Mapping the human Y chromosome by fingerprinting cosmid clones.

K Taylor, N Hornigold, D Conway, et al.

Genome Res. 1996 6: 235-248
Access the most recent version at doi:10.1101/gr.6.4.235

RESEARCH

Mapping the Human Y Chromosome by Fingerprinting Cosmid Clones

Kay Taylor,[1,2] Nick Hornigold,[2] Darren Conway,[2] Delyth Williams,[2] Zanna Ulinowski,[2] Mahima Agochiya,[2] Paolo Fattorini,[2] Pieter de Jong,[3,4] Peter F.R. Little,[5] and Jonathan Wolfe[2]

[2]The Galton Laboratory, University College London, London NW1 2HE, UK; [3]University of California, Lawrence Livermore National Laboratory (LLNL) Bio-Med L-452, Livermore, California; [4]Imperial College of Science, Technology, and Medicine, London SW7 2AZ, UK

We have used Y-specific cosmid clones in a random fingerprinting approach to build contigs on the human Y chromosome. Clones derived from two libraries have been analyzed. The construction of one library is described here, the second was the Y chromosome-specific library LLOYNC03 "M" (Lawrence Livermore National Laboratory). To date, we have fingerprinted 4430 cosmids: 377 contigs have been constructed containing from 2 to 39 clones. Along with the singletons, we estimate that we have covered 72.5% of the euchomatic portion of the Y chromosome with fingerprinted clones. Sequence tagged sites are being used to anchor cosmids and contigs onto the YAC framework.

As a first step toward the sequencing of whole genomes, several different strategies have been developed to produce an ordered set of overlapping clones or contigs. The underlying principle of many of these strategies is to create a diagnostic pattern or fingerprint for cloned random fragments that may then be compared to one another in a computer data base. Clones with significantly similar fingerprints are deemed to overlap. The first techniques of this kind were described by Coulson et al. (1986, 1988, 1991) and Olson et al. (1986), mapping the genomes of *Caenorhabditis elegans* and *Saccharomyces cerevisiae,* respectively. Similar approaches have been used in the mapping of the genome of *Escherichia coli* (Kohara et al. 1987; Knott et al. 1988,1989), *Drosophila* (Garza et al. 1989; Siden-Kiamos et al. 1990) and various human chromosomes: X (Wada et al. 1990), 16 (Stallings et al. 1990), 19 (Carrano et al. 1989), and 11 (Harrison-Lavoie et al. 1989, Heding et al. 1992). Most mapping projects proceed from yeast artificial chromosome (YAC) maps to cosmid maps. Although YAC clones cover large distances they are susceptible to rearrangements and chimerism. In addition to providing higher resolution, cosmids are easier to work with and are relatively easy to use for sequencing, exon trapping, and other gene isolation methods. We are thus aiming to produce a sequence-ready map of the Y chromosome.

The human Y chromosome is unusual compared to other human chromosomes. It has a relatively low gene density; approximately one-half to two-thirds of the entire length is compromised of two repetitive elements (DYZ1 and DYZ3, Cooke 1976), and the euchromatic portions contain a large number of lower copy number repetitive elements. Because the Y chromosome does not undergo recombination for the majority of its length, mapping by linkage analysis is not possible and physical methods must be used. Advantage has been taken of the haploid nature of the Y chromosome to assign sequences to chromosome intervals using panels of DNA from individuals with deletions (Affara et al. 1986; Vergnaud et al. 1986; Oosthuizen et al. 1990; Ma et al. 1992; O'Reilly et al. 1992). The most comprehensive deletion map was described by Vollrath et al. (1992), who ordered 132 loci including 104 sequence tagged sites (STSs) on a panel of 96 individuals with partial Y chromosomes. By a process of STS content mapping, 207 loci were used to construct a YAC contig map spanning 98% of the euchromatic region of the chromosome (Foote et al. 1992). A second YAC

[1]Corresponding author.
E-MAIL k.taylor@genetics.ucl.ac.uk; FAX 0171-383 2048.
[4]Present address: Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, New York 14263.

map, based on these and additional STSs but using larger YAC clones, has also been constructed (Jones et al. 1994). To make the transition between YACs and cosmids we have employed, first, a random fingerprinting strategy, and then, a more directed approach to generate cosmid contigs. The information supplied by the STS and YAC map allows these to be positioned on the chromosome.

RESULTS

Characterization of the 3E7 Y Chromosome Library

The library consists of 1728 human clones. The two repetitive elements that make up the large heterochromatic region of Yq, DYZ1 and DYZ2, appear to be very under-represented in the library; only 12% of the cosmids hybridize to DYZ1 and <1% to DYZ2, whereas the size of the heterochromatic block is estimated at 20 Mb (Smith et al. 1987; Schmid et al. 1990) and so contributes in the region of 30%–40% of the Y chromosome length. There is no real explanation for the paucity of these two repetitive elements in the library, which has been reported previously (Bishop et al. 1983; Wolfe et al. 1984a). There are two recognition sites for the cloning enzyme *Sau*3AI within the repeat unit so it should be possible for these fragments to be cloned and form part of the library. There are many *Hin*fI sites in the DYZ1 repeat producing fragments with an average size of 50 nucleotides. These small fragments can be seen in the fingerprinting gel in Figure 3 (below). Subtracting the number of clones containing these repeats from the number of clones in the library we estimate that the library represents ~1.75 chromosome equivalents (considering the portion of the chromosome excluding the Yq heterochromatin—estimated to be 30 Mb (Affara et al. 1994)—and assuming a cosmid insert size of 35 kb). The alphoid centromeric repeat DYZ3 (probe cY84) is somewhat over-represented in the cosmids. Given that in 3E7 the block is 470 kb long (Tyler-Smith and Brown 1987), we would expect (470/35)*1.75 = 24 of the clones in the library to be positive. Actually, 51 clones were identified. We initially screened the library with 26 different probes representing both single-copy and repetitive sequences and expressed sequences (results not shown). Twenty-four of these probes identified positives but clones containing the gene *MIC2* (Goodfellow et al. 1987) and the pseudogene *ASSP6* (Daiger et al. 1982) were not found.

Cosmid Fingerprinting

Cosmid clones from the library described above, together with clones from the Lawrence Livermore Library, have been used to build a contig map of the human Y chromosome. The clones were fingerprinted as described in Methods (Figure 1). Most fingerprints consist of from 30 to 40 bands. We reject bands that lie outside the range of the markers and also faint bands that may be caused by partial digestion. All overlaps that have

