Physical Mapping of the Human X Chromosome

By

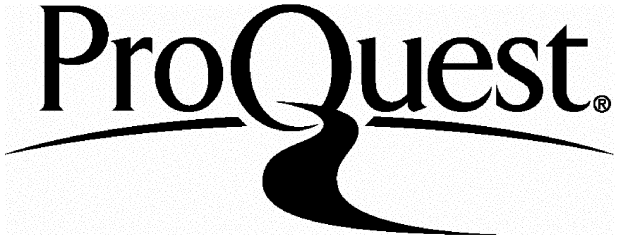
Hugues Roest Crollius

a thesis submitted for the degree of Doctor of Philosophy in the University of London

Department of Biology,
University College London,
London.

Genome Analysis Laboratory,
imperial Cancer Research Fund,
Lincoln's Inn Fields,
London
U.K.

Max-Planck-Institut für Molekulare Genetik
Ihnestrasse 73,
14195 Berlin
Germany



Acknowledgements

I would like to thank Hans Lehrach for giving me the opportunity to work on such an interesting and ambitious project. I am very grateful to Richard Mott for his patience in guiding me through my first steps in UNIX, to Mark Ross, Leo Schalkwyk and Marie-Laure Yaspo for their advice in many aspects of this work, to Igor Ivanov for his guidance in the DNA binding experiments, and to Karl Rak and Kieran Todd for their help in hybridisation experiments.

I am also especially grateful to Catherine Nguyen for her generosity and her energy, to Geneviève de Saint Basile for her trust and to Ulf Leser for his invaluable contribution to the X chromosome project.

My special thanks go to Fiona Francis for all her advice and support throughout the past four years, her patience and encouragement during the writing of this thesis and for devoting many hours to proofreading it.

This thesis is dedicated to my parents,

My mother and my late father,

Jan Roest Crollius

Abstract

The genetic analysis of the human X chromosome has evolved considerably since the beginning of the Human Genome Project. The involvement of the X chromosome in sex determination and its particular inheritance pattern are in a large part responsible for the strong interest that has motivated this progress.

This thesis describes the construction of a physical map of the X chromosome in YAC clones, important as a prerequisite to a deeper analysis of its gene content. As part of this work, some technological developments were investigated with the aim of accelerating the rate of data generation. Different hybridisation systems were assessed with radioactive and non-radioacttive labels, and with different supports for binding the target DNA. A new method was implemented to determine or confirm overlaps between YAC clones, which made use of a library of X chromosome Alu-PCR products. Ultimately a large and heterogeneous set of experimental results was analysed manually and with the assistance of computer software, resulting in a YAC contig map covering the majority of the X chromosome.

As part of this project, and in the wider context of the construction of a reference YAC collection for the X chromosome community, a database was developed to handle and distribute the information. The first implementation of this Integrated X chromosome Database (IXDB) was performed using the ACEDB software, and was mostly used internally to assist in the YAC map construction. The first version of the YAC map was publicly released in this system. In a second phase, IXDB was transferred to the ORACLE relational system to provide a more sophisticated and comfortable interface to the World Wide Web and a more robust data management system.

Table of Contents

CHAPTER ONE: Introduction

1. Physical Mapping of the Human Genome

1.1 The Human Genome Project

For several decades after the discovery of the structure of the deoxyribonucleic acid (DNA) molecule (Watson and Crick, 1953), a major area devoted to its study was concerned with the question as to how such a huge molecule could fit into such small cells. Gradually however, biologists realised that the information carried by the molecule was as important, if not more, than the molecule itself. Hints to this new concept came from the study of bacterial genetics (for instance (Jacob and Monod, 1961)) and gradually contributed to the emergence of a new discipline of biology, called molecular genetics. A combination of classical mendelian genetics and molecular biology, the field of molecular genetics has rapidly evolved to become one of the most influential sciences of the 20th century, both economically and socially. The reason for this prominent role is certainly due to its focus on the human genome, which considerably influences the state of our health, and has far reaching significance for human evolution and diversity. In 1988, the Human Genome Organisation (HUGO) was created to provide a support and coordinate various undertakings that became known as the Human Genome Project (McKusick, 1989). This project has two main motivations. The first, of major economical importance and consequence for human health, is the discovery of genes responsible for genetic disorders. In these diseases, whether mild or severe, the phenotype is often difficult or impossible to correct. Therefore the discovery of the molecular basis of diseases at the DNA level is often the only way leading to a treatment, or to the possibility of early diagnosis. The second motivation has its roots in scientific stimulus, the quest for knowledge, the understanding of how the information coded by the estimated 65.000 genes in the human genome (Fields et al., 1994) results in the making and functioning of a human being. The five year goals implemented by the NIH and DoE in the United States in 1990 and revised in 1993 (Collins and Galas, 1993) reflect the priorities that were set world wide in the framework of the human genome project. Of the eleven areas of genome research that were elaborated in the five year goals, physical mapping of the genome with a completed Sequence Tagged Site (STS) map at a resolution of 100 kb is certainly one of the most difficult and ambitious tasks. It illustrates the importance of high resolution physical maps, as a necessary complement to genetic maps. One of the essential tasks of HUGO has been the organisation of single chromosome workshops, which each year assemble scientists concerned with the genetic analysis

of a given chromosome. The output of these meetings is the establishment of consensus genetic and physical maps, which take into account experimental data from all participants. The need for such maps is obvious when considering positional cloning strategies for isolating disease genes. So far over 60 genes responsible for genetic disorders have been identified using this approach (http://www3.ncbi.nlm.nih.gov/Omim/searchomim.html). One of the most time consuming phases of a positional cloning project is the construction of a physical map of the region under investigation. The availability of physical maps of the genome in cloned DNA will therefore dramatically accelerate the isolation of genes. In addition, placing genetic markers on the physical maps confirms their order which, on the genetic map, is derived from statistical analysis of recombination events. Today, as the emphasis of the human genome project is moving towards the elucidation of the complete nucleotide sequence, the construction of robust physical maps has gained an increasing importance.

1.2 Genetic and physical maps

The first human genetic mapping studies began with the successful identification of X chromosome linkage for color blindness and haemophilia (Bell and Haldane, 1937). Routine linkage analysis was however limited by the small number of progeny in human populations, and by the outbred nature of human matings. The lod score (logarithm of the odds ratio for linkage) method overcame these difficulties and made the analysis of human linkage data practical (Morton, 1955). In the 1970's, existing statistical algorithms were implemented in efficient computer based analysis (Ott, 1974) and facilitated further the construction of genetic maps. However, linkage analysis approaches to human disorders was still limited by marker availability.

DNA-based polymorphisms allowed for a sudden expansion in the number of available markers. First, restriction fragment length polymorphisms (RFLPs) were proposed as a source of such DNA markers (Botstein et al., 1980), but suffered from low heterozygosity, which limited their use in families. The application of findings that variable number of tandem repeats (VNTR) (Jeffreys et al., 1985; Nakamura et al., 1987) and microsatellites of dinucleotide repeats (Weber and May, 1989) were highly polymorphic provided markers of higher information content. The first genome-wide human linkage map (Donis-Keller et al., 1987) made use of a pool of 61 families provided by the Centre d'Etudes du Polymorphisme Humain (CEPH). The latest genetic map of the human genome was constructed by the Généthon group and contains 5264 markers, providing a density of one marker every 1,6 cM (Dib et al., 1996). This map has recently been used as a framework for a first generation gene map of the human genome based on EST mapping using irradiation hybrid panels (Schuler et al., 1996). This data will greatly facilitate the identification of disease genes

by providing a set of candidate genes within relatively small and well defined regions. The usefulness of such transcript maps will be enhanced by the integration of the genetic positions with clone based physical maps, which allow a more direct way of studying the gene structure.

Physical maps can be constructed in the form of cloned contigs, long range restriction or cytogenetic maps. Contigs are contiguous, overlapping cloned DNA fragments arranged in the same order as they are found in the genome. Initial attempts to construct contigs were based on cosmids and were pioneered in studies on *Caenorhabditis elegans* (Coulson et al., 1986). After the advent of YAC cloning techniques (Burke et al., 1987), YAC clone contigs became the favoured method of constructing a first physical map in a region of interest, as in most positional cloning projects. In parallel to region specific map construction, several projects aim at a genome wide physical map, and have followed the proposition that Sequence Tag Sites (STS) be used as a common language for integrating different types of maps (Olson et al., 1989). This has resulted in YAC contig maps based on STSs that were ordered using genetic mapping (Chumakov et al., 1995) or using irradiation hybrid mapping (Hudson et al., 1995).

1.3 A series of technological breakthroughs in physical mapping

1.2.1 Molecular genetics

1.2.1.1 Pulse Field Gel Electrophoresis (PFGE)

Pulse Field Gel Electrophoresis (PFGE, Schwartz and Cantor, 1984) has been essential in the physical mapping of large chromosomal regions of the human genome. It is a technique that allows the separation of large DNA fragments of up to 9 Mb long by periodically alternating the electric field applied to an agarose gel. Long range restriction maps can be obtained by digesting genomic DNA with rare cutting enzymes, separating the fragments by PFGE, transferring the DNA to a filter membrane and hybridising with markers from the region of interest. This has been instrumental for example in the discovery of the cystic fibrosis gene (Riordan et al., 1989) and the Duchenne Muscular Dystrophy gene (Monaco et al., 1986). YAC clone inserts are typically between 100 and 1700 kb and therefore necessitate the use of PFGE for separating them from the natural yeast chromosome. This is essential when constructing YAC contigs in order to determine the size of the inserts, the presence of several clones within one yeast cell, or the occurrence of internal rearrangements after successive growth cycles.

1.2.1.2 Yeast Artificial Chromosome (YAC) cloning

The use of YACs for propagating DNA fragments several hundred kilobases in length was first described in 1987 (Burke et al., 1987). This new technology held two major promises. First, the ability to clone segments of DNA over a megabase in size, a capacity 25 times that of cosmids, represented a potential breakthrough in genome analysis and contributed to support proposals to map and sequence the human genome. For mammalian genetics, YAC cloning bridged the gap between genetic maps derived from linkage analysis, which have a resolution of about 2-10 Mb, and physical maps derived from fragments cloned in *E. coli* based vectors which have insert sizes in the 0.001-0.05 Mb range. It was also possible to envisage that YACs would allow the cloning of genomic regions which were so far refractory to cloning in *E. coli*, by providing a more suitable environment in an eukaryotic host. In addition, it was suggested that since YAC clone size could easily reach the megabase range, any gene or gene complex could be cloned along with its regulatory regions in a single microbial host, as a contiguous piece of DNA. This opened up new promises for the study of gene function and regulation.

Today, nearly ten years after the advent of YAC cloning, it is clear that most promises have been kept. YAC libraries have been constructed for many genomes including mouse (Larin et al., 1991), chicken (Toye et al., in press), *C. elegans, D. melanogaster*, rat, yeast and human (Larin et al., 1991), (Chumakov et al., 1995) (Anand et al., 1989). Two maps of the human genome in YAC clones have been reported so far (Hudson et al., 1995), (Chumakov et al., 1995). These are not complete by any means, but are supported by maps developed separately for chromosome Y (Foote et al., 1992), 21 (Chumakov et al., 1992), 22 (Collins et al., 1995) , 3 (Gemmill et al., 1995), 12 (Krauter et al., 1995), 16 (Doggett et al., 1995) and X (This thesis work and (Roest Crollius et al., 1996)). YACs have been attractive vectors for functional gene studies after transfer from yeast to mammalian cells. (Huxley et al., 1991). YACs have also been used to generate transgenic mice to complement mouse mutations (Forget, 1993) and to study X chromosome inactivation (Lee et al., 1996).

1.2.1.3 Radiation hybrids (RHs)

Radiation hybrids have made a large contribution to genetic and physical mapping projects. RHs are produced by subjecting donor cells (usually diploid human or single human chromosome on a rodent background) to lethal doses of irradiation, and then recovering the fragmented chromosomes by fusion to a recipient rodent cell line (Goss and Harris, 1975). Assaying marker content (STSs) or hybridisation probes in a panel of radiation hybrid lines can be used as a method for ordering DNA markers (radiation hybrid mapping, (Cox et al., 1990)), offering a complementary approach to classical genetic linkage. Hybrids are analysed for the presence or absence of specific

markers, and statistical methods are applied to estimate the frequency of breakage between them, hence an order can be determined. For long range mapping purposes, relatively low doses of irradiation are applied in an attempt to retain large fragments. In recent years, human-rodent hybrids containing single human chromosomes have been the favoured DNA source for mapping purposes. However it is now feasible to create high-resolution whole-genome radiation hybrid maps (Walter et al., 1994) by reverting to the original protocols of irradiating diploid human cells instead.

1.2.1.4 Fluorescent in situ hybridisation (FISH)

FISH is a sensitive method for localising probes to different chromosomes (Lichter et al., 1990). It enables the assignment of probes to a chromosome band (approx. 5-10 Mb resolution) when performed on metaphase spreads, or at higher resolution (down to a few kb) when the target DNA is interphase chromatin (Lawrence et al., 1988), reviewed in Houseal and Klinger, 1994). It has found many applications in the diagnosis of human diseases caused by chromosome rearrangements or aberrations (Pinkel et al., 1988) (Tocharoentanaphol et al., 1994), but it has also significantly contributed to the field of genome mapping. It is a rapid method for determining the human content of hybrid cell lines (Lichter et al., 1990) and can be used for localising cDNA or genomic clones. The latter application has proved especially useful with the increased used of YAC libraries in genome mapping. YAC clones are often chimeric (between 20 and 50 %) and FISH mapping is the easiest way to determine the chromosomal origins of each chimera. Using different dyes in the same hybridisation, clones such as cosmids or YACs have been mapped relative to each other on interphase chromatin (Haaf and Ward, 1994; Trask et al., 1991) (Senger et al., 1994). FISH is often used as the ultimate proof of a clone's integrity and location, because it allows the direct visualisation of the chromosomal position. It is however a technique difficult to automate since it requires human interpretation in each hybridisation. Attempts have been made in this direction however, with the development of methods to hybridise up to 24 different probes on a single microscope slide (Larin et al., 1994).

1.2.1.5 Polymerase Chain Reaction (PCR)

The PCR technology (Saiki et al., 1988) has almost certainly had an impact on every mapping project since its discovery, and might rightly be said to be revolutionary for genome analysis. For instance, it became possible to use PCR as a more rapid way of typing individuals in the construction of genetic maps based on Restriction Fragment Length Polymorphism (RFLP). In this case, the region surrounding a restriction site can be amplified from genomic DNA and digested with the enzyme that recognises the polymorphic site. Insertion, deletion, and variable-number tandem

repeats can also be detected by PCR. Short tandem repeat (microsatellite), composed of short stretches of di- tri, or tetra nucleotide repeats have now become a standard source of genetic markers for constructing genetic maps of the human genome (Dib et al., 1996; Gyapay et al., 1994; Weissenbach et al., 1992) and of the mouse genome (Dietrich et al., 1996; Dietrich et al., 1994) genomes. The PCR method has also greatly facilitated mutation detection in studies on genetic variation or in the search for specific sequence modifications associated with a change in phenotype. In particular, by allowing the same DNA segment from any individual to be isolated, it avoids the need to construct whole genomic libraries from that individual followed by the identification of clones with the mutation.

The possibility of cloning large fragments of DNA using the YAC system has made the construction of YAC contigs a favoured approach when initiating the physical characterisation of a particular region of the genome. YAC contig construction has been greatly facilitated by PCR, and more so when the physical map is integrated with genetic markers that can be assayed by this method. The identification of specific YAC clones from a complex library is possible using PCR to scan pools of clones in order to reduce them to a smaller number that can more easily be assayed by hybridisation (Green and Olson, 1990).

Primers can also be designed from sequences that occur more or less frequently in a given genome. Generally referred to as Interspersed Repeat Sequence PCR (IRS-PCR) this method generates a pool of products of varying complexity, that can be used as hybridisation probes, as templates for hybridisation, or as a fingerprint when separated into distinct bands by gel electrophoresis. This technique was first developed to amplify human DNA from rodent cell backgrounds in somatic cell hybrids (Nelson et al., 1989). In this case, primers were directed to the Alu repeat, which occurs on average every 4 kb in the human genome (Hwu et al., 1986). The method has found a wide range of applications in human genome mapping. These include the generation of fingerprints from radiation hybrids or YAC clones, enabling the detection of regions commonly retained in the former, or overlapping in the latter (Coffey et al., 1996). In addition, Alu-PCR has provided a rapid method to isolate a representative part of the YAC insert in useful quantities from crude lysate. This is important since YAC clones often co-migrate with natural yeast chromosome when separated by PFGE, preventing their purification. In the mouse, IRS-PCR utilises the B1 element and provides a similar method for generating probes from complex sources.

Further applications of PCR in clone mapping were developed that allow the isolation of YAC clone ends, essentially with the plasmid end-rescue (reviewed in Bates, 1996), inverse PCR (reviewed in Silverman, 1996), vectorette (Riley et al., 1990) and splint (Roux and Dhanaragan, 1990) methods. PCR is a technique that is well suited to automation, and several groups have developed large capacity thermocyclers that allow several hundred thousand reactions to be performed per day

(Hudson et al., 1995; Meier-Ewert et al., 1993). This automated technology was essentially developed in the context of whole genome studies.

The next phase of the human genome project is focused on genomic sequencing on a chromosome scale. Here also, the PCR method will play a major role: most sequencing projects now use a PCR based sequencing method (cycle sequencing, Carothers et al., 1989) and increasingly prepare template by PCR.

1.2.1.6 Genomic sequencing

Modern sequencing technology is based on the pionering work by F. Sanger (Sanger et al., 1977) who introduced the technique of dideoxy chain termination that is most often used today. At first limited due to the difficulties of obtaining single strand DNA and good primers, this technique became more widely applicable with the development of M13 vectors and oligonucleotide synthesisers. In order to automate certain areas of the sequencing process, such as gel electrophoresis, raw data acquisition and base calling, several models of automated sequencers were developed (e.g Ansorge et al., 1987; Smith et al., 1986). The chemistry of the sequencing reaction has evolved as well and now mostly relies on fluorescent labels either at the 5' end (dye primer) or at the 3' end (dye terminator) of the DNA molecules. This type of automated technology accounts for most of the results produced by large scale sequencing projects so far. Three bacterial genomes have been sequenced, *Haemophilus influenzae* (Fleischmann et al., 1995), *Mycoplasma genitalium* (Fraser et al., 1995) and *Methanococcus jannaschii* (Bult et al., 1996) and have provided a new approach to analysing genome information. Recently the sequence of the 12 Mb yeast *Saccharomyces cerevisiae* genome has been completed (Goffeau et al., 1996). It is the largest genome sequenced so far, and the first eukaryotic one. Out of the 5,885 predicted genes that this project uncovered, only 50% of the corresponding proteins could be ascribed a potential function. This demonstrates that although computational predictions provide guidance for designing experiments, the knowledge of complete gene sequences does not obviate from the need to carry out real experiments to determine protein function. Other large scale genomic sequencing projects are well underway, such as that of the fly *Drosophila melanogaster* and the worm *C. elegans.*

Sequencing the entire human genome has always been the ultimate purpose of the Human Genome Project, and this phase is now beginning. As a consequence, HUGO has recently set up a World Wide Web site that will monitor the amount of DNA that is sequenced, committed or projected for sequencing by the genome mapping community (Human Sequence Map Index http://hugo.gdb.org/hsmindex.htm). Approximately 1% of the human genome has been sequenced so far (November 1996, http://weber.u.washington.edu/~roach/human_genome_progress2.html) corresponding to approximately 30 Mb, about 8 Mb of which is on the X chromosome. This project currently raises important issues on the suitability of peer reviewing the

sequence data together with annotations that place the results in a broader biological context, compared to immediate release of assembled sequences (Adams and Venter, 1996; Bentley, 1996). The reason for this debate is due to the huge potential associated with human DNA sequence, especially for pharmaceutical companies with interests in new therapies for human diseases.

1.2.1.7 Mapping of Expressed Sequence Tags (ESTs)

The purpose of completly sequencing the human genome is to identify all the genes that it contains, and the generation of ESTs and their mapping should greatly facilitate this task (Adams et al., 1991). ESTs are short sequences, typically 300-400 base pairs long, determined from cDNA clones. Most 3' untranslated regions (3'UTR) of cDNA clones do not span introns and are therefore generally used for selecting primers, since they are more likely to give a product by PCR amplification from genomic DNA. In addition, 3'UTRs display less sequence conservation than do coding regions, making it easier to discriminate among gene family members that are very similar (Schuler et al., 1996; Wilcox et al., 1991). Over 600.000 EST sequences are currently available in DNA databases (75% of which are human) essentialy resulting from two large scale projects to generate very large numbers of ESTs from a variety of human cDNA libraries (Adams et al., 1995) (Hillier et al., 1996). Recently a consortium of laboratories has mapped over 16,000 ESTs on the human genome using radiation hybrid panels, and integrated it with the human genetic map (Schuler et al., 1996). This first generation transcript map of the human genome is an important resource for sequencing projects, as well as for positional cloning efforts. It also allows for the first time a global view to be taken of the gene distribution in the genome, and maybe for one to begin asking genome-wide questions on their biological function (Lander, 1996).

1.2.2 Bio-informatics

1.2.2.1 Computer aided data analysis

One impressive aspect of the Human Genome Project is the massive volume of information that has been generated since it started in 1988. In laboratories that are concerned with large scale genome analysis, the speed at which data can be generated has rapidly overtaken the capacity of human brains to process it, and has led to the development of computer software to take over the most demanding parts. Computer aided data analysis can be seen from two different angles. First, large scale experiments that generate huge amounts of results of identical nature, such as PCRs, hybridisations or gel fingerprints, need one type of algorithm to analyse the information over and over again, until a critical mass of data allows a sensible result to be obtained. This can be a contig map or a genetic map for instance. Such software

include for example the probeorder package (Mott et al., 1993) to build contig maps from hybridisation data, rhmapper to construct an STS map of the human genome (Hudson et al., 1995), or contigC for analysing clone fingerprints at the Sanger Centre (Coffey et al., 1996). Second, the volume and complexity of the information that is being accumulated in dedicated databases will make its interpretation impossible without powerful computer programs. Asking simple questions to a single database such as 'where is this clone on the map' or 'how many exons are in this gene' will always be possible and does not require much computer power. However, the purpose of the human genome project is to answer fundamental questions on gene function and regulation, their interactions and their effect on human development and health. The answers to such questions will require that terabytes of information be processed by 'intelligent' algorithms to deduce for instance that a given phenotype is most likely to be due to an alteration of a particular metabolic pathway involving a specific set of genes, one of which maps to the region segregating with the phenotype in an affected family. Another difficult problem is the dissection of polygenic traits, where the dysfunction of more than one gene in different parts of the genome is responsible for the disease (Lander and Schork, 1994; Weeks and Lathrop, 1995).

The human genome project is approaching this analysis phase, although many obstacles are already perceptible. The most critical one perhaps is the heterogeneous way in which data is stored, due to the existence of multiple databases with often completely diverging semantics.

1.2.2.2 A network of databases

The January issue of Nucleic Acid Research 1996 describes 58 different databases more or less related to human genetics and biochemistry. A number of databases were not reviewed, and new ones have appeared since. The abundance of such repositories for genes, clones, electrophoresis gels, proteins or mutations reflects the need to compile the vast amount of data in central electronic catalogues accessible publicly. It has been greatly facilitated by the widespread use of the World Wide Web in molecular biology (for a review see Harper, 1995). Most databases are now accessible in this way, superceding ftp, e-mails and gopher servers. The reason for the diversity of databases, and the absence of a few but exhaustive central databases, may be due to the fact that most are initiated through the need of a few specialists in a given field, for a uniform view of integrated data. In this context, it is often easier and faster to develop a new database rather than convincing an existing one to enlarge its focus to include the new theme. Since most databases are independent, each tends to describe the information in its own way, and this inevitably leads to heterogeneous formats, accession mode, and nomenclatures. One can distinguish three categories of approaches that attempt to solve this problem. First, by reformatting locally all the information in a single system (data warehouse);

second, by interconnecting databases via a common language (database federation) or third, by providing links to other databases in the WWW. The first strategy was for instance adopted by the Integrated Genome Database (IGD) group led by Otto Ritter at the DKFZ in Heidelberg. The main focus was on defining a unifying model of genomic data and related information, and applying it to a sample of existing databases. The system chosen for storing and re-distributing the 'integrated' information was ACEDB, partly because of its configurability, powerful graphical interface and ease of use. This work was supported by a close collaboration with the authors of the program (J. Thierry-Mieg and R. Durbin). The strategy was based on automatically downloading information from a series of genome related databases, locally reformating the data to a single format (ACEDB) and redistributing it via ftp. The project has now ceased to release data. Difficulties were essentially encountered with manipulating hundreds of gigabytes of data in ACEDB, and in keeping up with updates and consistency. New approaches are being investigated by this group, but the exercise has shown that pooling heterogeneous information from existing remote databases in a single data warehouse is complex and laborious. The second is based on the use of a common language for describing data models. One of these languages is based on the Object Protocol Model (OPM; Chen and Markowitz, 1995), which uses an essentially object-oriented approach to describe the structure of a set of related objects. The possibility of evaluating this system is currently restricted by the small number of cases where it has been adopted. GDB 6.0 (Fasman et al., 1996) is certainly the most complex genomic database so far. It has redesigned its schema to follow the OPM representation and has adopted this federation concept. This language (and therefore the GDB schema) is said to be easier to comprehend for the average user, and hence to query. The third mechanism is based on direct pointers between databases, via stored Hypertext Markup Language (HTML) links. The advantage of this approach is its simplicity and the speed at which it can be implemented technically. It also guarantees that information is up to date and in its original format. Links allow users to jump from one database to another by simply clicking with the mouse. Several databases make use of this system. For instance GDB has direct links to EMBL and OMIM, which also have links between them. This has the strong disadvantages of not providing a uniform view of the data and not allowing the selection of the information that is queried, nor its validation.

2. The Human X Chromosome

2.1 Introduction

A different dosage of X chromosome in male and females tends to reveal the presence of X-linked disease genes in males. It is also the origin of the intriguing

process of X chromosome inactivation, which inactivates one X in females in order to maintain gene expression dosage. Both of these features are responsible for the fact that the X chromosome is the most intensively studied human chromosome (reviewed by (Mandel et al., 1992)). It was the first to have a genetic map based on restriction fragment length polymorphism (RFLP) (Drayna and White, 1985). Genes for two X-linked diseases, chronic granulomatous disease (Royer-Pokora et al., 1986) and Duchenne Muscular Dystrophy (Monaco et al., 1986) were the first to be isolated b y positional cloning approaches. Figure 1.1 represents an ideogram of the X chromosome and some of its important features.

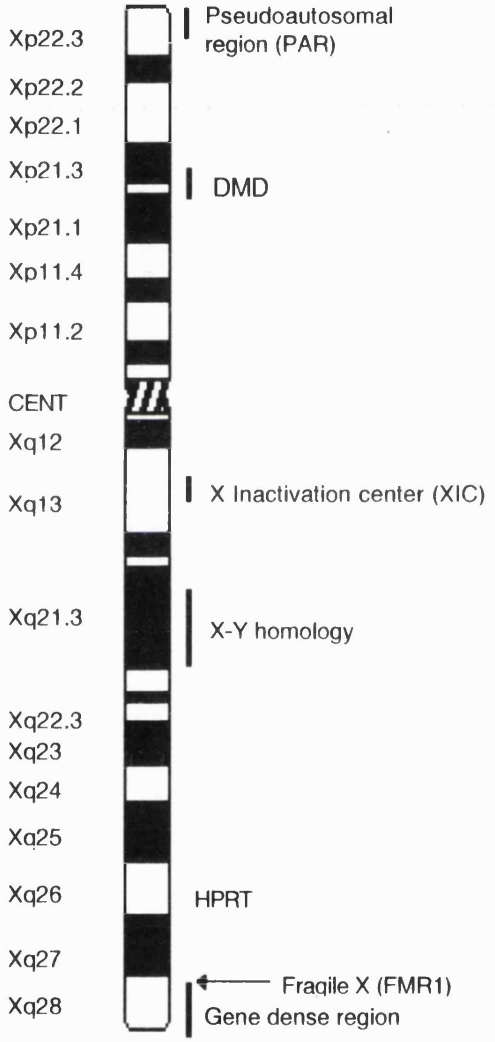


Figure 1.1 The human X chromosome. The pseudoautosomal region is a 2.6 Mb area at the tip of the chromosome identical to a region on the Y chromosome. A similar, smaller region is situated in Xq28. In Xq21.3, a large area has strong homologies to the Y but does not recombine with it. The DMD gene is the largest known human gene, spanning 2.6 Mb. The X inactivation centre controls X inactivation and contains a gene, XIST, that plays an important role in this phenomenon. The fragile X site in Xq27.3 is responsible for most cases of mental retardation in males.

2.2 X inactivation

Mary Lyon introduced the concept of mammalian X chromosome inactivation in 1961 (Lyon, 1961), with a hypothesis on dosage compensation to explain how males can survive with only one X chromosome. Using hybrids segregating rearranged X chromosomes that are inactivated, a region in Xq13.2-q13.3 called the X inactivation centre (XIC) was shown to be critical for controlling the process of X inactivation (Brown et al., 1991). The first gene found in this region of about 1000 kb was shown to be transcribed from the inactive X and not from the active one, in a non coding RNA that remains associated with the inactive X chromosome in female interphase nuclei (Brown et al., 1991; Brown et al., 1992). The map position, the expression pattern and the nature of the product of this gene called XIST (X-inactivation specific transcript) all indicate that it is likely to be involved in X inactivation. Further genetic evidence came recently from targeted mutation experiments (Penny et al., 1996) where null mutations were created in the mouse homolog Xist, showing that the Xist gene is required for X inactivation. A definite proof came very recently with the introduction of a 450 kb YAC containing the mouse XIC (including Xist) in male ES cells (Lee et al., 1996). The fragment, once inserted in autosomes, induces their inactivation in the ES differentiation pathway.

The mechanism which leads to inactivation is still unclear, although a model has been proposed in the mouse (Willard, 1996). During early development, both X chromosomes are in a state of 'pre-inactivation', and the Xist gene is expressed at low levels on both. The XIC guides the choice of which copy gets blocked, in a random fashion, thus down-regulating one Xist and leading to an active chromosome. The other Xist by default is up-regulated, the Xist RNA spreads along the chromosome, and inactivation occurs in most genes. The use of YAC transgenics has provided a new approach to manipulate the XIC and the XIST/Xist gene. This should lead in the near future to a better understanding of the molecular interactions which control X inactivation.

2.3 Evolution of the X and Y chromosomes

The differentiation of sex chromosomes in the course of evolution has followed a complex pattern which is not yet well understood. Increasing comparative mapping and genetic data offer some clues which shed light on part of this process. It is likely that the remote ancestor of most present days species with separate sexes may have been hermaphrodite, i.e. each individual was both male and female (Bull, 1983). Differentiation between males and females occurred by the appearance on the proto-Y chromosome of one or more genes that had consequences on gonad differentiation. This ancestral chromosome is likely to be different for birds, since no homologous genes are found between bird and mammalian sex chromosomes. From this point, the

ancestral autosome is likely to have evolved into the present X, while the mutated homolog took a different path and became the Y. The question as to how this happened is still the subject of much debate. Suggestions include either a major rearrangement, or the gradual degradation of the Y in combination with temporary expansions and attritions (Graves, 1995). In either case, the Y became protected from recombination after the acquisition of the sex determining factor(s) (Charlesworth, 1991). The X chromosome of monotremes and marsupials share the long arm and the pericentromeric region with their human homolog, indicating that most of human Xp was added after primates diverged from marsupials, about 130 millions years ago (Graves and Watson, 1991) (Watson et al., 1993). The boundary has recently been refined to Xp11.23 (Wilcox et al., 1996). This addition happened either as one fragment, or in separate stages but in either case occurred on both the X and the Y chromosome. The parts added to the Y chromosome would have then been subjected to the same forces of selection and drift that resulted in a progressive degradation of the original Y. Genes within this region became progressively inactivated, mutated or lost, while X homologs were simultaneously recruited by the X inactivation system to ensure a balanced expression dosage (Graves, 1995). A notable exception is the STS gene, which is active on both female X chromosomes, but is a pseudogene on the Y (Yen et al., 1988), implying that dosage imbalance can be tolerated in some cases. One region is still identical between the X and Y, and recombines in male meiosis in at least one point (Burgoyne, 1982) (Rappold et al., 1994). Situated at the telomeric end of the X and Y short arms, this region behaves as an autosome would, and is therefore called the pseudoautosomal region (PAR). It is known to contain 6 genes, which are expressed on both X and Y and therefore are predicted to escape inactivation in females (see (Rappold, 1993) for a review). The human PAR is not homologous to the mouse PAR, and ANT3, a human pseudoautosomal gene, is autosomal in prosimians (Toder et al., 1995). This suggests that the PAR represents a relic of the latest addition to the sex chromosomes recently in primate evolution (less than 40 million years ago). Due to these differences, the gene content of the pseudoautosomal region is not responsible for its function *per se.* However, the PAR must have had an important role in sex chromosome evolution, to ensure that each new addition is equally transferred to both the X and the Y. Studies on the STS and KAL genes which are just outside the PAR boundary on the X, show that these are part of a segment added even more recently, due to a very similar genomic structure of STSP and KALY, their respective Y pseudogenes (Del Castillo et al., 1992). The observation that both pseudogenes are in the same orientation, but on Yq and inverted compared to their X-linked counterparts suggests that the addition may have been followed by a pericentric inversion (Yen et al., 1988). The construction of a Yp deletion map has recently shown that there are two further regions of homology between the sex chromosomes: proximal Yp-Xq21 and distal Yp-Xq21 (Sargent et al., 1996).

Studies described above have shown that the long arm and the pericentromeric region of the human X chromosome are extremely well conserved across species, (reaching as far as monotremes) which implies a stable evolution during the last 170 million years approximately. This can partly be explained by Susumu Ohno's suggestion that the X is protected from rearrangements in order to avoid imbalance in dosage of gene expression (Ohno, 1967). Since the X lacks a counterpart to recombine with in males and is subject to inactivation in females, a rearrangement would create a double expression in the new autosomal location, which would presumably be lethal and therefore be selected against. However some exceptions exist to this theory, for instance the presence of genes in monotreme X chromosomes which pair with the Y and are presumably not inactivated (Graves and Watson, 1991). In addition, a recent report (Rugarli et al., 1995) shows that the CLNC4 gene in human Xp22.3 has a different localisation between two mouse species, near the pseudo-autosomal region for *Mus spretus* and on chromosome 7 in a laboratory inbred strain. Although inter-chromosomal rearrangements are limited when involving the X, it is clear that intra-chromosomal inversions and translocations have occurred. For instance in the mouse, 8 conserved linkage groups have so far been detected, which indicate a very complex series of rearrangements (Nelson et al., 1995).

2.4 Identifying X-linked disease genes

Because males are hemizygous for X, recessive diseases tend to be revealed following a typical pattern of female carriers and affected males. This accounts for the fact that more than half of the genes identified by positional cloning approaches so far are located on the X chromosome. The characteristic inheritance pattern has made the positioning of these genes obvious. Furthermore, several features of X chromosome genetics have assisted in their cloning. For instance, rare male patients who survive with large deletions spanning several megabases suffer from a contiguous gene syndrome. If several overlapping deletions are available, very accurate mapping of the genes can be performed (Ballabio et al., 1989; McCabe et al., 1992). Also rare females have been found with balanced X-autosome translocations where the intact X chromosome is inactive. If the breakpoint occurs within an X gene, it will lead to expression of the disease, and will provide a precise localisation for the gene. For X-linked diseases that severely affect reproductive fitness in males, unrelated patients can be expected to carry independent mutations (this is not the case for autosomal recessive disorders). Some of these mutations may be small deletions, and therefore it can be efficient to screen a panel of unrelated patients with probes from the critical region, provided it has been narrowed down to within a megabase. This strategy was first used for the DMD gene (Monaco, 1985). Identification of X-linked diseases has also been facilitated by the presence of selectable gene markers that ensure retention

of portions of the X in human/rodent somatic cell hybrids. For instance, both positive and negative selection for the hypoxanthine-guanine phosphoribosyltransferase (HPRT) locus in Xq26 allow the control of the retention of X chromosome fragments.

2.5 Status of the physical and genetic map

Genetic maps of the X chromosome present a special challenge because they can only be generated in female meioses, where two X chromosomes can pair and recombine. The latest whole-genome genetic map constructed by the Gén_thon group contains 216 markers mapped on the X chromosome (Dib et al., 1996) representing an average of 1.32 marker per Mb, the lowest density after chromosome 9. This was achieved by genotyping 20 families (against 8 for autosomes) representing 291 meioses (186 for autosomes). In this work, X chromosome markers showed also the lowest average heterozygosity (0.65). The shortage of markers which could be mapped to the X chromosome may be due to an actual deficit of $(CA)_n$-repeats on the X chromosome, as observed on the mouse X chromosome and other mammals (Dietrich et al., 1996). A radiation hybrid map was integrated with the above genetic map by screening the 91 radiation hybrids from the Genebridge 4 panel (Walter et al. 1994) with 6469 STSs which included most of those used for the genetic map (Hudson et al., 1995). As a result, an additional 272 markers were placed on the X chromosome.

Four separate groups are working on a whole X chromosome physical map in YAC clones. The two first ones are part of whole-genome efforts, and are based on STS screening of the CEPH 'mega' YAC library using two complementary approaches. In the first, genetic markers developed at Généthon (Dib et al. 1995) were screened against the YAC library (Chumakov et al., 1995). In the second, STSs developed by random sequencing of genomic clones, and a variety of other markers derived from microsatellite repeats as well as those from the above genetic map were first ordered via irradiation hybrid mapping and then screened against the YAC library (Hudson et al., 1995). In both cases the lower coverage of the X chromosome stands out compared to autosomes. This is partly due to the lower representation of X chromosome clones in the YAC library which was made from a male cell line, and to the lower number of polymorphic $(CA)_n$-repeats developed. Another effort focused on the X chromosome is led by the group of D. Schlessinger at the Washington School of Medicine in St. Louis. Here the strategy is essentially based on developing STSs from YAC ends and screening a selection of YAC libraries by PCR. It also incorporates YAC contigs developed separately by many other groups. This map was reported at the 7th X chromosome workshop in Hinxton (U.K.) to cover the chromosome at 85 kb inter-STS distance. The fourth X chromosome mapping project is described in this thesis and has so far resulted in 24 contigs covering 125 Mb or (80%) of the chromosome (Roest Crollius et al., 1996).

CHAPTER TWO: Material and Methods

1. Reagents

1.1 General reagents

Sigma Chemicals Co.
Trizma hydrochloride
Trizma base
bovine serum albumin (BSA)
dithiothreitol (DTT)
phenylmethylsulphonylfluoride (PMSF)
yeast t-RNA
salmon sperm DNA
sorbitol
antibiotics
markers for normal agarose gels
amino acid supplements for yeast work
ß-mercaptoethanol
human placental DNA for competition
hydrazine

BDH Laboratories
general salts and chemicals
proteinase K
ethanol
isopropyl alcohol
phenol
chloroform
isoamyl alcohol
N-lauroyl sarcosine
dichloromethane

MERCK
general salts and chemicals

Difco
bacto-tryptone
bacto-peptone
bacto-yeast extract
bacto-agar
yeast nitrogen base

Amersham International plc.
Hybond N+ membranes
[^{32}P-alpha] dATP (3000 mCi/mmol)
[^{32}P-alpha] dCTP (3000 mCi/mmol)

New England Biolabs
markers for normal gels
λ DNA

Pharmacia
dNTPs for PCR
dextran sulphate
hexanucleotides for random prime labelling

FMC
SeaKem GTG agarose
Saccharomyces cerevisiae marker chromosomes
lambda concatamer markers

Boehringer Mannheim
DIG-11-dUTP
DIG-11-UTP
anti-DIG alkaline phosphatase conjugate
Biotin-16-dUTP

JBL Scientific Inc.
Attophos

Bio-Rad
Acrylamide

LKB Broma
Bind silane

Fluka
ammonium persulfate

1.2 Enzymes (and Enzyme Buffers)

New England Biolabs
general restriction enzymes
T4 polynucleotide kinase

Boehringer Mannheim Biochemicals
Streptavidin-AP conjugate
Anti-DIG antibody Alkaline Phosphatase conjugate

Sigma Chemicals Co.
agarase
Nova Biolabs
Novozyme

ICN Biomedicals
Zymolase

Advanced Biotechnologies Ltd
Taq DNA polymerase

1.3 Other reagents and kits

Gibco BRL
CloneAmp UDG cloning kit

2. General solutions and media

(According to Sambrook et al., 1989)
LB media

bacto-tryptone	1%
bacto-yeast extract	0.5%
NaCl	0.1%
(bacto-agar	1.5%)

pH to 7.0

2x YT media

bacto-tryptone	1.6%
bacto-yeast extract	1%
NaCl	0.5%
(bacto-agar	1.5%)
pH to 7.0	

SOC media

bacto tryptone	2%
bacto-yeast extract	0.5%
NaCl	0.05%
Glucose	20 mM
KCl	2.5 mM
pH to 7.0	

10X HMFM

K_2HPO_4	360 mM
KH_2PO_4	132 mM
sodium citrate	17 mM
$MgSO_4$	4 mM
$(NH_4)_2SO_4$	68 mM
glycerol (w/v)	44 %

YPD media (non-selective)

yeast extract	1%
bacto-peptone	2%
D-glucose	2%
(bacto-agar	1.5%
ampicillin	50µg/ml

-U-T media (selective medium, SD)

D-glucose	2%
(+bacto-agar	2%)
yeast nitrogen base	0.67%
amino acid supplements	(final conc 1X)

20X amino acid supplements

adenine	400µg/ml
arginine	400µg/ml

isoleucine	400µg/ml
histidine	400µg/ml
leucine	1.2mg/ml
lysine	400µg/ml
methionine	400µg/ml
phenylalanine	1mg/ml
valine	3mg/ml
tyrosine	600µg/ml

Yeast nitrogen base

20X stock	13.4%

SCE

sorbitol	1.0M
sodium citrate pH5.8	0.1M
EDTA	10mM
DTT (added fresh)	10mM

1X TAE

Tris-acetate pH8.0	40mM
EDTA	1mM

0.5X TBE

Tris-borate pH8.0	45mM
EDTA	1mM

20X SSC

NaCl	3M
sodium citrate pH 7.0	0.3M

Denaturant

NaOH	0.5M
NaCl	1.5M

Neutralisation Solution

Tris-Cl pH7.0	1M
NaCl	1.5M

1XTE

Tris-Cl pH 8.0	10mM
EDTA	1mM

Church Buffer

Na_2HPO_4	0.25 M
SDS	5%
EDTA	10 mM

pH to 7.2

Stripwash Solution

TE	0.2X
SDS	0.1%

Alkaline Lysis Solution I

glucose	50mM
Tris-Cl pH8.0	25mM
EDTA	10mM

Alkaline Lysis Solution II

NaOH	0.2N
SDS	1%

Alkaline Lysis Solution III

potassium acetate 5M 60ml

glacial acetic acid	11.5ml
H_2O	28.5ml

(This solution is 3M with respect to potassium and 5M with respect to acetate).

