A genome-wide survey of segmental duplications that mediate common human genetic variation of chromosomal architecture

*Michael R. Mehan, Nelson B. Freimer and Roel A. Ophoff**

Department of Human Genetics and Center for Neurobehavioral Genetics, Neuropsychiatric Institute, University of California Los Angeles, Gonda Center, Room 3506, 695 Charles E. Young Drive South, Los Angeles, California 90095, USA.
**Correspondence to*: Tel: +1 310 794 9602; Fax: +1 310 794 9613; E-mail:ophoff@ucla.edu

Date received (in revised form): 20th May 2004

Abstract

Recent studies have identified a small number of genomic rearrangements that occur frequently in the general population. Bioinformatics tools are now available for systematic genome-wide surveys of higher-order structures predisposing to such common variations in genomic architecture. Segmental duplications (SDs) constitute up to 5 per cent of the genome and play an important role in generating additional rearrangements and in disease aetiology. We conducted a genome-wide database search for a form of SD, palindromic segmental duplications (PSDs), which consist of paired, inverted duplications, and which predispose to inversions, duplications and deletions. The survey was complemented by a search for SDs in tandem orientation (TSDs) that can mediate duplications and deletions but not inversions. We found more than 230 distinct loci with higher-order genomic structure that can mediate genomic variation, of these about 180 contained a PSD. A number of these sites were previously identified as harbouring common inversions or as being associated with specific genomic diseases characterised by duplication, deletions or inversions. Most of the regions, however, were previously unidentified; their characterisation should identify further common rearrangements and may indicate localisations for additional genomic disorders. The widespread distribution of complex chromosomal architecture suggests a potentially high degree of plasticity of the human genome and could uncover another level of genetic variation within human populations.

Keywords: *genomic architecture, segmental duplications, inversion polymorphism, genomic variation*

Introduction

Investigation of human genetic variation has focused mainly on single nucleotide polymorphisms (SNPs) and minisatellite and microsatellite repeat sequences. Although it has long been known that genomic rearrangements predispose to numerous disease phenotypes, it has only recently become apparent that some such rearrangements occur frequently in the general population. Investigation of specific loci or chromosomal regions has identified the few known common variations in genomic architecture. Now, however, the availability of bioinformatics tools, for searching for patterns in genome sequences, enables genome-wide surveys for particular types of higher-order structure predisposing to genomic variation. Segmental duplications (SDs) represent a form of genome architecture constituting up to 5 per cent of the human genome.[1–4] Non-allelic homologous recombination between these paralogous sequences results in changes of genomic structure creating inversions and other types of chromosomal rearrangements, sometimes leading to disease.[5–8] Recently, two genomic regions were identified that harbour a common inversion polymorphism. On chromosome 8p23, two large low-copy repeat regions containing olfactory-receptor (OR) gene clusters spanning approximately 350 kilobases (kb) each, and separated by approximately 4 megabases (Mb) of unique sequence, mediate recurrent genomic rearrangements.[9] An inversion polymorphism in this segment is present in heterozygous form in about 25 per cent of Europeans.[9] Additionally, a homologous structure of two pairs of OR gene clusters at 4p16, separated by almost 6 Mb, mediates a relatively common translocation between chromosomes 4p16 and 8p23.[10] More than 10 per cent of Europeans sampled are heterozygous for an inversion at 4p16.[10] Detailed examination of the regions at 4p16 and 8p23, which contain the submicroscopic genomic

inversions, revealed specific higher-order structures involving SDs termed palindromic segmental duplications (PSDs), which predisposes to inversions, duplications and deletions. A PSD consists of paired, *inverted* duplications within limited physical distance of each other. In addition to PSDs, non-homologous recombination between segmental duplications in *tandem* orientation (TSDs) are known to mediate duplications and deletions but not inversions, leading to changes in copy number of intervening DNA sequences.[11] Recurrent deletions and duplications are a known cause for genomic disorders and are observed relatively frequently,[11] whereas submicroscopic inversion events without change of copy number (of a gene) are hard to detect and may not necessarily lead to a distinct phenotype. We hypothesised that the genomic architecture containing PSDs associated with common inversion polymorphisms is not unique to the 8p23 and 4p16 regions. The existence in the human genome of recurrent SDs that mediate common inversion polymorphisms without known association with human diseases, raised the possibility that there are many more such PSD structures throughout the genome. In this paper we describe the results of a genome-wide database search for loci containing chromosomal architecture that could mediate genomic variation, with a special emphasis on PSD structures, and discuss the implications of these findings for our understanding of genome plasticity.

Materials and methods

Chromosome 8-4-11-3 PSD family analysis

Using documented markers from the SDs that mediate the genomic rearrangements on chromosomes 8p23 and 4p16, all four repetitive regions were downloaded from the National Center for Biotechnology Information (NCBI)'s public domain. Each pair of SDs was aligned. Each of the four SDs was compared with the others in six pair-wise alignments using Miropeats.[12] To eliminate alignment redundancy, we designed a Combine-and-Color (CC) algorithm that modifies the Miropeats output (see below). The parameters used for our CC algorithm, described below, were an internal spacing threshold of 50 base pairs (bp) and a maximum spacing difference of 75 bp.

CC algorithm

The purpose of the CC algorithm is to combine overlapping or closely neighbouring alignments into more comprehensive alignments, and to colour the corresponding alignment blocks for visualisation. The combining algorithm exhaustively compares all neighbouring local alignments from two sequences, two at a time. Using the relative start and stop locations of the alignment on each sequence, the algorithm computes the *internal spacing* on each sequence and the *spacing difference* between the two. The internal spacing, calculated once for each sequence, is the distance between the end of the first alignment and the beginning of the second. If the sequences overlap, the internal spacing would therefore be negative. Once the spacing between the alignments on both sequences has been calculated, the spacing difference is determined by calculating the absolute difference between the two internal spacings. The spacing difference ensures that the two alignments are uniformly spaced on both sequences. If both the internal spacings and the spacing difference are less than predefined thresholds, the two alignments are combined so that a new single alignment spans the regions on both sequences, defined by the previous two alignments. After all possible combinations have occurred, the alignment blocks are coloured according to size to aid in visualisation. Alignments of less than 100 bp are coloured yellow, and the colouration increases in darkness (ie orange, cyan, purple, green, red, blue) and ends in black as the alignment size increases to greater than 4 kb.

