European Journal of Human Genetics (2006) 14, 216–221
© 2006 Nature Publishing Group All rights reserved 1018-4813/06 $30.00
www.nature.com/ejhg

ARTICLE

Segmental duplication density decrease with distance to human-mouse breaks of synteny

Jesus Sainz*,[1], Pavol Rovensky[1], Sigurjon A. Gudjonsson[1], Gudmar Thorleifsson[1], Kari Stefansson[1,2] and Jeffrey R. Gulcher[1,2,]*

[1]*deCODE Genetics, Reykjavik, Iceland*

Segmental duplications are large genomic segments of recent origin and nearly identical sequence. Segmental duplications account for up to 5% of the human genome and they are often involved in genomic rearrangements and human disease. We developed a rapid computational method to characterize segmental duplications in the mouse and the human genomes according to four sequence assemblies for each species. Segmental duplication content in the mouse genome assemblies has largely changed over the four releases (from 0.2 to1.2%, 4.5 and 3.0%), while in the four human assemblies duplication content was 4.8, 3.5, 3.7 and 3.7%, respectively. This suggests that cataloguing and assembling duplications has been challenging in both genomes and any interpretation of comparative analyses of duplication content must keep this in perspective to avoid artifacts. Human and mouse segmental duplications are more frequent than expected in regions where there is a syntenic discontinuity and the duplication content in syntenic regions decreases significantly with distance from breakpoints of synteny. These observations indicate that in mouse and human the frequency of segmental duplications is strongly correlated with distance to human and mouse syntenic breaks or the most dynamic regions in evolution.

European Journal of Human Genetics (2006) **14**, 216–221. doi:10.1038/sj.ejhg.5201534; published online 23 November 2005

Keywords: segmental duplications; synteny; evolution

Introduction

For some time, whole-genome duplications have been proposed as a model of evolution.[1] More recently, segmental duplications have been shown to represent a large proportion of the human genome (up to 5% for some analysis) and have been characterized as an important feature of genome organization.[2–6] There is much evidence that implicate segmental duplications as one of the molecular mechanisms that lead to human diseases[7–11] and that associate duplicated segments with genome evolution.[12–14] Recent findings of a high rate of gene conversion in palindromic sequences of human and ape chromosome Y suggest that gene conversion is a more frequent event than previously suspected, particularly in palindromic and duplicated sequences.[15,16] This rapidly increasing body of data supports the notion that duplications play a very important role in the dynamics of genomic change.

Human chromosome 19 was the first chromosome analyzed with respect to duplications and human–mouse synteny and showed colocalization of duplicated genes and some breakpoints of synteny.[17] The first genome-wide analysis of duplications and human–mouse synteny compared human genome assembly build 30 with the mouse assembly MGSCv3.[18] The analysis showed dramatic enrichment of duplications at breakpoints of synteny in the human genome itself. They also analyzed the mouse X

*Correspondence: Dr J Sainz or Dr JR Gulcher, deCODE Genetics, Sturlugotu 8, IS-101 Reykjavik, Iceland. Tel: +354 570 1946;
Fax: +354 570 1903;
E-mail: sainz@decode.is or Jeffrey.gulcher@decode.is

[2]These authors contributed equally to this work

Received 4 January 2005; revised 29 August 2005; accepted 27 September 2005; published online 23 November 2005

chromosome from build 30 and confirmed that there was an increase in duplication but did not provide data for the mouse autosomes. Our study extends the previous studies by (1) using more updated sequence assemblies for mouse and human, (2) comparing four successive assemblies for each to ensure that the observations were not artifacts of sequence assembly errors, and (3) defining enrichment of duplications in relation to distance to the breaks of synteny.

