Human Molecular Genetics, 1996, Vol. 5, No. 5 659–668

Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the *RCP/GCP* and *G6PD* loci

Ellson Y. Chen[1,*], Massimo Zollo[1,+], Richard Mazzarella[2], Alfredo Ciccodicola[3], Chun-nan Chen[1], Lin Zuo[1], Cheryl Heiner[1], Frank Burough[2], Monica Repetto[1], David Schlessinger[2] and Michele D'Urso[3]

[1]Advanced Center for Genetic Technology, Applied Biosystems Division of Perkin Elmer Corp., 850 Lincoln Center Drive, Foster City, CA 94404, USA, [2]Department of Molecular Microbiology and Center for Genetics in Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA and [3]International Institute of Genetics and Biophysics, Via Marconi 10, 80125 Naples, Italy

Received December 15, 1995; Revised and Accepted February 8, 1996

DNA comprising 219 447 bp was sequenced in nine cosmids and verified at >99.9% precision. Of the standard repetitive elements, 187 Alus make up 20.6% of the sequence, but there were only 27 MERs (2.9%) and 17 L1 fragments (1.6%). This may be characteristic of such high GC (57%) regions. The sequence also includes an 11.3 kb tract duplicated with 99.2% identity at a distance of 38 kb. The region is 80–90% transcribed and 12.5% translated. Thirteen known genes and their exon–intron borders are all accurately predicted at least in part by GRAIL programs, as are six additional genes. From centromere to telomere, the orientation of transcription varies among the first eight genes, then runs centromeric to telomeric for the next five, and is in the opposite sense for the last six. Eighteen of the 19 genes are associated with CpG islands. Two islands are exact copies in the 11.3 kb repeat units, and could thus give rise to double dosage levels of an X-linked gene. Another island is associated with two genes transcribed in opposite directions. From the sequence data, three genes and their exon structure are inferred. One of them, previously associated with *HEX2*, is shown to be a different gene unrelated to hexokinases; a second gene, previously known by an EST, is plexin, from its 65.5% identity with the *Xenopus* analog; and a third is a subunit of a vacuolar H-ATPase, and is named *VATPS1*.

INTRODUCTION

Bernardi (1) was the first to suggest that gene concentration in the human genome is correlated with regional GC content in stretches of DNA ('isochores') that span as much as several megabases. The suggestion has been supported by the mapping of large numbers of CpG islands [regions enriched for CpG dinucleotides (2,3)] and high GC isopycnic fractions of DNA (4) to subregions of cytogenetic bands. Most extreme are the 'T' bands, usually subtelomeric (5), which contain the 3–4% of highest GC DNA (50–60%). Consistent with this enrichment, Antonarakis (6) has noted that 80% of 175 genes that have been mapped by linkage, cloned, and associated with known phenotypes are in high GC isochores.

In more direct regional analyses, chromosome 21 (7) and the long arm of the X chromosome (8) show subtelomeric regions of high GC and high gene content. In Xq28, a region of ~1.5 Mb of very high GC DNA (8) lies precisely in a region where genetic linkage mapping had indicated a high concentration of disease genes (9), and where a large number of CpG islands and corresponding genes had been recovered (10–12).

Because the yield of genetic information should be high, such regions become prime candidates for long-range sequencing. We have analyzed the most GC-rich portion of Xq28, 219.4 kb between the *RCP/GCP* (color vision) and glucose 6-phosphate dehydrogenase (*G6PD*) loci. The results show that the region is indeed highly transcribed and enriched for genes and certain repetitive elements, especially Alu sequences. The genes include 13 which are both predicted and have also been confirmed by reports of at least some cDNA sequence. Another six gene candidates are predicted by computer-assisted methods and CpG islands. Thus, the efficiency of gene detection through genomic sequencing now rivals other methods and provides additional candidates for the many diseases mapped to Xq28 (9).

RESULTS

With a sequencing precision >99.9% for random shotgun sequencing of a set of seven cosmids (see Fig. 1 and Materials and Methods), the combination of known and predicted repetitive sequences and genes provide a consensus view of the information content of the region.

*To whom correspondence should be addressed
+Present address: Telethon Institute for Genetics and Medicine, Via Olgetina 58, 20132 Milano, Italy

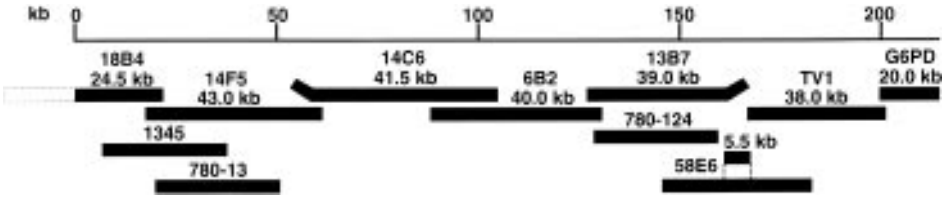


Figure 1. Cosmid contig across the sequenced region. The seven cosmids sequenced and three others (1345, 780-13 and 780-124; see Materials and Methods) are shown. Cosmid 18B4 was only partly sequenced; the other portion will be reported from sequencing efforts at the Sanger Center. Cosmid 58E6 was used to amplify a 5.5 kb bridge fragment between 13B7 and TV1.

Sequencing a high GC region: precision and verification

In a first effort, we sequenced and did a preliminary analysis of 52 kb in and centromeric to *G6PD* (13). Here we extend the analysis and carry it through the remaining 168 kb. In this region of high GC content, some templates showed the superposition of a series of peaks in sequencing chromatograms ('compression'). In some instances, the compression zone was resolved by readings from the complementary strand, which showed no comparable pileup of peaks. In extreme cases, compression was alleviated using dye-terminator sequencing with the nucleotide analog dITP replacing dGTP, or using dye-terminator sequencing with Sequenase and dNTPαS substituted for dNTP (14). During the editing of data, results from additional dye-terminator reactions also helped to resolve otherwise ambiguous base identifications for certain homopolymer-containing regions.

Stretches of poly(dA)/poly(dT) or regions containing short tandemly repeated sequences were often problematical. Presumably because polymerases sometimes stutter at repetitive sequences, the exact numbers of As or Ts were difficult to determine in extreme cases, and sequences beyond the homopolymer tracts were scrambled. Sequencing in both orientations and sequencing with oligo(dA) or oligo(dT) primers (see Materials and Methods) helped to resolve these problems. In addition, for such regions, some templates were easier to sequence than others. Typically, runs of 10–12 As or Ts could be fully resolved with PCR products as templates, runs of 20–25 with single-stranded M13 DNA, and runs of 40–50 with double-stranded plasmid DNA.

Comparisons of results with dye-primer and dye-terminator reactions were often useful, but did not always agree. For example, in the vicinity of base #145 214, dye-primer reactions inferred a run of 28 Ts, whereas dye-terminator data showed two Gs and a C interrupting the stretch of Ts. Using dye-primer chemistry, signals weaken progressively across such a region, but peak intensities are more even and are therefore less sensitive to contaminating noise. Thus, we opted for the sequence as inferred from dye-primer runs. In any case, essentially all such tracts occur at the end of Alu elements, and the precise sequence is thus likely to be relatively unimportant for biological function.