Genome-wide BLAST analysis

For the genome-wide detection of PSD and TSD pairs, all sequence data for each chromosome were downloaded from the University of California, Santa Cruz (UCSC) Genome July 2003 Freeze. To reduce computation time and background noise, chromosomes were 'fuguised'[1] by removing both repetitive elements masked by Repeat Masker (A.F.A. Smit and P. Green, unpublished), and unsequenced gaps from the July 2003 Freeze. PSD pairs were defined as two segmental duplications of at least 10 kb in length and with $\geq$ 90 per cent sequence identity, in *inverted* orientation, and with an internal spacing between the two members of the pair of a maximum of 8 Mb. TSD pairs were similarly defined as two segmental duplications in *tandem* orientation with identical criteria, as described for PSDs. We used the NCBI's stand-alone BLAST release 2.2.6 for the alignment of the chromosomes.[13] Using *bl2seq*, we aligned pairs of sequences consisting of a query sequence of 100 kb and a subject sequence of 8 Mb. The first pair consisted of the first 100 kb and first 8 Mb of the chromosome. The BLAST results from this pair constituted our first BLAST pair output. The process was repeated, stepping our 8 Mb window forward by 100 kb with each iteration. For our BLAST analysis, we implemented the restriction of a maximum expectation value (E-value) of e^{-20}. BLAST hits that met our criteria of 90 per cent identity and plus/minus and plus/plus orientation for PSDs and TSDs, respectively, were stored in a duplication file for annotation.

To annotate the PSD pairs, we used our CC algorithm to join neighbouring BLAST hits in our duplication file, which comprised the same PSD pair member. All BLAST hits that met our criteria were sorted by starting position of the query sequence to group neighbouring hits together. To preserve the structure of the PSD pair, neighbouring hits from both regions of the PSD pair were examined simultaneously. The thresholds for internal spacing and spacing difference were both 2 kb. After the combining algorithm was completed, the physical

locations of all PSD pairs whose two members were both greater than 10 kb were stored in a segment file. Annotation of TSD pairs was performed in identical fashion.

PSD pair BLAST database

Using the physical locations defined in the segment file, all PSD pair sequence data were extracted from the fugued chromosomes for preparation of a PSD pair BLAST database. To determine the sequence similarity between our PSD pair set, each PSD pair member was BLASTed against the database using *blastall*. For each PSD pair element queried, all PSD pair elements in our database containing at least one hit with an E-value of zero were considered significant, and the ranges of coverage were recorded in a coverage file.

Results

Structure of common inversion regions on 8p23 and 4p16

To assess the chromosomal architecture involved in the common inversions on chromosomes 8p23 and 4p16, we first compared the low copy repeats flanking the two inversion regions using the ICAass algorithm and graphically displayed them through our adaptation of the Miropeats program,[12,14] which reduces redundancy and clarifies the results through colourisation. From this analysis, we confirmed the high degree of sequence similarity between the two flanking duplicated segments for each inversion region (intra-chromosomal) as well as between the duplicated segments across the different chromosomes (inter-chromosomal). There was clearly recognisable sequence similarity for over 200 kb for duplicated segments flanking non-duplicated, unique sequences. We developed a probe from the 4p duplicated sequence and used it to search the genome with the online BLAST alignment tool[13] via the NCBI website. This led to the discovery of two other loci on human chromosomes 3 and 11, containing the same pairs of duplicated segments in inverse orientation with a similar internal spacing but unrelated sequence between them (Figure 1a–c). The duplicated sequences that are part of the PSDs at 8p23 and at 3q21 are in opposite orientation to the same segments of the PSDs at 4p16 and 11q13. We also noted three instances of single occurrences of these segments on chromosomes 7, 12 and 16, each containing at least 120 kb of the paralogous-4p sequence (data not shown).

