*4034–4041* *Nucleic Acids Research, 1996, Vol. 24, No. 20* 

# Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches

**Chun-Nan Chen, Ying Su, Primo Baybayan, Aleli Siruno, Ramaiah Nagaraja[1], Richard Mazzarella[1], David Schlessinger[1,*] and Ellson Chen**

Applied Biosystems Division (ABD), ACGT, Building 200, 850 Lincoln Centre Drive, Foster City, CA 94402, USA and [1]Molecular Microbiology and Center for Genetics in Medicine, Washington University School of Medicine, 660 South Euclid Avenue, Box 8232, St Louis, MO 63110, USA

Received May 31, 1996; Revised and Accepted August 12, 1996

DDBJ/EMBL/GenBank accession no. L78810

## ABSTRACT

**Ordered shotgun sequencing (OSS) has been successfully carried out with an Xq25 YAC substrate. yWXD703 DNA was subcloned into λ phage and sequences of insert ends of the λ subclones were used to generate a map to select a minimum tiling path of clones to be completely sequenced. The sequence of 135 038 nt contains the entire ANT2 cDNA as well as four other candidates suggested by computer-assisted analyses. One of the putative genes is homologous to a gene implicated in Graves' disease and it, ANT2 and two others are confirmed by EST matches. The results suggest that OSS can be applied to YACs in accord with earlier simulations and further indicate that the sequence of the YAC accurately reflects the sequence of uncloned human DNA.**

## INTRODUCTION

Mapping of the human genome has achieved long-range contiguity in cloned DNA primarily by the use of yeast artificial chromosomes (YACs; 1). The X chromosome, for example, has reached >95% coverage in YACs formatted with PCR-based markers (sequence-tagged sites or STSs; 2) and for most regions, YACs provide the only current coverage. If individual YAC clones could provide sequencing substrates, their large size and ready availability would reduce the number of sequencing subprojects to be done, with a corresponding decrease in logistical complexity.

Large insert bacterial clones are attractive alternative substrates for long-range sequencing (see Discussion), but sequencing of YACs or other large insert clones has a number of prerequisites. For example, computer-assisted assembly of sequences of subclones from a very large clone could be hampered by the number of repetitive elements that often occur in clusters and can be highly homologous. Also, in the case of YACs, human DNA represents only ~1% of the DNA content in a yeast cell and consequently sequencing substrates subcloned from YACs must be recovered against the high yeast background.

We sought to address these issues by deploying an 'ordered shotgun sequencing' strategy (OSS; 3) on a typical YAC. OSS attempts to facilitate sequencing a large clone by ordering a set of subclones based on paired end sequences (see below). Simulations (3–5) had suggested that such an approach could have several advantages. Here we report the first use of an OSS approach to obtain the full sequence of a YAC, which contains a 135 038 bp segment of Xq25. The results show that OSS is practicable for large scale sequencing, even by a small working group.

## MATERIALS AND METHODS

The strategy and implementation of OSS and the assembly of the map are outlined in Results, leading to Figures 1 and 2. Here we focus on the preparation of DNA substrates and the methods to analyze the resultant sequence.

### YAC subcloning into a λ-based vector

Agarose plugs were prepared by the method of Gnirke and Huxley (6) containing the yeast strain bearing YAC yWXD703 (DXS2277, from a human–hamster hybrid containing part of the X chromosome; further details for the YAC and STSs are given in the Genome DataBase and at the Washington University Genome Center WEB site, http://genome.wustl.edu/cgm/cgm.html . The plugs were loaded end-to-end across a 1% SeaPlaque agarose gel (FMC Bioproducts) and electrophoresed for 20 h with a 15–30 s switch interval in a DRII-CHEF gel unit (BioRad Laboratories). Five gels were run to obtain 2.5 µg purified YAC DNA.

To minimize the number of chimeric subclones, YAC DNA was partially digested with *Sau*3AI restriction endonuclease and the resulting overhangs were partially filled in with the Klenow fragment of DNA polymerase I and dGTP and dATP to prevent self-ligation of the fragments. The DNA was then fractionated on a 0.7% SeaPlaque agarose gel (FMC Bioproducts) and fractions of 6–9 and 9–12 kb were excised and cloned into λ to yield, respectively, 'series A' and 'series B' libraries. To generate the λ subclones, a preparation of λ BlueSTAR vector was digested with *Xho*I and partially filled in with dCTP and dTTP (Novagen) and each sized fraction of YAC fragments was ligated, packaged into

* To whom correspondence should be addressed

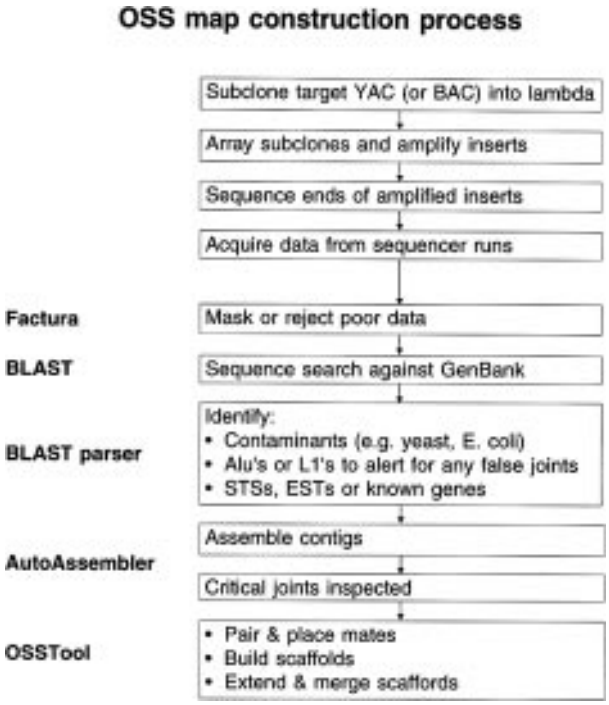


**Figure 1.** OSS map construction. At left are listed in bold type the software programs used in the map construction. Factura (Perkin-Elmer), AutoAssembler (Perkin-Elmer) and OSSTool are Macintosh-based software. OSSTool was developed by Harry Guiremand in the ACGT group at the Applied Biosystems Division of Perkin-Elmer Corporation. The GenBank search by BLAST was done on the National Center for Biotechnology Information (NCBI) network server. BLAST parser is a collection of UNIX shell scripts that organize the BLAST output to a table.

the vector and plated on ER1647 host cells. Individual plaques were picked from plates as agar cores and resuspended in SM buffer (7).