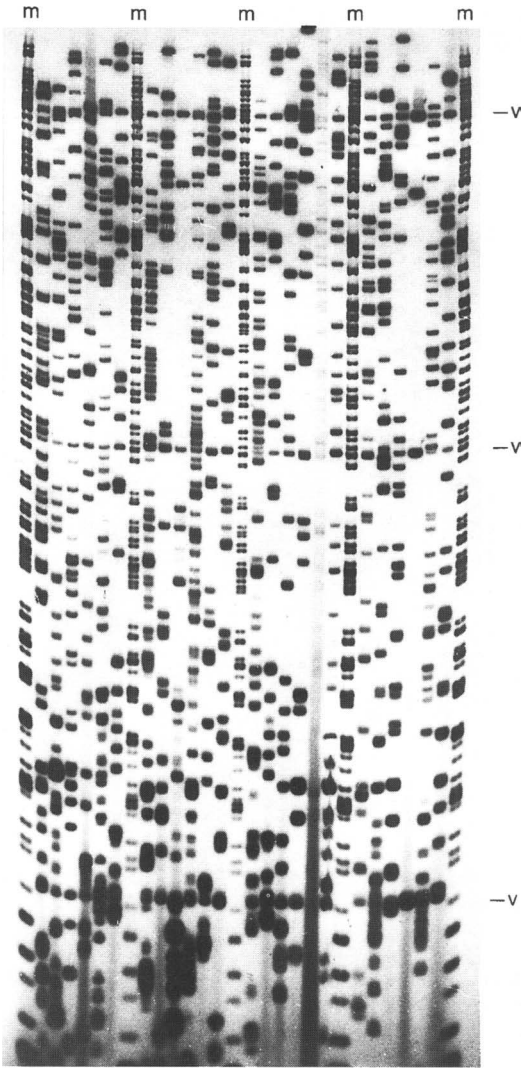

Figure 1 A fingerprinting gel. Twenty-four individual cosmid samples are shown. Lanes labelled m are the λ marker. Bands labeled v are derived from the vector.

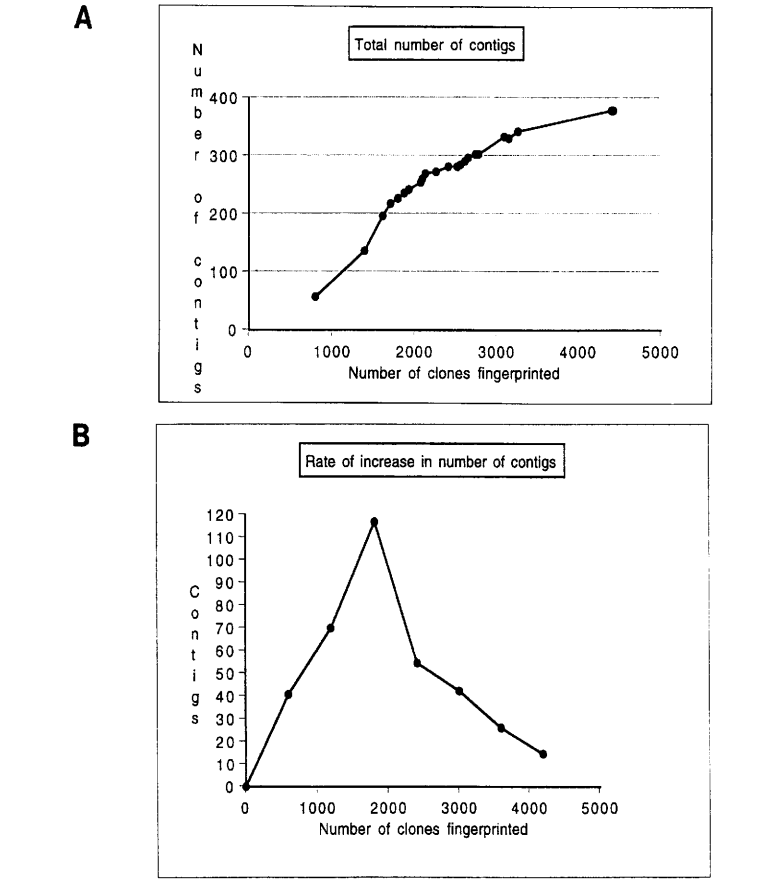

Figure 2 Various measurements of progress of the project. In these analyses the contigs containing repeats have been excluded. (*A*) The total number of contigs at various stages of the project. (*B*) The approximate rate of increase in contig number at various stages of the project. This was calculated every 600 fingerprints by counting the increase in the total number of contigs. (*C*) The number of contigs containing different numbers of clones at four stages of the project, after 1133, 2066, 3091, and 4362 clones had been fingerprinted.

been predicted by the computer analysis have been checked by manual examination of the original autoradiographs. This has allowed us to reject false contigs that arise when clones are mixed either in picking, during growth in the microtiter plate, or during gel loading. We also reject any match of <50% of bands when not supported by any other data (such as STS content).

To date, we have fingerprinted and analyzed 4430 clones (about five chromosome equivalents); 927 of these are from the 3E7 library and 3503 from the Lawrence Livermore Library. The clones fall into 377 contigs with a further 15 groups formed on the basis of repeats. Progress of the project is shown in Figure 2. We have not yet reached the point where we would anticipate that the total number of contigs will decrease. This is because the rate of joining is still exceeded by the rate of formation of contigs. However, the rate of increase in contig number is declining (Fig. 2A,B). The size of contigs, as determined by the number of clones in them, is also increasing as more clones are fingerprinted (Fig. 2C). As the random assembly began to proceed more slowly, we started to use approaches to select clones for fingerprinting. After 3143 clones, the data include some clones that have been preselected and so the data set is no longer entirely random. Preselection has produced contigs that are deeper than would have been generated by random fingerprinting.

Repeat Contigs

As the project progressed in its random phase, we noticed that some contigs contained a much larger number of clones than would be expected and it became impossible to order the clones in the contig. Some of these groups of clones could be recognized easily as containing tandem repeats by their fingerprint patterns, which con-

tained relatively few fragments of different sizes, but those that were present were intense relative to the vector bands suggesting many copies of the same size digestion product. Examples of the DYZ1- and DXYS20-containing clones are shown in Figure 3. In other cases, the fingerprint patterns did not look unusual. Such clones must contain larger, more complex repeated sequences. Examples of such groups include the clones containing the *RBM* (Ma et al. 1993) and *TSPY* (Arnemann et al. 1991) gene families. The *TSPY* or DYZ5 repeat (Manz et al. 1993) pseudocontig contains 78 clones that can neither be ordered nor divided into contigs from separate loci on the basis of their fingerprint patterns. At present, we have 15 groups of clones that either contain known or novel repeats.

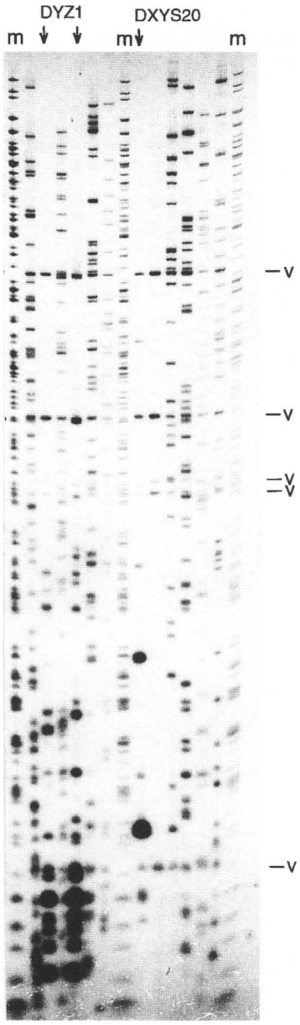

Figure 3 A gel including fingerprints of two clones that contain DYZ1 and one clone that contains DXYS20. DYZ1 clones were identified by hybridization and PCR. DXYS20 clones were identified on the basis of *Hin*fI restriction fragment size and confirmed by PCR and in situ hybridization to the pseudoautosomal region (data not shown).

