 Human Molecular Genetics, 2002, Vol. 11, No. 17 1987–1995

Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome

Xavier Estivill[1,2,*], Joseph Cheung[1], Miguel Angel Pujana[2], Kazuhiko Nakabayashi[1], Stephen W. Scherer[1] and Lap-Chee Tsui[1]

[1]Program in Genetics & Genomic Biology, Research Institute, The Hospital for Sick Children, and Department of Molecular and Medical Genetics, University of Toronto, Toronto M5G 1X8, Canada; and [2]Genes & Disease Program, Genomic Regulation Center, Passeig Marítim 37–49, 08003 Barcelona, Catalonia, Spain

Received April 5, 2002; Revised and Accepted June17, 2002

We have explored the National Center for Biotechnology Information (NCBI) single nucleotide polymorphisms (SNPs) database for a correlation between the density of putative SNPs, as well as SNPs that map to different chromosomal locations (ambiguously mapped SNPs), and segmental duplications of DNA in chromosome regions involved in genomic disorders. A high density of SNPs (14.4 and 12.4 SNPs per kb) was detected in the low copy repeats (LCRs) responsible for the chromosome 17p12 duplication and deletion that cause peripheral neuropathies. None of the SNPs at the *PMP22* gene were ambiguously mapped, but 93% of the SNPs at LCRs mapped on both LCR copies, indicating that they are in fact variants in paralogous sequences. Similarly, a high SNP density was found in the LCR regions flanking the neurofibromatosis type 1 (*NF1*) gene, with 80% of SNPs mapping on both LCR copies. A high density of SNPs was found within LCR sequences involved in the deletions that mediate contiguous gene syndromes on chromosomes 7q11, 15q11–q13 and 22q11. We have analyzed the whole sequence of chromosome 22, which contains 14% of ambiguously mapped SNPs, and have found a good correlation between these SNPs and segmental duplications detected by BLAST analysis. We have identified several segments of ambiguously mapped SNPs, four corresponding to LCRs involved in the chromosome 22q11 microdeletion syndromes. Our data indicate that most SNPs in LCR segments are in fact paralogous sequence variants (PSVs), and suggest that a significant proportion of the SNPs in the NCBI database correspond to PSVs within segmental duplications of the human genome sequence.

INTRODUCTION

Several million single nucleotide polymorphisms (SNPs) have been detected in the process of determining the sequence of the human genome (1). Many of these SNPs are being used in association and linkage studies to identify genomic regions that potentially contain genes that are involved in complex diseases (2). Segmental duplications [also known as low copy repeats (LCRs) or duplicons] account for at least 5% of the human genome sequence (3,4). Segmental duplications are genomic regions with a high DNA sequence identity that have arisen during the process of evolution of the genome (5). Some segmental duplications are involved in deletion, duplication, inversion or translocation of chromosomal material, leading to diseases that have been designated as genomic disorders (6–8). In recent years, an increasing number of monogenic diseases and contiguous gene syndromes have been found to be due to segmental duplications (7,8). These segmental duplications have different length, copy number and orientation along the same chromosome, and some are located on different chromosomes. In some cases, these segmental duplications show conformations that confer a predisposition to undergo specific genomic rearrangements in the offspring (9–11). Furthermore, segmental duplications have also been proposed as predisposing factors for

*To whom correspondence should be addressed at: Program in Genetics and Genomic Biology, Research Institute, The Hospital for Sick Children, Room 9107C, Elm Wing Annex, The Hospital for Sick Children, 555 University Ave., Toronto, ON M5G 1X8 Canada. Tel: +416 8132152; Fax: +416 8138319; Email: estivill@genet.sickkids.on.ca; xavier.estivill@crg.es

some complex genetic disorders (12). Therefore, the identification and characterization of new segmental duplications in the human genome should assist in defining chromosomal regions of susceptibility for human genetic disease.

In the process of large-scale identification of SNP sequences, differences between sequenced genomic clones have been entered in SNP databases (http://www.ncbi.nlm.nih.gov/SNP) as putative SNPs. Since segmental duplications have a high degree of sequence identity at a level of over 97%, we postulated that SNPs in these locations could be present at a higher density than in other genomic locations and could also be mapping to several regions, corresponding to segmental duplications. To evaluate this hypothesis, we have explored the frequency and mapping information of putative SNPs within chromosome regions that contain well-characterized genomic disorder mutations. We report here a tight correlation between the density of SNPs, as well as SNPs that have been mapped to different chromosomal locations in the NCBI SNP database (dbSNP) by BLAST or electronic PCR (e-PCR) (defined as SNPs with 'ambiguous map location'), and segmental duplications in the human genome. A high density of ambiguously mapped SNPs with different mapping locations accounts for the segmental duplications that flank regions containing genomic disorder mutations. We confirm that most of the SNPs that lie within segmental duplications are in fact paralogous sequence variants (PSVs), which have been entered in the databases as SNPs. We have also analyzed the whole sequence of chromosome 22 and have found a strong link between PSVs and segmental duplications. Our data suggest that a significant proportion of SNPs in the NCBI database correspond to PSVs within segmental duplications of the human genome sequence.

RESULTS

In the process of identifying segmental duplications that could be potentially involved in genomic disorders, we have explored the SNP database (SNP build102, draft sequence build28) at the NCBI (http://www.ncbi.nlm.nih.gov/SNP). This database contains 2.13×10^6 refSNPs. About 80% of these SNPs have been mapped onto the human genome sequence assembly. dbSNP denotes the mapping information of SNPs that hit the assembled genome sequence once or twice. RefSNPs hitting the human genome sequence twice are annotated with a 'warning of ambiguous location'. Despite the fact that 90% of the mapped refSNPs have unique positions within the human genome sequence, as detected by e-PCR and BLAST, about 5% have been mapped ambiguously in two locations, and another 5% appear to reside at more than two sites.

