

When all pairs of human subtelomeric blocks are compared, the vast majority have 99.0–99.9% identity (Supplementary Fig. S7). Pairwise comparisons of all interchromosomal segmental duplications in the genome peak at ~98% identity (ref. 1), indicating that most subtelomeric segmental duplications result from more recent events than other segmental duplications. Indeed, after correcting for redundancy, we find that subtelomeres account for 40% of all duplications in the latest genome assembly[3] that have a match of ≥98.7% identity on another chromosome. Remarkably, ~1 Mb (40%) of known subtelomeric terrain has a paralogous match of ≥99.5% identity, often rivaling the similarity of allelic copies.

We estimate conservatively that 49% (1.13 Mb) of known subtelomeric sequence was generated after humans and chimpanzees diverged (Supplementary Fig. S8). This amount equates to an observed rate of subtelomeric interchromosomal sequence duplication and/or transfer during the last 6.5 Myr of ~0.075 bases per site per Myr (Supplementary Table S9). We estimate from our cytogenetic analyses of the four blocks in Fig. 6 that these subtelomeric sequences relocated or changed in copy number at a rate of ~0.09 bases per site per Myr during the same time interval. The sequence- and cytogenetic-based estimation methods capture slightly different aspects of subtelomeric dynamics and underestimate the true rates of interchromosomal sequence transfer. Nevertheless, both estimates yield rates >60-fold higher than those of point mutations[28] or bases added by retrotransposon insertion[29] over the same evolutionary period.

Given the amount of new subtelomeric sequence apparently created during the past 5 Myr (1.0 Mb), we estimate that ~7 gene duplicates arose in human subtelomeres per Myr in recent times (Supplementary Table S9). Even if half of these genes are deceptively young owing to sequence transfers between pre-existing copies, the rate of gene duplication in subtelomeres (0.04 duplicates per gene per Myr) is fourfold higher than the genome-wide average[30]. The rate of gene creation in subtelomeres is matched only by that in pericentromeric regions, which, like subtelomeres, are hotbeds of segmental duplications[31].

Discussion

Here we demonstrate that a multitude of predominantly NHEJ-mediated translocations led to a complex patchwork of segmental duplications in human subtelomeres that exchange sequence at a remarkably high rate. The extraordinary recent dynamics of subtelomeres complicate the description of the human genomic landscape and its variation. Perhaps no chromosome-specific marker or block organization exists within subtelomeres, as they appear to evolve as a pool of variant allelic and paralogous structures. Moreover, inter-allelic subtelomeric recombination rates may be impossible to quantify owing to the high frequency of interchromosomal transfers.

Why are subtelomeres so plastic? Deviations in copy number of subtelomeric DNA might be better tolerated than segmental aneuploidy of other genomic regions. Furthermore, subtelomeres might be more susceptible to DSBs and/or more readily repaired through interchromosomal interactions than other regions[32]. Telomere clustering in meiotic cells[33] might favour exchange of chromosome ends during DSB healing. Gross allelic differences probably make some subtelomeres prone to mispairing at meiosis, catalysing further change.

Subtelomeric rearrangements might not be restricted to the germline, but could also arise in somatic cells during repair of DSBs or eroded telomeres. The resulting genotypic heterogeneity might affect fitness at the cellular and/or individual levels. Indeed, subtelomeres coalesce with telomeres in DNA-repair foci in naturally senescent cells[34,35] and in cells with artificially induced telomere dysfunction[36]. Furthermore, the high level of apparent sister chromatid exchanges observed at chromosome ends (10^{-2} per Mb per generation)[37] signals a high DSB rate and could subsume interchromosomal subtelomeric exchanges.