In agreement with the size selection of cloned DNA, gel electrophoresis of λ inserts showed that 285 series A subclones were 6–9 kb and 350 series B were 9–12 kb. Thus the libraries made with the partial fill-in of restriction fragment ends were unlikely to contain chimeric subclones and, in fact, only one possible chimeric B clone (B203, Fig. 2) was detected during subsequent sequencing.

To characterize the subclone library, both positive probing with total human placental DNA and negative probing with total yeast DNA were used. Series A contained 46% yeast inserts, series B 41%. Archived subclones were arrayed on fresh plates and lifts were prepared for hybridization by standard methods (7). The probes were labeled with either a Genius Kit (BMB) or with [$^{32}$P]dATP (Amersham).

## PCR-based preparation of λ insert DNAs

To permit efficient and potentially automatable recovery of end sequences from inserts, we have successfully replaced further phage purification with PCR amplification from the λ subclones. As templates for PCR, supernatants from a 4 h miniculture in 96-well plates were consistently better than phage stored in SM buffer. Two microliters of phage suspension were incubated with reagents from an XL-PCR kit (Perkin-Elmer) in 50 µl reaction mixtures. PCR was performed in a GeneAmp 9600 (Perkin-Elmer) apparatus, using T3 (5′-ATTAACCCTCACTAAAGGGA-3′) and T7 (5′-TAATACGACTCACTATAGGG-3′) primers. Conditions were 94°C for 30 s for the first cycle followed by 94°C for 15 s, 52°C for 20 s, 68°C for 5 min for 15 cycles; 94°C for 15 s, 52°C for 20 s, 68°C for 5 min plus extensions in 15 s increments for another 15 cycles; a final incubation at 72°C for 10 min.

Two to three microliters from each of the total 50 µl PCR reactions were analyzed on a 0.7% agarose gel to size and verify the quality of the amplification products. To prepare the PCR products for sequencing, we tried dilution (8), ammonium acetate precipitation (9), Microcon purification (10), PEG precipitation (7) and combined treatment with exonuclease I and shrimp alkaline phosphatase (ExoI/sAP; 11,12) The last method gave the highest yields and best sequencing results and is well suited to the processing of large numbers of samples. An 8 µl portion of each PCR product was treated with 10 µl containing 4 U each of ExoI and sAP (Amersham) at 37°C for 1 h, then at 72°C for 15 min to inactivate the enzymes. Without further manipulation, samples were then sequenced by dye–primer methods (13) on CATALYST (Perkin-Elmer) robotic workstations.

## OSS map building

An OSSTool program was created. As outlined in Figure 1, contig building was initiated using the Perkin-Elmer AutoAssemble program to find overlaps among end sequences and build a framework physical map by connecting groups of clones using the pairwise relationships of end sequences from individual λ clones. Full sequencing was then initiated for advantageous λ clones, usually those anchored to known parts of the YAC such as vector arms or other mapped markers like STSs or ESTs. Repeats and markers were identified by BLAST searches and noted on the nascent map.

## Complete sequencing of amplified λ inserts

PCR products from selected λ clones were sonicated and the 1.2–1.5 kb fractions purified on agarose gels were further subcloned into an M13 vector for complete random shotgun sequencing. The procedures employed are similar to those described previously (14), including the use of a CATALYST workstation and a 373A or 377 Automated Sequencer (Applied Biosystems), except that blunt-end ligation was used to construct the M13 shotgun library. In general, a set of 96 samples processed from a single microtiter plate was sufficient to assemble a 7–10 kb fragment using FACTURA and INHERIT Autoassembler programs (Applied Biosystems), with an average of 5-fold coverage (see Table 1). Data editing, gap closure and problem solving with dye terminator reactions were all as described (14).

In later experiments, M13 template preparation was replaced by PCR amplification of inserts with reduced primer and dNTP concentrations and no ExoI/sAP treatment, similar to that reported (15).

## Computer analysis of sequence data

The analysis closely followed the procedure described earlier (14). Repetitive elements were identified and then masked and the unique sequence tracts were screened for potential coding regions and other signature sequences, such as TATA boxes and promoter boxes. Exon candidates were sought by the GRAIL1.2