TEN9

Tris-Cl pH 9.0	50 mM
EDTA pH 8.0	100 mM
NaCl	200 mM

3. Experimental procedures

3.1 Hybridisation fingerprinting of YAC clones

3.1.1 Preparation of genomic DNA from cell lines

The 578 cell line (~1.10^8 cells) was obtained as a monolayer culture from the ICRF Cell Production unit. Cells were detached from the flask using trypsin-versene and pelleted in a Beckman J6-HC centrifuge for 10 min at 1000 rpm. Cells were resuspended in 20 ml Ten9 and RNase to 100 µg/ml final. After a 10 min incubation Proteinase K was added to 1 mg/ml and SDS to 1%. The cell suspension was incubated 12 hrs at 50°C in a plastic tube on a rocker. The preparation was extracted once with phenol and once with chloroform. In each case, the emulsion was placed on a rocker for 2 hrs and centrifuged at 3000 rpm (Beckman J6-HC) in a Falcon tube for 5 min at room temperature to separate the phases. The aqueous phase was transferred to a fresh tube, and NaAC was added to 100 mM, followed by the addition of 0.6 volume of isopropanol at room temperature to precipitate the DNA. The sample was placed on a rocker until a precipitate formed, which was spooled out with a glass rod. The DNA was washed by dipping the glass rod in 70% ethanol, then transferred to 2 ml of 1xTE for rehydration. The DNA was left to dissolve at 50°C with gentle agitation and stored at -20°C.

3.1.2 Preparation of whole yeast DNA from YAC clones

Whole yeast DNA was extracted by a modification of the procedure of Chumakov et al. (Chumakov et al., 1992) and used as template for large scale inter Alu PCR of YAC clones. Frozen glycerol stock YAC libraries were replicated into 96-well microtitre plates containing 100 µl selective medium (Anand et al., 1989) (SD -ura, -trp) , and cultures were grown for 3 days at 30 °C. Cells were pelleted for 10 min at 2000 rpm (Beckman J6-HC), and supernatants were removed by inversion. Cell pellets were washed in 50 µl SCE buffer, then cells were harvested as before. Yeast cells were converted to spheroplasts by incubation for 1 hour at 37 °C in 25 µl SCE containing 4 mg/ml novozym and 10 mM dithiothreitol. Then, 60 µl 0.14 N NaOH were added to each well and plates were incubated for 7 min at room temperature. DNA extracts were neutralised with 60 µl of 1 M Tris-HCl (pH 8.0) and stored at -20 °C.

3.1.3 Preparation of plasmid DNA

Plasmid DNA from the XPL library was prepared by the alkaline lysis miniprep procedure described in (Sambrook et al., 1989).

3.1.4 Alu PCR reactions

3.1.4.1 from total genomic DNA

Conditions for amplifying inter-Alu PCR fragments were identical for both genomic DNA from the cell line 578, or genomic DNA from yeast cultures containing YAC clones. Inter Alu PCR was carried out in 67mM Tris-HCl (pH 8.8), 16.7 mM $(NH_4)_2SO_4$, 6.7 mM $MgCl_2$, 0.5 mM each dNTP, 170 µg/ml BSA, 10 mM 2-mercaptoethanol, 1.3 µM primers (Cole et al., 1991) and 0.6 U Taq polymerase. Approximately 50 ng of template genomic DNA was added. Primers were either Ale1, Ale3 or HL1, or a combination of the above. For the amplification of DNA for cloning in the pAmp1 vector, primers were a 1:1 mixture of $(CAU)_4$-Ale3 and $(CUA)_4$-Ale3. Cycling conditions were an initial denaturation step at 94°C for 4 min, followed by 30 cycles at 94°C for 30 sec, 68°C for 1 min, 72°C for 1 min, and a final extension step at 72 °C for 4 min, in an MJ-PTC100 thermocycler. The presence of products was verified by running 1/10 of the reaction on a 1.8% agarose gel in 1xTAE for 1 hr.

Ale3: CCA CTG CAC TCC AGC CTG GG
Ale1: GCC TCC CAA AGT GCT GGG ATT ACA G
HL1: CAT GGC ACA TGT ATA CAT ATG TAA CWA ACC (W=75% A, 25%T)

3.1.4.2 from the entire XPL and cX libraries

Prior to spotting on nylon membranes, the XPL and cX libraries were amplified by Alu-PCR. A water-bath PCR robot built in our laboratory was used for this purpose (Meier-Ewert et al., 1993), which allows entire libraries to be amplified in microtitre plate format. Conditions for YAC clone amplification were as follows. The PCR buffer was as above, except that a mixture sufficient for 5000 reactions was prepared and dispensed in 50 µl aliquots into the wells of 384-well polypropylene microtitre dishes (Genetix). DNA from individual clones in 96 well plates was transferred to the reaction plate using a 96-pin plastic device (Genetix), and plates were heat sealed with a plastic film. PCR was carried out using the large capacity waterbath robot for 30 cycles of 3 min at 94 °C followed by 6 min at 65 °C, with an initial denaturation of 10 min and a final extension of 10 min. The XPL library was amplified in an identical way, except for the components of the PCR reaction which were 75 mM Tris-HCl (pH 9), 20 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.1 % (w/v) Tween, 0.2 mM each dNTP, 1.5 µM primer Ale3, and 0.5 U Taq polymerase. Frozen glycerol stocks of the plasmid clones were directly used as template for amplification, by transferring a small amount of frozen cells from the 384 well storage plates to the reaction plates with a 348 pin gadget (Genetix).

3.1.5 Preparation of electro-competent cells

The strain of *E. coli* chosen for cloning the X chromosome Alu-PCR products was DH5α (Genotype: F^- φ80d*lac*ZDM15, Δ(lacZYA-argF)U169, deoR, recA1, endA1, hsdR17(r_K^-, m_K^+), gal^-, phoA, supE44, l^-, thi-1, gyrA96, relA1). One liter of L-broth was inoculated with 1/100 volume of a fresh overnight culture. Cells were grown at 37°C with vigorous shaking to an absorbance of 0.7 at 600 nm. Cells were harvested by chilling the flask on ice for 15 min, before centrifuging in a cold rotor at 4000 x g for 15 min. The pellet was resuspended in a total of 1 liter ice cold water, and centrifuged as above, then resuspended in 500 ml of ice cold water, and centrifuged again as above. The pellet was resuspended in 20 ml 10 % glycerol, centrifuged as above, and resuspended in a final volume of 2 ml 10 % glycerol. The cell concentration was approximately 3 x 10^{10} cells/ml. The suspension was aliquoted in small polypropylene tubes (40 µl), snap frozen at -70 °C and kept for up to 6 months.

3.1.6 Cloning with the pAMP1 system

For the annealing reaction, approximately 50 ng of Alu PCR products from cell line 578, 50 ng of pAMP1 vector (Gibco BRL), 1 U Uracil DNA glycosylase were added together in a final volume of 20 µl 1 x annealing buffer (20 mM Tris-HCl (pH 4.8), 50 mM KCl, 1.5 mM $MgCl_2$). The components were mixed, incubated 30 min at 37 °C, and placed on ice. The cloning method is based on the degradation of the uracil bases at the extremities of each PCR product (see 3.1.4.1), creating an overhang that is complementary to the cloning site of the vector. The annealing reaction was used directly for transformation without purification. Electro-competent DH5α cells were thawed on ice and immediately mixed with 1 to 2 µl of the annealing reaction. After 1 min incubation on ice, cells were then subjected to electroporation (Dower et al., 1988) in a BioRad Gene Pulser apparatus. The cell suspension was transferred to a 0.1 cm cuvette and a 2.5 kV pulse was applied (25 µF, Pulse Controller to 200 Ω). The time constant was between 4.5 to 5 msec (field strength approximately 12.5 kV/cm). The cuvette was immediately removed from the chamber and 1 ml of SOC medium was added. Cells were allowed to recover by incubating the suspension at 37 °C with agitation for 1 hr, before plating on selective medium (2 YT agar + ampicillin). Clones were picked in 384 well microtitre plates, containing 60 µl of 2YT medium + ampicillin per well, with a picking robot constructed in our laboratory

3.1.7 Preparation of library filters

The Alu-PCR products of each clone of the XPL library were robotically spotted on nylon filters (see Chapter 3). A few nanolitres of DNA were transferred by the pins of the robot gadget in a regular array. After spotting, filters were gently submerged in denaturing solution for 2 min and transferred to a large volume of

neutralising solution for 2 min. DNA was cross-linked by UV irradiation and filters were air dried.

3.1.8 Hybridisation of YAC Alu PCR probes

Several steps of this protocol were modified in the course of the project, to reflect improvements aimed at streamlining the procedure. The development and testing of the optimisations are described in detail in chapter 4. The following protocols are those used for producing the majority of the data.

3.1.8.1 Radioactive hybridisations

Approximately 25 ng of YAC Alu-PCR products were labelled by random priming (Feinberg and Vogelstein, 1983) with α^{32}P-dATP (Amersham). Incorporation was checked by separating 1µl of the labelling reaction on polyethyleneimide (PEI) chromatography paper in 750 mM KH_2PO_4 (pH3.5). Repetitive sequences were blocked by hybridising the probe in solution to an excess of denatured total human placental DNA (SIGMA). The probe and 500 ng/ml total human DNA were denatured together in 100 µl of 180 mM sodium phosphate, and allowed to renature for 3 hrs at 65°C (Sealey et al., 1985). In parallel, filters were pre-hybridised in 5 to 10 ml of hybridisation buffer (5% SDS, 1.0 M sodium phosphate) for 3 hours at 65 °C The probe was added to the hybridisation buffer, and left to hybridise for 12 to 16 hrs at 65 °C. Filters were washed together first in 2xSSC. 0.1 % SDS at room temperature for 20 min, then once in pre-warmed 0.1 SSC, 0.1 % SDS at 65 °C for 20 min with agitation. Filters were wrapped in Saran film and exposed to X ray film for 16 to 48 hrs at -70 °C.

3.1.8.2 Non-radioactive hybridisations with biotinylated probes

Primers used for Alu-PCR amplification were labelled with a molecule of biotin at the 5' end using biotin amidite (ICRF Central Services, Clare Hall). Subsequently, each PCR product generated was assumed to be labelled with the same efficiency as the primers were. Approximately 50 ng of DNA was pre-annealed in solution in order to block repetitive sequences, as for radioactive probes. Hybridisation and washing steps were also identical. After the last wash, filters were incubated 5 min in PBS + 5% SDS, and then incubated 45 min in a minimum volume of the same solution supplemented with 0.3 U Strepavidin Alkaline Phosphatase/ml. Filters were washed 4 times for 10 min in PBS, and once in 0.1 M diethanolamine (pH 9.0), 1 mM $MgCl_2$. Approximately 25 µl/cm^2 of Attophos substrate (0.6 mg/ml in 2.4 M diethanolamine) was sprayed or dabbed directly on the membrane. After 1 hr incubation at room temperature, images of the entire filters were captured under U.V. light (365 nm) with a

CCD camera (Pulmix, 740x572 pixels or Photometrics, 1317x1035 pixels) and stored in a Unix workstation as TIF files (Tagged Image Format).

3.1.8.3 Non-radioactive hybridisations with DIG labelled probes.

YAC Alu-PCR products were labelled during the PCR reaction by incorporation of DIG-11-dUTP. For this, the dTTP nucleotide was replaced with a 65:35 ratio of dTTP:DIG-dUTP. Approximately 125 ng of PCR products were used for pre-annealing, in the same conditions as for radioactive probes, and then added to 3 ml of hybridisation buffer (5% SDS, 0.5 M sodium phosphate) pre-warmed at 65 °C. The probe solution was sprayed with an air-gun on the filters and left to hybridise 12 to 16 hrs at 65 °C. Several filters were generally stacked on top of each other and wrapped in Saran film to avoid evaporation. Filters were washed as for radioactive hybridisations, and incubated 5 min in PBS + 5% low fat dehydrated milk (Tesco). Filters were incubated 45 min at room temperature in a minimum volume of the same solution supplemented with 0.1 u/ml anti-DIG antibody conjugated to alkaline phosphatase. The following washes and detection steps are as described above.

3.1.8.4 Strip washing of filters.

After hybridisation and detection, radioactive probes were removed by placing the filters in strip wash solution (0.2xTE, 0.1%SDS) heated to 90°C and incubating for 45 min at 65°C. Non radioactive probes were removed by first rinsing the filters in 0.4M NaOH, to ensure that alkali-labile digoxygenin was removed, followed by the same treatement as for radioactive probes. Filters were stored in 50 mM Sodium phosphate, 10 mM EDTA at room temperature, or air dried and wrapped in Saran film for longer periods of time.

3.2 Development of alternative hybridisation systems

3.2.1 Preparation of polyacrylamide coated glass plates

Microscope slides were used throughout all the test experiments. Slides were thoroughly washed in 1M NaOH for 60 min, rinsed with redistilled water and dried at 80 °C for at least 3 hr. The side of one slide was wiped with Replicone ($(CH_3)_2SiCl_2$), followed by Triton X-100. One face of a second glass plate was wiped with a 0.1% solution of Bindsilane (3-Trimethoxysilylpropyl methacrylate) in order to get high binding capacity. The two slides were then clamped together, treated sides facing each other, and spaced with one layer of Scotch tape (appr. 30 µm thick). An 8% acrylamide solution (acrylamide:bis-acrylamide=30:1) was prepared separately. To

500 µl of this solution, 1µl of TEMED (tetramethyl ethylene diamide) and 1 µl of freshly made ammonium persulfate ($(NH_4)_2S_2O_8$) were added just before use. Approximately 200 µl of this solution was then pipetted between the 2 glass plates, avoiding any bubbles in the acrylamide. After 20-30 min of polymerisation, a layer of polyacrylamide was bound to the face treated with Bindsilane. The other slide was removed, and the 'acrylamide' slide washed with redistilled water and dried at room temperature. The slide was placed in a 50% hydrazine (N_2H_4) solution for 1 hr, followed by thorough rinses in redistilled water. The slide was finally dried at room temperature, ready for use. Prior to spotting, DNA was always denatured by addition of NaOH to 0.2 M to the DNA solution. The DNA was then manually spotted with a micropipette, generally as 1 µl drops, and the glass plate was air dried for at least 2 hrs.

3.2.2 Hybridisation to polyacrylamide coated glass plates

Several protocols have been used in the course of this project, to hybridise probes to DNA fixed on acrylamide coated glass plates. The varied parameters essentially included hybridisation temperature and buffer, and washing conditions after hybridisation. Details of the different tests performed are described in chapter 4. The protocol outlined here is the procedure that gave the best results, taking into account all parameters tested.

Probes were XPL clone inserts amplified by Alu-PCR from *E. coli* colonies. Approximately 20 ng of DNA was labelled by random prime labelling (Feinberg and Vogelstein, 1983). Glass plates prepared as described above were incubated in hybridisation buffer (10 mM sodium phosphate, 1.0 M NaCl, 10 mM EDTA, 10 % PEG) or modified Church buffer (0.25 M Na_2HPO_4, 5% SDS) for 10 min prior to the addition of heat denatured probe (typically 0.2. 10^6 cpm/ml). The probe was left to hybridise for 1 hr, and glass plates were washed in 2 x SSC for 10 min, wrapped in Saran film and exposed to a Phosphorimager screen (Molecular Dynamics) for 1 hr.

4. Computer software

4.1 Scoring hybridisation results

4.1.1 Radioactive hybridisations

Hybridisation results generated using radioactive probes were obtained on X-ray films. The positions of the signals on the filters were copied manually with a marker pen on transparencies where a scaled grid was printed representing an array of all DNA spots on a filter. This step allowed the positive signals to be recorded on a hard copy support, without filter background and hybridisation artefacts. Transparencies

were scanned in a Unix Sparc2 workstation using an X-ray film scanner (Amersham). The resulting TIFF images were scored using Acepro (Huw Griffith). This program first locates the boundary of the rectangle representing the filter, identifies the grid of DNA spots and searches in the area corresponding to each predicted DNA spot for a peak of pixel intensity. Each peak in principle is the result of a recorded positive signal made by the marker pen when scoring the X-ray film. The output of the program is the position of the intensity peaks within the grid, in ASCII. Each 'positive' is identified by the microtitre plate that it originates from and the well within the plate.

4.1.2 Fluorescent probes

Hybridisation results using fluorescent probes were directly recorded on a computer disk via a CCD camera (Pulmix, 740 x 572 pixels or Photometrics, 1317 x 1035 pixels). The water cooled Photometrics camera was controlled with the IPLab Spectrum PVCam Support software (version 2.5.6k, Signal Analytics, Vienna, Virginia, USA) from a Power Macintosh computer. Images were transferred by ftp (file transfer protocol) to a Unix workstation (Sun Sparc2 or DEC alpha 150) and scored with spotter16, an image analysis program written by R. Mott. It uses a two-stage algorithm to locate the positive spots on the image. First, it identifies the central guide dot spotted in the center of each 384 block of DNA spots arrayed on the filter (because images are captured under U.V. light, all guide dots are visible, (see figures 3.4, 3.10b and c, chapter 3). For this, the program sums the pixel values of the rows and columns over the image, and creates a grid in which the rows and columns intersect on the guide dots. Next, a search is made for a local maximum of pixel value in the square corresponding to each DNA spots around the central guide dot. To take advantage of the fact that each positive clone should appear twice within a block, the local maximum of each duplicate pair is computed simultaneously. The signal intensity is then quantified by integrating around each spot centre, and adjusted by substracting the local background around each block. A log file is finally printed that contains a list of the grid coordinates for each spot found positive, together with their signal intensity before and after background subtraction.

4.2 Analysis of hybridisation results

Generating data by hybridisation of DNA probes to high density filters can be very fast and can rapidly overcome the capacity of investigators to analyse the results manually. Our laboratory has used this method first to generate maps in YACs, P1 and cosmids clones of *Schizosaccharomyces pombe*, and consequently developed computer programs to analyse the raw hybridisation data (Mott et al., 1993). These programs were further developed and optimised by their authors to adapt to more noisy and complex dataset such as that produced by the X

chromosome project. This dataset consisted essentially of YAC Alu-PCR hybridisations to YAC high density DNA filters (approximately 1000 hybridisations), and to XPL library filters (196 hybridisations). Analysis of the data was part of this thesis project as described in chapters 3 and 5.

Files containing the raw hybridisation results are simple ASCII text files named after the probe and the filter used in a given experiment, and listing the clones found positive on the filter (described in § 4.1, this chapter). These files are either generated manually, or are the direct output of image analysis software. A database is first created and maintained on a Unix workstation using the program well2clone (R. Mott). Its basic function is to merge into a single flat file the results of all individual hybridisation experiments held in separate files. Well2clone also performs a number of related tasks, such as changing clone and probe names into their canonical forms and correcting clone names according to the order of plates spotted on filters. Well2clone produces two files, that are suitable for contig building software. First a list of all probes used in the project so far, unordered. Second, a file containing a matrix with clones listed as rows, and probes listed as columns, and filed with a single value at the intersection between each clone-probe pair. The value is 1, 2, or 3 if the probe actually hybridises to the clone with a weak, medium or strong signal intensity respectively. A value of 0 indicates a negative, and 9 indicates a missing value, eg the probe was not hybridised to a filter containing this clone.

The main program of the contig-building package written in our laboratory (R. Mott, A. Grigoriev) is probeorder. Its main purpose is to order probes and clones using an algorithm based on simulated annealing. The program assumes that the probes are single-copy. The output is the reordered probes and clones, plus a logfile of diagnostics. Probeorder precalculates a distance between each pair of probes, which is by default the percentage of clones hit by one probe in a pair but not by both. It then uses simulated annealing to find that order of the probes that has the minimum sum of inter probe distances. Once the annealing has produced an order of probes, the program calculates how best to fit the clones under probes. An order can be suggested to the program for ordering probes along the map, based for example on FISH results or marker content. Probeorder produces a log file (in ASCII text) that contains a list of contigs and possible connections between them. This file is then the basis for a more refined manual analysis, which can take into account additional mapping data, on a contig by contig basis. (Probeorder and the suite of analysis tools is available by anonymous ftp from ftp.lif.icnet.uk in directory icrf-public/GenomeAnalysis/contig.tar.Z).

4.3 Digitising maps

A large number of clone and marker maps were entered into databases in the course of this project, which justified the writing of dedicated software tool to assist the

process. Xcontigview, written by Huw Griffith (ICRF Genome Analysis lab), is an X window based program written in C and using the X toolkit widget set. Maps are scanned from the paper support in a TIFF image file that is read and displayed by Xcontigview. An image overlay is created by the user via the graphical interface. The boundaries of items such as clones, probes or genes are set by mouse clicks and recorded by the program. The number of pixels between two clicks is converted into kb using a scale and origin that was entered beforehand. Clones that overlap with a probe are cross-referenced together with their respective coordinates in a text output, that is converted to ace files, suitable for parsing directly in an acedb database. Xcontigview is available as source code by anonymous ftp from ftp.mpimg-berlin-dahlem.mpg.de in pub/lehrach/users/crollius/xcontigview-1.5.tar.Z

4.4 Databases

4.4.1 Acedb

Acedb is an object-oriented database software with a strong emphasis on graphical maps for navigating in the data. It was written in C by Richard Durbin (The Sanger Centre, Hinxton, UK) and Jean-Thierry Mieg (CNRS, Montpellier, France), and includes a number of features contributed by many users of the system. It was first developed to visually represent physical maps constructed with the contig9 program written by J. Sulston, for the *C. elegans* mapping project. It is now a general tool for storing genomic and biological information and over 40 known projects are using it to maintain and distribute data. The acedb code and executables for most platforms are available by anonymous ftp from:
ftp.sanger.ac.uk in directory pub/acedb
lirmm.lirmm.fr in directory genome/acedb
ncbi.nlm.nih.gov in directory repository/acedb.

4.4.2 ORACLE.

ORACLE is a relational database management system available commercially from ORACLE corporation, Redwood City, USA.

CHAPTER THREE: Hybridisation Fingerprinting of YAC Clones

1. Introduction

Several methods exist which can reliably demonstrate the overlap between two YAC clones. They can be divided in three classes: hybridisations of labelled probes, STS detection by PCR, and fingerprinting, each very different in their experimental design and in the type of result generated.

The hybridisation of DNA probes to YAC DNA fixed on a solid support followed by the detection of the bound probe, is an extremely robust technique since it relies on the exact homology of long DNA molecules (between a few hundred bp to several kb) with their target YAC DNA. In this sensitive technique, where picograms of DNA can be detected on Southern blots (Southern, 1975), no prior knowledge of DNA sequence is necessary. The yield of the experiment is proportional to the number and diversity of target DNAs screened in parallel and in large scale physical mapping in particular, this latter feature can be exploited by using libraries of large number of clones gridded at high densities on a solid support (Lehrach et al., 1990). The detection of STSs (Olson et al., 1989) in YAC clones by PCR (Saiki et al., 1988), relies on the enzymatic amplification of a specific and unique sequence from the YAC DNA template. Here, a small stretch of DNA sequence must be determined before two primers can be designed for the PCR reaction. Although this is the most costly and time consuming part of the procedure, it is also this feature which allows a very rapid transfer of information between investigators, since only the primer sequence is needed to reproduce the experiment. Combined with dedicated efforts to generate polymorphic STS markers (Dib et al., 1996) and central repositories of mapping information such as the Genome Database (Fasman et al., 1996), this has established 'STS content mapping' of YAC clones as the most widely used technique for establishing genetic and physical maps of the human genome (Chumakov et al., 1995) (Hudson et al., 1995). Both probe hybridisation and STS mapping methods provide a direct evidence for the presence of an overlap between two YAC clones, respectively a specific hybridisation signal or a specific band on an electrophoresis gel.

The fingerprinting method, on the other hand, only indirectly detects overlaps. Mapping by clone fingerprinting is generally based on the separation by electrophoresis of a pool of sub-fragments obtained for each individual clone, followed by the estimation of the different fragment sizes. For each clone, the list of fragment sizes, or fingerprint, is then compared to others in the hope that overlapping clones will share a significant number of bands. For clones that can be separated away from the hosts' genome, restriction digests with frequently cutting enzymes can be used for

generating the fragments. Applied to cosmids, it was one of the methods used for mapping the 80 Mb of the *Caenorhabditis elegans* genome (Coulson et al., 1986). The size of YAC clones however prevents the use of such restriction digests as a means to obtain fragment pools for each clone insert. Typically between 200 kb and 1.5 Mb, YACs fall within the range of the yeast chromosomes themselves and therefore can not be purified easily. An alternative method is to use Interspersed Repeat Sequence-PCR (IRS-PCR), where primers are derived from a sequence likely to occur a number of times within a YAC clone. The Alu sequence from the SINE family of repeats was first used for this purpose (Nelson et al., 1989) due to its convenient frequency of approximately once every 4 kb of genomic DNA (Hwu et al., 1986). Two strong disadvantages of fingerprinting methods are first that they depend on a large overlap between clones in order to be reliable, and secondly that the statistical nature of the results necessitates that complementary techniques are used to confirm them. However, it requires very little starting DNA, which can be isolated as a crude preparation of whole yeast genomic DNA, and is very easy and fast to implement (Coffey et al., 1996).

This chapter describes a new method which combines the high degree of sensitivity and reliability associated with hybridisation based experiments, with the speed and ease of implementation of Alu-PCR. The principle is based on the cloning of a pool of Alu-PCR products from the X chromosome, followed by the arraying of the clones at high density on nylon filters, in the form of PCR products. The YAC clones to be fingerprinted are first amplified by Alu-PCR and then hybridised to the filters. Cloned PCR products identified in common between two YAC probes indicate an overlap (Figure 3.10A). In addition, the cloned PCR products are readily available as mapping reagents for further experiments. This approach was used to confirm overlaps between YAC clones that were already assembled into clusters in separate experiments (performed by M. Ross), to construct new contigs between previously unclustered YAC clones, and to place the cloned Alu PCR products on the emerging X chromosome YAC contig map, as a reservoir of new markers. Since the technique was to be used to fingerprint several hundred YAC probes by one person, different aspects of the method have been optimised with a view to facilitating the large scale of the experiment, raising the throughput of data production and facilitating the acquisition of the results. A streamlining of the hybridisation procedure was accomplished which included the evaluation of new non-radioactive detection systems and these aspects of the project are described in chapter 4.

2. Strategy

2.1 Basic protocol

Figure 3.1 summarises the strategy described in this chapter. DNA was prepared from a hybrid cell line containing a single copy of the human X chromosome on a hamster (CHO) background and used for an Alu PCR reaction with the Ale3 primer. After cloning in pAmp1 by annealing and transforming in DH5α, each colony was robotically picked and inoculated into 384 well microtitre plates. Each clone was then amplified by Alu-PCR, using the Ale3 primer, in a robotic waterbath PCR machine (Meier-Ewert et al., 1993). The PCR products were spotted onto nylon filters (figure 3.3 and 3.4) at a high density by a spotting robot. This library of Alu PCR products of the human X chromosome is hereafter referred to as XPL (X PCR Library).

YAC clones to be fingerprinted were selected from a collection of clones already assigned to the X chromosome (cX library), mainly in separate experiments performed by M. Ross. YACs were amplified by Alu-PCR using the primer Ale3 and after labelling, hybridised to the XPL filters. After scoring of the positive XPL clones, the data were entered in a UNIX database and analysed with probeorder and well2clone (Mott et al., 1993).

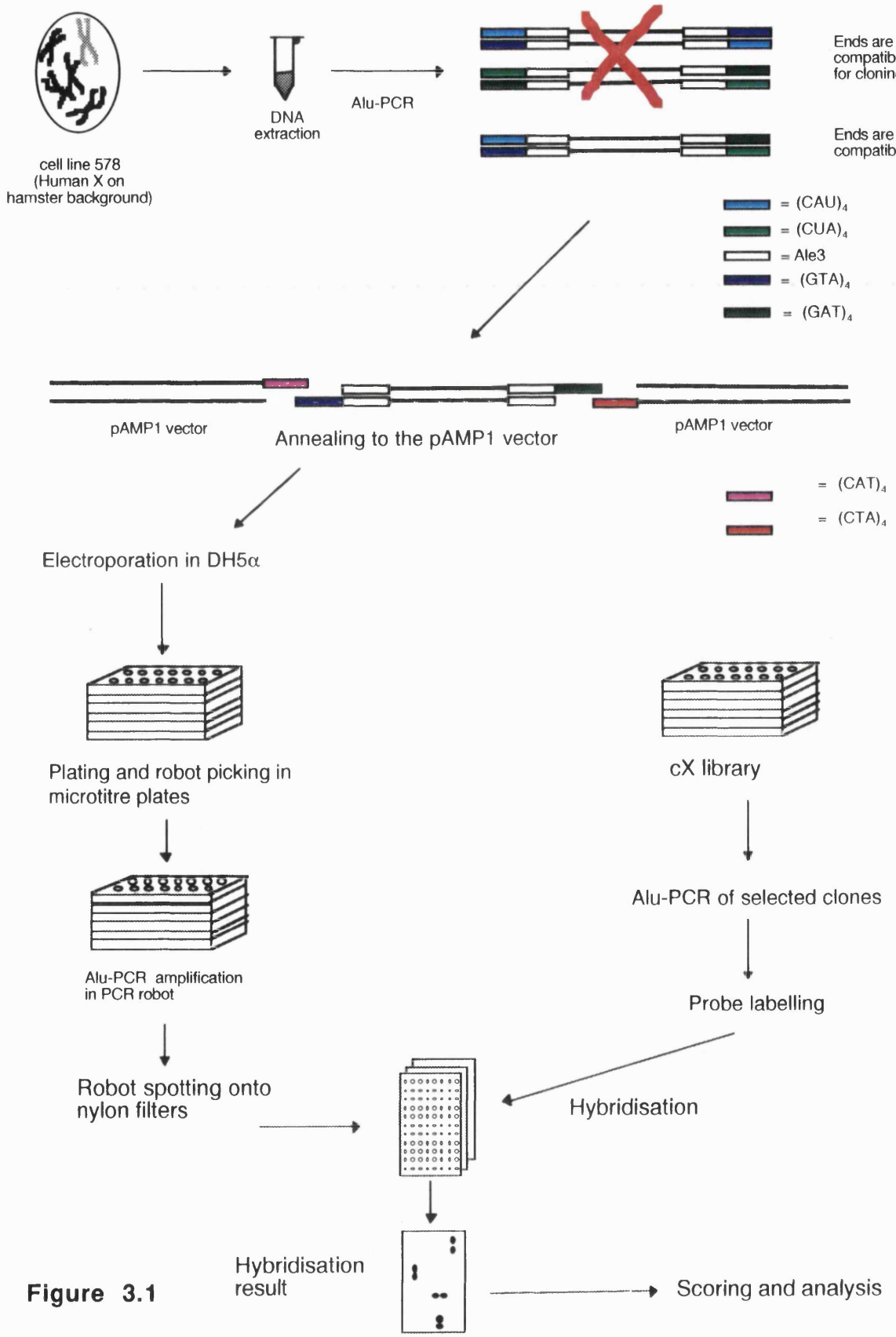

cell line 578
(Human X on
hamster background)
DNA
extraction
Alu-PCR
Ends are not
compatible
for cloning
Ends are
compatible
= (CAU)4
= (CUA)4
= Ale3
= (GTA)4
= (GAT)4
pAMP1 vector
Annealing to the pAMP1 vector
pAMP1 vector
= (CAT)4
= (CTA)4
Electroporation in DH5α
Plating and robot picking in
microtitre plates
cX library
Alu-PCR amplification
in PCR robot
Alu-PCR of selected clones
Probe labelling
Robot spotting onto
nylon filters
Hybridisation
Hybridisation
result
Scoring and analysis

Figure 3.1

2.2 Theoretical considerations

An important factor that was considered before starting this project was the number of XPL clones that were required for screening by each YAC probe, so that sufficient information could be generated in each experiment for detecting overlaps. The lower limit is the minimum number of clones necessary to detect all overlaps that exist in the set of YACs assigned to the X chromosome. There is no real upper limit, but factors such as filter size and spotting density were limiting, considering that high throughput needed to be achieved by one person.

The collection of YAC clones that were gradually assigned to the human X chromosome in our laboratory originated from 3 libraries: the ICRF (Larin et al., 1991), the CEPH 'mega' library (Chumakov et al., 1995) and an X chromosome specific YAC library (HHMI library, Lee et al., 1992). The average insert sizes for the three libraries is 670, 1054 and 250 kilobases respectively (see table 1 chapter 5 for a more complete description). In the course of the X chromosome mapping project, 1,727 ICRF YACs, and 643 CEPH YACs were identified. The HHMI library, where in principle all clones are from the X chromosome, contains 3,150 clones. The average insert size of the clones mapped to the X chromosome and constituting the pool of probes used for the fingerprinting project is therefore:

$$\frac{(1727 \times 670) + (643 \times 1054) + (3150 \times 250)}{(1727 + 643 + 3150)} = 475 \text{ kb}$$

In order to estimate the minimum number of XPL clones necessary to detect all overlaps, each YAC clone was assumed to overlap by 50 % with each neighbour for which an overlap must be detected. In practice, considering the number of YAC clones described above (5520 clones of average size 475 kb), this value is expected to be closer to 90 %. The fingerprinting approach however was based on pre-selecting most of the YAC probes from clusters that were assembled in previous experiments, and from which a minimal number of clones could be extracted that covered the cluster. The region of overlap between two clones is therefore smaller in this case, and a figure of 50% seemed a conservative and reasonable estimate.

Therefore the minimum number of XPL clones necessary to detect all overlaps is the number of intervals, on the entire chromosome, represented by 50% of an average YAC length:

$$\frac{160.10^3}{(475 / 2)} = 673 \text{ clones}$$

Where 160.10^3 is the size of the X chromosome in kilobase pairs.

In such an idealised model (figure 3.2), assuming that each cloned Alu-PCR product is evenly spaced along the chromosome, any YAC clone placed randomly on the map would at least contain 2 XPL clones, sufficient to detect its neighbours.

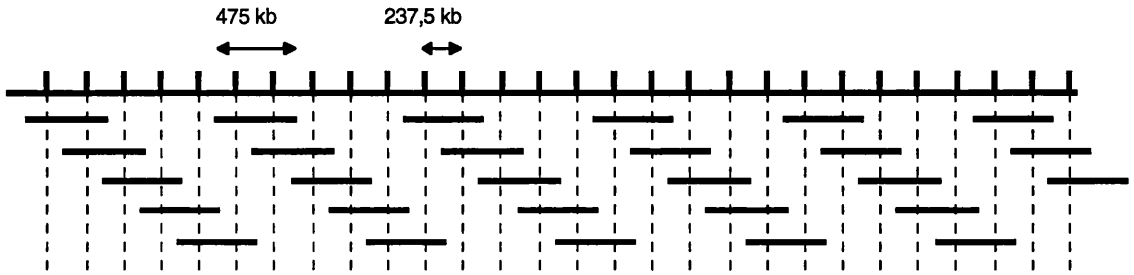


Figure 3.2. Idealised representation of a portion of the X chromosome, where XPL clones are evenly spaced every 237,5 kb (half an average YAC size).

Several known features of the experimental procedure and of the X chromosome mapping project were then taken into account in an attempt to adjust this minimum number to the real case:

- this model, although suitable for clones originating from the ICRF and CEPH mega libraries, would not allow most of the HHMI clones (average size 250 kb) to detect enough XPL clones to be linked to adjacent YAC clones.
- when amplifying Alu sequences from a very large molecule such as the entire X chromosome, one must expect a number of PCR artefacts to interfere, in particular a preferential amplification of some sequences over others, leading to a biased representation in the library and to a certain extent of redundant cloning.
- Alu sequences are not randomly distributed in the human genome. Some regions of the X chromosome will therefore have a lower probability of being represented in the library, leading to an impossibility to detect overlaps between clones mapping in those regions.

All these factors indicate that the minimum number of 673 XPL clones predicted by the above model would be insufficient. However, the threshold above which all overlaps between YAC clones can be detected, regardless of their size, of cloning artefacts and of Alu sequence distribution is impossible to predict, apart from the obvious fact that it must be higher than 673. Therefore the number of XPL clones to be picked must be set to the upper limit, which is fixed by requirements such as the highest density of clones that can be spotted on a membrane of suitable size for high throughput experiments.

At the time when the experimental procedure was designed, spotting robots designed in our laboratory had an accuracy of the order of 0.1 mm. Empirical results obtained so far by other members of the department showed that the highest density of spotted and grown colonies that allowed unambiguous scoring of positive clones

was 36864 on one membrane measuring 22 x 22 cm. Here, each XPL clone was amplified by PCR and therefore could be spotted in the form of naked DNA. The size of each DNA spot is dependent upon the size of the pins of the gadget which transfers the DNA from the microtitre plate to the nylon membrane, and is therefore much smaller than bacterial colonies. Consequently, the spotting density of the XPL library could be raised compared to colony filters and a new spotting pattern was designed to this effect. The gadget used for the DNA transfer is made of 384 pins, so that one robot movement from the plate to the filters can transfer the content of an entire 384 microtitre plate at once. Spotting patterns therefore must follow this format, and those used so far represented a grid of evenly spaced colonies at 1,125 mm intervals. To increase this density for the XPL library, a new spotting program was written by the engineer responsible for automation development (A. Ahmadi). The spacing between DNA spots decreased to 0.9 mm, allowing 55296 individual spots on one 22 x 22 filter.

However, with the current hybridisation protocols, it was decided that such filters were too large to be handled on a large scale and the spotting area of the complete library was therefore restricted to one sixth of the filter. This meant that instead of producing the normal 12 filters, a single robot run could produce 72 filters of the XPL library, each filter carrying one sixth of the total number of spots, or 9216. In order to reduce hybridisation artefacts consisting of dark spots on X-ray films which could be taken as a positively hybridising probes, each XPL clone was spotted in two positions on each membrane, therefore producing two hybridisation signals when identified by a YAC Alu-PCR probe. The spotting pattern 'in duplicate' was carefully designed so that the orientation of the two spots relative to each other is specific to the plate from which the XPL clone is issued (Figure 3.4)

Figure 3.3 shows an XPL filter hybridised with biotinylated Ale3 primers, and detected with strepdavidin-alkaline phosphatase conjugates and Attophos. It shows the array of 9200 DNA samples as bright spots, and the india ink guide dots as black spots on the filter background.