Materials and methods

Genomic assemblies and synteny

All genomic sequence assemblies and syntenic maps were downloaded from the University of California at Santa Cruz (UCSC) Genome Bioinformatics. Pericentromeric regions were defined by adding 1 Mb to each side of the intervals annotated by UCSC Genome Bioinformatics as centromeric and their adjacent heterochromatic intervals in the 'chromAgp' files. Telomeric regions were defined as the 2 Mb at the beginning (except in the acrocentric chromosomes) and the 2 Mb at the end of the chromosomal sequences.

Detection of segmental duplications

We have developed an annotation package of programs and scripts to detect genomic segmental duplications. The package utilizes BLAT[19] as the alignment algorithm and allows the construction of segmental duplication detection in large genomes in a quick and efficient manner. The simplest model to detect segmental duplications would be to align the whole genome against itself but the current algorithms and computer memory available are limited to a much smaller query sequence sizes. Hence, we decided to divide the task into steps, which can be executed with moderate computing power using generally accepted tools and algorithms for alignment. The package cuts the genomic sequence into 1 kb consecutive segments, executes the alignment algorithm and analyzes the resulting alignments to define the duplications. We selected BLAT as the alignment program to use because it is fast and has the advantage that the masking of the DNA for high copy repeats is not necessary.

Analysis of the alignments to detect duplications

The identity of alignment is defined as 100(matches/query size) (%). Query size is 1 kb. Alignments that are not self-hits and have a minimum identity of 90% are selected. Only blocks of the alignment that are located within a 1000 bp window of the target are used. A recursive procedure assembles the duplicated genomic intervals for all queries using the following criteria:

a. The minimum length of the duplication is 5000 bp.
b. The total length of the gaps (in the target or the query) is less than 5% of the duplication.

Post-processing

Duplication intervals overlapping a minimum of 90% of their length with a single high-copy repetitive element are considered false positives and removed. A duplication interval is defined as an interval created by the set of overlapping or adjacent query sequences that constitute the duplication map. Duplications are defined as a query interval that has homology with a corresponding target interval. All the analyses in this manuscript use data from the query interval.

Statistical methods

We tested the significance of the observed difference in the number of direct *versus* inverted duplications and in the number of intrachromosomal *versus* interchromosomal duplications using a χ^2 test. The difference in the average identity and average size of inter- *versus* intrachromosomal duplication intervals was tested using a randomization procedure, that is, randomly interchanging elements in the two groups. This procedure was repeated 100 000 times and the *P*-value calculated as the fraction of randomization tests that yielded a difference in the average identity or average size that was equal or greater to the observed differences. All *P*-values presented are two-sided.

Results

Human segmental duplications

The criterion used to construct the duplication maps were set to a minimum of 5 kb duplication and 90% identity for each of the 1 kb fragments. The human segmental duplication maps were constructed using the National Center for Biotechnology Information (NCBI) build 31, NCBI build 33, NCBI build 34 and NCBI build 35 genomic assemblies. We will refer from now to the latest assembly, build 35, unless otherwise stated. Using an identity threshold of 90%, we detected segmental duplications in 3119 human genomic intervals spanning 105.250 Mb (3.7% of the genome) with an average identity of 96.5% (Table 1). These intervals can include one or more segmental duplications.

The largest duplicated interval in human spans 1.5 Mb on chromosome Y and the average duplicated interval is 33.7 kb. Chromosomal distribution of duplications is not uniform, with chromosome Y and 9 having the greatest duplication content (53.0 and 9.0%, respectively), and chromosome 3 having the least duplication content (0.7%) (Figure 1a). Among the intrachromosomal duplications, inverted duplications are more frequent than direct duplications (54 *versus* 46%, $P = 1.3 \times 10^{-13}$) in build 35. The ratio of inverted duplications *versus* direct duplications has increased in the finished assemblies, from 0.8 in build 31 and 1.0 in build 33 to 1.2 in build 34 and build 35 (Table 1).