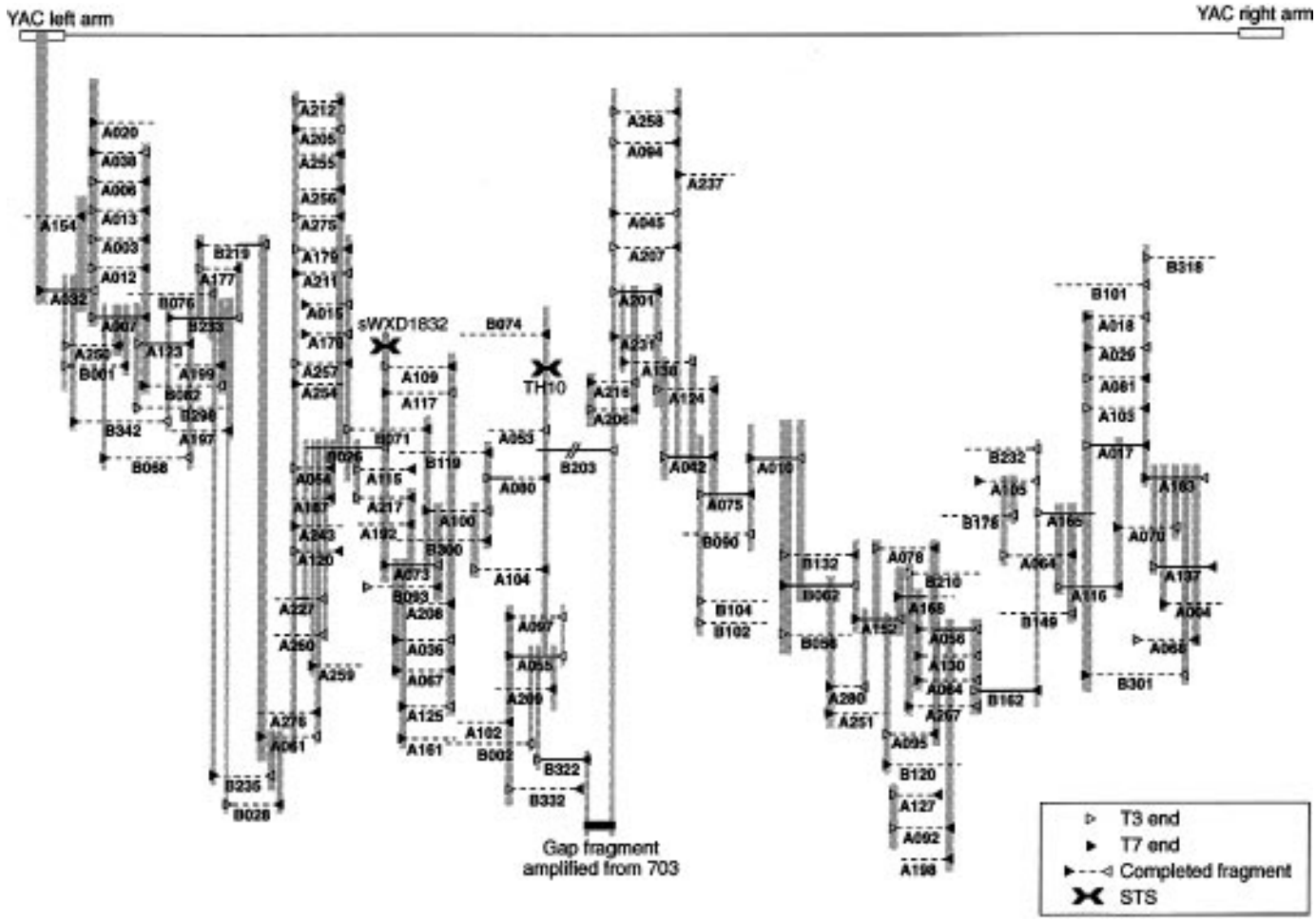


**Figure 2.** OSS map showing coverage of yWXD703 by λ subclones. Subclones are labeled with their names from series A or B (see text). Sequenced ends are indicated by an open (T3 end) or a closed triangle (T7 end). No triangle indicates that there is no sequence information available for that end. Selected subclones with inserts sequenced in their entirety are shown as solid lines; clones with only insert ends sequenced are shown as dashed lines. Shaded vertical bars show segments that overlap between subclones. Also indicated are two additional STSs, sWXD1832 and TH10, that were identified during the end sequencing phase (see Discussion) and the position of a single fragment (middle of the figure) that was amplified directly from YAC DNA to close a gap not covered by the available λ subclones.

(Gene Recognition and Analysis Internet Link) program on the server at Oak Ridge National Laboratory and also by comparison with Genbank entries for ESTs (16) and cDNAs. The exons predicted by GRAIL and the results of comparisons are indicated in Figure 3 (see text).

## RESULTS

YAC yWXD703 has been mapped in a region of Xq25 where STS content is concordant in a number of YACs (17; Nagaraja *et al.*, manuscript in preparation). Furthermore, STSs made from the ends of the YAC map in the contig and the YAC had tested positive for an STS made from the ANT2 gene, for which a cDNA sequence has been published (18). Thus, several preliminary criteria of quality and verifiability were satisfied.

### Implementation of OSS

The success of an OSS approach depends on the ability to obtain sequence information efficiently from both ends of subclones studied. Using optimized conditions for PCR amplification we were able to amplify all of the human DNA inserts from archived λ subclones. In the current work, >90% of the amplified inserts yielded usable end sequence tracts. Two thirds of those are longer than 450 bases and even the much shorter tracts are useful for OSS map construction (see below).

To construct a partial map (Fig. 1), the initial set of insert end sequences are searched: (i) against Genbank as a first screen to identify possible genes which may provide further ordering information; (ii) against consensus Alu and L1 sequence elements. Inferred clone overlaps involving Alu or L1 sequences are flagged to watch out for similar but non-identical repetitive elements, but the scoring of any dubious overlaps becomes definitive during the full sequencing of overlapping λ subclones, since essentially every clone that ends in a repetitive element also includes unique sequences.

The project began with 76 series A subclones of 6–9 kb. In general, the recursive generation of OSS maps revealed problem areas early on, permitting focused attention while the rest of the region was systematically finished. In particular, when it became clear that the A clones were not random enough to cover the entire YAC extent (see Discussion), the end sequences were supplemented