Genome-wide search for PSDs and TSDs

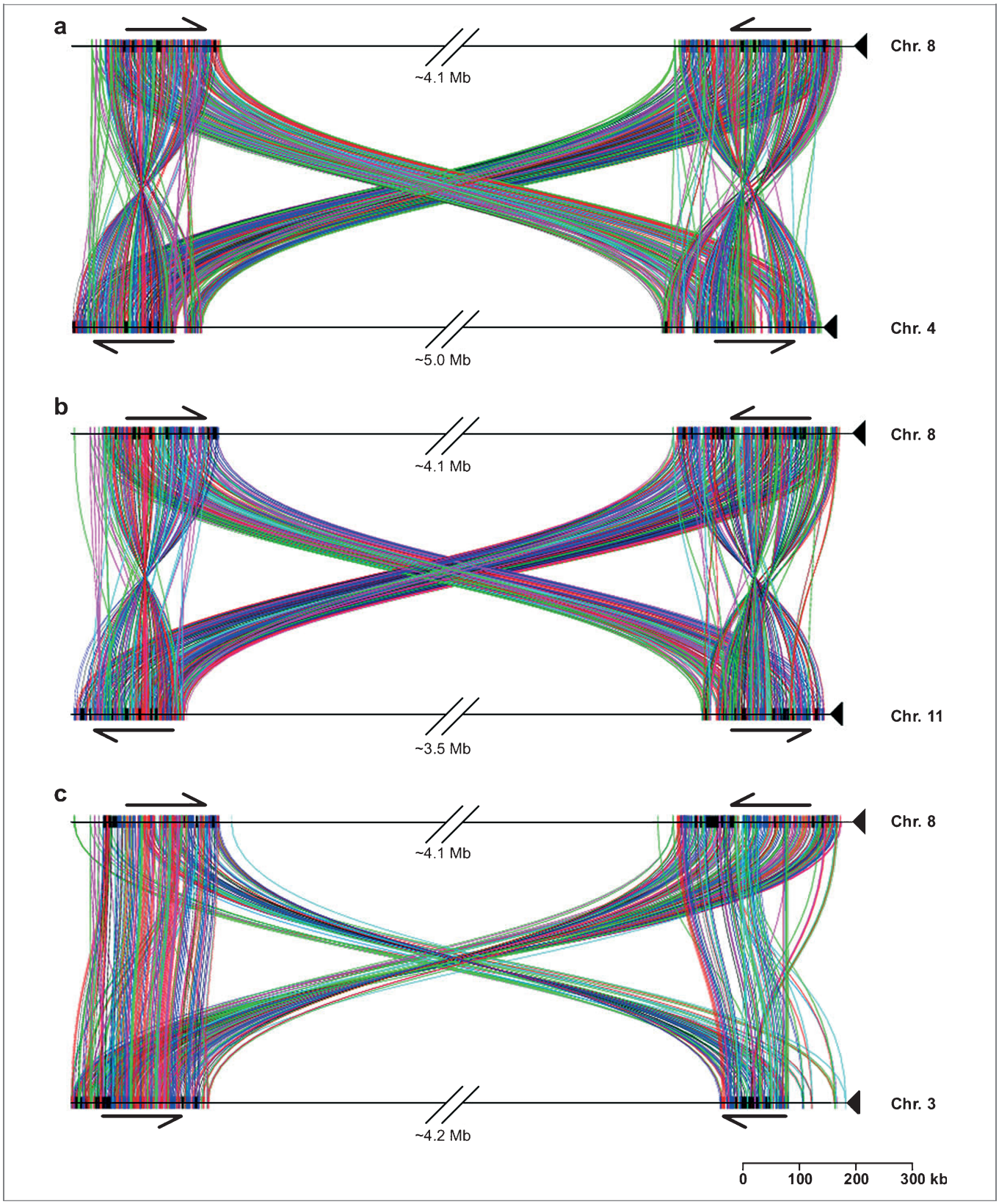
To identify additional PSDs, as well as TSDs, in the human genome, we performed a genome-wide database search using the July 2003 Freeze (UCSC Genome Browser) of the human genome and BLAST-based tools. We searched for PSDs based on a minimal length of 10 kb for each member of the PSD showing greater than 90 per cent sequence identity, and based on a maximum distance of 8 Mb between each member of the pair. Both lengths were based on 'fuguised' sequences of each chromosome, that is, with repetitive and ambiguous sequences removed.[1] On average, the fuguisation process removed 52 per cent of the genome (July 2003 Freeze), indicating that the original minimal length for each member was approximately 20 kb, with a maximum spacing of approximately 16 Mb. Fuguisation of the Y chromosome removed about 81 per cent of the total sequence, reflecting the abundance of highly repetitive sequences and large gaps of unknown sequences on this chromosome.[15] The search identified 861 PSDs throughout the human genome (Figure 2), with a concentration of PSDs near centromeres; 50 per cent of the 861 PSDs occur within 4.4 Mb of a centromere. We observed no over-representation of PSDs in subtelomeric regions. Similarly, we identified 705 TSD pairs throughout the human genome, with 50 per cent within 4.6 Mb of a centromere; no such correlation was found with subtelomeric regions. The results of the genome-wide search for PSDs and TSDs are available online (http://SDbrowser.genetics.ucla.edu) in a dynamic browser providing coordinates with reference gene sequences from NCBI's RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq) as well as genetic markers from the deCODE genetic map.[16]

Since PSDs seem to cluster in specific regions, and multiple PSD hits may represent single instances of duplicated segments in inverted orientation, we assessed the redundancy of PSDs from our screen. This was done by combining overlapping pairs of PSDs into one, as well as joining PSD pairs that were within 100 kb. We estimated that there are 179 distinct PSD-containing regions. Likewise, 144 distinct TSD loci were identified. For the combined PSD and TSD data, we identified 233 distinct regions in the genome containing large blocks of duplicated sequences in opposite or tandem orientation. From these 233 regions, 86 loci contain exclusively PSDs whereas about half that number (46) feature only TSDs. These numbers not only show that there are more PSD than TSD structures in the human genome but also indicate that PSDs and TSDs frequently co-localise within the same region of the genome.

The initial finding of the 8-4-11-3 PSD family led to the suggestion that perhaps more distinct PSD families could be identified. We therefore performed a sequence analysis between each of the PSDs using the BLAST algorithm.[13] For 550 out of the 861 total PSD pairs (69.3 per cent) we found no inter-chromosomal hit, and for 118 of the 861 PSD pairs (13.7 per cent) we found neither an inter- nor an intra-chromosomal hit. The vast majority of hits (94 per cent) were intra-chromosomal. Further sequence comparisons of all PSD sequences against the human genome revealed a large number of unpaired duplicated segments throughout the genome (data not shown).

Known genomic regions with PSDs and TSDs

We tested the ability of our database search method to detect PSDs by analysis of several loci in the genome already known


a
Chr. 8
~4.1 Mb
Chr. 4
~5.0 Mb
b
Chr. 8
~4.1 Mb
Chr. 11
~3.5 Mb
c
Chr. 8
~4.1 Mb
Chr. 3
~4.2 Mb
0
100
200
300 kb

to harbour these structural features. First, we showed that the search could detect the 8p-4p inversion polymorphism regions and the related loci at 11q13 and 3q21, confirming both the genomic architecture of PSD pairs at these sites and the clustering of sequence similarities with each other (Figure 3). Secondly, we examined whether our thresholds could detect regions known to be associated with genomic disorders mediated by PSDs. We matched a number of genomic disorders from different regions in the genome to corresponding PSD locations; three of these locations are shown in more detail in Figure 4. The genomic disorders and their locations are (i) the Williams-Beuren syndrome (WBS; MIM 194050) region at 7q11, containing an inversion polymorphism mediated by PSDs with complex repetitive structure surrounding the elastin (ELN) gene[17] (Figure 4a); (ii) the Angelman syndrome (AS; MIM 105830) region on chromosome 15q11-q13, which is often deleted in patients with AS and inverted in 4.5 per cent of the chromosomes in the general population,[18] characterised by a PSD (Figure 4b); and (iii) Sotos syndrome (SoS; OMIM 117550), a recently reported neurological disorder characterised by cerebral overgrowth caused by haploinsufficiency of the NSD1 gene at chromosome 5q35.[19] A large proportion (up to 50 per cent) of Japanese patients revealed a common 2.2 Mb deletion which was suggested to be mediated by the presence of duplicated segments of highly homologous sequences.[19,20] Figure 4c shows the results of the survey of the chromosome 5q35 region with the NSD1 gene located within a PSD pair. No inversion polymorphism has been reported for this region so far.