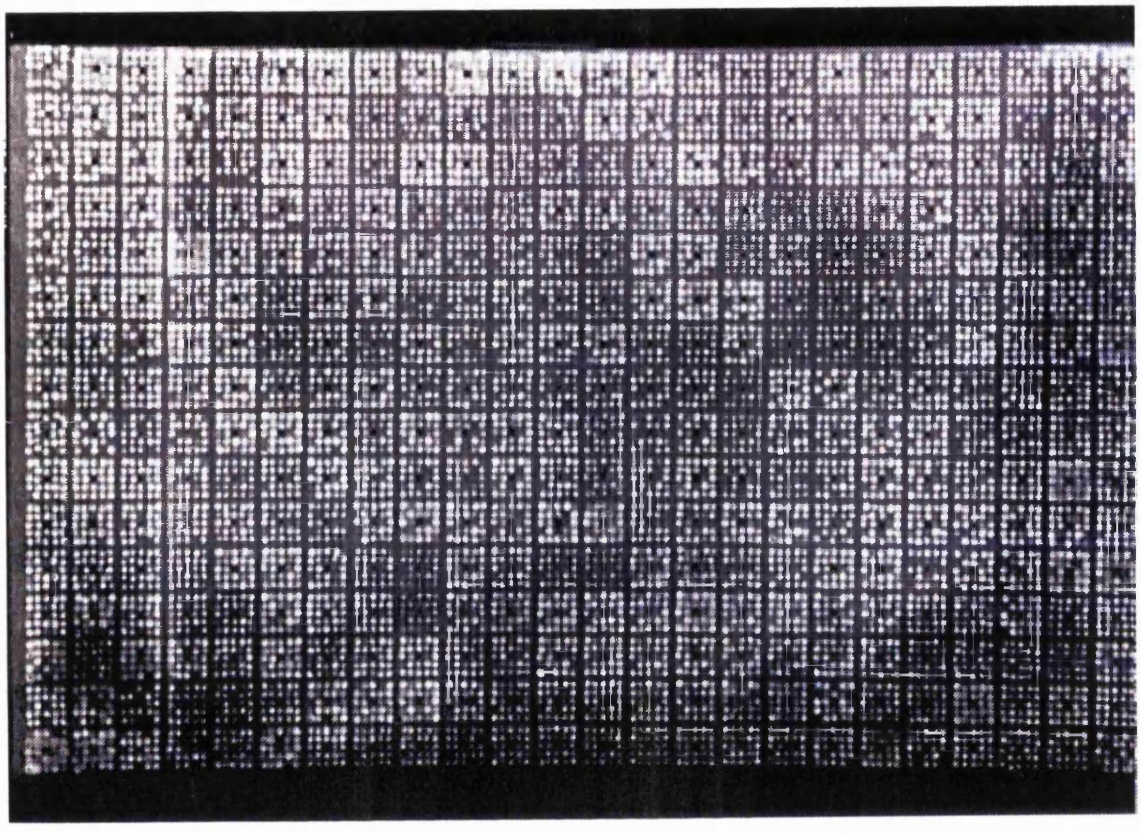


Figure 3.3

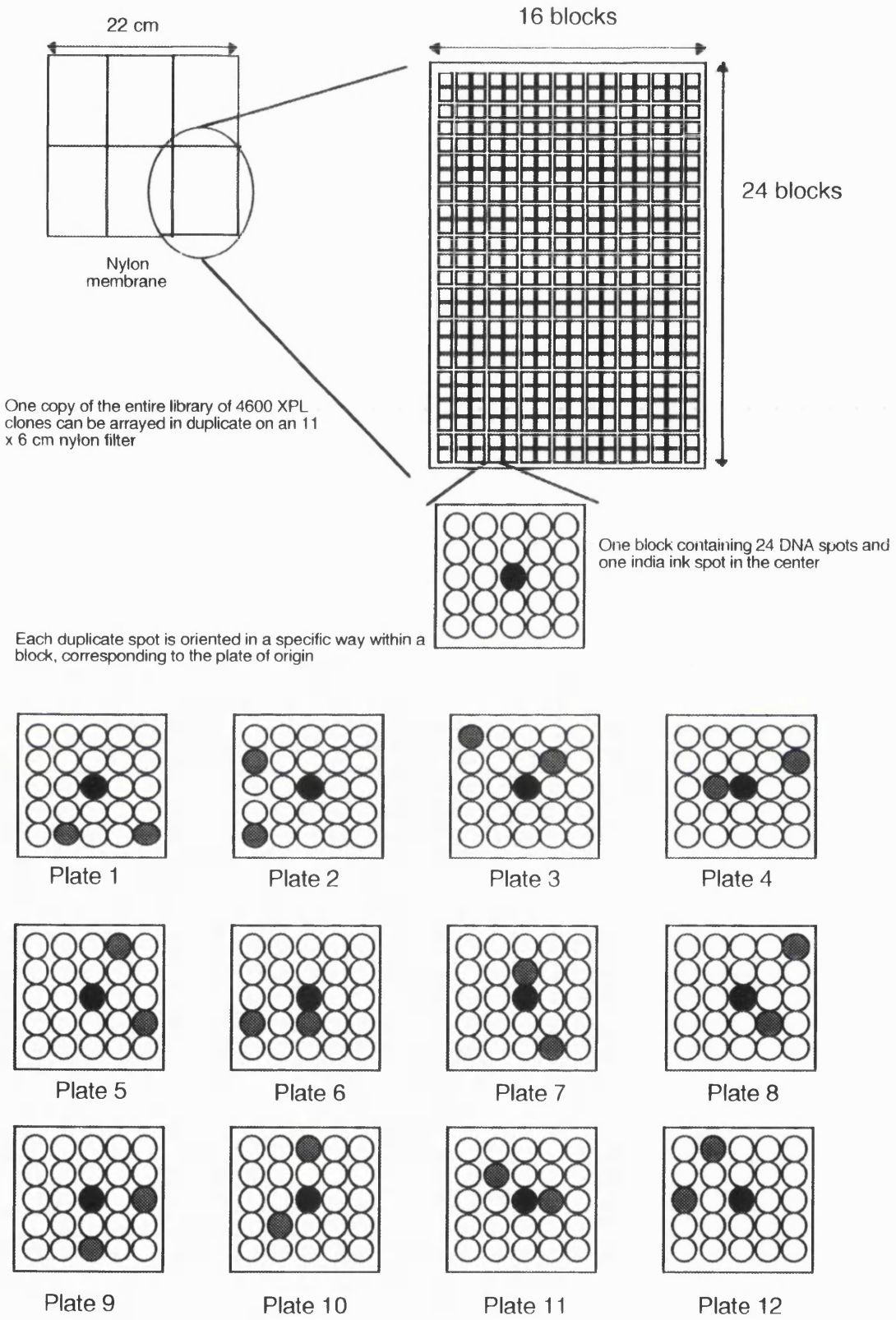

22 cm
16 blocks
24 blocks
Nylon membrane
One copy of the entire library of 4600 XPL clones can be arrayed in duplicate on an 11 x 6 cm nylon filter
One block containing 24 DNA spots and one india ink spot in the center
Each duplicate spot is oriented in a specific way within a block, corresponding to the plate of origin
Plate 1
Plate 2
Plate 3
Plate 4
Plate 5
Plate 6
Plate 7
Plate 8
Plate 9
Plate 10
Plate 11
Plate 12

Figure 3.4

In this format, a total of 4108 individual XPL clones could be spotted on one small filter (11 x 7 cm) or more than six times the minimal number of clones required by the ideal model described above. Therefore each YAC clone of mean size 475 kb should hybridise to an average of 12.2 XPL clones. This figure was judged sufficient to counter-balance most factors such as redundant cloning and uneven Alu sequence distribution, and to allow small YAC clones to find a target in the XPL library.

3. Construction of a library of X chromosome Alu PCR products

3.1 Preparation of insert DNA

3.1.1 Cell line

The choice of cell line 578 as a source of human X chromosome DNA was driven by the fact that it had been well characterised previously in the context of the establishment of a whole X chromosome linkage map (Wieacker et al., 1984) and by members of our laboratory in a project focusing on mapping in the DMD region. It was therefore known to contain a single X chromosome on a hamster background, and FISH experiments performed in the above projects showed that the X chromosome was complete. DNA was prepared from cultured cells, and used as template in Alu-PCR reactions.

3.1.2 Primers

The X chromosome YAC mapping project relied at several stages on the Alu PCR technique and early on it was decided that for the hybridisation fingerprinting part described here, the same primers would be selected as had been used previously. These were Ale1 and Ale3, first used by Cole et al. 1991 and designed to amplify outwards from the end of Alu sequences, therefore producing inter-Alu PCR products. In an attempt to increase the complexity of the pool of PCR products to be cloned from cell line 578, the use of primers specific for the L1 repeat element was also tested.

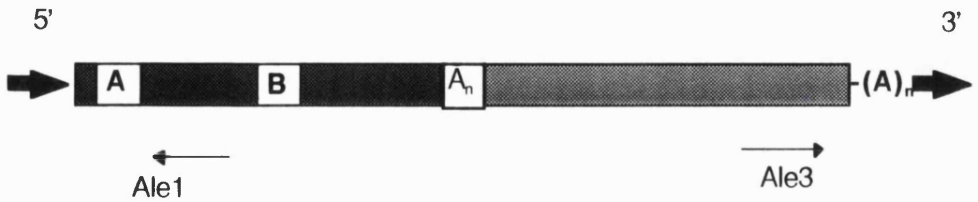


Figure 3.5 Schematic structure of the human Alu repeat and position of the Ale1 and Ale3 primers. The human Alu sequence is about 300 bp long and consists of two

directly repeating monomer units. The first on the left hand side, of approximately 120 bp, contains two internal RNA polymerase III promoter sequences (A and B), and is separated from the second one by an adenine rich region. The second monomer contains a 31 bp insert and therefore measures approx. 150 bp, and is terminated by a highly polymorphic adenine rich region. The sequence is flanked by two direct repeats represented by thick arrows (Kariya et al., 1987) (Kass and Batzer, 1995).

An essential requirement for the selection of primers is to ensure that they are human-specific, in order to avoid cloning rodent DNA from the cell line. This was tested for Ale1 since previous work in our laboratory suggested that this primer may amplify from hamster DNA. It was also tested with HL1, the primer selected from the human L1 consensus sequence, on which human specificity had never been tested. Ale1 was tested by amplifying in separate reactions hamster genomic DNA, human DNA, and a mix of hamster and human DNA in proportions similar to those present in cell line 578 (20:1) (Figure 3.5). The annealing temperature of the PCR reaction was set to 73°C, five degrees higher than the predicted melting temperature, in order to ensure that primers would specifically bind to the template DNA. Discrete bands in lane 1 show that Ale1 is not human specific. This result is in contradiction with data presented in a review on Alu-PCR fingerprinting (Parrish and Nelson, 1994) where Ale1 is indicated as being human-specific.

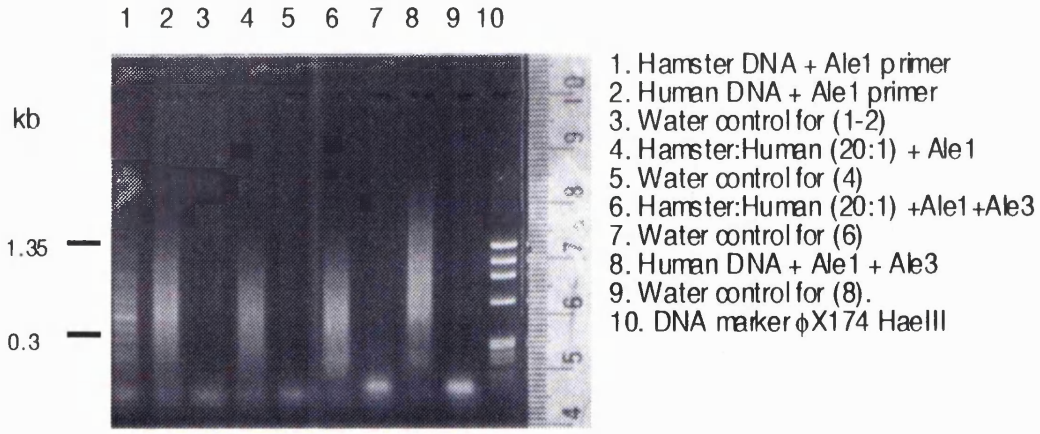
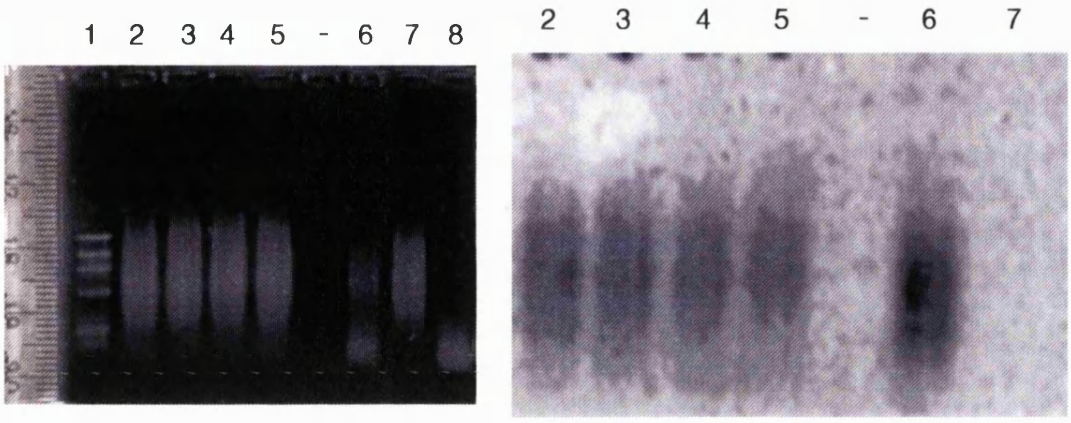


Figure 3.6. Control PCR to assess the specificity of primer Ale1. All PCR reactions were performed using 50 ng template genomic DNA. Distinct bands in lane 1 clearly indicate that this primer amplifies from hamster genomic DNA.

The specificity of the primer HL1 was tested in a different way. In this experiment, HL1 could not be tested alone, since the average distance between two L1 sequences in the human genome is often beyond the polymerisation capacity of Taq Polymerase. When mixed in equimolar proportions with Ale3, which was known to

be human-specific (data not shown), it appeared that hamster genomic DNA gave amplification products (Figure 3.7A). A gel containing these products was blotted and hybridised with labelled PCR products from sample 1 (cell line 578 amplified with Ale3 and HL1). Results show that the probe binds to products amplified from hamster DNA, but not to human DNA, demonstrating that the products present in lane 1-4 are hamster DNA amplified between two L1 sequences (Figure 3.7B).



Legend figure 3.7A:
1. DNA marker ϕX174 HaeIII
2-5. Cell line 578 + primer Ale3+HL1
6. Hamster DNA + Ale3 + HL1
7. Human DNA + Ale3 + HL1
8. Water control

Figure 3.7A Gel electrophoresis of of L1 and Alu-PCR products from cell line 578, hamster and human DNA with a 1:1 mix of Alu (Ale3) and L1 (HL1) primer. **3.7B.** Southern blot of this gel hybridised with PCR products from cell line 578 (sample from lane 1).

Reasons for the lack of human-specificity of primer Ale1 and HL1 could not be determined from the literature or the sequence databases, and this was not investigated further. It is presumably due to the ability of the primers to amplify from Alu-like and L1-like sequences in the hamster genome. It was therefore decided that Ale3 only should be used in the construction of the XPL library.

3.1.3 Size selection

In order to control the insert size of XPL clones, attempts were made to select different fractions of inter-Alu amplification products from cell line 578. Similar size selection experiments were being performed by co-workers, while constructing a

library of mouse inter-B1 amplification products (McCarthy et al., 1995). Following their approach, amplification products from 578 were slowly separated overnight by gel electrophoresis, followed by careful slicing of the corresponding gel lane. Approximately 24 slices could be made, each containing a different population of 578 amplification products based on their size. Each slice was then agarased, and after precipitation, DNA was re-amplified with Ale3 as for the primary amplification. It was however never possible to obtain the expected size gradient across the 24 samples, as expected from the experiment performed on mouse DNA. All samples showed a size distribution ranging from 3 kb to 100 bp. An effort was made to reproduce the experiments performed with mouse DNA, but it is possible that gel overloading affected the migration of DNA according to its size. The different result could also be explained by the different nature of the DNA. In particular, it was speculated that contamination with genomic human DNA in the size selected samples was difficult to avoid, due to its omnipresence in the laboratory environment. Possible PCR controls that could have been included in this experiment to test this hypothesis would have been the inclusion of a sample of digested agarose from the gel used for size selection, outside of the DNA lane, as well as starting buffer from the gel tank. However, these problems were not investigated further after a preliminary cloning experiment showed that clone sizes were representative of the distribution of insert DNA sizes without prior size selection.

3.2 Cloning of Alu-PCR products

3.2.1 Choice of vector

Two different vectors are available from GIBCO BRL which are suitable for cloning with the Uracil DNA Glycosylase system, depending if the PCR products are to be cloned directionally (pAMP1) or non-directionally (pAMP10). In the case of Alu-PCR products amplified with a single primer, DNA fragments are not locus specific and it is therefore not necessary to clone the products directionally. Directional cloning is in fact less efficient since only 50% of the PCR products have the correct combination of primers and are suitable for annealing to the vector. However in this project, a high cloning efficiency was not required since insert DNA was not limiting and a relatively small number of clones was necessary. Consequently, in this particular case either pAMP1 or pAMP10 may be used and since pAMP1 was available at the time in our laboratory, it was used for the construction of the XPL library.

3.2.2 Annealing and cloning

Several annealing reactions were set up, with varying vector:insert ratios within the range suggested by the manufacturer. Each ligation was electroporated at

different ligation:cell ratio in electro-competent DH5α cells, and 20 random colonies were re-amplified by Alu-PCR for each ligation test. Cloning efficiencies were variable (between 2.10^6 and $8.2\ 10^6$ cfu/µg insert DNA) but did not appear to be related to the above parameters (insert:vector or ligation:cell ratios). Each PCR was checked on an agarose gel in order to assess the size distribution and check for absence of cloning artefacts (e.g. cloning of primer-dimers). In about 15 % of the cases, more than one band appeared in a given lane, which might indicate cloning artifacts. However, the amplification of multiple fragments from a single clone was not reproducible, indicating that this phenomenon is likely to have been due to a PCR artefact. To test this, six clones which gave several bands after Alu-PCR were chosen and re-amplified several times. Each clone showed at least one band (the one of higher molecular weight), and inconsistently one to two additional bands of lower molecular weight. It was suspected that one or two additional sites could be present within the inserts, complementary for Ale3 with the exception of perhaps one base that would create a mismatch in stringent PCR conditions, but allow annealing of Ale3 in non-stringent reactions. Since the test PCR reactions described here were performed with excess primer and template, it is possible that Ale3 could anneal in the first cycles of the reaction to one of those sites, and produce additional bands of smaller size. To verify this hypothesis, plasmids were prepared by alkaline lysis from the six clones and digested by NotI and SalI to release the insert. In each case, a single insert of the expected size was present, demonstrating that the presence of multiple bands in those clones was not due to cloning artefacts, but to PCR artefacts. Since the DNA that was amplified in such PCRs all originated from one insert, it was presumed that such clones would not be a source of error in future experiments involving the XPL library

Annealed material that showed the best cloning efficiency and insert size distribution were electroporated on a larger scale and clones were robotically picked into 384 well microtitre plates. A total of 5376 clones were picked, in 14 384-well plates. Several replicas were prepared from the original and stored at -70°C.

3.3 Alu-PCR and filter preparation from the XPL library

A master mix was prepared for 5000 Alu-PCR reactions of 50 µl each, and manually aliquoted in 12 polypropylene 384 well plates. Colonies from each well of the XPL library were then transferred with a 384-pin plastic device to the PCR plates, which were heat sealed with a sheet of acrylic and placed in a mobile cage inside the PCR robot. The waterbath PCR machine was then programmed for 35 cycles, with 3 minutes denaturing time at 96°C and 6 minutes annealing-extension time at 68°C. A single temperature was chosen for both the annealing and the extension steps since, in the case of the Ale3 primer, the difference between the two (2°C) is within the

accuracy range of the thermostat of the PCR robot (+/-1°C). In a benchtop commercial thermocycler (e.g. PTC-100, MJ Research) conditions for the same reaction would be 94 °C for 30 sec, and 68°C for 30 sec, for 30 cycles. Extended times were programmed for each step in the waterbath PCR robot, due to the volume of plastic (plates) and metal (cage to hold the plates) that was transferred between each waterbath, which had to be equilibrated to temperature in addition to the reactions themselves. After the cycling steps, 126 random samples were selected across the library and electrophoresed on an agarose gel to assess the quality of the amplification (Figure 3.8). Approximately 9% of the clones did not produce an insert, either due to a cloning or to a PCR artifact. This was considered sufficiently low that it would not hamper the use of the library.

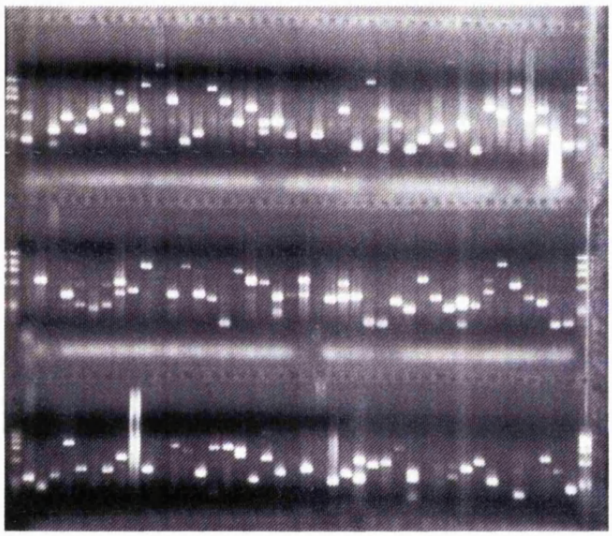


Figure 3.8. Sample of 126 XPL clones randomly selected from the library after PCR amplification in the robotic waterbath PCR machine.

The PCR products were then spotted on nylon filters as described in paragraph 2 (Strategy), after which plates were stored at -20°C and re-used when new spotted filters were required.

4. Hybridisation of X chromosome YAC Alu PCR probes

This section describes the hybridisation of YAC Alu-PCR products to XPL filters, in the context of the construction of an X chromosome YAC map that involved a number of people in our laboratory.

The main method used to detect overlaps between YAC clones was the direct hybridisation of YAC Alu-PCR products to entire YAC libraries spotted on nylon filters, in the form of Alu-PCR products. The fingerprinting technique described here was essentially used as a complement, to confirm overlaps between the YAC clones and support the contigs. Probes to be fingerprinted were therefore selected from these contigs, according to a number of criteria: YACs situated at the end of contigs, YACs which together represented a minimum span of the contigs, and YACs for which additional evidence was needed to confidently place them in a contig. After selecting 560 YAC clones from the contigs according to the above criteria, a DNA sample from each clone was transferred from the cX library to a new microtitre plate, and amplified by Alu-PCR using the primer Ale3. A sample of each PCR reaction was then electrophoresed on an agarose gel to verify the presence and quality of the products (Figure 3.9).

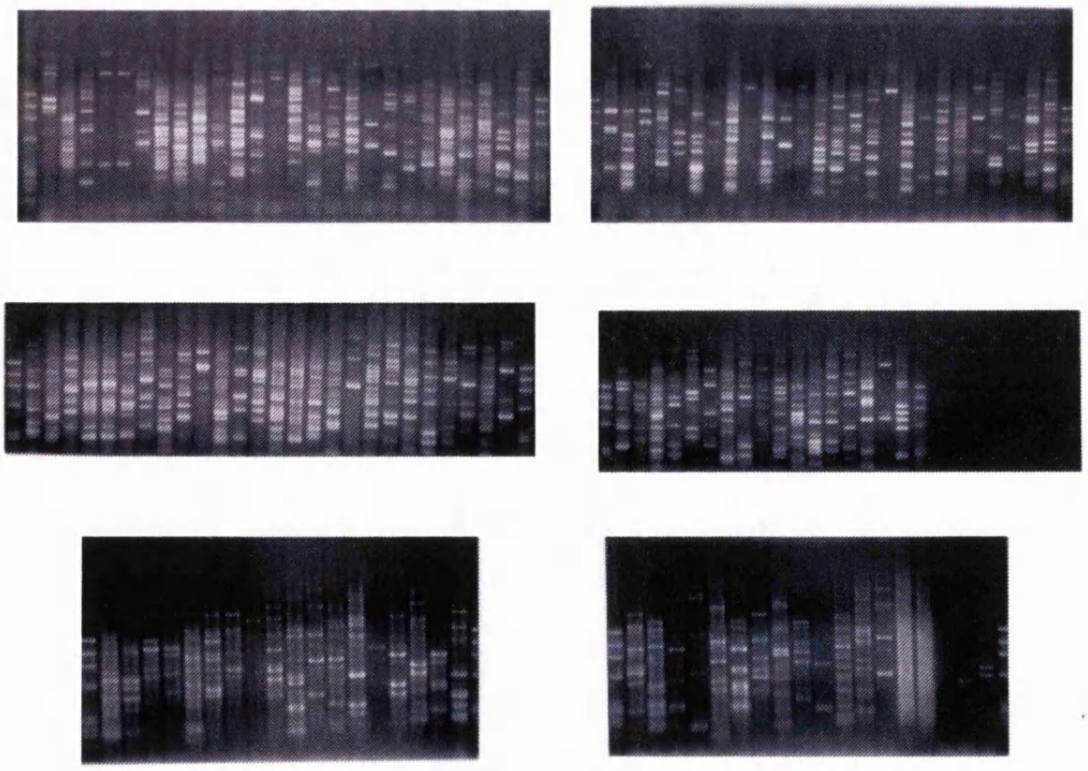


Figure 3.9. Sample of 126 YAC Alu-PCR reactions separated on agarose gels by electrophoresis.

To increase the throughput of the work described below, several new techniques for binding DNA to solid supports and new detection methods were investigated, and this part is described in chapter 4. In the course of this project, different labelling and detection methods were used, which include:

- incorporation of radioactive nucleotides in the PCR products by random priming, followed by direct detection on X-ray films.
- attachment of a molecule of biotin to the Ale3 primer prior to the amplification of the YAC clones, and detection of hybrids by the binding of a streptavidin-alkaline phosphatase conjugate (Str-AP). The substrate for the AP was Attophos.
- incorporation of digoxygenin (DIG) labelled nucleotides during the PCR reaction, and detection of hybrids by the binding of an anti-DIG antibody conjugated to alkaline phosphatase (AP). As above, the substrate for the AP was Attophos.

Regardless of the hybridisation system used, YAC Alu-PCR products were competed with sheared total human DNA in order to block repetitive sequences, which were mostly contributed by the Ale3 primer and the 3' end of the Alu sequence (approximately 80 bp on each product). The protocol used for this was based on studies performed by Sealey et al. (1985), which have shown that pre-reassociation of Alu sequences was optimally carried out at Co x t = 100, at 68°C in 0.18 M sodium, where:

Co = concentration of driver DNA (total human here) in mg/ml

t = time during which the pre-reassociation is performed, in minutes.

Independently of the hybridisation and detection methods used, results consisted of a set of positive XPL clones, each represented by a duplicate signal on a filter. The first YAC probes that were hybridised to the XPL library were ICRFy900A0472 and ICRFy900B1239. These were selected from a contig that was constructed in the course of a positional cloning project (Francis et al., 1994a) performed in our laboratory. A strong overlap between the two clones had therefore been previously demonstrated by a number of experiments including single-copy probe hybridisations and restriction mapping, and therefore was a good test case for the fingerprinting experiments. The two clones were previously shown to measure approximately 650 kb and to overlap across most of their respective lengths (~600 kb). Results obtained with the fingerprinting method show that ICRFy900B1239 and ICRFy900A0472 hybridise to 18 and 16 clones respectively, and share 14 of them. This result is highlighted in figure 3.12. The proportion of shared clones (70%) is close to the expected value of 85% based on the extent of the overlap. A second test was performed that involved two YAC clones which overlap across a smaller region. The two YACs (ICRFy900A0201 and ICRFy900G0732) had been characterised in the course of a second mapping project on the X chromosome across the ZFX-POLA loci (Francis et al., 1994b), and were known to span 800 kb, 350 kb of which was the

overlap region (45%). Each YAC was fingerprinted by hybridisation to the XPL library and identified 22 XPL clones in total, 14 out of which were shared (63% of the total). Both test hybridisations were in good agreement with expected figure. This encouraging result suggested that the fingerprinting method could reliably detect overlaps, and was therefore used on a larger scale.

A total of 196 YAC probes were hybridised to the XPL library in the course of the project, identifying 1332 XPL clones. In parallel to the last 50 hybridisations, XPL clones were also labelled and hybridised back to the XPL. Figure 3.10a shows the result of 3 radioactive YAC Alu-PCR probes hybridised separately to three XPL library filters. The 3 probes identify several XPL clones in common, indicating that all 3 overlap. Figure 3.10b and 3.10c show examples of YAC Alu-PCR probes detected using fluorescence (biotin and digoxigenin systems respectively). Different YAC Alu-PCR probes were used in each case.

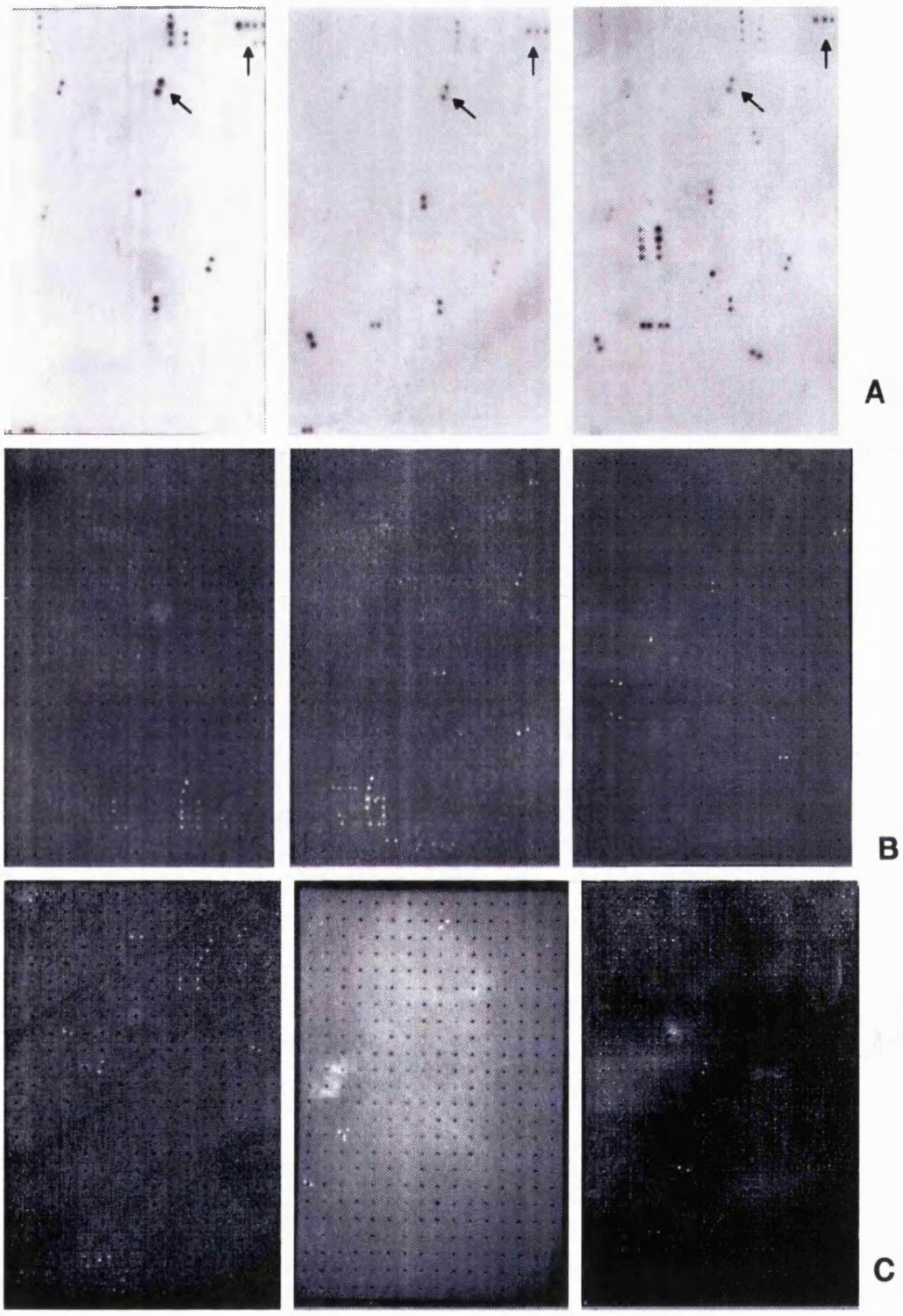


Figure 3.10 Example of 9 different YAC Alu-PCR probes fingerprinted by hybridisation to the XPL library. The three probes in 'A' were radiolabeled, in 'B' were biotinylated and in 'C' were DIG labeled. The three YACs in A share a number of positive XPL clones (arrows) therefore demonstrating an overlap. This example shows a gradient of hybridisation quality in B and C, ranging from clear and unambiguous signals (e.g. B right or C middle) to hybridisation of perhaps more repetitive probes, leading to more background (e.g. C right)

5. Analysis of results

Considering the number of YAC probes used and especially the number of XPL clones that were identified, computer software were necessary to process and analyse the data. As hybridisations were performed, results were transferred to a Unix workstation and formatted in the well2clone environment (see chapter 2, ß 4.2). Data files were processed by probeorder each time new hybridisation results were entered. The logfile produced by probeorder provided a number of statistics and suggested possible links between contigs, and were used to monitor the progress and detect possible inconsistencies. A Postscript file displayed the contigs graphically, in a way that highlighted potential connections between different contigs, and also assisted in detecting inconsistencies. Figure 3.11 shows a section of the log file that is produced by probeorder.

Figure 3.11 Section of the log file produced by probeorder and showing the contig containing the two YAC probes ICRFy900A0472 and ICRFy900B1239 first hybridised to the XPL library as a test for the experimental approach (underlined and indicated with arrows). The first line indicates the contig number, its size (no. of probes) and if known, its location on the chromosome, based on probe mapping information (here FISH mapping and marker content). The probes are listed below, with a series of 5 numbers and a summary of the available position information. The numbers after the probe include the distance to the next probe in the contig (as used in the annealing), the numbers of clones linking the probe to the next, and the number of clones spanning the probe. The list of fitted XPL clones (e.g. ICRFp600H218) are printed underneath, with their hybridisation results to the probes as a matrix. For example, the first XPL clone only hybridises to the first probe (ICRFy900A01126), while the second hybridises to the first and the second probe. The number 3 indicates the intensity of the hybridisation signal (3=strong, 2=medium, 1=faint), and a dot indicates a negative result. A # sign means that this clone hybridises only to this YAC probe in the entire dataset. The last section shows a list of possible connections from this contig to others, based on hybridisation results between a clone from the contig above to probes belonging to other contigs.

```
contig   47 size:    6 location:  Xp22.2 - Xp22.3  FISH ( SP )

! probes:

ICRFy900A01126  42 12 13 19 37
ICRFy900B1239   38 14 18 20 30 mapped at Xp22.2 - Xp22.3  FISH ( SP )
ICRFy900A0472   29 13 16 17 24 mapped at Xp22.13  rfh 25.17-28.48
CEPHy904B03906  18  3 14 15 13 mapped at Xp22.13  DXS274 @24.5 ( e )
hhmi7c2         78  3 14 59 78
CEPHy904D10892                 mapped at X 0.0  AFM203wf4 (DXS993) ( I )

! length: 1.82 clone lengths (0.91)

! fitted clones:

! ICRFp600H218     3.....#
! ICRFp600I053     33....
! ICRFp600B0310    333...
! ICRFp600H214     333...
! ICRFp600H1512    33.333
! ICRFp600F0611    333333
! ICRFp600I103     333333
! ICRFp600L153     333333
! ICRFp600E144     333333
! ICRFp600C2210    333333
! ICRFp600L164     333333
! ICRFp600F0411    333333
! ICRFp600F216     333333
! ICRFp600L116     .3....
! ICRFp600G162     .3....
! ICRFp600E112     .3....
! ICRFp600H109     .33...
! ICRFp600I2210    .333.3
! ICRFp600H215     .33333
! ICRFp600K0611    ..3333
! ICRFp600G039     ..3333
! ICRFp600A118     3..333
! ICRFp600H2210    ....3.
! ICRFp600N164     .....3#
! ICRFp600L055     .....3#
! ICRFp600G244     .....3#
! ICRFp600L035     .....3#


! connections to other contigs:

! ICRFp600H214 (3) -> contig   52 probe ICRFy900B08146    Xp22.13  rfh 25.17-28.48
! ICRFp600H109 (3) -> contig   23 probe hhmi30D7
! ICRFp600K0611 (3) -> contig   52 probe hhmi29E3         XP22.33  BC 5513.76 X
P22.3 * FISH ( NC ) CORRECTED FROM XQ22.3 18/2/94
! ICRFp600A118 (3) -> contig   34 probe ICRFy900E1127
! ICRFp600A118 (3) -> contig   76 probe ICRFy900B0824     Xq28.0  FISH ( RV )
! ICRFp600H198 (3) -> contig   23 probe hhmi30D7
! ICRFp600J0111 (3) -> contig   36 probe ICRFy900B04163
! ICRFp600M142 (3) -> contig   52 probe ICRFy900E0101
```

Figure 3.11

In the probeorder contig building procedure, an order of YACs along the chromosome was suggested to the program, but not forced upon it. This order was given to the program as a list of YAC clone names ordered from Xpter to Xqter, based on experimental data such as FISH mapping, marker content, or radiation hybrid mapping data (see chapter 5). The result summarised on figure 3.12 shows a diagonal with Xpter towards the top left corner, and Xqter towards the bottom right corner. When the data indicates an overlap between two YACs, their respective list of positive XPL clones share at least one element between them. This can be seen when the blocks corresponding to two or more YAC probes overlap horizontally. In all cases, the strength of the overlap increases with the number of XPL clones shared between two probes. For instance, probes 10 to 15 show a very strong overlap as well as probes 114 to 117, while the overlap between probes 54 and 55 is suggested by only one shared XPL clone. A number of scattered XPL clones are visible outside of the diagonal on figure 3.11, representing approximately 5,5 % of the total (75 out of 1332). These are clones that have been identified by two YAC clones that belong to different contigs. The distance between an XPL clone and the diagonal is proportional to the distance between the two contigs to which it belongs. In an ideal situation such clones should be absent, and their presence can be taken as a measure of the 'noise' in this experimental dataset. Out of the 196 YAC probes that were hybridised, 133 (~68%) identified at least one XPL clone, and were therefore useful for contig construction by probeorder.

Figure 3.12 Probes are aligned horizontally on the top and bottom axis. Due to space constraints, the names of the positive XPL clones are not indicated, but otherwise would be listed vertically on the left and right hand side of the drawing. When a YAC clone identifies an XPL clone, a black box is drawn, the shade of grey indicating the intensity of the signal (black, strong; grey, medium; light grey, faint).

clones: 1332 probes: 133

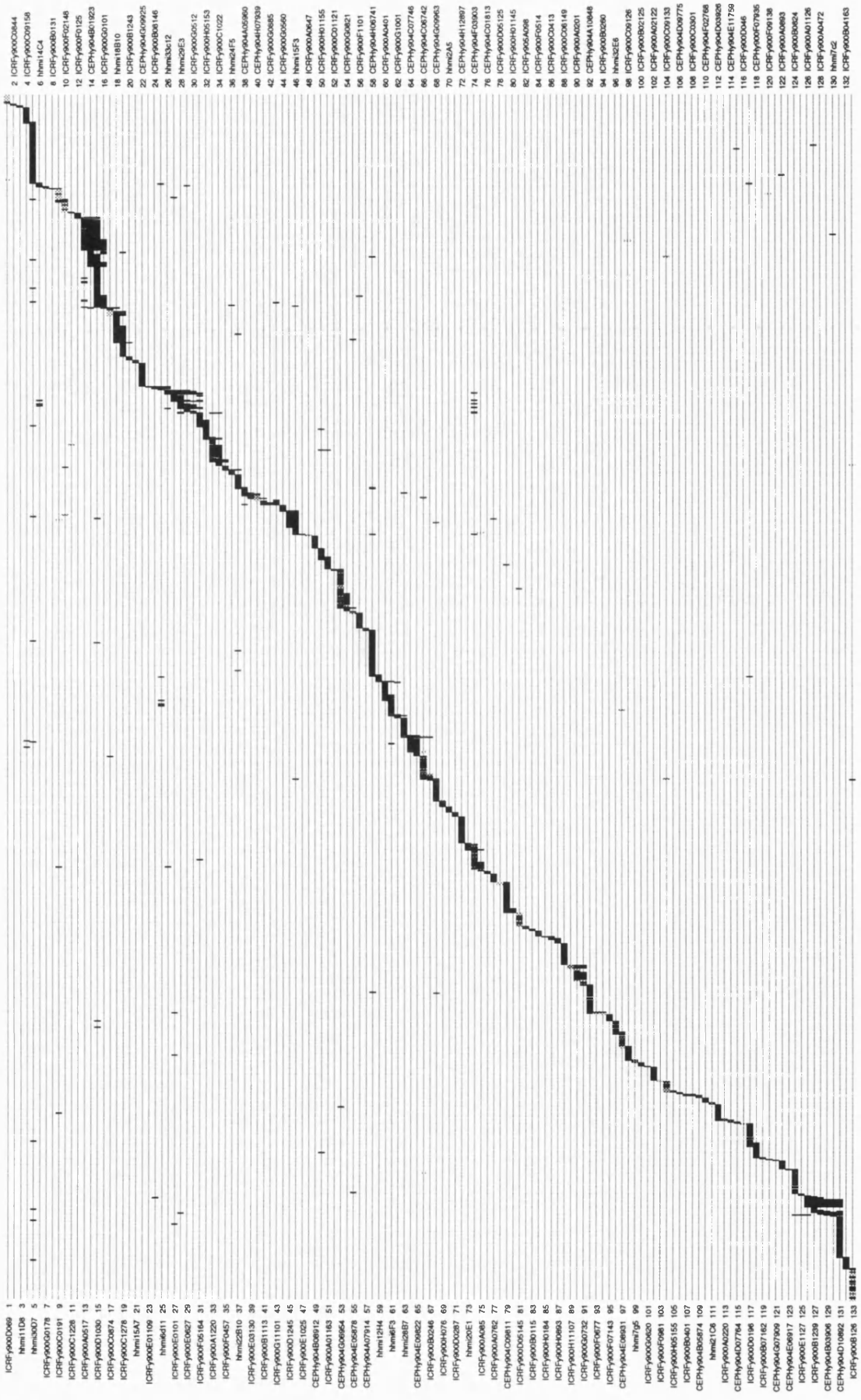


Figure 3.12

The distribution of the hybridisations is shown in figure 3.13a and 3.13b. On average, each YAC probe hybridises to 10 XPL clones, ranging from 0 to 97. Most probes (75%) hybridise to less than 20 XPL clones, and about half (52%) to less than 10 XPL clones.

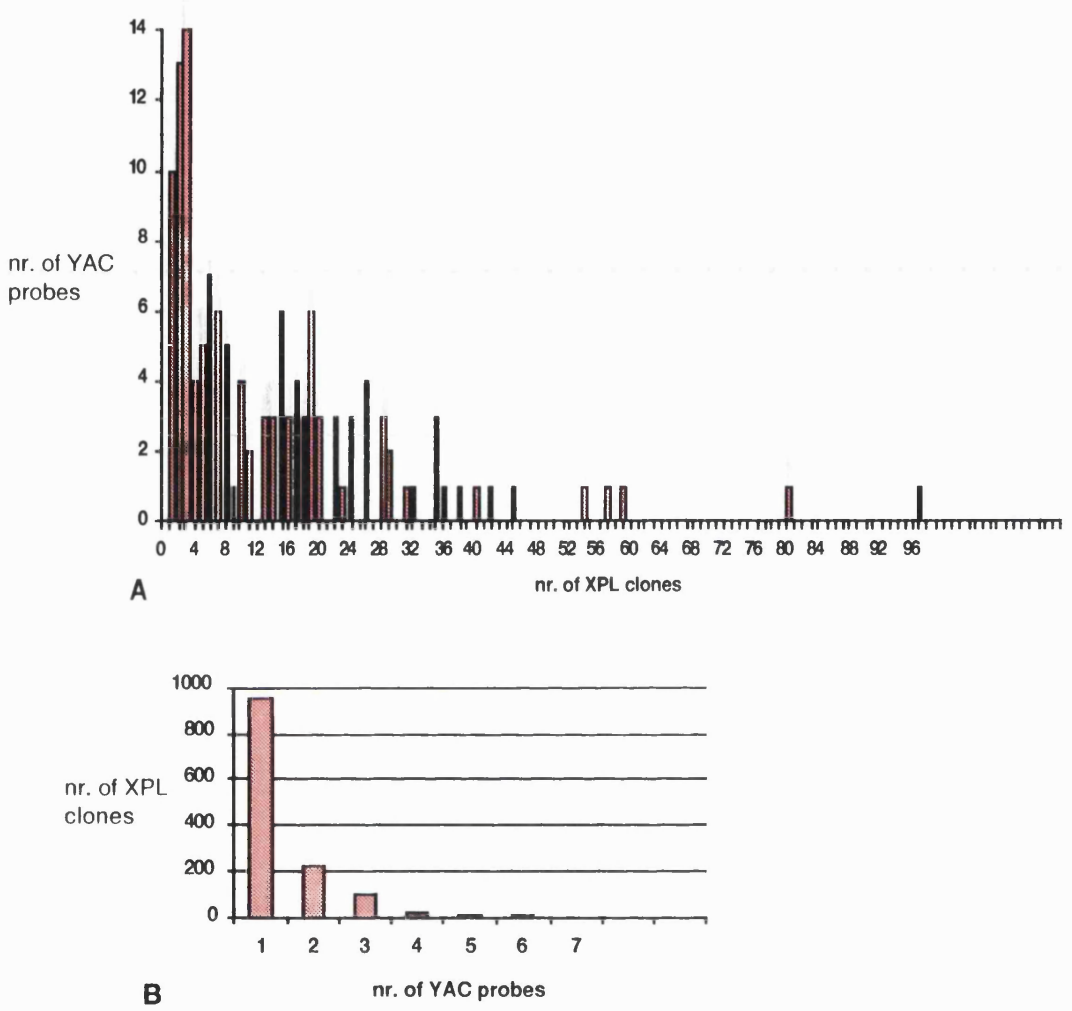


Figure 3.13. A: Graph representing the distribution of hybridisation results, indicating the number of YAC probes (Y axis) that hybridise to any given number of XPL clones (X axis). A single YAC hybridises to 97 XPL clones, while 14 probes (11%) hybridise to 3 XPL clones (highest peak). **B:** Graph representing the number of XPL clones that hybridise to any number of probes, ranging from most clones (952, or 71%) hybridising to a single YAC, to one XPL clone that hybridises to 7 YAC probes.

