Chapter 3

Construction of a Sequence-Ready Bacterial Clone Contig

3.1 Introduction

The generation of clone maps covering large regions of the human genome has evolved significantly over the last five years. The international collaboration to map and sequence the human genome has brought about the adaptation of existing methods and the development of new ways to generate bacterial clone maps to sequence large genomes accurately and efficiently. Some of these developments, such as bacterial clone fingerprinting were pioneered in the mapping and sequencing of small genomes (*C. elegans* Sequencing Consortium, The, 1998; Coulson, A.*, et al.*, 1986; Goffeau, A.*, et al.*, 1996; Olson, M. V.*, et al.*, 1986). This chapter will describe the application and evaluation of large-scale mapping techniques and describe how they evolved along with the available resources during the construction of a sequence-ready bacterial clone map covering approximately 6 Mb of human chromosome Xq22 between DXS366 and DXS1230.

A 6.5 Mb YAC map was previously constructed in Xq22 between DXS366 and DXS87 (Vetrie, D.*, et al.*, 1994) and included four genes and fifteen previously mapped genetic markers (Dib, C.*, et al.*, 1996) (see Figure 3.1). The genes had previously been identified because of their role in a variety of diseases and included the PLP gene, defects in which cause Pelizaeus-Merchbacher Disease (PMD) (Hudson, L. D.*, et al.*, 1989; Trofatter, J. A.*, et al.*, 1989). There are still a number of diseases for which no gene has been cloned, but for which the critical regions include the region of interest in this chapter between DXS366 and DXS1230. For instance, genetic analysis of a family with X-linked megalocornea showed close linkage to DXS94 and DXS87 (Mackey, D. A.*, et al.*, 1991).

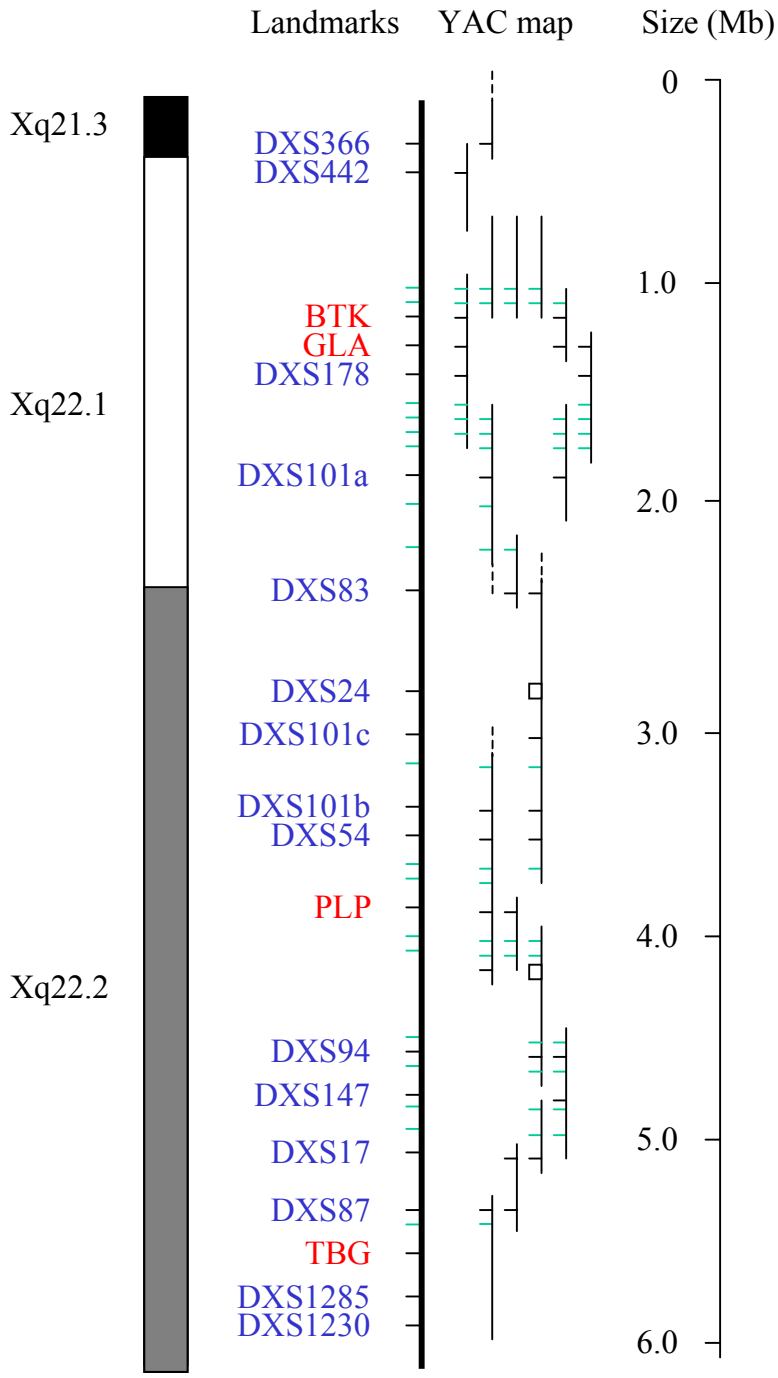


Figure 3.1: *The status of the region of interest before the generation of the bacterial clone contig began (modified from Vetrie, D., et al., 1994). The YAC map was constructed using known genes (shown in red), genetic markers (shown in blue) and additional end probes (shown as green lines). The region is approximately 6 Mb based on the sizing of YACs by pulsed-field gel electrophoresis.*

The generation of a sequence-ready bacterial clone contig in Xq22 would provide the basis for a more detailed study of the genes and sequence contained within the region, particularly with respect to the diseases for which the gene remains uncloned.

RESULTS

3.2 Contig construction

The strategy to construct a bacterial clone contig within Xq22 was based on using the available YAC map to generate initial coverage in bacterial clones (see Figure 3.2). Prior to the start of the project, a subset of the YACs from the available YAC contig (Vetrie, D., *et al*., 1994) were pooled and used as probes by Elaine Kendall and Dave Vetrie (Guy's Hospital) to screen gridded arrays of two cosmid libraries (LLNX01 and GHc, see Section 2.7 in M&M's), which represented the best available sources of X chromosome enriched bacterial clones at the time. A subset of the YAC clones were used individually as probes to screen the cosmids. All positive cosmids were rescreened with individual YAC clones.

At the start of the project I was provided with a total of 1400 cosmids which were fingerprinted using a radioactive label, the method developed during the mapping of the *C. elegans* genome (Coulson, A*., et al.*, 1986) (see Section 2.12.1). The fingerprints for all the cosmids were digitised using IMAGE (see Section 2.23.1) and contigs were assembled using FPC (see Section 2.23.2). A total of 26 contigs covering 3 Mb or 50% of the region were generated, an example of which is shown in Figure 3.3. A summary of the status of the mapping after the cosmid fingerprinting is shown in Figure 3.16a.

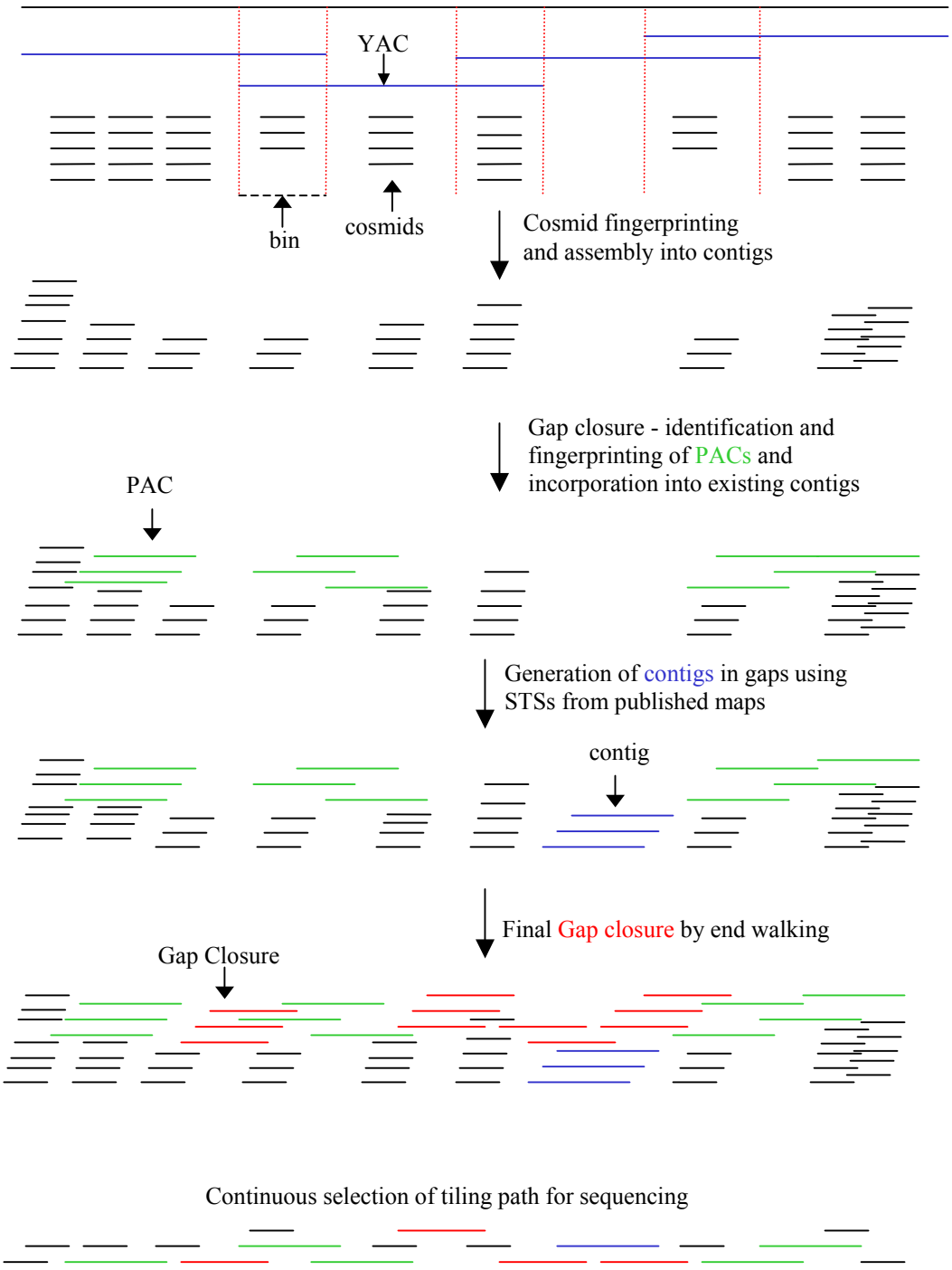


Figure 3.2: *Strategy for the construction of the bacterial clone contig. Binned cosmids are fingerprinted and assembled into contigs. Whole cosmid hybridisation identifies PACs to close gaps and extend contigs. New contigs are generated in gaps using STSs from published maps before final gap closure. At each stage, clones are chosen for genomic sequencing.*