Where Alu repetitive elements occurred in tandem or in clusters, sequence assembly was more difficult but was generally possible based on detailed comparisons of Alu sequences, monitoring branches in the assembled sequence and using additional mapping information such as the locations of known restriction sites. At one location, however, at the junction of cosmids 6B2 and 13B7 (base #126 300–128 100 in the sequence), a group of Alu elements produced a variety of apparent branches and necessitated extraordinary effort to achieve closure. In a first effort, sequence redundancy was increased, but the array of sequences only became progressively more divergent. The sequence was finally resolved by amplifying an additional 1.4 kb PCR product that bridges the unstable area, subcloning it into a pUC plasmid vector and sequencing it by primer walking. It then became clear that a number of the M13 subclones had been rearranged. The instability and variable sequence content of clones can probably be attributed to the tendency of tandem and inverted Alu repeats to recombine.

Of the group of 12 cosmids sequenced, six were overlapped to obtain the consecutive 219 447 bp region schematized in Figure 1. Unexpectedly, despite considerable earlier characterization of the cosmid clones, two of those six contained chimeric segments at one end: 14C6 had an adventitious 3.1 kb at its left (centromeric) end; 13B7, 1.3 kb at its telomeric end. These insert-end fragments clearly branched outside of the consensus contigs, and presumably derived from cloning artifacts during the preparation of cosmid libraries.

Clone instability was even more patent among three other cosmids that were sequenced. Cosmids 1345, 780-13 and 780-124 (see Fig. 1) have been proven to contain substantial rearrangements, two of them (1345 and 780-13) mediated by recombination involving the near-perfect 11.3 kb repeats (Zollo *et al.*, in preparation).

The precision (reproducibility) of the sequencing results was earlier assessed at 99.9% for the 52 kb in cosmid TV1 (13). Multiple cosmid coverage and higher sample purity increased overall precision to >99.9% for the other 168 kb. Consistent with this level of precision, sequence comparisons of several kilobases of regions of overlap between cosmids, or between cosmid and PCR-amplified DNA segments, have detected no differences. Comparable concordance was also found in comparisons of the genomic sequence with encoded cDNA segments.

Computer-assisted sequence analysis

Repetitive sequences and GC/CpG content. Runs of more than four consecutive di, tri, tetra and pentanucleotide repeats were noted throughout the region, and are tabulated in the GenBank entry for this sequence [(poly(A) tracts, for example, are found at the end of each of the large number of Alu elements]. We note that repeat stretches of the dinucleotide GT, the prime repetitive sequence used in developing markers that detect polymorphism (15), exceed $(GT)_5$ at four sites in the region [at residues 38 762 (six repeats), 126 352 (seven repeats), 197 031 (10 repeats) and 197 051 (seven repeats)].

Alu sequences (16) are very abundant. Of the 231 repetitive elements detected, 187 are half ('monomer') or whole Alus, summing to 20.4% of the 220 kb. They include representatives of six subfamilies (16), and are tabulated in the GenBank entry. In contrast, only 17 L1 fragments were seen, deriving from seven of the

17 subfamilies (17), and adding up to 1.6% of the total sequence content. The positions of Alu (blue) and L1 sequences (orange), and their forward or reverse directions, are indicated in Figure 1.

The 27 repetitive sequences that are not Alu- or L1-like include 16 types of elements (18), with MER2 (five occurrences), MER22 (1), MER3 (1), MER 33 (1), MER 42C (1), MER MIR (2), MER MIR2 (1), MER MLR (3), MER MLT1 (one type A, three type C, and two type F), MLT2C2 (2), MSTC (1), SVA (1), THR (1) and LTR5 (1).

In addition, an 11.3 kb sequence is duplicated (99.2% identity), as described further below.

Distribution of GC and CpG islands. In contrast to other gross features of sequence, those related to GC distribution are of direct interest for the determination of gene content. More than 90% of the sequence contains 50–60% GC, which is characteristic of the most GC-rich 'H3' fraction of the genome (1). At intervals, however, sharp peaks with values of GC in excess of 70% over at least 1 kb were observed, and indicated the likely locations of 'CpG islands' (2,3).

The probable sites of CpG islands were even more sharply delineated in plots of the local concentrations of CpG dinucleotide (Fig. 2). The CpG content in typical portions of the genome is 10-fold less than that expected from overall GC content [i.e. 10-fold less than the product of (C)(G) in a sequence tract]. Peaks observed above a background level 3-fold less than the theoretical expectation were taken as an indication of a concentration of CpG dinucleotides. Because the peaks are relatively discrete, 17 could be scored with ease (numbered in Fig. 2).

CpG islands were also assessed by the clustering of restriction sites for enzymes with CpG in their recognition sequences (19), which cut DNA at relatively rare sites. The predicted restriction sites for five such enzymes, *Bss*HII, *Mlu*I, *Eag*I, *Sac*II and *Not*I, are given in the GenBank entry. Based on the near coincidence of at least two rare-cutter sites, 15 islands might be detected in the sequence in this way. At least 12 were actually found in an analysis of YACs in the region (12), and 16 in a study of smaller clones from the region (20).

Content of known and predicted genes

The region contains 13 genes, shown along the map in Figure 2, for which at least some cDNA sequence has been reported. References and GenBank accession numbers for the most complete analyses are given below, including those that provide full cDNA sequences for *G6PD, GdX, P3, GDI, QM,* and *FLN*.

The 13 proven genes are all associated with CpG islands. One island (CpG island 8 in Fig. 2) lies at the 5′ ends of two genes, *XAP-2* and the DNase I-like transcript, which are transcribed in opposite directions (Fig. 2 and see below).

Five CpG islands that are not associated with previously known genes include the first (CpG island 1 in Fig. 2); one in each of the two 11.3 kb repeat regions (CpG islands numbers 2 and 6 in Fig. 2); one between *EMD* and *FLN* (CpG island number 4 in Fig. 2); and one between *XAP-6* and *GdX* (CpG island number 13 in Fig. 2).

Computer-assisted predictions of gene content (Materials and Methods and below) are in agreement with the estimates obtained by counting CpG islands. Combining the positions of known and predicted genes, the cohort of known and predicted exons are displayed in Figure 2. The census of putative genes rises to 19 with the addition of one that is not associated with any CpG island (see below). In all, six new gene candidates are predicted for further testing. They are referred to in the tally below in order from centromere to telomere, and are labeled in Figure 2 as *CVG-1–CVG-6.*

CVG-1. BLAST searches identify a region with some homology to G-CSF and several other proteins near the very beginning of the sequence. This gene is associated with the weak CpG island number 1 in Figure 2.

CVG-2. Associated with CpG island 2, three putative exons lie in the first of the duplicated 11.3 kb sequences. Predicted exons are scored as only moderate to good, but such predictions are usually substantiated (for example, four of the six exons accurately predicted in the emerin gene had 'good' scores).