Table 1 Segmental duplications in four human genomic assemblies

Human duplications	*Build31*	*Build33*	*Build34*	*Build35*
Duplicated genome length (bp)	136 600 000	99 831 000	104 407 000	105 250 000
Duplicated genome fraction (%)	4.80	3.52	3.67	3.69
Number of duplicons	4486	2875	2960	3119
Average identity (%)	96.60	96.30	96.34	96.51
Average identity of intrachromosomal (%)	97.68	97.24	97.33	97.50
Average identity of interchromosomal (%)	95.45	95.42	95.50	95.61
Ratio intra/inter chromosomal (length)	2.52	1.96	1.77	1.74
Ratio inverted/direct duplications (number)	0.83	0.98	1.18	1.18
Largest duplicon length (bp)	2 829 000	3 028 000	1 526 000	1 526 000
Average duplicon length (bp)	30 450	34 723	35 272	33 744
Median duplicon length (bp)	12 000	13 000	12 000	12 000
Average intrachromosomal duplicon length (bp)	30 720	35 988	35 084	33 500
Average interchromosomal duplicon length (bp)	22 586	23 102	24 203	24 163

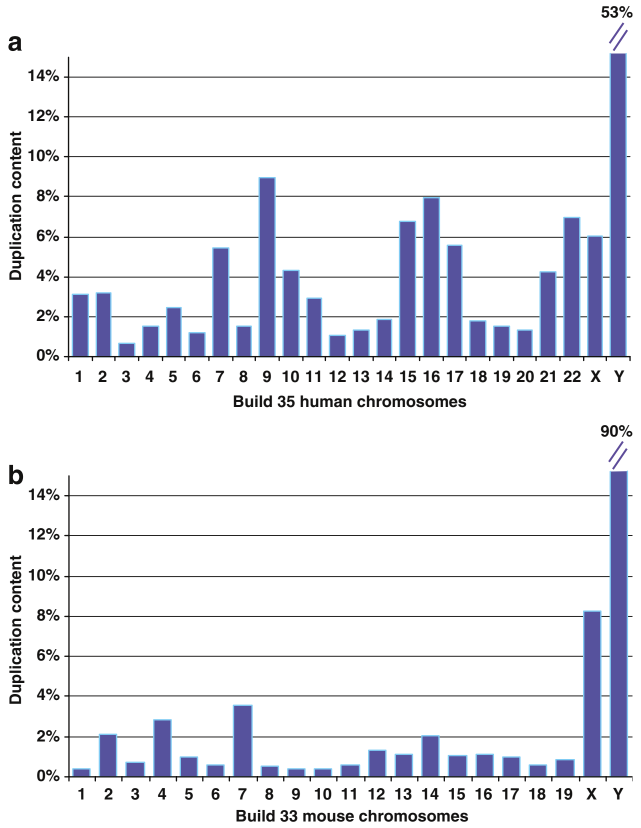

Figure 1 Duplication content of mouse and human genomes. (**a**) Human duplication content per chromosome. (**b**) Mouse duplication content per chromosome.

Intrachromosomal duplications are 1.7 times more abundant than interchromosomal duplications ($P<10^{-100}$), have a higher average identity (97.5% compared to 95.6%, $P<0.00001$), and have a larger average size (33.5 Kb compared to 24.2 Kb, $P<0.00001$). The ratio of intrachromosomal duplication *versus* interchromosomal duplication content has diminished in the consecutive assemblies (Table 1).

Mouse segmental duplications

The segmental duplication maps were constructed using the mouse assemblies MGSCv3, NCBI build 30, NCBI build 32 and NCBI build 33. We will refer from now on to the latest mouse assembly, build 33, unless indicated otherwise. The mouse duplications span a total of 74.2 Mb or 2.98% of mouse genome. The number of duplicated intervals is 2807 and their average size is 26.4 kb with an average identity of 97.3%. The largest duplication spans 564 kb on chromosome Y.