Directed Approaches

The YAC framework and ordered STSs produced by Vollrath et al. (1992), Foote et al. (1992), and Slim et al. (1993a) have provided anchor points with which to position our randomly generated contigs and new loci at which we can select clones.

We have screened the libraries with 132 STSs either by hybridization of PCR products or by amplification of pooled cosmid DNA and have identified at least one cosmid containing each of 94 of these (Table 1). Clones that were positive by hybridization were tested by PCR. Not all clones identified by hybridization produced an amplification product. This could be explained in two ways. Either related loci exist that are sufficiently similar to cross-hybridize but are sufficiently diverged that the primers will not anneal, or although the PCR primers amplify a specific fragment, there is a repeated sequence somewhere in the product. For example, sY66, sY67, and sY69 all hybridize to the same, large set of cosmids. However, by PCR of pools of cosmid DNA, different single clones were identified for each of sY66 and sY67 out of 4000 cosmids screened. Table 1 includes only clones that would amplify by PCR and thus were confirmed as containing the STS.

Where clones are already members of a contig, work is in progress to check other clones in that contig by PCR to localize the STS within it. Alternatively, if the clones have not already been fingerprinted, this is being done and new contigs formed. In the cases where we have used clones obtained by screening to build contigs, they tend to be short and deep. Randomly fingerprinted clones are important in extending contigs and achieving wider coverage of the chromosome.

Of 46 STSs in the pseudoautosomal region, 40 (87%) identified at least one cosmid of which 38 (95%) had already been or have since been placed into contigs. In the Y-specific region, 5 of the 85 STSs tested were not tied to a single clone at this stage because of their repetitive nature (see Table 1 footnote). Of the remaining 80 STSs, 56 (70%) identified one or more cosmids. Twelve of these clones have not been fingerprinted and so we cannot yet say whether they are in contigs or not, and two STSs identify clones containing the

Table 1. Loci Used to Screen Y Libraries

Y region	STS	Locus	Cosmid	Contig[a]
	STS from Slim et al. (1993)			
Pseudoautosomal region Yp		DXYS129	19 F8	+
		DXYS60	118 H3	+
		DXYS153	17 F8	–
		DXYS130	68 A6	+
		DXYS131	103 F9	+
		DXYS132	103 F9	+
		DXYS59	111 B11	–
		DXYS133	52 C12	+
		DXYS134	52 C12	+
		DXYS135	7 D8	+
		DXYS136	52 C12	+
		DXYS137	60 D11	+
		DXYS138	70 C11	+
		DXYS139	75 C9	+
		DXYS140	75 C9	+
		DXYS141	75 C9	+
		DXYS142	37 C11	+
		HIOMT	82 C9	+
		DXYS144	60 D4	+
		DXYS143	94 C9	+
		DXYS17	109 G9	+
		XE7	79 H5	+
		DXYS145	10 A5	+
		DXYS146	—	–
		DXYS148	1 B2	–
		DXYS147	129 E6	+
		DXYS149	—	
	cDNA	MIC2	130 A8	+
		DXYS151	2 H10	+
		DXYS152	48 G4	+
		DXYS150	115 A3	+
	STS from Vollrath et al. (1992)			
	sY1	DXYS14	40 H2	+
	2	DXYS20	8 A10	+
	4	DXYS87	—	–
	5	DXYS86	130 B5	+
	3	DXYS28	126 C2	+
	6	DXYS15	—	–
	179	DXYS103	—	–
	7	DXYS85	67 B11	+
	cDNA	ANT3	54 E5	+
	9	CSF2RA	53 B5	+
	10	DXYS91	5 G5	+
	175	DXYS99	34 F8	+
	180	DXYS104	—	–
	11	DXYS92	117 C4	+
	12	DXYS93	117 C4	+
	178	DXYS102	35 H1	+
	13	DXYS77	110 E5	+
	cDNA	XG	37 E8	+

(*Table 1 continued on following pages.*)

Table 1. *(Continued)*

Y region	STS	Locus	Cosmid	Contig[a]
	Deletion intervals from Vollrath et al. (1992)			
Y-specific region				
1A1A	14	SRY	37 E8	+
1A1B	15	DYS234	71 B5	+
1A1B	16	RPS4Y	—	–
1A1B	17	RPS4Y	71 B5	+
1A2	18	DYS251	—	–
1A2	cDNA	ZFY	P 8H	–
1B	19	DYS252	K 10H	–
1C	20	DXYS42	K 10H	–
1D	21	DXYS69	72 G5	–
1E	22	DXYS106	72 G5	–
1E	23	DXYS107	6 G5	nfp
1E	24	DXYS5	—	–
2A	25	DXYS108	1 E1	nfp
2A	26	DXYS109	—	–
2A	28	DXYS110	40 A1	nfp
2A	29	dXYS111	3 F2	–
2A	30	DXYS112	30 B11	nfp
2A	31	DXYS113	—	–
2A	32	DXYS114	—	–
2A	33	DXYS115	11 D10	nfp
2A	34	DXYS253	—	–
2A	35	DXYS116	—	–
2A	36	DXYS2	repeat	–
2A	37	DXYS6	repeat	–
2A	38	DXYS117	—	–
2A	39	DXYS118	—	–
2C	46	DXYS122	1 F8	+
2C	47	DXYS9	39 H1	+
3A	48	DXYS12	29 H4	nfp
3A	50	DXYS124	23 E5	+
3C	57	DYS257	64 F8	nfp
3D	66	DYS261	5 C4	+
3E	67	DYS262	13 H3	+
3F	68	DYS263	—	–
3G	69	DYS264	C 9E	+
4A	70	AMG	C 9E	+
4A	71	DYS265	16 C4	+
4A	72	DYS266	27 E5	+
4A	73	DXYS1	16 C4	+
4A	74	DXYS267	—	–
Centromere	78	DYZ3	K 3H	repeat
5A	81	DYS271	37 H6	+
5B	82	DYS272	12 E5	+
5C	83	DYS11	6 D8	+
5C	84	DYS273	—	–
5C	85	DYS274	15 G9	+
5C	86	DYS148	—	–
5D	87	DYS275	4 F9	+
5D	165	DYS248	20 G4	+
5E	89	DYS277	32 F4	nfp
5E	90	DYS278	—	–

Table 1. *(Continued)*

Y region	STS	Locus	Cosmid	Contig[a]
5E	182	KAL	56 D1	+
5F	151	KAL	H 3G	
5G	91	DYS136	—	–
5G	92	DXYS126	29 C5	+
5G	93	DXYS127	10 E6	–
5G	94	DYS279	10 E6	–
5H	95	DYS280	25 H2	nfp
5I	96	DXYS3	—	–
5I	16	DYS243	13 F12	+
5I	97	DYS281	13 F12	+
5I	98	DYS282	—	–
5I	99	STS	—	–
5I	100	DYS196	28 G6	+
5I	101	DYS197	17 F11	+
5J	102	DYS198	2 E11	nfp
5J	103	DYS199	—	–
5J	104	DYS200	—	–
5K	105	DYS201	L 8C	nfp
5K	106	DYS202	12 E1	–
5M	113	DYS205	—	–
5M	114	DYS206	30 B4	+
5M	115	DYS207	30 B4	+
5M	116	DYS208	6 F8	+
5M	117	DYS209	6 F8	+
5M	118	DYS210	repeat	–
5N	121	DYS212	—	–
6A	132	DYS7	repeat	–
6A	134	DYS224	repeat	–
6A	135	DYS225	C 7H	+
6B	136	DYS226	Q 3G	–
6E	146	DYF52S1	C 11D	+
6F	156	DYS239	—	–
		CD24	35 F6	+
	cDNA	XGPY	37 5A	+
	cDBA	RBM	A 5F	repeat
	cDNA	TSPY	K 8D	repeat
		DAZ	6 B7	nfp
		160		
7		DYZ1	M 2A	repeat