We aimed to use the dbSNP information to identify segmental duplications in the draft sequence of the human genome. Our working hypothesis was that high densities of SNPs at specific chromosomal regions, and segments of DNA containing SNPs that have been mapped to more than one chromosomal region, might point to regions containing segmental duplications. We first analyzed the density and mapping information of SNPs in chromosomal regions that contain known segmental duplications, previously shown to be involved in human diseases. We found that segmental duplications that are responsible for well-defined Mendelian genomic disorders [the Charcot–Marie–Tooth type 1 (CMT1A) duplication, and the hereditary neuropathy with liability to pressure palsies (HNPP) deletion (13); and the deletions that cause about 15% of cases of neurofibromatosis type 1 (NF1) (14)] have high densities of SNPs, as compared to the regions that are rearranged in these diseases. In the case of the CMT1A/HNPP region, the total number of SNPs at proximal REP and distal REP are 291 and 338, respectively (SNP densities of 14.4 and 12.4 per kb, with an average distance between consecutive SNPs of 75 nucleotides); while for the genomic region that contains the *PMP22* gene, 50 SNPs have been reported (SNP density 1.4/kb, or one SNP every 700 nucleotides) (Fig. 1 and Table 1). While none of the SNPs at the *PMP22* gene were ambiguously mapped, 272 (93%) SNPs at proximal REP and 315 (93%) at distal REP showed ambiguous locations. Table 2 shows a segment of 4 kb of the proximal REP and distal REP containing 44 consecutive SNPs, of which only five and six are specific for each chromosome location respectively, while 33 of them map to both chromosome regions, indicating sequence divergence between the two LCR copies. Considering the whole sequence of these LCRs, about 90% of the reported SNPs are in fact PSVs, which have accumulated in these two genomic locations during evolution.

Similar results were obtained for the duplicons that flank the *NF1* gene. A 7-fold higher SNP density was detected in the duplicon regions, as compared to the *NF1* genomic region, with an average distance between SNPs of 280 nucleotides at the NF1 LCRs and 1800 nucleotides within the *NF1* gene. Also, more than 80% of the SNPs at the *NF1* duplicon regions were ambiguously mapped (Table 1), most corresponding to PSVs.

The SNP densities at the duplicon regions of CMT1A/HNPP and NF1 are much higher than those reported by large-scale SNP discovery projects (1,15), describing SNPs at average distances in a range between 900 and 1400 nucleotides. These high densities further confirm that most of these SNPs are PSVs. In fact, the proportion of bona fide SNPs (those that are not PSVs) at these LCRs is in the range of 1 every 500–1000 nucleotides, in agreement with the findings in inter-LCR regions and for the whole human genome sequence.

High densities of SNPs, most of them ambiguously mapped, were also detected within the regions that contain the LCR sequences involved in the deletions that mediate contiguous gene syndromes [Williams–Beuren syndrome (WBS) on 7q11 (11); Prader–Willi syndrome and Angelman syndrome (PWS/AS) (16) on 15q11–13; and DiGeorge syndrome (DGS) and velo-cardio-facial syndrome (VCFS) on 22q11 (17)] (Table 1). Despite the fact that most SNPs at the WBS LCRs were ambiguously mapped, the SNP density for the whole LCR region (over 400 kb) is not higher than the average for other regions in the genome. This is probably because some of these LCRs map to several regions along chromosome 7 (more than 10 locations), some corresponding to sequences related to the *PMS2* gene (not shown). Thus, SNPs that show multiple locations have probably not been reported and annotated in the NCBI dbSNP. Remarkably, between 80% and 93% of the SNPs that map to duplicon-containing regions of contiguous gene syndromes show ambiguous map locations by either BLAST