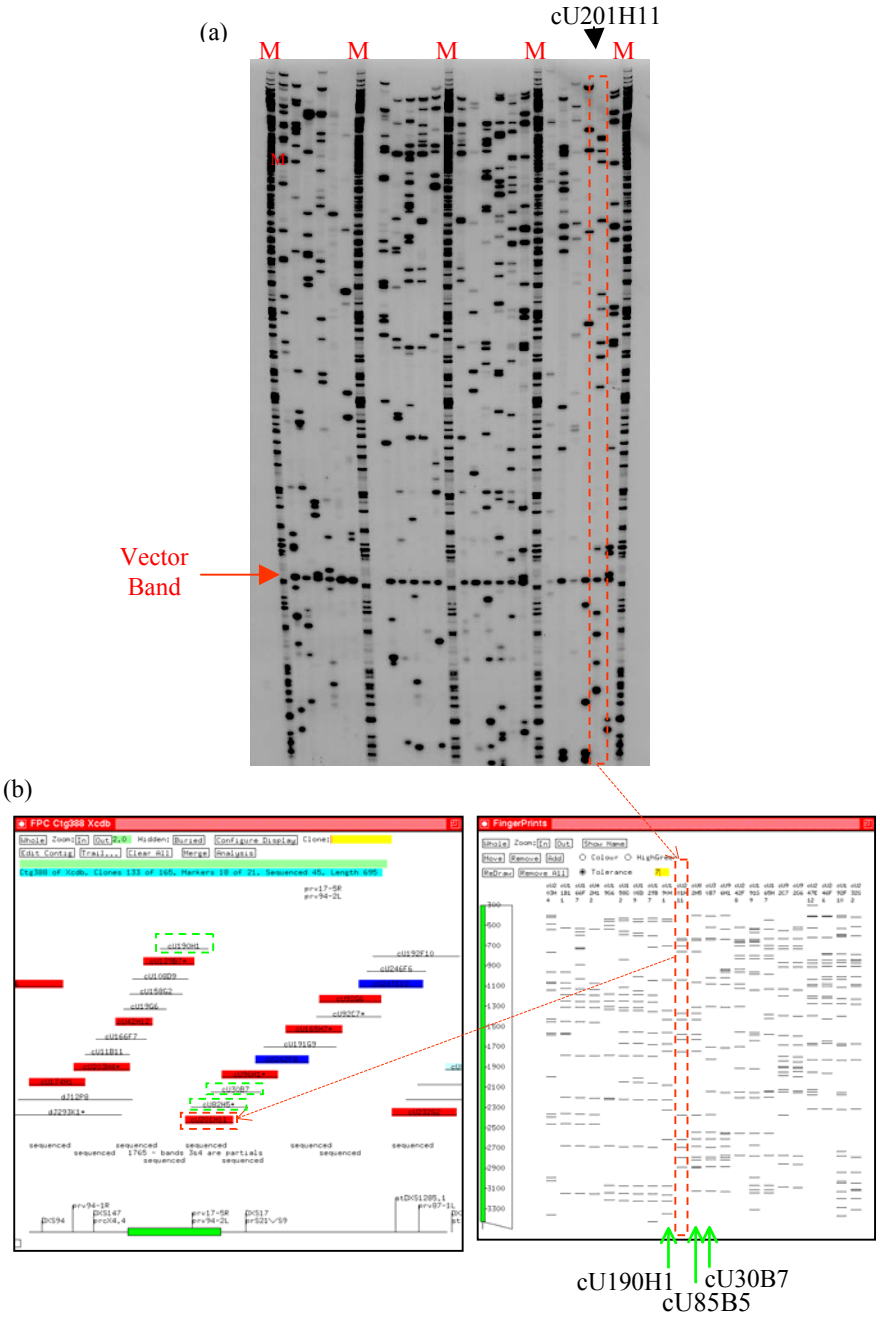


Figure 3.3: *Cosmid fingerprinting and assembly (a) An autoradiograph of the fingerprints of 24 cosmids. A marker (M) is run every seventh lane and a common vector band is seen in all sample lanes as indicated. (b) A section of one of the contigs constructed and their fingerprints. For example, lane 23 contains the fingerprint for cU201H11, which was found to overlap significantly (greater than or equal to 1e –04 - see Section 2.23.2) with cU190H1, cU82H5 and cU30B7 when compared to all other fingerprints in the FPC database.*

It had been reported that DXS101 had five loci in the region (Vetrie, D.*, et al.*, 1994), two copies each at DXS101a and DXS101b, and a single copy at DXS101c. The five DXS101 loci can be distinguished by digestion of the DNA with *Eco*R1, Southern blotting of the digested DNA and probing with the DXS101 plasmid cX52.5. Each locus generates specific size fragments (5.5 kb and 7.0 kb for DXS101a, 6.0 kb and 11.5 kb for DXS101b and 13.0 kb for DXS101C). Hybridising cX52.5 (the DXS101 probe) to the available cosmids identified all those that contain the DXS101 loci (work carried out by Elaine Kendall). From the work carried out in this thesis, the fingerprinting and analysis of the DXS101-positive cosmids assembled the cosmids into three contigs. Based on the original binning of the cosmids, two of the contigs represented three of the DXS101 loci, and one contig appeared to contain both copies of DXS101 present at DXS101c (see Figure 3.4).

At this time the first of the large-insert bacterial clone libraries (PACs) became available. In order to close gaps between existing contigs, radioactively labelled *Hin*d III-digested cosmids were pooled and used as probes to hybridise to gridded arrays of PAC clones from RPCI-1 (Ioannou, P. A.*, et al.*, 1994) (see Figure 3.5). A total of 149 PACs were identified with 33 cosmids. Seven cosmids failed to identify any PACs, based on the lack of overlapping PAC clones when the fingerprints were compared to those of the cosmids. It was estimated that the RPCI-1 library represented three genome equivalents and the screening carried out at this stage showed that, on average four PACs were identified with each cosmid probe, which was roughly equivalent to what was expected (three PACs for each cosmid probe).

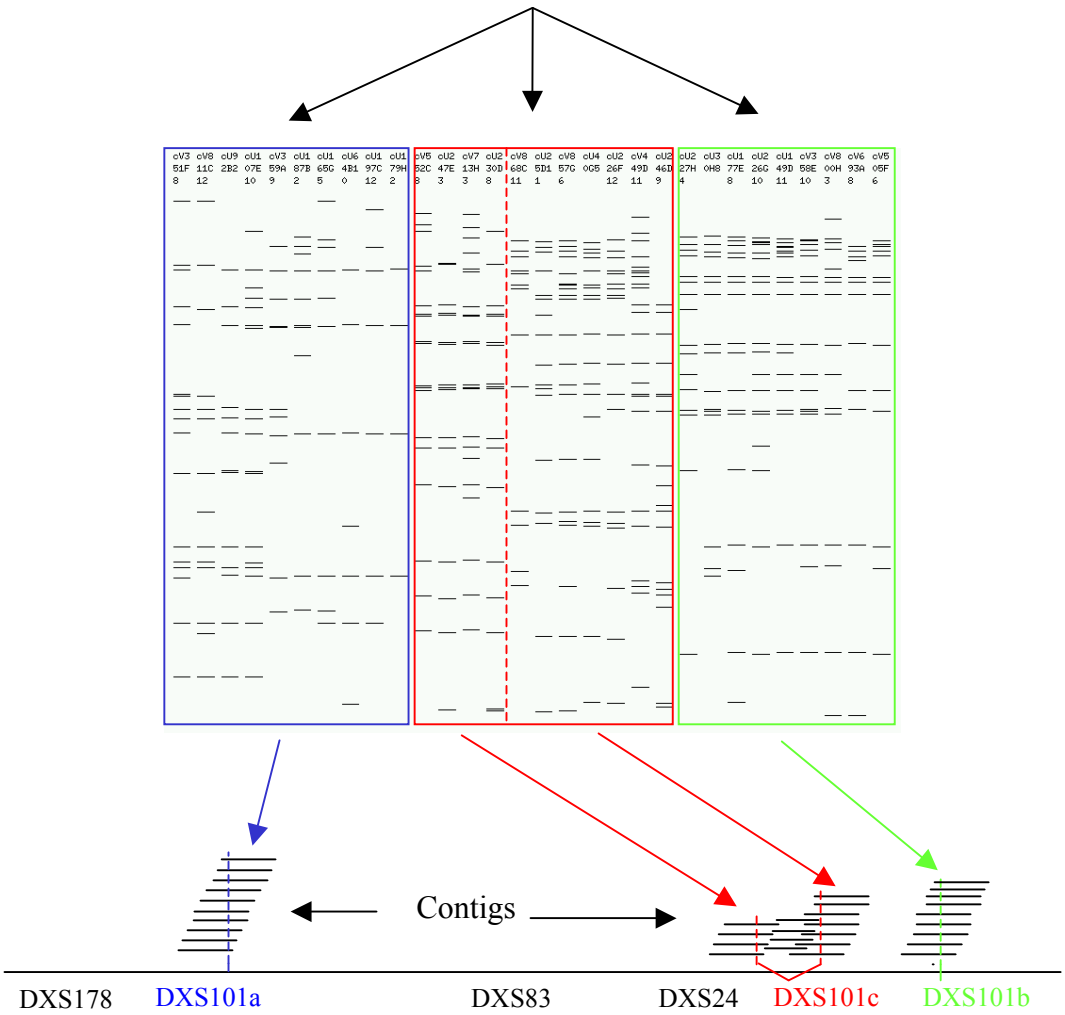
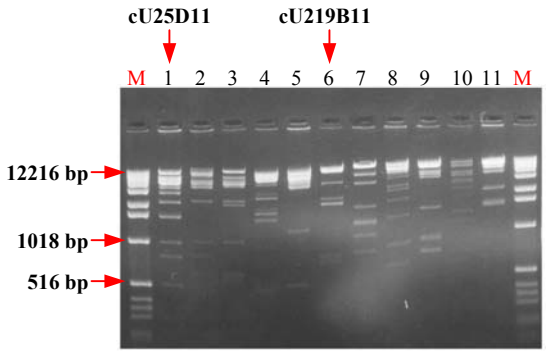


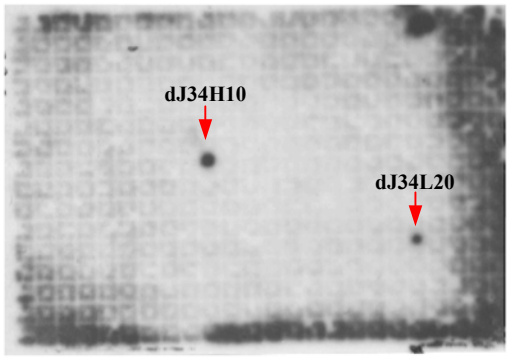
Figure 3.4: *Fingerprinting of DXS101-positive cosmids. The DXS101 positive cosmids were assembled into three contigs, based on fingerprinting, representing four of the five different loci for DXS101. Fingerprinting could not distinguish the two loci within DXS101b*

Figure 3.5*: (see over) PAC isolation by whole cosmid hybridisation (a) A photograph of a gel showing 11 cosmids digested with Hind III. Marker lanes are indicated (M). (b) An autoradiograph of one filter from the gridded PAC library showing the positive PACs identified when 6 of the 11 cosmids, including cU25D11 and cU219B11, were hybridised as a pooled probe. (c) The section of the contig showing cU25D11 overlapping with dJ34H10 and their fingerprints (dJ1055C14 and dJ593F11 were identified with an STS designed to the end of cU35G3 later in the project– data not shown). As described in Section 2.23.2, overlaps between clones are based on the number of bands they have in common. The bands in the fingerprint of dJ34H10 are shown in blue, and equivalent bands in the fingerprints of other clones are shown in red. Black bands in dJ34H10 do not match any other bands, and black bands in other clones do not match any bands in dJ34H10. The 3 black bands in cU25D11 were not seen in any other clones in the contig and were supposed to be false positives.*