Loci are presented in order along the chromosome from short arm teleomere to long arm. Only those loci with which we have screened the library are shown. Where one or more positive clones were identified, one clone name is given (tray number, column, row). Numerical trays are from the Lawrence Livermore Library; alphabetical trays are from our library. In some cases, a cosmid has not been assigned to an STS (–). Between 3840 and 5376 clones were screened by PCR, representing three to five chromosome equivalents; however, because we did not carry out the screening exhaustively a negative result here does not necessarily mean that the STS is absent from the library. Some STSs from repetitive regions of the genome amplify most or all of the master pools (repeat); and we have not gone on to identify single cosmids at these loci. [a](+) A contig associated with an STS. (–) Singletons. If a clone has not yet been fingerprinted (nfp) we cannot tell whether it is in a contig or not; additionally, groups of clones (rather than contigs) at repeated loci are indicated.

DYZ1 and DYZ3 repeats, which cannot be sorted into contigs. However, of the remaining 42 clones, 33 (79%) are in contigs. The higher proportion of STSs tied to cosmids in the pseudoau-

tosomal region reflects a more thorough screening strategy in this region rather than bias in the library. In the pseudoautosomal region the whole library was screened by hybridization with STS PCR products, and the positive clones subsequently confirmed by PCR. In the Y-specific region approximately one-third of the library was screened by PCR in a "single-pass" strategy, and no attempt was made at this stage to follow up negative or ambiguous results.

Where an STS identified two or more cosmids, this overlap was usually confirmed by fingerprinting. In two cases, this was not so, with the clones containing sY81 and sY101 being spread throughout several different contigs. Presumably these two STSs are also repetitive. In addition to the STSs already described, which include 11 gene-specific loci, we have used 6 cDNA probes to screen the library by hybridization and have identified cosmids containing 16 of 17 known Y chromosome genes and pseudogenes. One pseudogene, steroid sulfatase (*STSP*), was not identified. Contigs were built around all of the others except for *ZFY*, which remains as a singleton from the 3E7 library, *DAZ* (clones not yet fingerprinted), and the *RBM* and *TSPY* gene families. Thus, as far as genes are concerned, 16 of 17 (94%) of probes identified a cosmid, and all but one of the nonrepeated, fingerprinted clones are in contigs. Contigs around *AMELY* and *XG/SRY* are shown in Figure 4, A and B. This paper does not present details of the contigs, as our mapping information is available as an ACeDB file (Durbin and Thierry Mieg 1991) accessible by anonymous ftp from ftp://diamond.gene.ucl.ac.uk/pub/chrom.Y. This file can also be read directly as a text file and can be searched using text editors.

DISCUSSION

Although formulae exist to calculated the extent of genome coverage in a random fingerprinting project (Lander and Waterman 1988), our data, in common with some other projects (Harrison-Lavoie et al. 1989), have never conformed to their predictions. Now that we have included nonrandom data, these formulae are even less applicable. Simple calculations based on the size of contigs are affected both by the lack of direct measurement of contig size and by the possibility of undetected overlaps involving both contigs and singletons. A crude estimate of our coverage based on contig size and number is 23 Mb in contigs—~77% coverage assuming a genome size of 30 Mb—but singletons will increase this value by an unknown amount.

A better indication of the extent of coverage is obtained from the proportion of newly fingerprinted, randomly chosen clones that either fall into existing contigs or form a new contig with a singleton. This "hit rate" must be adjusted to account for the number of hamster and mouse contaminants in the library, which is 13% for the Lawrence Livermore Library. The unadjusted hit rate was 0.28 after 798 clones had been fingerprinted and rose steadily to 0.63 (average figure over the last 75 randomly fingerprinted clones). These figures include hits by clones that fall into repeat contigs. From this, we estimate that 72.5% of the Y chromosome is covered in fingerprinted clones. This estimate of the coverage excludes any regions of the chromosome that do not exist in the library.

A further estimation of the coverage of the chromosome in contigs that is not merely based around the proportion of the chromosome present in our library would be to take independent loci from another source and ask whether they exist in contigs. To a certain extent we have done this by assessing how many of the STSs fall into contigs. If we consider the total number of STSs and cDNAs (138) and the total number of these in contigs (75) it appears that 54% of the chromosome is covered in contigs. However, this experiment is not perfect for several reasons: We have not exhaustively looked for the STSs in the library so some that have no clone attached may actually be present; some of the STSs are repeated and have not been tied to one cosmid and some of the clones have not been fingerprinted. Taking only the single-copy STSs that have been attached to a fingerprinted clone, the coverage for the chromosome as a whole rises to well over 80%. The true value lies somewhere between these upper and lower limits.

The STSs may not be a particularly good indication of coverage, as they themselves appear to be quite clustered, representing selected regions of the chromosome. In several instances, more than one STS is found per clone, and although the 22 anchored contigs in the Y-specific region contain 44 STSs, they only span ~1.75 Mb, indicating, in this region, an average spacing of 1 locus per 40 kb rather than the 1 per 140 kb indicated by Foote et al. (1992).

The average size of contigs is ~60 kb (although the average size of the *anchored* contigs in the Y-specific region is 80 kb). This does not seem