The largest and most extended PSD and TSD structures are found on the Y chromosome at the azoospermia factor c (AZFc) region (Figure 4d). This 3.5 Mb region is known to consist of massive palindromes with uniform breakpoints that are frequently deleted in infertile men.[21] Recently, this region was reported to harbour a deletion polymorphism possibly affecting spermatogenesis.[22] Lastly, a known inversion polymorphism that features a PSD structure is located at the emerin (EMD) locus at chromosome Xq28 (Figure 5).[23] This region, almost 50 kb in length, has been found to be inverted in about 20 per cent of X chromosomes mediated by flanking inverted repeats, each being 11.3 kb in length and with >99 per cent sequence identity. The pair of duplicated segments that are part of this PSD do not meet our stringent criteria of minimal length of 10 kb for each PSD member after 'fuguisation'. Thus, the EMD locus is an example of a PSD region with smaller segments that mediate inversion polymorphisms.

Discussion

We have performed a genome-wide survey of specific patterns of chromosomal architecture in the human genome, based on results of our initial analysis of two regions containing inversion polymorphisms on chromosomes 8p23 and 4p16.[9,10] Close examination of these two loci revealed the presence of PSDs, consisting of paired, inverted duplications, flanking the inversion region with unique sequence. Moreover, we confirmed the high degree of sequence similarity between these PSDs across the different chromosomes and identified two other loci on chromosomes 3q21 and 11q13, containing almost identical PSDs in inverse orientation, with similar spacing but without known genomic variation. This result showed that the genomic architecture of the 8p23 and 4p16 inversion regions is not unique. Moreover, the existence of recurrent SDs in the human genome, mediating inversion polymorphisms, raised the possibility that there are many more such PSD structures throughout the human genome; this possibility led us to perform a genome-wide survey of PSDs that can mediate genomic variation of chromosomal architecture. The genome-wide analysis also included a survey of TSDs, which can also mediate ectopic sequence exchange, causing deletions or duplications. We focused particularly on regions with PSDs that could mediate 'balanced' inversion events, that is, without loss or gain of intervening sequence. Our results revealed a large number of loci harbouring these structural features; most of these were previously unknown and await further confirmation and characterisation.

Distribution of PSDs and TSDs

The chromosomal distribution of PSDs, as well as TSDs, shows distinctive patterns. Segmental duplications within

Figure 1. Pair-wise alignments using the ICAass algorithm and our adaptation of Miropeats (reducing redundancy and clarifying alignments via colouration) revealing a high degree of sequence similarity and structure between the chromosome 8p23 inversion region and duplicated segments from other chromosomes. Coloured lines depict the degree of alignment; darker coloured lines represent longer alignments. Colours range from yellow (<100 bp of sequence) through to orange, cyan, purple, green, red, blue and black (>4 kb of sequence). The arrows indicate the relative orientation of the repeats at the different loci. Figures 1a and 1b show the alignment from the 8p23 inversion region, with sequences from 4p16 and 11q13, respectively, showing the entirety of the segmental duplications in opposite orientation with respect to chromosome 8p23 low-copy repeats. Figure 1c shows a similar alignment, with sequences from 3q21 revealing the same orientation and structure as chromosome 8p23 low-copy repeats but with one duplicated segment not completely present, partly due to the complete sequence not being available for that genomic region.