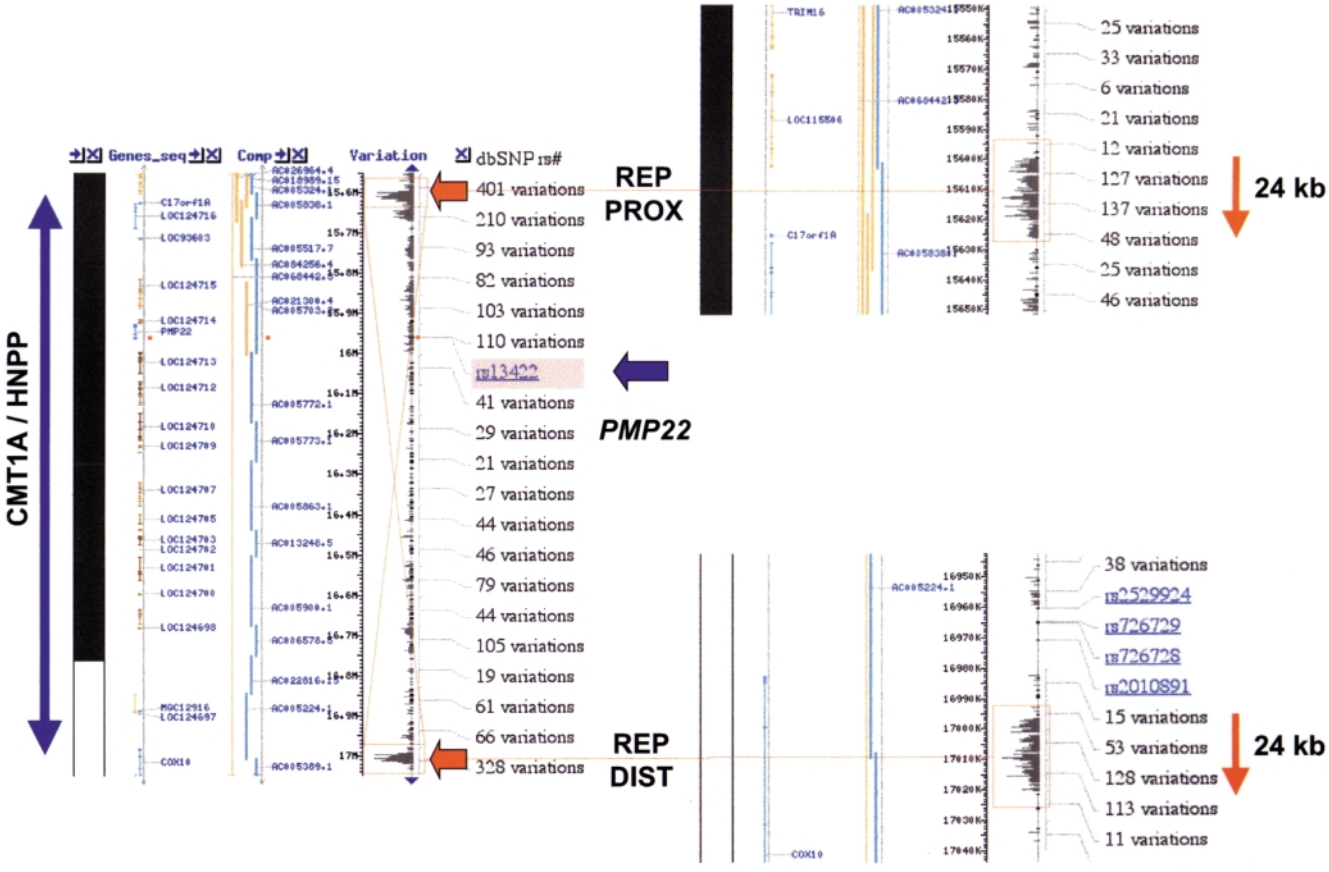

Figure 1. Map view (NCBI, build28) of human chromosome 17p12, showing the high density of SNPs within the proximal and distal REPs that flank the region that is duplicated or deleted in patients with Charcot–Marie–Tooth type 1A (CMT1A) or hereditary neuropathy with liability to pressure palsies (HNPP), respectively. The regions containing the REPs are boxed. Red arrows indicate the orientation of the REP sequences; the blue arrow shows the region that is either deleted or duplicated in patients with HNPP or CMT1A, respectively.

(18) or by e-PCR (19). Most of these putative SNPs should actually correspond to PSVs.

To further explore the relationship between regions containing segments of ambiguously mapped SNPs (or PSVs) and segmental duplications, we have analyzed the NCBI dbSNP for the entire sequence of human chromosome 22 (20). Figure 2 shows the plots corresponding to the SNP density and SNP mapping ambiguity for the whole sequence of the long arm of chromosome 22. Thirty-one DNA segments with SNP densities higher than 40 SNPs/10 kb were identified, showing an average distance between SNPs of 250 nucleotides. Remarkably, four segments have SNP densities of over 10 SNPs/kb (segments 9, 11, 14 and 15 in Fig. 2A).

Eighteen segments containing stretches of at least 10 ambiguously mapped SNPs per 10 kb were detected. We have found that some of these segments were clustered in several contiguous regions of ambiguously mapped SNPs. This is the case for the LCRs that cause the DGS/VCFS deletions (named A, B, C and D in 17), corresponding to segments 2–5 in Figure 2B and segments 6–11 in Figure 2A. Most SNPs in segments containing high densities of SNPs were ambiguously mapped, but some segments of high SNP density were not detected as ambiguously mapped in the NCBI dbSNP. This could be because they map to several regions, and their ambiguous chromosomal locations have not been annotated. Table 3 shows the characteristics of different segments containing ambiguously mapped SNPs. While the SNP density for the whole of chromosome 22 (only the 34 710 kb sequenced region was considered) is 1 SNP/kb (21), the SNP density in these segments ranges between 0.6 and 10.2/kb. These segments contain between 18.6% and 85.9% of ambiguously mapped SNPs, which is much higher than the 13.9% for the whole of chromosome 22.

We have used PipMaker (22) analysis to confirm the presence of segmental duplications in regions containing segments of ambiguously mapped SNPs. This allows us to define the orientation of the segmental duplications and the number of copies within each segment. Figure 3 shows three examples of the plots of the PipMaker analysis of three chromosome regions containing several segments of PSVs, which correspond to several complex segmental duplications. Segments 6–9 in Figure 2B, spanning about 900 kb between nucleotide positions 19 000 000 and 20 100 000, have a high number of copies of repeat sequences that are specific to this region of chromosome 22. Within the repeats that are scattered along this region of chromosome 22, there are several duplicated segments organized in tandem. This large genomic region contains all the genes of the variable, constant and joining segments of the immunoglobulin lambda segments (23). Interestingly, a

sequence located just at the end of segment 9 has one of the highest PSV densities of chromosome 22, with over 200 PSVs clustered in a region of less than 20 kb (Fig. 3, left panel). This corresponds to the constant regions of immunoglobulin lambda. Despite the presence of segmental duplications in sequences with high densities of PSVs, as detected by PipMaker analysis, some sequences located between these segmental duplications also display high densities of SNPs. Finally, a 70 kb sequence located between nucleotides 19 400 and 19 470 maps to at least 10 other chromosomal locations but not on chromosome 22 (not shown).

Segment 17 contains several LCRs organized in tandem within a region of less than 200 kb. Genes located in this region correspond to several protein sequences that show similarities with APOBEC1 (24). Segment 18 has five pairs of DNA LCRs, all in a tandem orientation and spanning 150 kb. These duplicated segments contain two different copies of the gene CGI-96 (25). In both cases, the segmental duplications contain large numbers of PSVs, while the sequences between and flanking the segmental duplications have the average density of SNPs for chromosome 22.

To further characterize the intrachromosomal segmental duplications of chromosome 22, its sequence assembly was masked for repeats and was compared against itself by BLAST analysis. The q11.1–q11.22 region contains a high density of intrachromosomal segmental duplications. While some correspond to the DGS/VCFS LCR sequences, several other segmental duplications have been identified in other regions of the chromosome. We have compared the genomic regions of chromosome 22 containing high number of SNPs and ambiguously mapped SNPs with segmental duplications identified by BLAST (Fig. 2C). Remarkably, all regions containing PSVs were identified by BLAST, and most regions identified by BLAST corresponded to PSVs. Additional regions predicted by the SNP analysis, but not detected by BLAST using the chromosome 22 DNA sequence against itself, likely correspond to interchromosomal segmental duplications. To evaluate these interchromosomal duplications, the chromosome 22 sequence was analyzed by MegaBLAST against each NCBI build28 chromosome sequence. All regions with ambiguously mapped SNPs not mapping to chromosome 22 were shown to map to other chromosome regions (Table 3).