FLN. An analysis of exon–intron borders has been published for filamin (*FLN*, X53416; ref. 21), an actin-binding protein. It is associated with CpG island 3. GRAIL analysis predicts all 48 exons already reported (22), and suggested another exon in the first intron that might occur in alternatively spliced forms. Compared with the cDNA sequence, one additional exon, number 30, just 24 bp long, was neither detected in earlier determinations nor predicted by GRAIL.

CVG-3. Associated with CpG island 4 and three exons predicted by GRAIL.

EMD. The emerin gene, associated with CpG island 5, has been completely analyzed in published reports (X82434; ref. 23).

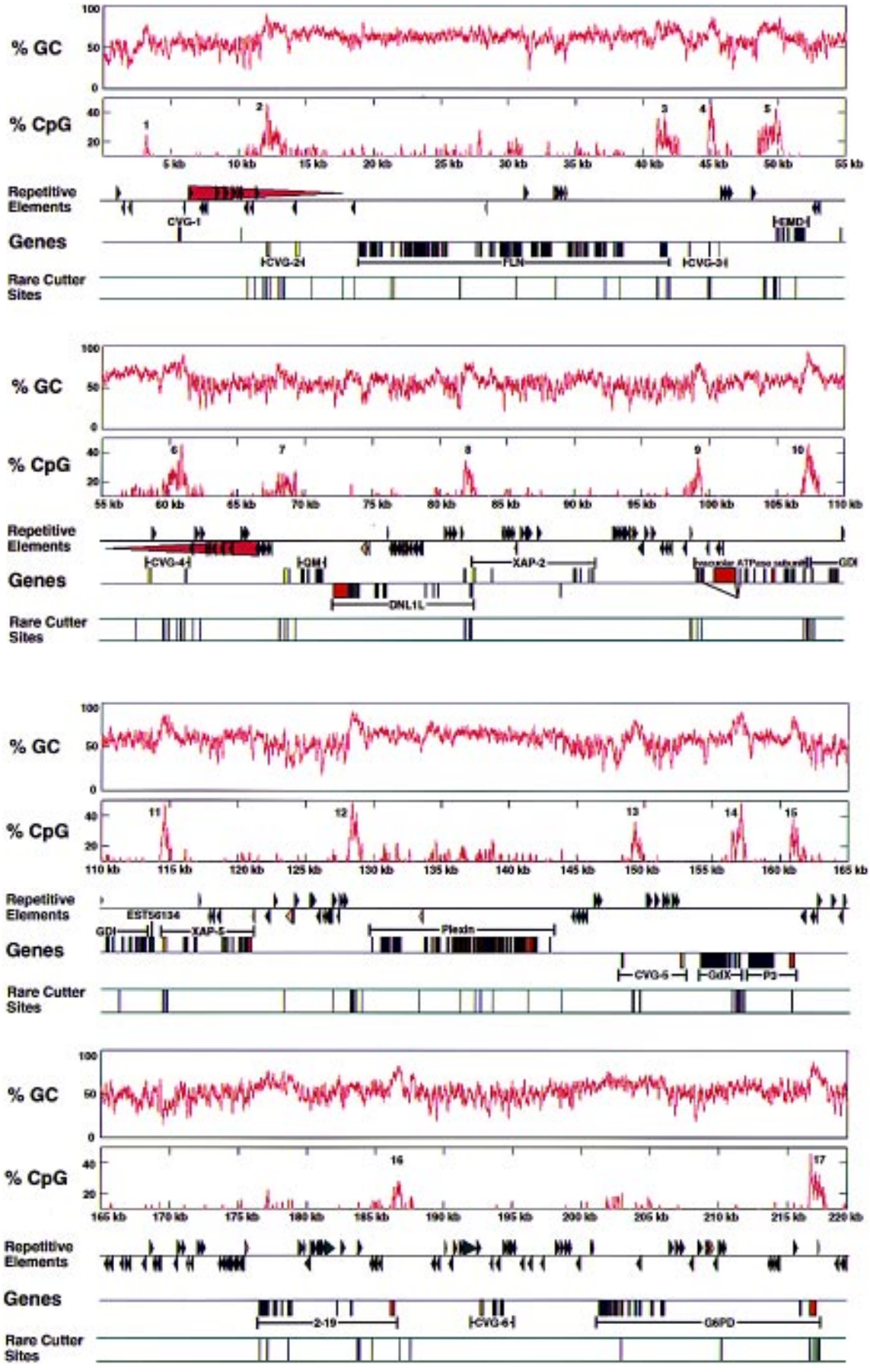
CVG-4. The second 11.3 kb repeat contains the predicted gene *CVG-4*, associated with CpG island 6. Because these two repeat segments have inverted orientations, *CVG-4* would be transcribed off the opposite strand from *CVG-2*.

QM. Starts at CpG island 7. M73791 (M. Kroepelin, submitted to GenBank) completely defines the exon structure, though once again two additional potential exons are predicted 5′ of the currently known exon 1 (Fig. 2).

DNL1L. Begins at CpG island 8. This encodes a DNase I-like protein [L40823 of ref. 24; the gene is the same as *XAP-1* (X74606; all *XAP* genes are in ref.11); and *G4.8* (X87196, ref. 10)]. It includes the nine exons in the cDNA sequence reported as L40823 and one candidate exon predicted in the promoter region, two in intron 1, and one in intron 2. Another possible exon in intron 1 has a GRAIL score of good, but runs in the sense opposite to the *DNL1L* cDNA and is likely to be a false positive.

XAP-2. This 16.6–17 kb portion of DNA encodes a series of exons reported in cDNA *XAP-2* [X74607; also *G4.5* (X87198, ref. 10)]. Like *DNL1L*, the transcript would begin from CpG island 8, with <80 nucleotides between the two start sites. The region also includes four additional candidate exons. One of those lies in the CpG island at the 5′ end and is a likely part of the gene; the other three, at the 3′ end, could easily fit into a putative extension of the transcript. (The first intron of *XAP-2* is >3 kb in length, and contains a fifth possible exon, but it is encoded on the opposite strand and is likely to be a false positive.)

VATPS1. Telomeric to *XAP-2* is a gene starting at CpG island 9. We identify this gene here as a subunit of the vacuolar H-ATPase (D16469; ref. 25). The sequence matches that of a microclone


% GC
% CpG
Repetitive Elements
Genes
Rare Cutter Sites
CVG-1
HEMD-I
HCVG-2I
FLN
HCVG-3I
HCVG-4
HOM
XAP-2
GDI
DNL1L
EST56134
XAP-5
Plexin
CVG-5
GdX
P3
2-19
HCVG-6I
G6PD

previously derived from Xq28. The earlier general localization to Xq28 is thus specified precisely. The cDNA sequences confirm the GRAIL predictions of nine exons with an additional possible exon at the 3′ end. Thus the exon–intron borders are now established.

This gene can be linked to a previously reported expressed sequence tag (EST), for the gene *XAP-3* (X74605; also *16A*, X87196; ref. 10). From the genomic sequence, the EST is found to lie in the 3′-untranslated region of the putative transcript, in a portion which was not part of the cDNA sequence previously studied.