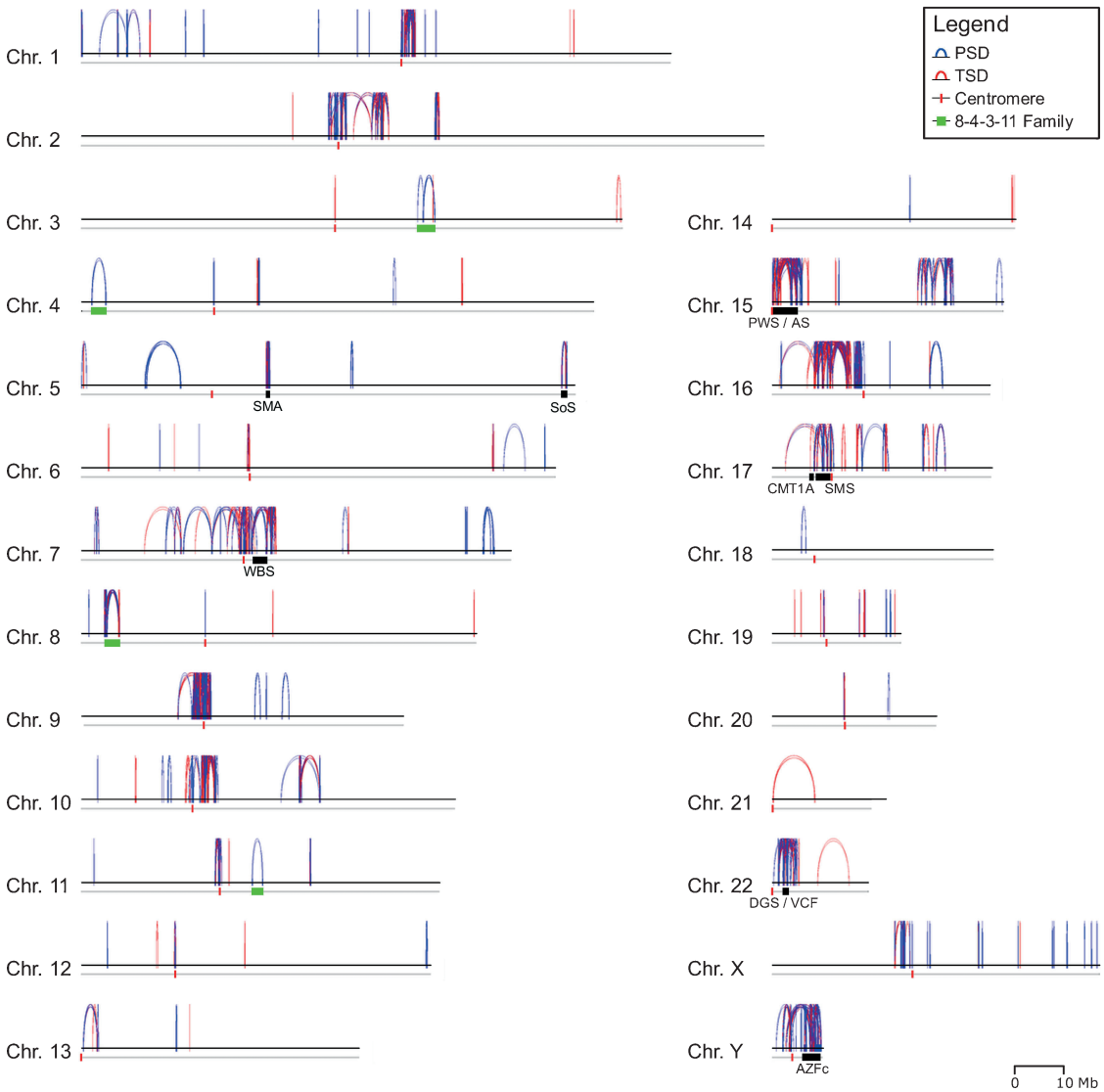


Figure 2. Results from a genome-wide survey for a complex type of higher-order chromosomal structure which may mediate genomic variation. Distribution of all 861 palindromic segmental duplications (PSDs) (blue lines) and 705 SDs in tandem orientation (TSDs) (red lines) that met our criteria of a minimum length of 10 kb and maximum spacing of 8 Mb are displayed on the 'fugued'[1] scale for each chromosome. Red boxes indicate locations of centromeres, removed during the fuguisation process. Green bars represent the inversion regions from chromosomes 8p23 and 4p16, as well as related regions from chromosomes 11q13 and 3q21; detailed results for these regions are shown in Figure 3. Chromosomal locations (labelled WBS, AS, SoS and AZFc) associated with genomic disorders are represented as black bars: Williams–Beuren-syndrome (WBS) on chromosome 7q11, Sotos syndrome (SoS) on chromosome 5q35, Angelman syndrome (AS) on chromosome 15q11-q13 and the azoospermia factor c (AZFc) region on the Y chromosome; each of these regions is detailed in Figure 4. Additional hotspots for a number of known genomic disorders are labelled as follows: SMA for spinal muscular atrophy, PWS for Prader–Willi syndrome, CMT1A for Charcot–Marie–Tooth disease type 1A, SMS for Smith–Magenis syndrome and DGS/VCF for DiGeorge/velocardiofacial syndrome.

pericentromeric and subtelomeric regions are well documented[24–26] but their distribution and number vary by chromosome.[2] Our results demonstrated an over-representation of PSDs and TSDs near centromeres, although not all centromeres are characterised by a high density of these kinds of structures. The reason for the abundance of PSDs and TSDs near some centromeres could lie in the fact that these pericentromeric regions harbour greater overall plasticity.[27,28] The scarcity of these structures in subtelomeric regions, which are also recognised as sites of rapid genomic change, however, may suggest differences in higher-order structure and/or type of plasticity between centromeres and telomeres in the human

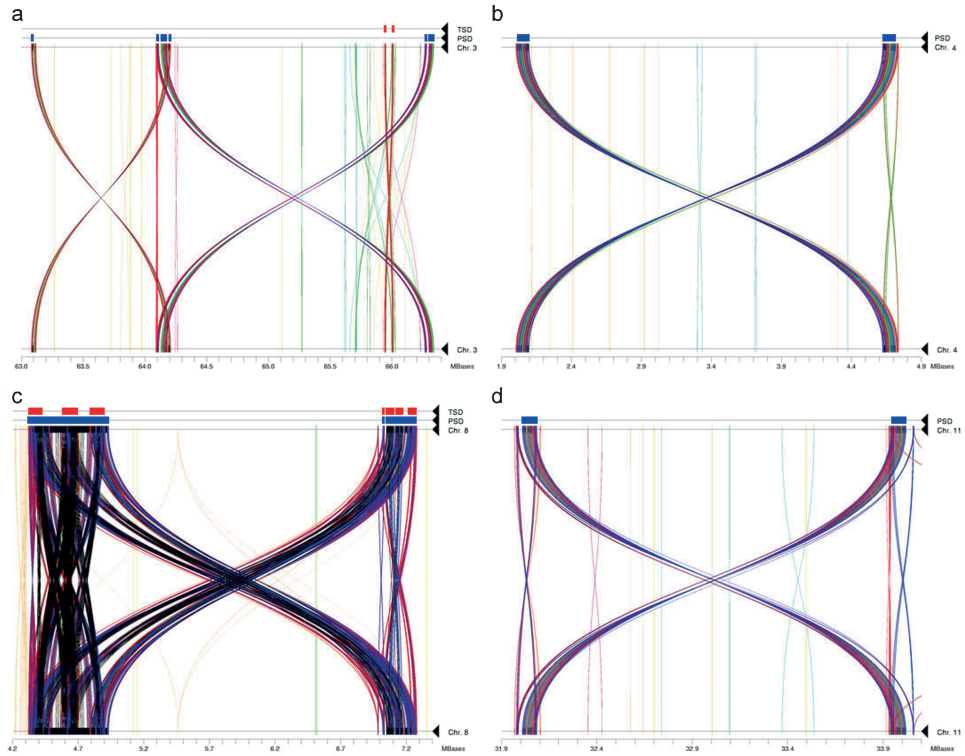


Figure 3. Graphical adaptation of BLAST alignments of palindromic segmental duplication (PSD) regions from chromosomes 8p23, 4p16, 11q13 and 3q21 showing high sequence similarity between inverted members of PSD pairs with relatively equal internal spacing between the loci. Horizontal bars containing graphical alignment represent identical sequences from each 'fugued' chromosomal region, as indicated. All sequence identities with an E-value of e^{-20} and ≥ 90 per cent identity are displayed graphically between the horizontal bars, colourised based on alignment length, as described in Figure 1. The bar directly above the alignment represents either PSDs or SDs in tandem orientation (TSDs). (a) Two overlapping PSDs are present on chromosome 3q21, one showing a similar but unique structure to 8-4-3-11 family-related PSDs and one internally flanking chromosome 3-specific PSD; no common inversion polymorphism is reported for this region. (b) PSD mediating the inversion on chromosome 4p16 with less complex structural features. (c) PSD mediating the common inversion on 8p23 harbouring multiple internal TSDs, especially on the left arm. (d) PSD at chromosome 11q13, with sequence structure and internal spacing similar to PSD regions at 8p23 and 4p16 but without known common inversion polymorphism.

genome. Comparative studies of these regions are required to ascertain whether this is a particular structural characteristic for centromeres and telomeres in general, or is possibly restricted to the human genome.