DISCUSSION

The results presented here show that chromosome regions containing high densities of putative SNPs and ambiguously mapped SNPs correspond to segmental duplications of the human genome. We have demonstrated this association in the analysis of segmental duplications involved in Mendelian genomic disorders and contiguous gene syndromes, and in the analysis of the whole of chromosome 22, identifying several intrachromosomal segmental duplications.

All segmental duplications detected here have a high density of SNPs, most showing ambiguous locations in the NCBI dbSNP. Some of these regions correspond to previously described LCRs flanking regions involved in genomic disorders. Most of these SNPs are nucleotide variants within parologous sequences and are not truly single nucleotide polymorphisms. Accordingly, we have designated these SNPs as PSVs (paralogous sequence variants, as defined in 26).

Table 1. SNP density and ambiguously mapped SNPs (paralogous sequence variants or PSVs) at five chromosomal regions containing genomic disorder mutations

Disorder locus/gene	NT_Contig (Build 28)	Chromosome region	Length (kb)	Total SNPs *n*	SNPs/kb	Ambiguous SNPs *n* (%)
CMT1A/HNPP[a]						
Proximal REP	NT_010718	17p12	23.5	338	14.4	315/338 (93.2)
Distal REP	NT_010718	17p12	23.4	291	12.4	272/291 (93.5)
PMP22	NT_010718	17p22	34.8	50	1.4	0/50 (0)
NF1[b]						
LCR proximal	NT_024897	17q11	56.7	193	3.4	158/193 (81.9)
LCR distal	NT_010799	17q11	36.1	138	3.8	117/138 (84.8)
NF1	NT_010799	17q11	280.3	155	0.55	1/155 (0.6)
WBS[c]						
Proximal LCR	NT_007867	7q11	424.0	245	0.58	215/245 (87.8)
MDH2	NT_007867	7q11	600.0	137	0.23	1/137 (0.7)
PWS/AS[d]						
Proximal LCR	NT_010362	15q11	390.3	741	1.90	629/741 (84.9)
LOC123111-UB3A	NT_010178	15q11	129.8	32	0.25	0/32 (0)
DGS/VCFS[e]						
LCR-D	NT_011520	22q11	313.4	639	2.04	548/639 (85.8)
SERPIND1-LZTR1	NT_011520	22q11	219.7	359	1.63	0/359 (0)

[a]The duplicated or deleted region in CMT1A and HNPP is about 1.4 Mb; the length of the duplicon is 24 kb; other duplicon sequences are located telomeric and centromeric to these duplicons.
[b]The deleted region in NF1 is about 1.5 Mb; the length of the duplicon has been estimated at 85 kb, but only the regions with identity between the two LCR sequences were considered for calculations.
[c]The deleted region in WBS is about 1.6 Mb; the length of the duplicon is estimated to be 450 kb; only one of the duplicons has been considered for calculations.
[d]The deleted region in PWS/AS is about 3.5 Mb; the length of the duplicon has been estimated to be 350 kb; only one of the duplicons has been considered for calculations.
[e]The deleted region in DGS/VCFS ranges between 1.5 and 3 Mb; the lengths of the duplicons range between 15 and 350 kb; only one of the duplicons (LCR-D) has been considered for calculations.

The high densities of PSVs identified in some regions of chromosome 22 contrast with the average densites of SNPs detected in this chromosome of one SNP every 1000 nucleotides (Table 2) (21). Despite the strong concordance between density and ambiguity mapping of putative SNPs along this chromosome, this relationship was not complete, as some segments with high densities of SNPs were not described as mapping to several regions. These segments probably correspond to regions of the genome that contain sequences that have been under positive selection (1), to regions that have been studied in great detail in the search for genes involved in complex disorders, leading to the detection of a large number of SNPs (2), or to segmental duplications that have erroneously been assembled in the same genomic region. We have not performed a complete survey to clarify this issue, but this should be resolved as the assembly of the sequence is improved and these putative SNPs are characterized.

Table 2. Variations at 44 consecutive refSNPs (NCBI dbSNP) in a 4 kb region of the proximal and distal LCRs of the CMT1A/HNPP locus

Variation[a] (proximal REP)	Variation[a] (distal REP)	db_xref dbSNP	Variants Alleles
2687109	4083983	2856182	A/G
2687159	4084033	2530370	G/A
2687167	4084041	2530369/2601991	A/T
2687190	4084064	2530368	G/A
2687211	4084087	2259118	T/G
2687454	4084330	2257782	G/C
2687525	4084401	2257784	A/T
2687591	4084467	2257786	T/G
2687830	4084706	2257822	G/A
2688149	4085025	2257828	G/A
2688269	4085145	2257830	G/A
2688494	4085370	2530367	T/G
2688517	4085393	2530366	G/A
2688520	—	2530365	G/A
2688548	4085424	2530364	C/A
—	4085798	2515772	C/A
2689332	4086211	2530363	G/A
2689343	4086222	2906923	G/A
2689389	4086268	2601993	A/G
2689978	4086861	2856225	T/C
2690009	4086892	2530384	G/C
2690018	4086901	2856224	T/C
2690019	4086902	2530383	G/A
2690157	4087040	187344	G/C
2690174	4087057	2856223	G/A
2690234	4087117	3020880	A/G
2690426	4087309	2079617	G/C
—	4087436	2532450	T/C
—	4087437	2532449	G/A
—	4087785	2856128	G/T
2690536	—	2856125	A/T
2690564	—	2856126	C/T
2690846	—	2860279	T/C
2690907	4087785	2856128	G/T
2691114	4087992	2856129	A/T
2691196	4088074	2323742	C/T
2691309	4088187	2323743	C/T
—	4088188	2323744	A/G
2691327	4088210	2323745	G/T
2691331	4088214	2323746	C/T
2691348	4088231	2323747	C/T
2691357	4088240	2323748	A/G
—	4088270	2654302	A/G
2691393	—	2257237	C/T

[a]Nucleotide positions in *Homo sapiens* chromosome 17 working draft sequence segment NT_010718 (version NT_010718.8 GI:1860394).