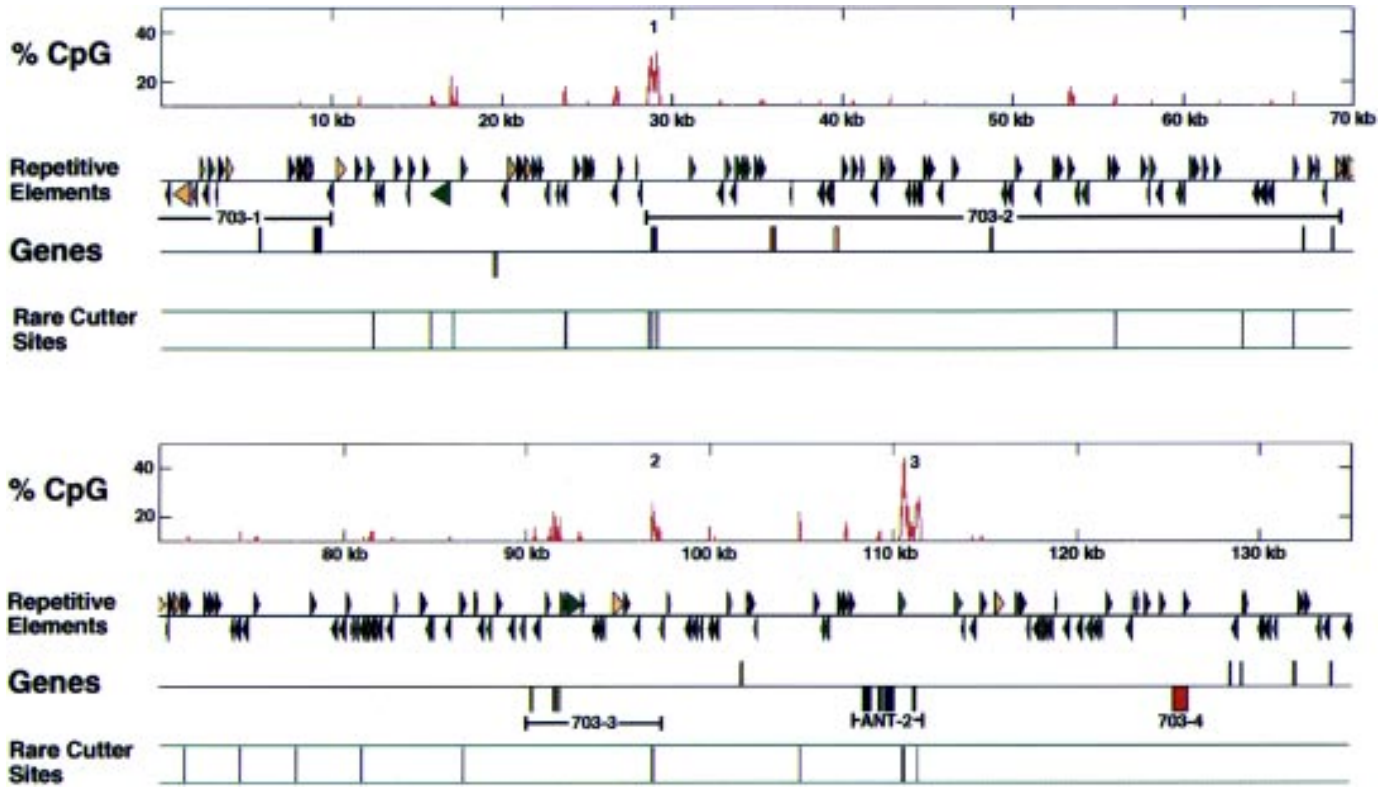


**Figure 3.** Representation of repetitive elements and possible cues for genes in the sequence of yWXD703. The top row represents the content of CpG dinucleotides with a sliding 50 bp window across the region. In addition (bottom row), the predicted location of sites for any of a group of five restriction enzymes (*Bss*HII, *Mlu*I, *Eag*I, *Sac*II and *Not*I) that have CpG sequences in their recognition sites were plotted and are detailed in GenBank entry L78810. These plots permitted the identification of (numbered) putative CpG islands by inspection. Then Alu (SINE), L1 (LINE) and moderately repetitive elements (MERs) were identified by the CENSOR software program developed by J.Jurka. The repetitive elements were further classified into subgroups in the GenBank entry and are plotted according to the strand on which they are encoded, color coded blue (Alu), orange (L1) and green (MER). Putative exon content and genes are shown on the 'GENE' line: exons in yellow are predicted by GRAIL but not confirmed; exons in blue are predicted and confirmed; exons in red are confirmed though not predicted by GRAIL (see text).

from 27 of the larger B subclones. The amplification efficiency for B clone inserts is only slightly lower (>90% in recent experiments) and the distribution of B end sequences tended to complement the A series, connecting many small contigs and thus carrying the OSS map toward closure (Fig. 2). For this project, finishing required the filling of one gap (Fig. 2) by sequencing a 2 kb fragment amplified from the YAC with primers at the edges of the two remaining contigs. In addition, fragment A152 contained a poorly resolved sequence associated with apparent polymerase slippage at a poly(T) or poly(A) tract; it was clarified using a (T)16V primer (as in a comparable case; 14).

Table 1 summarizes features of the shotgun sequencing of 21 λ inserts from 18 A and three B clones. For two reasons, an overall 5-fold redundancy of sequence coverage (96 sequencing reactions; Table 1) was attempted for each λ subclone across the YAC. First, this number of sequences was sufficient to lower the average number of gaps/lambda to 0 or 1, a level that made it straightforward to finish the sequence by primer walking or other methods. Second, ~1 in 3000 bases were idiosyncratically different in individual reads, a variation that we attribute to misincorporations that occurred during PCR amplification of the λ inserts and were preserved in individual M13 subclones. About five readings of the same sequence were sufficient to identify the odd readings and infer a consensus sequence.

The precision (reproducibility) of the sequencing results was earlier assessed at >99.9% for the 219 kb DNA in Xq28 (14). With multiple subclone coverage and higher sample purity, a comparable or greater level was reached here. Consistent with high precision, sequence comparisons of several kilobases of regions of overlap between λ subclones or between λ and PCR-amplified DNA segments have detected no differences. Comparable concordance was also found in comparisons of genomic sequence with encoded cDNA segments.