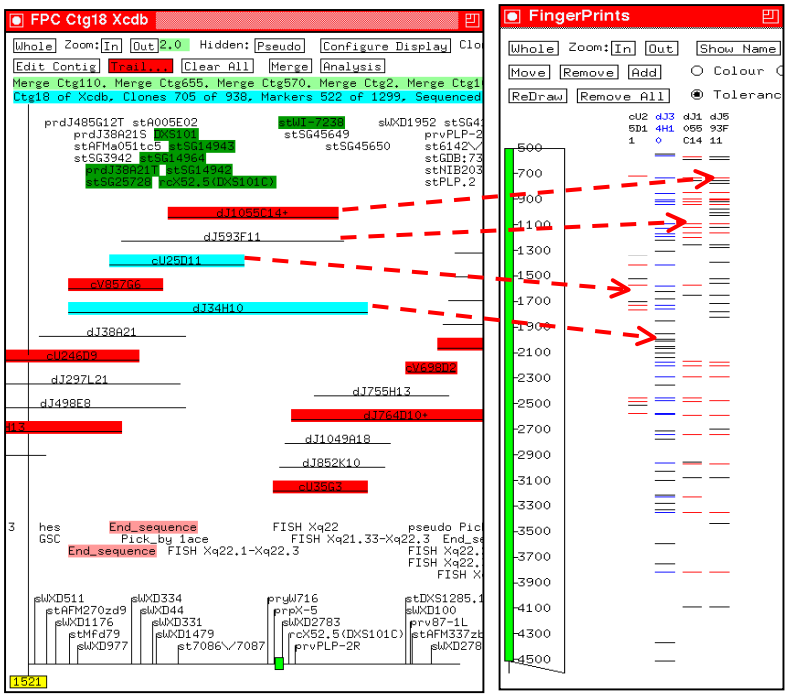
(a)


cU25D11
cU219B11
M 1 2 3 4 5 6 7 8 9 10 11 M
12216 bp
1018 bp
516 bp

(b)


dJ34H10
dJ34L20

(c)

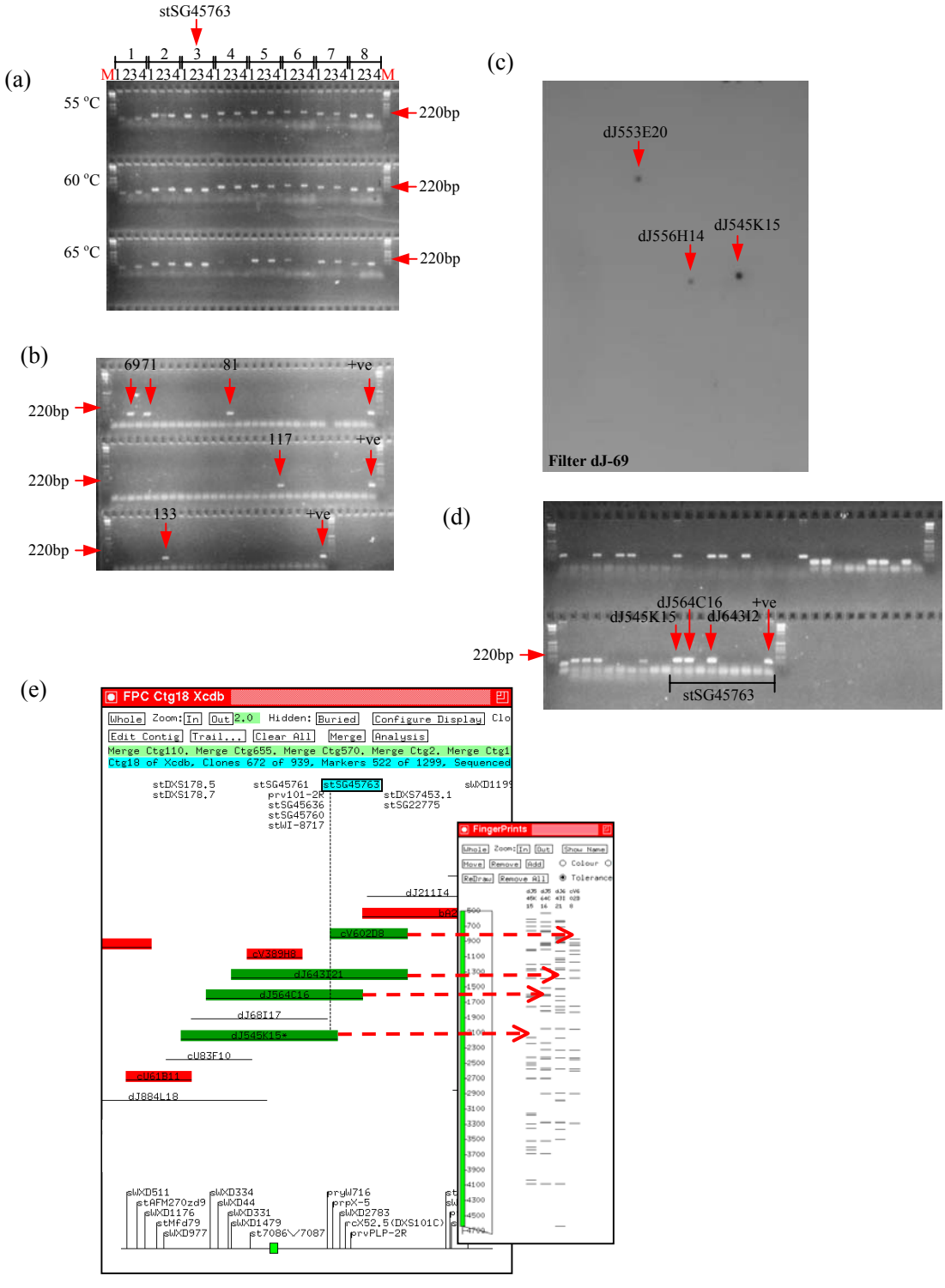

FPC Ctg18 Xcdb
FingerPrints
dJ1055C14+
dJ593F11
cU25D11
cV957G6
dJ34H10
dJ38A21
cU246D9
dJ297L21
dJ498E8
dJ1049A18
dJ852K10
cU35G3
dJ755H13
dJ764D10+
End_sequence
FISH Xq22.1-Xq22.3
FISH Xq22
pseudo Pic
1521

The PACs were fingerprinted and 80 of the 149 (55%) were incorporated into the existing contigs thus closing fifteen gaps and extending seven ends of contigs. At the end of this second stage there were ten contigs covering 3.9 Mb or 65% of the region (see column (b) of Figure 3.16). At this time, fingerprinting by fluorescent labelling was developed (Gregory, S. G.*, et al.*, 1997). In order to benefit from the increased throughput and safety of this technique, a subset of clones (all cosmids identified for sequencing and all PAC clones) were fingerprinted using this method and assembled into the same set of contiguous blocks. Other cosmids that were radioactively fingerprinted and assembled into contigs, but did not form part of the minimum set chosen for genomic sequencing, were not re-fingerprinted.

At this point in the project, the clones which extended the ends of the contigs furthest were the larger insert PAC clones, and whole cosmid hybridisation was therefore no longer useful to identify further bacterial clones for gap closure. Hybridisation of probes derived from whole PAC clones (similar to whole cosmid hybridisation) to the PAC library would have been problematical given the cross hybridisation of vector DNA between probe and target clones. Insert-specific amplification, e.g *Alu* PCR, is limited by the fact that only a fraction of the insert is amplified. At this time an STS-based YAC map was published across the region (Srivastava, A.K., *et al*., 1999), containing additional STSs that had not been available previously. Twenty-seven STSs thought to lie in gaps between existing bacterial clone contigs (see Table 2.4) were screened against the sections of the PAC library (RPCI-1, 2, 3). This identified 50 new PAC clones which were fingerprinted and assembled into three new contigs. An example of clone isolation using these novel STSs is shown in

Figure 3.6 (see also Figure 3.16c). At the end of the third stage there were 16 contigs covering 5 Mb or 80 % of the region.

Figure 3.6: *(see over) PAC isolation using STSs taken from YAC map of Srivastava, A. K., et al. (1999) (a) Eight STSs (1-8) designed from sequence generated at the ends of 10 clones were tested for their ability to amplify unique sequence in human genomic DNA at three different temperatures of the PCR. Templates included human DNA (1), X-chromosome hybrid (2), hamster genomic DNA (3) and $T_{0.1}E$ (4). (b) One of the STSs, stSG45763 designed to one end of cV602D8 (see Table 2.5) was used to amplify DNA of pools 67-150 (each containing 2912 PAC clones) from RPCI-3 library. Five positive pools were detected, as indicated. A positive control (human genomic DNA) was run in parallel. (c) The product of amplification of genomic DNA using stSG45763 was labelled, pooled with 9 other products, and used as a hybridisation probe to screen gridded filters representing clones of each pool (1 filter represents 1 pool). The filter shown represents pool 69, and 3 positives were identified as marked. (d) Positive clones detected on filters representing the pools shown positive in (b) were streaked and individual colonies tested against stSG45763. The two other positives on filter 69 (dJ556H14 and dJ553E20) were found to be positive for another unrelated STSs. The other positive clones were identified from other filter hybridisations. (e) The fingerprints of the 3 clones show good correspondence of fingerprint patterns confirming overlap and were integrated into the contig by comparison with other fingerprints previously in the database (e.g cV602D8).*


(a)
stSG45763
1 2 3 4 5 6 7 8
M12341234123412341234123412341234M
55 °C
60 °C
65 °C
220bp
220bp
220bp
(b)
6971
81
+ve
220bp
117
+ve
220bp
133
+ve
220bp
(c)
dJ553E20
dJ556H14
dJ545K15
Filter dJ-69
(d)
dJ564C16
+ve
dJ545K15
dJ64312
220bp
stSG45763
(e)
FPC Ctg18 Xcdb
Whole Zoom:In Out 2.0 Hidden: Buried Configure Display
Edit Contig Trail... Clear All Merge Analysis
Merge Ctg110, Merge Ctg655, Merge Ctg570, Merge Ctg2, Merge Ctg1
Ctg18 of Xcdb, Clones 672 of 939, Markers 522 of 1299, Sequenced
stDXS178.5
stDXS178.7
stSG45761
stSG45763
prv101-2R
stSG45636
stSG45760
stWI-8717
stDXS7453.1
stSG22775
FingerPrints
dJ211I4
cV602D8
cV389H8
dJ643I21
dJ564C16
dJ68I17
dJ545K15*
cU83F10
cU61B11
dJ884L18
stAFM270zd9
stMfd79
prpX-5
cX52.5(DXS101C)
prvPLP-2R

Closure of the remaining gaps was completed using either probes generated by vectorette end rescue (see Sections 2.19 and 2.20) from the ends of clones at the ends of contigs, or STSs generated after directly sequencing the ends of the cloned PAC inserts (sequencing was carried out by Elizabeth Huckle). DNA of clones chosen for vectorette end rescue was prepared using a standard alkaline lysis and subsequent phenol chloroform extraction, and digested with *Rsa*I (see Section 2.11.1). Vectorette 'bubbles' (see Table 2.3) were ligated on to the ends of the restriction fragments and amplification of each end of the insert of each clone was carried out using vector specific primers. The amplification products were resolved on agarose gels, excised and stored to be used as templates for probe generation.

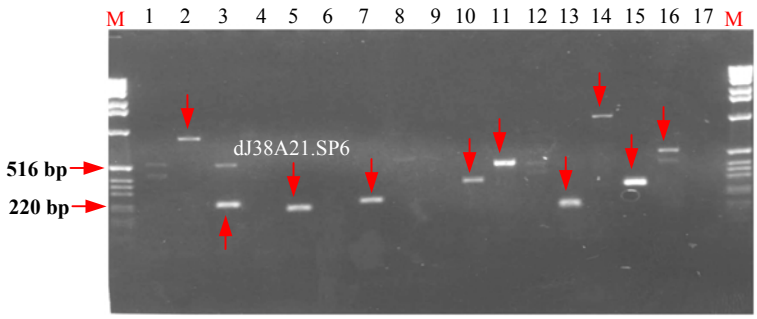
The vectorette probes or STSs from end sequencing were used to screen six bacterial clone libraries that were now available (RPCI-1, 3, 4, 5, 11 and 13) (see Figure 3.7). In total, 131 PACs and 101 BACs were identified using 20 probes and 103 STSs. All newly identified clones were fingerprinted and incorporated into the contigs and all remaining gaps were closed.