As in the human, chromosomal distribution of duplications in the mouse is not uniform, ranging from 0.38% on chromosome 10 to 8.2% on chromosome X and 90.2% on chromosome Y (Figure 1b), inverted duplications are more frequent than direct duplications (51.1 and 48.9%, $P=3.4\times10^{-10}$) and the ratio of inverted *versus* direct duplications has increased in the two latest assemblies (Table 2).

As in human, intrachromosomal duplications are more abundant (6.45-fold more) than interchromosomal duplications ($P<10^{-100}$), have a higher average identity (97.3% compared to 96.9%, $P<0.00001$), and the intervals containing intrachromosomal duplications have a larger average size (30.6 *versus* 9.2 kb, $P<0.00001$). Intrachromosomal duplications are more abundant and their identity higher than interchromosomal duplications in all mouse assemblies but the ratio has fluctuated greatly (Table 2).

Analysis of mouse and human duplications show no sequence similarity as would be expected for duplications that have arisen since the lineage of these species separated 75 million years ago.

Synteny and segmental duplications

Long-range chromosomal organization in human and mouse is known to be similar and conserved syntenic

Table 2 Segmental duplications in four mouse assemblies

Mouse duplications	*MGSCv3*	*Build 30*	*Build 32*	*Build 33*
Duplicated genome length (bp)	4 396 000	28 648 000	111 526 000	74 200 000
Duplicated genome fraction (%)	0.18	1.19	4.47	2.98
Number of duplicons	633	1879	5539	2807
Average identity (%)	96.39	97.07	97.16	97.26
Average identity of intrachromosomal (%)	96.41	97.86	97.53	97.31
Average identity of interchromosomal (%)	96.39	96.85	96.68	96.86
Ratio intra/inter chromosomal (length)	1.23	5.86	2.85	6.45
Ratio inverted/direct duplications (number)	0.60	0.37	0.94	1.05
Largest duplicon length (bp)	114 000	228 000	430 000	564 000
Average duplicon length (bp)	6944	15 246	20 134	26 433
Median duplicon length (bp)	6000	8000	10 000	8000
Average intrachromosomal duplicon length (bp)	7374	17 315	21 327	30 620
Average interchromosomal duplicon length (bp)	6282	7170	16 238	9236

regions have been reported.[2,3,20,21] To further investigate a possible role of segmental duplications in chromosomal rearrangement during speciation, we placed the locations of segmental duplications on human/mouse synteny maps generated by the UCSC. The latest syntenic map has high resolution and identifies 1111 regions of human synteny with mouse that account for 92.4% of the human sequence (here called 'syntenic genome'). The nonsyntenic intervals (here called 'nonsyntenic genome') account for 7.6% of the human sequence. Human duplications are 7.2 times more frequent than expected by chance in genomic regions nonsyntenic with mouse: 55.1% of duplicated sequences locate in the 7.6% nonsyntenic genome. The human nonsyntenic genome spans 218.6 Mb and 26.6% of it (58.0 Mb) is comprised of segmental duplications. In contrast the syntenic genome spans 2650 Mb with segmental duplications making up only 1.8% (47.2 Mb). However, analysis of the syntenic 50 kb adjacent to the breaks of synteny indicates that a higher fraction is duplicated: 17.5 Mb of a total of 87.2 Mb (20.1%). This indicates that not only the nonsyntenic regions but also the adjacent syntenic regions have a much higher frequency of duplications than the rest of the genome.

To determine if the duplication content is related to the distance from the junction of syntenic discontinuity, we analyzed genomic intervals on both sides of all junctions in the genome. The duplicated fraction of the syntenic genome at a distance to the junction greater than 100 kb, from 100 to 50 kb, from 50 to 20 kb and from 20 to 0 kb is 1.0, 8.8, 15.6 and 25.0%, respectively. Duplication content has a strong negative correlation with the distance to the breaks of synteny in the four intervals analyzed (distance as the median for each class of interval; correlation coefficient = −0.82). Duplication content of the nonsyntenic intervals from 0 to 20 kb and >20 kb to the synteny junction is 24.7 and 27.0%, respectively. These results show that the duplication content increases with the proximity to the nonsyntenic regions and reaches maximum values in the internal part of the nonsyntenic region (Figure 2). The higher duplication content in the syntenic genome near the break of synteny junctions is observed for all the human chromosomes with the only exception of chromosome Y.