Furthermore, additional entries in GenBank assign a function to the gene. They identify it by comparison with a subunit of the bovine vacuolar H-ATPase cDNA (GenBank U10039, ref. 26) which is >85% identical and 90% homologous to the cDNA sequence (the direct comparison is included in the annotation for GenBank entry L44140). The bovine sequence lacks the first (5′) exon reported for the human cDNA, but contains an additional two exons found upstream of that point in the human sequences. Thus, the transcripts from this gene apparently exhibit differential splicing, and it is tentatively assigned the acronym *VATPS1* (vacuolar ATPase subunit 1).

GDI. Starts from CpG island 10. A complete cDNA, containing all 11 exons, has been deposited in GenBank as D45021 by N. Nishimura, K. Sano and H. Nakamura [the gene is also *XAP-4* (X74608) and *1A* (X87194, ref. 10)]. The predicted transcription unit also includes EST56134.

XAP-5. A coding region starting from CpG island 11 contains the published sequence tag for *XAP-5* [X74611, which is also *9F* (X87199, ref. 10)]. The EST coincides with what would be two exons in the 3′-untranslated region of the possible transcription unit, predicting no homology to any known protein or open reading frame. Other putative segments of the transcript, however, find significant homologies with two entries in GenBank.

One entry, an open reading frame from yeast, has homology to the predicted *XAP-5* mRNA. In fact, it strengthens the inference of exonic structure through putative exons 12 and 13. Those two exons, predicted with excellent and good scores by GRAIL, fall in a segment of ~1500 bp at the end of the *XAP-5* region.

A second homology showed essentially complete overlap of 1069 bp with a segment of a reported cDNA sequence, and confirms more of the GRAIL exon predictions.

The sequence submitted earlier as Z46376 was reported to lie upstream of the *HEX2* gene, but for several reasons the connection to *HEX2* is likely to be spurious. First, *HEX2* lies on chromosome 2 rather than X, with the genomic sequence reported here deviating totally from the reported cDNA just after a *Sma*I recognition site at nucleotide 1068. Second, an additional reported *HEX2* sequence agrees with the *HEX2* portion of the first earlier report, but contains no '*XAP-5*-like' stretch. Third, the divergent further sequence in the reported tract, which indeed matches the *HEX2* sequence, runs off the opposite DNA strand (i.e. 3′ to 5′, or head-to-head with the *XAP-5* gene; the orientation is confirmed by the directionality of GT/AG splice junction sequences in the genomic DNA). Thus, the first sequence reported is most likely a chimera, bringing together a segment of the *HEX2* gene from chromosome 2 and a portion of the *XAP-5* region of Xq28. *XAP-5* has no sequence relationship to *HEX2*.

Other significant homologies to the putative XAP-5 protein were observed, but were related primarily to similar patterns of predicted amino acid charge, and have not been further analyzed.

Plexin. The subsequent 15 kb, starting from CpG island 12, contain the human homolog of plexin, a neuronal cell surface molecule earlier reported from *Xenopus laevis.* The inference of the entire putative 'cDNA' for the gene is relatively difficult because the two sequences are identical at only 65% of the residues, and because greater identity is seen at the 3′ ends and no genomic sequence is available for *Xenopus.* Assuming that the two coding sequences are of essentially the same length, and that exon–intron junctions use the most likely signals, Figure 3 presents a probable protein sequence. It includes 33 putative exons, two of which correspond to the EST sequence reported for *XAP-6* [X74609, also referred to as *6.3* (X87197, ref. 10)]. GRAIL programs predicted 30 of the 33 exons, 28 of them identified as 'excellent' (annotated in GenBank entry L44140). The 33 exons add up to ~6.8 kb, which is very similar to the size observed with EST X74609 as a probe in Northern analyses. The two *XAP-6* exons have significant homology to a megakaryocyte-associated tyrosine kinase, and the predicted exon 3 for this gene shows homology to two additional tyrosine kinases, a c-*met* proto-oncogene tyrosine kinase and a c-*sea* tyrosine kinase.

CVG-5. Telomeric to plexin is the large CpG island 13 (already noted in ref. 9), near three putative exons that could be associated in *CVG-5.* All three exons run in the same direction, but two are 5′ (scored as good) and one (scored excellent) is 3′ to the island. This CpG island might thus be one of those within a gene (27).

GdX, P3. Like the putative *CVG-5*, the next two genes are associated with CpG islands (numbers 14 and 15), and are small, each containing only a few exons. Both are known in cDNA form [*GdX* as J03589 (28); *P3* as X12458 (29)].

2–19. The *2–19* gene, starting from CpG island 16, was predicted earlier on the basis of sequence (13), and was then confirmed to exist both by PCR tests on cDNA libraries and by the recovery of a corresponding cDNA by direct selection methods [X87193 (9,13); equivalent to *XAP-7* (X74610)].

CVG-6. CVG-6 was also predicted in the earlier study (13).

G6PD. G6PD has been sequenced from several types of clone, including cosmids and lambdas (ref. 30; X55448).

Figure 2. Features of the 219.4 kb sequence. The top rows show the percentage GC and CpG dinucleotide in a moving window of 100 bp across the whole sequence. CpG islands are numbered 1–17 in the CpG dinucleotide panel. The next row represents the Alu sequences (blue arrowheads) and L1 sequences (orange arrowheads) with their direction indicated as forward on the top strand and reverse on the bottom strand; MERs are indicated by green arrowheads. Two copies of an 11.3 kb repeat are shown as magenta arrowheads across their extent. The row indicating genes shows predicted exons (yellow rectangles); GRAIL2-predicted and already demonstrated exons (blue rectangles); and demonstrated exons that were not predicted by GRAIL2 as red rectangles. Each transcribed gene region is bracketed and named, including six putative genes *CVG-1–CVG-6.* The bottom line shows the expected restriction sites for any of a group of five enzymes with at least one CpG dinucleotide in its recognition site (see Materials and Methods).

Figure 3. Plexin, putatively expressed from CpG island 12. The human sequence is shown. Comparison with the encoded protein with the*Xenopus* sequence deduced from its reported cDNA sequence (see text) indicates identical positions with solid black lines; dots indicate positions at which the human sequence must be gapped to align it with the frog sequence.

Performance of gene prediction methods

The genes already known in whole or in part have provided an experimental test of the predictive power of current versions of GRAIL (31) and other programs (see Materials and Methods). The exon predictions by GRAIL were encouragingly accurate. Among 19 predicted gene candidates, 13 were coincident with the positions of reported cDNAs and, wherever it was available, the predictions agreed with reported cDNA sequences. One of the 13, the H-ATPase, had been localized roughly to Xq28, but is here assigned both a precise location and a functional identity.

XGRAIL1.2 found 148 of a possible 162 known exons, or a 91% success rate for the 13 known genes. This predictive accuracy is comparable with that determined for a different group of high GC genes (ref. 32). Since GRAIL evaluates genomic DNA for protein coding potential, and since 5′ and 3′ exons frequently contain regions that are not translated, one would expect more errors in those exons. Indeed, if they are excluded from the analysis, XGRAIL would have a 95% success rate. Furthermore, of all the correctly predicted exons, 83.5% (101) were scored as ‘excellent’, 14.0% (17) as ‘good’ and 2.5% (3) as ‘marginal’. The 11 failures (false positives) included seven ‘good’ and four ‘marginal’ cases.