The final bacterial clone contig covers approximately 6 Mb of Xq22 between DXS366 and DXS1230 and contains 92 cosmids, 211 PACs and 101 BACs (see Figure 3.8). A total of 44 probes (24 positioned by Elaine Kendall and Dave Vetrie, 20 vectorette end probes positioned duing this project) and 130 STSs (103 STSs designed to sequence generated at the end of the clones, 27 STSs from YAC map of Srivastava, A.K., *et al*., 1999) have been used in the construction of the contig. A minimum set of clones from the contig was chosen for genomic sequencing (carried

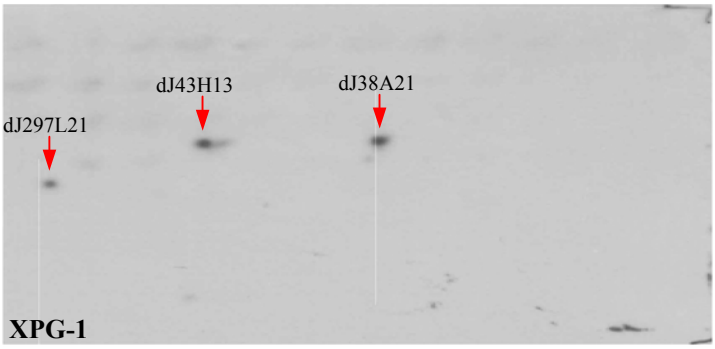
out by the Sanger Centre sequencing teams) and resulted in one contiguous segment of sequence covering 6.0 Mb.

Figure 3.7: *(see over) PAC isolation using STSs generated by vectorette PCR or end sequencing (a) Generation of 10 products from the SP6 ends of 17 PAC clones by vectorette PCR (successful amplification is indicated by an arrow). Lanes 3 and 16 contain two bands, in each case the stronger one was excised and used to generate a probe for walking. (b) One product, dJ38A21.SP6, was labelled and used as a hybridisation probe to screen two filters representing an X chromosome-specific collection of PAC and cosmid clones (XPG-1 and XPG-2). 6 positive clones were identified, 3 of which were present on the filter shown and are indicated. (c) The fingerprints of the 6 clones show good correspondence of fingerprint patterns confirming the overlap. The 5 clones (highlighted in green) were integrated into the bacterial clone map by comparing the fingerprints with other fingerprints previously in the database (e.g dJ1143D15).*

(a)


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 M
dJ38A21.SP6
516 bp
220 bp

(b)


dJ43H13
dJ38A21
dJ297L21
XPG-1

(c)

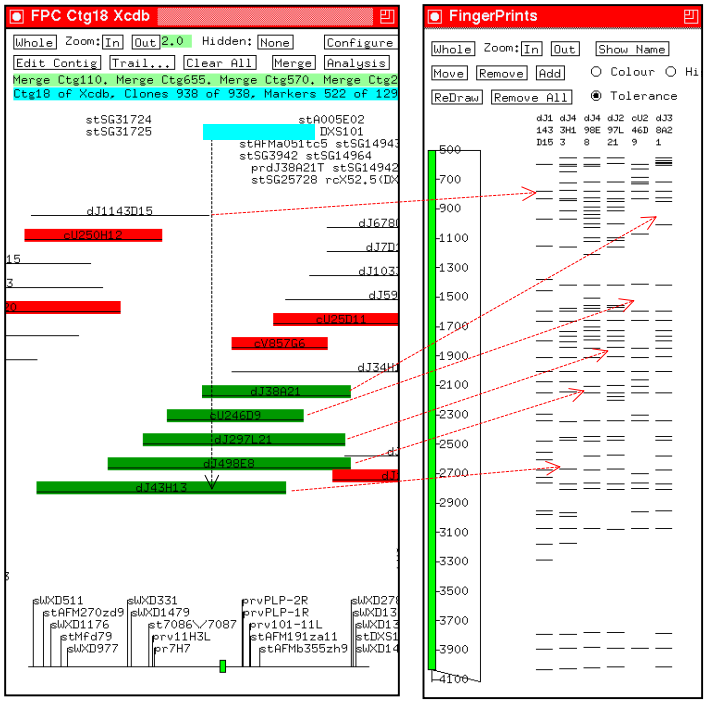
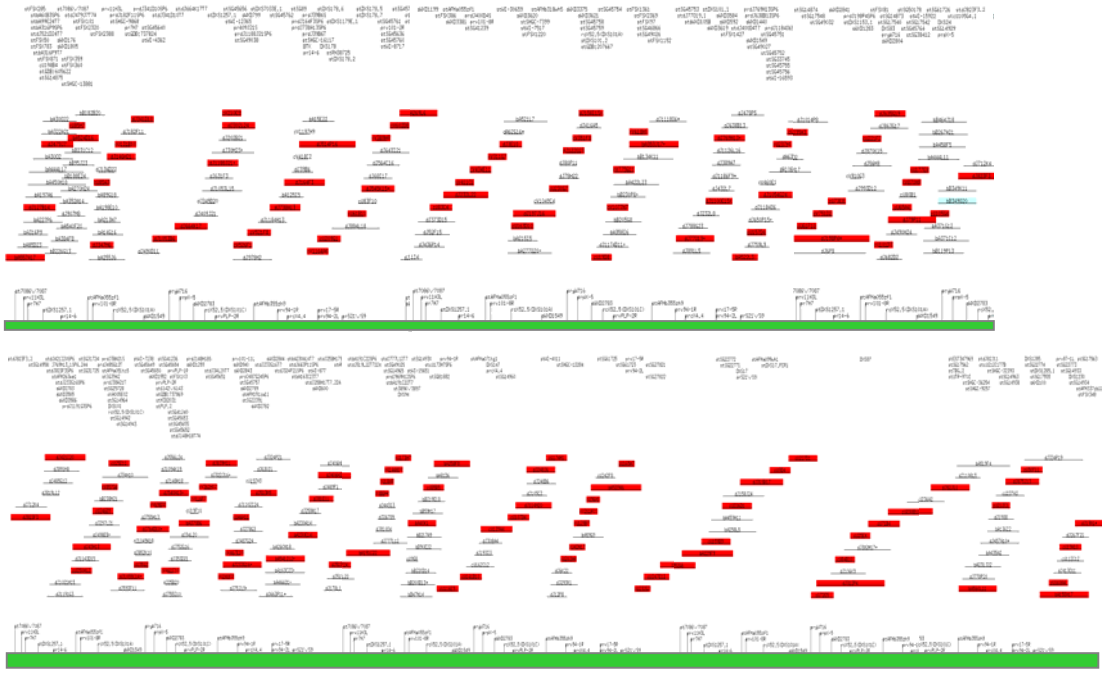

FPC Ctg18 Xcdb
FingerPrints
Whole
Zoom: In Out
Hidden: None
Configure
Edit Contig
Trail...
Clear All
Merge
Analysis
Show Name
Move
Remove
Add
Colour
ReDraw
Remove All
Tolerance
dJ1143D15
cU250H12
dJ67B0
dJ59
cU25D11
cV957G6
dJ34H1
dJ38A21
cU246D9
dJ297L21
dJ498E8
dJ43H13
stSG31724
stSG31725
stA005E02
DXS101
stSG3942 stSG14964
stSG25728
stSG14942
stSG14943
Merge Ctg110. Merge Ctg655. Merge Ctg570. Merge Ctg2
Ctg18 of Xcdb, Clones 938 of 938, Markers 522 of 129
sWXD511
stAFM270zd9
sWXD1176
stMfd79
sWXD977
sWXD331
sWXD1479
st7086\/7087
prv11H3L
pr7H7
prvPLP-2R
prvPLP-1R
prv101-11L
stAFM191za11
stAFMb355zh9
sWXD13
stDXS1
sWXD14
500
700
900
1100
1300
1500
1700
1900
2100
2300
2500
2700
2900
3100
3300
3500
3700
3900
4100

Figure 3.8: *(see over) FPC diagram of bacterial clone contig between DXS366 and DXS1230. The markers used to identify the clones and confirm contig order during earlier stages of the project are shown at the top. A subset of the markers, chosen as framework markers are indicated at the bottom. Clones in the contig are indicated by horizontal black lines, the length of the clone is determined by the number of bands in the fingerprint. The overlap between clones is determined by the number of bands pairs of clones have in common. The clones shown in red were identified for the minimum tiling path for genomic sequencing.*



For full diag see Xq22FPCctg
Diag 3 (horizontal page set

3.3 Comparison of the published maps

The generation of the complete sequence of the region of interest allows for the study of the accuracy of previously published maps in order to verify marker order and placement and identify conflicts with the final sequence map. It also allows for a comparison of physical distance with genetic distances. There are two types of maps available covering the region of interest between DXS366 and DXS1230: a genetic map (Dib, C.*, et al.*, 1996) and three physical maps; the RH map (electronic version released in 1999, updated from Deloukas *et al* (1998) and two YAC/STS based-map (Srivastava, A. K.*, et al.*, 1999, Vetrie, D.*, et al.*, 1994). All STSs on these published maps have been accurately positioned on the sequence and the order and distance from neighbouring STSs compared.

3.3.1Genetic Map

The order and physical distances of markers on the final sequence map were compared with the order and genetic distances of markers on the available genetic map (Dib, C., *et al*., 1996) (see Figure 3.9). The region between DXS366 and DXS1230 contains eight genetic markers that are placed within 2.5 cM of each other on the genetic map. Three markers (shown in grey in Figure 3.9), placed within the same region on the genetic map could not be identified in the sequence between DXS366 and DXS1230. All available X chromosome finished and unfinished sequence was searched using BLAST (Altschul, S. F.*, et al.*, 1990) and the results revealed that two markers (AFMb083yb5 and AFMa052xc1) are located approximately 500 kb distal to DXS1230 and the third (AFMa162yc9) is located approximately 4 Mb distal to DXS1230.

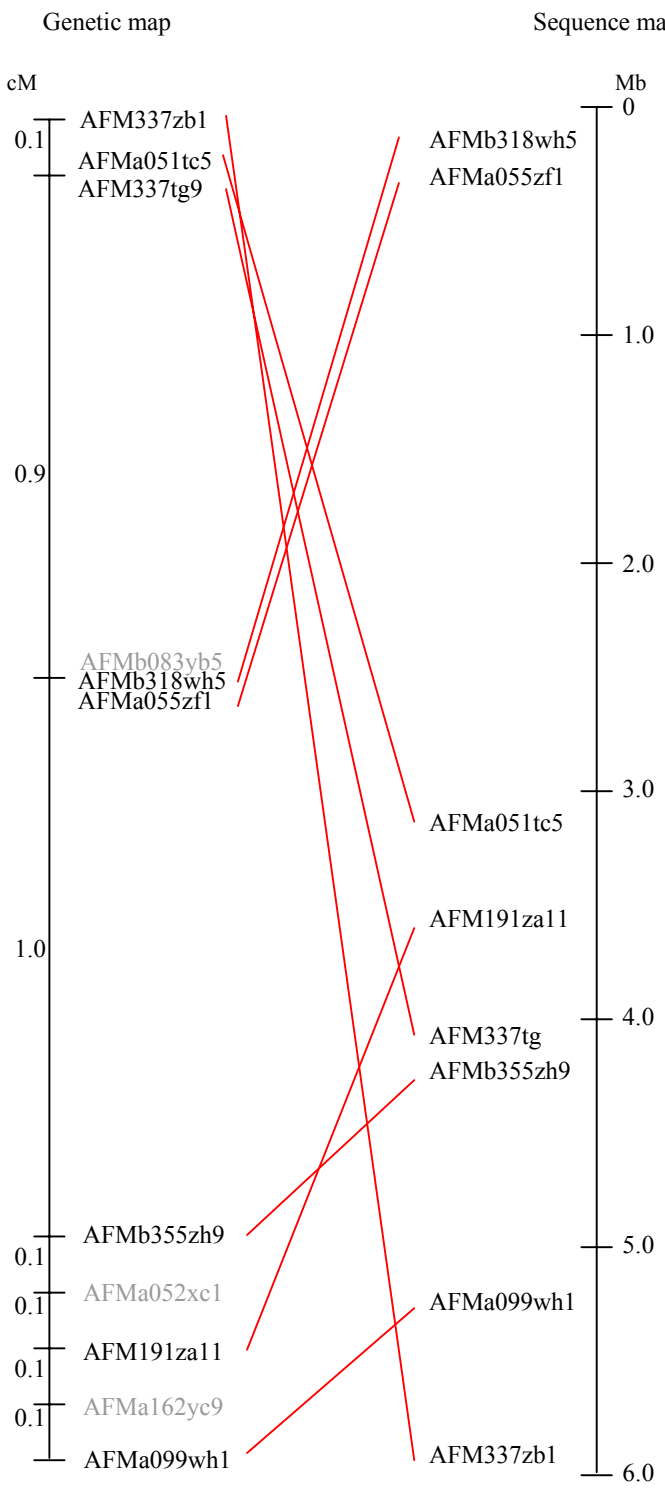


Figure 3.9: *Comparison of the genetic map. A comparison of the STS order and genetic versus physical distance, between the published genetic map (on the left) and the final sequence map (on the right). Red bars link the same genetic marker and marker names in grey have not been able to be placed within the final sequence map.*