We observed that SDs, including both PSDs and TSDs, are usually arranged in a complex structure consisting of multiple modules, some in direct orientation and others in inverted orientation. Of the relatively few SDs that are uniquely PSDs or TSDs, there is a preponderance of PSD regions. The high prevalence of PSDs suggests that SDs within close proximity preferentially have occurred in inverted orientation. It is unclear why this preference would occur, as there is no apparent advantage for inverted duplications over tandem duplications. It is possible, however, that a preponderance of PSDs exists because of three-dimensional structural advantages at the chromatin level for these events to occur. If this is true, one might expect that comparison of genomic structures of SDs across species will reveal the same bias in distribution of PSDs versus TSDs. Further study is required to confirm this hypothesis.

The observation that the 8p23 and 4p16 inversion-mediating PSDs are members of a family with at least two additional, nearly identical loci at 11q13 and 3q21, led to the suggestion that perhaps more distinct PSD families could be identified. Sequence comparison between all identified PSDs revealed that the vast majority of PSDs were related to at least one other PSD sequence in the genome, with almost all sequences found on the same chromosome. This high rate of intra-chromosomal hits suggests a closer relationship between PSDs on the same chromosome than between PSDs on different chromosomes. This strong bias is not surprising, since any PSD pair, by definition, consists of an *intra*-chromosomal duplicated segment in close proximity. Similar numbers have been found for paralogous sequences in general,[1,2] suggesting that PSDs are not an inherently different group of duplicated segments within the human genome.

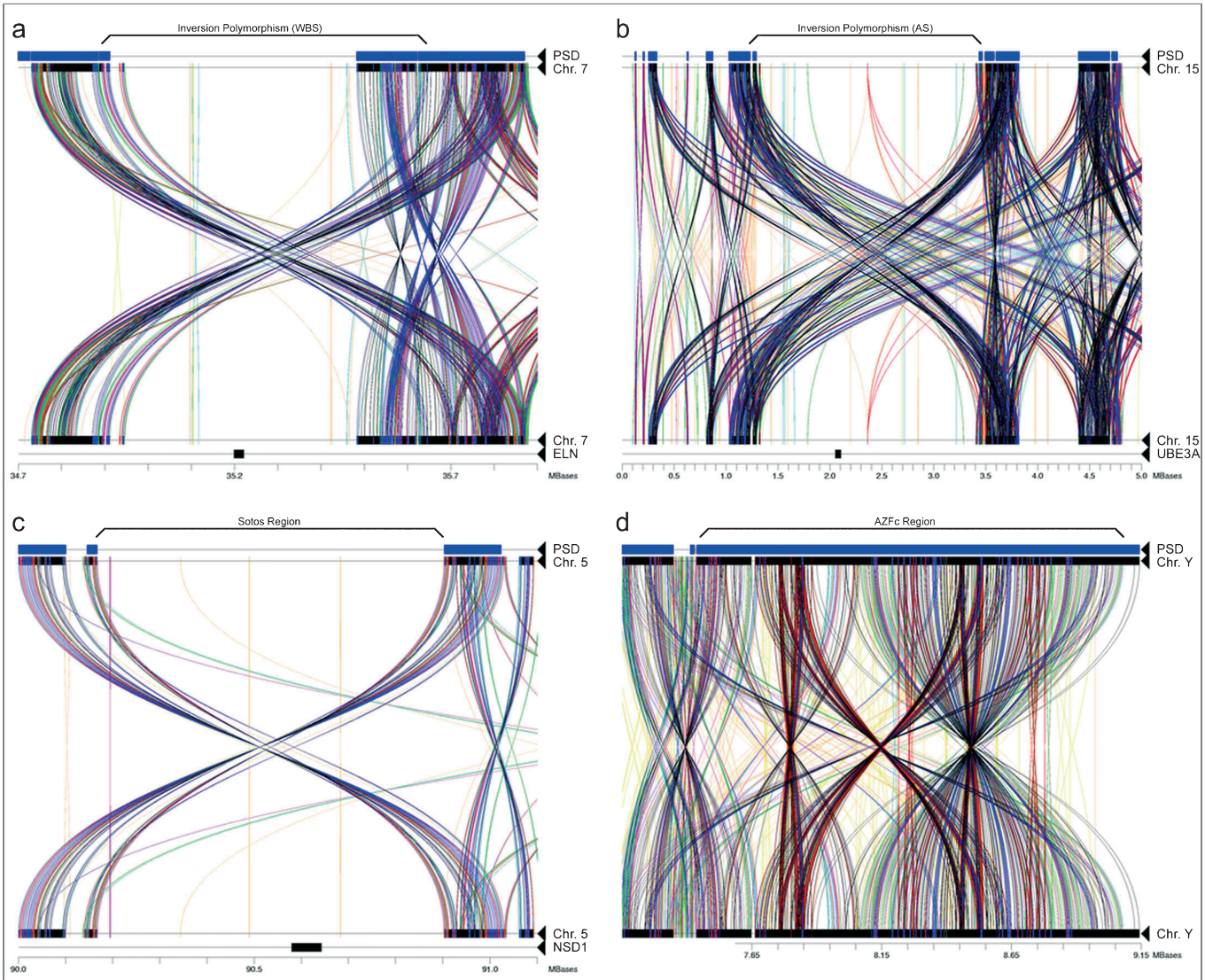


Figure 4. Graphical adaptation of BLAST alignments of four regions associated with chromosomal rearrangements producing genomic disorders. The approximate locations of primary genes associated with the respective diseases are indicated with a black bar. (a) Williams-Beuren syndrome (WBS) region at 7q11 containing inversion polymorphism with the elastin (ELN) gene located as indicated.[17] (b) The palindromic segmental duplication (PSD) region on chromosome 15q11-q13, inverted in 4.5 per cent of the chromosomes in the general population[18] and often deleted in patients with Angelman syndrome (AS). The location of the ubiquitin protein ligase e3a (UBE3A) gene is indicated. (c) The region at chromosome 5q35 surrounding the NSD1 gene is frequently deleted in Japanese patients with Sotos syndrome and is characterised by a PSD. (d) The largest PSD region of the human genome is located in the AZFc region on the Y chromosome, featuring massive palindromes with minimal internal spacing. This region is known to harbour recurrent deletions and a deletion polymorphism that both may be associated with spermatogenic failure in infertile men. [21,33]