The identification of genomic segments containing PSVs should be extremely useful as a first step in detecting segmental duplications. In the chromosome 22 analysis performed here, several segmental duplications were identified, in addition to those known to cause DGS/VCFS, inv dup(22), cat eye syndrome and the recurrent constitutional translocation t(11; 22) (17, 28–30). For instance, segment 10 corresponds to a segmental duplication located near the *BCR* gene. The links between the other segmental duplications and diseases or chromosome rearrangements have yet to be established. Even for some of the well-defined segmental duplications in the region of DGS/VCFS, it is not well established how the orientation of the duplicons predisposes to the different chromosome rearrangements. It is possible that specific orientations of LCRs in this region predispose to the 22q11 deletions that have been found in cases of schizophrenia (31). Detailed analysis of the location and orientation of these segmental duplications using BLAST and PipMaker should lead to the identification of potential targets for genomic mutations on this chromosome. The chromosome 22 BLAST results obtained in this study are similar to those reported previously (27) and correlate with the regions containing PSVs.

Mapping information of SNPs is not uniformly presented in different SNP databases. In addition to information for SNPs that map to a single and to two locations (ambiguously mapped SNPs), NCBI also provides a list of SNPs that have been mapped to more than two locations in the human genome (ftp://ftp.ncbi.nih.gov/snp/human/chr_rpts/Multi.txt.gz). These SNP entries were excluded from annotation to chromosomes in their dbSNP. Although we have not examined this dataset thoroughly, we suspect that these SNPs mapping to several locations would likely link to regions of segmental duplications and should therefore be referred to as PSVs as well. We have examined the Celera SNP database and found that allele frequencies were not assigned where duplicated regions are located. This is likely due to the detection of multiple alleles without knowledge that they are, in fact, sequences from highly paralogous sequences. Also, we have observed that Celera scaffolds often collapse in regions containing large LCRs. Other SNP databases specifically exclude all the SNPs that map to different chromosomes (http://snp.ims.u-tokyo.ac.jp/), although not those with multiple intrachromosomal locations, resulting in a loss of information regarding segmental duplications (32). Interestingly, about 1% of the SNPs reported on chromosome 21 identify more than two alleles, and they have not been included in the SNP database of this chromosome (16).

While the NCBI dbSNP is extremely useful in pointing to potential segmental duplications in the human genome characterization, SNPs at these locations should be used with caution when mapping disease loci. It has been suggested that at least 5% of the human genome contains segmental duplications (3,4). This figure is not far from the 10% of refSNPs that map to ambiguous locations (http://www.ncbi.nlm.nih.gov/SNP). Thus, a significant proportion of the SNPs in the NCBI database are likely to correspond to PSVs located within segmental duplications. These PSVs should be