## Analysis of the sequence of yWXD703

Initial analysis with the CENSOR program was done with default settings, but in general with a threshhold score of 35. This eliminated the identification as 'repetitive elements' of exonic sequences that are paucirepetitive in the genome. With these settings a total of 210 repetitive sequences were detected. They included 129 Alus, nearly all of them full sequences (i.e. 'dimers' rather than 'monomers'), accounting for 24.9% of the total sequence. They are roughly equally distributed throughout the region and are indicated in Figure 3 (middle rows) in blue. The complete tabulation also included nine L1 matches of at least 400 bp and 17 shorter sequences that were judged to be fragments of L1s, comprising 5.6% of the total sequence (orange in Fig. 3). The

**Table 1.** Sequencing statistics for YAC 703

| Clone # | No. sample sequenced | No. samples in assembly | Efficiency (%) | Failure | Sequencing success rate (%) | Average read length | Fold coverage | Size (kb) |
|---|---|---|---|---|---|---|---|---|
| A010 | 72 | 51 | 70.8 | 7 | 90.3 | 531 | 4.3 | 6.3 |
| A056 | 96 | 67 | 69.8 | 16 | 83.3 | 511 | 5.0 | 6.8 |
| A075 | 72 | 46 | 63.9 | 25 | 65.3 | 534 | 3.5 | 7.1 |
| A183 | 94 | 65 | 69.1 | 15 | 84.0 | 523 | 5.2 | 6.5 |
| A061 | 96 | 84 | 87.5 | 8 | 91.7 | 515 | 5.8 | 7.4 |
| A007 | 96 | 85 | 88.5 | 5 | 94.8 | 500 | 5.8 | 7.3 |
| A017 | 81 | 59 | 72.8 | 15 | 81.5 | 528 | 4.0 | 7.7 |
| A032 | 96 | 78 | 81.3 | 4 | 95.8 | 488 | 5.2 | 7.3 |
| A042 | 96 | 81 | 84.4 | 6 | 93.8 | 523 | 6.3 | 6.8 |
| A055 | 96 | 81 | 84.4 | 0 | 100 | 528 | 6.7 | 6.4 |
| A073 | 72 | 66 | 91.7 | 2 | 97.2 | 555 | 6.1 | 6.0 |
| A100 | 96 | 70 | 72.9 | 10 | 89.6 | 515 | 4.8 | 7.5 |
| A105 | 96 | 74 | 77.1 | 15 | 84.4 | 520 | 5.9 | 6.5 |
| A116 | 72 | 58 | 80.6 | 7 | 90.3 | 524 | 4.2 | 7.3 |
| A123 | 96 | 85 | 88.5 | 6 | 93.8 | 428 | 6.3 | 5.8 |
| A152 | 96 | 80 | 83.3 | 4 | 95.8 | 494 | 5.6 | 7.0 |
| A165 | 96 | 81 | 84.4 | 6 | 93.8 | 504 | 5.4 | 7.5 |
| A201 | 96 | 87 | 90.6 | 3 | 96.9 | 531 | 6.6 | 7.0 |
| B026 | 96 | 75 | 78.1 | 11 | 88.5 | 505 | 3.8 | 10.0 |
| B062 | 96 | 73 | 76.0 | 21 | 78.1 | 534 | 3.7 | 10.5 |
| B233 | 96 | 81 | 84.4 | 4 | 95.8 | 512 | 4.9 | 8.5 |
| Average: | 91 | 73 | **80.0** | 9 | **89.7** | 514 | **5.2** | 7.3 |

For each of 21 λ clones the numbers of samples sequenced, the average read length and the resultant redundancy are recorded. The fourth column from the left, Efficiency, gives the percentage of sequence runs that contributed to the contig (column 3) among the total number of samples for a particular λ subclone (column 2)

other 55 repetitive sequences included 21 classes of MERs, totalling 9.3% of the sequence content (green in Fig. 3). Thus, overall the region was composed of 40.9% repetitive sequences.

In the first step to detect candidates for genes, CpG islands (19) were indicated by telltale concentrations of GC and CpG dinucleotides. In Figure 3, three strong CpG islands, coinciding with at least two restriction sites for enzymes with CpG in their recognition sites, are indicated by the numbers 2–4. Another weaker possible island, numbered 1, coincided with only a single rare-cutter restriction site and is not further considered here. CpG island association and other evidence for each of five gene candidates, ANT2 and 703-1 to 703-4 (Fig. 2), is summarized as follows.

*703-1*. Two exons are predicted by GRAIL at nt 5754–5838 and 9072–9326, encoded on the top strand. The latter exon overlaps rat EST L105369, with 80% identity over 255 bp. Like the predicted exons, the EST is transcribed from the top strand, left to right in Figure 3. Thus, the 5′-end of the gene presumably lies more centromeric in Xq25. Neither the EST nor the genomic sequence resemble any known gene/protein, so that the gene remains anonymous.

*703-2*. This gene, initiated at CpG island 2, includes a first exon that is predicted by GRAIL at nt 28 802–29 076 and is also highly homologous to a reported mitochondrial gene (M31659; 20; see also Discussion). Furthermore, although it is not predicted by GRAIL, a second exon in an open reading frame 7000 bp toward the right arm (nt 35 858–36 081) shows a putative correct splice junction that would continue precisely with further sequence homologous to the same mitochondrial gene (Fig. 3). Four additional exons predicted by GRAIL and extending through nt 65 907 may represent additional coding sequence of the 703-2 gene.

*703-3*. Three exons transcribed toward the left arm are predicted by GRAIL and could be associated with CpG island 3. In this case, there is as yet no verification of the exons by detection of cDNA, transcribed RNA or a gene of known function.

*ANT2*. This gene, associated with CpG island 4, contains four exons that are predicted by GRAIL (nt 108 408–108 831, 109 212–100 352, 109 578–110 066 and 111 101–111 281). All are confirmed by comparison with the published cDNA sequence.

We note that the YAC sequence deviates from the published cDNA sequence for ANT2 at several residues, but sequence tracts identical to those in the YAC occur in many ESTs in dbEST that are homologous to the 3′-end of ANT2, including 18 in the Merck-Washington University database.