Assuming that genes do not overlap or interrupt one another, the false positive rate of XGRAIL could be ~6.9%, based on 11 putative exons falling in intronic regions of the known genes. None of these are candidates scored as excellent, but it is still possible that they represent alternatively spliced forms which might be observed in cDNA isoforms.

As for the boundaries of exons, the 5′ and 3′ boundaries of the 13 known genes are not included in the group, since those are not ‘exon junctions’. Of the 190 known junctions centromeric to the *2–19* gene, 145 (76%) were accurately predicted. The other 45 junctions differed significantly from the XGRAIL prediction.

We have separated out the performance for the most telomeric known genes, *2–19* and *G6PD*, since the predictions of exonic borders were distinctly worse in those two instances. Only 2/34 junctions (6%) were accurately predicted.

Thus, the recognition of exon borders is still rather variable from gene to gene, and reached an overall value of only ~75% for the known genes; but the overall success rate of GRAIL in identifying possible exons has encouraged us to use the predictions of excellent, good and marginal exons as one of the starting points for further analyses.

DISCUSSION

The results strongly support the primacy of regional GC content (‘isochores’; ref. 1) as an organizing principle for much of the substructure of chromosomes. The 57% GC content of this region places it in the 3–4% of the human genome with levels >50%. Furthermore, the level of GC is relatively uniform (Fig. 2): the dispersion around the average exceeds 4% only in the delimited segments in CpG islands, where the GC content can exceed 90% . Sequence analysis confirms in detail the expectation from the

work of Pilia *et al*. (8) and direct mapping studies (10–12) that this region is very high in GC content and, as predicted for such regions earlier (33), is also very rich in gene content.

GC-rich, gene-rich regions of the genome are thus likely to be high priority substrates for sequence analysis. The difficulties imposed by compression zones of high GC and poly(dA)·poly(dT) tails of Alu elements are superable; and current computer programs are already adequate to predict genes and other notable structural features with reasonable accuracy. The inferences can then serve as a guide for the evaluation and recovery of candidate genes. These are of special interest, since two of the genes in this region, those encoding emerin and G6PD, are already associated with diseases, and a number of other disease genes also map to distal Xq28 (9).

Selective content of repetitive sequences, and resolution of Alu/L1 paradox

The content of repetitive sequences is also probably idiosyncratic for high GC regions. The repertoire of moderately repetitive sequences includes four tracts of the $(GT)_n$ dinucleotide, roughly in accord with overall estimates for its frequency of occurrence in the genome (34). The representation of other repetitive complex sequence elements is, however, very skewed. Earlier studies have shown that several viruses [Rous sarcoma (35) and hepatitis B (36)] tend to integrate in high GC cellular DNA. Among the most highly repetitive sequences, Alus and half-Alus predominate in this region [0.85/kb, or ~3-fold the concentration estimated for the entire genome (16)]. In contrast, L1 elements, which occupy a comparable fraction of genomic sequence (17), were detected only at low levels.

These results suggest a resolution of a puzzling discordance between different assays for the regional genomic representation of Alu and L1 elements. Cytogenetic *in situ* analyses with Alu and L1 probes had shown a distinctive pattern of relative distribution in certain regions in metaphase spreads (37,38,59). However, the pattern was not obviously correlated with standard G and R bands; and direct assays of YACs across 50 Mb of Xq24-qter showed a broad range of ratios of Alu/L1 in every cytogenetic band compartment (39,60). In fact, the extreme bias toward selective Alu hybridization obtained in *in situ* experiments correlates well with the locations of very high GC DNA in the subbands that have been called 'T' bands (5). The very same cytogenetic regions include the region of Xq28 analyzed here (12). Thus, high Alu content may be characteristic of very high GC isochores. The hybridization of Alu probes would then be most marked in T bands (1).

Consistent with the relatively high content of Alu compared with L1 sequences in high GC DNA, one Alu per 1.2 kb and one L1 per 13 kb was also observed in the other longest reported sequence of high GC DNA, 106 kb of 52.7% GC from 19q13.3 (41) with one Alu per 1.4 kb and one L1 per 30 kb. However, trends are difficult to discern in lower GC DNA. Fifty six kilobases of lower GC (48%) DNA at the *HPRT* locus (42) contained one Alu per 1.2 kb and one L1 per 11 kb, much like the values reported here in 57.2% GC DNA; whereas 130 kb of still lower GC (45%) in Xq28 (43) contained one Alu per 7 kb and one L1 per 4 kb. Finally, 685 kb of the lowest GC DNA (42.5%) reported, at the T cell receptor locus (44), again contained one Alu per 6 kb but only one L1 per 24 kb. Thus, it may be premature to make strong inferences about the relative levels of repetitive elements as a function of GC content.

Very little variation is seen in the 11.3 kb segment that occurs in inverted orientation at two sites, separated by 38 kb. The two repeats units are identical over >99.2% of their sequence. This near-identity could result if the duplication had simply occurred recently in human evolution; but it is notable that the lack of divergence extends to the repeated very long stretches of $(GT)_{51}$.

Gene content and orientation in the region

The persuasiveness of sequencing as an approach to gene finding is influenced by its efficiency compared with other methods. For this region, extensive work by a number of centers and laboratories with methods like the hybridization of genomic DNA clones to cDNA collections, have recovered some or all of 13 genes. All of those were also identified (and one predicted before it had been found independently; ref. 13) by computer-assisted gene-finding methods. In addition, up to six more genes are suggested here based on the deductions from sequence. In other words, up to 30% of the gene content in a much-studied region has still been more accessible based on sequencing. In addition, exon–intron borders have been determined for a number of genes, and their nature has been clarified.

The 13 known genes and five of the six newly predicted ones are associated with CpG islands, which have thus far been telltales for genes in every case examined. One of them (island 8) is 'double-headed', marking the apparent initiation site of transcription for two genes on opposite strands. This is a rare but already observed phenomenon (27); a similar case has been reported nearby in Xq28 for the adrenoleukodystrophy gene (45) and a neighboring gene (46).

Because the prediction of exons from known genes was highly accurate, an estimate of the extent of transcription based on the overall predictions may be reliable. The region would be 80–90% transcribed, and can be subdivided into three parts with respect to the orientation of transcription of groups of genes. In the most centromeric portion, the directions of transcription (Fig. 2) tend to alternate. *CVG-1* and *CVG-2* run Cen–Tel; *FLN* and *CVG-3*, Tel–Cen; *EMD*, *CVG-4* and *QM*, Cen–Tel; and *DNL1L*, Tel–Cen. A second portion then contains a set of five genes (*XAP-2*, *H-ATPase*, *GDI*, *XAP-5* and *XAP-6*) which are all transcribed in the same sense, Cen–Tel. The third segment includes five genes (*CVG-5*, *GdX*, *P3*, *CVG-6*, *2–19*, *G6PD*) all transcribed Tel–Cen. Thus, the region includes two blocks of genes transcribed in the same direction, as noted earlier by Bione *et al.* (9); but there is a more centromeric region where the direction of transcription seems to vary much more. Whether this reflects features of the higher order organization of GC-rich chromatin remains to be seen.