There is some disagreement between the order of the eight genetic markers identified in the sequence and their order on the genetic map. For instance, AFMb318wh5 and AFMa055zf1 are placed 0.9 cM distal to AFM337zb1, AFMa051tc5 and AFM337tg9, but are 3 Mb more proximal in the sequence. Analysis of the draft sequence showed that for long chromosome arms 1 cM equals approximately 1 Mb whereas for the shortest chromosome arms 2 cM equals approximately 1 Mb. Recombination is also not uniform across these regions. There are regions where recombination is less frequent (e.g. towards centromeres) and other regions that appear to have a higher recombination frequency (e.g. towards telomeres) (IHGSC, 2001). Three markers (AFM337zb1, AFMa051tc5 and AFM337tg9) have been placed within 0.1 cM on the genetic map, but cover a distance of 3 Mb on the sequence map, which may represent a region of low level recombination. Although the genetic mapping has been able to cluster the eight markers in one region of the genome, it has not been able to identify their correct order. The eight markers have been positioned on the genetic map with odds of greater than 1000:1 that there is no other more likely position. However, the markers are all positioned within a 3 cM interval, which is reaching the limitations of resolution for genetic mapping. This may account for the differences in the marker order based on genetic mapping, and the actual marker order identified from the sequence.

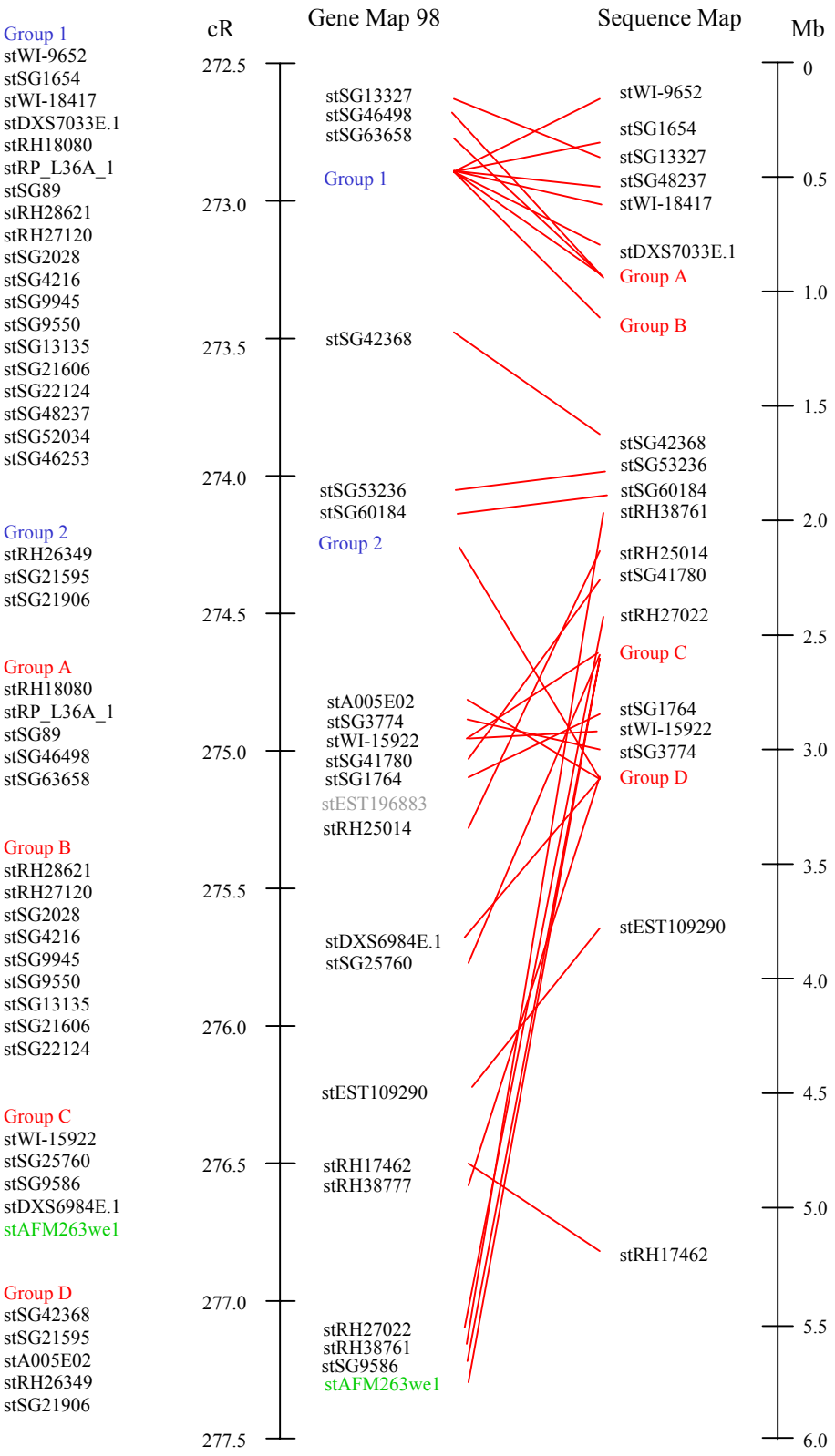
3.3.2 RH Map

The order and physical distances (given in centiRays (cR) of markers placed on the RH map were compared to the order and distance of markers on the final sequence map (see Figure 3.10). The region of interest between DXS366 and DXS1230 is located within one bin of the RH map between the two framework markers DXS990

and DXS1106 (the STS for DXS1106, stAFM263we1, is positioned within the region and is shown in green in Figure 3.10). The 6 Mb region is estimated to be 5 cR according to the RH map, which is similar to the reported figure that on average, 1 cR is equivalent to 1 Mb (Deloukas, P.*, et al.*, 1998). A total of 44 markers have been positioned within the region of interest by RH mapping and 43 of them have been located within the final sequence map (see Figure 3.10). The sequence for the one marker that has not been placed in the final sequence map (stEST196883 – shown in grey on Figure 3.10) did not match any human genome sequence currently available. In one case, stWI-15922 appears once on the RH map but twice on the final sequence map.

The main difference between the two maps lies between 275.0 cR and 277.5 cR on the Gene Map (a region of 2.5 cR) and between 2.5 Mb and 3.0 Mb on the sequence map (a region of 0.5 Mb) where there is also some discrepancy between the marker order and marker distances. The markers in the region of RH map between 275.0 cR and 277.5 cR are clustered within the 0.5 Mb region of the final sequence map.

Figure 3.10: *(see over) Comparison of the gene map. A comparison of the STS order and centi-Ray versus physical distance between the published RH Map (on the left) and the final sequence map (on the right). Red lines link the same markers. Names in grey indicate those markers that could not be placed within the final sequence map. Groups 1, 2, A, B, C and D represent clusters of markers positioned too closely on either the RH map or in the sequence map to be resolved on the diagram.*


Group 1
stWI-9652
stSG1654
stWI-18417
stDXS7033E.1
stRH18080
stRP_L36A_1
stSG89
stRH28621
stRH27120
stSG2028
stSG4216
stSG9945
stSG9550
stSG13135
stSG21606
stSG22124
stSG48237
stSG52034
stSG46253
Group 2
stRH26349
stSG21595
stSG21906
Group A
stRH18080
stRP_L36A_1
stSG89
stSG46498
stSG63658
Group B
stRH28621
stRH27120
stSG2028
stSG4216
stSG9945
stSG9550
stSG13135
stSG21606
stSG22124
Group C
stWI-15922
stSG25760
stSG9586
stDXS6984E.1
stAFM263we1
Group D
stSG42368
stSG21595
stA005E02
stRH26349
stSG21906
cR
Gene Map 98
272.5
273.0
273.5
274.0
274.5
275.0
275.5
276.0
276.5
277.0
277.5
stSG13327
stSG46498
stSG63658
Group 1
stSG42368
stSG53236
stSG60184
Group 2
stA005E02
stSG3774
stWI-15922
stSG41780
stSG1764
stEST196883
stRH25014
stDXS6984E.1
stSG25760
stEST109290
stRH17462
stRH38777
stRH27022
stRH38761
stSG9586
stAFM263we1
Sequence Map
Mb
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
stWI-9652
stSG1654
stSG13327
stSG48237
stWI-18417
stDXS7033E.1
Group A
Group B
stSG42368
stSG53236
stSG60184
stRH38761
stRH25014
stSG41780
stRH27022
Group C
stSG1764
stWI-15922
stSG3774
Group D
stEST109290
stRH17462

3.3.3 YAC maps

There are two published YAC maps that include the region of interest between DXS366 and DXS1230. As mentioned in the introduction to this chapter, one of the published YAC maps was used as the basis for generating the initial coverage in bacterial clones (Vetrie, D*., et al.*, 1994). The marker order in this YAC map was consistent with the order on the final sequence map (data not shown). Although the markers in the YAC map were used to identify and orient bacterial clone contigs during the construction of the final sequence-ready contig, the order of the markers was confirmed independently through genomic sequencing and there are no inconsistencies between the bacterial clone contig and the sequence map.

The order of the markers on the second published YAC map (Srivastava, A. K*., et al.*, 1999) was compared to the order in the final sequence (see Figure 3.11). There are four regions that show inconsistencies, primarily through inversions of groups of markers. Although the order generated by Srivastava, A.K., *et al*. differed from the final order in the sequence map, the YAC map was a valuable resource during this project as a source of STSs.

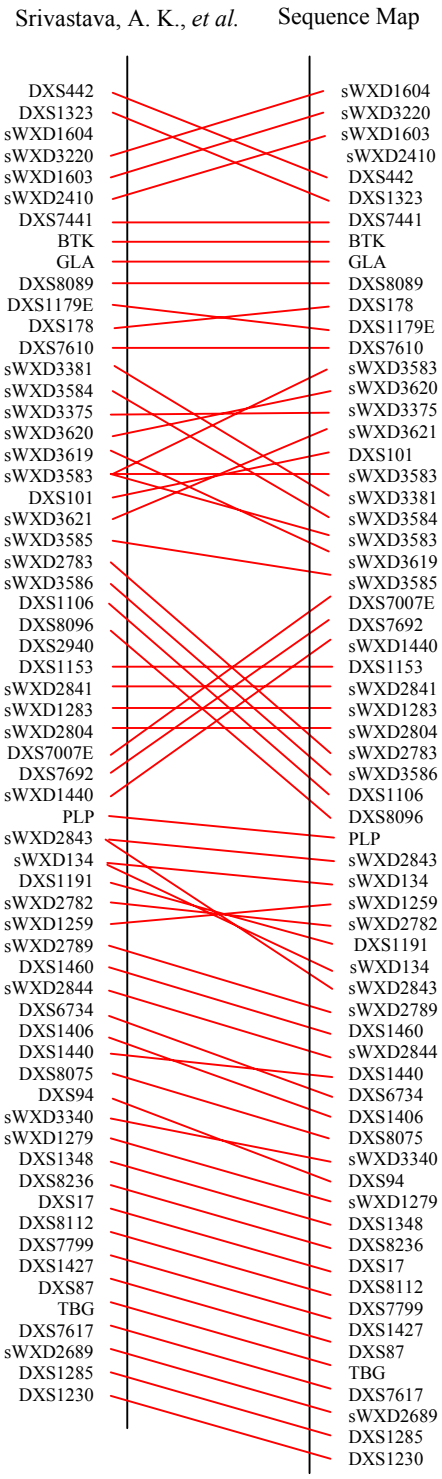


Figure 3.11: *Comparison of the YAC map. A comparison of STS order between the published STS-based YAC contig of Srivastava, A.K., et al (1999)(on the left) and the final sequence map (on the right). Red lines link the same marker.*

3.4 Sequence composition and repeat content analysis

The complete sequence provides the opportunity to analyse the base composition and the repeat of the region, and in some cases identifying specific sequences that generated conflicting data which required resolution before the completion of the sequence.