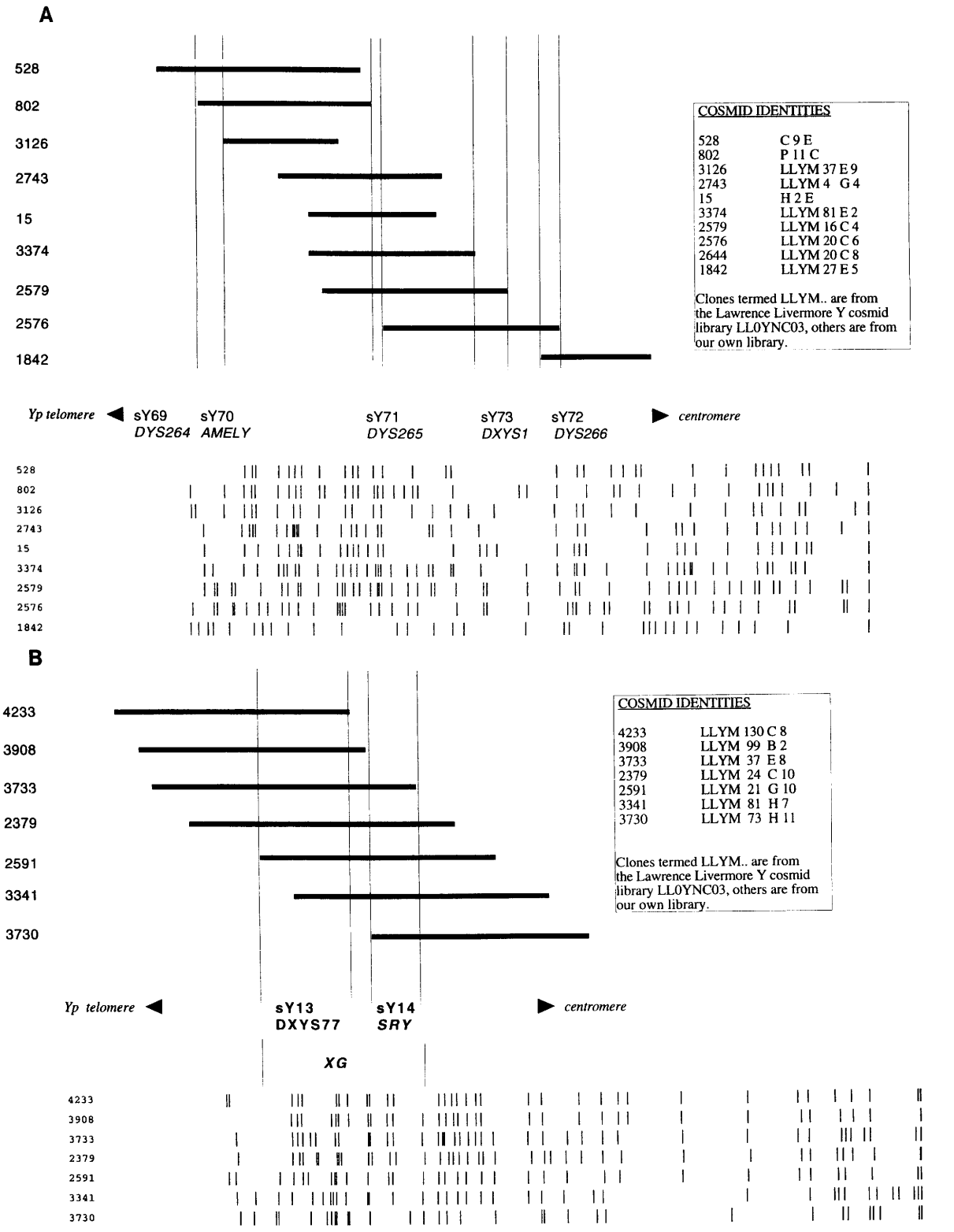

Figure 4 Contig diagrams. Each horizontal line represents a cosmid, and the length of the lines is proportional to the number of fingerprint bands. Numbers beside cosmids correspond to clone identities in the data base (a key is provided that gives the library identifications). Below each contig is shown the computer-generated (ContigC) representation of the clone fingerprints. (*A*) Region around amelogenin including four other STSs. (*B*) A contig spanning the Yp pseudoautosomal boundary containing the gene *SRY* and the 5′ end of *XG* along with the STS sY13.

very large considering that a cosmid is 35–40 kb long but is attributable to the high stringency conditions that we use to determine whether clones really overlap one another. Usually clones overlap by ~75% of their bands, and thus each new clone contributes ~10 kb to a contig. This high degree of overlap means that we can be very confident that clones placed in contigs belong there. Overlaps involving fewer (50%–70%) bands are checked by looking at the autoradiographs by eye and only accepted if there is no doubt. Of course, this stringency means that overlaps will be missed, but this is preferable to forming inaccurate contigs. The redundancy of the library (~12-fold) should mean that there will be sufficient clones at each locus to extend the contigs in this way. Each contig is verified again each time a new clone is entered and some adjustments of position may be made depending on which existing clones are hit by an incoming clone. We usually believe that the clones in a contig come from that locus, and this is confirmed when a probe hits several clones in a contig; however, the order of clones is not necessarily absolutely correct. In large contigs small position errors are magnified, and the computer autoalignment (used when two clones are so similar that they are not aligned by hand) generates its own small errors. Exceptions to this are contigs that contain repeats. These are completely unsorted and may well contain clones from more than one (albeit related) locus. Cosmid deletions do not appear to be a problem in general, and any deleted clones would not be present in the contigs because the band patterns would not line up properly; similarly clones with major rearrangements (e.g., chimeric clones) would not fit into a contig.

The STS content analysis has allowed us to position the contigs on the YAC map (Foote et al. 1992), which was in turn generated by ordering the STSs on deletion panels (Vollrath et al. 1992) However, we have found discrepancies between the order of STSs in our contigs and those in the YAC map; for example, in the amelogenin contig, sY72 and sY73 are inverted compared to the YAC map. Sometimes YAC ends cannot create enough deletion intervals to order a group of STSs, and our cosmids provide a useful way of doing so. Where more than one STS is found in a single cosmid, PCR of the other clones in that contig can separate the STSs into regions of the contig. This ordering of more than one STS in a contig allows the orientation of the contig to be determined (Fig. 4A,B). Approximately half of the contigs in the pseudoautosomal region can be oriented in this way, as they contain more than one STS due to the higher density of these markers in the region.

Currently we have 42 contigs at known positions on the chromosome. Future work will direct our cosmid selections to find clones that will bridge the gaps between adjacent contigs and look carefully for undetected overlaps between contigs thought to be very close to one another.

At present, more contigs are produced as more clones are analyzed, but after a certain number of clones have been added to the data base, contigs should join together faster than new contigs are made and the total number of contigs should decrease. Taking the *C. elegans* project as a guide, we would expect to reach a point where the random strategy is no longer effective when we have analyzed 6000 random clones. None of the mapping projects mentioned in the introductory section have achieved completion by purely random methods. We are therefore employing other strategies to complete the map, not least because of the problems that we have encountered attributable to the high incidence of low copy repeat number elements on the Y chromosome. The overriding problem with this project has been the presence of a large number of Y-linked repetitive sequences. These result in clones from different loci that contain copies of the repeat being placed into large contigs with many ends. STSs are useful as identifiers of single-copy loci, but when used as hybridization probes even these were seen to contain repetitive elements. Our overwhelming impression is that no single method is sufficiently powerful to produce a reliable map of this chromosome but that fingerprinting in conjunction with PCR and hybridization of STSs is capable of rendering a large portion of the chromosome ready for sequencing. However, there is no way of using cosmids alone to overcome the problem of repeats, and because of this we believe that a map based solely on cosmid contigs can never be completed. It will remain a series of islands that are surrounded by a sea of repetition.