With one gene per 11.6 kb, this portion of the human genome approaches the gene density observed in the nematode. In that organism, extensive sequencing has found a density of one gene per 5 kb in the central regions of chromosomes, which are relatively enriched for genes, and an estimated overall gene density of ~1 per 8–10 kb when the rest of the chromosomes are included (47).

If the entire human genome had a gene content comparable with this region, it would contain 260 000 genes. That number is about four times the current estimates (3,48). Therefore, the region sequenced contains ~4-fold more genes than the average for the genome.

The gene content in this region can be compared with other extensive sequence tracts discussed above, including distal Xq28, 19q13.3 and the *HPRT, IDS* and T cell receptor loci. Again consistent with the dependence on GC level, regions of 57, 52.7, 48, 45 and 42% GC have one gene every 11, 17, 57, 43 and 170 kb, respectively.

The estimate of gene number includes the possible gene in the 11.3 kb repeated sequence. This would then be an instance of effective 'dosage compensation' for an X-linked gene compared with autosomes. The presence of a single active X chromosome in somatic cells would be compensated by duplication of the gene on the X. This would be somewhat analogous to the case of the gene in intron 22 of the Factor VIII gene, which occurs in a number of copies in the region (49).

Although five of the six predicted genes are associated with CpG islands, most of the putative exons remain hypothetical at present. Thus far, primer pairs have been developed for predicted exons for three genes in the region. Two of them (*CVG-6*, the only gene whose existence is not also supported by an associated CpG island, and *2–19*) were found to amplify PCR products from pools of cDNAs (14,50) and in RT-PCR experiments with cellular RNA preparations (work in progress). The *2–19* transcript has been independently cloned as a cDNA (14). In ongoing work, the *CVG-1* gene has been similarly verified by PCR-based tests; and similar approaches should permit the full elucidation of transcription units for all of the genes, including *XAP-6, XAP-5* and *XAP-2*. It will be particularly interesting to see if primer pairs for the putative exons in the 11 kb repeats detect corresponding sequences expressed in mRNA.

MATERIALS AND METHODS

Sequencing strategy

Nine cosmids (18B4, 14F5, 14C6, 6B2, 13B7, TV1, 1345, 780-13 and 780-124, see Fig. 1) were obtained from four collections [Bione *et al*. (9) for the first five; ref. 51 for TV1; ref. 52 for 1345; and cosmids 780-13 and 780-124, subclones of a YAC made by Dr A. Gnirke]. The cosmids were sequenced by random shotgun methods (13,44,53) supplemented by PCR-based DNA fragment amplification and primer-walking strategies (except for cosmid 780-13, which was completed by sequencing 13 *Eco*RI subclones). Cosmids 1345, 780-13 and 780-124 exhibited deletions that will be discussed in detail elsewhere (Zollo *et al*., in preparation). A tenth cosmid, 58E6 (from the chromosome-specific library prepared at the Lawrence Livermore Laboratory), was used to obtain a 5.5 kb fragment that bridges between cosmids 13B7 and TV1. For the region sequenced earlier in TV1 (13), this study includes only further analysis of summarized previous data.

Sequencing methods

Cosmid DNA was sonicated and the ends repaired with T4 DNA polymerase. A pair of adaptor sequences, ATCTCGAGCTCTAGAG and pCTCTAGAGCTCGA, were added to the sonicated fragments, and 1–2 kb fragments were recovered from the mixture after agarose gel electrophoresis and subcloned into the M13mp9 vector. The vector was prepared by digestion with *Bam*HI followed by single base (dGTP) addition catalyzed by the Klenow DNA polymerase fragment. This procedure produced a three base (ATC) overhang at the 5′ end of the DNA fragments to be cloned and a complementary three base (GAT) protrusion at the 5′ end of the vector DNA. Because the ends of the DNA fragments and the vector DNA are compatible, but the fragment and vector DNAs are not self-compatible, this strategy avoids the self-ligation and double or multiple insert ligations that can occur in standard blunt-end ligation methods.

DNAs cloned into M13 were prepared for sequencing by a streamlined manual method (54) and sequenced using dye-primers and cycle sequencing chemistry with *Taq* polymerase on an Applied Biosystems Catalyst 800 work station and 373A automated sequencers (14). Sequence tracts across most of the cosmid were assembled using Applied Biosystem's FACTURE™ and INHERIT™ program, with 650–850 M13 subclones processed per cosmid (13) (with the exception of cosmid 13B7, for which a specific assembly problem led to the sequencing of >1000 clones; see Results). With average useable sequence tracts of 450 bp, and 90% of the samples included in the finished sequence, coverage was between 6- and 8-fold. Most (99%) of the sequence was determined on both strands but, at some locations, no opposite strand clone had been recovered. In those cases, consensus sequence was derived by sequencing templates in the same orientation with two different sequencing chemistries (dye-primer and dye-terminator; ref. 14).

Gap closure

In areas that were problematic or were covered only once, the assembly program was used to infer clones likely to extend across a gap. The inserts of those M13 clones were amplified by PCR and sequence was then derived with dye-primers from the opposite end of the clone insert (14). A second approach to gap closure required the synthesis of additional primers to extend sequences ('walking') from the edges of contigs using dye-terminator reactions. Finally, the remaining gaps seemed to represent DNA that was not recoverable in the M13 clones. They were closed using PCR primers to amplify the intervening material, which was then sequenced by dye-terminator reactions (14).

Problem areas

Whenever necessary, results from regions with high GC contents (e.g. >90% in the vicinity of base #107 290) or homopolymer tracts [e.g. 20–30 bases of poly(A) or poly(T) in a row] were confirmed by additional experiments with different sequencing chemistries (dye-primer and dye-terminator) and different enzymes (*Taq* polymerase and sequenase; ref. 14). For particular regions containing stretches of poly(dA), additional readings beyond the homopolymer run were obtained using dye-terminator reactions with a mixture of primers that contained T16 followed by one C, A or G (53).

To sequence a sizeable gap between 13B7 and TV1, and to resolve an area where assembly showed an apparently branched sequence (see Results) at the junction of 6B2/13B7, PCR products of 5.5 and 1.4 kb, respectively, were amplified with flanking primer pairs, and the amplification products were used as a template source for sequencing. Each was subcloned into pUC vectors and sequenced after subcloning into M13 (the 5.5 kb fragment) or using primer sequences inferred from a succession of tandem sequencing steps (the 1.4 kb fragment).

Computer-assisted sequence analysis

Programs based on the GCG package (55) and simple derivatives were used to infer the content of repetitive sequences. Using a sliding window of 100 bp, the overall concentration of GC, and the specific level of the dinucleotide CpG were determined (Fig. 2). Runs of more than four consecutive di, tri, tetra and pentanucleotide repeats and rare-cutter restriction sites containing CpG dinucleotides in their recognition sites (Fig. 2) were also assessed.

More complex repetitive elements, including Alu, L1 and moderately repetitive (MER) sequences were identified using the CENSOR program and a database of repetitive sequences [(18,56) see Fig. 2].