The results of ectopic repair of subtelomeres could have advantages beyond the healing of damaged ends. With their propensity to duplicate and exchange, subtelomeres could serve as a nursery for new genes and a place where haplotypes can diversify faster than in single-copy genomic regions. Sequence transfer between paralogous genes (as in Fig. 5) has the potential to create advantageous new combinations of sequence variants, aiding adaptive peak shifts[38]. Indeed, subtelomeric genes are associated with adaptive processes in other organisms (refs 39–41 and citations in ref. 5). However, subtelomeric dynamics are a double-edged sword. Some DSB repair events could result in loss or gain of dosage-sensitive genes in the most distal single-copy DNA, or in contextual changes with adverse effects on gene regulation. The sequence analyses presented here contribute to a developing framework that will enable exploration of the roles of subtelomeric dynamics in normal variation, adaptive change and clinically manifested disorders.

The translocation-based model developed here to explain subtelomeric segmental duplications could be broadly applicable to other interchromosomal segmental duplications (Fig. 2). The first step in this model is a reciprocal translocation (Fig. 2a), which arises *de novo* in ~1/2,000 concepti[42]. One in 500 healthy individuals carries a cytogenetically visible, balanced translocation[43]. A second interchromosomal exchange between the translocation derivatives (Fig. 2b) is likely to be selectively favoured if it reduces the risk of passing a grossly imbalanced chromosomal complement to gametes. Duplicated segments, particularly when present on just one allele, can in turn promote translocations through NAHR (Fig. 2d). Furthermore, a DSB occurring in a hemizygous region stemming from an unbalanced translocation has increased probability of causing another translocation, inversion or intrachromosomal deletion, owing to the absence of a homologous template for its repair (Fig. 2e). Thus, segmental polymorphisms predispose to further rearrangements, which in turn lead to new segmentally polymorphic structures. This cycle of segmental polymorphism and gross genomic rearrangement is particularly obvious in subtelomeres and could underlie structural variation[44–46] and genomic disorders[4] arising at many other locations in the human genome.

METHODS

Additional results and methodological details, including the basis for all rate calculations, are provided as Supplementary Methods.

Sequence collation and analysis. Details of the iterative search for finished subtelomeric sequences are provided in the Supplementary Methods. Sequences with continuous overlap of >99.8% nucleotide identity were merged into contigs (Supplementary Table S1) and assumed to represent the same genomic region or an allelic variant. We used a combination of approaches to establish or verify the chromosome location of contigs (Supplementary Table S1), including PCR of a monochromosomal hybrid panel (Supplementary Table S3), FISH (Supplementary Table S10), and matches to half-YAC (yeast artificial chromosome) vector-insert junction sequences[10] (Supplementary Table S11). Regions of similarity were identified from pairwise sequence alignments made by BLAST2[47], without masking repeats. Blocks of paralogy were delineated when one or more contigs showed a break in homology, except where paralogy adjoined a gap in available sequence. Block colour/number are changed in Fig. 1 if similarity is lost on one or more subtelomeres, except when loss of homology occurs within 3 kb of another breakpoint. However, all breakpoints were evaluated for mechanistic signatures (see below). Blocks from different chromosomal contigs were aligned using cross_match (http://www.phrap.org/) and MAVID[48]. Per cent identities of block copies were calculated without insertions or deletions and with Jukes-Cantor correction for multiple substitutions. From 1,438 alignments (26.8 Mb total aligned sequence), a best-matching partner was identified for each block in each chromosomal contig (Supplementary Fig. S7). To remove redundancy in cases of reciprocal best matches, only one of the two alignments was included in the estimation of the amount of recently generated sequence (see Supplementary Methods and Supplementary Table S9). We also calculated the sum of non-overlapping interchromosomally duplicated bases with paralogous match of ≥98.7% in subtelomeres or elsewhere in the latest genome assembly (Build 35), as outlined in Supplementary Methods.