METHODS

Library Preparation

A library of 750,000 clones was prepared in a way similar to that described in Wolfe et al. (1984b). DNA was prepared from the somatic cell hybrid 3E7 [which contains

the human Y chromosome as the major component of human material on a mouse background (Marcus et al. 1976)], partially digested with *Sau*3AI, size selected, and ligated into the *Bam*HI site of Lorist B (Cross and Little 1986). Clones containing human inserts were selected by hybridization to radiolabeled human male DNA: Colonies were grown on Hybond N membrane (Amersham), and then DNA was attached to the membrane by denaturation in alkali (1.5 M NaCl, 0.5 M NaOH), followed by neutralization (1.5 M NaCl, 0.5 M Tris at pH 8.0), and remval of bacterial debris by washing in 2× SSC. Filters were baked for 2 hr at 80°C, prehybridized in Church solution (0.5 M NaPi at pH 7.2, 7% SDS, 1 mM EDTA) and hybridized in a smaller volume of the same solution. The probe was sonicated human male DNA that was radiolabeled with [α-^{32}P] dCTP using Amersham oligolabeling kits. Filters were washed in Church wash (0.04 M NaPi at pH 7.2, 1% SDS) for 5 min at room temperature and then in 0.5× SSC for 2 × 20 min at 65°C. Clones giving a positive signal on duplicate filters were picked into gridded arrays and screened for a second time as described above. Clones representing 1728 independent signals were arrayed in microtiter plates. A second cosmid library (LL0YNC03 "M"), prepared in the vector Lawrist 16 using Y chromosomes flow-sorted from the somatic cell hybrid J640-51, which contains 3 intact human chromosomes Y, 22, and 9, has also been used. This library has ~13,000 members of which 82% are human, 13% are hamster, and 5% are nonrecombinants.

Dot Blot Hybridizations

DNA was prepared from clones in 96-well microtiter trays (Gibson and Sulston 1987) and used to produce DNA dot blots that were probed with a number of cloned DNA fragments localized previously to regions of the chromosome (Affara et al. 1986): DNA from a single micropreparation was used to produce four dot blot filters. DNA was denatured by the addition of 300 μl of 0.4 M NaOH, 10 mM EDTA, for 10 min at room temperature. Seventy-five microliters was then transferred by multichannel pipette to a dot blot apparatus fitted with Zetaprobe membrane (Bio-Rad) that had been prewet in distilled water. After vacuum suction the wells were washed with 300 μl of the same solution, which was also taken across the membrane by vacuum. Filters were washed in 2× SSC, air-dried, and baked for 1 hr at 80°C. Prehybridization was carried out in 5× SSC, 0.033 mg/ml sonicated herring sperm DNA, 2% SDS, and 10× Denhardt's solution at 65°C. Hybridization was carried out in the same solution with the addition of 5% dextran sulfate. Filters were washed for 15 min in 2× SSC/0.1% SDS at room temperature and then twice in 0.5× SSC/0.1% STS at 65°C usually for 20 min. Removal of bound probe was achieved by washing the filters twice in 0.1× SSC/0.5% STS at 95°C. Autoradiography was done overnight at −70°C.

Colony Filter Hybridizations

Cosmids were picked onto nylon membrane using 96-pin devices and lysed in situ as described above. Filters were hybridized to the PCR products from various STSs. PCR products were obtained using the primers and conditions described by Vollrath et al. (1992). Products were separated on agarose gels to confirm specificity, bands were excised, and the DNA was retrieved by centrifugation through siliconized glass wool followed by phenol extraction and ethanol precipitation. Products were radioactively labeled using Amersham random-prime kits.

Cosmid Fingerprinting

DNA isolated from 100 μl cultures of clones using the method of Gibson and Sulston (1987) was resuspended in 7.5 μl of TE buffer [10 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0)] and digested with the addition of the following reaction mix: 1 μl of ddH_2O, 1 μl of 10× buffer (as supplied by New England Biolabs), and 0.5 μl of *Hin*fI (10 U/μl; New England Biolabs) for 3 hr at 37°C. The resulting fragments were labeled by the addition of 2 μl of TE buffer, 1.5 μl of dATP (200 mM), 0.3 μl of Klenow enzyme (5U/μl, Boerhinger), and 1 μl [α-^{32}P]dCTP (3000 Ci/mmole, 0.2 μCi/μl in TE). The reaction was allowed to proceed for 15 min at room temperature and was terminated by the addition of 2 μl of gel-loading solution. Samples were electrophoresed in a 21 × 40-cm apparatus (Bio-Rad) in nondenaturing 4% acrylamide gels in TBE [gel = 4% 19:1 acrylamide/bisacrylamide in 1× TBE; 10× TBE = 0.89 M Tris base (pH 8.0), 0.89 M boric acid, 20 mM EDTA]. Gels were bonded to the front plate as described by Coulson et al. (1986). Samples were run alongside standards of λ DNA digested with *Hin*fI and end-labeled using [α-^{35}S]dATP and Kenow: 20 μl (10 μg) λ DNA (*c*1857 *ind1sam7*, New England Biolabs) was digested in 4 μl of 10× buffer supplied with the enzyme by New England Biolabs, 2 μl (20 units) of *Hin*fI, and 14 μl of H_2O for 3 hr at 37°C. The fragments were labeled by diluting the digestion reaction with 30 μl of TE and adding 4 μl of [α-^{35}S]dATP (1000 mCi/mmole, 12.5 μCi/μl) and 4 μl (20 units) of Klenow DNA polymerase. After incubation at room temperature for 15 min, the reaction was diluted further with three volumes of TE and terminated with 28 μl of gel-loading solution. Samples (2 μl of cosmid digests and 0.7 μl of markers) were loaded onto gels that were run at 50 W for 1 hr and 10 min, dried at 80°C, and exposed to Amersham Hyperfilm βmax overnight. It was important for computer analysis of the results that gels were run in uniform conditions.

Data Analysis

The analysis of the data was semiautomatic but required manual interaction. The autoradiograph was scanned by an Amersham gel reader that digitized the image. The information was then transferred to a VAX3100 workstation where image processing programs detected the bands, arranged them into lanes, and recorded the position of each band in relation to the standards (Sulston et al. 1988, 1989). For each clone pair, matches between bands were ranked in terms of the probability that the match was attributable to chance, based on the number of bands in each matching pair and the tolerance limit of band positions (0.7 mm). [See Sulston et al. (1988) for detail of the format of data generated by the computer.] The computing system used in this project was described in Heding et al. (1992). Later analyses were performed on a SUN IPX workstation using the programs Image and ContigC. (For

information about these programs, contact http://www/sanger/ac/uk/~fw/ContigC/index.html.)

Cosmid Library DNA Pools

Plates 1–40 from the Lawrence Livermore Library and 16 plates from our library were pooled. Primary pools consisted of colonies from eight consecutive trays. DNA was prepared from these using maxiprep columns (Promega) and resuspended in 2 ml of TE. DNA from subpools from individual plates, and rows and columns from the eight trays (28 subpools per primary pool) was prepared by pelleting the colonies, resuspending in 1 ml of TE, boiling for 5 min to release DNA, and then pelleting cell debris. One microliter of the supernatant was used in each PCR reaction.

PCR Primers

The primers used in the selection of cosmids were mainly those from Slim et al. (1993a) and Vollrath et al. (1992), and the conditions used were essentially those of Vollrath et al. Additional PCR primers were designed to the genes *HIOMT, CD24, XE7,* and *RPS4Y*. Primers were from Rodriguez et al. (1994), Hough et al. (1994), Ellison et al. (1992), and Fisher et al. (1990), respectively.