In parallel to the last 50 YAC hybridisations, XPL clones were also individually hybridised back to the XPL library. Hybridisations to the XPL library enabled the assessment of its redundancy, by providing a measure of the number of identical

clones. Out of the 50 XPL clones randomly selected from the library, 8 did not identify any positives, and the 42 remaining ones identified 225 clones (average 5.3).

As described in chapter 5, overlaps between YAC clones that were generated as a result of this project were used in a more complex analysis of various experimental data. In this analysis, YAC clusters were first assembled by probeorder based on YAC to YAC hybridisations, and refined into contigs using FISH, marker content, gel fingerprints and the results described in this chapter. The latter detected overlaps between 113 YAC clones, and left 20 YAC clones as singletons.

6. Discussion and conclusions

The hybridisation fingerprinting approach described in this chapter did not play a major role in the construction of a YAC contig map of the X chromosome. Its impact was minimised mainly by the small number of hybridisations performed compared to the number of YACs necessary to build the map. It was crucial in two instances to close gaps that could not otherwise be bridged, and in general provided confirmatory evidence for linking YAC clones together, which was its major intended role. In this project a very large number of direct YAC to YAC hybridisations were performed, making it very rare to find two clones that overlapped by hybridisation fingerprinting but not by direct YAC to YAC hybridisation.

About 32 % of all YAC clones did not hybridise to any XPL clone. This can be explained by two factors. First, it is likely that among the ~3000 clones constituting the cX library from which YACs were selected as probes, a certain percentage were not from the X chromosome. In fact, about 30% (89 out of 301) of all YACs that were mapped by FISH during the X chromosome mapping project were assigned to one or more autosomes, and had no position on the X. It is not possible to correlate these two figures directly since there may have been a bias towards selecting non-X clones for FISH mapping due to conflicting data and the need to confirm their position. A second reason may be a low representation of X chromosome Alu-PCR products in the XPL library. Hybridisation of 50 XPL clones back to the XPL library indicate that it is about 5 times redundant. This results implies that although the library contains 4600 clones, they represent only ~900 unique X chromosome Alu-PCR products. This is approximately 1,5 times more than the minimum number which would theoretically be sufficient to detect all overlaps (see § 2.2, this chapter). However, this may not be sufficient in regions poor in Alu repeats such as G dark bands. About 70% of all XPL clones have not been identified in hybridisations, leaving a pool of approximately 600 unique Alu-PCR products that remain to be positioned.

The small number of hybridisations that were performed did not make use of the full potential of this fingerprinting technique. However, the method has proven that it can reliably detect overlaps between YAC clones, in a way that is very complementary to direct YAC to YAC hybridisations. The latter are extremely efficient

to detect a maximum of overlaps with a minimum of effort. A single hybridisation will in principle detect at once all possible overlaps between YAC clones that are arrayed and the YAC clone used as probe. It does not however provide any indication as to the strength of the overlaps. In contrast, the fingerprinting method only detects overlaps indirectly. It also requires that all YAC probes from a given collection are hybridised to the XPL library in order to detect all possible overlaps between any given YAC and all others. In the results described in this chapter, each YAC clone detects on average 10 XPL clones and an overlap is represented on average by 4 shared clones. This means that the information is distributed over several positive clones, instead of one for a YAC to YAC hybridisation, hence limiting the risk of misinterpreting the results.

An advantage of this method over other fingerprinting techniques based on fragment separation in gels for instance, is that each cloned Alu-PCR product is readily available as new marker for further studies. This is particularly useful for constructing higher resolution maps with clones of smaller insert size than YAC clones. With the availability of robotic spotting machines, the most efficient way of approaching such project is first to hybridise YAC probes to arrayed libraries such as PACs or cosmids, in order to rapidly identify and map clones of interest. In a second step, XPL clones that are already mapped to the YAC contigs provide a source of hybridisation probes to confirm overlaps between YACs and underlying PAC/cosmid bins. A beneficial side effect of this strategy is that overlaps between the clones propagated in *E. coli* start to be refined as well.

CHAPTER FOUR: Development of alternative hybridisation systems

1. Introduction

Labelling and detecting nucleic acids is a fundamental method in molecular biology. The use of radioactive isotopes (^{32}P, ^{33}P, or ^{35}S) has been predominant in this technique and it is probably still so today. The main reasons are the simplicity, high sensitivity and robustness of experiments such as hybridisations in which this method is an essential part. The major disadvantages associated with radioactive isotopes are hazards to health and environment, extensive safety regulations and costs for handling, and short half-life time of the probes. These factors have stimulated numerous attempts to develop alternative systems with equal sensitivity and convenience, without the above drawbacks. Many have been partially successful, and have since been commercially available (e.g. the DIG system from Boehringer Mannheim, or the ECL system from Amersham). Most hybridisation results published in the literature today are still performed with radioactive probes, thus demonstrating that none has reached the performances of isotopic labelling methods. A shift in the application of hybridisation (and hence labelling and detection) methods is however noticeable today. The increased use of automated methods in the Human Genome Project and the strong support it has enjoyed from funding agencies have yielded huge amounts of resources in biological reagents such as ESTs, genomic clone libraries, genetic markers, radiation hybrid panels etc. Automation and miniaturisation are therefore increasingly necessary in order to efficiently make use of these resources.

Many processes routinely performed in a laboratory working on the Human Genome Project are repetitive. This is especially the case when large scale experiments are planned, for instance when work is carried out at the level of entire genomes or entire chromosomes. One aspect of the project described in this thesis is an example, since it involved the hybridisation of large numbers of YACs covering most of the human X chromosome to a clone library. The repetitive operations in this case are the hundreds of hybridisations which had to be performed in identical conditions. Like many other processes, such as picking and spotting of clones, and PCR amplification (Lehrach et al., 1990) (Meier-Ewert et al., 1993) (Hudson et al., 1995), hybridisations could be at least partially automated.

Different systems have been developed in this direction on a small scale and for specialised applications (Alderton et al., 1994) (Cherry et al. 1994). At the start of

this project the standard hybridisation protocols in use in our laboratory were the random-prime labelling of long DNA molecules (Feinberg and Vogelstein, 1983), followed by their hybridisation to template DNA which had been bound to nylon filters either by Southern blotting, robotic arraying or *in situ* lysis of robotically arrayed colonies. Two major aspects of this protocol had so far hampered the development of an automated system. The first is the handling of the nylon membranes which, due to their soft and loose texture, need extra manual care at several stages in the hybridisation process. The second is the hazard associated with isotopes. The most commonly used is ^{32}P, which requires that all manipulations are performed behind at least 1 cm thick perspex screens in a contained working environment. This chapter describes work which attempted to address these two issues. First, the use of a rigid support such as glass was explored as a possible material to bind DNA molecules. A different approach, consisting of the permanent attachment of a standard nylon membrane to a sheet of acrylic was also investigated. Second, two non-radioactive labelling systems were tested and optimised, based on the fluorescent emission of positive signals.

2. Rigid supports for binding DNA

2.1 DNA attachment to polyacrylamide coated glass plates

2.1.1 Introduction

A good immobilisation support for DNA should have a high binding capacity, be durable and should not interfere with hybridisation and quantification processes. The use of radioactive probes generally solves the problem of background interferences, and nylon has proved to be an excellent support for such probes, meeting all of the above criterias. At a time when fluorescent detection of hybridisation signals was being investigated in this project, this was however a concern since nylon was the standard immobilisation support and nylon has a strong inherent fluorescence. Glass is a rigid support that shows virtually no fluorescent background. An added advantage of glass over nylon is its rigidity, which would make automated manipulations straightforward. Previous research in this field (Khrapko et al., 1991; Khrapko et al., 1989) had shown that glass coated with a thin layer of polyacrylamide was suitable for binding short oligonucleotides (8-mer). The attachment of longer DNA fragments (100-1200 bp) is tested here, and the system is evaluated with regards to its suitability as a support for repetitive hybridisations.

2.1.2 Binding tests

Glass in itself does not bind DNA efficiently, and therefore must be chemically modified. The first modification involves the formation of a stable amide bond between glass and an acryl group, by treating the glass surface with 3-(trimethoxysilyl)propyl methacrylate (Bindsilane)(figure 4.1).

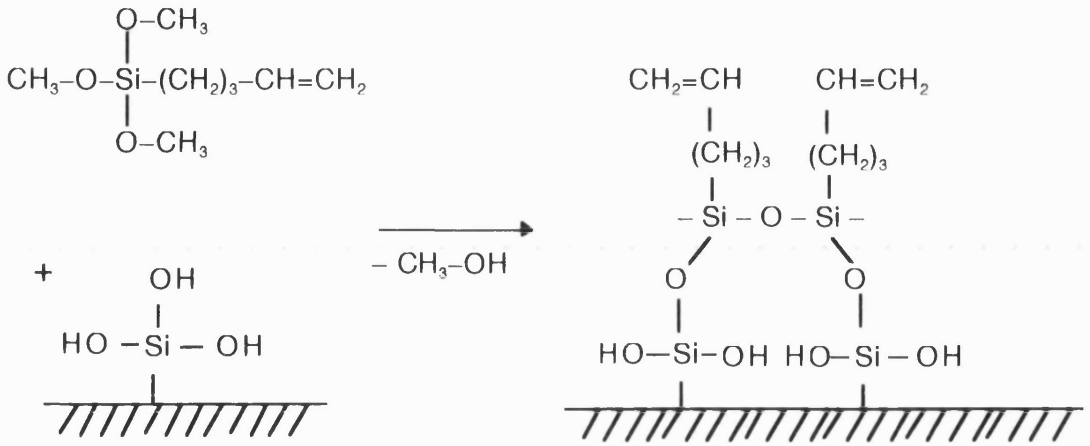


Figure 4.1 Chemical modification of the glass surface with Bindsilane creates reactive groups that will form covalent bonds with the polyacrylamide layer

Secondly, a thin layer (30 μm) of 8% acrylamide, 30:1 bisacrylamide, ammonium persulfate ($Na_2S_2O_8$) and TEMED (as for PAGE gels) is polymerised on this activated glass surface. The polyacrylamide matrix is then activated by a short treatment with hydrazine, which substitutes some amide groups with hydrazide groups (Figure 4.2)

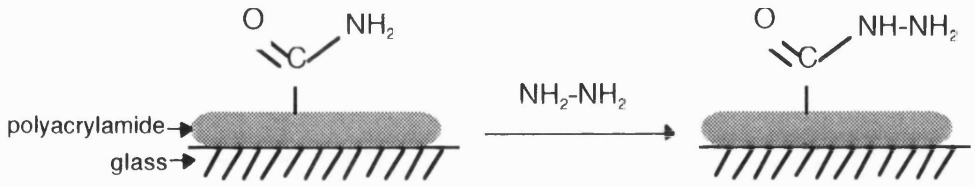


Figure 4.2 Amino groups on the surface of the polyacrylamide layer react with hydrazine to form hydrazide groups.

DNA does not contain groups which are expected to react effectively with hydrazide groups, and therefore requires to be activated prior to binding to the glass plate. Glutaraldehyde has been known to efficiently cross-link proteins, most probably

via amino groups (Richards and Knowles, 1968). The reaction is generally fast (less than 1 hr at room temperature) and requires approximately 1% (9mM) glutaraldehyde or less. More recently, glutaraldehyde has been used to cross-link DNA with proteins such as histones (Kuykendall and Bogdanffy, 1992) or biotin (Al-Hakim and Hull, 1988). It was therefore considered that glutaraldehyde may also efficiently cross-link long-chain DNA molecules to a polyacrylamide support carrying highly reactive hydrazine groups, and tests designed to verify this hypothesis are described below.

Experiments were performed in order to first compare the binding capacity of a polyacrylamide glass plate with and without glutaraldehyde, and to compare the results with the performance of a positively charged nylon membrane such as Hybond N+. Different dilutions of a 100 bp radiolabelled PCR product amplified from an XPL clone were spotted in triplicate on an polyacrylamide slide with and without prior treatment by glutaraldehyde (0.25 %), and on a positively charged (Hybond N+) nylon membrane, without activation but with U.V. cross-linking. The remaining radioactivity of each dot was then measured after successive washes of increasing strength, using a Minimonitor 125 ("Victoreen") equipped with a home-made counts integrator (I. Ivanov), through a 5 mm aperture in a lead screen. The counts were measured during 15 seconds for each DNA spot, and the average between three triplicate spots was plotted against the intensity of the washes (figure 4.3). After the last wash, the three supports were autoradiographed for 1 hr (figure 4.4).

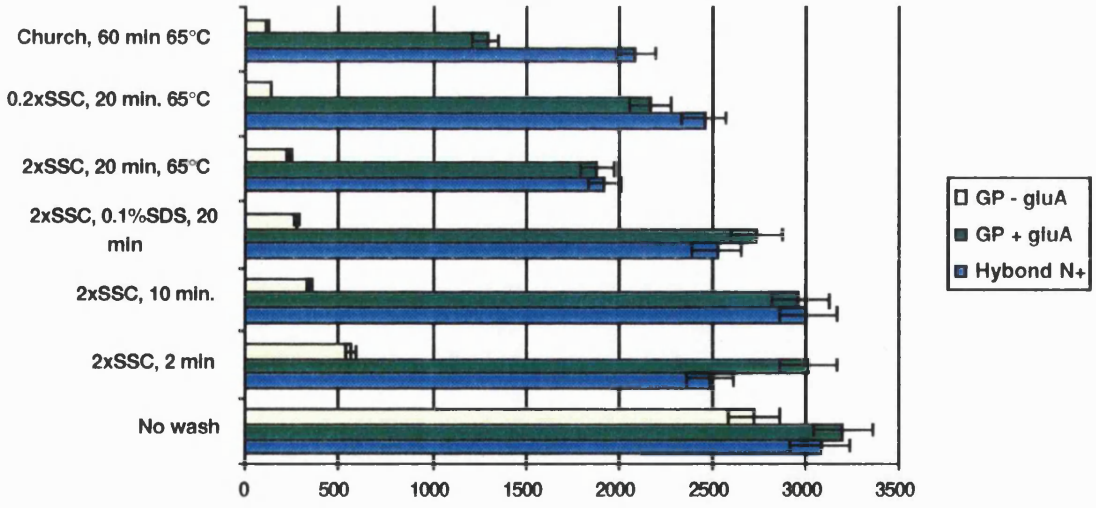


Figure 4.3. Chart indicating the amount of remaining radioactivity (disintegrations per 15 sec.) (X axis) on the three different supports after successive washes of increasing strength (Y axis), for approx 1 ng of DNA initially spotted. (Abbreviations: GP=Glass Plate, gluA=Glutaraldehyde)

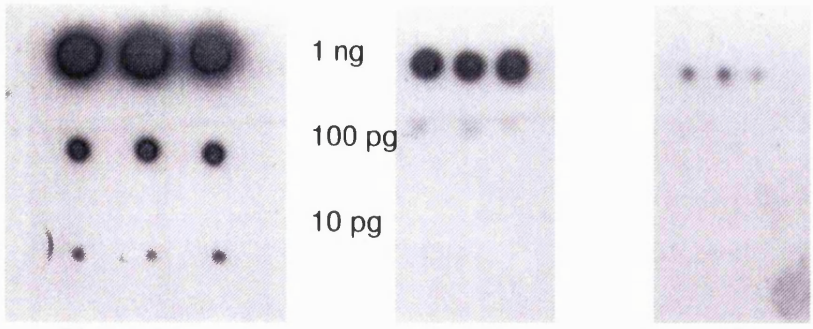


Figure 4.4

Figure 4.4. Autoradiogram of the 3 supports after the last wash of 1 hour in Church buffer at 65°C.

Results show that a 100 bp DNA molecule can be attached to a polyacrylamide matrix and withstand strong washing conditions, up to those encountered in hybridisation experiments. The binding capacity of the polyacrylamide layer is comparable to that of the nylon membrane used as reference, until the Church wash at 65 °C after which a sharper loss of DNA is observed, compared to the Hybond N+. The results also indicate that the binding of DNA to the polyacrylamide layer is due to the action of the glutaraldehyde, since 80 % of the DNA which was not treated is lost after the first mild wash, compared to 6% of treated DNA.

Next, the amount of glutaraldehyde sufficient to cross-link DNA to the polyacrylamide layer was tested since unwanted modifications in the DNA molecule may occur which could affect its hybridisation properties. In particular if too many adenine and cytosine bases (which carry the amino groups likely to react with glutaraldehyde) are affected and used for binding to the glass plate, this may prevent DNA:DNA hybrids to form. It was therefore necessary to test the binding of DNA with reduced amounts of glutaraldehyde. A 0.5 ng/µl solution of the labelled 100 bp XPL clone PCR product was incubated for 10 min with different amounts of glutaraldehyde, ranging between 2% and 0.0675 %. One nanogram of each sample was then spotted in duplicate on an acrylamide coated glass plate, and washed for 10 min at room temp with 2xSSC, followed by an overnight incubation in Church buffer at 65°C with agitation. One nanogram of the same DNA was spotted in duplicate on a piece of Hybond N+ as reference. Figure 4.5 shows a 4 hr autoradiography of the two supports.

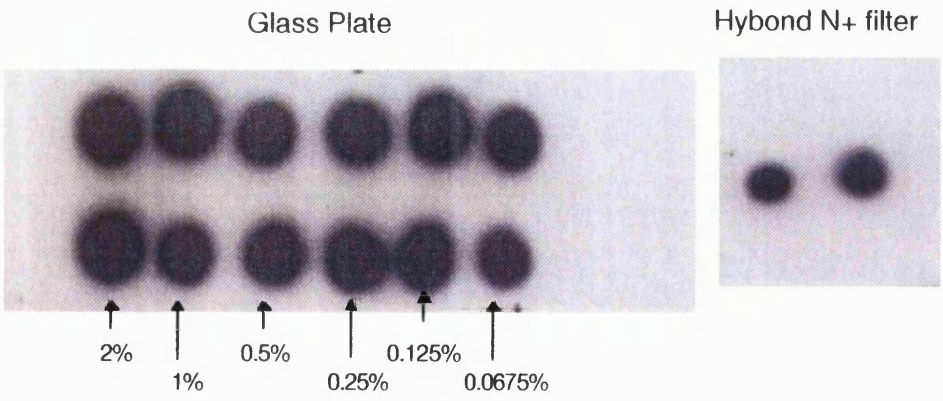


Figure 4.5. Autoradiography of an acrylamide coated glass plate, activated with hydrazine and spotted with 1 ng of a radioactively labelled 100 bp XPL clone insert incubated with varying amounts of glutaraldehyde. A Hybond N+ membrane spotted with the same DNA but without glutaraldehyde is shown on the right.

No significant variation in the amount of remaining DNA can be observed, either within the dots on the glass plate, or between any of the samples on the glass plate and the samples on the nylon membrane. It is likely that a further reduction in glutaraldehyde concentration is possible without affecting the DNA binding capacity. Nevertheless, since a concentration of 0.0675% glutaraldehyde was later shown not to affect hybridisations, lowered concentrations were not tested further. The observation that results on the glass plate and the nylon membrane are comparable contrasts slightly with those stemming from the previous experiment, where already after 1 hr of wash at 65°C in Church buffer, a marked difference is observed between the 1 ng spots on the glass plate and the nylon membrane. A hypothesis which could explain this difference is based on the mechanic fragility of the polyacrylamide matrix. It is conceivable that the upper part of the layer is more fragile that the part closer to the glass plate, which is also covalently bound to the glass. A relatively short wash in Church buffer would rapidly remove the fragile upper part of the polymer, including the bound DNA, while this effect would not be visible on the more resistant nylon membrane, resulting in the marked difference between the DNA spots in figure 4.4. However, a 12 hr wash in Church buffer would have a continuous scrubbing effect on the DNA bound to the nylon membrane, while the damage to the polyacrylamide matrix would to a large extent cease after the uppermost fragile part is removed. Therefore after approximately one hour, DNA is not further removed from the glass plate while it is continually washed away from the membrane, resulting in the latter catching up with the former. This was not investigated further, since ultimately this observation does not detract from the result that under hybridisation conditions, an activated polyacrylamide layer has the same ability to bind a 100 bp PCR product compared to a positively charged nylon membrane.

The next step was therefore to assess the properties of the acrylamide coated glass plates as a support for hybridisations.

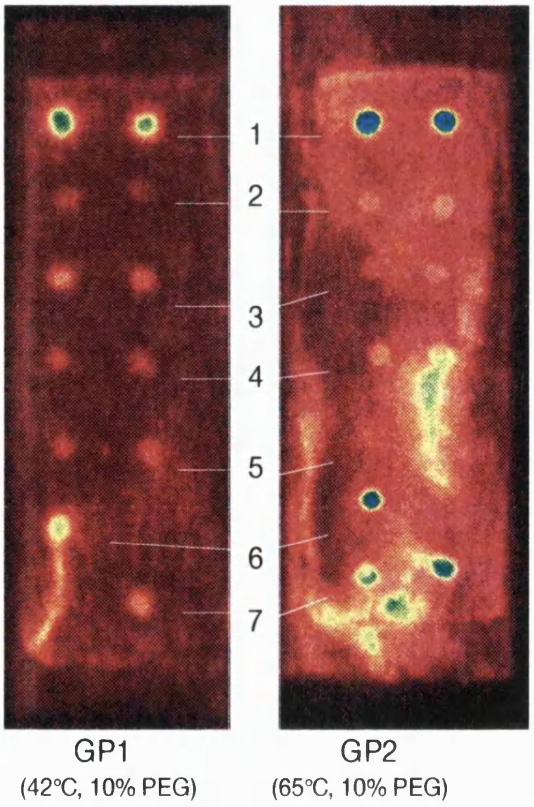
2.1.3 Hybridisation tests

In the tests described above, where only the binding capacity of the glass plates was evaluated, it was noticed that all layers of polyacrylamide were fragile to a certain degree after a prolonged incubation at 65°C in a buffer containing SDS. The edges of the layer were sometimes rubbed away, and the polyacrylamide was easily scraped off the glass plate by an inadvertent movement. It was therefore expected that during the hybridisation tests, where the plates would remain several hours in hybridisation buffer (5% SDS, 0.5 M sodium phosphate), this effect would be very detrimental to the lifespan of the glass plates. An adapted hybridisation buffer was therefore assessed, based on previous studies with such supports. Khrapko et al (1991) suggest the use of 1M NaCl, 10 mM sodium phosphate pH7.0, and 0.5 mM EDTA as hybridisation buffer for short probes (up to 20mer). The difference in sodium concentration compared to the buffer used so far on nylon membrane is not significant for the kinetics of the hybridisation (Meinkoth and Wahl, 1984). SDS is used as blocking agent on nylon membranes , and was therefore not considered necessary in hybridisations on polyacrylamide. Therefore the above buffer was used as a basis for the following experiments. In order to minimise the effect of prolonged exposure of the polyacrylamide layer to a 65°C temperature, two further modifications to the standard hybridisation protocol with nylon membranes were made. First, the addition of polyethylene glycol 6000 (PEG) was tested. PEG is known to accelerate the rate of renaturation of a probe to immobilised nucleic acids (Amasino, 1986). This effect is probably due to the exclusion of probe molecules from the volume occupied by the polymer, resulting in an effective increase in probe concentration. If the kinetics of the hybridisation can be enhanced, then the time during which the glass plates must stay at 65°C can be reduced in the same proportions. Amasino (1986) showed that an optimal concentration of 10 % PEG reduced the hybridisation time of a probe DNA to a target bound to nylon by at least 75%, and these conditions were followed here. In a parallel hybridisation test, an identical glass plate was hybridised with the same probe, in the same buffer but at 42 °C, in order to assess the effects of the temperature on the acrylamide layer and its influence on hybridisation specificity in these conditions.

Five different XPL clones were selected for the hybridisation tests. Approximately 1 ng of each clone insert amplified by Alu-PCR was spotted in duplicate on 2 glass plates. Also, 100 ng of Alu PCR products from cell line 578 and 100 ng of sheared salmon sperm DNA were spotted on the same supports. One clone (A8) was labelled by random priming and hybridised to each plate. Plate 1 was hybridised at 42°C, while the hybridisation on plate 2 was performed at 65°C both in

a hybridisation buffer containing 10 % PEG. Probe concentration was $0.2.10^6$ cpm/ml, and was left to hybridise for 1 hr. The first wash was very mild (2xSSC, 10 min at room temperature) in order to measure the effects of washing in a progressive manner. Plates were then exposed to a PhosphorImager screen for 1 hr (Figure 4.6).

Results clearly show that in both cases the probe hybridises more strongly to itself, compared to the other 4 XPL clones, present as negative controls. It also hybridises to the pool of X chromosome products from cell line 578. This DNA spot contains in principle a few copies of the clone A8 and the result is therefore expected. However it also contains 100 times more Ale3 sequence present at the 5' ends of each PCR product (100 ng spotted, versus 1 ng for the clones), which means that the signal seen for this DNA spot is probably not specific to the A8 insert. The salmon sperm signal is only present on plate 2, although weaker than the A8 or total human DNA spots. Overall the signals on this plate are much stronger than on plate 1, and this may explain the results seen for the salmon sperm DNA. Considering that it also contains 100 times more DNA than on the 4 XPL clones used as controls, one might expect a significant increase in background.



1= 1ng of clone A8 PCR product
2= 1ng of clone D1 PCR product
3= 1ng of clone D2 PCR product
4= 1ng of clone D3 PCR product
5= 1ng of clone D4 PCR product
6= 100 ng '578' Alu PCR products
7= 100 ng salmon sperm DNA

Probe = clone A8 PCR product
(0.2 10^8 cpm/ml)

GP1= 42 °C hybr. 10 % PEG
GP2= 65 °C hybr. 10 % PEG

Exposure= 1 hr PhosphorImager screen.

Figure 4.6. Hybridisation results to glass plate. Clone A8 was hybridised to 2 polyacrylamide coated glass plates containing duplicate spots of the same DNA (position 1), 4 different PCR products used as negative controls (position 2-5), total X

chromosome Alu-PCR products (position 6, single spot) and sheared salmon sperm DNA (position 7). Colors are false, but no change in contrast or intensity was made.

A difference in the intensity of signals and background can be seen between the two plates. Signals are stronger on the plate which was hybridised 65°C. On this plate, smears are visible around positions 7 and 4-5, which correlate with damaged areas of the polyacrylamide layer. Hybridisation signals on GP1, although indicating the same specificity as GP2, are weaker. These observations suggest that hybridising at 65°C is potentially damaging for the acrylamide layer, but also allows DNA to bind more efficiently to its target, or to reach its target more efficiently in the time.

If this type of support was to be used in large scale hybridisation experiments, then to be practical it must allow for multiple hybridisations to be performed on the same glass plate, and for DNA to be detectable when spotted with metallic pins as with a robot gadget. These conditions were tested by firstly reproducing the hybridisation as on GP2, and removing the bound probe by a 10 min incubation in 100% formamide. After checking that no probe remained by a 1 hr exposure to a phosphorimager screen (data not shown), the same probe was re-hybridised in identical conditions (GP4, fig 4.7). In parallel, a glass plate was spotted with the same DNA but instead of depositing the liquid as a 1 µl drop, the DNA was spotted with a metallic pin which carried approximately 10 nl of PCR reaction (GP3). The amount of DNA on each spot is difficult to quantitate, since it depends strongly on the amount of liquid transferred by contact between the pin and the acrylamide layer. Previous work in our laboratory indicates that the volume typically transferred with a 400 µm pin is approximately 10 nl (E. Maier, pers. comm.). Assuming this, the amount of DNA which is deposited by the pin on GP3 is therefore approximately 250 pg.

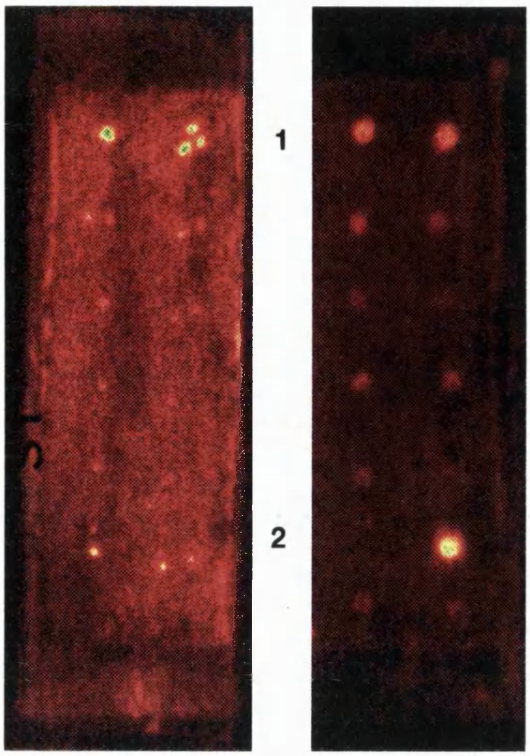


Figure 4.7 Hybridisation results on a glass plate spotted with metallic pins (GP3) and on a glass plate which had been hybridised, strip eashed and then re-hybridised. On GP3, the DNA (≅ 250 pg) was spotted with a 400 µm metallic pin. On GP4, the probe was previously hybridised, stripped and re-hybridised after removal of the probe. In both cases position 1 correspond to the probe DNA and position 2 corresponds to total X chromosome Alu-PCR products. Other duplicate spots seen in the background betweeen positions 1 and 2 correspond to unrelated XPL clones used as negative controls. Salmon sperm DNA was spotted below position 2.

Scattered signals are visible on GP3, instead of two single bright spots, at position 1 and 2, and in the background for some negative controls. This is due involuntary multiple transfers of DNA with the metallic pin, during manual spotting. Hybridisation signals on GP3 shows the same specificity as on GP1 and GP2 (fig 4.6). Only the DNA used as probe is strongly visible while the 4 other control DNAs are almost in the background. Again the X chromosome Alu-PCR product pool is clearly visible. This result suggests that if DNA is be robotically spotted using a 384 pin gadget (400 µm), the amount of DNA bound to the acrylamide layer would not be limiting. Results on GP4 however reveals a strong limitation for the use of such supports in successive hybridisations. After stripping of the bound probe, a second hybridisation shows that the intensity of the positive signals decreases dramatically compared to an identical hybridisation on GP3 and GP2.

2.1.4 Conclusions

The above experiments were designed to bind DNA fragments to a solid support made of an activated polyacrylamide layer attached to a glass plate, and to detect this DNA by means of a two phase hybridisation with a radio-labelled or fluorescent probe. The intention was to compare the properties of such a support with those of a positively charged nylon membrane (Hybond N+, Amersham), and assess its potential for large scale hybridisations. Results show that this support can bind a 100 bp DNA fragment, treated with glutaraldehyde, as efficiently as a nylon membrane

within the limits of one hybridisation, but that performances as a hybridisation support decrease below an acceptable level afterwards. Although the reagents required to prepare a glass plate are cheap, the time and manual care needed to prepare each support mean that it cannot for this application be considered as a 'single use' hybridisation template. Moreover, since the spotted DNA is first amplified by PCR, extra costs would be involved due to the increased amounts of Taq polymerase required for preparation of single use hybridisation template. Finally the fragility of the polyacrylamide layer would inevitably lead to damages during the multiple steps where the supports must be handled manually. For these reasons, the possibility of using glass plates as a replacement for nylon membranes was rejected. This does not detract from the fact that glass plates can be used as a support for hybridisation in other applications, and that different binding chemistries should be investigated. In the specific context of this thesis project however it proved unsuitable and was not investigated further. Instead alternatives were developed using nylon fixed to a rigid support.

2.2 Attachment of nylon membranes to sheets of acrylic

Attaching nylon membranes directly to a rigid support seemed an attractive idea since it would combine the exceptional binding properties of nylon for DNA, with the advantage of a solid support in repetitive manual handling, and perhaps in automatic handling. Previous attempts in this direction had been performed in our laboratory, with the use of solvents to dissolve the surface of polypropylene sheets before applying the nylon membrane. In this approach, while the solvent evaporates, the polypropylene re-polymerises, gradually embedding the nylon fibres into the plastic. However prolonged incubations at 65°C in a solution containing SDS easily removed the membrane from the support. Experiments were therefore performed to investigate this approach further.

Chloroform is used in many DNA purification protocols and therefore does not affect the structure and integrity of the DNA molecule. It is also known to efficiently dissolve acrylic (perspex). For instance, chloroform is often used to bind two pieces of acrylic by first dissolving the surfaces to be bound, in the construction of simple laboratory equipment. Chloroform was therefore first selected to test different rigid polymers for its performances as a dissolving agent. Commonly available plastics were tested, although only polycarbonate (PC), polypropylene(PP) and acrylic were sensitive to the action of the solvent. Of the three polymers, only PP and acrylic could allow a sheet of nylon to be so strongly bound that only tearing the nylon would separate it from the support. The effect seemed permanent since a 30 min incubation in 100°C Church buffer did not affect the binding of the nylon to the support. Hybridisation tests were therefore performed (overnight in Church buffer at 65°C) with nylon membranes spotted with DNA samples and attached to either PP or acrylic

sheets (1 mm thick), as well as to a free nylon membrane in parallel. Results were identical between PP and acrylic bound membranes, and both showed the same signal intensities compared to the unbound nylon (data not shown). However both rigid supports showed an increased background, probably due to labelled probe DNA caught in the mesh of nylon fibres and not effectively removed during the washing steps. This was expected since by binding one side of the membrane to a support, its ability to let liquids flow through the fibres is affected. Consequently the washing steps are less efficient and this may be a cause for the increased background. However this did not affect significantly the signal to noise ratio of the positive signals, and was not considered a disadvantage of the system. One noticeable difference between the PP and acrylic supports was the 'bowed' aspect of the PP sheet after the hybridisation, while the acrylic sheet remained unchanged. As a consequence, the nylon membrane was slightly creased in its middle part, as a response to the change in shape of the support.

One disadvantage of using chloroform in this system is the high speed at which it evaporates. As a consequence, the acrylic or PP re-polymerises rapidly and often does not allow the membranes to be uniformly applied to the surface with the necessary pressure. It was found (Geoffrey Glayzer, ICRF, pers. com.) that dichloromethane (CH_2Cl_2) has a similar dissolving effect on acrylic, but with a much slower evaporation time. Its effect on DNA was tested by first attaching a nylon membrane spotted with DNA samples to a 1 mm sheet of acrylic with this solvent, and hybridising it in parallel to the same samples on a membrane fixed with chloroform. No difference could be seen by autoradiography. Therefore all high density gridded membranes carrying XPL PCR products, and used to generate results described in chapter 3, were bound to 1 mm acrylic sheets with dichloromethane.

3. Hybridisation methods

3.1 Radioactive methods

3.1.1 Probe preparation and labelling

The labelling, hybridisation and detection protocols used in this project are mostly adaptations of standard protocols used in our laboratory, and did not necessitate any particular optimisations. Probe DNA was generated by Alu-PCR from crude lysis of yeast cultures containing YAC clones (figure 4.8).

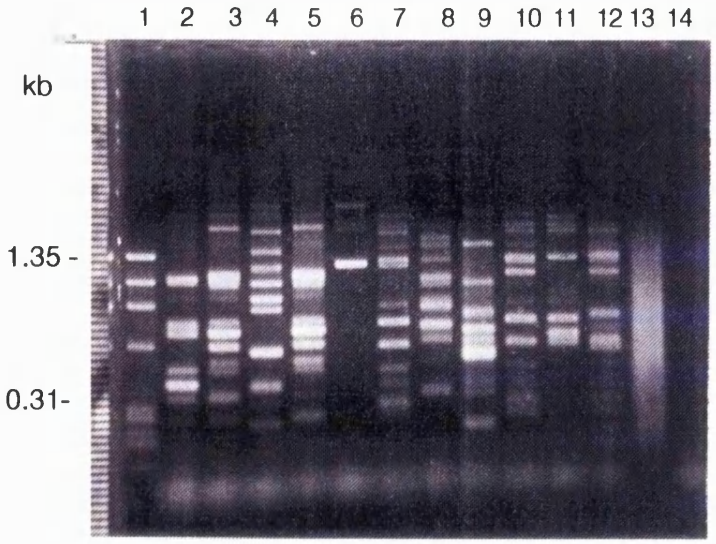


Figure 4.8. Sample of 11 YAC clones amplified by Alu-PCR using the Ale3 primer, and electrophoresed on an 1.8 % agarose gel. YACs originate from a single contig in Xp22. YAC clones 3 and 5 show a similar band pattern. Approximately 25 ng of this DNA was directly used for labelling, without purification.

Most labelling experiments involved between 10 and 20 probes and were therefore performed in 96-well microtitre plates. The DNA was heat denatured in a thermocycler, chilled on ice and the random prime labelling mix was then added to the wells. After an overnight incubation at room temperature, all probes were checked for incorporation using polyethyleneimide (PEI) chromatography paper (figure 4.9).

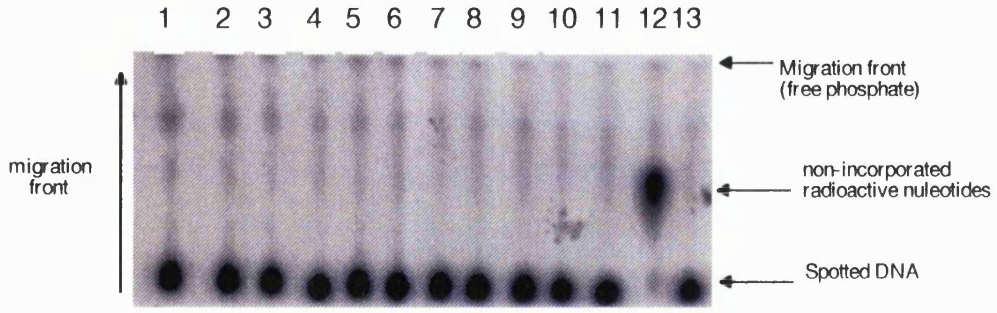


Figure 4.9. Example of 13 YAC Alu-PCR probes labelled by random priming and separated by chromatography to check the incorporation of isotopic nucleotide. In this series, sample 12 only did not label.

3.1.2 Hybridisation, detection and scoring.

In order to take advantage of the fact that membranes were fixed to a rigid support, a special hybridisation box was designed in collaboration with other members of the laboratory (L. Schalkwyk, L. McCarthy) and constructed first by the local workshop and later by a commercial company (Engineering & Design Plastics, Cambridge, UK). This box was designed to allow 48 filters of 11 x 7 cm bound to a rigid support to be vertically inserted in 48 slots. Each slot, or hybridisation chamber, was then filled with sufficient hybridisation buffer to cover the filter (10-15 ml), to which probes were then added. The box was then placed in a 65°C incubator overnight, and the filters were then pooled for all the subsequent washing steps. Detection was performed by autoradiography and the scoring was done manually. For this, a grid the same size as the membrane, printed on a transparency was placed over the X-ray films, and positive signals were marked with a marker pen. Transparencies were then scanned into a computer, and a software written in our laboratory (Acepro, H. Griffith) was then used to assign to each mark the name of the corresponding XPL clone, based on the spotting order of the clones on the membrane.

Approximately 110 YAC probes were hybridised in this way. Although this system allowed higher throughput to be achieved compared to the traditional hybridisations in plastic bags, the use of large amounts of isotope per experiment, and the manual scoring were strong limitations of this system. For these reasons alternative, non-radioactive, methods were developed which would circumvent these two difficulties.