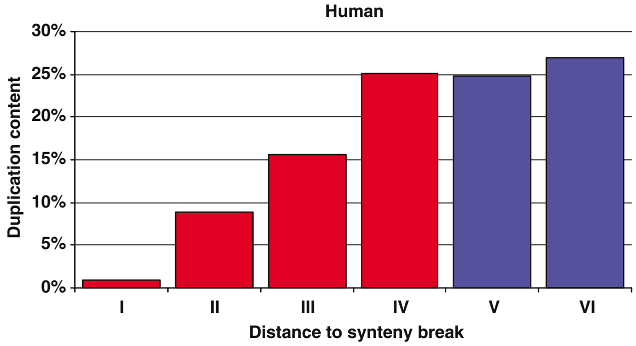

Figure 2 Duplication content in human build 35 according to the distance to synteny breaks with mouse. Blue bars represent nonsyntenic genome and red bars syntenic genome. Bar numbers: I. In synteny and at more than 100 kb to the junction. II. In synteny and from 100 to 50 kb to the junction. III. In synteny and from 50 to 20 kb to the junction. IV. In synteny and from 20 to 0 kb to the junction. V. In nonsynteny and from 0 to 20 kb to the junction. VI. In nonsynteny and at more than 20 kb to the junction.

Analysis of the mouse synteny show, as in human, that duplications are 8.0 times more abundant than expected in sequences that are nonsyntenic with human: 63.4% of the duplications are in the 7.9% nonsyntenic mouse genome. The mouse nonsyntenic genome spans 202.9 Mb and contains 47.1 Mb of segmental duplications or 23.2% of its length while the syntenic genome spans 2366 Mb with 27.1 Mb of segmental duplications or 1.1% of the length. As in human, duplication density in mouse decreases with distance to the breaks of synteny. For the syntenic intervals at greater than 100 kb, from 100 to 50 kb, from 50 to 20 kb and from 20 to 0 kb, the duplication fraction is 1.0, 2.2, 3.5 and 4.9%, respectively. As in human, the duplication

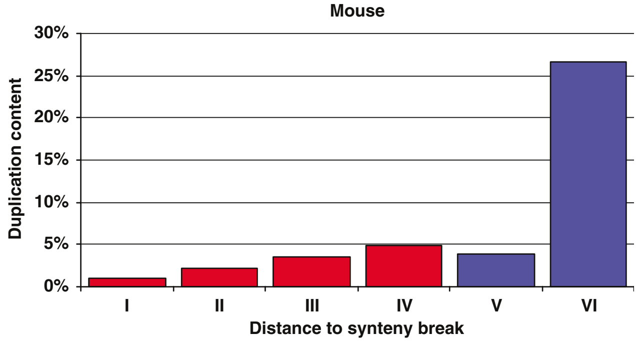

Figure 3 Duplication content in mouse build 33 according to the distance to synteny breaks with human. Blue bars represent nonsyntenic genome and red bars syntenic genome. Bar numbers: I. In synteny and at more than 100 kb to the junction. II. In synteny and from 100 to 50 kb to the junction. III. In synteny and from 50 to 20 kb to the junction. IV. In synteny and from 20 to 0 kb to the junction. V. In nonsynteny and from 0 to 20 kb to the junction. VI. In nonsynteny and at more than 20 kb to the junction.

content has a strong negative correlation with the distance to breaks of synteny of the four intervals analyzed (distance as the median for each class of interval; correlation coefficient = −0.76). For the nonsyntenic intervals from 0 to 20 kb and >20 kb the duplicated fraction is 3.8 and 26.7%, respectively (Figure 3). In mouse, we observe a much greater density of duplications in the nonsyntenic intervals at a distance of more than 20 kb of the syntenic regions than the density in the 20 kb intervals adjacent to the synteny/nonsynteny boundaries (Figure 3). This difference is not observed in human (Figure 2). It is not clear to us the reason for this difference.