HIOMT	5′TCAGCCTAGCCTCGTGTTTT3′
	5′GGAGTTACGACTCAGCGAGG3′
and also	5′ACCAGAATGCCACCACCTAG3′
	5′GACGTTGGCACAGCCTCT3′
CD24	5′AGGGGACATGGGCAGAGCAA3′
	5′AAGAGACTGGCTGTTGACTG3′
XE7	5′GACGTCCTGGTCAAGGTGTT3′
	5′GAAGCCCATGTACTCACGGT3′
RPS4Y	5′ATCCTGTCATCAAGGTGAACG3′
	5′GCCATTGGCATCCTTCAC3′

Probes L24A and pMF1 were used to screen for*ZFY* [North et al. (1991) and Palmer et al. (1991), respectively]. *ANT3* cDNA was a gift from Gudrun Rappold (Slim et al. 1993b), *MIC2* probes p44C1.2, 0.65E, and 3′Rsa were obtained from Peter Goodfellow (Smith et al. 1993), the *XG* probe BM22Y was obtained from Polly Weller (Ellis et al. 1994), *RBM* cDNA MK5 was obtained from Ma Kun (Ma et al. 1993), and *TSPY* cDNA was a clone generated in this laboratory.

ACKNOWLEDGMENTS

We thank John Sulston for advice and for alteration of some of his programs, Nabeel Affara for DNA probes, and Maria Lioumi, Anjala Pradhan, Patricia Cogram, and Alec Ostler for pinning some STSs to cosmids. The chromosome-specific gene library LL0YNC03 used in this work was constructed at the Biomedical Sciences Division, Lawrence Livermore National Laboratory, under the auspices of the National Laboratory Gene Library Project sponsored by the U.S. Department of Energy. The work was funded by Medical Research Council Human Genome Mapping Project (MRC HGMP) grants G9119139 and 9226072 to J.W., a collaborative grant from the EEC (now the European Union) to J.W., Drs. N. Affara, C. Tyler-Smith, and J. Weissenbach (PL 910006). D.C. was funded by an MRC HGMP studentship (G78/1098), and M.A. received a Wellcome summer studentship (VS/95/UCL/002/GM/jj).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Affara, N.A., M.A. Ferguson-Smith, J.L. Tolmie, K. Kwok, M. Mitchell, D. Jamieson, A. Cooke, and L. Florentin. 1986. Variable transfer of Y-specific sequences in XX males. *Nucleic Acids Res.* **14:** 5375–5387.

Affara, N.A., Y.-F.C. Lau, H. Briggs, P. Davey, M.H. Jones, O. Khwaja, M. Mitchell, and C. Sargent. 1994. Report of the first international workshop on Y chromosome mapping. *Cytogenet. Cell. Genet.* **67:** 360–402.

Arnemann, J., S. Jakubiczka, S. Thuring, and J. Schmidtke. 1991. Cloning and sequence analysis of a human Y-chromosome-derived, testicular cDNA, TSPY. *Genomics* **11:** 108–114.

Bishop, C.E., G. Guellaen, D. Geldwerth, R. Voss, M. Fellous, and J. Weissenbach. 1983. Single-copy DNA sequences specific for the human Y chromosome. *Nature* **303:** 831–832.

Carrano, A.V., J. Lamerdin, L. Ashworth, B. Watkins, E. Branscomb, T. Slezak, M. Raff, P. De Jong, D. Keith, L. McBride, S. Meister, and M. Kronick. 1989. A high resolution, fluorescence-based semiautomatic method for DNA fingerprinting. *Genomics* **4:** 129–136.

Cooke, H. 1976. Repeated sequence specific to human males. *Nature* **262:** 182–186.

Coulson, A., J. Sulston, S. Brenner, and J. Karn. 1986. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **83:** 7821–7825.

Coulson, A., R. Waterson, J. Kiff, J. Sulston, and J. Kohara. 1988. Genome linking with yeast artificial chromosomes. *Nature* **335:** 184-186.

Coulson, A., Y. Kozono, B. Lutterbach, R. Showkeen, J. Sulston, and R. Waterston. 1991. YACs and the *C. elegans* genome. *BioEssays* **13(8):** 413–417.

Cross, S.H. and P.F.R. Little. 1986. A cosmid vector for systematic chromosome walking. *Gene* **49:** 9–22.

Daiger, S.P., R.S. Ildin, and T.-S. Su. 1982. Sequences on the human Y chromosome homologous to the

autosomal gene for arginosuccinate synthetase. *Nature* **298:** 682–684.

Durbin, R. and J. Thierry Mieg. 1991. A C. elegans database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk. and ncbi.nlm.nih.gov.

Ellis, N.A., T.Z. Ye, S. Patton, J. German, P.N. Goodfellow, and P. Weller. 1994. Cloning of PBDX, an MIC2-related gene that spans the pseudoautosomal boundary on the chromosome Xp. *Nature Gentet.* **6:** 394–400.

Ellison, J.W., C. Ramos, P.H. Yen, and L.J. Shapiro. 1992. Structure and expression of the human pseudoautosomal gene *XE7. Hum. Mol. Genet.* **1:** 691–696.

Fisher, E.M.C., P. Beer-Romero, L.G. Brown, A. Ridley, J.A. McNeil, J. Lawrence, H.F. Willard, F.R. Bieber, and D.C. Page. 1990. Homologous ribosomal protein genes on the human X and Y chromosomes: Escape from X inactivation and possible implications for Turner syndrome. *Cell* **63:** 1205–1218.

Foote, S., D. Vollrath, A. Hilton, and D. Page. 1992. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258:** 60–66.

Garza, D., J.W. Ajioka, D.T. Burke, and D.L. Hartl. 1989. Mapping the *Drosophila* genome with yeast artificial chromosomes. *Science* **246:** 641–646.

Gibson, T.J. and J. Sulston. 1987. Preparation of large numbers of plasmid DNA samples in microtitre plates by the alkaline lysis method. *Gene Anal. Technol.* **4:** 41–44.

Goodfellow, P.J., S. Darling, G. Banting, B. Pym, C. Mondello, and P.N. Goodfellow. 1987. Pseudoautosomal genes in man. *Development* (Supppl.) **101:** 119-125.

Harrison-Lavoie, K.J., R.M. John, D. Porteous, and P.F.R. Little. 1989. A cosmid clone map derived from a small region of human chromosome 11. *Genomics* **5:** 501–509.

Heding, I.J.P., A.C. Ivens, J. Wilson, M. Strivens, S. Gregory, J.M.N. Hoovers, M. Mannens, B. Redeker, D. Porteous, V. Van Heyningen, and P.F.R. Little. 1992. The generation of ordered sets of cosmid DNA clones from human chromosome region 11p. *Genomics* **13:** 89–94.

Hough, M.R., P.M. Rosten, T.L. Sexton, R. Kay, and R.K. Humphries. 1994. Mapping of CD24 and homologous sequences to multiple chromosomal loci. *Genomics* **22:** 154–161.

Jones, M.H., O.S. Khwaja, H. Briggs, B. Lambson, P.M. Davey, J. Chalmers, C.Y. Zhou, E.M. Walker, Y. Zhang, C. Todd, M.A. Ferguson Smith, and N.A. Affara. 1994. A set of ninety-seven overlapping yeast artificial chromosome clones spanning the human Y chromosome euchromatin. *Genomics* **24:** 266–275.

Knott, V., D.J.G. Rees, Z. Cheng, and G.G. Brownlee. 1988. Randomly picked cosmid clones overlap the *pyr*B and *ori*C gap in the physical map of the *E. coli* chromosome. *Nucleic Acids Res.* **16:** 2601–2612.