3.4.1 Sequence composition analysis

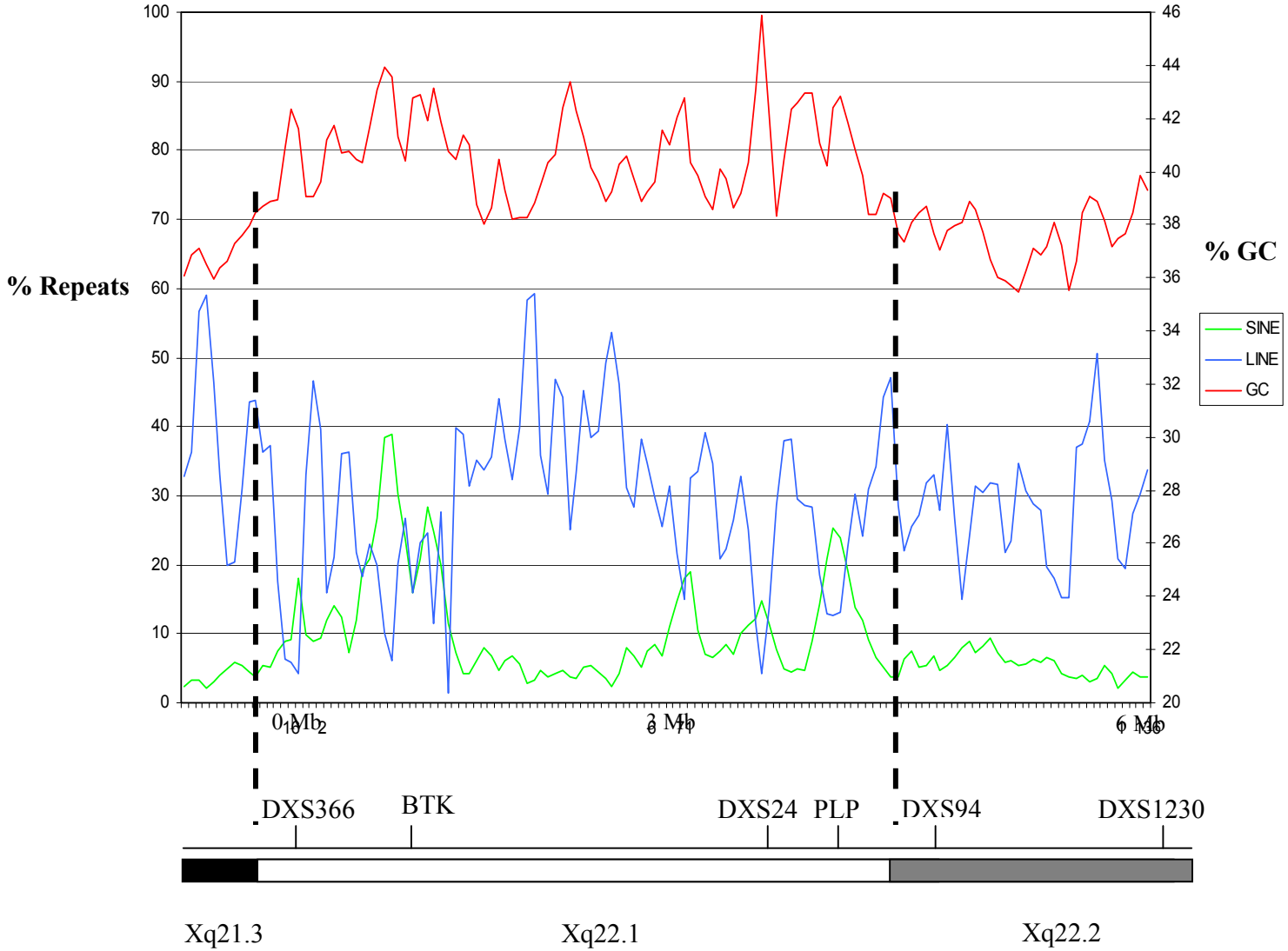
The boundaries of the contig constructed in this chapter are identified as being DXS366 at the proximal end and DXS1230 at the distal end. DXS366 was identified as a genetic marker (variable number of tandem repeats – VNTR) and originally placed broadly on the cytogenetic map between Xq21.2 and Xq24, based on a series of X chromosome translocation breakpoints (Dietz-Band, J. N.*, et al.*, 1990). It was more recently placed in proximal Xq22 by YAC mapping (Vetrie, D.*, et al.*, 1994). DXS1230 was also identified as a genetic marker (dinucleotide repeat) and had been localised to approximately 6 Mb distal to DXS366 by YAC mapping (Vetrie, D.*, et al.*, 1994). Cytogenetic bands are sized as a fraction of the total length of the chromosome. In the case of Xq22.1, it is estimated to be approximately 5 Mb given the X chromosome is 164 Mb in size and Xq22.1 is approximately 32 times smaller than the total length of the X chromosome. Based on these size estimates, this would place DXS1230 in Xq22.2.

The evidence for the localisation of the flanking markers DXS366 and DXS1230 would suggest that the contig described in this chapter spans part of Xq22.1 and part

of Xq22.2. Xq22.1 is a light band, whereas Xq22.2 is a dark band. Dark bands are associated with high AT (or low GC) due to the fact that certain chromosome stains such as DAPI bind AT rich regions preferentially (Schnedl, W.*, et al.*, 1977) and subsequently, light bands are associated with lower AT (or higher GC).

In order to analyse the sequence content in the region of interest between DXS366 and DXS1230, the available genomic sequence from approximately 650 kb upstream of DXS366 to DXS1230 was divided into 100 kb segments, overlapping by 50 kb. The GC content of each 100 kb segment was then analysed using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and each result plotted as single point on a linear scale. In order to identify any correlation between GC content and repeat content in the region the SINE content and LINE content of the same 100 kb segments of sequence were also analysed (see Figure 3.12).

Figure 3.12: *(see over) A graph showing the relative abundance of the GC content (red), LINES (in blue) and SINES (in green) across the region of interest. The scale for GC content is given on the right side of the graph, and the scale for SINE and LINE content is given on the left side of the graph. The position of markers previously placed on the cytogenetic map is also indicated.*


% Repeats
% GC
SINE
LINE
GC
100
90
80
70
60
50
40
30
20
10
0
46
44
42
40
38
36
34
32
30
28
26
24
22
20
0 Mb
3 Mb
6 Mb
DXS366
BTK
DXS24
PLP
DXS94
DXS1230
Xq21.3
Xq22.1
Xq22.2

The GC plot shows that for the region predicted to be within Xq22.1, the GC content remains above 38%, whereas the regions flanking Xq22.1, namely Xq21.3 and Xq22.2 (both dark bands) the GC content drops below 38%. This is consistent with the assumption that light bands are GC richer than dark bands. Xq22.1 is consistently higher than the genome average of 41% (figure taken from the analysis of the draft sequence (IHGSC, 2001). In general, Xq21.3 and Xq22.2 appear to be higher in LINE and lower in SINE, but the SINE and LINE content of Xq22.1 is much more variable.

3.4.2 Analysis of previously identified low copy repeats

It is well known that common repeats such as SINES are widely dispersed in the human genome. As discussed in Section 3.2, it had been reported that there were five copies of DXS101, a low copy repeat specific to Xq22. The five copies had been placed in three different loci, DXS101a, DXS101b and DXS101c (DXS101a and DXS101b each contain two copies of DXS101) (Vetrie, D*., et al.*, 1994). The sequence of each locus has not previously been determined and the probe used to identify the DXS101-positive cosmids was not available for this project. In order to identify the positions of DXS101 within the sequence, genomic sequences thought to include each copy of DXS101 (based on previous hybridisation to the available cosmids carried out by Elaine Kendall) were compared to each other by BLAST. Five regions of approximately 900 bp have been identified that appear to represent the previously reported DXS101 repeat. The results are summarised in Table 3.1.

Table 3.1: *Position of the DXS101 loci in the genomic sequence*

DXS101 Locus	Genomic sequence	Position in sequence	Size (bp)	Reported Restriction Fragment Size (kb)	Actual Restriction Fragment Size (kb)
DXS101a_1	dJ122O23	21992-22909	917	7.0	7.5
DXS101a_2	cV351F8	12407-13313	907	5.5	5.5
DXS101b_1	cV857G6	16018-16898	881	11.5	11.6
DXS101b_2	cV857G6	38408-39297	890	6.0	6.3
DXS101c	cU177E8	35575-36468	894	13.0	12.9

Each of the five sequences identified were located within *Eco*RI restriction fragments (see actual restriction fragment size in Table 3.1) that corresponds to those identified previously when the DXS101 probe was hybridised to *Eco*RI digested DXS101-positive YAC clones (Vetrie, D*., et al.*, 1994) (see reported restriction fragment size in Table 3.1). The five sequences were aligned using CLUSTALW and appear to cluster into two groups (data not shown). Group 1 contains DXS101a_1, DXS101b_1 and DXS101b_2 and are greater than 85 % identical to each other. The second group contains DXS101a_2 and DXS101c and are 80 % identical to each other. Within the 900 bp there is a region of approximately 100 bp that is greater than 90 % identical in all five sequences and would account for the ability to identify all loci by hybridisation with the DXS101 probe.

3.4.3 Analysis of previously unidentified low copy repeats

The sequence of the region allows for previously unidentified low copy repeats to be characterised. The 6 Mb of sequence was analysed for repeats using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) to remove previously

characterised common repeats and compared to itself by BLAST. All self-matches were removed and the remaining results viewed in ACT (see http://www.sanger.ac.uk/Software/ACT) (see Figure 3.13). The results are summarised in Table 3.2. There are six duplications greater than one kb in length which are greater than or equal to 99% percentage identical.