*703-4*. A sequence not suggested by GRAIL as a putative exon nevertheless matched both ends of a clone arising from the 3′-end of a cDNA. The EST, again from the Washington University/Merck consortium project (clone ID158302) is essentially identical to nt 125 258–125 914 in the yWXD703 sequence. The clone is polyadenylated, so that it most likely represents a cDNA rather than a contaminating genomic DNA or unspliced message sequence, but the extent of the corresponding gene is unknown.

Four exons are predicted by GRAIL as transcribed toward the left arm distal to 703-4, but none of them is homologous to any known expressed sequence or gene. In the absence of any additional evidence at this time, they are not suggested as an additional gene.

The YAC thus contains two regions (Fig. 3), each extending about half of its length, with the first half transcribed toward the right arm and the rest transcribed toward the left arm.

## DISCUSSION

This study was designed: (i) to test the OSS formulation, with some comparison with the current standard of random shotgun sequencing; (ii) to assess the feasibility of the YAC as a sequencing substrate; (iii) to analyze the sequence and begin to assess the fidelity of the YAC compared with uncloned DNA.

### Evaluation of OSS

OSS is designed with a small workgroup in mind, with the goal of sequencing a large target in an easily manageable way. However, implementation of OSS imposes requirements that include the subcloning of a large clone into 7–10 kb fragments and the efficient sequencing of the ends of the cloned fragments. The requirements have been successfully met in this test.

As a method to generate fragments from a large target, partial restriction digestion with *Sau*3A1 has the advantages that it generates cohesive ends which can be ligated more efficiently than blunt ends and the ends can be partially filled in to prevent self-ligation and improve cloning efficiency. The disadvantages are that there is preferential cleavage at some sites and some regions may lack *Sau*3A1 sites.

One way to compensate for local deviations in *Sau*3A1 site distribution uses a mixture of insert sizes. As an example in this study, the B subclone (average ~10 kb) library complemented coverage by the A library (average 7 kb). A strategy using a mixture of clones of more than one size range can help achieve overall coverage and also can counter the problem of non-random distribution of subclones. In fact, simulations indicate that even a small fraction of larger inserts can provide such benefits (4). The fraction here (1:6 or 14%) is in the range recommended from simulations, but because large clones are harder to amplify, the practical case may deviate from models and much more work will be required to optimize the approach.

PCR amplification easily provides increasingly reliable clone inserts of pure human DNA as substrates for the generation of end sequences. As an alternative to clone DNA preparation, amplification of subclone inserts of up to 10 kb is currently robust and straightforward with long-range PCR kits. Essentially all the inserts are single bands on agarose gels and after treatment with ExoI and sAP, to purify the products (see Materials and Methods), direct sequencing of the ends of amplified inserts can be efficiently performed with large numbers of samples.

The OSS approach has another requirement, recovery of sequence from both ends of the majority, if not all, of the subclones. If there are appreciable numbers of subclones for which sequence is available for only one end, the efficiency of OSS map construction declines sharply, since connections to other clones cannot be established. In this project the fraction of clones for which both ends yielded sequence was >65% (which was improved to >80% in later experiments), high enough to sustain map construction. Furthermore, we have routinely recovered sequence from an apparently poor run even when it was as short as 30 bases. Sequences that short can nevertheless help to make a map (3) and are an important aid in mapping gaps and confirming overlaps.

Long read lengths are critical, however, at the level of random shotgun sequencing of the λ intermediate clones and thus determine their optimum size. Current read lengths of the order of 550 bases can easily sustain the analysis of λ inserts of 10–20 kb. Ten kilobases is now the upper limit for high efficiency long-range PCR, but as long-range PCR continues to improve, a smaller number of larger λ substrates should suffice, simplifying the project.

The rate of a project depends on the rate of sequencing of subclones. A single technician can now sequence up to two 10 kb λ clones/week. Each 150 kb YAC or other large insert clone thus requires ~2 person-months of effort to reach the gap closure phase. The rate of an individual project can be increased by analyzing more λ subclones simultaneously, but this increases the chance that a region will unintentionally be sequenced more than once. For example, when several λ subclones from yWXD703 were simultaneously chosen for minimal overlaps and sequenced completely, the group containing λ A055 (Fig. 2) merged with that containing A080, with a substantial overlap in sequenced area. Therefore, to avoid duplicated effort, each project can be assigned to only one person at a time and the rate of very large projects can rather be increased by expanding the number of large insert clones being analyzed.

### Comparison of random and ordered shotgun sequencing

In critical respects ordered and random shotgun sequencing are equivalent. For example, comparable precision requires comparable sequencing redundancy and both methods require finishing efforts to fill gaps. But each approach has advantages.

In current large scale sequencing projects, random shotgun sequencing is much further along the learning curve, so that costs have been progressively reduced by incremental improvements and a stable and increasing level of efficiency has been demonstrated. Also, random shotgun methods employ only a single subcloning step from the large insert clone to sequencing substrates, whereas in the approach to OSS used here, two successive cloning steps are used. The bridge to smaller clones at each step is, however, made more efficient by replacing clone DNA preparation with automatable PCR amplification, and the OSS approach affords several possible compensating advantages.

First, OSS is relatively forgiving of substantial levels of contamination. With BACs or PACs as substrates, contamination levels from bacterial DNA can be as low as <5% or as high as >25% in different preparations (work in progress), but even higher levels require little effort to identify and discard contaminating subclones at the end sequencing stage of OSS. For example, single-pass pulsed-field gel purification had enriched yWXD703 DNA ~50-fold over the starting DNA. With these preparations, there is still close to 50% contamination with yeast, but most of those clones could be detected by probing with yeast DNA. Furthermore, extraneous clones can easily be identified by sequencing one end of each 7–12 kb subclone and finding yeast sequences by comparison with the yeast sequence in GenBank. In the instant case, <10% of the sequencing reactions in the project (180 end sequences out of 2400 total reactions for this project) would be required to discard all vector and yeast subclones without prior screening. In contrast, in random shotgun sequencing, every contaminating clone receives the same attention as each

authentic clone, so that half the reactions would be required to eliminate 50% contamination.