To look for genes, repetitive sequences were masked. The unique portions, indexed to their positions in the total sequence, were then checked for any clusters of a selected group of promoter elements (13) and analyzed by a group of programs including FASTA (57), GRAIL1 and GRAIL1.2 (31) and BLAST (58). The sequence and annotation with repetitive sequences and genes, including lists of known and predicted exons and the coordinates of repetitive sequences and subfamilies of Alu and L1 sequences, are summarized here with graphic representations in Figure 2, and are detailed in GenBank (accession no. L44140).

ACKNOWLEDGEMENTS

We thank Primo Baybayan, Rose Bello and Alessandro Arcucci for their technical assistance in obtaining and editing data, and Tim Burcham for improving the INHERIT program during the project. Supported by grants from the NIH (HG00201), the CNR Progetto Finalizzato Ingegneria Genetica and Telethon Italy (Number 417).

REFERENCES

1. Bernardi, G. (1993) The human genome organization and its evolutionary history: a review. *Gene*, **135**, 57–66.
2. Craig, J.M. and Bickmore, W.A. (1994) The distribution of CpG islands in mammalian chromosomes. *Nature Genet.*, **7**, 376–382.
3. Antequera, F. and Bird, A.A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
4. Saccone, S., DeSario, A., Wiegant, J.R., Raap, A., Della Valle, G. and Bernardi, G. (1993) Correlation between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA*, **90**, 11929–11933.
5. Dutrillaux, B. (1973) Nouveau systeme de marquage chromosomique: les bands T. *Chromosoma*, **41**, 395–402.
6. Antonarakis, S. (1994) Genome linkage scanning: systematic or intelligent? *Nature Genet.*, **8**, 211–212.
7. Gardiner, K, Aissani, B. and Bernardi, G. (1990) A compositional map of human chromosome 21. *EMBO J.*, **9**, 1853–1858.
8. Pilia, G., Little, R.D., Aissani, B., Bernardi, G. and Schlessinger, D. (1993) Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics*, **17**, 456–562.
9. Mandel, J.L., Monaco, A.P., Nelson, D., Schlessinger, D. and Willard, H. (1992) Genome analysis and the human X chromosome. *Science*, **258**, 103–109.
10. Bione, S., Tamanini, F., Maestrini, E., Tribioli, C., Poustka, A., Torri, G., Rivella, S. and Toniolo, D. (1993) Transcriptional organization of a 450-kb region of the human X chromosome in Xq28. *Proc. Natl Acad. Sci. USA*, **90**, 10977–10981.
11. Sedlacek, Z., Korn, B., Konecki, D.S., Siebenhaar, R., Coy, J.F., Kioschis, P. and Poustka, A. (1993) Construction of a transcription map of a 300 kb region around the human G6PD locus by direct cDNA selection. *Hum. Mol. Genet.*, **2**, 1865–1869.
12. Palmieri, G., Romano, G., Ciccodicola, A., Casamassimi, A., Campanile, C., Espostio, T., Cappa, V., Lania, A., Johnson, S., Reinbold, R., Poustka, A., Schlessinger, D. and D'Urso, M. (1994) YAC contig organization and CpG island analysis in Xq28. *Genomics*, **24**, 149–158.
13. Zollo, M., Mazzarella, R., Bione, S., Toniolo, D., Schlessinger, D., D'Urso, M. and Chen, E.Y. (1995) Sequence and gene content in 52 kb including and centromeric to the G6PD gene in Xq28. *DNA Sequence*, **6**, 1–11.
14. Chen, E.Y. (1994) Automated DNA sequencing and analysis. In Adams, M.D., Fields, C. and Venter, J.C. (eds), *Automated DNA Sequencing and Analysis Techniques.* Academic Press, Ltd, pp. 3–10.
15. Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44**, 388–396.
16. Shen, M.R., Batzer, M.A. and Deininger, P.L. (1991) Evolution of the master Alu genes. *Mol. Evol.*, **33**, 311–320.
17. Smit, A.F.A., Toth, G., Riggs, A.D. and Jurka, J. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.*, in press.
18. Jurka, J., Kaplan, D.J., Duncan, C.H., Walichiewica, J., Milosavljevic, A., Murali, G. and Solus, J.F. (1993) Identification and characterization of new human medium reiteration frequency repeats. *Nucleic Acids Res.*, **21**, 1273–1279.
19. Bickmore, W.A. and Bird, A.P. (1992) Use of restriction endonucleases to detect and isolate genes from mammalian cells. *Methods Enzymol,*. **216**, 224–244.
20. Maestrini, E., Tamanini, F., Kioschis, P., Gimbo, E., Marinelli, P., Tribioli, C., D'Urso, M., Palmieri, G., Poustka, A. and Toniolo, D. (1992) An archipelago of CpG islands in Xq28: identification and fine mapping of 20 new CpG islands of the human X chromosome. *Hum. Mol. Genet.*, **1**, 275–280.
21. Gorlin, J.B., Yamin, R., Egan, S., Stewart, M., Stossel, T.P., Kwiatkowski, D.J. and Hartwig, J.H. (1990) Human endothelial actin-binding protein (ABP-280, nonmuscle filamin): a molecular leaf spring. *J. Cell Biol.*, **111**, 1089–1105.
22. Patrosso, M.C., Repetto, M., Villa, A., Milanesi, L., Frattini, A., Faranda, S. Mancini, M., Maestrini, E., Toniolo, D. and Vezzoni, P. (1994) The exon–intron organization of the human X-linked gene (*FLN1*) encoding actin-binding protein 280. *Genomics*, **21**, 71–76.
23. Bione, S., Maestrini, E., Rivella, S., Mancini, M., Regis, S., Romeo, G. and Toniolo, D. (1994) Identification of a novel X-linked gene responsible for Emery–Dreifuss muscular dystrophy. *Nature Genet.*, **8**, 323–326.
24. Parrish, J.E., Ciccodicola, A., Wehnert, M., Cox, G.F., Chen, E. and Nelson, D.L. (1995) A muscle-specific DNase I-like gene in human Xq28. *Hum. Mol. Genet.*, **4**, 1557–1564.
25. Yokoi, H., Hadano, S., Kogi, M., Kang, X., Wakasa, K. and Ikeda, J.-E. (1994) Isolation of expressed sequences encoded by the human Xq terminal portion using microclone probes generated by laser microdissection. *Genomics*, **20**, 404–411.
26. Supek, F., Supekova, L., Mandiyan, S., Nelson, H., Pan, Y.E. and Nelson, N. (1994) A novel accessory subunit for vacuolar H^+ ATPase from chromaffin granules. *J. Biol. Chem.*, **269**, 24102–24106.
27. Lar, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
28. Toniolo, D., Persico, M.G. and Alcalay, M. (1988) A 'housekeeping' gene on the X chromosome encodes a protein similar to ubiquitin. *Proc. Natl Acad. Sci. USA*, **85**, 851–855.
29. Alcalay, M. and Toniolo, D. (1988) CpG islands of the X chromosome are gene associated. *Nucleic Acids Res.*, **16**, 9527–9543.
30. Chen, E.Y., Cheng, A., Lee, A., Kuang, W.-J., Hillier, L., Green, P., Schlessinger, D., Ciccodicola, A. and D'Urso, M. (1991) Sequence of human glucose 6-phosphate dehydrogenase cloned in plasmids and a yeast artificial chromosome (YAC). *Genomics*, **10**, 792–800.
31. Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.
32. Lopez, R.,Larsen, F. and Prydz, H. (1994) Evaluation of the exon predictions of the GRAIL software. *Genomics*, **24**,133–136.
33. Aissani, B. and Bernardi, G. (1991) CpG islands, genes and isochores in the genome of vertebrates. *Gene*, **106**, 185–195.
34. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **79**, 6465–6469.
35. Rynditch, A., Kadi, F., Geryk, J., Zoubak, S., Svoboda, J. and Bernardi, G. (1991) The isopycnic, compartmentalized integration of Rous sarcoma virus sequences. *Gene*, **106**, 165–172.