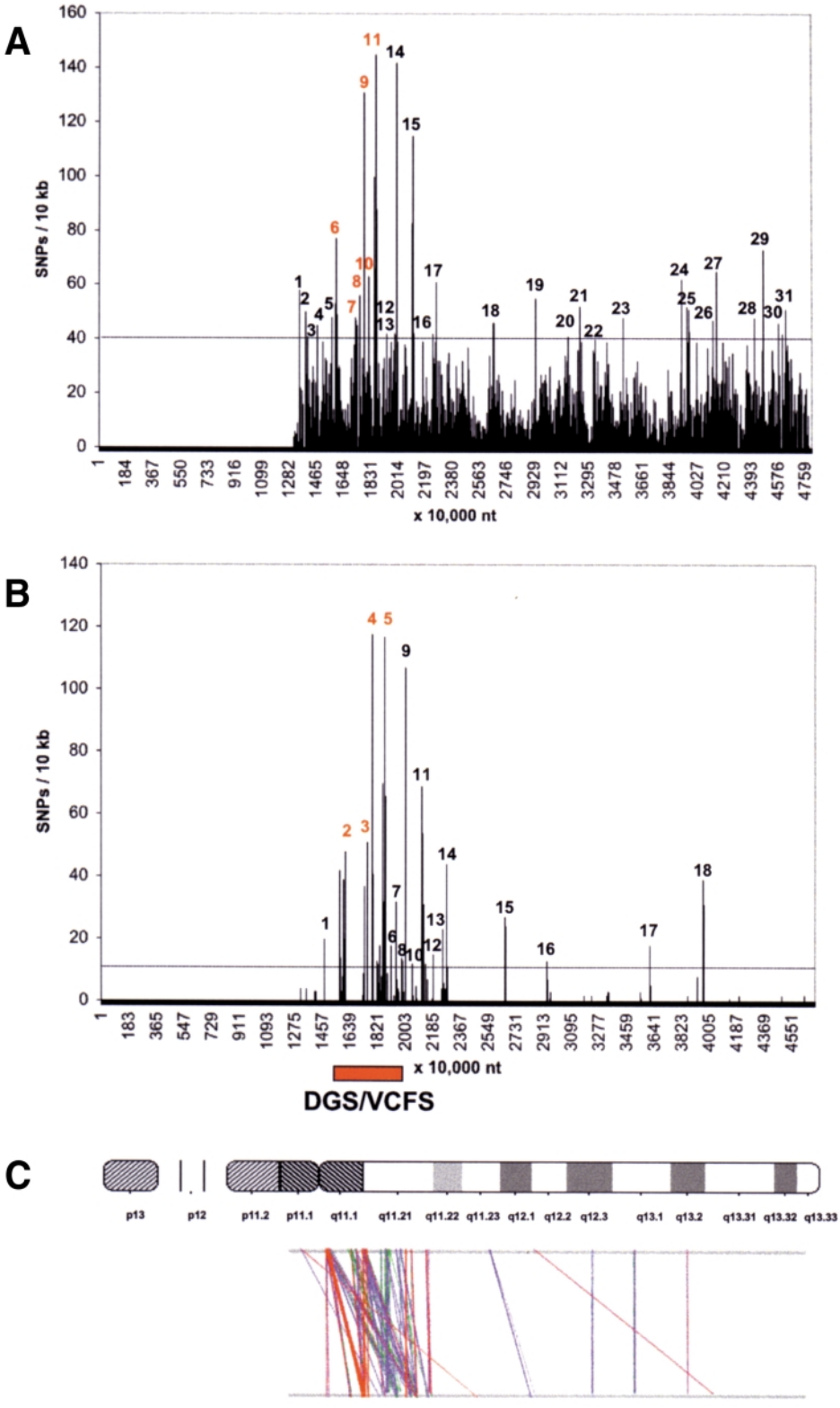

Figure 2. Identification of segmental duplications of human chromosome 22 by SNP and BLAST analysis. (**A**) Graphical view of the density of SNPs. A threshold of 40 SNPs/10 kb was used to define 31 segments of high SNP density (about 4-fold the average in the human genome). (**B**) Graphical view of ambiguously mapped SNPs along the chromosome. The numbers indicate 18 segments, each having at least 10 contiguous ambiguously mapped SNPs per 10 kb. The region containing the segments that correspond to the DGS/VCFS LCR sequences is boxed. (**C**) GenomePixelizer display of the intrachromosomal BLAST results from chromosome 22 (red, 100% sequence identity; purple, 95–99%; green, 90–94%). The correspondences between locations at the panel are only approximate. The region of the DGS/VCFS deletions is shown as a red bar, and SNP segments corresponding to LCRs involved in deletions are shown in red (2 to 5).

useful, in conjunction with BLAST approaches (3,27), to further characterize segmental duplications in the human genome. Finally, since some of these segmental duplications are likely polymorphic, PSVs could also be useful in tracing variability in copy number at these loci and in studying their potential link to human genetic disease.

MATERIALS AND METHODS

SNP viewing and density

NCBI MapViewer (http://www.ncbi.nlm.nih.gov//cgi-bin/Entrez/hum) was used to examine high-SNP-density regions from the build28 (February 2002) NCBI assembly, displaying the corresponding NT_contigs, sequence component, gene sequence and variation. A more detailed analysis was done for specific chromosomal locations that are known to contain well-characterized genomic disorder mutations and for the entire human chromosome 22 by using the NCBI human SNP dataset (http://www.ncbi.nlm.nih.gov/SNP). This dataset was obtained from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/snp/human/chr_rpts/) and was then imported into Microsoft Excel, as tables containing information about SNPs and their genomic positions. The total numbers of SNPs, as well as ambiguously mapped SNPs, from regions of our analysis were sorted by chromosome positions and then tabulated along a 10 kb window using the Excel commands ROUNDDOWN and COUNTIF. A threshold of 40 SNPs/10 kb was used to define regions with high SNP density (4-fold the average in the human genome). In the graphical view of ambiguously mapped SNPs along chromosome 22, a threshold of 10 contiguous ambiguously mapped SNPs per 10 kb was considered.

Chromosome 22 segmental duplications

The build28 NCBI Chromosome 22 assembly was obtained through the NCBI ftp site, which lacks the first 13 Mb of ambiguous nucleotides of the unknown p arm and centromere sequences. The entire chromosome sequence was subsequently masked for repeats using RepeatMasker (Smit and Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html). The masked chromosome sequence was compared against itself using MegaBLAST2 (http://www.ncbi.nih.gov/BLAST/) running on a local Unix server, and a BLAST report table was generated (using the –D option in the BLAST2 command). Modules of alignment, despite deletions and insertions caused by masked repetitive sequences, together forming a large contiguous alignment (to detect duplicons with sizes larger than 10 kb), were kept in the analysis. The results were parsed under the following criteria: BLAST results having >90% DNA sequence identity over >80 bp in length with expected value $<e^{-30}$. The parsed results were converted into a coordinate file as an input for display using GenomePixelizer (32). Regions containing segmental duplications identified by the BLAST method were compared to regions with high densities of SNPs and those with high numbers of ambiguously mapped SNPs by examining their corresponding coordinates in the same chromosome 22 sequence. The analysis only considered the intrachromosomal segmental duplications of chromosome 22. A graphical view of ambiguously mapped

Table 3. Segments of human chromosome 22 DNA sequence containing at least 10 contiguous ambiguously mapped SNPs (PSVs)

Segment	Region	Beginning (kb)	End (kb)	Length (kb)	Total SNPs	Total SNPs/kb	Ambiguously mapped SNPs		
							n	%	Location
1	22q11.1	14 670	14 680	10	44	4.4	20	50.0	Inter
2	22q11.1	15 680	16 070	390	579	1.5	301	52.0	Intra
3a	22q11.21	17 290	17 350	60	189	3.1	88	46.6	Intra
3b	22q11.21	17 500	17 540	40	85	2.1	73	85.9	Intra
4	22q11.21	17 820	17 910	90	369	4.1	285	77.2	Intra/Inter
5a	22q11.21	18 160	18 170	10	62	6.2	13	21.0	Intra
5b	22q11.21	18 200	18 230	30	83	2.8	22	26.5	Intra/Inter
5c	22q11.21	18 270	18 720	450	1313	2.9	1036	78.9	Intra
6	22q11.21	19 100	19 120	20	58	2.9	32	55.2	Intra
7	22q11.21	19 370	19 380	10	41	4.1	32	78.0	Intra
8	22q11.21	19 760	19 850	90	182	2.0	56	30.8	Intra
9	22q11.21	20 030	20 050	20	204	10.2	130	63.7	Intra
10	22q11.21	20 440	20 490	50	71	1.4	27	38.0	Inter
11a	22q11.22	21 070	21 210	140	545	3.9	392	71.9	Intra
11b	22q11.22	21 370	21 380	10	16	1.6	12	75.0	Intra
11c	22q11.22	21 440	21 500	60	35	0.6	17	48.6	Intra
12	22q11.22	21 810	21 870	60	56	0.9	34	60.7	Intra
13	22q11.22	22 420	22 520	100	191	1.9	74	38.7	Inter
14	22q11.22	22 600	22 810	210	340	1.6	76	22.3	Intra
15	22q12.1	26 550	26 630	80	167	2.1	73	43.7	Inter
16	22q12.2	29 340	29 400	60	145	2.4	27	18.6	Intra
17	22q13.1	36 070	36 170	100	107	1.1	62	57.9	Inter
18	22q13.31	39 580	39 680	100	250	2.5	162	64.8	Intra
All	22q	13 110	47 820	34 710	34 491	1.0	4800	13.9	–

Some segments shown in Figure 2 are grouped in subsegments (3a, 3b, etc.). Intra, intrachromosomal; Inter, interchromosomal.