3.2. Non-radioactive hybridisations

3.2.1 Introduction

Most non-radioactive methods for detecting a probe bound to immobilised nucleic acids work on the same model: the probe is labelled with a molecule which is recognised with high affinity by a second molecule, itself conjugated with an enzyme such as alkaline phosphatase (AP) or horseradish peroxidase (HP). The enzyme then reacts with a substrate uniformly applied to the membrane, and produces either a coloured substance visible directly (Leary et al., 1983) or a chemiluminescent molecule detected by an X-ray film or a CCD camera (Voyta et al., 1988) (Bronstein et al., 1993). Alkaline phosphatase is most commonly used because, among other attributes, of its high thermal stability (Jablonski et al., 1986). A broad range of parameters influence the choice of one type of labelling and detection techniques versus another. Radioactive labels are very sensitive and exhibit low background noise, but are expensive and hazardous to manipulate. Colourimetric methods generally use 5-

bromo-4-chloro-3-indolyl phosphate (BCIP) and/or 4-nitro blue tetrazolium (NBT) which form a purple insoluble precipitate upon action of AP. Although of lower cost, the method is not quantitative since the product partially inhibits the reaction, and precludes multiple use of the same membrane since the precipitate can not be removed. Some such substrates are also carcinogenic. Several chemiluminescent systems have been known for some time. In particular, 1,2-dioxetane compounds (e.g. AMPPD) are attractive molecules due to the long persistence of the signal, often hours or days after the addition of the substrate (Beck and Köster, 1990). The direct incorporation and detection of fluorescent dyes is not possible because the quantum yield is too low. An alternative is the use of a substrate for AP which produces a fluorescent molecule, thus amplifying the fluorescent signal. A widely used substrate is beta-methylumbelliferyl phosphate (MUF-P). The dephosphorylated molecule is optimally excited at 370 nm, and detected at 440 nm. Since the substrate itself also emits at 440 nm, a strong background would be expected in filter hybridisations. More recently however, a fluorogenic substrate called Attophos (2'-(2-benzothiazolyl)-6'-hydroxybenzothiazole phosphate or BBTP) has been developed [Cano, 1992 #1320]. Attophos yields upon hydrolysis the highly fluorescent 2'-(2-benzothiazolyl)-6'-hydroxybenzothiazole (BBT), which emits at 560 nm when excited at 420 nm. The long Stokes shift (140 nm) combined with low emission of the substrate at 560 nm makes it an ideal system for filter hybridisations. It was tested in combination with two different DNA labelling methods (digoxigenin and biotin) for its suitability in the large scale hybridisation scheme described in this project (Fig 4.10)

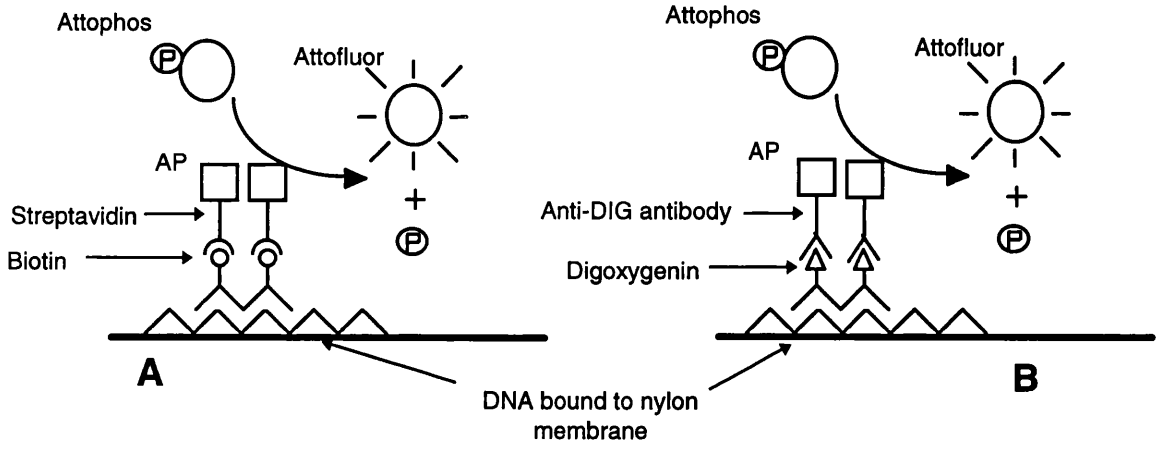


Figure 4.10. Principle of the detection of a probe bound to nylon using biotin (A) or digoxigenin (B) labelled probes. In both cases the reporter molecule is the same (AP, Alkaline Phosphatase) but it is either bound to streptavidin (A) or to an anti-DIG antibody (B).

3.2.2 Biotin labelled probes

For practical reasons, the tests designed to assess the suitability of Attophos for detecting a DNA target on high density grids were performed with biotinylated DNA probes. In the X chromosome mapping project, all probes (from YAC or XPL clones) are amplified by PCR, and biotinylated primers were easily available from ICRF Central Services, with a known percentage of attached biotin. This is important, since it guarantees a constant labelling efficiency of the probe, and eliminates a potentially variable step which would otherwise complicate the interpretation of the final results. Alu-PCR amplification of YAC clones was performed with Ale3 primers which were biotinylated at 89% (source: ICRF Central Services). It was therefore assumed that 89% of all PCR products had a biotin molecule attached to their 5' end.

The sensitivity of Attophos was first tested by spotting known amounts of biotinylated DNA on a nylon membrane, followed by its detection using streptavidin bound alkaline phosphatase. DNA amounts ranging from 5 to 50 attomoles (PCR products C1=600 bp and D1=750 bp) were spotted on a small filter with the same amounts of lambda DNA as negative controls. The detection protocol was based on (Cherry et al., 1994), apart for the optical set up. A simple glass filter with a narrow band pass at 430 nm (+/- 5 nm, Schott) was used in combination with a 100 Watt lamp mounted in a slide projector. Data acquisition was performed with a standard Polaroid camera normally used for ethidium bromide stained agarose gels, with a 530 nm cut-off filter (Kodak) placed in front of the objective. Results showed that DNA spots in the whole range of concentrations are visible one hour after the substrate is applied.

Following this, the same products C1 and D1, but non labelled and non diluted were spotted in triplicate on nylon filters also with lambda DNA as negative control. Here, DNA was manually spotted with robot gadget pins of two different sizes, which deposited approximately 200 attomoles and 70 attomoles each (250 and 90 ng respectively). Filters were hybridised over-night in Church buffer at 65°C, in standard conditions. Probes were biotinylated PCR products from clones C1 and D1, in separate hybridisation bags. Results indicate that hybridisations are specific, since 1 hour after the substrate is applied, only the DNA corresponding to the probe is detectable (Figure 4.11A). The specificity is maintained after 16 hr (Figure 4.11B), although at this stage all DNA spots are visible.

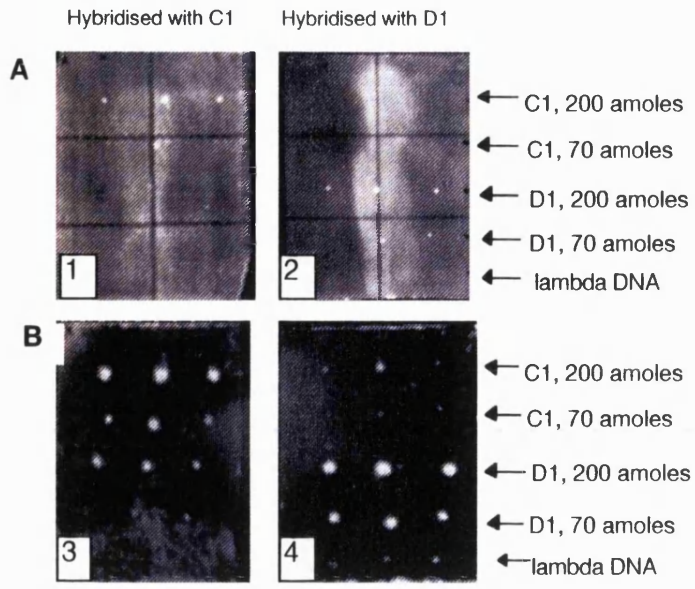


Figure 4.11. Hybridisation test. Two identical filters containing different amounts of XPL PCR products C1 (600 bp) and D1 (750 bp) were separately hybridised with C1 and D1. Photographs were taken after 1 hr (A) and 16 hr (B). All DNA spots are visible in (B), including λ DNA, but the difference between signal intensities still allows the clear differentiation between positives and negatives.

Following this positive result, more tests were performed with complete XPL filters containing 9250 DNA spots, and the same level of specificity was observed. In the course of the various test experiments, it was realised that the narrow band pass (10 nm) of the optical filters employed to excite the fluorescent molecules meant that the intensity of the incident light was very low, and this bears on the resulting intensity of the hybridisation signals. It was found empirically that 'long wave' UV light commonly used for visualising DNA stained with ethidium bromide was much more suitable. The light emitted by such U.V. bulbs follows a wide distribution of wavelengths, with a maximum at 365 nm. Although this is not optimal for Attophos (max. at 430 nm), the high energy carried by U.V. light largely compensates for the lack of specificity. The background produced by the nylon filter itself was stronger than with the narrow band-pass filter, but this was likely to be the result of the increased intensity of the incident light. In any case, the intensity of the light emitted by positively hybridising probes was at least proportionally greater. The use of long wave U.V. as excitation light was therefore adopted for all subsequent experiments involving Attophos.

Several test hybridisations were made to directly compare radioactive probes versus biotinylated probes. YAC Alu-PCR products were used, since these were to be the typical probes analysed in this project. Among the 10 probes tested with both systems, small variations between the number and respective intensities of the positive signals could be observed (data not shown). Over 90 % of XPL clones detected by radioactive probes could be detected by fluorescence. Also, 5% of fluorescent signals were not present on the X ray films. It is likely that these variations

are due to the presence or absence of a given DNA spot on one filter compared to another. Once it was established that this system produces equivalent results compared to radioactive hybridisations, an improved system was designed in order to streamline the acquisition of results and the scoring of positive signals. Together with electronic engineers (A. Ahmadi, J. Curtis), a scanning station was constructed with 2 x 6 watt long wave U.V. bulbs and a CCD camera equipped with a 530 nm cut-off filter. The complete system was enclosed in a light-tight enclosure, and a mobile platform was set to automatically fetch a filter outside of the enclosure, bring it under the camera before moving out again to take the next filter. Images were directly transferred from the computer controlling the station to a Unix platform, where images were automatically analysed. For this a program called spotter16 written by R. Mott in our laboratory was used. Initially written for a separate project, this program was modified to handle small filters with increased density (see chapter 2). Each filter could be automatically scored in a few minutes, with far less errors than with manual scoring. Results were a list of positive clones for each filter analysed, in a format suitable for analysis software that were required to continue the data processing. The combination of fast image capture and reliable image scoring was a major advantage of the fluorescent system.

Approximately 40 YAC probes were ultimately hybridised using this system. In general however, this method proved to lack some of the reproducibility that is normally associated with radioactive hybridisations. A number of probes would only successfully hybridise after several negative assays, although no parameters were intentionally changed between assays. The overall appraisal is therefore unclear. The fluorescent detection of biotinylated probes is easier and results are obtained faster than with radiolabeled probes. However the lack of reproducibility between hybridisations meant that overall at least as much work is necessary to analyse the same number of probes. It was important to find an alternative to the biotin/streptavidin system, which would increase the robustness of the hybridisations, while still allowing the use of fluorescence for the detection part.

3.2.3 Digoxygenin labelled probes

Digoxigenin, a hapten derived from Digitalis plants, is the core element of a wide range of reagents and kits commercialised by Boehringer Mannheim generally referred to as the 'DIG system' (reviewed in (Höltke et al., 1995)). DNA probes labelled with digoxigenin are recognised by an anti-DIG antibody conjugated to alkaline phosphatase, which turns Attophos into the fluorescent Attofluor. There are no obvious reasons why this system should be superior to the biotin/streptavidin system. The K_d for biotin/streptavidin interactions is approximately 10^{-15}, corresponding to an affinity 5 order of magnitudes higher than for interactions between digoxigenin and anti-digoxigenin antibodies. Based on this observation, the specificity of the

DIG/antiDIG system is not expected to be higher than for biotin/streptavidin. It has been said that streptavidin has a tendency to bind non-specifically to solid phases like nylon membranes, or that biotin can be found endogenously in biological samples, both factors which may increase background. However this never seemed to create a problem in the application described above. In contrast, the main problem was due to irreproducibility, characterised by an absence of any signals on most filters in a given hybridisation batch. Preliminary tests by other members of the laboratory showed that the DIG system had a similar sensitivity to the biotin/strepdavidin system, and it was therefore assessed in this application, and compared directly to biotin in parallel experiments. Ten YAC probes for which corresponding XPL clones were known from previous experiments were amplified by Alu-PCR using biotinylated Ale3 primer. Samples were then split in two and one half was labelled with DIG-11-dUTP by random priming, using a 35:65 ratio of dUTP:dTTP. Each set of probes were hybridised on XPL library filters and detected using the appropriate protocol, given the labelling method. Out of the 10 probes tested, one did not give any positives with either DIG or biotin labelling, 8 gave the required positives with biotin and one did not, while all remaining 9 probes gave the expected positives with the DIG system (data not shown). In this test, the difference in reproducibility is not significantly different between the two systems. However the intensity of the signals generated via the DIG system was visibly higher, enabling an easier scoring of positive clones. This may have been due to the difference in labelling methods, since biotin was only present once per DNA molecule, while DIG was incorporated in several positions by random priming. However this tendency of the DIG system to produce stronger signals was reproduced in another parallel test where YAC Alu PCR probes were generated with either biotinylated or DIG labelled nucleotides. For this reason, all further hybridisations of YAC Alu-PCR products were labelled with DIG-11-dUTP during the PCR, and detected with the DIG system. In total approximately 40 DIG labelled YAC clones were hybridised to the XPL library.

4. Discussion and conclusions

In this chapter, the use of two different supports for immobilising DNA, and three different labelling systems have been assessed for the particular application of hybridising Alu PCR products to high density grids of PCR products. The combination which proved the most convenient, sensitive and robust consisted of positively charged nylon membranes bound to sheets of acrylic, hybridised with DIG labelled probes, and detected with Attophos. It is probable that this combination would need to be modified for different applications. For instance, The XPL PCR library was arrayed on small (7x11 cm) filters, and larger membranes may not be suitable for binding via a solvent to acrylic sheets. The generation of large amounts of labelled probe is not

limiting since it is the result of an amplification reaction. Other applications may have limited sources of probe DNA, hampering the use of the DIG system which often requires a high probe concentration. However, further developments in streamlining the hybridisation protocol have since generalised its potential use to a wider range of applications. For instance in our laboratory, DIG labelled probes are now sprayed on membranes with an air gun (H. Hummerich, unpublished), avoiding the need to add probes to hybridisation bottles or bags. In addition, filters are 'laminated' on thin plastic sheets in a simple commercial heat laminator used for covering documents with plastic (D. Bancroft, unpublished). Both improvements allow for several hundred large high-density filters to be hybridised per person per week, even without automation, provided filters and probes are not limiting. Laminated filters and DIG labelled probes therefore constitute a good alternative for high throughput hybridisation schemes compared to radioactive methods. Applications are however limited, and exclude those where a quantitative result is required, or where complex probes are involved. Different avenues are being explored further, such as the use of direct labelling of DNA probes with fluorescent dyes. This would ideally require that a different support than nylon be used for binding the target DNA, due to its high inherent fluorescence. This includes attachment of the target DNA to glass plates using a different chemistry as that described above. An example of such development is the use of cDNA microarrays 'printed' on glass slides (Schena et al., 1995) (Schena et al., 1996) to measure the expression level of a complex mixture of transcripts. In this type of technology DNA is bound to a poly-L-lysine coated glass plate, and DNA probes are directly labelled with fluorescent dyes, a method which is definitely superior to isotopic labels. However here also the glass plates are single-use, a problem not yet resolved which may hamper a wider use of this technique.

CHAPTER FIVE: Construction of a YAC Contig Map of the X Chromosome

1. Introduction

YAC maps have now been reported covering most of chromosome Y (Foote et al., 1992), 21 (Chumakov et al., 1992), 22 (Collins et al., 1995), 3 (Gemmill et al., 1995), 12 (Krauter et al., 1995) and 16 (Doggett et al., 1995). Three whole X chromosome YAC mapping studies, apart from the one carried out in this thesis work, are underway. The first two are part of whole genome mapping studies at the CEPH (Chumakov et al., 1995) and Whitehead Institute/MIT (Hudson et al., 1995), where the X chromosome is poorly represented compared to the autosomes. The third study combines contigs from many groups (Nagaraja et al. 1995), and was reported at the 6th and 7th X chromosome workshops (Nelson et al., 1995)(Nelson et al. 1995). This latter map, based on STS content information, is estimated to cover 70% of the chromosome. In addition to these global efforts, numerous YAC contigs have been established in smaller regions defined by genetic mapping or by cytogenetic abnormalities. These have often been a template for the cloning of disease genes by positional cloning (recently HYP (The HYP consortium, 1995), OA1 (Bassi et al., 1995)) and sometimes have evolved into physical and transcriptional maps of larger regions (Ferrero et al., 1995). These regional efforts have often utilised common sets of markers and library clones, and it has been possible to establish "consensus" YAC maps over still larger tracts of the chromosome (Nelson et al., 1995).

The strategy most commonly employed for constructing YAC maps uses STS screening of YAC libraries by PCR. STSs are derived from genetic markers (e.g. AFM markers) or from physical landmarks such as YAC ends. When an STS is found to be contained in two different YAC clones, these are then assumed to overlap. In addition, if the STS is genetically mapped, it provides an indication of the position of the YACs. In contrast, the approach chosen in our laboratory relied on the direct detection of overlaps between clones, by hybridising YAC Alu-PCR products to complete YAC libraries (Figure 5.1).

This chapter describes the construction of a YAC contig map of the human X chromosome. The data generation was carried out by a team of people in several laboratories and various aspects have already been described in previous chapters of this thesis. The construction of clone maps at the megabase scale requires the generation of many types of experimental data such as clone overlap, position, size

and integrity. The information produced is vast and diverse, and ultimately must be combined, analysed and translated into contig maps which serve as a basis for further investigations. Data analysis for this project was carried out using a suite of programs written in our laboratory, although ultimately the map was built manually taking into consideration all available data. This chapter therefore summarises the project in its entirety (naming the individuals involved in each aspect, in the paragraph titles) and presents the YAC contig map (Roest Crollius et al. 1996).

Figure 5.1 Schema of the strategy used to construct a YAC contig map of the X chromosome. Three YAC Libraries (HHMI, ICRF and CEPH) were spotted as Alu PCR products on nylon membranes and a selection of probes from the X specific library were used for hybridisations. The positive clones were re-picked in a collection of X chromosome YAC clones (cX library) and more hybridisation were carried out with probes from the HHMI, cX and Alu PCR product library. After filtration to remove cross contamination and obvious false positives, the experimental data was combined with YAC mapping data from a separate radiation hybrid project, from the RLDB and Infoclone databases, and from FISH mapping experiments. This was done using the program probeorder, which was used to suggest contigs and to display all the information in one format. This information was combined with the gel fingerprinting data and analysed manually to build the final contigs.

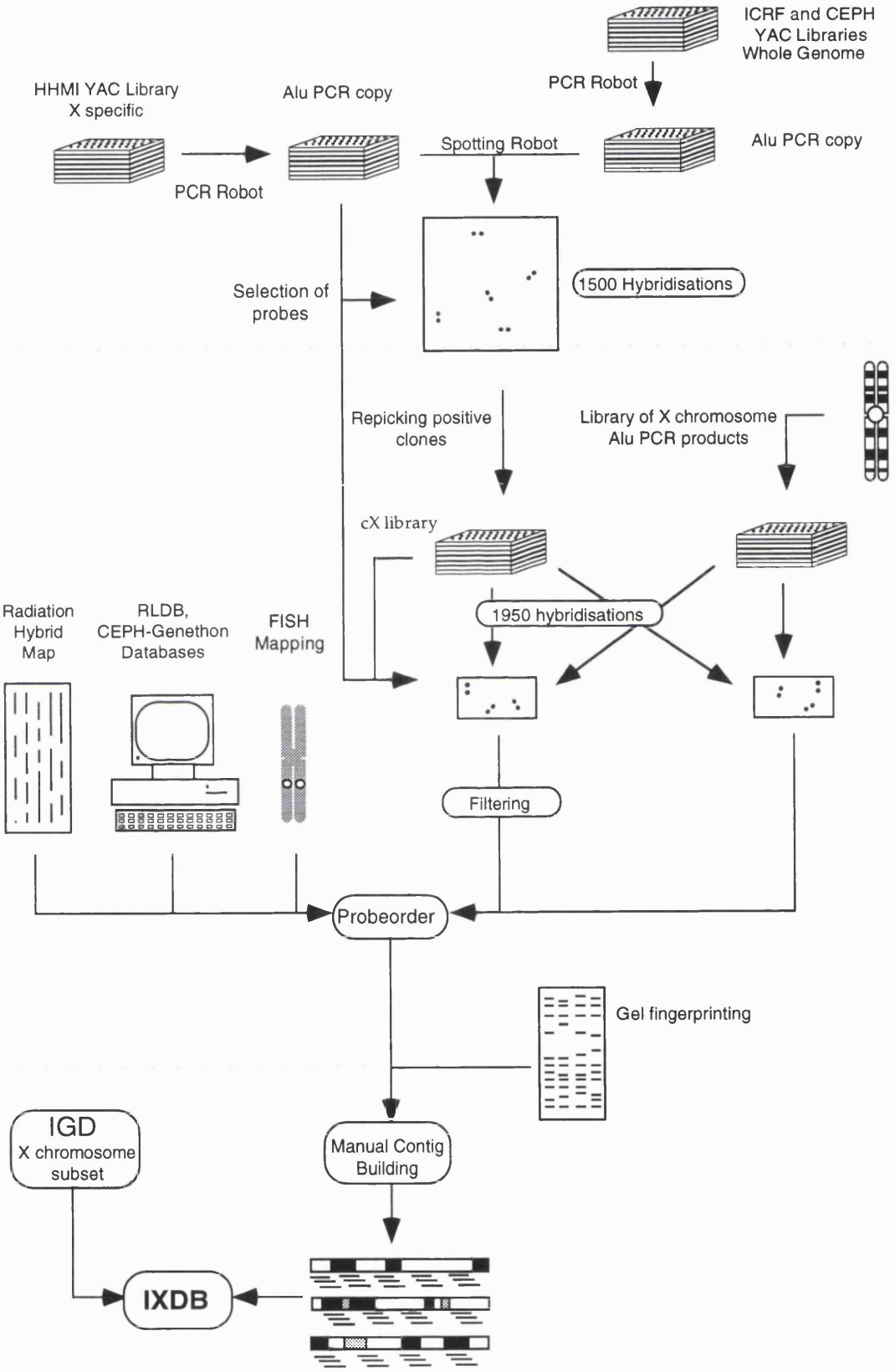

ICRF and CEPH
YAC Libraries
Whole Genome
PCR Robot
HHMI YAC Library
X specific
Alu PCR copy
Spotting Robot
Alu PCR copy
PCR Robot
Selection of
probes
1500 Hybridisations
Repicking positive
clones
Library of X chromosome
Alu PCR products
cX library
1950 hybridisations
Radiation
Hybrid
Map
RLDB,
CEPH-Genethon
Databases
FISH
Mapping
Filtering
Probeorder
Gel fingerprinting
IGD
X chromosome
subset
Manual Contig
Building
IXDB

Figure 5.1

2. Generation of YAC overlap data

2.1 Direct YAC to YAC Alu PCR hybridisations (M. Ross et al. ICRF)

Identification of X chromosomal YACs and primary YAC overlap information were derived by YAC to YAC hybridisation experiments. Multiple entry points were established along the chromosome by random probe selection from an X chromosome specific YAC library (Lee et al., 1992) (HHMI hereafter). Probes were derived from individual YACs by inter Alu PCR (Nelson et al., 1989) using a combination of the primers Ale1 and Ale3 (Cole et al., 1991). Hybridisation targets were also inter Alu PCR products, derived both from the HHMI library and from the whole genomic ICRF and CEPH "mega" YAC libraries. Together these libraries contain a theoretical fifteenfold coverage of the X chromosome (Table 1). Each of the 50,000 clones was amplified using a microtitre plate PCR robot and the products were gridded on nylon membranes in high-density arrays. The inter Alu PCR products from the YAC probes were radiolabelled and hybridised to the gridded arrays. In a total of 543 successful hybridisations out of 565 performed with HHMI YAC probes to these arrays, 3,978 different clones were identified (1,727 ICRF, 643 CEPH, 1,608 HHMI).

Library	DNA source	number of clones	X chr. coverage	Av. insert size	Reference
ICRF	GM1416B (48, XXXX)	20,000	8.9	670 kb	(Larin et al., 1991)
	OXEN (49, XYYYY)	400			(Monaco et al., 1991)
	HD1 (46, XX, homozygous for Huntington disease)	400			M. T. Ross, unpublished
CEPH 'mega'	Boleth (46, XY lymphoblasts)	23,700	4.1	1054 kb	(Chumakov et al., 1995)
HHMI	Micro 21D (Xpter-q27.3 on CHO background)	3,150	5	250 kb	(Lee et al., 1992)

Table 1. YAC libraries used in the X chromosome mapping project

The 2,370 positive clones from the ICRF and CEPH libraries were re-picked into microtitre dishes to create a collection of X chromosomal YAC clones (the cX library). Hybridisation filters were generated from the cX and HHMI libraries as described above. Inter Alu PCR products from 316 clones in the cX library and a further 124 clones in the HHMI library were hybridised to these filter arrays, thus generating additional clone overlap information.

2.2 YAC Alu PCR Gel-Fingerprinting (S. Gregory, D. Bentley, The Sanger Centre)

As data analysis proceeded 1,149 clones were selected for Alu-PCR gel fingerprinting. YAC clones were individually amplified by Alu-PCR, and the products separated on polyacrylamide gels. Fingerprints were analysed automatically by the program contigC (Sulston et al., 1988)(derived from Contig9, Sulston et al. 1988), which yields a probability of overlap based on sharing of product sizes. Subsequently, a detailed manual comparison of fingerprints was performed to confirm potential overlaps and to provide a suggested order of clones based on subsets of shared and non-shared bands.

2.3 YAC hybridisation fingerprinting

Results generated by YAC hybridisation fingerpinting are described in detail in Chapter 3. A total of 196 YAC probes were hybridised to cloned Alu-PCR products from the X chromosome (XPL library). Results were analysed by probeorder to assemble YACs into contigs based on the XPL clones they share.

2.4 YAC end mapping (M. Ross, C.J. Knight, ICRF)

After partial analysis of the YAC to YAC hybridisation data by probeorder as described below, 110 YAC clones were selected from the emerging contigs to be mapped more precisely. Criteria for selection were based on the importance of these clones for connecting contigs, where these connections were based on very few clones and therefore required additional information to be strengthened. All YAC ends were isolated by the vectorette method (Riley et al. 1992), and used as hybridisation probes to screen the cX YAC library.

3. Collection of positional information

3.1 The IXDB database

The development of the IXDB database is described in details in chapter 6. Its main purpose was to provide a repository of information collected both from the public domain and directly from RLDB participants. The ACEDB database system was chosen at an early stage due to its powerful graphical interface and to its configurability. Data description files ('models') were designed, initially by expanding from those which are distributed with the ACEDB sofware (designed for the *C.*

elegans database). At a later stage, the model files distributed by the Integrated Genome Database consortium (to which our laboratory belongs) were used and customised to handle the X chromosome data.

Information was collected from a large number of YAC clones used in this project, predominantly from two sources: the Reference Library Database (RLDB, (Zehetner and Lehrach, 1994)) and the CEPH/Généthon database (Chumakov et al., 1995). The RLDB is a repository of mapping information obtained via the distribution of many types of reference libraries (YAC, cosmid, PAC, cDNA etc.) including the ICRF and CEPH YAC libraries. The RLDB was queried for all human YAC clones previously mapped to the X chromosome, and 1,723 records were retrieved, of which 10 % had also been confirmed by secondary screening. In addition, 28 RLDB participants were directly contacted and often visited, and provided 42 contigs in candidate regions for disease genes. Although there was some overlap between these datasets, information was collected on the marker content of 1,181 clones. From the CEPH-Genethon database, 711 YAC clones associated with an X chromosome marker, and derived from the whole genome map, were retrieved.

In total, these two sources provided marker information on 3,074 YAC clones. Of these, 1,150 were also identified in our hybridisations and were used to annotate the contig map with marker information.

3.2 The radiation hybrid map (J. Kumlien, A. Grigoriev, ICRF)

A radiation hybrid map of the X chromosome constructed in our laboratory was also used for localising contigs lacking markers relative to each other. The map, comprising 72 hybrids, was constructed using 50 STSs spread evenly along the chromosome. Inter Alu PCR products from the hybrids were hybridised to filters of the cX library, and, conversely, 450 YACs were hybridised to the hybrid panel. This allowed 971 YACs to be placed with confidence in approximately 3 Mb intervals (average distance between the STSs used).

3.3 FISH mapping (C.G. See, S. Povey, U.C.L.; N. Carter, The Sanger Centre; R. Vatcheva, ICRF)

FISH mapping experiments were performed with YAC clones belonging to unanchored contigs or with clones for which confirmation was needed before placing them on the map. Out of 301 clones selected, 212 were assigned to the X chromosome, of which 48 were X-autosome chimerae. The 89 non-X clones were mainly in singleton contigs or in very short contigs which could not be linked to other contigs.

4. Construction of the YAC contig map.

Probeorder (see Chapter 2) was used continuously during the time experimental data was accumulated. Regularly, log files produced by the program were examined manually in their entirety. At a certain stage in this procedure an increasing number of inconsistencies became apparent in the YAC to YAC hybridisations, when compared to other mapping data such as FISH or marker content. In addition, the results disagreed significantly from predictions based on mapping algorithms (Grigoriev, 1993), which indicated that until 200-300 probes were hybridised, most of them should be unlinked. Probeorder was however able to construct very large clusters involving the majority of the YAC probes and clones much before that stage. In these large clusters, clones were in most cases not consistently originating from the same region on the chromosome. This result also diverged from expectations based on the apparent quality of most hybridisation results, where excellent signal to noise ratio were observed, which made the selection of positive clones an easy task. As the volume of data increased, islands of strongly linked clones gradually emerged within the giant clusters, connected together by weaker links. Such islands or 'sub-clusters' were much more compatible with FISH and marker content information, although connections between them were clearly wrong.

The first approach that attempted to solve this problem was based on two assumptions. First, it was believed that regardless of the inconsistencies observed, the volume of hybridisation results should at some stage reach a critical mass that would allow the map to be built by the contig building software. In parallel, further improvements on the latter would contribute to make the necessary amount of data as small as possible. Such improvements included the introduction of constraints in the contig building process. For instance in one variant which proved the most successful, two YAC probes were not allowed to be linked unless they shared a minimum number of positive clones in their hybridisations. This strategy only allowed strong links to be used by probeorder. When this minimum number of shared clones was set to two or three, most of the strong clusters described above reappeared, but unlinked. This allowed probeorder to list them in the order corresponding to the map position attached to clones inside the cluster. Another approach consisted of removing from the dataset hybridisation results which did not meet some pre-defined conditions. Such a condition would be that a probe could not specifically hybridise to more clones than it could possibly overlap with, considering the depth of the libraries that were screened. For instance, a probe selected from the HHMI library could not overlap with (and hence hybridise to) more than 20 clones from the CEPH and ICRF libraries. However both approaches had the immediate consequence of reducing the size of each cluster considerably, since a large number of probes were rejected from the analysis by probeorder. When constraints were applied to the contig building process, the number of probes and clones was reduced by 20 to 50 %. This did not necessarily imply that

an equivalent proportion of hybridisation results were not satisfactory, since many probes could be genuinely linked by only one clone, or could realistically hybridise to an exceptionally high number of clones. Such a drastic reduction of the dataset was more a consequence of the lack of flexibility of the algorithms. A set of rules or constraints could only be applied to the entire dataset at once, and could not distinguish between genuine weak links and true inconsistent data.

Consequently, it was decided to reduce the influence of the contig building software in the analysis, and to increase the importance of human decisions in isolating 'good' from 'bad' experimental results. In this approach, the first step was to analyse the entire dataset (2700 hybridisation results) using probeorder, without applying any constraints. This produced a log file which contained 113 clusters, comprising 4,087 clones. In 38 cases, clusters contained a single probe (singletons) and therefore were not useful for contig construction. The remaining 75 clusters containing 3973 clones were used for constructing contigs manually. A walking procedure was initiated from pter to qter which consisted of examining each cluster in turn, in combination with any other information available associated with YACs from the clusters. Attempts to analyse the set of hybridisations automatically as described above showed that information derived from different experimental sources did not agree in many places. Therefore in the manual analysis, it became necessary to define a hierarchy between the different type of data, so that in case of conflict decisions could be made in a consistent manner. FISH and marker content information were considered together first; fingerprints (gel and hybridisation based) were then compared in order to confirm the overlaps and determine a relative order of clones within a cluster.

An example is illustrated in figure 5.2. Contig 55 reveals two clusters mapping to Xq13 and to Xp22, and joined by hhmi25h4 that hybridises to both ICRFy900F1236 (Xq13) and to ICRFy900H01149 (linked to the Xp22 cluster). This connection is however weak (italics in figure 5.2), since it is only supported by one faint (intensity 1) YAC to YAC hybridisation result. Both clusters agree well internally when mapping data associated with the YACs are compared. In the Xp22.33 cluster, an ICRF YAC was mapped by FISH to this cytogenetic band, 3 hhmi clones have been positioned across the pseudoautosomal boundary (PABX) by a collaborative group, a CEPH YAC contains DXS1233 which maps approximately 500 kb from PABX (Information from CEPH Infoclone database), and finally an ICRF YAC contains MIC2, a pseudoautosomal gene, and has been mapped by irradiation hybrid to the first interval (0.00-6.93) on Xpter. The Xq13 cluster contains 4 clones that contain closely linked genetic markers (DXS131, DXS559, PHKA1 and RPS4X, RLDB database). The clone ICRFy900B0862 from this cluster has however been assigned to Xp11.21 by radiation hybrid mapping, with a probability of 0.9. This information can be rejected on the basis that all the other pieces of data converge to a single location on the X chromosome. Both clusters were placed on the consensus X chromosome

map in the Xq13.1 and Xp22 bands where the markers are located (Xq13.1 cluster is shown in figure 5.4).

Figure 5.2. Sample of the log file produced by probeorder on the YAC to YAC hybridisations, when no constraints have been applied. This cluster contains 8 probes and is divided in two sub-groups of YACs mapping to Xq13 (top) and Xp22 (bottom), linked by a weak connection. The format and nomenclature of this figure is identical to that of figure 3.11.

```
contig   55 size:    8 location:  Xp22.3 FISH (RV) 3 q13.1 3 q13.2

! probes:

hhmi19c6          66    1    3
ICRFy900B0862 77    1    5 mapped at Xq13 DXS131, Xp11.21 rfh 66.13-74.35
ICRFy900F1236     80    1    3
hhmi25h4          80    1    4
ICRFy900B0246     43    9   11 mapped at Xp22.3 FISH(RV) Yp11 3q27 14 p11.2
hhmi19b4          66    5   13
ICRFy900A0101     93    1    8 mapped at Xp22.32 MIC2 Xp22.33 *rfh 0.00-6.93
hhmi30e4

! length: 5.49 clone lengths (2.75)

! fitted clones:

! ywxD792          1''''.'.   mapped at Xq26.0 DXS874
! hhmi2D10         1..'....#
! hhmi1D6          12.'....
! hhmi19C5         '1.'.'.'
! ICRFy900F065     '1.'.'.'   mapped at Xq13 DXS131, DXS559
! ICRFy900C09133   '3.'.'.'   mapped at Xq13 PHKA1, RPS4X
! ICRFy900B0862    '32'.'.'   mapped at Xq13 DXS131, Xp11.21 rfh 66.13-74.35
! CEPHy904E06917   '.2''''.'  mapped at X    (FISH)
! ICRFy900F1236    ..32'...   mapped at Xq13 DXS131
! ICRFy900H02111   ...1'...
! CEPHy904E05824   ...1'
! ICRFy900H01149   ...12...
! hhmi9B1          '..'3'.'
! hhmi8E2          ...'22..
! ICRFy900B0246    ....33..   mapped at Xp22.3 FISH(RV)+ Yp11.3 3q27 14p11.2
! ICRFy900H12112   ....33..
! hhmi19B4         '..'23.'
! hhmi14E9         ...'12..   mapped at Xp22.33 BC 2185.87 PABX
! hhmi15H2         ...'333.   mapped at Xp22.33 BC 2187.74 PABX
! hhmi6F12         ...'231.   mapped at Xp22.33 BC 2411.03 PABX
! CEPHy904C06742   ....331.   mapped at X 0.0  AFMa141xe5 (DXS1233)
! hhmi23B5         ...'2311
! hhmi20B4         ...'.3..
! hhmi14D6         ...'.3..
! hhmi9H8          ...'.2..
! ICRFy900A0101    ...''33.   mapped at Xp22.32 MIC2 X p22.33 *rfh 0.00-6.93
! hhmi21D2         ...'..1.
! ICRFy900H12146   '..'.'2'
! CEPHy904D05860   '..'.'1'
! hhmi26H6         ...'...1#
! CEPHy904F12770   .......3
! hhmi30E4         .''''.'3#
! hhmi16F4         ...'...3#
! hhmi17D2         ...'...3#
! CEPHy904F12783   .......3#

! connections to other contigs:
!  ywxD792 (1) -> contig   76 probe ICRFy900F0725     Xq28.0 FISH 9p25 (NC)
!  hhmi1D6 (2) -> contig   76 probe hhmi19e8
! CEPHy904E06917 (2) -> contig   24 probe CEPHy904F01816
! CEPHy904E06917 (3) -> contig   59 probe ICRFy900B0446
! ICRFy900H02111 (3) -> contig   57 probe hhmi24h1
! ICRFy900H02111 (2) -> contig  113 probe ICRFy900F0457
! CEPHy904C06742 (1) -> contig   68 probe ICRFy900D0236     Xp11.4 FISH (SP)
! CEPHy904C06742 (3) -> contig   84 probe ICRFy900H0410
! CEPHy904F12770 (2) -> contig   65 probe ICRFy900F1101
```

weak connection

Xq13 cluster

Xp22 cluster

Figure 5.2

The next cluster in this log file (contig 56) allowed another group of YACs to be placed about 300 kb telomeric from this position, in the Xq13.2 band. Analysis of hybridisation fingerprinting data showed that ICRFy900C09133, from the centromeric cluster, overlaps with ICRFy900H05155 (figure 5.3).

```
! probes:

ICRFy900C09133 ~       86    2   13   13  85
ICRFy900H05155

! length: 0.85 clone lengths (0.42)

! fitted clones:

! ICRFp600M245     3.#
! ICRFp600J175     2.#
! ICRFp600N215     3.#
! ICRFp600P1812    2.#
! ICRFp600J1811    2.#
! ICRFp600K169     2.#
! ICRFp600N161     1.
! ICRFp600M201     3.#
! ICRFp600L222     1.#
! ICRFp600O1510    3.#
! ICRFp600H161     2.#
! ICRFp600G2310    23
! ICRFp600P1311    23
```

Figure 5.3 Sample of the same log file as illustrated in figure 3.11 (hybridisation fingerprinting). The YAC probe ICRFy900C09133 is part of the Xq13 cluster in figure 5.2. Positive XPL clones for this YAC probe are indicated with a signal intensity (1, 2 or 3) in the first column, and positives for ICRFy900H05155 in the second column. ICRFy900C09133 is shown here to overlap with a YAC that contains PHKA1 (ICRFy900H05155): the two YAC probes share two XPL clones (underlined here, and red arrows in figure 5.4). PHKA1 is a marker present in an adjacent contig, in Xq13.2 (contig 56) . This result allowed the bridging of the gap between the two contigs.

ICRFy900H05155 contains the marker PHKA1 (underlined in figure 5.4), also contained in ICRFy900B0871 which belongs to the telomeric Xq13.2 contig. This information allowed to bridge the gap by demonstrating an overlap between two YAC clones belonging to the two adjacent contigs. The telomeric contig, which spans the marker DXS227, contains a group of YAC clones that have been gel fingerprinted (figure 5.5), allowing a relative order to be determined.

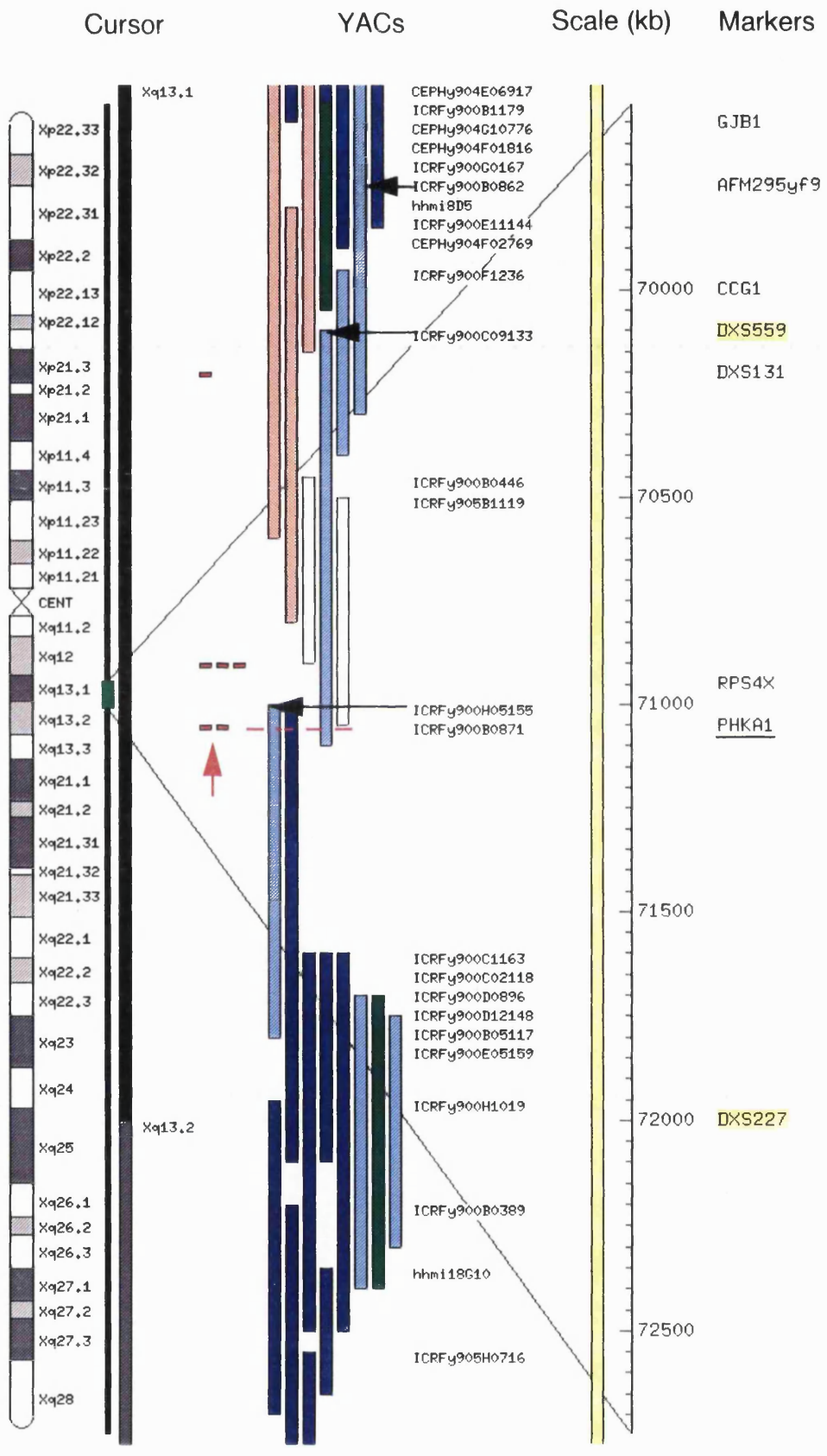

Cursor
YACs
Scale (kb)
Markers
Xq13.1
Xp22.33
Xp22.32
Xp22.31
Xp22.2
Xp22.13
Xp22.12
Xp21.3
Xp21.2
Xp21.1
Xp11.4
Xp11.3
Xp11.23
Xp11.22
Xp11.21
CENT
Xq11.2
Xq12
Xq13.1
Xq13.2
Xq13.3
Xq21.1
Xq21.2
Xq21.31
Xq21.32
Xq21.33
Xq22.1
Xq22.2
Xq22.3
Xq23
Xq24
Xq25
Xq26.1
Xq26.2
Xq26.3
Xq27.1
Xq27.2
Xq27.3
Xq28
Xq13.2
CEPHy904E06917
ICRFy900B1179
CEPHy904G10776
CEPHy904F01816
ICRFy900G0167
ICRFy900B0862
hhmi8D5
ICRFy900E11144
CEPHy904F02769
ICRFy900F1236
ICRFy900C09133
ICRFy900B0446
ICRFy905B1119
ICRFy900H05155
ICRFy900B0871
ICRFy900C1163
ICRFy900C02118
ICRFy900D0896
ICRFy900D12148
ICRFy900B05117
ICRFy900E05159
ICRFy900H1019
ICRFy900B0389
hhmi18G10
ICRFy905H0716
70000
70500
71000
71500
72000
72500
GJB1
AFM295yf9
CCG1
DXS559
DXS131
RPS4X
PHKA1
DXS227

Figure 5.4

Figure 5.4 Graphical representation of the two Xq13 contigs joined by experiments described in figures 5.2, 5.3, and 5.4. The figure is a screen dump of a map from the Integrated X chromosome database (ACEDB version, see chapter 6), focusing on the Xq13 band. The YAC clones that proved instrumental for joining the two contigs are indicated by black arrows. The red arrow shows the two XPL clones that indicated the overlap between YACs of the two contigs, one of which contains the PHKA1 marker underlined on the right. Markers highlighted in yellow are polymorphic. Color codes for the YAC clones are as follows: green, chimeric; pink, FISH mapped; clear blue, contains markers; dark blue, has been gel fingerprinted; white, only YAC to YAC hybridisations. Most YAC clones are associated with a combinations of the above data although a single color was chosen for each clone, based on a hierarchy that follows the above list. All clones were involved in YAC to YAC hybridisation and/or hybridisation fingerprinting experiments (i.e. are extracted from the probeorder log files, figure 5.3 and 5.4).