Genome assembly

It is important to note that the current sequence assembly of the human genome, even though in its advanced stages, is still incomplete and contains gaps. This could lead to under-representation of duplicated segments, misalignments of some sequence data[1,29] and uncertainties in the proper orientation especially of low-copy repeat sequences. For this reason, independent molecular confirmation is required for any region identified in our survey that may underlie common genomic variations. The combination of SD data in inverted *and* tandem orientation (PSD and TSD, respectively) is thus necessary to identify these sites. Moreover, under-representation of SDs in the human genome assembly will lead to an underestimation of loci with higher-order structure that are potentially involved in genomic variation. These factors certainly limit our survey, in that PSDs may be recognised as TSDs or vice versa, or alternatively that the actual number of these types of loci in the human genome is higher than our

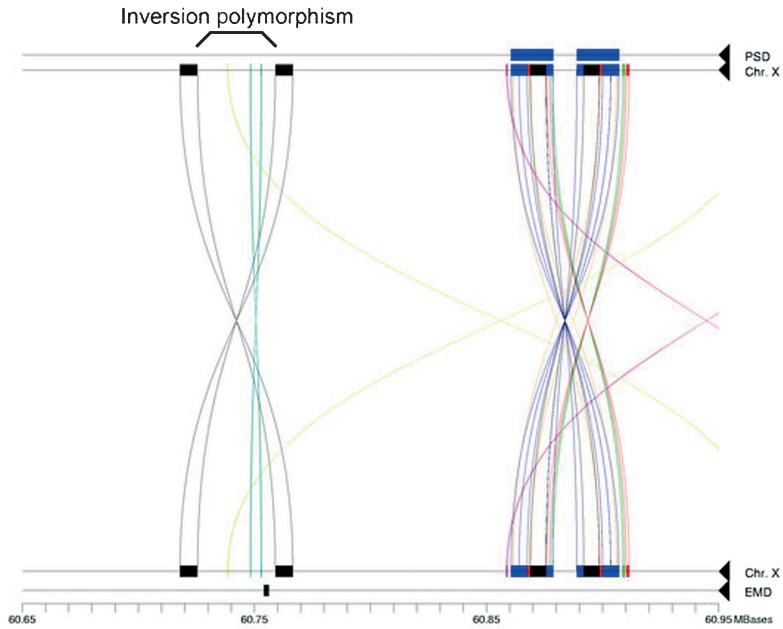


Figure 5. Graphical adaptation of BLAST alignments of the region containing a common inversion polymorphism surrounding the emerin (EMD) gene associated with Emery–Dreifuss muscular dystrophy.[23] This is an example of a palindromic segmental duplication (PSD) associated with a common inversion that did not meet our stringent criteria of minimum length of 10 kb for each duplicated segment after 'fuguisation'. The location of the EMD gene is indicated below. The yellow and purple lines represent short sequence overlaps of <100 and <200 bp, respectively, mapping to the telomeric region just outside the window range on the X chromosome.

data suggest. Nevertheless, our effort to identify specific sequence structures involving SDs in the human genome is an important step to corroborate the concept that human genome plasticity is probably very substantial and is not limited to the pericentromeric and subtelomeric regions.

Genome plasticity

One example of a region associated with a genomic disorder is that of 17p11, associated with the Smith-Magenis syndrome (SMS; MIM 182290). This region comprises three SDs (distal, middle and proximal) combined into two PSDs forming one TSD structure, commonly deleted in subjects with SMS.[30] Interestingly, within this SMS region the breakpoint region for a common isochromosome 17q [i(17q)] in human neoplasia was recently reported;[31] i(17q) is associated with loss of 17p, which includes the tumour-suppressor gene TP53. The recurrent breakpoint of i(17q) was described as a PSD locus. This example suggests, again, that somatic rearrangements are not random but that genome architecture, such as PSDs and TSDs, may also be important in chromosomal rearrangements associated with human neoplasia. Moreover, a different study identified an abundance of SDs in this 17p region and also reported that the particular genomic architecture is involved in non-recurrent chromosomal rearrangements and unusual-sized deletions.[30]

The presence of a great number of regions in the human genome harbouring higher-order structures predisposing to genomic variation also implies that the chromosomal structure of these loci may vary between human populations. It is interesting, for example, that for Sotos syndrome, microdeletions are commonly observed in Japanese patients but only in a very small fraction of non-Japanese patients.[19,20,32] The reason for this large difference in frequency of microdeletions could be due to a patient-selection bias but one could also argue that some 5q35 alleles with a particular variation in genomic architecture predisposing to these events in the respective populations are the basis for the observed differences. Even though the latter may be incorrect for the 5q35 Sotos syndrome region, it may be the correct scenario for one of the many other regions in the genome containing complex chromosomal architecture.

Our results indicate that the human genome harbours a considerable number of regions whose higher-order structure may vary within human populations. The approach that we employed, however, was restricted to a single set of criteria, focusing on PSDs with a minimum length for each segment and with a given maximum spacing. These criteria were applied to maximise the chances of identifying regions that predispose to recurrent inversions such as those seen in 8p and 4p. There are, however, already examples of PSD regions with smaller segments that mediate inversion polymorphisms, for example at the EMD locus on chromosome Xq28. This, in addition to the limitations of the current genome assembly, as previously mentioned, suggests that further investigation may reveal

additional regions in which there is common variability of genomic structure. Such variability in higher-order structure of the genome could also alter our interpretation of genetic maps and haplotypes, especially at high resolution. Under the assumption of uniform architecture, we consider maps to represent a fixed order of markers, although we recognise that the genetic distance between two markers is variable (eg between males and females) and therefore represents an average. Similarly, we may need to consider that the order also represents an average for specific regions in the genome. Indeed, unrecognised variability in the order of markers could increase uncertainty in estimates of the distances between them. Systematic identification of relatively widespread genetic variations in genome structure may be important for comparative genomic studies, for analysis of recombination in the human genome and, in particular, for mapping phenotypes. While only a small proportion of SNPs may have a functional effect, it is likely that a relatively high proportion of variants in higher-order structure have either direct or indirect effects on the function of one or more genes, given the large amount of genome sequence incorporated within each variant.