Table 3.2: *Low copy duplications between DXS366 and DXS1230*

Repeat	Type	Length (kb)	Identity	Starting Positions in sequence map (kb)
1	Inverted	5	99	1168 1187
2	Inverted	1	100	1411 1891
3	Direct	1	99	1412 1939
4	Inverted	140	99	1769 1920
5	Direct	18	99	3518 3536
6	Inverted	12	100	5804 5838

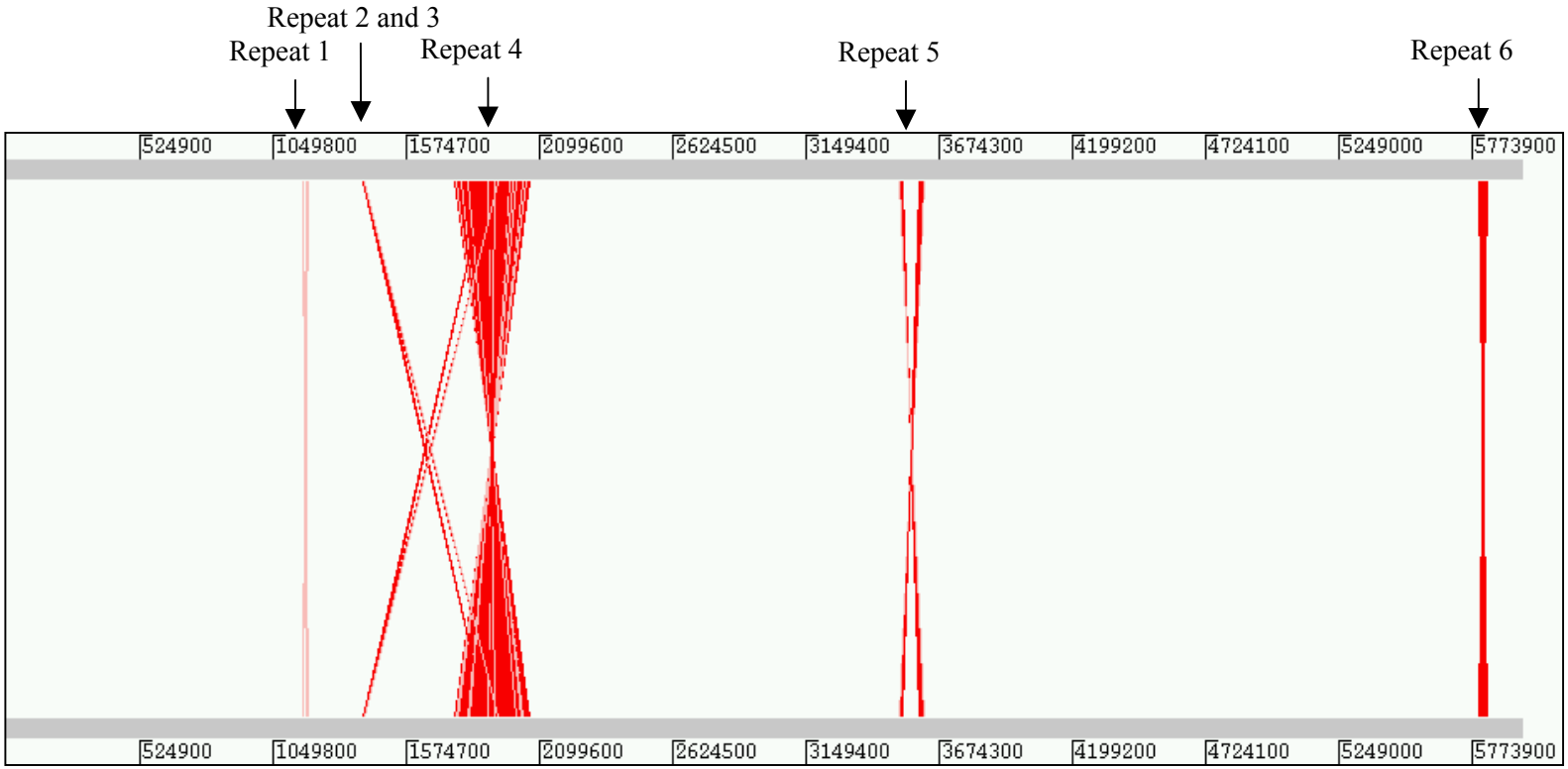
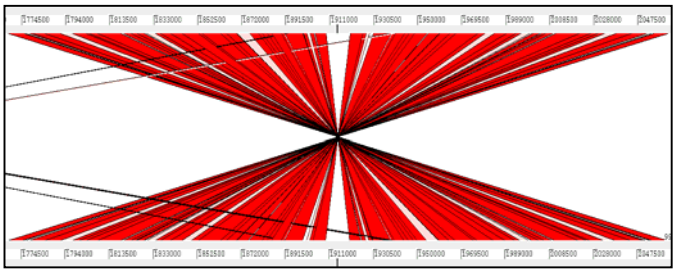


Figure 3.13 *An image from the computer program ACT, showing the position of low copy repeat sequences within the final sequence map. The grey numbers indicate a base pair scale. The region was compared to itself by BLAST, and red bar links similar sequences. The region is represented twice, along the top and along the bottom of the diagram, therefore each repeat is present twice. The threshold is set to show repeats greater than 1kb in length and greater than 99% identical.*

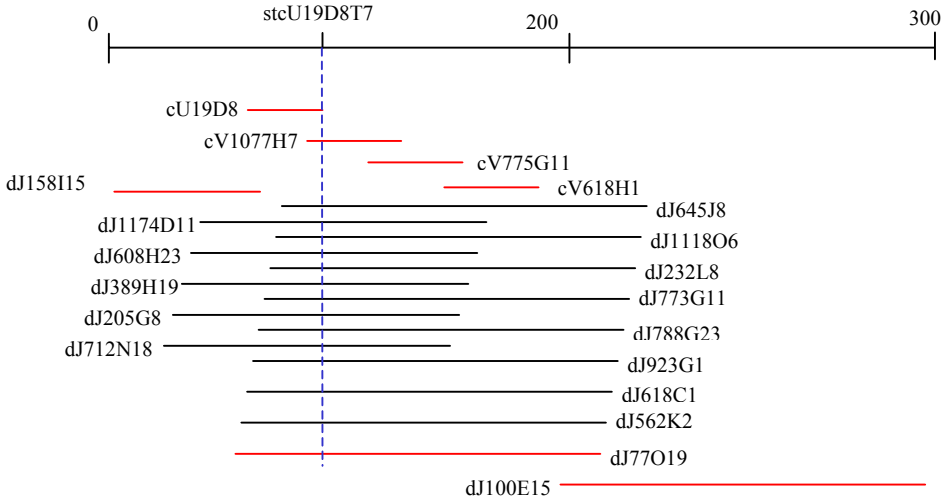
The largest repeat identified is a 140 kb inverted repeat separated by less than 10 kb (see Figure 3.14a). During the construction of the 6 Mb contig, clones mapping to this 300 kb region were thought to overlap based on fingerprinting and STS content (see Figure 3.14b). In particular an STS designed to the end of cU19D8 (stcU19D8T7) mapped to a series of PAC clones including dJ77O19. However, the sequence of dJ77O19 did not contain the sequences of cV1077H7, cV775G11 and cV618H1, but did contain a portion of sequence that matched part of the sequence of dJ158I15 and cU19D8. The regions of overlap were 100% identical but could not be a true overlap because the match was inverted. This left a gap in the contig between cV618H1 and dJ77O19 which was closed by the identification of a BAC contig constructed using stFSX2369, an X chromosome specific STS. The sequence of bA422L23 was found to contain the sequence of cV1077H7, cV775G11 and cV618H1. The sequence of bA353J17 closed the remaining gap between cV618H1 and dJ77O19 (see Figure 3.14c). An STS designed outside the region of duplication (stdJ77O19.1) confirmed the correct placement of the clones.

Figure 3.14: (see over) *Analysis of 140 kb indirect repeat (a) An enlargement of a section of Figure 3.13 showing the region of the sequence containing the repeat 4. (b) The contig as constructed before the identification of the duplicated repeat. An STS designed to the end of cU19D3 (stcU19D3T7) identified 14 clones whose overlap was confirmed by fingerprinting. The clones shown in red were identified for sequencing. (c) The final contig constructed using genomic sequence information confirmed by STS content. The clones for which genomic sequence was available are shown in red. Clones identified by the STS stFSX2369 were incorporated into the contig by sequence comparison. bA422L23 overlapped with four cosmids, cU19D8, cV1077H7, cV715H1 and cV618H1. bA353J17 overlapped bA422L23, cV618H1 and dJ77O19. Two positions for stFSX2369 were also identified. stdJ77O19.1 was designed outside the duplication and used to confirm the position of the remaining PACs.*

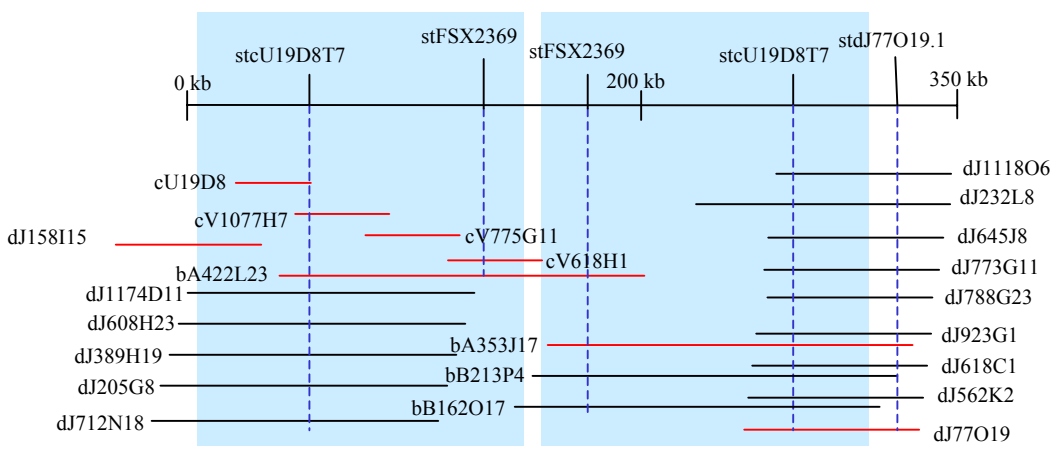
(a)



(b)


0
stcU19D8T7
200
300
cU19D8
cV1077H7
cV775G11
dJ158I15
cV618H1
dJ645J8
dJ1174D11
dJ1118O6
dJ608H23
dJ232L8
dJ389H19
dJ773G11
dJ205G8
dJ788G23
dJ712N18
dJ923G1
dJ618C1
dJ562K2
dJ77O19
dJ100E15

(c)


stcU19D8T7
stFSX2369
stFSX2369
stcU19D8T7
stdJ77O19.1
0 kb
200 kb
350 kb
cU19D8
cV1077H7
dJ158I15
cV775G11
cV618H1
bA422L23
dJ1174D11
dJ608H23
dJ389H19
dJ205G8
dJ712N18
bA353J17
bB213P4
bB162O17
dJ1118O6
dJ232L8
dJ645J8
dJ773G11
dJ788G23
dJ923G1
dJ618C1
dJ562K2
dJ77O19

3.4.4 Analysis of clone instability

The final gap to be closed in the sequence covered a region of DNA around DXS24 that had proven to be unstable in YACs (see Figure 3.15). The region containing DXS24 had previously been placed between DXS83 and DXS101C by genetic mapping. Data from the YAC map (Vetrie, D.*, et al.*, 1994) gave conflicting evidence to suggest either it mapped outside the region, or that YACs containing DXS83 and DXS101C were deleted for the region containing DXS24. A single YAC was identified with DXS24, and estimated to be approximately 50 kb. The region containing DXS24 was anchored to a region distal to DXS83 by bacterial clone mapping when cU105G4, a DXS24-positive cosmid, was shown to overlap with dJ79P11. The other end of dJ79P11 was mapped to a DXS83-containing contig. This evidence agreed with both the previous placement of DXS24 by genetic mapping, and the subsequent hypothesis by Vetrie, D., *et al*. (1994) that YACs positive for DXS83 and DXS101c were deleted for DXS24.

Linking DXS24 to DXS101C proved more difficult. The clone dJ823F3 was identified using STSs designed to the ends of both cU177E8 and dJ42120 (see Figure 3.15) and selected for genomic sequencing. However the sequence of dJ823F3 revealed that although it showed overlap with a portion of dJ79P11, no common sequence was found with cU105G4. The overlap with dJ79P11 was in the same orientation and fibre-fish was carried out (by Pawandeep Dhami, data not shown) to confirm a deletion in dJ823F3 rather than a direct repeat.