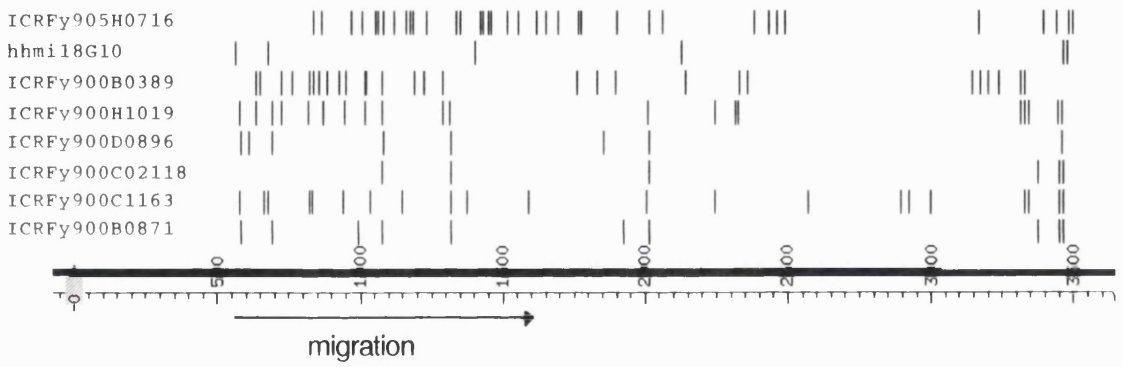


Figure 5.5 Schematic representation of Alu-PCR gel fingerprints of 8 YAC clones present in the Xq13.2 contig on figure 5.5. Some bands are clearly shared between clones, supporting the presence of overlaps. The YAC ICRFy900B0871 (bottom) contains the marker PHKA1, also contained in ICRFy900H05155, which is shown in figure 5.3 to overlap with a YAC from the Xq13.1 contig, thus joining the Xq13.1 and the Xq13.2 contigs. The scale is in mm, and indicates the corrected migration distance of the DNA fragments in the electrophoresis gel.

This type of analysis was followed for the entire list of clusters, thus progressively extracting reliable contigs from the Alu PCR hybridisation dataset generated by probeorder. Links between clusters were identified in the output of probeorder, by seeking YAC probes hybridising consistently to clones present in two different clusters. Identification of such links reduced the total number of contigs from 75 to 24. When available, results stored in IXDB (generated by RLDB collaborators) were compared and used to orientate contigs and confirm orders and overlaps

between clones. Since all of the RLDB collaborators which indirectly contributed to the project also participated in the establishment of the yearly consensus map, the integration of the contigs with the consensus map was greatly facilitated. The analysis described above was performed over a period of 6 months. The result of this strategy is a YAC contig map of the X chromosome integrating 655 genetic markers with 906 YAC clones, organised into 24 contigs (Figure 5.6). The total coverage is estimated to be 80 % of the length of the chromosome, or 125 Mb of DNA. Based on the hybridisation fingerprints results described in chapter 3, 79 intervals could be defined in the YAC contigs, in which 1,420 XPL were placed.

Figure 5.6 Integrated YAC map of the human X chromosome, slightly modified from the IXDB (acedb version 4.3) map view. The scale is based on a 160 Mb chromosome and each graduation represents 5 Mb. A chromosome ideogram is drawn on the left, and each colored box to the right represents a YAC clone (see color code in expanded view on the right of the figure). The yellow boxes to the right of the clones show the extent of the contigs. These cover approximately 125 Mb of the chromosome (80 %), and include 906 YAC clones. A magnified view of 10 Mb in Xp22 is schown to the right. The color code is only an indication of one of the techniques which contributed to the positioning of a given clone on the map. All clones shown have either been hit or used as probe in a hybridisation experiment. White clones have no other evidence for their position. Clear blue clones contain the markers (DXS, genes) indicated to the right of the scale. Pink clones have been mapped by FISH. When the FISH experiment indicates a chimeric clone, the latter is shown in green. An Alu gel-fingerprint is stored in IXDB for the dark blue clones and available in a dedicated viewing tool for comparison between clones (Figure 5.5). However when a clone has been anaiysed by more than one method (e.g. by gel fingerprinting and mapped by FISH) only one technique is indicated by the color code. In IXDB, a single click of the mouse produces a window where the complete set of information attached to a clone is displayed. Small red boxes between the clones and the chromosome bands represent cloned Alu PCR products identified in hybridisation fingerprint experiments. More than one Alu clone in a single position indicates that the order between the clones could not be resolved. A maximum of 3 clones are shown, while the average number of clones per position is 10. Units in the expanded view are in kilobases, starting from 0 to 160,000 (pter to qter). The scale is only indicative, and facilitates the comparison with the X community consensus map.

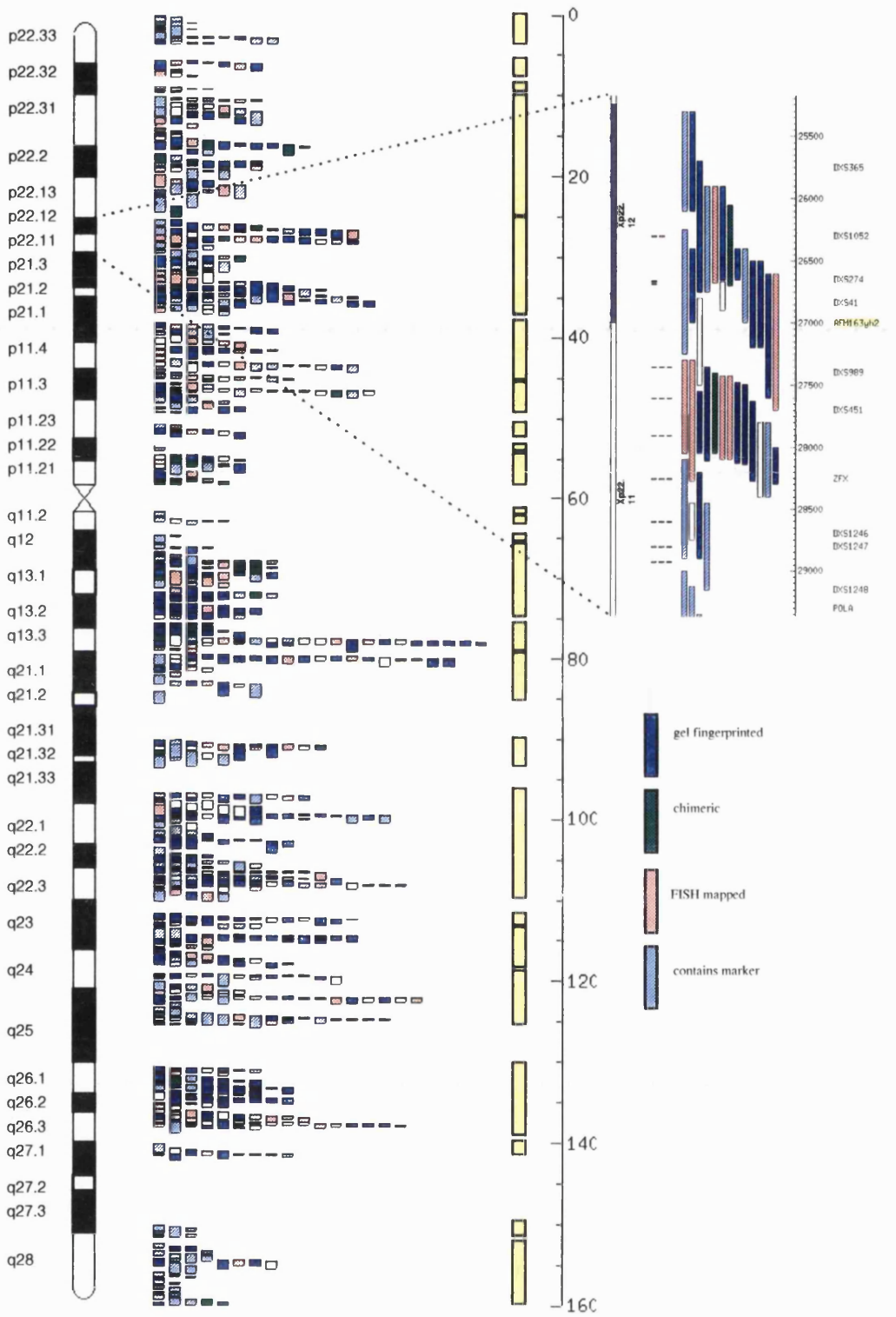

p22.33
p22.32
p22.31
p22.2
p22.13
p22.12
p22.11
p21.3
p21.2
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
q11.2
q12
q13.1
q13.2
q13.3
q21.1
q21.2
q21.31
q21.32
q21.33
q22.1
q22.2
q22.3
q23
q24
q25
q26.1
q26.2
q26.3
q27.1
q27.2
q27.3
q28
0
20
40
60
80
100
120
140
160
gel fingerprinted
chimeric
FISH mapped
contains marker

Figure 5.6

5. Discussion and Conciusions

A predominantly hybridisation based experimental approach has been applied to establish YAC clone contigs covering approximately 80% (125 Mb) of the human X chromosome in 24 contigs. The map comprises some 750 discrete markers of all types (genetic, vectorette, inter Alu-PCR products). A large experimental dataset was generated which was first processed with computer programs in order to lower its complexity. A stringent manual analysis was then performed on each YAC cluster, using all available information.

In the Alu-PCR hybridisation data, 22 % of the YAC probes did not hybridise to themselves. The hybridisation data generated by these probes was still considered in the analysis, for the following reasons: the most frequent source of false negative can be ascribed to the absence of the probe DNA on the filter, either because the robot pins did not transfer liquid on this particular spot, or because the YAC did not amplify properly in the waterbath PCR robot. In these cases connections between the probe and the clones it identifies are still correct. Alternatively, it is possible that probes were accidentally mixed up, leading to apparent false negatives. These results could easily be detected in the redundant dataset, as they present a clear aberrant hybridisation pattern in the later analysis. Based on this observation, 51 probes were removed from the dataset (7 % of all probes). False positive hybridisation results, on the other hand, can introduce false connections between clones. It is impossible to accurately measure the rate of false positives, but based on the number of links between probes that had to be ignored in the manual analysis, it is estimated at 10-15 %. The reason for their presence can either be due to human error (typing, scoring, sample handling) which is particularly acute in a large scale project, and to non-specific sequence similarities between clones. Mis-scoring was limited by the fact that each X-ray film was scored by two different persons independently. Human error was corrected further by scanning the dataset with programs to detect specific patterns of e.g. cross-contamination (probes in adjacent wells with identical hybridisation patterns) or non-removal of a probe from a filter (same clones positive in two successive hybridisations). Non specific sequence similarities are a well known problem in mapping large regions of the human genome, and come in addition to the high level of chimerism observed in YAC libraries (30 %). Whether due to repeat sequences or gene families, this problem can only be avoided by using complementary techniques to help make decisions. This problem was addressed by complementing the Alu PCR data with a battery of different types of data (fingerprints, marker content, FiSH, end mapping, radiation hybrids) in a stringent manual analysis.

The map covers 80 % of the chromosome, with 25 gaps. The depth of coverage is uneven (figure 5.6) with up to a 20 fold difference within 2 Mb around the Menkes syndrome locus. Long range coverage is however balanced, with 38 % of the YACs localised on the short arm (36 % of chromosomal length), and with the remaining

62 % on the long arm. The largest gap in the YAC contig map is in Xq27 where almost the complete band, which measures approximately 11 megabases, is not represented. Since a complete lack of Alu sequences over such a large region can be excluded, the reason for this under representation must be ascribed to an unfortunate absence of probes mapping to this region in our random selection of HHMI clones. The region is at least partially represented in the target libraries, since YAC contig construction has been reported in this cytogenetic band (Zucchi et al., 1996). Nevertheless in some cases we do observe a correlation between large gaps and the presence of a G-dark band. This is consistent with studies showing that these regions are relatively poor in Alu sequences (Korenberg and Rykowski, 1988). However not all G-dark bands are poorly represented (for example Xp22.2, Xq13.1 and Xq23).

This map was compared with X chromosome YAC contig maps built as part of whole genome efforts by CEPH (Chumakov et al., 1995) and by the Whitehead/MIT (Hudson et al., 1995) groups. In both cases, the X chromosome stands out due to its poor coverage compared to the average of the autosomes. This is principally due to the low representation of the X chromosome in the CEPH library made from a male cell line, which was the only substrate for the construction of the physical maps. Also the human and mouse X chromosomes contain either less CA repeats, or less polymorphic CA repeats (Dietrich et al., 1996). This leads to a lower density of genetic markers available for whole genome physical maps based on this type of STSs. Therefore, an independently derived map of the X chromosome in YAC clones using libraries enriched for X chromosome DNA and independent from CA repeat content is particularly complementary. The approximate coverage of the X chromosome in the CEPH map, calculated with markers in common with the workshop consensus map, is 52 Mb (32 %). The marker order between our map and the CEPH map agrees well, except in one instance, where a group of markers is clearly misplaced in the CEPH map (XIST is placed in Xp11). Comparison with the Whitehead/MIT map of the X chromosome is more difficult, since the majority of markers have been developed very recently by this group and are therefore not placed on the map described here. It was possible to find 35 DXS markers common to both maps, for which the order broadly agrees, except for the first half of the short arm. In that region, the order of the 9 common DXS markers strongly disagrees with the X chromosome community consensus map and with the map presented here, over a 30 Mb region. The Whitehead/MIT order used for comparison is extracted from the radiation hybrid/STS content map. Based again on the physical distances between common markers in both maps, the estimated coverage of the Whitehead/MIT map is approximately 50 Mb. This is also sustained by the maximal length of the contigs presented, based on the average length of a YAC clone. The consensus map established at the 6th X chromosome workshop (Nelson et al., 1995) and updated at the 7th, reported an almost complete coverage of the chromosome in YAC contigs and the presence of approximately 10 gaps in the map. This consensus was derived by collating the

contigs from more than 50 different groups, and concentrates on marker order rather than attempting to present YAC clone organisation. Therefore, the estimation of the size and number of gaps and the YAC coverage has to be taken with caution.

Gap closure is undertaken using two strategies, bypassing the use of inter Alu PCR and establishing useful landmarks for a cosmid/P1/PAC map of the chromosome. First, these *E. coli* based clones are being identified using genetic and physical STS markers developed in the CEPH and Whitehead/MIT mapping efforts which are likely to be positioned in the gaps. The STSs are amplified from total human DNA and used as hybridisation probes. The positive clones in turn are used to screen the genomic YAC libraries to identify clones missed by Alu PCR YAC probes. Secondly, a combination of L1 (Line repeat) and Alu primers are being used to amplify YAC clones from the ends of existing contigs. These are used to screen the same cosmid and PAC libraries, therefore identifying cosmids and PAC clones at the edges of the gaps. When used to screen against the YAC genomic libraries, these probes can identify new clones extending from the original contig.

Arising out of the X chromosome workshop was a common accord that a repository of all YAC clones known or supposed to map to the X chromosome must be established, in combination with a dedicated database which would make available all the published mapping information. This project was taken on by our laboratory, which has now distributed 7 copies of a 9,000 clone collection to genome centres worldwide. Also, high density gridded YAC colony filters of the collection are made available, as well as DNA pools for PCR screening. This will increase the value of the X chromosome YAC resources available worldwide, and will allow verification and completion of the existing consensus YAC map. The clone collection includes the cX library described here and clone sets from groups based at the Sanger Centre, The Baylor College of Medicine, the Washington University School of Medicine and many others.

Clearly, the mapping of the X chromosome is reaching a stage where increasing efforts will be put into the construction of higher resolution maps in bacterial cloning systems (cosmids, P1, PACs, BACs), in which the YAC clone resources and maps will play a central role (see chapter 7). These bacterial clones will be essential for the large scale genomic sequencing and transcriptional mapping of the chromosome. Using the YAC contig map presented here, a systematic identification of X chromosome PAC and cosmid clones has been initiated. This is the next logical step towards a high resolution ('sequence ready') clone and transcript map of the chromosome, itself the consummate template for large scale sequencing projects.

CHAPTER SIX: IXDB, The Integrated X Chromosome Database

1. Introduction

One characteristic of biological research, especially in molecular genetics, is that scientists continuously generate resources in the form of clone libraries, cell lines, antibodies, etc. Research would be difficult without the availability of such samples. Investigators spend further time and energy characterising the samples in their own experiments, and it is these studies which lead to scientific discoveries, deduction of new biological mechanisms or simply to maps of the genome. The truly marvellous feature of this process is that a sample that once was with a thousand almost identical other samples, may overnight acquire an immense value simply because an experimental evidence shows that it contains a long sought after gene. Pieces of knowledge attached to a biological sample therefore gives the latter its true value.

In the course of the Human Genome Project, millions of such samples have been generated, distributed and characterised, a task which has required unprecedented manpower and funding from the scientific community. It has become obvious that the only way to keep duplication to a minimum is to maintain databases in which results are recorded and made publicly available. The John Hopkins University School of Medicine in Baltimore was designated to be the first home of such a repository, and developed the Genome DataBase (GDB) (Fasman et al., 1996). The GDB relies on submissions from individual scientists, and has now become a huge repository. Unfortunately the amount of information that has poured into the GDB has now overwhelmed the capacity 'GDB editors' to curate the data and establish a consensus. One contributing factor to this increased complexity stems from the behavior of scientists themselves. Investigators want credit for the work they have done, and legitimately so. However, biologists have a well known tendency to rename any sample that has been previously characterised under a different name, which allows new entries to be created in the GDB. The consequence of this is an increased and hidden redundancy. In addition, because the GDB essentially contains information derived and interpreted from experimental results, but not the latter themselves, it is often problematic to establish the true quality of a piece of data.

The Reference Library System (RLS, Zehetner and Lehrach, 1994) in contrast, has been a repository of both experimental data and biological samples. The system provides the means to generate results, by distributing high density gridded filters and individual clones. In order to interpret the results of hybridisations performed on the filters, scientists must submit the raw experimental results, enabling the Reference

Library DataBase (RLDB) to store them. A large pool of 'reference' clone libraries is available to all scientists, and a strong emphasis is placed on keeping the names of individual clones constant in the entire process. One implied duty of investigators who make use of resources provided by the RLS is to provide feed-back information on the samples that are obtained. The ICRF YAC library (Larin et al., 1991) has been part of the RLS since 1990 and has therefore seen an increased amount of mapping information attached to a large number of clones. The CEPH 'mega' YAC library has also been available via the RLS, although the main source of information available on this library is the result of the CEPH's human genome mapping project itself. The Quickmap Infoclone database (Chumakov et al., 1995) maintained by CEPH is similar in essence to the RLDB, except that it concentrates on a single library, and that results are essentially generated in house by the CEPH group itself. As for the RLDB, it makes available raw experimental results and makes no attempts at polishing or interpreting them.

The X chromosome mapping project initiated in our laboratory did not include any experimental strategy to associate YAC clones with known genetic or physical markers, although FISH experiments were performed on selected clones. There are two reasons for this. First, as opposed to many other mapping projects, detection of YAC overlaps did not rely on marker content (STS, genetic or physical markers, etc.) but on direct YAC to YAC hybridisations. Second, the map was mainly built with the ICRF and CEPH YAC libraries which were already deeply characterised for marker content, for the reasons described above. A separate attempt by our group would have lead to a large amount of unnecessary duplication. This information is however vital in order to place the YAC contigs emerging from the hybridisation work on the X chromosome genetic and physical map. This chapter first describes the strategy employed to collate data on marker content of the YACs used in the project in ACEDB.

A European consortium of 17 groups has since been created to develop a transcriptional map of the human X chromosome. To adapt to this new challenge of distributing and updating data between 17 major groups, the database system was changed from ACEDB to a relational system, namely ORACLE. Both databases currently co-exist, and have been made publicly accessible over the World Wide Web (WWW). The development of IXDB in ACEDB, the delineation of the concepts followed by the second version of IXDB in ORACLE, the design of the data structure, integration in the work-flow and in the X chromosome project at large are part of this thesis work. The technical implementation of IXDB in ORACLE, the parsing of data and the implementation of the database server are however the work of Ulf Leser at MPIMG.

2. Collection of data

2.1 Marker content of YAC clones

2.1.1 The Reference Library DataBase

2.1.1.1 Origin of the data

The Reference Library System was used in two stages. First, SQL (Structured Query Language) queries for all clones that had been identified by probes mapped to the X chromosome retrieved 1,723 records. Most of these clones were identified in primary screens of library filters, and in 90% of the cases the results had not been confirmed by a secondary hybridisation to the clones themselves. These results were difficult to use directly. Since most probes were named using an internal laboratory nomenclature, it was not possible to relate the results to each other. The stated localisation of the probes on the chromosome were often relative to a cytogenetic band and not to neighbouring loci, which is not sufficient for placing clones on a map. It was therefore necessary to go further upstream and contact directly the scientists who submitted their results to RLDB. A list of 21 laboratories was drawn up which had made use of RLS resources for building YAC contigs on the X chromosome. In the documentation accompanying the original RLS material sent out (filters and clones), investigators were informed of their responsibility to return all mapping data derived from these samples. Each was therefore individually contacted by letter, FAX, telephone calls or personal visits. In order to facilitate the process of returning substantial information, extensive submission forms were sent to each group. In most cases, it was however easier to simply collect laboratory notes and maps. Furthermore, attendance at three successive International workshops on X chromosome mapping (St. Louis, USA 1993; Heidelberg, Germany 1994; Banff, Canada, 1995) allowed a large amount of results to be collected and updated in a concise format. At these occasions, it became clear that many investigators had screened copies of the ICRF YAC library that had been distributed by our laboratory to other groups. These results had bypassed the RLS system and therefore were not listed in the RLDB records studied in the first phase. Results from 28 groups were finally collected, which together represented 42 contigs. All were established in the context of positional cloning projects. A number of these projects overlapped and therefore some maps were constructed over identical regions.

2.1.1.2 Data entry in ACEDB

The advantage of collecting information actively by directly contacting laboratories and gathering tables and maps is that each dataset represents a

snapshot of the current results on a particular mapping project. Information was generally condensed on one map representing the clones graphically in the relevant genomic region, with their attached genetic and physical markers. An accompanying table summarised the experimental results, by plotting the list of probes against the YAC clones and indicating a positive or a negative match. The main disadvantage however of actively collecting results is that data is presented in a very heterogeneous format, in contrast to results provided via submission forms. In order to make sense out of the information, to compare results and ultimately use them for our mapping project, it was therefore necessary to translate the heterogeneous set of data into one format. ACEDB was chosen at an early stage of the project to be a repository of this information. For undertakings of this scale and nature, ACEDB is a well suited database system that provides extensive graphical features, is very easy to customise to a given set of genomic data, and has a simple way of importing and exporting information. The database that progressively emerged from compiling diverse X chromosome data in ACEDB (and later in ORACLE) was named the Integrated X chromosome Database (IXDB)

Maps provided by collaborators in the RLS were hand or computer drawn on paper and it rapidly became obvious that manually converting them to ace format (the ASCII text format required to enter data in ACEDB) would be a tedious and error prone process. A software called xcontigview was therefore written by Huw Griffith in our laboratory to partially automate this operation. Maps were first converted to TIFF images (Tagged Image Format) using either a CCD camera normally used for ethidium bromide stained gels, or an X-ray film scanner (Amersham). Images were loaded in the program and appeared on the computer screen. After setting some parameters such as the origin, scale, source of data etc., the content of the map (clones and probes) was digitised using the computer mouse. This was simply done by clicking on the beginning and the end of each clone, probe or gene and typing in the name of each object. The program then automatically converted this information into ace format. In a few minutes a complicated map could therefore be parsed into IXDB and represented in the physical map display. Experimental results accompanying the maps and often represented as tables, were directly typed in the database.

2.2 Public domain data

Our laboratory is part of the Integrated Genome Database consortium, which in the first phase of the project attempted to compile in a single format (ACEDB) data from a variety of independent databases such as GDB, OMIM, RLDB, EMBL, etc. It was clear that when IGD would release its first set of data, the X chromosome section could be directly imported into our ACEDB database. This would ensure that an exhaustive sampling of public domain information related to the X chromosome would be integrated with our experimental dataset. In the early stages of the X chromosome

project however, the IGD project was still in its early development and it was therefore necessary to start translating publicly available data in-house. The main focus was put on the Genome Database, since it provides a way of identifying the official D-number of numerous probes used by collaborators from the RLS. The Human Genome Mapping Project Resource Centre (HGMP-RC) situated at the time in Harrow (UK) was a mirror site of the GDB and provided a convenient way of accessing data from the ICRF. A simple user interface was available for registered users via a telnet session. The GDB was queried for all loci and genes localised on the X chromosome, and approximately 2500 entries were retrieved the first time. Three updates were performed in the course of the following 18 months and the number of entries reached approximately 3000 for the last update. Complete entries were downloaded to a local Unix station (Sun Sparc 2) via electronic mail, in batches of 100 kilobytes. The files were then concatenated and processed by an awk program to translate the information into ace files. These were then directly parsed into IXDB.

Approximately 24 months after the start of the X chromosome project, IGD released its first set of integrated data. Data was available via ftp from the Deutsches Krebsforschungszentrum (DKFZ) in Heidelberg and was sorted either according to the database of origin (GDB, EMBL, RLDB, etc.) or according to the relevant human chromosome. The X chromosome dataset was downloaded and regular updates were made every few weeks. The emerging YAC map under construction in our laboratory was already annotated with approximately 600 genetic and physical markers. Incorporating the IGD X chromosome section allowed each locus and gene to be fully described with up-to-date information. The first phase of the IGD project has now ended and consequently the release of data has been terminated as well.

A consensus marker map is established for each chromosome once a year based on information collected at single chromosome workshops, and is a reference for all investigators in the field. The YAC map constructed in our laboratory is also strongly correlated with the consensus maps. Xcontigview was used to parse in IXDB the consensus maps from the 1994 and 1995 workshops (Heidelberg and Banff respectively (Nelson et al., 1995; Willard et al., 1994)), providing a convenient way to compare the order of markers with other maps as well as a position for approximately 500 loci described by the IGD data. By combining experimental data from our group and 28 collaborators from the RLS with the exhaustive dataset from IGD, IXDB gradually became an electronic catalogue of a large fraction of the data generated so far on the X chromosome physical map by the community.

3. IXDB in ACEDB

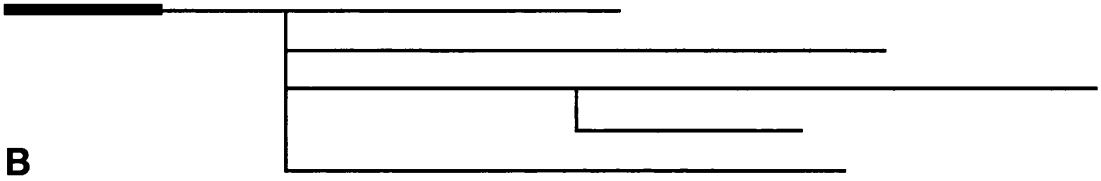
3.1 The ACEDB system

ACEDB is a system close to the object-oriented system, although some features differ from it and make ACEDB a system of its own. ACEDB works with information that is stored into the main memory of the computer and maintains it there, a feature that gives excellent results in terms of efficiency of data location and retrieval. From a biologist's point of view, ACEDB is a graphical program which works using a windowing system, presenting data in different types of windows according to the different types of map. Maps and windows are linked in a hypertext fashion, so that clicking on an object will display further information in a separate window. The main disadvantages of ACEDB concern data integrity and sharing. There are few if no mechanisms to ensure that data values are valid and secure. Only a single user may have write access to the database at one time, and the security on the database disk files for multiple users with write permission is minimal. The query language provided with ACEDB is limited, although this is not an issue for most biologists, since retrieval of information is seldom performed via the query facility.

3.2 Data structure and database content

The structure of the data in ACEDB closely fits the biology that it is modelling, and therefore allows an intuitive design of the schema. A consequence of this is an enhanced data independence (i.e. independent of the way the database physically stores the information) and strong connectivity (relationships and interdependencies among the data). Each object is represented by a unique identifier, its name, followed by an ensemble of attributes organised into a tree. Branches of this tree may contain pointers to other objects or data. Objects are members of classes, and each class has a model which specifies the maximal extension of the branches, and the type of data (text, numerical value, etc.) which is allowed at each position. In general, each object stored in the database only has a part of the branching pattern permitted by the model. Another type of class is based on arrays rather than trees, which allows a more rigid but more efficient storage of data such as DNA sequences. The number of classes, their name, and their relations and attributes are defined to represent the domain in the most efficient way and are the responsibility of the database developer. The ensemble of class descriptions is contained in a single text file called the 'models' file (figure 6.1).

```
?Laboratory    is      UNIQUE ?Object XREF is
                       external_methods Pick_me_to_call Text Text
                       specific_properties Colleague ?Colleague XREF Laboratory
                                           contact ?Contact
A                      generic_properties  #Generic_properties
```



B

Figure 6.1 Example of a class described in the model file (A), with a schema of its 'tree' representation (B). The class name is preceded by a '?' sign, and is followed by a list of attributes ('tags'), each on a different line (branch). Each tag specifies the type of information which can be used, in the database, to describe the object 'Laboratory'. For instance on the third branch, a specific property of the class "Laboratory" is that it contains 'Colleague' and 'Contact'. Both tags are followed by the class name that describes the information 'Colleague' or 'Contact' (name, address, etc.). In addition, the class ?Laboratory is cross-referenced (XREF) to the class ?Colleague, so that the addition of a laboratory name to a colleague's name stored in the class ?Colleague automatically updates the class ?Laboratory.

In a first phase, IXDB was developed with ACEDB version 1.7, using the models distributed with the software, as a template for development. These were initially customised by the authors of the software for the data of the *C. elegans* genome project, and many parts were therefore directly available for our X chromosome project. Small modifications were however performed to customise the schema to this project. This essentially included the addition of the YAC and Cosmid classes as a replacement for the more general Clone class. The first version contained approximately 30 classes, of which only 6 (YAC, Map, Author, Cosmid, Locus, Contig) were used to store data. The remaining classes were either specific to the *C. elegans* project or required by the ACEDB code. At this stage, IXDB only stored information derived from collaborators in the RLS system, and data from GDB (Loci). In the second phase IXDB moved to ACEDB version 3 and at the same time started using the models from the IGD group. Since our laboratory was a member of the IGD consortium, we influenced to some extent the models that were written by the group of Otto Ritter to accommodate the various types of data from the databases that were involved in this project. Here, the schema was much more complex, with 104 different

classes available. This upgrade was timely, since it came at a time where the YAC map under construction was reaching a level of complexity that required an adequate environment to manipulate it. The contig map generated internally was therefore transferred from the probeorder text outputs into IXDB. In a last phase, IXDB was upgraded to ACEDB version 4 as it became available, still with models derived from IGD. This coincided with the publication of the YAC map, and provided a convenient way of making the huge volume of information publicly available. In its current status (IXDB-v1.0 in acedb-v4.3, Dec. 1996) IXDB contains information on over 20.000 different objects arranged in 58 classes. These include 4,200 YAC clones, 12,490 loci, and 3,705 sequences. Data on YAC clone overlaps has only been generated internally within the X chromosome project. Sequences are the complete set of X chromosome sequences available from EMBL as of 15-08-96, and represent approximately 7,5 Mb of DNA, 6 Mb of which are in fragments larger than 10 kb.

4. IXDB in ORACLE

4.1 The ORACLE system

ORACLE (version 7.3) is a commercial relational database management system (DBMS) that is widely used in industry and administration applications. The relational DBMS that it is based upon organises the data into tables. Entities such as clones or persons are stored in separate tables, that can be interconnected by referential integrity links. Such entities can be linked to each other by so called bridge tables. This uniform and clear organisation principle allows the efficient storing and querying of very large amount of information in a secure and comfortable environment. However, a drawback of the relational DBMS is the difficulty to translate many of the biological concepts and applications into the required table structure. Thus, a relational DBMS does not allow data independence. In addition, it makes development, maintenance and modifications time consuming. Some features of the ORACLE system are nonetheless essential for a database such as IXDB, and justify the change from the ACEDB system. These are:

- the possibility of concurrent access to write to and read from the database, for several users at the same time without affecting consistency.
- the ability to recover from network problems using a transaction concept.
- the possibility to adapt to growing amounts of data and users.
- efficient and comfortable retrieval mechanisms

Compared to ACEDB, a weak point of the ORACLE system is its complete lack of comfortable and user-friendly interface. The interface therefore must be designed and developed, which adds to the already time consuming task of developing the database structure. However, this can be considered as a gain in

flexibility, since the interface can be customised to the specific applications that the database was designed for.

	Strength	Weaknesses
ACEDB (~ object oriented)	• Data independence • Data connectivity • Performance • Powerful graphical interface	• Data integrity • Query language • Data sharing • No WWW server
ORACLE (relational DBMS)	• Data integrity • Data sharing • Query language • WWW server	• Data independence • Performance • No end-user interface

Table 2. Comparison of the strength and weaknesses of the ACEDB and ORACLE systems (relational DBMS in general).

4.2 Data structure and database content

Figure 6.2 shows a graphical representation of the database design for the X chromosome project, and the relations between the different tables.

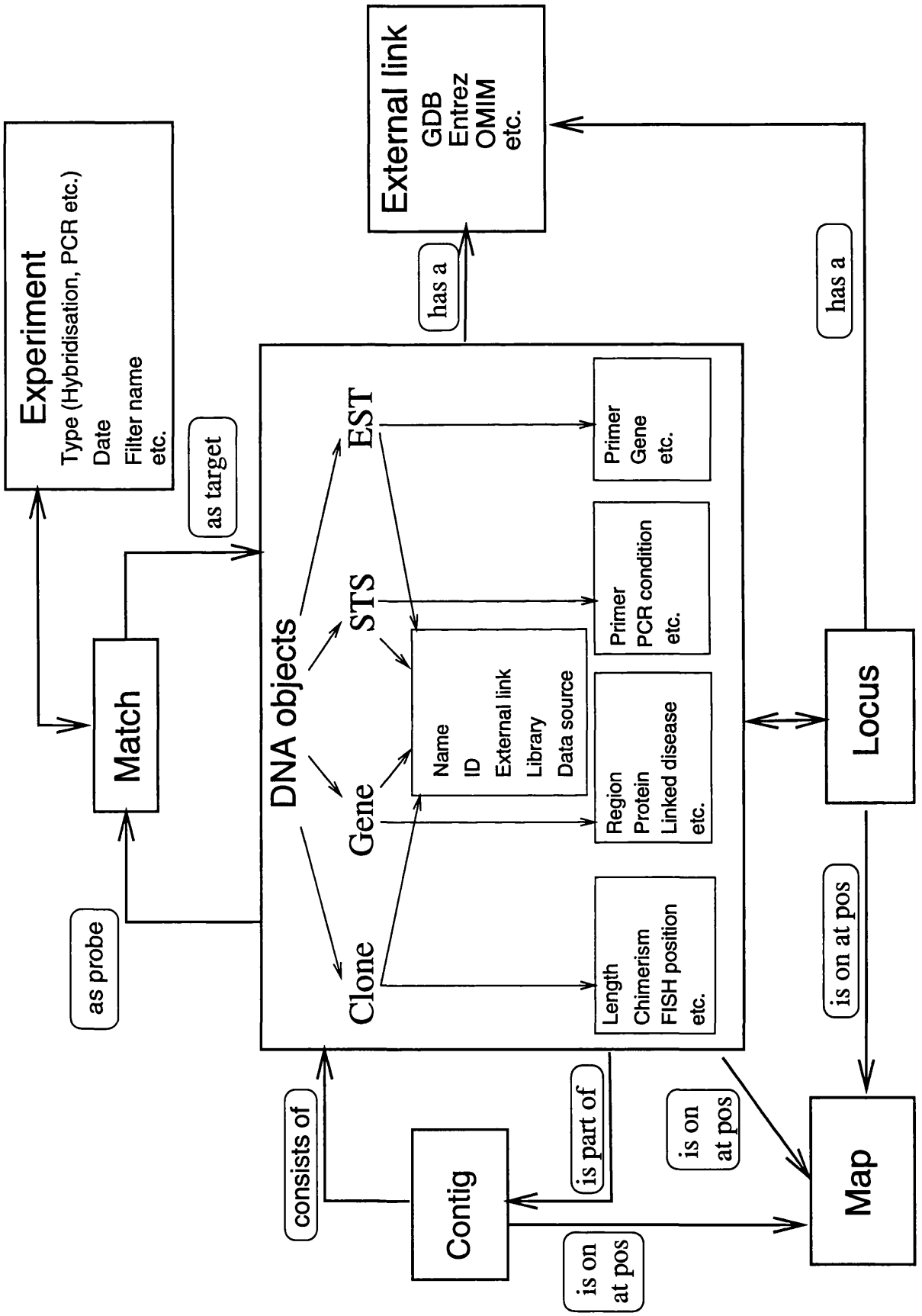

Experiment
Type (Hybridisation, PCR etc.)
Date
Filter name
etc.
External link
GDB
Entrez
OMIM
etc.
has a
has a
as target
Match
as probe
DNA objects
EST
STS
Gene
Clone
Name
ID
External link
Library
Data source
Primer
Gene
etc.
Primer
PCR condition
etc.
Region
Protein
Linked disease
etc.
Length
Chimerism
FISH position
etc.
Locus
is on at pos
consists of
Contig
is part of
is on at pos
is on at pos
Map

Figure 6.2

The purpose of IXDB in ORACLE is different to that of the ACEDB version. It is first designed to act as a central repository for information on the cX collection that was assembled in our group. It contains nearly 9,000 YAC names from sets of clones provided by 14 laboratories world-wide. Together with the clones, all groups provided mapping information in various types and format, which are now integrated in IXDB, and made publicly available. The second purpose of IXDB is to act as a core database for the X chromosome European consortium, that our group is coordinating. The aim of this project is to construct a transcript map of the X, in a combination with a large scale hybridisation based mapping strategy by our group and small scale regional approaches by all partners. The key element of such a project is the pooling of results, in order to avoid duplication of effort, and to benefit from the sum of all resources and information that is generated. The third purpose of IXDB is to offer an integrated view of a maximum amount of related information on the X chromosome, to assist our internal research project. This includes data from past workshops, from collaborators, external databases, literature, and from internal experimental work (Figure 6.3)

To date, IXDB in ORACLE is in its initial phase of development. The first objective described above has been reached, with information on the 9,000 YAC clones from the cX collection, and 3,000 additional ones (which were not included in the collection but for which data was available) stored in IXDB and accessible to the public. All loci mapped to the X chromosome and stored in GDB are present at least with hyperlinks, but often carry additional information. The same applies to all X chromosome genes stored in OMIM. IXDB is designed to store multiple instances of the same object, with the possibility of attaching separate, and even conflicting information to each instance. This allows the same YAC clone, distributed to several laboratories, to be associated with information from each source separately. In addition, each instance of the original clone may be stored under the alias name that it is generally given by the laboratory where it is being characterised. However, when a clone name is used as query, information on this clone and all its aliases is retrieved at the same time and presented with equal importance. The database therefore can deal with contradictory data, and leaves it to the user to resolve conflicts on a case by case basis.

The most important step when designing IXDB was to describe in an abstract way all the DNA objects, samples, experiments and data that are the tools and the purpose of a genome research project. Much work had already been performed in this field, for instance by the GDB team at Johns Hopkins University or the IGD group at DKFZ. Possibly the most difficult type of object to describe is the locus, due to the

WORLD
In-House-processing
ACE-WWW
Client-server
Selected dataset
X-YAC-Collection
Distributed hybridisation of
cX-YAC-collection.
ACEDB
Used as graphical interface
Maps - references
PhysicalMapper ?
Other tools
X-consortium
All kinds of experiments
All kind of clones
Hybridisation
results
Certain
Query
Ability
Special
Export
filters
Map-processing
Access to integrated dataset
for map-construction
All kind
of results
Certain
Query
Abilities
WWW
interface
ACE
dumps
Spec.
format
for
analysis
Day-to-day work
All kinds of querys for
- experimental design
- information demands
Full range
of queries
graphical interface
IXDB
Chromosome X - Hybridisation data
Research
Full SQL-possibilities for
- advanced querys
Permanent automatic data-exchange
in distributed environment
Primary database
German genome project - complete database, all chromosomes

Figure 6.3

variety of meanings that have been attached to this term. In genetic mapping, a locus is a point in relation to another point, from which it is separated by a recombination distance. In physical mapping, a locus may be a genomic region (gene), a cloned piece of DNA assigned to a region ranging from a chromosome to a few kb, an STS etc. The GDB has dealt with this situation by introducing the concept of DNA segments (D-segments) and their associated 'D-number', which described any standard genomic landmark that could be mapped (Gusella et al., 1980). The latest version of GDB (GDB 6.0, Fasman et al., 1996) has moved away from this system, replaced by the use of unique GDB identification numbers ('GDBids') to describe all objects in the database. The term *'Locus'* has also been removed, replaced by *'Genome region'*. This evolution has met with some resistance from the mapping community (Kingsbury, 1996). In IXDB, the term locus has been maintained in a first phase, despite its lack of rigorous definition. It allows the database to start storing information immediately, and the biologist to find a familiar place to look for genes and D-number markers. IXDB is a database that stores experimental results, and this concept in principle should resolve ambiguities. In a second phase IXDB has therefore also moved away from the term 'Locus', by assigning each such object to a more specific class. For instance, some laboratories have submitted YAC clones to the GDB as loci and hence have been given D-numbers. These objects are now stored in IXDB as YAC clones under the original YAC name, and the D-number has become a synonym. Another example are the numerous STS markers that have been assigned D-numbers by GDB, and can more conveniently be stored in IXDB as STSs under their original name, with the D-number as synonym.