Second, a similar consideration applies to sequencing of the vector component of subclones. BACs or PACs contain 8–16 kb of vector sequence in random subclones, whereas those segments are again largely discarded at an early stage in OSS.

Third, and related to the other potential advantages, OSS offers added flexibility in the choice of subclones or regions for discretionary sequencing. For example, regions of a YAC or other large substrate that overlap a neighboring YAC are again quickly found by end sequencing and need not be completely redone. If one estimates that the two neighboring clones would each overlap a sequencing substrate by the order of 20% of its length, then a corresponding fraction of the total sequencing effort might be saved using OSS rather than a random shotgun approach. One can even imagine approaches in which the early identification and mapping of large L1 elements (or other repetitive sequenes in tandem) could activate the option of less thorough sequencing there than in regions of unique DNA (21). For the human genome project, precise end-to-end sequencing may preclude such selectivity, but sequencing of other organisms might make it more attractive, especially when funding is limited.

As a subsidiary feature, gaps in the overall map are localized at an early stage in such a project, permitting focused attention. Similarly, within each subfragment, the average number of gaps (0.4) is relatively simple to fill and, because repetitive elements are segregated into subclones, they are easier to discriminate during sequence assembly. Perhaps because of the added information provided by the compartmentalization, the overall redundancy of sequencing was ~6-fold (see Table 1), compared with 8- to 10-fold in the sequencing of a comparable amount of DNA by random shotgun methods. This comparison, however, is contingent, since various approaches are becoming increasingly efficient and approaches have not been compared with the same clones.

It is also possible that OSS will have advantages for a small group. In ongoing work a group of three staff with some informatics support has been able to reach a level of ~1 Mb of sequence throughput per year. This level would be enough to manage many large scale sequencing projects and compares favorably with the per capita throughput of large scale random shotgun sequencing by large groups. It remains untested, however, whether random shotgun sequencing on a comparable scale can also be efficiently adapted to a small group.

## YACs as long-range sequencing substrates for OSS

Based on this study and simulations, subcloning of a single YAC can provide material for projects across at least 150 kb and likely for much larger stretches of DNA.

In ongoing work, OSS has also been applied to bacterial clones like BACs or PACs with ease. Bacterial clones show much lower levels of co-cloning and are easier to prepare free of contaminating host DNA than are YACs, though as we comment above, the sequencing required to discard adventitious yeast clones is a very small fraction of total effort and could be reduced further, for example by a second pulsed-field gel electrophoresis (22) or by passage of the YAC through a 'window' strain (23).

As an additional consideration, there are many regions that are only available in YACs. (Even in the model eukaryotic organism *Caenorhabditis elegans*, 10–20% of the genome has thus far been cloned only in YACs; 24.) YACs may thus be the substrate of choice (or necessity) for long-range sequencing of certain portions of genomes. It is therefore encouraging that OSS can already be applied to YACs, with an efficiency that can obviously be improved. If subclones are randomly derived from a starting target clone, sequences from their ends indeed produce a map (Fig. 2), as had been anticipated (3), and sequencing effort would be expected to scale linearly with the size of the clone.

## Analysis of sequence and gene content

*Repetitive sequences.* The suite of programs assembled for the analysis of an earlier region of 220 kb (14) has been useful in this case as well. In the earlier case, in a region of very high GC (57.2%), the repetitive sequences were heavily weighted toward Alu, with only a marginal contribution of L1 and MER. Here again, in a region of moderate GC content (45.5%), Alus dominate. The corresponding percentages of sequence that are Alu, L1 and MERs were 20.4, 1.6 and 3% in the high GC region and 24.9, 5.7 and 9.3% here. Thus, as first indicated in the comparison of the Xq28 segment with other reports of long-range sequencing (see Discussion; 14), the results show no trend in the proportions, for example, of Alu:L1. Rather, they are in agreement with the earlier analysis of Alu- and L1-containing restriction fragments in YACs from Xq24-qter, which indicated that both Alus and L1s were widely distributed with variable densities across nearly all of the 50 Mb region (25).

*Gene number and characteristics.* Gene content, however, continues to follow the trend observed for large scale sequence tracts, that coding capacity is proportional to overall GC content in 'isochores' (26), zones of up to 1 Mb length with characteristic roughly constant GC. Like other regions with 42–45.5% GC (27,28), yWXD703 contains about four genes in 140 kb [including three candidate genes (703-2, 703-3 and ANT2) and two partial genes (703-1 and 703-4)], or about 1/32 kb.

It is of interest that of the five gene candidates in the YAC, four have already been supported by EST content (703-1, 703-2, ANT2 and 703-4) or homology to known genes (ANT2 and 703-2). (Furthermore, the EST homologous to 703-4 derives from fetal kidney and the sequence has also now been detected in fetal kidney RNA in Northern analysis; work in progress.)

Gene 703-2 is especially interesting, since, as Figure 3 shows, it is highly homologous to 'GDC', a carrier protein that has been associated with Graves' disease (29). GDC is a mitochondrial protein and one can wonder whether the X-linked gene detected here is a form that is functionally related. It is unlikely that the gene product of 703-2 is also a constituent mitochondrial protein, since it would then be expressed in every cell, but no corresponding cDNA was present in any of 12 cDNA libraries assayed by PCR-based tests.

## Fidelity of the YAC sequence to uncloned human DNA

The increasing number of ESTs available from systematic cDNA projects and the consituent ANT2 gene provide alternative determinations of sequence and particularly of the gene segments that are usually considered the most important part of genomic sequence. Comparisons permit some inferences about the fidelity of the YAC sequence to Xq25 genomic DNA.

The simplest criterion for fidelity is co-linearity. Genomic DNA should show no deletion of entire exons or within exons of

cDNA. The orders of exons and their sequences are co-linear in yWXD703 and all expressed sequences available thus far, including every EST assigned to the region, a portion of a gene similar to one associated with Grave's disease and all of the ANT2 gene. In addition, the primer sequences for three other STSs developed from genomic DNA [registered in the WU Genome Center database as sWXD1832 (2736L; DXS7340), sWXD774 (TH4) and sWXD770 (TH10)] are all contained in yWXD703 at the proper interprimer distances.