We asked the question whether duplications in nonsyntenic intervals were clustered in pericentromeric or telomeric regions and we found that this was not the case: the percentage of all segmental duplications detected in nonsyntenic regions when we exclude the pericentromeric or telomeric regions in the analysis (51% in human and 62% in mouse) is not very different from the percentage detected including these regions (55% in human and 63% in mouse).

We found large differences in segmental duplication content between the four mouse sequence assemblies analyzed. Even with the large range of variation in segmental duplication content for the mouse assemblies, duplication content in synteny breaks higher than expected by a random distribution was found in all mouse assemblies. Thus, duplications correlate with syntenic breaks. These results confirm previous studies that showed colocalization of duplicated genes and breakpoints of synteny, specifically in human chromosome 19[17] and in the human genome in general by comparing human assembly build 30 and mouse assembly build 30.[18] Our study shows that these observations hold up even in the most recent sequence assemblies and when performed at high resolution. We furthermore found that the duplication prevalence shows a gradient that peaks at the breaks in synteny.

Discussion

A possible caveat could be that a large fraction of the breaks of synteny used might have been generated mainly by lack of sequence alignment between both genomes and do not reflect rearrangements originated in the evolution from an ancestral genome. We believe that our evidence supports that this is not the case. We analyzed the breaks of synteny and in at least 85% of the cases they corresponded to clear genomic rearrangements caused by breaks joining different chromosomes, by intrachromosomal inversions, or by intrachromosomal translocations. Restricting the analysis to the breaks of synteny clearly originated by genomic rearrangements indicates that their duplication content is similarly high than the one observed analyzing all breaks of synteny (26.9 *versus* 26.6% for all breaks in human and 22.0 *versus* 23.2% in mouse).

One may argue that the human genome that appears to be nonsyntenic to mouse, represents parts of the mouse genome yet to be sequenced or vice versa and that segmental duplications are more difficult to sequence. However, in our analyses of the four human genome sequence assemblies, we observe consistent results. Therefore, even if there are limitations in the mouse sequence assemblies that create collapses of real duplications and pseudo-duplications or other errors, it would be hard to support that these limitations generate the striking correlation to syntenic discontinuities, particularly when in the most recent and perfected assemblies the correlation becomes stronger.

We observe that even with large changes in duplication content from previous mouse assemblies to the latest one, the duplication content is always higher than expected at the breaks of synteny. In the latest mouse assembly (build 33), the duplication characteristics are more consistent with the ones in the finished human genome than it was in older mouse assemblies. The latest mouse assembly has a similar percentage of duplicated genome to that of the human, a higher content of duplications in the regions were the synteny breaks as in human and a higher duplication content near the breaks of synteny as in human. Also the average identity of interchromosomal duplications and intrachromosomal duplications is similar in the last mouse assembly and in the last human assembly. It is important to note that mouse build 32 and build 33 used a clone-based tiling path file method for the assembly, a more robust method to prevent the collapse of duplications than the whole genome shotgun sequence method used in the previous assemblies.[22]

We observe that inverted duplications are more frequent than direct duplications in human build 34 and build 35 while they are similarly frequent in build 33 and they are less frequent in build 31. The increase in the ratio inverted/direct duplications is also observed in the last three mouse assemblies.