Knott, V., D.J. Blake, and G.G. Brownlee. 1989. Completion of the detailed restriction map of the *E. coli* genome by the isolation of overlapping cosmid clones. *Nucleic Acids Res.* **17:** 5901–5912.

Kohara, Y., K. Akiyama, and K. Isono. 1987. The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50:** 495–508.

Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Ma, K., A. Sharkey, S. Kirsch, P. Vogt, R. Keil, T.B. Hargreave, S. McBeath, and A.C. Chandley. 1992. Towards the molecular localization of the *AZF* locus: Mapping of microdeletions in azospermic men within 14 subintervals of interval 6 of the human Y chromosome. *Hum. Mol. Genet.* **1:** 29–33.

Ma, K., J. Inglis, A. Shakey, W.A. Bickmore, R.E. Hill, E.J. Prosser, R. Speed, E.J. Thomson, M. Jobling, K. Taylor, J. Wolfe, H.J. Cooke, T.B. Hargreave, and A.C. Chandley. 1993. A Y chromosome gene family with RNA-binding protein homology: Candidates for the azoospermia factor *AZF* controlling human spermatogenesis. *Cell* **75:** 1287–1295.

Manz, E., F. Schneiders, A.M. Brechlin, and J. Schmidke. 1993. TSPY-related sequences represent a microheterogeneous gene family organized as constitutive elements in DYZ5 tandem repeat units on the human Y chromosome. *Genomics* **17:** 726–731.

Marcus, M., R. Tantravahi, V. Dev, D. Miller, and O. Miller. 1976. Human-mouse cell hybrid with human multiple Y chromosomes. *Nature* **262:** 63–65.

North, M., C. Sargent, J. O'Brien, K. Taylor, J. Wolfe, N.A. Affara, and M.A. Ferguson-Smith. 1991. Comparison of ZFY and ZFX gene structure and analysis of alternative 3′ untranslated regions of ZFY. *Nucleic Acids Res.* **19:** 2579–2586.

Olson, M.V., J. Dutchik, M. Graham, G. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman, and T. Frank. 1986. Random clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci.* **83:** 7826–7830.

Osthuizen, C.J.J., J.S. Herbert, L.K. Vermaar, J. Brusnicky, J. Frike, L. Du Plessis, and A.E. Retief. 1990. Deletion mapping of 39 randomly isolated Y-chromosome DNA fragments. *Hum. Genet.* **85:** 205–210.

O'Reilly, A.J., N.A. Affara, E. Simpson, P. Chandler, E. Goulmy, and M.A. Ferguson-Smith. 1992. A molecular deletion map of the Y chromosome long arm defining X and autosomal homologous regions and the localization

of the HYA locus to the proximal region of the Yq euchromatin. *Hum. Mol. Genet.* **1:** 379–385.

Palmer, M.S., P. Berta, A.H. Sinclair, B. Pym, and P.N. Goodfellow. 1990. Comparison of human ZFY and ZFX transcripts. *Proc. Natl. Acad. Sci.* **87:** 1681–1685.

Rodriguez, I.R., K. Mazuruk, T.J. Schoen, and C.J. Chader. 1994. Structural analysis of the human hybdroxyindole-O-methyltransferase gene. Presence of two distinct promoters. *J. Biol. Chem.* **269:** 31969–31977.

Schmid, M., M. Guttenbach, I. Nanda, R. Studer, and J.T. Epplen. 1990. Organisation of DYZ2 repetitive DNA on the human Y chromosome. *Genomics* **6:** 212–218.

Siden-Kiamos, I., R. Saunders, L. Spanos, T. Majerus, J. Treanear, C. Savakis, C. Louis, D.M. Glover, M. Ashburner, and F.C. Kaftos. 1990. Towards a physical map of the *Drosophila melanogaster* genome: Mapping of cosmid clones within defined genomic divisions. *Nucleic Acids Res.* **18:** 6261–6270.

Slim, R., D. Le Paslier, S. Compain, J. Levilliers, P. Ougen, A. Billault, S.J. Donohue, D.C. Klein, L. Mintz, A. Bernheim, D. Cohen, J. Weissenbach, and C. Petit. 1993a. Construction of a yeast artificial contig spanning the pseudoautosomal region and isolation of 25 new sequence tagged sites. *Genomics* **16:** 691–697.

Slim, R., J. Levilliers, H.J. Judecke, U. Claussen, V.C. Nguyen, N.M. Gough, B. Horsthemke, and C. Petit. 1993b. A human pseudoautosomal gene encodes the ANT3 ADP/ATP translocase and escapes X-inactivation. *Genomics* **16:** 26–33.

Smith, K.D., K. Young, C. Conver Talbot, and B.J. Schmeckpeper. 1987. Repeated DNA of the human Y chromosome. *Development* (Suppl.) **101:** 77–92.

Smith, M.J., P.J. Goodfellow, and P.N. Goodfellow. 1993. The genomic organisation of the human pseudoautosomal gene MIC2 and the detection of a related locus. *Hum. Mol. Genet.* **2:** 417–422.

Stallings, R., D. Torney, C.E. Hildebrand, J.L. Longmire, L. Deaven, J. Jett, N. Doggett, and K. Moyzis. 1990. Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci.* **87:** 6218–6222.

Sulston, J., F. Mallett, R. Staden, R. Durbin, T. Horsnell, and A. Coulson. 1988. Software for genome mapping by fingerprinting techniques. *Compost. Appl. Biosci.* **4:** 125–132.

Sulston, J., F. Mallet, R. Durbin, and T. Horsnell. 1989. Image analysis of restriction enzyme fingerprint autoradiograms. *Compost. Appl. Biosci.* **5:** 101–106.

Tyler-Smith, C. and W.R. Brown. 1987. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* **195:** 457–470.

Vergnaud, G., D.C. Page, M-C. Simmler, L. Brown, F. Rouyer, B. Noel, D. Botstein, A. De La Chapelle, and J. Weissenbach. 1986. A deletion map of the human Y chromosome based on DNA hybridization. *Am. J. Hum. Genet.* **38:** 109–124.

Vollrath, D., S. Foote, A. Hilton, L.G. Brown, P. Beer-Romero, J.S. Bogan, and D. Page. 1992. *Science* **258:** 52–59.

Wada, M., R.D. Little, F. Abidi, G. Porta, T. Labella, T. Cooper, G. Della Valle, M. D'Urso, and D. Schlessinger. 1990. Human Xq24-Xq28: Approaches to mapping with yeast artificial chromosomes. *Am. J. Hum. Genet.* **46:** 95–106.

Wolfe, J., R.P. Erickson, P.W.J. Rigby, and P.N. Goodfellow. 1984a. Regional localisation of 3 Y-derived sequences on the human X and Y chromosomes. *Ann. Hum. Genet.* **48:** 253–259.

Wolfe, J., R.P. Erickson, P.W.J. Rigby, and P.N. Goodfellow. 1984b. Cosmid clones derived from both euchromatic and heterochromatic regions of the human Y chromosome. *EMBO J.* **3(9):** 1997–2003.

Received November 17, 1995; accepted in revised form March 1, 1996.