36. Zerial, M., Salina, J., Filipski, J. and Bernardi, G. (1986) Genomic localization of hepatitis B virus in a human hepatoma cell line. *Nucleic Acids Res.*, **14**, 8373–8386.
37. Korenberg, J.R. and Rykowski, M.C. (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell*, **53**, 391–400.
38. Filatov, L., Mamaeva, S. and Tomilin, N. (1987) Non-random distribution of Alu family DNA repeats in human chromosomes. *Mol. Biol. Rep.*, **12**, 117–121.
39. Porta, G., Zucchi, I., Hillier, L., Green, P., Nowotny, V., D'Urso, M. and Schlessinger, D. (1993). Alu and L1 sequence distribution in Xq24-q28, and their comparative utility in YAC contig assembly and verification. *Genomics*, **16**, 417–425.
40. Braaten, D.C., Thomas, J.R., Little, R.D., Dickson, K.R., Goldberg, I., Schlessinger, D., Ciccodicola, A. and D'Urso, M. (1988) Locations and contexts of sequences that hybridize to poly(dG-dT)·(dC-dA) in mammalian ribosomal DNAs and two X-linked genes. *Nucleic Acids Res.*, **16**, 865–881.
41. Martin-Gallardo, A., McCombie, W.R., Gocayne, J.D., FitzGerald, M.G., Wallace, S., Lee, B.M.B., Lamerdin, J., Trapp, S., Kelley, J.M., Liu, L-I., Dubnick, M., Johnston-Dow, L.A., Kerlavage, A.R., de Jong, P., Carrano, A., Fields, C. and Venter, J.C. (1992) Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genet.*, **1**, 34–39.
42. Ansorge, W., Caskey, C.T., Erfle, H., Zimmermann, J., Schwager, C., Stegemann, J., Civitello, A., Rice, P., Voss, H. and Edwards, A. (1990) Automated DNA sequencing of the human HPRT locus. *Genomics*, **6**, 593–608.
43. Timms, K.M., Lu, F., Shen, Y., Pierson, C.A., Muzny, D.M., Gu, Y., Nelson, D.L. and Gibbs, R.A. (1995) 130 kilobases of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2-sulfate sulfatase locus. *Genome Res.*, **5**, 71–78.
44. Hood, L., Rowen, L. and Koop, B.F. (1995) Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Ann. N.Y. Acad. Sci.*, **758**, 390–412.
45. Contreras, M., Mosser, J., Mandel, J.L., Aubourg, P. and Singh, I. (1994) The protein coded by the X-adrenoleukodystrophy gene is a peroxisomal integral membrane protein. *FEBS Lett.*, **344**, 211–215.
46. Mosser, J., Sarde, C.O., Vicaire, S., Yates, J.R. and Mandel, J.L. (1994) A new human gene (DXS1357E) with ubiquitous expression, located in Xq28 adjacent to the adrenoleukodystrophy gene. *Genomics*, **22**, 469–471.
47. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connel, B.M., Copsey, T., Cooper, J., Coulson, A., Craxton, M., Dear, S., Du, Z., Durbin, R., Favello, A., Fraser, A., Fulton, L., Gardner, A., Green, P., Hawkins, T., Hillier, L., Jler, M., Johnston, L., Jones, M., Kershaw, J., Kirsten, J., Laisster, N., Latrelle, P., Lightning, J., Lloyd, C., Mortimore, B., O'Callaghan, M., Parsons, J., Percy, C., Rifken, L., Roopra, A., Saunders, D., Shownkeen, R., Sims, M., Smaldon, N., Smith, A., Smith, M., Sonnhammer, E., Staden, R., Sulston, J., Thierry-Mieg, J., Thomas, K., Vaudin, M., Vaughan, K., Waterston, R., Watson, A., Weinstock, L., Wilkinson-Sproat, J. and Wohldman, P. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C.elegans*. *Nature*, **368**, 32–38.
48. Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994) How many genes in the human genome? *Nature Genet.*, **7**, 345–346.
49. Naylor, A.A., Buck, D., Green, P., Williamson, H., Bentley, D. and Giannelli, F. (1995) Investigation of the factor VIII intron repeated region (int22h) and the associated inversion junctions. *Hum. Mol. Genet.*, **4**, 1217–1224.
50. D'Esposito, M., Mazzarella, R., Pengue, G., Jones, C., D'Urso, M. and Schlessinger, D. (1994) PCR-based immortalization and screening of hierarchical pools of cDNAs. *Nucleic Acids Res.*, **22**, 4806–4809.
51. Martini, G., Toniolo, D., Vulliamy, T., Luzzatto, L., Dono, R., Viglietto, G., Paonessa, G., D'Urso, M. and Persico, M.G. (1986) Structural analysis of the X-linked gene encoding human glucose 6-phosphate dehydrogenase. *EMBO J.*, **5**, 1849–1855.
52. Kaneko, K., Warren, S.T., Miyatake, T. and Tsuji, S. (1993) Isolation of 353 *Not*I linking clones and 62 DNA markers (DXS607–668) at human Xq24-qter. *Cytogenet. Cell Genet.*, **64**, 5–8.
53. Chen, E.Y., Kuang, W.-J. and Lee, A. (1991) Overview of manual and automated DNA sequencing by the dideoxy chain termination method. *Methods*, **3**, 3–19.
54. Zollo, M. and Chen, E.Y. (1994) A manual high-throughput M13 DNA preparation. *BioTechniques*, **16**, 370–372.
55. Genetics Computer Group, Program Manual for the GCG package, Version 7 (1991) 575 Science Drive, Madison, WI, 53711.
56. Jurka, J., Walichiewicz, J. and Milosavljevic, A. (1992) Prototypic sequences for human repetitive. *DNA. J. Mol. Evol.*, **35**, 286–291.
57. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
58. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.S. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
59. Gardiner, K., Aissani, B. and Bernardi, G. (1990) A compositional map of human chromosome 21. *EMBO J.*, **9**, 1853–1858.
60. Saccone, S., DeSario, A., Della Valle, G. and Bernardi, G. (1992). The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. *Proc. Natl Acad. Sci. USA*, **89**, 4913–4917.