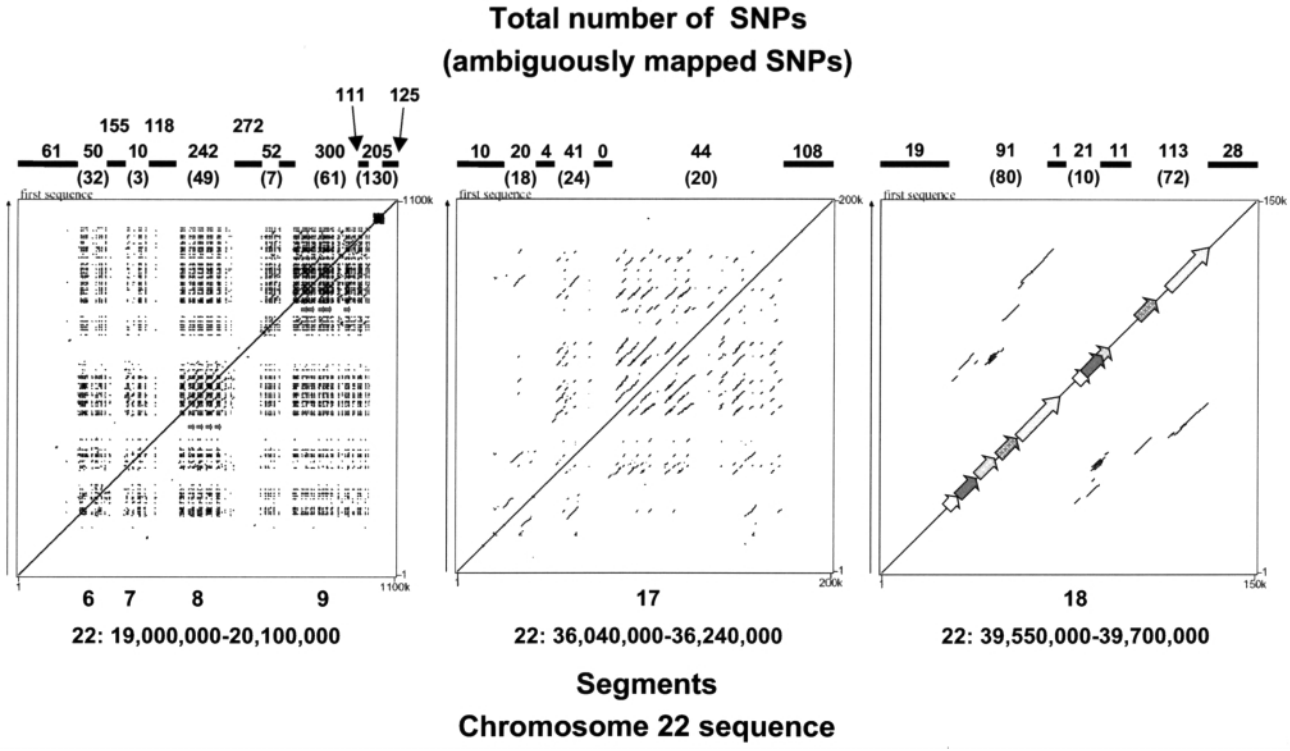

Figure 3. Dot-plot of the PipMaker alignment of three genomic sequences of human chromosome 22 with high densities of PSVs, showing regions that contain segmental duplications. The left panel shows a region of 1100 kb (nucleotides 19 000 000–20 100 000), corresponding to segments 6–9 in Figure 2B. The high density of segments showing alignment is mainly due to the presence of a repeat sequence that is specific to this chromosome region. Within the region, there are several segmental duplications oriented in tandem. The end of segment 9 shows a small segment with a high level of SNPs, most ambiguously mapped. The middle panel shows the plot of sequence 36 040 000–36 240 000, corresponding to segment 17 in Figure 2B. The right panel illustrates the plot of sequence 39 550 000–39 700 000, corresponding to segment 18 in Figure 2B. The horizontal black bars correspond to the regions without ambiguously mapped SNPs.

SNPs along the chromosome was compared with the GenomePixelizer display of the intrachromosomal BLAST results from chromosome 22 (red, 100% sequence identity; purple 95–99%; green, 90–94%).

Interchromosomal analysis

The build28 NCBI human genome assembly is available through the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). Each of these chromosome assemblies was reordered according to the assembly table provided by NCBI from the file seq_contig.md, since the sequences were stored in unordered multi-fasta files. Then, pairwise comparisons were made between the masked version of chromosome 22 and the masked version of each of the NCBI chromosome assemblies by MegaBLAST2. All results were parsed and generated using the same criteria as for intrachromosomal detection.

Characterization of duplicons

Sequences of specific regions containing segmental duplicons with high SNP densities were obtained from their respective NT_contigs or assembled from BAC/PAC sequences from Genbank. Each sequence was then repeat-masked and aligned against itself using PipMaker (22) (tables containing information on repeat sequences were submitted to the PipMaker server). The size, orientation and structure between segmental duplications can be interpreted by using the PIP and Dot-plot output generated by PipMaker.