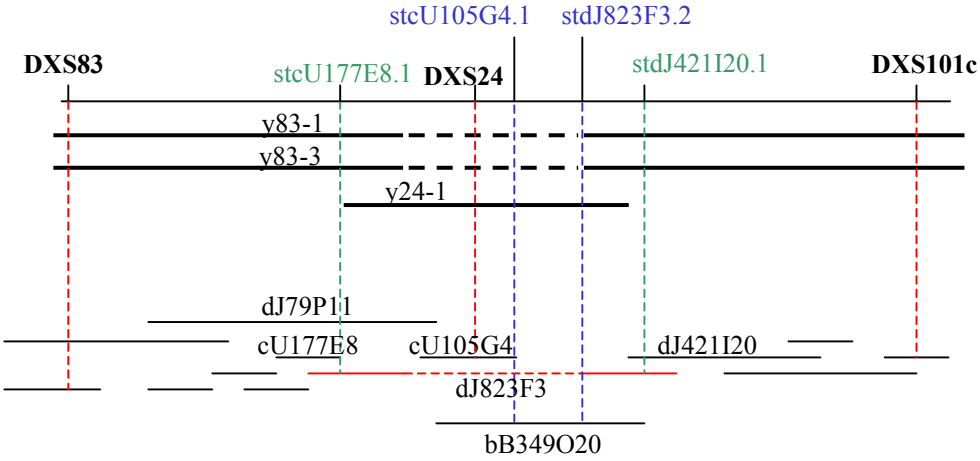


Figure 3.15: *Analysis of clone instability showing the region around DXS24 and the status of the mapping. The YAC map published by Vetrie, D., et al (1994) showed YAC clones, y83-1 and y83-2, were deleted for DXS24 (indicated by thick dotted lines), and only one YAC, y24-1, was only positive for DXS24, and no other surrounding markers. T he sequencing of dJ823F3 (shown in red), isolated with stcU177E8.1 and stdJ421I20.1 appeared to reveal a similar deletion (indicated by a thin dotted line) to that seen in the YACs. The sequencing of BAC bB349O20, identified with stcU105G4.1 and stdJ823F3.1, closed the gap in the sequence. Unlabelled thin black lines represent surrounding clones anchoring DXS24 to both DXS83 and DXS101c.*

The deletion is a similar phenomenon to that which had been seen in the YAC clones. Two further STSs (stcU105G4.1 and stdJ823F3.2) were used to identify bB349O20, which when sequenced, bridged the remaining sequence gap. Analysis of the region deleted in dJ823F3 revealed that the proximal and distal boundaries of the deletion were positioned within *Alu* repeats, which contain 12 bp of identical sequence.

3.5 Discussion

This chapter describes the construction of a 6 Mb sequence-ready bacterial clone map covering the region of Xq22 between DXS366 and DXS1230. The contig was built in four stages using the best resources available at the time for each stage (see Figure 3.16). Initial coverage across the region was gained using overlapping cosmid clones, which have insert sizes of approximately 40 kb. Although these have been used extensively in the past to map regions of the human genome and genomes of other organisms (Coulson, A.*, et al.*, 1988, Doggett, N. A.*, et al.*, 1995), larger insert bacterial clones are a more suitable reagent for sequencing as they enable a greater amount of sequence to be generated using a fewer number of clones. The PACs have an average insert size of 120 kb (Ioannou, P. A.*, et al.*, 1994) and the improvements in cloning in the latest BAC libraries mean these have an average insert size of about 180 kb (Shizuya, H.*, et al.*, 1992). Therefore the later stages of the construction of the contig reflect these improvements. Early walking between cosmid contigs identified PACs, but the later gaps were closed using BACs.

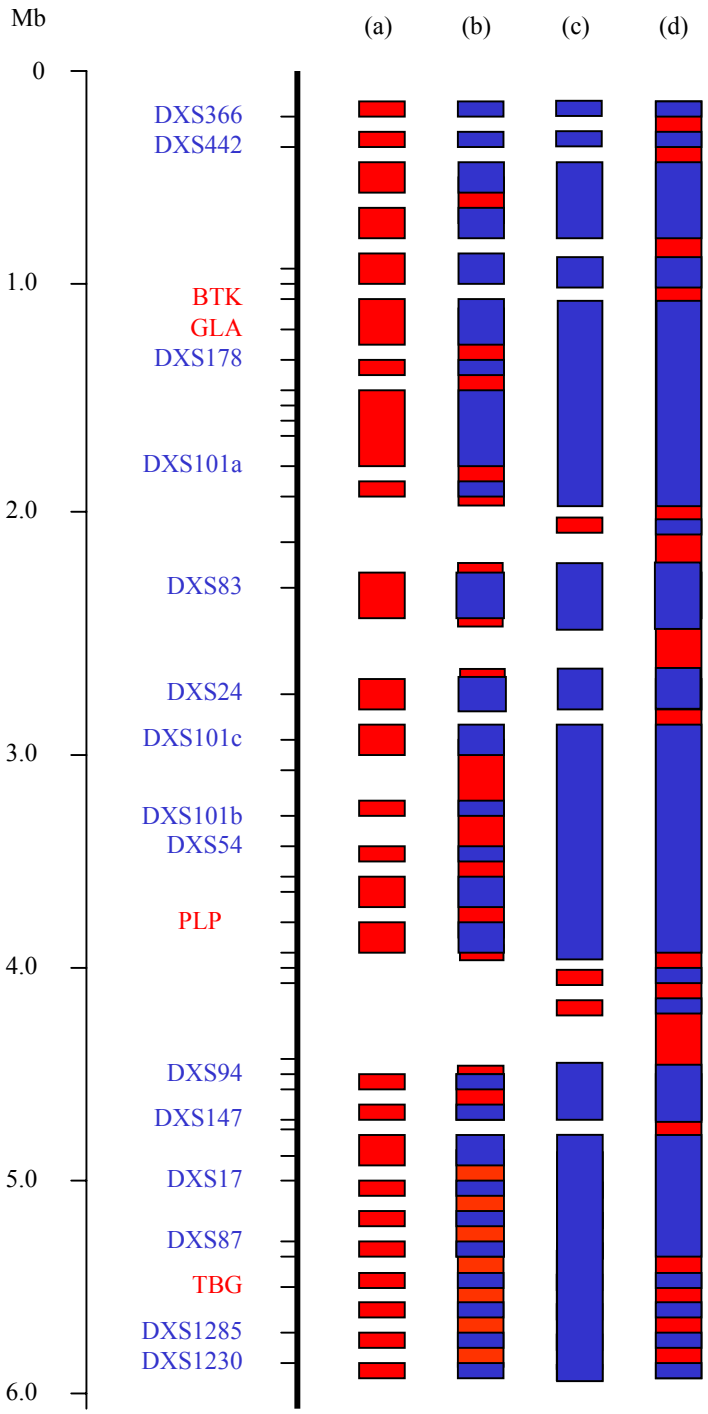


Figure 3.16: *The status of the mapping at each stage of contig construction. Bars represent the contigs; new coverage gained at each stage is shown as red bars and existing coverage is shown as blue bars. The four stages are (a) cosmid fingerprinting, (b) PAC identification using whole cosmid hybridisation, (c) seeding contigs in gaps and (d) final gap closure.*

The bacterial clone maps covering large portions of the human genome (Bentley, D. R.*, et al.*, 2001; Bruls, T.*, et al.*, 2001) have been constructed using the equivalent of the last two stages of this project. Initial coverage was gained using a high density of markers across a given region to identify bacterial clones and gap filling was achieved with STSs from the ends of clones at the ends of contigs. A density of markers greater than fifteen per megabase has proved sufficient to cover 80% of large regions, eg, whole chromosomes, in sequence-ready contigs (Bentley, D. R.*, et al.*, 2001). If the contig described in this chapter was generated today, the start point would be the identification of BAC clones using the available STS landmarks as probes (see Figure 3.16c and d).

The integration of cosmids with PACs and BACs to form a single map had two major problems. The first is the difficulty of identifying significant overlaps between a cosmid and a larger insert clone such as a PAC or BAC using fingerprinting. FPC reports overlaps based on the probability of two clones overlapping by chance, the lower the score, the less likely the match is a random event (see Section 2.23.2). In calculating the probability score, the number of bands in common between the two clones is taken into account relative to total number of bands in each clone fingerprint. Probability of overlap is then assessed on this basis. Table 3.3 shows a set of FPC probability scores between a series of clones in the final contig.

Table 3.3: *Example of probability of overlaps, comparing clones of different sizes*

Clone 1	Number of Bands	Clone 2	Number of Bands	Matches	Probability	Actual Overlap
cV467E10	20	cU46H11	25	11	2e-05	30 kb
cV362H12	18	dJ839M11	35	10	6e-04	25 kb
dJ3E10	44	dJ197J16	55	21	1e-10	25 kb
bA269L6	39	dJ409F10	28	14	7e-04	0 kb

The results show that even though the cosmid cV362H12 and the PAC dJ839M11 overlap by 25 kb, the reported probability is the same as that reported between two non-overlapping large insert clones bA269L6 and dJ409F10. In this way, a significant overlap between a cosmid and PAC or BAC could be missed. In general a threshold of probability can be set when contigs are constructed using similarly sized clones e.g. 1e-04 for cosmids and 1e-10 for PACs/BACs.

The second major problem with constructing a contig using clones of different lengths is identifying the minimum set for sequencing. After the initial assembly of cosmid clones, a minimum set of clones from each contig was identified and sequenced. The larger insert PAC or BAC clones were added at the ends of contigs to either close gaps or extend coverage into the gaps. In most cases, when a PAC or BAC clone was sequenced, at least one cosmid at the end of each contig became redundant. A more efficient procedure for selecting a minimum set of clones for sequencing is to generate a complete contig in similarly sized clones and then choose the clones for sequencing.

The availability of the genomic sequence between DXS366 and DXS1230 allowed for a comparison of previously published maps and an assessment of the accuracy of the different types of maps. The physical maps that were compared in Section 3.3 were generated as part of the overall aim to map, sequence and analyse large regions of the human genome and have proven vital in generating the final bacterial clone contig map described in this chapter. Although differences in marker order were observed between the published maps, these can be accounted for by the limitations of each method used. Genetic mapping and radiation hybrid mapping rely on

recombination and radiation-induced DNA breakage respectively. These events do not necessarily occur randomly with respect to physical distance. The refinement of both maps and the accurate positioning of the markers on the genomic sequence is important for future study of the region. The refinement of the genetic map and the positioning of the genetic markers on the sequence are important as these markers are still being used to define critical regions for as yet uncloned diseases. The genomic sequence will also allow for the identification of new sequences that may be useful for genetic mapping such as previously unidentified dinucleotide repeats (e.g CA) that may be polymorphic in the population. The RH map contains STS generated from EST sequences and the placement of these STSs on the genomic sequence will aid the identification of the genes within the sequence.

The genomic sequence generated from clones selected from the bacterial clone contig described in this chapter provides the basis for a higher resolution analysis of the region than has previously been possible. One of the genes contained within the genomic sequence is PLP, and duplication of the region including the PLP gene is the primary cause of Pelizaeus Merchbacher Disease (PMD; Hudson, L. D.*, et al.*, 1989; Trofatter, J. A.*, et al.*, 1989), causing a gene dosage effect that is thought to lead to increased expression of the protein product and a disturbance of development or maintenance of myelin (Inoue, K.*, et al.*, 1999). Work is currently underway with Karen Woodward (Institute of Child Health, London) to map the breakpoints of these duplications in different PMD families onto the sequence map, and to determine whether features within the sequence make the region susceptible to duplication.

A comprehensive transcript map that includes the region between DXS366 and DXS1230 is also being constructed using the available genomic sequence (Ian Barrett, The Sanger Institute). A systematic scanning of the available genomic sequence, using a combination of *de novo* gene prediction and similarity searches are placing previously known but unlocalised genes and identifying novel genes that are experimentally confirmed. A number of genetic disorders, for which the gene responsible has been localised to the region, remain uncloned and the sequence and ultimately the genes in the region will allow for the systematic screening for the genes responsible.