Subtelomeric block analysis by PCR and FISH. The subtelomeric content of 24 individual human chromosomes isolated in a hybrid panel was analysed by PCR using 160 primer pairs (Supplementary Fig. S1 and Table S3). FISH was performed as detailed in the Supplementary Methods, using block-specific probes generated by long-range PCR (blocks 20, 5 and 2) or cosmid f7501 (block 3; ref. 6) on primary cultures of three unrelated Caucasians (2 males and one female) and cell lines of male chimpanzee, orangutan and gorilla. The assumptions used to conservatively estimate the rate with which these blocks changed copy number or location since the divergence of humans and chimpanzees are given in Supplementary Methods. Note that this rate excludes homology-based sequence transfers among pre-existing copies, whereas the sequence-based estimate includes duplications and homology-based sequence transfers, but not changes in segment location.

Breakpoint analyses. We identified homology breakpoints from all pairwise subtelomeric sequence alignments and evaluated a nonredundant set for mechanistic signatures as described in the Supplementary Methods (Supplementary Table S5). All remaining block junctions in Fig. 1 are nearly identical replicas of members of this junction set, owing to their duplication within larger segments. The number of independent DSBs was counted as two for each deduced NHEJ event and one for each NAHR event in the non-redundant set. We queried the human genome by BLAT with the 200 bp surrounding each NHEJ breakpoint lacking a gene or known repeat, and we found no novel repeats.

Detection of homology-based transfer. Changes in per cent identity along pairwise sequence alignments were determined using the percentIDplot program (E. M. W. and E. V. L, unpublished data). The best-matching pair in each 5-kb and 2-kb window in each sequence was identified from a multiple sequence alignment generated using MAVID[48] (Supplementary Fig. S5). Phylogenetic trees were constructed using PAUP[49].

Received 7 April; accepted 5 July 2005.

1. Samonte, R. V. & Eichler, E. E. Segmental duplications and the evolution of the primate genome. *Nature Rev. Genet.* **3,** 65–72 (2002).
2. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
3. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
4. Shaw, C. J. & Lupski, J. R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13** (review issue 1), R57–R64 (2004).
5. Mefford, H. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* **3,** 91–102 (2002).
6. Trask, B. J. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7,** 13–26 (1998).
7. Monfouilloux, S. *et al.* Recent human-specific spreading of a subtelomeric domain. *Genomics* **51,** 165–176 (1998).
8. Martin, C. L. *et al.* The evolutionary origin of human subtelomeric homologies—or where the ends begin. *Am. J. Hum. Genet.* **70,** 972–984 (2002).
9. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E. M. & Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1. *Genome Res.* **12,** 1651–1662 (2002).
10. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409,** 948–951 (2001).
11. Linardopoulou, E. *et al.* Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location. *Hum. Mol. Genet.* **10,** 2373–2383 (2001).
12. Knight, S. J. & Flint, J. The use of subtelomeric probes to study mental retardation. *Methods Cell Biol.* **75,** 799–831 (2004).
13. Brown, W. R. *et al.* Structure and polymorphism of human telomere-associated DNA. *Cell* **63,** 119–132 (1990).
14. de Lange, T. *et al.* Structure and variability of human chromosome ends. *Mol. Cell. Biol.* **10,** 518–527 (1990).
15. Flint, J. *et al.* Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum. Mol. Genet.* **6,** 1305–1313 (1997).
16. Riethman, H. *et al.* Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.* **14,** 18–28 (2004).
17. Smit, A. F. & Riggs, A. D. Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA* **93,** 1443–1448 (1996).
18. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9,** 1048–1058 (1999).
19. Pfeiffer, P., Goedecke, W. & Obe, G. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis* **15,** 289–302 (2000).
20. Rothkamm, K., Kruger, I., Thompson, L. H. & Lobrich, M. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Mol. Cell. Biol.* **23,** 5706–5715 (2003).
21. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin

and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73,** 823–834 (2003).