ACKNOWLEDGEMENTS

This work was supported by the Departament d'Universitats Recerca i Societat de la Informació (DURSI) and the Departament de Sanitat to X.E. and the Canadian Institutes of Health Research (CIHR) to S.W.S. and L.-C.T. X.E. is Senior Scientist of the Centre de Regulació Genòmica (CRG) and a Visiting Scientist of the Hospital for Sick Children Research Institute. L.-C.T is a Distinguished Scientist of CIHR and Sellers Chair of Cystic Fibrosis Research. S.W.S. is a Scholar of CIHR and International Scholar of the Howard Hughes Medical Institute.

REFERENCES

1. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
2. Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S. *et al.* (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.*, **29**, 223–228.

3. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. and Eichler, E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
4. Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., Furey, T.S., Kim, U.J., Kuo, W.L., Olivier, M. *et al.* (2001) Integration of the cytogenetic landmarks into the draft sequence of the human genome. The BAC resource consortium. *Nature*, **409**, 953–958.
5. Eichler, E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.*, **17**, 661–669.
6. Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.*, **14**, 417–422.
7. Stankiewicz, P. and Lupski, J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, **18**, 74–82.
8. Emanuel, B.S. and Shaikh, T.H. (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat. Rev. Genet.*, **2**, 791–800.
9. Giglio, S., Graw, S.L., Gimelli, G., Pirola, B., Varone, P., Voullaire, L., Lerzo, F., Rossi, E., Dellavecchia, C., Bonaglia, M.C. *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.*, **68**, 874–883.
10. Jobling, M.A., Williams, G.A., Schiebel, G.A., Pandya, G.A., McElreavey, G.A., Salas, G.A., Rappold, G.A., Affara, N.A. and Tyler-Smith, C. (1998) A selective difference between human Y-chromosomal DNA haplotypes. *Curr. Biol.*, **8**, 1391–1394.
11. Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe. T., Cox, S., Tsui, L.C. and Scherer, S.W. (2001) A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nat. Genet.*, **29**, 321–325.
12. Gratacos, M., Nadal, M., Martin-Santos, R., Pujana, M.A., Gago, J., Peral, B., Armengol, L., Ponsa, I., Miro, R., Bulbena, A. and Estivill, X. (2001) A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell*, **106**, 367–379.
13. Reiter, L.T., Murakami, T., Koeuth, T., Pentao, L., Muzny, D.M., Gibbs, R.A. and Lupski, J.R. (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat. Genet.*, **12**, 288–297.
14. Lopez-Correa, C., Dorschner, M., Brems, H., Lazaro, C., Clementi, M., Upadhyaya, M., Dooijes, D., Moog, U., Kehrer-Sawatzki, H., Rutkowski, J.L. *et al.* (2001) Recombination hotspot in NF1 microdeletion patients. *Hum. Mol. Genet.*, **10**, 1387–1392.
15. Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
16. Amos-Landgraf, J.M., Ji, Y., Gottlieb, W., Depinet, T., Wandstrat, A.E., Cassidy, S.B., Driscoll, D.J., Rogan, P.K., Schwartz, S. and Nicholls, R.D. (1999) Chromosome breakage in the Prader–Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.*, **65**, 370–386.
17. Shaikh, T.H., Kurahashi, H., Saitta, S.C., O'Hare, A.M., Hu, P., Roe, B.A., Driscoll, D.A., McDonald-McGinn, D.M., Zackai, E.H., Budarf, M.L., Emanuel, B.S. (2000) Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.*, **9**, 489–501.
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Schuler, G.D. (1998) Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol.*, **11**, 456–459.
20. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
21. Dawson, E., Chen, Y., Hunt, S., Smink, L.J., Hunt, A., Rice, K., Livingston, S., Bumpstead, S., Bruskiewich, R., Sham, P. *et al.* (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res.*, **11**, 170–178.
22. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.
23. Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J.L., Wang, J. and Shimizu, N. (1997) One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.*, **7**, 250–261.
24. Madsen, P., Anant, S., Rasmussen, H.H., Gromov, P., Vorum, H., Dumanski, J.P., Tommerup, N., Collins, J.E., Wright, C.L., Dunham, I. *et al.* (1999) Psoriasis upregulated phorbolin-1 shares structural but not functional similarity to the mRNA-editing protein apobec-1. *J. Invest. Dermatol.*, **113**, 162–169.
25. Lai, C.H., Chou, C.Y., Ch'ang, L.Y., Liu, C.S. and Lin, W. (2000) Identification of novel human genes evolutionarily conserved in Caenorhabditis elegans by comparative proteomics. *Genome Res.*, **10**, 703–713.
26. Horvath, J.E., Schwartz, S. and Eichler, E.E. (2000) The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Res.*, **10**, 839–852.
27. Bailey, J.A., Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocchi, M. and Eichler, E.E. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.*, **70**, 83–100.
28. Mears, A.J., el-Shanti, H., Murray, J.C., McDermid, H.E. and Patil, S.R. (1995) Minute supernumerary ring chromosome 22 associated with cat eye syndrome: further delineation of the critical region. *Am. J. Hum. Genet.*, **57**, 667–673.
29. McTaggart, K.E., Budarf, M.L., Driscoll, D.A., Emanuel, B.S., Ferreira, P. and McDermid, H.E. (1998) Cat eye syndrome chromosome breakpoint cluster: identification of two intervals also associated with 22q11 deletion syndrome breakpoints. *Cytogenet. Cell Genet.*, **81**, 222–228.
30. Kurahashi. H. and Emanuel, B.S. (2001) Long AT-rich palindromes and the constitutional t(11;22) breakpoint. *Hum. Mol. Genet.*, **10**, 2605–2617.
31. Karayiorgou, M., Morris, M.A., Morrow, B., Shprintzen, R.J., Goldberg, R., Borrow, J., Gos, A., Nestadt, G., Wolyniec, P.S., Lasseter, V.K. *et al.* (1995) Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl Acad. Sci. USA*, **92**, 7612–7616.
32. Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. and Nakamura, Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.*, **30**, 158–162.
33. Kozik, A., Kochetkova, E. and Michelmore, R. (2002) GenomePixelizer—a visualization program for comparative genomics within and between species. *Bioinformatics*, **18**, 335–336.