A more stringent requirement for fidelity is based on the detailed comparison of sequence. That YAC sequences can indeed be faithful to genomic DNA has been indicated in another case (9), in which G6PD sequence was derived with 99.99% precision and complete agreement for both a YAC and a cDNA. For segments of yWXD703, disregarding a low level of polymorphic variation in sequence, the maximum agreement cannot exceed the accuracy of the determination of cDNA sequences, or ~96% (EST sequences in GenBank are also less accurate from 3′-ends of cDNAs and toward the end of long sequence tracts). Since the genomic sequences are ~95–97% identical to the independently determined cDNA sequences, the results are consistent with a YAC sequence that is likely to be faithful to uncloned DNA. However, the comparisons are of course still limited. More stringent evidence could be obtained by comparative sequencing of tracts sampled from uncloned DNA by PCR-based methods at statistically chosen intervals or by the comparison of Southern hybridization patterns of genomic and YAC DNA when probed with labeled YAC DNA.

In conclusion, this 135 kb segment of Xq25 has been analyzed in a single YAC clone using OSS. The approach worked relatively smoothly, though there is still room for improvement. The sequence analysis programs also worked to give accurate gene predictions and comparisons to ESTs, indicating that the human DNA in the YAC likely retains the sequence of genomic DNA.

## ACKNOWLEDGEMENTS

We thank Warren Regala, Sandra MacMillan and Patricia Taillon-Miller for technical help with sequencing, YAC preparations and subcloning. Support for this project came from the NIH (GESTEC grant HG00201).

## REFERENCES

1 Burke,D.T., Carle,G.F. and Olson,M.V. (1987) *Science*, **236**, 806–812.
2 Olson,M., Hood,L., Cantor,C. and Botstein,D. (1989) *Science*, **245**, 1434–1435.
3 Chen,E.Y., Schlessinger,D. and Kere,J. (1993) *Genomics*, **17**, 651–656.
4 Roach,J.C., Boysen,C., Wang,K. and Hood,L. (1995) *Genomics*, **26**, 345–353.
5 Singh,G.B. and Krawetz,S.A. (1995) *Genomics*, **25**, 555–558.
6 Gnirke,A. and Huxley,C. (1991) *Somat. Cell Mol. Genet.*, **17**, 573–580.
7 Sambrook,J, Fritsch,E.F. and Maniatis,T. (eds) (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
8 Ruano,G. and Kidd,K.K. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2815–2819.
9 Chen,E.Y., Kuang,W-J. and Lee, A. (1991) *Methods, Companion to Methods Enzymol.*, **3**, 3–19.
10 Mihovilovic,M. and Lee,J.E. (1989) *BioTechniques*, **7**, 14–16.
11 Werle,E., Schneider,C., Renner,M., Volker,M. and Fiehn,W. (1994) *Nucleic Acids Res.*, **22**, 4354–4355.
12 Hanke,M. and Wink,M. (1994) *BioTechniques*, **17**, 858–859.
13 Chen,E.Y. (1994) In Adams,M.D., Fields,C. and Venter,J.C. (eds), *Automated DNA Sequencing and Analysis Techniques.* Academic Press, London, UK, pp. 3–10.
14 Chen,E.Y., Zollo,M., Mazzarella,R., Ciccodicola,A., Chen,C., Zuo,L., Heiner,C., Burough,F., Ripetto,M., Schlessinger,D. and D'Urso,M. (1996) *Hum. Mol. Genet.*, **5**, 659–668.
15 Trower,M.K., Burt,D., Purvis,I.J., Dykes,C.W. and Christodoulou,C. (1995) *Nucleic Acids Res.*, **23**, 2348–2349.
16 Adams,M.D. *et al.* (1991). *Science*, **252**, 1651–1656.
17 Schiebel,K., Mertz,A.,Winkelmann,M., Nagaraja,R. and Rappold,G. (1994) *Genomics*, **24**, 605–606.
18 Battini,R., Ferrari,S., Kaczmarek,L., Calabretta,B., Chen,S.T. and Baserga,R. (1987) *J. Biol. Chem.*, **262**, 4355–4359.
19 Antequera,F. and Bird,A. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 11995–11999.
20 Zarrilli,R., Oates,E.L., McBride,O.W., Lerman,M.I., Chan,J.Y.C., Santisteban,P., Ursini,M.V., Notkins,A.L. and Kohn,L.D. (1989) *Mol. Endocrinol.*, **3**, 1498–1508.
21 Nurminsky,D.I. and Hartl,D.L. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 1694–1698.
22 Vaudin,M., Roopra,A., Hillier,L., Brinkman,R., Sulston,J., Wilson,R.K. and Waterston,R.H. (1995) *Nucleic Acids Res.*, **23**, 670–4.
23 Hamer,L., Johnston,M. and Green,E.D. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 11706–11710.
24 Coulson,A., Kozono,Y., Lutterbach,B., Shownkeen,R., Sulston,J. and Waterston,R. (1991) *Bioessays*, **13**, 413–417.
25 Porta,G., Zucchi,I., Hillier,L., Green,P., Nowotny,V., D'Urso,M. and Schlessinger,D. (1993) *Genomics*, **16**, 417–425
26 Bernardi,G. (1993) *Gene*, **135**, 57–66.
27 Hood,L., Rowen,L. and Koop,B.F. (1995) *Annls NY Acad. Sci.*, **758**, 390–412.
28 Timms,K.M., Lu,F., Shen,Y., Pierson,C.A., Muzny,D.M., Gu,Y., Nelson,D. and Gibbs,R.A. (1995) *Genome Res.*, **5**, 71–78.
29 Fiermonte,G., Runswick,M.J., Walker,J.E. and Palmieri,F. (1992) *DNA Sequence*, **3**, 71–78.