A common feature in all human and mouse assemblies analyzed is that intrachromosomal duplications had a higher identity than interchromosomal duplications. In the current assemblies the identity is 1.9% higher in human and 0.5% higher in mouse. Given that it is unlikely that intrachromosomal duplications are on average more recent than interchromosomal duplications, we think that molecular mechanisms such as gene conversion could help to preserve the sequence identity of duplications within the same chromosomes in a higher degree than between different chromosomes.

The common characteristics of segmental duplications in the most recent mouse and human assemblies, particularly if we consider that mouse and human duplication sequences do not share any significant homology, suggest that the mechanisms that generate and preserve segmental duplications in mouse and human have similar molecular bases and that these mechanisms are, at least in great extent, independent of the sequence content of the duplicated segment.

In human and in mouse, the duplication content of the syntenic genome increases with proximity to the junctions where the synteny breaks and the regions of discontinuity in the synteny have the highest content of duplications. It has been reported that rearranged chromosomes associate with an accelerated rate of evolution.[23] Considering that segmental duplications tend to be located where mouse and human ancestral chromosomes have been rearranged, we can hypothesize that segmental duplications are a driver for genomic and chromosomal evolution in man and mouse.

Acknowledgements

We thank Dr Augustine Kong for helpful comments.

References

1 Ohno S, Wolf U, Atkin NB: Evolution from fish to mammals by gene duplication. *Hereditas* 1968; **59**: 169–187.
2 Lander ES, Linton LM, Birren B *et al*: Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.
3 Waterston RH, Lindblad-Toh K, Birney E *et al*: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; **420**: 520–562.
4 Cheung VG, Nowak N, Jang W *et al*: Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 2001; **409**: 953–958.
5 Bailey JA, Gu Z, Clark RA *et al*: Recent segmental duplications in the human genome. *Science* 2002; **297**: 1003–1007.
6 Cheung J, Estivill X, Khaja R *et al*: Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol* 2003; **4**: R25.
7 Gratacos M, Nadal M, Martin-Santos R *et al*: A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell* 2001; **106**: 367–379.
8 Mazzarella R, Schlessinger D: Pathological consequences of sequence duplications in the human genome. *Genome Res* 1998; **8**: 1007–1021.
9 Emanuel BS, Shaikh TH: Segmental duplications: an ‘expanding’ role in genomic instability and disease. *Nat Rev Genet* 2001; **2**: 791–800.
10 Stankiewicz P, Lupski JR: Genome architecture, rearrangements and genomic disorders. *Trends Genet* 2002; **18**: 74–82.
11 Singleton AB, Farrer M, Johnson J *et al*: alpha-Synuclein locus triplication causes Parkinson’s disease. *Science* 2003; **302**: 841.
12 Samonte RV, Eichler EE: Segmental duplications and the evolution of the primate genome. *Nat Rev Genet* 2002; **3**: 65–72.
13 Eichler EE: Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* 2001; **17**: 661–669.
14 Goidts V, Szamalek JM, Hameister H, Kehrer-Sawatzki H: Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Hum Genet* 2004; **115**: 116–122.
15 Skaletsky H, Kuroda-Kawaguchi T, Minx PJ *et al*: The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 2003; **423**: 825–837.
16 Rozen S, Skaletsky H, Marszalek JD *et al*: Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 2003; **423**: 873–876.
17 Dehal P, Predki P, Olsen AS *et al*: Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 2001; **293**: 104–111.
18 Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X: Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet* 2003; **12**: 2201–2208.
19 Kent WJ: BLAT—the BLAST-like alignment tool. *Genome Res* 2002; **12**: 656–664.
20 Bailey JA, Yavor AM, Viggiano L *et al*: Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am J Hum Genet* 2002; **70**: 83–100.
21 Gregory SG, Sekhon M, Schein J *et al*: A physical map of the mouse genome. *Nature* 2002; **418**: 743–750.
22 She X, Jiang Z, Clark RA *et al*: Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 2004; **431**: 927–930.
23 Navarro A, Barton NH: Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science* 2003; **300**: 321–324.