22. Zhou, Y. & Mishra, B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc. Natl Acad. Sci. USA* **102,** 4151–4156 (2005).
23. Nergadze, S. G., Rocchi, M., Azzalin, C. M., Mondello, C. & Giulotto, E. Insertion of telomeric repeats at intrachromosomal break sites during primate evolution. *Genome Res.* **14,** 1704–1710 (2004).
24. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6,** 526–538 (1989).
25. Mefford, H. C., Linardopoulou, E., Coil, D., van den Engh, G. & Trask, B. J. Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Hum. Mol. Genet.* **10,** 2363–2372 (2001).
26. Der-Sarkissian, H., Vergnaud, G., Borde, Y. M., Thomas, G. & Londono-Vallejo, J. A. Segmental polymorphisms in the proterminal regions of a subset of human chromosomes. *Genome Res.* **12,** 1673–1678 (2002).
27. Lemmers, R. J. *et al.* Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nature Genet.* **32,** 235–236 (2002).
28. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68,** 444–456 (2001).
29. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13,** 358–368 (2003).
30. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
31. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430,** 857–864 (2004).
32. Ricchetti, M., Dujon, B. & Fairhead, C. Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast. *J. Mol. Biol.* **328,** 847–862 (2003).
33. Bass, H. W. Telomere dynamics unique to meiotic prophase: formation and significance of the bouquet. *Cell. Mol. Life Sci.* **60,** 2319–2324 (2003).
34. d'Adda di Fagagna, F. *et al.* A DNA damage checkpoint response in telomere-initiated senescence. *Nature* **426,** 194–198 (2003).
35. Zou, Y., Sfeir, A., Gryaznov, S. M., Shay, J. W. & Wright, W. E. Does a sentinel or a subset of short telomeres determine replicative senescence? *Mol. Biol. Cell* **15,** 3709–3718 (2004).
36. Takai, H., Smogorzewska, A. & de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr. Biol.* **13,** 1549–1556 (2003).
37. Cornforth, M. N. & Eberle, R. L. Termini of human chromosomes display elevated rates of mitotic recombination. *Mutagenesis* **16,** 85–89 (2001).
38. Hansen, T. F., Carter, A. J. & Chiu, C. H. Gene conversion may aid adaptive peak shifts. *J. Theor. Biol.* **207,** 495–511 (2000).
39. Halme, A., Bumgarner, S., Styles, C. & Fink, G. R. Genetic and epigenetic regulation of the *FLO* gene family generates cell-surface variation in yeast. *Cell* **116,** 405–415 (2004).
40. Fabre, E. *et al.* Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol. Biol. Evol.* **22,** 856–873 (2005).
41. De Las Penas, A. *et al.* Virulence-related surface glycoproteins in the yeast pathogen *Candida glabrata* are encoded in subtelomeric clusters and subject to *RAP1*- and *SIR*-dependent transcriptional silencing. *Genes Dev.* **17,** 2245–2258 (2003).
42. Warburton, D. *De novo* balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* **49,** 995–1013 (1991).
43. Genetics and Public Policy Center. Genetics Information: Translocations. ⟨http://www.dnapolicy.org/genetics/translocations.jhtml⟩ (2004).
44. Wong, Z., Royle, N. J. & Jeffreys, A. J. A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7,** 222–234 (1990).
45. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
46. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
47. Tatusova, T. A. & Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174,** 247–250 (1999).
48. Bray, N. & Pachter, L. MAVID multiple alignment server. *Nucleic Acids Res.* **31,** 3525–3526 (2003).
49. Swofford, D. L. *PAUP**. Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer Associates, Sunderland, 2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the many contributors to the Human Genome Project who generated the sequences that made this study possible, and to the Eichler and Haussler laboratory groups for making data on segmental duplications readily accessible. Our work was supported by the NIH. We thank E. Eichler, H. Malik, D. Gottschling, J. Gogarten, K. Rudd and M. Schlador for comments on the manuscript, and J. Felsenstein for advice on phylogenetic analyses.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.J.T. (btrask@fhcrc.org).