Acknowledgments

We thank Susan Service and York Mararhens for comments. M.R.M. was supported by NSF IGERT Training Award #DGE-9987641. This work was funded by a NARSAD Young Investigator Award to R.A.O. and by grants from the US National Institutes of Health; (R01 MH 49499) to N.B.F. and (R01 GM 068875-01) to R.A.O.

References

1. Bailey, J.A., Yavor, A.M., Massa, H.F. *et al.* (2001), 'Segmental duplications: Organization and impact within the current human genome project assembly', *Genome Res.* Vol. 11, pp. 1005–1017.
2. Bailey, J.A., Gu, Z., Clark, R.A. *et al.* (2002), 'Recent segmental duplications in the human genome', *Science* Vol. 297, pp. 1003–1007.
3. Cheung, J., Estivill, X., Khaja, R. *et al.* (2003), 'Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence', *Genome Biol.* Vol. 4, pp. R.25.
4. Lander, E.S., Linton, L.M., Birren, B. *et al.* (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
5. Samonte, R.V. and Eichler, E.E. (2002), 'Segmental duplications and the evolution of the primate genome', *Nat. Rev. Genet.* Vol. 3, pp. 65–72.
6. Stankiewicz, P. and Lupski, J.R. (2002), 'Molecular-evolutionary mechanisms for genomic disorders', *Curr. Opin. Genet. Dev.* Vol. 12, pp. 312–319.
7. Emanuel, B.S. and Shaikh, T.H. (2001), 'Segmental duplications: An 'expanding' role in genomic instability and disease', *Nat. Rev. Genet.* Vol. 2, pp. 791–800.
8. Mazzarella, R. and Schlessinger, D. (1998), 'Pathological consequences of sequence duplications in the human genome', *Genome Res.* Vol. 8, pp. 1007–1021.
9. Giglio, S., Broman, K.W., Matsumoto, N. *et al.* (2001), 'Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements', *Am. J. Hum. Genet.* Vol. 68, pp. 874–883.
10. Giglio, S., Calvari, V., Gregato, G. *et al.* (2002), 'Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation', *Am. J. Hum. Genet.* Vol. 71, pp. 276–285.
11. Inoue, K. and Lupski, J.R. (2002), 'Molecular mechanisms for genomic disorders', *Annu. Rev. Genomics Hum. Genet.* Vol. 3, pp. 199–242.
12. Parsons, J.D. (1995), 'Miropeats: Graphical DNA sequence comparisons', *Comput. Appl. Biosci.* Vol. 11, pp. 615–619.
13. Altschul, S.F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.* Vol. 215, pp. 403–410.
14. Parsons, J.D. (1995), 'Improved tools for DNA comparison and clustering', *Comput. Appl. Biosci.* Vol. 11, pp. 603–613.
15. Skaletsky, H., Kuroda-Kawaguchi, T. *et al.* (2003), 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes', *Nature* Vol. 423, pp. 825–837.
16. Kong, A., Gudbjartsson, D.F., Sainz, J. *et al.* (2002), 'A high-resolution recombination map of the human genome', *Nat. Genet.* Vol. 31, pp. 241–247.
17. Osborne, L.R., Li, M., Pober, B. *et al.* (2001), 'A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome', *Nat. Genet.* Vol. 29, pp. 321–325.
18. Gimelli, G., Pujana, M.A., Patricelli, M.G. *et al.* (2003), 'Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions', *Hum. Mol. Genet.* Vol. 12, pp. 849–858.
19. Kurotaki, N., Imaizumi, K., Harada, N. *et al.* (2002), 'Haploinsufficiency of NSD1 causes Sotos syndrome', *Nat. Genet.* Vol. 30, pp. 365–366.
20. Kurotaki, N., Harada, N., Shimokawa, O. *et al.* (2003), 'Fifty microdeletions among 112 cases of Sotos syndrome: Low copy repeats possibly mediate the common deletion', *Hum. Mutat.* Vol. 22, pp. 378–387.
21. Kuroda-Kawaguchi, T., Skaletsky, H., Brown, L.G. *et al.* (2001), 'The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men', *Nat. Genet.* Vol. 29, pp. 279–286.
22. Repping, S., Skaletsky, H., Brown, L. *et al.* (2003), 'Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection', *Nat. Genet.* Vol. 35, pp. 247–251.
23. Small, K., Iber, J. and Warren, S.T. (1997), 'Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats', *Nat. Genet.* Vol. 16, pp. 96–99.
24. Eichler, E.E. (2001), 'Recent duplication, domain accretion and the dynamic mutation of the human genome', *Trends Genet.* Vol. 17, pp. 661–669.
25. Mefford, H.C. and Trask, B.J. (2002), 'The complex structure and dynamic evolution of human subtelomeres', *Nat. Rev. Genet.* Vol. 3, pp. 91–102.
26. Guy, J., Hearn, T., Crosier, M. *et al.* (2003), 'Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p', *Genome Res.* Vol. 13, pp. 159–172.
27. Ventura, M., Archidiacono, N. and Rocchi, M. (2001), 'Centromere emergence in evolution', *Genome Res.* Vol. 11, pp. 595–599.
28. Amor, D.J. and Choo, K.H. (2002), 'Neocentromeres: Role in human disease, evolution, and centromere study', *Am. J. Hum. Genet.* Vol. 71, pp. 695–714.
29. Eichler, E.E. (2001), 'Segmental duplications: What's missing, misassigned, and misassembled — and should we care?', *Genome Res.* Vol. 11, pp. 653–656.
30. Stankiewicz, P., Shaw, C.J., Dapper, J.D. *et al.* (2003), 'Genome architecture catalyzes nonrecurrent chromosomal rearrangements', *Am. J. Hum. Genet.* Vol. 72, pp. 1101–1116.
31. Barbouti, A., Stankiewicz, P., Nusbaum, C. *et al.* (2004), 'The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats', *Am. J. Hum. Genet.* Vol. 74, pp. 1–10.
32. Douglas, J., Hanks, S., Temple, I.K. *et al.* (2003), 'NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes', *Am. J. Hum. Genet.* Vol. 72, pp. 132–143.
33. Repping, S., Skaletsky, H., Brown, L. *et al.* (2003), 'Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection', *Nat. Genet.* Vol. 35, pp. 247–251.