Relationships between former 'Loci' and other objects such as clones or genes are clarified in the same procedure. Such relationships can be classified into the following types: 'is equal to', 'overlaps with', 'is derived from', 'contains', and 'is contained in'. Assigning a relationship between two objects requires that the types of the objects are known. For instance, a relationship is known to exist between a locus (an STS in this case) and a YAC clone, and must be classified into one of the above type. All STSs can only be contained in YAC clones (or in addition be derived from the YAC clone). Therefore the knowledge that the locus is an STS immediately allows the deduction that the type of relationship is: Locus 'is contained in' YAC. Furthemore, it is possible to systematically assign this relationship to all STSs and all YACs, provided a relationship is known to exist. This kind of assignment is impossible if several object types (genes, clones, STSs) are stored under the term locus, and if in addition a single locus name can correspond to an STS, and a cloned DNA fragment simultaneously.

The strategy consisting of storing experimental results, and of using the results to navigate in the data, contains at least temporarily an inherent drawback. In the X chromosome project described here, the most important information to help construct the YAC map concerned the marker content of YAC clones. Although a description of the experimental procedure that led to assigning the markers to the clones was

available (mostly on paper), only the final result - clone X contains marker Y- was entered in the ACEDB database, and hence was transferred to the ORACLE version. Therefore the experimental link between the clone and the marker (e.g. which type of probe was used, in which type of experiment) is missing. This situation was acceptable as long as the data was used internally for map construction. It is however a problem when the information is transferred into a publicly available database such as the ORACLE version of IXDB. Here, experimental links are explicitly described, and are systematically used to link a marker to a clone. In the current version of IXDB (version 1.0, Dec. 1996) many such links are empty, and replaced by an "Unknown" tag to reflect this situation. This situation will affect the database temporarily, until the complete details are translated from their current paper format to IXDB.

4.3 Future developments

The next phase of development will only concern the ORACLE version of IXDB. Its focus will be placed on establishing a network between the 17 participants of the X chromosome European consortium. Several new features will be built to deal with the situation where several users will need access at the same time for write and read, and will need the possibility to keep information confidential. As well as being a database, IXDB is meant to be a working environment to build physical and transcript maps. To this end, a graphical tool has been written by A. Grigoriev (MPIMG) in JAVA (Gosling and McGilton, 1995) that enables users to visualise maps locally but which reads data from the database at the MPIMG. Extensions to this tool are being written that will allow users to edit the maps, and for the changes to be saved in IXDB.

Other projects within the laboratory are in the process of implementing specific databases in ORACLE based on the IXDB model. This will be done in a coordinated fashion in order to maintain compatibility. These include the mouse mapping, gene catalogue (oligo-fingerprinting) and chromosome 21 mapping projects.

5. The World Wide Web (WWW) and the X chromosome project

The need for a comprehensive WWW site to present our X chromosome project publicly, arose when it became clear that it would not be possible to publish the YAC contig map on paper. Consequently, three different representations of the map have been made publicly available. First, images of a complete representation of the chromosome with individual clones is on the server, and can be expanded in a region of interest by clicking with the mouse. Secondly, the complete map and associated data, formatted for ACEDB, is placed on the MPIMG anonymous ftp

server, and can easily be reached via the WWW site (figure 6.3A). In addition to the map, information on the cX YAC collection is available (history, description of content and availability). IXDB in ORACLE has recently been made available from the same location (figure 6.3B). It provides a third way to view the map, through the JAVA applet (DerBrowser) described above. The WWW site was created in March 1996 and as of 07-12-1996 has been accessed over 3100 times by more than 600 different visitors.

6. Discussion and conclusion

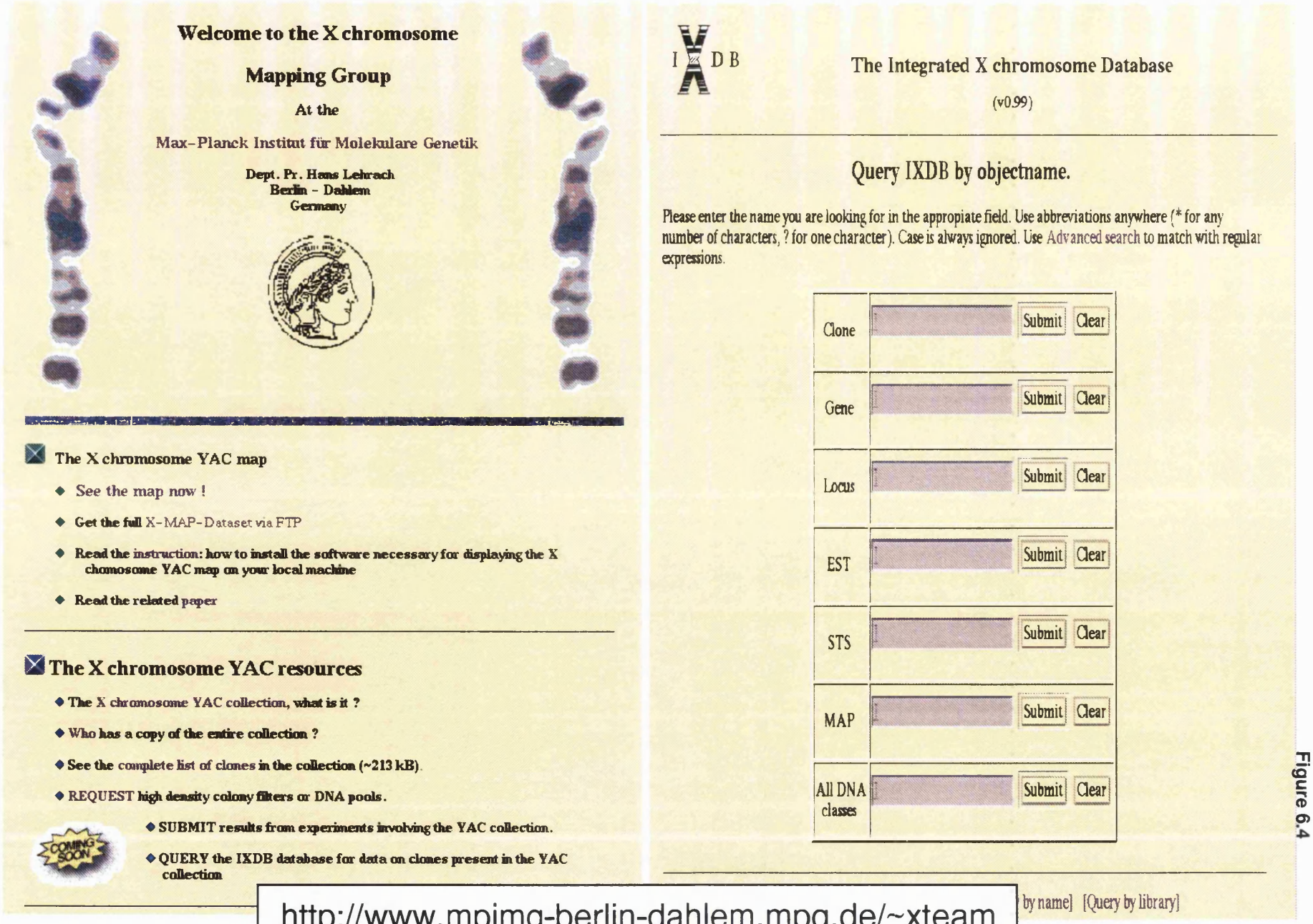
IXDB was first developed in ACEDB as an internal lab database, and in a second step it became a means to release the YAC map of the X chromosome publicly. Its content has now been transferred to a second version of IXDB in ORACLE with an interface on the WWW, and which is used for the same purpose as the first IXDB, in addition to storing data from the cX YAC collection and interconnecting members of the European X consortium.

ACEDB is a software designed to be used primarily by biologists, and an effort has been made by the authors of the program to make this possible without the need for an extensive knowledge of Unix or any programming language. The ACEDB database management system is intuitive, and so is the navigation in the database, since it essentially relies on graphical displays and mouse clicks. Due to these features, ACEDB has had a real impact on the way many biologists handle their mapping projects. Over 50 different ACEDB databases exist, mostly in plant genome mapping (source: ACEDB frequently asked questions: http://probe.nalusda.gov:8000/acedocs/acedbfaq.html). Another advantage of the ACEDB system is the close interaction that exists between ACEDB curators and more importantly between the latter and the authors of the program, R. Durbin and J. Thierry-Mieg. A bionet newsgroup exists (bionet.software.acedb) which provides a platform for discussion, and often acts as a 'helpdesk' for solving any problem related to the software. The authors are continuing to develop ACEDB, for instance with an improved query language (The Sanger Centre) or a JAVA applet as interface for ACEDB on the WWW (JADE, J. Thierry-Mieg, L. Stein, http://alpha.crbm.cnrs-mop/fr/jade/jade.html).

For the past 6 years ACEDB has been an important mechanism for the Sanger Centre to distribute data on the *C. elegans* mapping project to members of the scientific community. This is however a 'one way' system, where users download regular updates but cannot contribute or annotate data except on their local copy, or by contacting the curators who will incorporate the new data in the public release. ACEDB therefore meets its first limits when a database project has to expand from being an internal lab tool to being a communication tool between several laboratories, where both concomitant read/write access must be allowed over the internet. Since the

JAVA applet for the WWW interface is still in development, ACEDB also currently lacks any form of interactive query interface on the WWW, which is the most popular and convenient way of accessing information for today's biologist. In addition, if the current version of the ACEDB code would allow concomitant read/write access for multiple users and support a WWW interface, it would lack a number of features that would make these tools safe from the database management system point of view.

For these reasons IXDB has been re-implemented in ORACLE in order to adapt to the needs of the X chromosome project. Until the ORACLE version is fully operational, the ACEDB version will be maintained and will remain accessible but not updated. The advantages of using ORACLE lies in its very robust safety features and built-in routines to handle multiple users at the same time. It is also extremely well supported, since it is a commercial product. The number of users of this software are countless, since approximately 60 % of all commercial and administrative entities that use a database in Europe use ORACLE (source: ORACLE Corporation, Redwood City, USA). As for ACEDB, electronic newsgroups for ORACLE exist (comp.databases.oracle) that provide online support for and by users. It could be said that ORACLE has its strength where ACEDB has its weaknesses (Table 2). On the other hand ORACLE lacks many of the features that have made ACEDB such a successful software for biologists. Developing a database in ORACLE requires a good background knowledge of the Unix environment as well as programming abilities. It has no built-in graphical displays or graphical navigation tools, and the relational model that it is built upon does not follow a logical structure from a biological point of view. However, as opposed to ACEDB, developing such tools is possible without modifying the ORACLE code, and can therefore be performed independently. The possibility of developing new tools and customising ORACLE for the X chromosome project, is the real advantage of ORACLE over ACEDB. For instance, a graphical browser that displays maps stored in IXDB has been released (DerBrowser, A. Grigoriev, MPIMG) and provides this essential feature that was lacking since IXDB moved to ORACLE. The object-oriented description of genomic data using the OPM system (Chen and Markowitz, 1994) is likely to become a convenient way of enabling independent databases to interconnect their content. It is planned in the future of IXDB to represent its content in this way.



Welcome to the X chromosome

Mapping Group

At the

Max-Planck Institut für Molekulare Genetik

Dept. Pr. Hans Lehrach
Berlin - Dahlem
Germany

The X chromosome YAC map

- See the map now !
- **Get the full** X-MAP-Dataset via FTP
- **Read the instruction: how to install the software necessary for displaying the X chomosome YAC map on your local machine**
- **Read the related paper**

The X chromosome YAC resources

- **The X chromosome YAC collection, what is it ?**
- **Who has a copy of the entire collection ?**
- **See the complete list of clones in the collection (~213 kB).**
- **REQUEST high density colony filters or DNA pools.**
- **SUBMIT results from experiments involving the YAC collection.**
- **QUERY the IXDB database for data on clones present in the YAC collection**

COMING SOON

IXDB

The Integrated X chromosome Database

(v0.99)

Query IXDB by objectname.

Please enter the name you are looking for in the appropiate field. Use abbreviations anywhere (* for any number of characters, ? for one character). Case is always ignored. Use Advanced search to match with regular expressions.

Clone		Submit	Clear
Gene		Submit	Clear
Locus		Submit	Clear
EST		Submit	Clear
STS		Submit	Clear
MAP		Submit	Clear
All DNA classes		Submit	Clear

by name] [Query by library]

http://www.mpimg-berlin-dahlem.mpg.de/~xteam

Figure 6.4

CHAPTER SEVEN: Discussion and Future Perspectives

The work described in this thesis was designed to contribute to the physical mapping of the human X chromosome. A fingerprinting method for YAC contig construction was developed, which included the assessment of different DNA hybridisation systems. Analysis of experimental data from several sources (including data generated by others), lead to a first generation YAC contig map of the chromosome. This map was supported by collating information related to the X chromosome, which necessitated the development of two databases dedicated to this chromosome.

1. Large scale projects versus small scale projects

As reviewed in Parish and Nelson (1994), Kass and Batzer (1995), and Coffey et al. (1996), Alu-PCR has often been used as a rapid way to isolate DNA from complex sources such as radiation hybrid cell lines (Nelson et al., 1989) (Monaco et al., 1991) or YAC clones (Nelson et al., 1991) (Chumakov et al., 1992; Chumakov et al., 1995), or to generate a unique fingerprint specific to the DNA source (Coffey et al., 1996). Alu PCR products are often used as hybridisation probes. One problem associated with the use of these PCR products in hybridisations is the presence of part of the Alu sequence at the extremities of each product, which is a potential source of non-specificity. The Ale3 primer used in this study is 20 bases long, and its 3' end is situated 23 bases away from the 3' end of the consensus Alu sequence (Kariya et al., 1987) (figure 3.4). Each PCR product therefore contains 86 bp of sequence contributed by the Alu repeat. On average, approximately 20 distinct PCR products could be amplified from each YAC clone using Ale3, ranging from approximately 60 to 3000 bp in size. Visual inspection of a sample such as that shown in figure 3.8 indicates that the products are evenly distributed in this range. The total amount of DNA typically amplified in a YAC Alu-PCR is therefore approximately 30 kb, of which 1.7 kb are contributed by the Alu repeat, or 5.7%. This figure is about half of the predicted 10-15 % for human genomic DNA (Britten et al., 1988). In this context, a YAC Alu-PCR probe is not expected to be more difficult to hybridise or lead to a higher background than a genomic probe of similar length (e.g. cosmid insert) provided it has been properly pre-reassociated with non-labelled DNA.

Large scale experiments that have been performed or described in this thesis are strongly focused on the use of Alu-PCR. All hybridisations of YACs, XPL clones, or radiation hybrids were based on this method, and even fragments for gel fingerprints were generated in this way. A second characteristic of these experiments is the relatively high level of equivocal data that was generated, and that had to be identified and removed from the analysis. It is therefore tempting to associate these two features, and conclude that experiments based on Alu-PCR are not robust. As discussed above, there are no reasons why this should be, and one should therefore look elsewhere for an explanation as to why these large scale experiments produced a high level of unsatisfactory data. Other large scale projects do not escape this problem. For instance, a rapid examination of the raw data produced by the CEPH-Genethon YAC mapping projects (http://www.cephb.fr/bio/infoclone.html) show that each experiment (YAC Alu-PCR hybridisation, STS mapping, L1 fingerprint) also yields a high percentage of inconsistent information. A similar review of raw data produced by the Whitehead/MIT groups when constructing another STS/YAC map of the human genome shows the same 'noise' in the data (http://www-genome.wi.mit.edu:80/cgi-bin/contig/yac_info). In both cases this is reflected by acknowledged difficulties to analyse the data, and in some inaccuracies in published results ([Chumakov, 1995 #1237; Hudson, 1995 #1351] respectively; see discussion in chapter 5 for X chromosome related errors). These examples support the view that problems of data robustness in large scale experiments is not due to the techniques employed, but to the large scale aspect itself.

When small scale mapping achievements are reported in the literature, such as YAC contigs over a few megabases, it is rare that any finding is latter refuted by a different approach. The small scale of the approach allows for experiments to be repeated when results are ambiguous, or for extra experiments to be performed in weak regions. In practice, small scale experiments generate the same amount of noisy data, but problems are solved as the investigation progresses. When a publication is finally released, it does not contain any of the inconsistent data. In a large scale project, repeating experiments that produce ambiguous results, or verifying a weak result with revised experiments is often not possible. The reason for this lies in the difficulty in performing targeted experiments in such projects, as opposed to systematically applying a few techniques to a large number of samples. In addition, large scale enterprises tend to release a polished and condensed summary of their results in the literature, while the complete raw data is also available in electronic form, revealing the gradient of quality. Therefore, large scale projects do not generate more inconsistent data than any other small or medium scale projects. The reason why it is often seen as such is therefore likely to be due to the fact that in the former, equivocal data is acknowledged and revealed, while in the latter it is kept in the lab book.

2. STS versus hybridisation mapping

The use of Sequence Tagged Sites (STSs) as a general mapping technique was first proposed in 1989 (Olson et al. 1989), and has since been taken on by the U.S. Human Genome Project as the core technique for physical and genetic mapping of the human and mouse genomes. Mostly due to the strong influence of the US project on the global, world-wide Human Genome Project, this approach has been naturally followed by most other genome centres in Europe and Asia. The main advantages of this approach are its simplicity and reproducibility. Amplifying by PCR a short fragment of DNA with a unique set of primers is technically straightforward and primer sequence can be easily placed in common repositories, printed in publications, or send by FAX and mail thus virtually acting as a 'common language' between scientists, as was originally proposed.

Due to this simplicity it is amenable to automation and hence applicable to large scale mapping projects. This is exemplified by recent achievements from the Whitehead/MIT and CEPH/Généthon groups (Hudson et al., 1995, Chumakov et al. 1995, Dib et al. 1996). The cost of the technique is relatively high, since it requires Taq polymerase and the synthesis of many different oligonucleotides, but it is still well within the price range of most molecular biology protocols. The same principle is now being applied to cDNA clones, from which ESTs are derived (Adams et al. 1991), and has recently yielded the first human genome transcript map (Schuler et al. 1996).

Considering these impressive achievements and the widespread use of the STS/EST mapping technique, it is difficult to find many disadvantages to this method. In physical mapping, it is however clear that the use of STSs, each representing a single point in the genome, is a very inefficient and time consuming way of screening a clone library. For instance, to map a single STS to a YAC clone that is part of a whole genomic YAC library, as in the large scale projects mentioned above, over 50 PCR reactions must be performed. In addition, there is no guarantee that a positive clone will be unambiguously identified, even if such a positive clone exists in the library. The information yield of such a screen is low, since it only results in matching one probe with one clone.

Hybridisation-based screening of libraries has been proposed (Lehrach et al. 1990) as an alternative to PCR screenings with STSs. This method requires that libraries are arrayed as colonies or DNA spots on a solid support (e.g. nylon filters), ideally at high densities and with the use of robotic spotting machines. Since the preparation of such material is not possible for every laboratory, this scheme has been placed in a broader context of a 'Reference Library System' (Zehetner and Lehrach, 1994) where such library filters are distributed to the community, and results pooled in a public database.

Hybridisation-based screening with a single-copy probe is theoretically much more efficient and cost-effective than PCR screening with an STS. In a single

hybridisation, all potential positives present in the library can be identified. This approach is also amenable to the use of probes of varying complexities, from locally complex probes (pools of Alu-PCR products from a YAC clone) to dispersed complex probes (e.g. pools of cDNAs). It is also possible to combine the advantages of the fingerprinting techniques with the efficiency of hybridisations (Lehrach et al, 1990), for instance in sequencing by hybridisation strategies ((Drmanac et al., 1993), Meier-Ewert et al. unpubl.). Hybridisations on library filters also allow the retrieval of genomic clones of one species with a probe from a related but different species (e.g (Baxendale et al., 1995)). Although the hybridisation approach is elegant, versatile and in principle efficient, it suffers from two disadvantages. First, it is a technically demanding method, which is still mostly based on the use of expensive and hazardous isotopic labels. Technical difficulties include the presence of frequent and moderate repeats in human genomic DNA, that must be blocked with non-labelled DNA prior to hybridisations, and which are a source of potential false positives. Second, reproducing a hybridisation requires that the DNA probe be physically shared between two laboratories. More importantly it requires an abundance of library filters, a bottleneck that has been an important factor responsible for the restricted use of the hybridisation method as a general mapping technique.

However, hybridisation protocols have been simplified and optimised, and many national resource centres and commercial companies now distribute high density library filters. A gradual shift from YAC based mapping to the construction of maps using clones propagated in *E. coli* may also lead to a wider use of the hybridisation technique. With the availability of good quality YAC maps of most human chromosomes, screening PAC, P1, BAC or cosmid libraries with mapped YAC probes is a straightforward method to identify redundant sets of clones useful to derive such maps. The increased emphasis placed on comparative genomics is also likely to increase the use of library filter hybridisations in genome mapping.

Both the STS-based and hybridisation-based screening approaches have advantages and disadvantages, depending on the application and laboratory set-up. Consequently, many laboratories use both approaches in a complementary way.

3. The X chromosome map: from YACs to PACs

3.1 Introduction

The X chromosome mapping project described in this thesis resulted in 24 YAC contigs spanning 125 Mb of the chromosome, and is therefore not complete. However, while gap closure is ongoing, the current status of the map is extremely useful to start constructing higher resolution maps. Considering the ultimate goals of mapping the human genome, such maps are essential for a variety of reasons. Maps are tools that

enable first the positioning and then identification of human genes. By extension, maps are also the basis for completely sequencing the human genome. For both purposes, YAC clone maps can only be intermediates. YAC clones can accommodate very large fragments of human DNA, and are therefore well suited to cover large regions of the genome with a minimum of effort. However studying a 100 kb human gene with a YAC clone that is five to ten times larger requires manipulating a large excess of DNA. In addition, YAC clones are often subject to internal rearrangements and suffer from a high level of chimerism in libraries. YACs have also a very similar structure to the natural yeast chromosomes, and this prevents the purification of the insert DNA easily. Although sequencing directly from a shotgun library made from a YAC clone is possible (Chen et al. 1996) their large insert size can also make assembling of shotgun clones difficult. In addition, due to the high level of internal rearrangements observed (e.g. Bates et al. 1992), it is questionable whether YAC clones reflect sufficiently well human genomic DNA to be a reliable source of material for sequencing.

3.2 Other cloning systems

Several other systems have been designed to accommodate pieces of DNA which are smaller than YAC inserts, but are propagated in *E. coli* for ease of manipulation. These can be called intermediate-capacity cloning systems, since they bridge the gap between the 40 kb cosmid insert size and the larger YAC insert which can be greater than 1Mb. The P1 cloning system (Sternberg, 1990) has been developed more recently than either the YAC or cosmid systems, and two human genomic libraries have been constructed with this system (Francis et al., 1994c) (Shepherd et al., 1994). The P1 system packages linear recombinant DNA into phage particles, followed by injection into *E. coli* and circularisation of the DNA using P1 loxP recombination sites and a host expressing the P1 Cre recombinase. A P1 plasmid replicon is used for maintaining the vector at one copy per cell thus avoiding high-copy-number instability of clones. A different cloning system in *E. coli* is the bacterial artificial chromosome (BAC, Shizuya et al., 1992). It uses an F-factor based vector and can accommodate inserts up to 300 kb. BACs are transformed into *E. coli* as circular molecules by electroporation, and are similarly maintained at a low copy number in the cell. More recently a system has been described which combines several of the features of the P1 and BAC cloning systems, called PAC (P1-derived artificial chromosome, Ioannou et al., 1994). The vector retains most of the properties of the P1 cloning vector, however circular recombinant DNA is electroporated into *E. coli* cells. The phage headful constraints on insert size of the P1 system are thus eliminated, and average insert sizes of 130-150 kb have been attained. A large human library has been constructed with this system (P. de Jong, unpublished) and has been freely and widely distributed to many research centres. The library now

comprises approximately 450.000 clones and represents a 7.5 fold coverage of the human X chromosome. It has a disadvantage over BAC libraries as a source of material for sequencing, since the 16 kb vector is large compared to the 5 kb of the BAC vector (although comparable to sequencing in cosmids), and is therefore an added cost in shotgun sequencing strategies.

3.3 Construction of an X chromosome map in PAC clones

The PAC library described above has many advantages over other existing genomic libraries that have contributed to selecting it for the X chromosome mapping project in our laboratory. It is used in parallel by several other groups in the X chromosome community, which greatly facilitates the integration of results. It has been made available at an early stage enabling its use in a wide range of mapping projects, which have confirmed several properties of the system, including low chimerism, low frequency of recombination, and good representation of genomic DNA used for cloning (P. de Jong, pers. comm).

The strategy adopted to construct a PAC map of the X chromosome relies on the YAC contig map described in this thesis work. YAC Alu-PCR products selected from a minimum tiling path of the contigs are amplified using a combination of 2 Alu and one Line primer, and radioactivelly labelled. Probes are competed to block repetitive elements and hybridised to high density colony filters of the PAC library. In this way, PAC clones are rapidly mapped to defined regions of the X chromosome. In addition, each YAC clone produces a different pattern of positive PAC clones, which can be compared to confirm overlaps. This project is currently ongoing, and will produce a 'pocket' map of the chromosome, similar to that already constructed for human chromosome 21 (Nizetic et al., 1994). In parallel, end probes prepared from the PAC clones are hybridised back to the PAC library and to X chromosome cosmid libraries ((Nizetic et al., 1991) and Lawrence Livermore Laboratories) to refine the pockets into contigs and enrich the coverage in *E. coli* propagated clones.

The construction of such map is an intermediate step towards the development of a transcript map of the chromosome, and ultimately deciphering its complete DNA sequence.

4. The need for databases in the Human Genome Project

In a first approximation, developing a database can be merely compared to the construction of shelves for a library. It facilitates the physical arrangements of objects, allows them to be separated according to pre-established criteria and ultimately is designed to facilitate their display and their retrieval. In this respect, developing a specific database using existing software tools is not a subject of scientific interest but

rather can be considered as a technical achievement. However, when the task of a database is to represent a 'domain' that is concerned with the biology of the human organism at the molecular level, this view may need to be revised.

Molecular biology is a fast moving field which has earned a great deal of attention in the past ten years since the first outline of the Human Genome Project was drawn up. It is a field that is concerned with answering questions related to human diseases, biological diversity and the functioning of the human body. It may even answer questions on brain functions, such as the biological basis for processes like memory. It is currently possible to relate various pieces of information and make deductions based on results that have been generated so far, although it is mostly restricted to very specialised questions, in specialised areas. For instance, the search for a biologically active molecule in the treatment of a monogenic disease can be considered as a relatively specialised problem. It is influenced by a limited number of factors, such as the possible interactions with the endogenous molecules involved in the metabolic pathway affected by the disease, and the secondary effects of candidate molecules etc. Another example is the search for a gene responsible for a monogenic disease. It follows one or more strategies but all are concerned with a very specific phenotype, set of patients, region of the genome, etc. Other questions however tackle more global problems or problems that have several causes that can not be separated from each other in the first place. An example is the mapping of polygenic diseases, such as hypertension and diabetes, which are caused by multiple genes interacting with each other and with environmental factors to create a gradient of susceptibility to the disease. Human geneticists confronted with such tasks must consider, measure and weigh a series of factors that span more than one field of biology, and take in account a large volume of data to obtain statistical significance. Another example is the analysis of protein function, once the corresponding gene has been identified and decoded. It will increasingly involve the comparison of protein sequences across many species for which representative model organisms have been studied in much greater detail than the human. With the recent publication of the complete sequence of the yeast genome, and the growing amount of characterised mutations in the bacteria, fruit fly, mouse or zebrafish this becomes increasingly possible. The colossal amounts of information that biologists have generated so far already makes answers to some of these complex questions a realistic prospect.

In this context, the development of a database that allows investigators to take a global look at a particular field takes much more importance. When the field that must be covered is in constant and rapid evolution, it becomes a challenge. When in addition the database is to be a working tool for analysis programs whose tasks will be to match objects and make deductions, it becomes a scientifically meaningful problem.

The X chromosome mapping project and its associated database which are described in this thesis follow such a pattern. It is clear that without the help of

electronic storage and retrieval facilities, it would not have been possible to construct the YAC map that is described here. On the other hand, it is also clear that analysis software that attempted to make sense of the experimental data reached their limits early in the project. The main reason for this is the heterogeneous format in which much of the data was stored, ranging from lists on paper and remote database information to ACEDB graphics. IXDB in its ORACLE version is aimed at solving this problem, by providing the means to store and classify information in a single homogeneous format. Only then will it be possible to ask the global questions that will take into account more than one data type, from more than one source. Another reason for the inability of computer software to construct the map is also linked to the complexity and heterogeneity of the data itself, a factor that is likely to become worse as progress is made, further justifying the need for databases in the Human Genome Project.

5. Future Development of IXDB

IXDB currently stores information on 12.000 YAC clones mapped to the X chromosome, and has links to several other major databases such as GDB and OMIM. The X chromosome project is progressing from the YAC map to a higher resolution map in clones propagated in *E. coli* (see above), on which a transcript map will be developed. The volume and the complexity of the information increases rapidly, and is more and more the result of combined efforts from many groups rather than a single large scale enterprise. In this context it is not yet clear in which direction IXDB will evolve. There are several alternatives that can be considered, and a combination of many factors will decide which path IXDB will follow. At the 7th XCW, the five editors appointed by HUGO to establish a consensus marker map and write the workshop report have acknowledged that the volume of mapping data on the X chromosome has increased to an extent that curating becomes an impossible task. This is partly due to the fact that no repository exists for this mapping information in an electronic format. The GDB partially fulfils this role but the nature of the data stored in GDB and the way submissions are handled makes the integration of data and the establishment of a consensus extremely difficult. IXDB, which is only concerned with the X chromosome, has put a strong focus on storing experimental data with no attempts at resolving inconsistencies, and is freely accessible. In addition IXDB contains already a large amount of information contributed by external laboratories, and several marker and clones maps covering the entire chromosome. It has therefore the potential to become a 'community database' that would be curated by a number of scientists with interests and experience in specific regions of the chromosome. For this to happen however, it will be necessary to find the necessary resources in manpower for curating and maintaining the database. Also the X chromosome community must reach an agreement that such a community database is needed. This may not be obvious in the light of the motivations that drive the majority of the groups working on

the X chromosome. Most investigators are concerned with small regions containing a gene of interest, and may be satisfied with the current situation of multiple databases and heterogeneous accession modes. A unified repository of information that encompasses the entire X chromosome has an immediate value only to those groups with a global interest or with very large regions under investigation. Such a project certainly would obtain the support of the Genome Database, since it would guarantee a minimum of quality and consistency in a large volume of data that could be integrated regularly with the rest of the information on the human genome. At the other extreme, and in case the X community feels that a community database is not needed, IXDB would remain an essential tool for a more restricted number of laboratories. These would be the groups that participate in the European consortium and some close collaborators. Maintenance and curation would be performed essentially by our group, although data would also be submitted directly by individual labs. In this model, the only public part would be the information on the YAC collection, the consensus maps and related entities. The rest, which would be provided directly by members of the consortium, would only be visible by the latter. Between these two extremes, there may be room for alternatives that combine a part of both models.

6. The X chromosome transcript map and the European Consortium

Our laboratory is coordinating a consortium of 17 independent European groups with strong interests in defined regions of the X chromosome. This project has recently started (July 1996) and aims at constructing a transcript map of the X chromosome. It is based on a strong interaction between regional expertise provided by laboratories with long term interests in specific regions, and a large scale approach adopted by our laboratory. Some resources will be shared, such as clone libraries, facilitating the integration of results from each partner on a single map. Gene identification will be performed by exon-trapping and cDNA selection methods, as well as targeted genomic sequencing. One strong element of the project is a cDNA library enriched for X chromosome genes in construction in the laboratory of A. Poustka in Heidelberg. It will be distributed to all partners, and is expected to contain most X chromosome genes in a compact format. The task of our group is to produce high density filters of the enriched cDNA library and to screen it with genomic fragments selected from the PAC and cosmid contigs. In addition, the cDNA library will be characterised by oligo-fingerprinting, a technique which involves the hybridisation of short oligomers to arrayed cDNA libraries (Meier-Ewert et al., unpublished). This will generate information such as partial sequences useful for low stringency homology searches in databases, and cDNA clustering to reduce problems of over-representation of cDNA clones in the library.

The project will closely interact with other international efforts such as the EST mapping projects. ESTs assigned to the X chromosome will be placed on the PAC and cosmid map by hybridisation. Sequencing the X chromosome is another project that the community has recently initiated in a coordinated fashion.

7. Sequencing on the X chromosome

At the 7th International workshop on X chromosome mapping (7th XCW, Hinxton, UK, 1996), a concerted plan to divide the chromosome between different groups aims at completing most of the sequence within a few years. Some groups such as the Sanger Centre, the St. Louis sequencing centre and the Baylor College of Medicine have taken it up on them to sequence approximately 80-90 Mb, while the rest was shared among smaller groups. To date (December 1996) approximately 1 % of the human genome has been sequenced, most of which on the X chromosome (7.5 Mb as of August 1996, Source EMBL and IXDB).

8. Conclusions

The human X chromosome has a number of characteristics that make its study particularly challenging but also exciting. Every month, reports in the literature unveil new advances on X-inactivation, detailed transcript maps, cloning of disease genes, or sequence analysis of ever larger regions. The level of competition is high in these areas, and stimulates progress even further. The complexity and volume of the information that is generated also means that laboratories must stay somehow connected to keep track of this progress, and often turn competition into collaboration.

The next few years will see the fruits of the long era of physical and genetic mapping of the genome. As the resolution of these maps increase, regions of homology between the Y and the X chromosome, and comparative studies with other mammals and chordates will reveal the key evolutionary steps that have shaped the animal world. More importantly, the ~ 65.000 genes that lie in the human genome will be deciphered. The real challenge will start then, with the problems of analysing and using the information. Social and commercial issues have already been raised as to how and for what purpose data from the Human Genome Project should be used (for a review see Knoppers and Chadwick, 1994) Lapham et al. 1996). National and international ethical committees have been created to promote discussion and establish guidelines for the creation of laws that will hopefully provide the required legislative background for the use of genetic information on human beings. It would not do justice to the complexity of these issues to review them here. The only clear principle that has reached a common agreement is the respect for the integrity and

dignity of the human individual. How this will be interpreted across national cultures and religions is another matter.

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., Mccombie, W. R., and Venter, J. C. (1991). Complementary-DNA sequencing - expressed sequence tags and human genome project. Science *252*, 1651-1656.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., and Sutton, G. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequences. Nature *377 Suppl.*, 3-174.

Adams, M. D., and Venter, J. C. (1996). Should non-peer-reviewed raw DNA sequence data release be forced on the scientific community. Science *274*, 534.

Al-Hakim, H. H., and Hull, R. (1988). Chemically synthesized non-radioactive biotinylated long-chain nucleic acid hybridisation probes. Biochemical Journal *251*.

Alderton, R. P., Kitau, J., and Beck, S. (1994). Automated DNA hybridization. Analytical Biochemsitry. *218*, 98-102.

Amasino, R. M. (1986). Acceleration of Nucleic Acid Hybridisation Rate by Polyethylene Glycol. Analytical Biochemsitry *152*, 304-307.

Anand, R., Villasante, A., and Tylersmith, C. (1989). Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed-field gel-electrophoresis. Nucleic Acids Research *17*, 3425-3433.

Ansorge, W., Sproat, B., Stegemann, J., Schwager, C., and Zenke, M. (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Research *15*, 4593-602.

Ballabio, A., Bardoni, B., Carrozzo, R., Andria, G., Bick, D., Campbell, L., Hamel, B., Fergusonsmith, M. A., Gimelli, G., Fraccaro, M., Maraschio, P., Zuffardi, O., Guioli, S., and Camerino, G. (1989). Contiguous gene syndromes due to deletions in the distal short arm of the human X-chromosome. Proceedings Of The National Academy Of Sciences Of The United States Of America *86*, 10001-10005.

Bassi, M. T., Schiaffino, M. V., Renieri, A., De Nigris, F., Galli, L., Bruttini, M., Gebbia, M., Bergen, A. A., Lewis, R. A., and Ballabio, A. (1995). Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome. Nature Genetics *10*, 13-19.

Bates, G. P., Valdes, J., Hummerich, H., Baxendale, S., Le, Paslier Dl, Monaco, A. P., Tagle, D., MacDonald, M. E., Altherr, M., Ross, M., et al., (1992) Characterization of a yeast artificial chromosome contig spanning the Huntington's disease gene candidate region. Nature Genetics *1* ,180-7

Bates, G. (1996). Isolation of YAC ends by plasmid rescue. Methods in Molecular Biology *54*, 139-44.

Baxendale, S., Abdulla S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., Beck, S., Lehrach, H. (1995). Comparative sequence analysis of the human and pufferfish Huntington's disease genes. Nature Genetics *10*, 67-76.

Beck, S., and Koster, H. (1990). Applications of dioxetane chemiluminescent probes to molecular biology [published erratum appears in Anal Chem 1991 Apr 15;63(8):848]. Analytical Chemistry *62*, 2258-70.

Bell, J., and Haldane, F. R. S. (1937). Proc. R. Soc. London Ser. B. *123*, 119.

Bentley, D. R. (1996). Genomic Sequence Information should be released immediately and freely in the public domain. Science *274*, 533.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. [Review]. American Journal of Human Genetics *32*, 314-31.

Britten, R. J., Baron, W. F., Stout, D. B., and Davidson, E. H. (1988). Sources and evolution of human Alu repeated sequences. Proceedings of the National Academy of Sciences of the United States of America *85*, 4770-4.

Bronstein, I., Voyta, J. C., Murphy, O. J., Tizard, R., Ehrenfels, C. W., and Cate, R. L. (1993). Detection of DNA in Southern blots with chemiluminescence. Methods in Enzymology *217*, 398-414.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human Xist: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542.

Brown, C. J., Lafreniere, R. G., Powers, V. E., Sebastio, G., Ballabio, A., Pettigrew, A. L., Ledbetter, D. H., Levy, E., Craig, I. W., and Willard, H. F. (1991). Localization of the x-inactivation center on the human x-chromosome in xq13. Nature *349*, 82-84.

Bull, J. J. (1983). Evolution of Sex determining mechanisms (Menlo Park, CA, USA: Benjamin Cummings).

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Venter, J. C., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science *273*, 1058-73.

Burgoyne, P. S. (1982). Genetic homology and crossing over in the X and Y chromosomes of mammals. Human Genetics *61*, 85-90.

Burke, D. T., Carle, G. F., and Olson, M. V. (1987). Cloning of large segments of exogenous dna into yeast by means of artificial chromosome vectors. Science *236*, 806-812.

Cano, R. J., Torres, M. J., Klem, R. E., and Palomares, J. C. (1992). DNA hybridization assay using Attophos, a fluorescent substrate for alkaline phosphatase. Biotechniques *12*, 264-267.

Carothers, A. M., Urlaub, G., Mucha, J., Grunberger, D., and Chasin, L. A. (1989). Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. Biotechniques *7*, 494-6.

Charlesworth, B. (1991). The evolution of sex chromosomes. Science *251*, 1030-33.

Chen, C., Su, Y., Baybayan, P., Siruno, A., Nagaraja, R., Mazzarella, R., Schlessinger, D., Chen, E. (1996) Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches. Nucleic Acid Research *24,* 4034-41

Chen, I. A., and Markowitz, V. M. (1995). An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools. Information Systems *20.*

Cherry, J. L., Young, H., Di Sera, L. J., Ferguson, F. M., Kimball, A. W., Dunn, D. M., Gesteland, R. F., and Weiss, R. B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. Genomics *20*, 68-74.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., Gervy, P., Legall, I., Soularue, P., and Grinas, L. (1992). Continuum of overlapping clones spanning the entire human chromosome-21q. Nature *359*, 380-387.

Chumakov, I. M., Rigault, P., Le Gall, I., Bellané-Chantelot, C., Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros, I. et al. (1995). A YAC contig map of the human genome. Nature *377 Suppl.*, 175 - 297.

Coffey, A., Gregory, S., and Cole, C. G. (1996). Alu-PCR fingerprinting of YACs. In Methods in Molecular Biology, D. Markie, ed. (Totowa, N.J.: Human Press Inc.), pp. 97-114.

Cole, C. G., Goodfellow, P. N., Bobrow, M., and Bentley, D. R. (1991). Generation of novel sequence tagged sites (STSs) from discrete chromosomal regions using Alu-PCR. Genomics *10*, 816-826.

Collins, F., and Galas, D. (1993). A new five-year plan for the U.S. Human Genome Project. Science *262*, 43-46.

Collins, J., Cole, C., Smink, L. J., Garett, C. L., Leversha, M. A., Soderlund, C. A., Maslen, G. L., Everett, L. A., Rice, K. M., Coffey, A. J., Gregory, S. G., and Gwilliam, R. (1995). A high-density YAC contig map of human chromosome 22. Nature *377*, 367 - 379.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986). Toward a physical map of the genome of the nematode caenorhabditis-elegans. Proceedings Of The National Academy Of Sciences Of The United States Of America *83*, 7821-7825.

Cox, D. R., Burmeister, M., Price, R., Kim, S., and Myers, R. M. (1990). Radiation Hybrid Mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. Science *250*.

Del Castillo, I., Cohen-Salmon, M., Blanchard, S., Lutfalla, G., and Petit, C. (1992). Structure of the X-linked Kallmann syndrome gene and its homologous pseudogene on the Y chromosome. Nature Genetics *2*, 305-310.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature *380*, 152-154.

Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M. J., Ingalls, K. A., TJ, O. C., Evans, C. A., DeAngelis, M. M., Levinson, D. M., Kruglyak, L., Goodman, N., Copeland, N. G., Jenkins, N. A., Hawkins, T. L., Stein, L., Page, D. C., and Lander, E. S. (1996). A comprehensive genetic map of the mouse genome. Nature *380*, 149-52.

Dietrich, W. F., Miller, J. C., Steen, R. G., Merchant, M., Damron, D., Nahf, R., Gross, A., Joyce, D. C., Wessel, M., Dredge, R., Marquis, A., Stein, L. et al. (1994). A Genetic Map of the Mouse With 4,006 Simple Sequence Length Polymorphisms. Nat. Genet. *7*, 220--225.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H., Marrone, B. L., Longmire, J. L., Lane, S. A., Whitmore, S. A., Lowenstein, M. G., Sutherland, R. D., Mundt, M. O., Knill, E. H., Bruno, W. J., Macken, C. A., Torney, D. C., Wu, J. R., Griffith, J., Sutherland, G. R., Deaven, L. L., Callen, D. F., and Moyzis, R. K. (1995). An integrated physical map of human chromosome 16. Nature *377 Suppl.*, 335-366.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S., and et al. (1987). A genetic linkage map of the human genome. Cell *51*, 319-37.

Dower, W. J., Miller, J. F., and Ragsdale, C. W. (1988). High-efficiency transformation of *Escherichia coli* by high-voltage electroporation. Nucleic Acids Research *16*, 6127-6145.

Drayna, D., and White, R. (1985). The genetic linkage map of the human X chromosome. Science *230*, 753-758.

Drmanac, R., Drmanac, S., Strezoska, Z., Paunesku, T., Labat, I., Zeremski, M., Snoddy, J., Funkhouser, W. K., Koop, B., Hood, L., and et al. (1993). DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing [published erratum appears in Science 1994 Feb 4;163(5147):596]. Science *260*, 1649-52.

Fasman, K. H., Letovsky, S. I., Cottingham, R. W., and Kingsbury, D. T. (1996). Improvements to the GDB Human Genome Data Base. Nucleic Acids Research *24*, 57-63.

Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Analytical Biochemistry *132*, 6-13.

Ferrero, G. B., Franco, B., Roth, E. J., Firulli, B. A., Borsani, G., Delmas-Mata, J., Weissenbach, J., Hailey, G., Schlessinger, D., Chinault, A. C., Zoghbi, H. Y., Nelson, D. L., and Ballabio, A. (1995). An integrated physical and genetic map of a 35 Mb region on chromosome Xp22.3-Xp21.3. Human Molecular Genetics *4*, 1821-1827.

Fields, C., Adams, M. D., White, O., and Venter, J. C. (1994). How many genes in the human genome. Nature Genetics *7*, 345-346.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., and et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Science *269*, 496-512.

Foote, S., Vollrath, D., Hilton, A., and Page, D. C. (1992). The human Y chromosome-overlapping DNA clones spanning the euchromatic region. Science *258*, 60-66.

Forget, B. G. (1993). YAC transgenes: bigger is probably better. Proceedings Of The National Academy Of Sciences Of The United States Of America. USA *90*, 7909-7911.

Francis, F., Benham, F., See, C. G., Fox, M., Ishikawa-Brush, Y., Monaco, A. P., Weiss, B., Rappold, G., Hamvas, R. M. J., and Lehrach, H. (1994a). Identification of YAC and cosmid clones encompassing the ZFX-POLA region using irradiation hybrid cell-lines. Genomics *20*, 75-83.

Francis, F., Rowe, P., Econs, M. J., See, C. G., Benham, F., Oriordan, J., Drezner, M. K., Hamvas, R., and Lehrach, H. (1994b). A YAC contig spanning the hypophosphatemic rickets disease gene (HYP) candidate region. Genomics *21*, 229-237.

Francis, F., Zehetner, G., Höglund, M., and Lehrach, H. (1994c). Construction and Preliminary Analysis of the ICRF Human P1 Library. Genetic analysis: Techniques and Applications *11*, 148-157

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., and et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. Science *270*, 397-403.

Gemmill, R. M., Chumakov, I., Scott, P., Waggoner, B., Rigault, P., Cypser, J., Chen, Q., Weissenbach, J., Gardiner, K., Wang, H., Pekarsky, Y., Le Gall, I., Le Paslier, D., Guillou, S., Li, E., Robinson, L., Hahner, L., Todd, S., Cohen, D., and Drabkin, H. A. (1995). A second-generation YAC contig map of human chromosome 3. Nature *377 Suppl.*, 299-320.

Goffeau, A., Barrel, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 Genes. Science *274*, 546-567.

Gosling, J., and McGilton, H. (1995). The Java Language Environment. Sun Microsystem White Paper *http://java.sun.com/whitePaper/java-whitepaper-1.html*.

Goss, S. J., and Harris, H. (1975). New methods for mapping genes in human chromosomes. Nature *255*, 680-684.

Graves, J. A. M. (1995). The origin and function of the mammalian Y chromosome and Y borne genes - an evolving understanding. BioEssays *17*, 311-317.

Graves, J. A. M., and Watson, J. M. (1991). Mammalian sex chromosomes: evolution of organisation and function. Chromosoma *101*, 63-68.

Green, E. D., and Olson, M. V. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain-reaction. Proceedings Of The National Academy Of Sciences Of The United States Of America *87*, 1213-1217.

Grigoriev, A. V. (1993). Theoretical predictions and experimental-observations of genomic mapping by anchoring random clones. Genomics *15*, 311-316.

Gusella, J. F., Keys, C., Varsanyi-Breiner, A., Kao, F. T., Jones, C., Puck, T. T., and Housman, D. (1980). Isolation and localisation of DNA segments from specific human chromosomes. Proceedings Of The National Academy Of Sciences Of The United States Of America *77*, 2829-2833.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. (1994). The 1993-94 Généthon human genetic-linkage map. Nature Genetics *7*, 246-339.

Haaf, T., and Ward, D. C. (1994). High-resolution ordering of YAC contigs using extended chromatin and chromosomes. Human Molecular Genetics *3*, 629-633.

Harper, R. (1995). World Wide Web resources for the biologist. Trends in Genetics *11*, 223-228.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M. et al. (1996). Generation and analysis of 280,000 Human Expressed Sequence Tags. Genome Research *6*, 807-827.

Höltke, H. J., Ankenbauer, W., Mühlegger, K., Rein, R., Sagner, G., Seibl, R., and Walter, T. (1995). The Digoxigenin (DIG) system for non-radioactive labelling and detection of nucleic acids - an overview. Cellular and Molecular Biology *41*, 883-905.

Houseal, T. W., and Klinger, K. W. (1994). Commentary : What's in a spot ? Human Molecular Genetics *3*, 1215-1216.

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., Slonim, D. K., Baptista, R., Kruglyak, L., Xu, S., Hu, X., Colbert, A. M. E., Rosenberg, C. et al. (1995). An STS-Based Map of the Human Genome. Science *270*, 1945-1954.

Huxley, C., Hagino, Y., Schlessinger, D., and Olson, M. V. (1991). The human hprt gene on a yeast artificial chromosome is functional when transferred to mouse cells b y cell-fusion. Genomics *9*, 742-750.

Hwu, H. R., Roberts, J. W., Davidson, E. H., and Britten, R. J. (1986). Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proceedings of the National Academy of Sciences of the United States of America *83*, 3875-9.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and Dejong, P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nature Genetics *6*, 84-89.

Jablonski, E., Moomaw, E. W., Tullis, R. H., and Ruth, J. L. (1986). Preparation of oligodeoxynucleotide-alkaline phosphatase conjugates and their use as hybridization probes. Nucleic Acid Research *14*, 6115-6128.

Jacob, F., and Monod, J. (1961). Genetic mechanisms in the synthesis of proteins. Journal of Molecular Biology *3*, 318-356.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. Nature *316*, 76-9.

Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., and Matsubara, K. (1987). Revision of consensus sequence of human Alu repeats - a review. Gene *53*, 1-10.

Kass, D. H., and Batzer, M. A. (1995). Inter-Alu polymerase chain reaction: advancements and applications. Analytical Biochemistry *228*, 185-93.

Khrapko, K. R., Lysov Yu, P., Khorlin, A. A., Ivanov, I. B., Yershov, G. M., Vasilenko, S. K., Florentiev, V. L., and Mirzabekov, A. D. (1991). A method for DNA sequencing by hybridization with oligonucleotide matrix. DNA Sequence *1*, 375-88.

Khrapko, K. R., Lysov Yu, P., Khorlyn, A. A., Shick, V. V., Florentiev, V. L., and Mirzabekov, A. D. (1989). An oligonucleotide hybridization approach to DNA sequencing. FEBS Letters *256*, 118-22.

Kingsbury, D. T. (1996). Consensus, common entry, and community curation. Nature Biotechnology *14*, 679-680.

Knoppers, B. M., and Chadwick, R. (1994). The human genome project: under an international ethical microscope. Science *265*, 2035-2036.

Korenberg, J. R., and Rykowski, M. C. (1988). Human genome organization - Alu, Lines, and the molecular-structure of metaphase chromosome bands. Cell *53*, 391-400.

Krauter, K., Montgomery, K., Yoon, S., LeBlanc-Straceski, J., Renault, B., Marondel, I., Herdman, V., Cupelli, L., Banks, A., Lieman, J., Menninger, J., Bray-Ward, P., Nadkarni, P., Weissenbach, J., Le Paslier, D., Rigault, P., Chumakov, I., Cohen, D., Miller, P., Ward, D., and Kucherlapati, R. (1995). A second-generation YAC contig map of human chromosome 12. Nature *377 Suppl.*, 321-334.

Kuykendall, J. R., and Bogdanffy, M. S. (1992). Efficiency of DNA-histone cross-linking induced by saturated and unsaturated aldehydes in vitro. Mutation Research *283*, 131-136.

Lander, E. S. (1996). The new genomics: Global views of biology. Science *274*, 536.

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. Science *265*, 2037-48.

Larin, Z., Fricker, M. D., Maher, E., Ishikawa-Brush, Y., and Southern, E. (1994). Fluorescence *in situ* hybridisation of multiple probes on a single microscope slide. Nucleic Acid Research *22*, 3689 - 3692.

Larin, Z., Monaco, A. P., and Lehrach, H. (1991). Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proceedings Of The National Academy Of Sciences Of The United States Of America *88*, 4123-4127.

Lawrence, J. B., Villnave, C. A., and Singer, R. H. (1988). Sensitive, high-resolution chromatin and chromosome mapping *in situ*: presence and orientation of two closely integrated copies of EBV in a lymphoma line. Cell *52*, 51-61.

Leary, J. J., Brigati, D. J., and Ward, D. C. (1983). Rapid and sensitive colorimetric method for visualizing biotin-labeled DNA probes hybridized to DNA or RNA immobilized on nitrocellulose: Bio-blots. Proceedings of the National Academy of Sciences of the United States of America *80*, 4045-9.

Lee, J. T., Murgia, A., Sosnoski, D. M., Olivos, I. M., and Nussbaum, R. L. (1992). Construction and characterization of a yeast artificial chromosome library for Xpter-Xq27.3 - a systematic determination of cocloning rate and X-chromosome representation. Genomics *12*, 526-533.

Lee, J. T., Strauss, W. M., Dausman, J. A., and Jaenisch, R. (1996). A 450 kb transgene displays properties of the mammalian X-inactivation center. Cell *86*, 83-94.

Lehrach, H., Drmanac, R., Höheisel, J., Larin, Z., Lennon, G., Monaco, A. P., Nizetic, D., Zehetner, G., and Poustka, A. (1990). Hybridization fingerprinting in genome mapping and sequencing. Genome Analysis *1*, 39-81.

Lichter, P., Ledbetter, S. A., Ledbetter, D. H., and Ward, D. C. (1990). Fluorescence insitu hybridization with Alu and L1 polymerase chain-reaction probes for rapid characterization of human-chromosomes in hybrid cell-lines. Proceedings Of The National Academy Of Sciences Of The United St *87*, 6634-6638.

Lichter, P., Tang, C., Call, K., Hermanson, G., Evans, G. A., Housman, D., and Ward, D. C. (1990). High-resolution mapping of human chromosome-11 by insitu hybridization with cosmid clones. Science *247*, 64-69.

Lyon, M. F. (1961). Gene action in the X-chromosome of the Mouse (*Mus musculus* L.) Nature *190*, 373-374.

Mandel, J. L., Monaco, A. P., Nelson, D. L., Schlessinger, D., and Willard, H. (1992). Genome analysis and the human X-chromosome. Science *258*, 103-109.

McCabe, E. R., Towbin, J. A., van, d. E. G., and Trask, B. J. (1992). Xp21 contiguous gene syndromes: deletion quantitation with bivariate flow karyotyping allows mapping of patient breakpoints. American Journal of Human Genetics *51*, 1277-85.

McCarthy, L., Hunter, K., Schalkwyk, L., Riba, L., Anson, S., Mott, R., Newell, W., Bruley, C., Bar, I., Ramu, E., Housman, D., Cox, R., and Lehrach, H. (1995). Efficient high-resolution genetic mapping of mouse interspersed repetitive sequence PCR products, towards integrated genetic and physical mapping of the mouse genome. Proc. Natl. Acad. Sci. USA *92*, 5302-5306.

McKusick, V. A. (1989). The Human Genome Organisation: History, Purpose and Membership. Genomics *5*, 385-387.

Meier-Ewert, S., Maier, E., Ahmadi, A., Curtis, J., and Lehrach, H. (1993). An automated approach to generating expressed sequence catalogs. Nature *361*, 375-376.

Meinkoth, J., and Wahl, G. (1984). Hybridisation of nucleic acids immobilized on solid supports. Analytical Biochemistry *138*, 267-284.

Monaco, A. P. (1985). Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment. Nature *316*, 842-845.

Monaco, A. P., Lam, V. M., Zehetner, G., Lennon, G. G., Douglas, C., Nizetic, D., Goodfellow, P. N., and Lehrach, H. (1991). Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products. Nucleic Acids Research *19*, 3315-8.

Monaco, A. P., Muller, U., Larin, Z., Meier-Ewert, S., and Lehrach, H. (1991). Isolation of the human sex determining region from a Y-enriched yeast artificial chromosome library. Genomics *11*, 1049-1053.

Monaco, A. P., Neve, R. L., Collettifeener, C., Bertelson, C. J., Kurnit, D. M., and Kunkel, L. M. (1986). Isolation of candidate cDNAs for portions of the duchenne muscular-dystrophy gene. Nature *323*, 646-650.

Morton, N. E. (1955). American Journal of Human Genetics *7*, 277.

Mott, R., Grigoriev, A., Maier, E., Hoeisel, J., and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schyzosaccharomyces pombe*. Nucleic Acid Research *21*, 1965-1974.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and et al. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. Science *235*, 1616-22.

Nelson, D. L., Ballabio, A., Cremers, F., Monaco, A. P., and Schlessinger, D. (1995). Report of the sixth international workshop on X chromosome mapping 1995. .

Nelson, D. L., Ballabio, A., Victoria, M. F., Pieretti, M., Bies, R. D., Gibbs, R. A., Maley, J. A., Chinault, A. C., Webster, T. D., and Caskey, C. T. (1991). Alu-primed polymerase chain-reaction for regional assignment of 110 yeast artificial chromosome clones from the human X-chromosome - identification of clones associated with a disease locus. Proceedings Of The National Academy Of Sciences Of The United States Of America *88*, 6157-6161.

Nelson, D. L., Ledbetter, S. A., Corbo, L., Victoria, M. F., Ramirezsolis, R., Webster, T. D., Ledbetter, D. H., and Caskey, C. T. (1989). Alu polymerase chain-reaction - a method for rapid isolation of human-specific sequences from complex DNA sources. Proceedings Of The National Academy Of Sciences Of The United States Of America *86*, 6686-6690.

Nizetic, D., Gellen, L., Hamvas, R. M. J., Mott, R., Grigoriev, A., Vatcheva, R., Zehetner, G., Yaspo, M. L., Dutriaux, A., Lopes, C., Delabar, J., Van Broeckhoven, C., Potier, M., and Lehrach, H. (1994). An integrated YAC-overlap and cosmid-pocket map of the human chromosome 21. Human Molecular Genetics *3*, 759-770.

Nizetic, D., Zehetner, G., Monaco, A. P., Gellen, L., Young, B. D., and Lehrach, H. (1991). Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries Proceedings Of The National Academy Of Sciences Of The United States Of America. *88*, 3233-7.

Ohno, S. (1967). Sex chromosomes and sex linked genes. In Momographs on Endocrinology, L. A. and e. al., eds. (New-York: Springer-Verlag).

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. Science *245*, 1434-1435.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. American Journal of Human Genetics *26*, 588-97.

Parrish, J. E., and Nelson, D. K. (1994). Practical aspects of fingerprinting human DNA using Alu polymerase chain reaction. Methods in molecular and cellular biology *5*, 71-77.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J., and Gray, J. (1988). Fluorescence *in situ* hybridisation with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. Proceedings Of The National Academy Of Sciences Of The United States Of America *85*, 9138-9142.

Rappold, G. A. (1993). The pseudoautosomal regions of the human sex-chromosomes. Human Genetics *92*, 315-324.

Rappold, G. A., Klink, A., Weiss, B., and Fischer, C. (1994). Double crossover in the human Xp/Yp pseudoautosomal region and its bearing on interference. Human Molecular Genetics *3*, 1337-1340.

Richards, F. M., and Knowles, J. R. (1968). Glutaraldehyde as a protein cross-linking reagent. Journal of Molecular Biology *37*, 231-233.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J. C., and Markham, A. F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acid Research *18*, 2887-2890.

Riordan, J. R., Rommens, J. M., Kerem, B. S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J. L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S., and Tsui, L. C. (1989). Identification of the cystic-fibrosis gene - cloning and characterization of complementary-DNA. Science *245*, 1066-1072.

Roest Crollius, H., Ross, M. T., Grigoriev, A., Knights, C. J., Holloway, E., Misfud, J., Li, K., Playford, M., Gregory, S. J., Humphray, S. J., Coffey, A., See, C. G., Marsh, S., Vatcheva, R., Kumlien, J., Labella, T., Lam, V., Rak, K. H., Todd, K. A., Mott, R., Graeser, D., Rappold, G., Zehetner, G., Poustka, A., Bentley, D. R., Monaco, A. P., and Lehrach, H. (1996). An integrated YAC map of the human X chromosome. Genome Research *6*, 943-955.

Roux, K. H., and Dhanaragan, P. (1990). A strategy for single site PCR amplification of dsDNA: Priming digested cloned of genomic DNA from an Anchor Modified Restriction site and a short internal Sequence. Biotechniques *8*, 48-57.

Royer-Pokora, B., Kunkel, L. M., Monaco, A. P., Goff, S. C., Newburger, P. E., Baehmer, R. L., Cole, F. S., Curnette, J. T., and Orkin, S. H. (1986). Cloning the gene for an inherited human disorder, chronic granlomatous disease on the basis of its chromosomal location. Nature *322*, 32-38.

Rugarli, E., Adler, D. A., Borsani, G., Tsuchiya, H., Franco, B., Hauge, X., Disteche, C., Chapman, V., and Ballabio, A. (1995). A different chromosomal localization of the Clcn4 Gene in *Mus spretus* and C57BL/6J Mice. Nature Genetics *10*, 466-471.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science *239*, 487-91.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, New York.: Cold Spring Harbor Laboratory Press).

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings Of The National Academy Of Sciences Of The United States Of America. *74*, 5463-5467.

Sargent, C. A., Briggs, H., Chalmers, I. J., Lambson, B., Walkers, E., and Affara, N. A. (1996). The sequence organisation of Yp/proximal Xq homologous regions of the human sex chromosomes is highly conserved. Genomics *32*, 200-209.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-70.

Schena, M., Shalon, D., Heller, R., Chai, A., P.O., B., and Davis, R. W. (1996). Parallel human genome analysis - microarray-based expression monitoring of 1000 genes. Proceedings of the National Academy of Sciences of the United States of America *93*, 10614- 10619.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguey-Tom, P., Aggarwal, A., Bajorek, E., Bentolila, S., et al. (1996). A gene map of the human genome. Science *274*, 540-546.

Schwartz, D., and Cantor, C. R. (1984). Separation of chromosome-sized DNAs by pulse field gel electrophoresis. Cell *37*, 67-75.

Sealey, P. G., Whittaker, P. A., and Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. Nucleic Acid Research *13*, 1905-1922.

Senger, G., Jones, T. A., Fidlerova, H., Sanseau, P., Trowsdale, J., Duff, M., and Sheer, D. (1994). Released chromatin - linearized DNA for high-resolution fluorescence *in-situ* hybridization. Human Molecular Genetics *3*, 1275-1280.

Shepherd, N. S., Pfrogner, B. D., Coulby, J. N., Ackerman, S. L., Vaidyanathan, G., Sauer, R. H., Balkenol, T. C., and Sternberg, N. (1994). Preparation and screening of an arrayed human genomic library generated with the P1 cloning system. Proceedings Of The National Academy Of Sciences Of The United States Of America *91*, 2629-2633.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector Proceedings Of The National Academy Of Sciences Of The United States Of America *89*, 8794-8797.

Silverman, G. A. (1996). End-rescue of YAC clone inserts by inverse PCR. Methods in Molecular Biology *54*, 145-55.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674-9.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology *98*, 503-17.

Sternberg, N. (1990). Bacteriophage-P1 cloning system for the isolation, amplification, and recovery of dna fragments as large as 100 kilobase pairs. Proceedings Of The National Academy Of Sciences Of The United States Of America *87*, 103-107.

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. (1988). Software for genome mapping by fingerprinting techniques. Computer and Applied Bioscience *4*, 125-132.

The HYP consortium (1995). A gene (PEX) with homologies to endopeptidases is mutated in patients with X-linked hypophosphatemic rickets. Nature Genetics *11*, 130-136.

Tocharoentanaphol, C., Cremer, M., Schrock, E., Blonden, L., Kilian, K., Cremer, T., and Ried, T. (1994). Multicolor fluorescence *in situ* hybridization on metaphase chromosomes and interphase Halo-preparations using cosmid and YAC clones for the simultaneous high resolution mapping of deletions in the dystrophin gene. Human Genetics *93*, 229-35.

Toder, R., Rappold, G. A., Scheibel, K., and Schempp, W. (1995). ANT3 and STS genes are autosomal in prosimians lemurs: implications for the evolution of the pseudoautosomal region. Human Genetics *95*, 22-8

Toye, A.A., Schalkwyk, L., Lehrach, H. Bumstead, N. A yeast artificial chromosome (YAC) library containing 10 haploid chicken genome equivalent. Mammalian Genome (in press).

Trask, B. J., Massa, H., Kenwrick, S., and Gitschier, J. (1991). Mapping of human-chromosome Xq28 by 2-color fluorescence *in situ* hybridization of DNA-sequences to interphase cell-nuclei. American Journal of Human Genetics *48*, 1-15.

Voyta, J. C., Edwards, B., and Bronstein, I. (1988). Clinical Chemistry *34*, 1157.

Walter, M. A., Spillett, D. J., Thomas, P., Weissenbach, J., and Goodfellow, P. N. (1994). A method for constructing radiation hybrid maps of whole genomes. Nature Genetics *7*, 22-28.

Watson, J. D., and Crick, F. H. C. (1953). Molecular structure of nucleic acids- a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Watson, J. M., Frost, C., Spencer, J. A., and Graves, J. (1993). Sequences homologous to the human-X-borne and human-Y-borne zinc finger protein genes (ZFX/Y) are autosomal in monotreme mammals. Genomics *15*, 317-322.

Weber, J. L., and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. American Journal of Human Genetics *44*, 388-96.

Weeks, D. E., and Lathrop, G. M. (1995). Polygenic disease: methods for mapping complex disease traits. Trends in Genetics *11*, 513-9.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A 2nd-generation linkage map of the human genome. Nature *359*, 794-801.

Wieacker, P., Davies, K. E., Cooke, H. J., Pearson, P. L., Bhattacharya, S. S., Zimmer, J., and Ropers, H. H. (1984). Towards a complete linkage map of the human X chromosome: regional assignements of 16 cloned single copy DNA sequences employing a panel of somatic cell hybrids. American Journal of Human Genetics *36*, 265-276.

Wilcox, A. S., Khan, A. S., Hopkins, J. A., and Sikela, J. M. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. Nucleic Acids Research *19*, 1837-43.

Wilcox, S. A., Watson, J. M., Spencer, J. A., and Graves, J. A. M. (1996). Comparative mapping identifies the fusion point of an ancient mammalian X-autosomal rearrangement. Genomics *35*, 66-70.

Willard, H. F. (1996). X chromosome Inactivation, XIST, and pursuit of the X inactivation center. Cell *86*, 5-7.

Willard, H. F., Cremers, F., Mandel, J. L., Monaco, A. P., Nelson, D. L., and Schlessinger, D. (1994). Report of the fifth international workshop on human X chromosome mapping 1994. Cytogenetics and Cell Genetics *67*, 295-358.

Yen, P. H., Marsh, B., Allen, E., Tsai, S. P., Ellison, J., Connolly, L., Neiswanger, K., and Shapiro, L. J. (1988). The human X-linked steroid sulphatase gene and Y-encoded pseudogene: evidence for an inversion of the Y chromosome during primate evolution. Cell *55*, 1123-1135.

Zehetner, G., and Lehrach, H. (1994). The reference library-system - sharing biological-material and experimental-data. Nature *367*, 489-491.

Zucchi, I., Mumm, S., Pilia, G., MacMillan, S., Reinbold, R., Susani, L., Weissenbach, J., and Schlessinger, D. (1996). YAC/STS Map across 12 Mb of Xq27 at 25 kb Resolution merging Xq26-qter. Genomics *34*, 42-54.

RESEARCH

An Integrated YAC Map of the Human X Chromosome

Hugues Roest Crollius,[1,2,8,9] Mark T. Ross,[2,3,8] Andrei Grigoriev,[1,2] Catherine J. Knights,[2] Ele Holloway,[2,3] Joseph Misfud,[2] Kim Li,[2] Martin Playford,[2] Simon G. Gregory,[3] Sean J. Humphray,[3] Alison J. Coffey,[3] Chee Gee See,[4] Sharon Marsh,[3] Radost Vatcheva,[2] Johan Kumlien,[2] Tullio Labella,[2] Veronica Lam,[2] Karl H. Rak,[1] Kieran Todd,[1,2] Richard Mott,[2,3] D'vorah Graeser,[2] Gudrun Rappold,[7] Gunther Zehetner,[1,2] Annemarie Poustka,[5] David R. Bentley,[3] Anthony P. Monaco,[6] and Hans Lehrach[1,2]

[1]Max-Planck-Institut für Molekulare Genetik, 14195 Berlin, Germany; [2]Imperial Cancer Research Fund, London WC2A 3PX, UK; [3]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; [4]University College London, London NW1 2HE, UK; [5]Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany; [6]Wellcome Trust Centre for Human Genetics, Oxford OX37BN, UK; [7]Institut für Humangenetik und Anthropologie, 69120 Heidelberg, Germany

The human X chromosome is associated with a large number of disease phenotypes, principally because of its unique mode of inheritance that tends to reveal all recessive disorders in males. With the longer term goal of identifying and characterizing most of these genes, we have adopted a chromosome-wide strategy to establish a YAC contig map. We have performed >3250 inter Alu-PCR product hybridizations to identify overlaps between YAC clones. Positional information associated with many of these YAC clones has been derived from our Reference Library Database and a variety of other public sources. We have constructed a YAC contig map of the X chromosome covering 125 Mb of DNA in 25 contigs and containing 906 YAC clones. These contigs have been verified extensively by FISH and by gel and hybridization fingerprinting techniques. This independently derived map exceeds the coverage of recently reported X chromosome maps built as part of whole-genome YAC maps.

The establishment of clone maps for each human chromosome is a prerequisite for transcript mapping and genomic sequencing. This goal became feasible following the development of the yeast artificial chromosome (YAC) cloning system (Burke et al. 1987). Several human YAC libraries have been made available (Anand et al. 1989; Albertsen et al. 1990; Larin et al. 1991; Chumakov et al. 1995), and YAC maps have now been reported covering most of chromosome Y (Foote et al. 1992), 21 (Chumakov et al. 1992), 22 (Collins et al. 1995), 3 (Gemmill et al. 1995), 12 (Krauter et al. 1995), and 16 (Doggett et al. 1995).

[8]These authors contributed equally to this work.
[9]Corresponding author.
E-MAIL roest@mpimg-berlin-dahlem.mpg.de; FAX +49 30 8413 1380.

The X chromosome is one of the most intensively studied of all human chromosomes. A reason for this interest is that males are hemizygous for X chromosome loci, and hence more disease phenotypes have been revealed on the X chromosome than on any autosome (McKusick 1994). The mapping of disease genes on the X chromosome is facilitated by their characteristic phenotypic pattern (female carriers, affected male offspring) and by the manifestation of maternal meiotic recombinations between X chromosomal loci in male offspring.

Three whole X chromosome YAC mapping studies are under way. The first two are part of whole-genome mapping studies at the Centre d'Etudes du Polymorphisme Humain (CEPH) (Chumakov et al. 1995) and Whitehead Institute/

6:943–955 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

Massachusetts Institute of Technology (MIT) (Hudson et al. 1995), where the X chromosome is poorly represented compared with autosomes. The third study combines contigs from many groups (R. Nagaraja, S. MacMillan, J. Miao, C. Jones, B. Cho, B. Eble, G. Halley, M. Masisi, J. Terrell, M. Trusgnich, et al., pers. comm.) and was reported at the sixth X chromosome workshop (D.L. Nelson, A. Ballabio, F. Cremers, A.P. Monaco, and D. Schlessinger, pers. comm.). At that time this map, based on sequence-tagged site (STS) content information, was estimated to cover 70% of the chromosome. In addition to these global efforts, numerous YAC contigs have been established in smaller regions defined by genetic mapping or by cytogenetic abnormalities. They have often been a template for the cloning of disease genes by positional cloning [recently HYP (The HYP consortium 1995), OA1 (Bassi et al. 1995)] and sometimes have evolved into physical and transcriptional maps of larger regions (Ferrero et al. 1995). These regional efforts have often used common sets of markers and library clones, and from these it has been possible to establish "consensus" YAC maps over still larger tracts of the chromosome (Nelson et al. 1995).

Our goal is to establish a physical map of the whole X chromosome, and here we report the results of our efforts to establish a map in contiguous YAC clones (Fig. 1). Our efforts have yielded currently 25 contigs covering an estimated 125 Mb (80%) of the X chromosome (Fig. 1). These contigs have been established primarily by direct hybridizations between YAC clones, and are supported by two fingerprinting methods, YAC end mapping and fluorescence in situ hybridization (FISH) localizations (see Fig. 2). The map contains 906 clones known to cover 655 genetic marker loci and a further 192 discrete marker loci (YAC end, cloned inter Alu-PCR product). This result comes at a time when comparisons between autosomes and the X chromosome indicate a shortage of polymorphic STS markers on the latter (Chumakov et al. 1995; Hudson et al. 1995). The present study, based on different techniques, avoids the pitfalls of STSs mapping and draws together global and regional expertise to integrate resources on the X chromosome.

RESULTS

Large-scale Generation of Overlap Information

Identification of X chromosomal YACs and primary YAC overlap information were derived by YAC to YAC hybridization experiments (Fig. 3). Multiple entry points were established along the chromosome by random probe selection from an X chromosome-specific YAC library (Lee et al. 1992) (HHMI hereafter). Probes were derived from individual YACs by inter Alu PCR (Nelson et al. 1989) using a combination of the primers ALE1 and ALE3 (Cole et al. 1991), which recognize the most conserved regions of the human Alu repeat and direct amplification outward from its left and right ends, respectively. Hybridization targets were also inter Alu-PCR products, derived both from the HHMI library and from the whole-

Figure 1 Integrated YAC map of the human X chromosome, slightly modified from the IXDB (acedb version 4.3) map view. The scale is based on a 160-Mb chromosome and each graduation represents 5 Mb. A chromosome ideogram is drawn on the left, and each colored box to the right represents a YAC clone (see color code). The yellow boxes to the right of the clones show the extent of the contigs. These cover ~125 Mb of the chromosome (80%) and include 906 YAC clones. A magnified view of 10 Mb in Xp22 is shown to the right. The color code is an indication of only one of the techniques that contributed to the positioning of a given clone on the map. All clones shown have been either hit or used as probe in a hybridization experiment. White clones have no other evidence for their position. Clear blue clones contain the markers (DXS, genes) indicated to the right of the scale. Pink clones have been mapped by FISH. When the FISH experiment indicates a chimeric clone, the latter is shown in green. An Alu gel fingerprint is stored in IXDB for the dark blue clones and available in a dedicated viewing tool for comparison between clones. However, when a clone has been analyzed by more than one method (e.g., by gel fingerprinting and mapped by FISH), only one technique is indicated by the color code. In IXDB, a single click of the mouse produces a window where the complete set of information attached to a clone is displayed. Small red boxes between the clones and the chromosome bands represent cloned Alu-PCR products identified in hybridization fingerprint experiments. More than one Alu clone in a single position indicates that the order between the clones could not be resolved. A maximum of three clones are shown, and the average number of clones per position is 15. Units are in kilobases, starting from 0 to 160,000 (pter to qter). The scale is only indicative and facilitates the comparison with the X community consensus map.

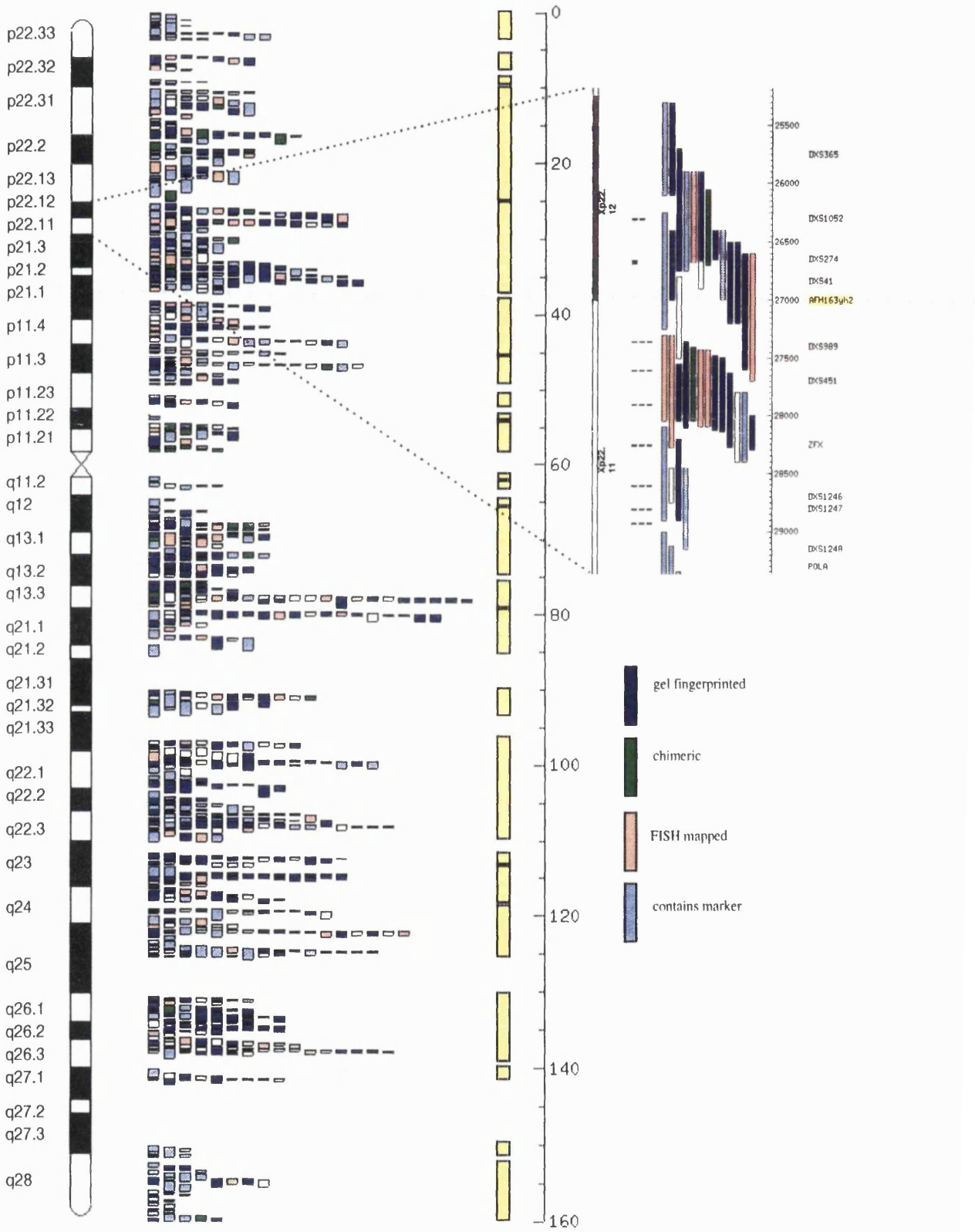


Figure 1 (*See facing page for legend.*)

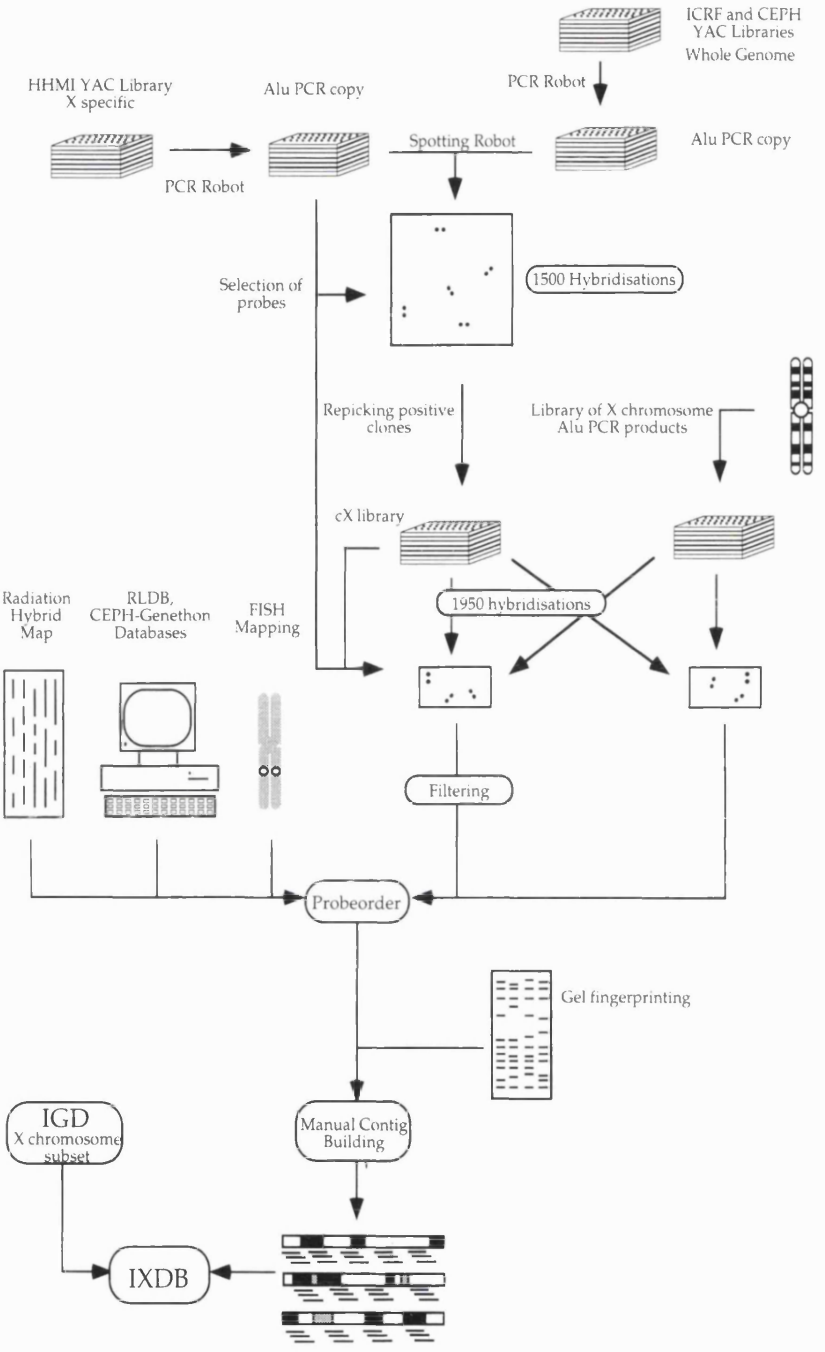


Figure 2 Schema of the strategy used to construct a YAC contig map of the X chromosome. Three YAC Libraries (HHMI, ICRF, and CEPH) were spotted as Alu-PCR products on nylon membranes, and a selection of probes from the X-specific library was used for hybridizations. The positive clones were repicked in a collection of X chromosome YAC clones (cX library), and more hybridizations were carried out with probes from the HHMI, cX, and cloned Alu-PCR product library. After scanning the data to remove cross-contamination and obvious false positives, the experimental data were combined with YAC mapping data from a separate radiation hybrid project, from the RLDB and CEPH-Genethon data bases, and from FISH mapping experiments. This was done using the program Probeorder, which was used to construct YAC clusters and to display all the information in one format. This information was combined with the gel and hybridization fingerprinting data and analyzed manually to build the final contigs. The resulting map and all the experimental data are combined with the X chromosome subset of IGD in IXDB.

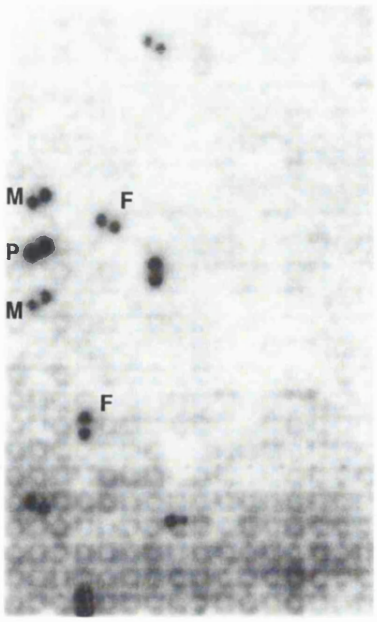


Figure 3 Example hybridization of a YAC Alu-PCR probe to a gridded array of YAC Alu-PCR products. The probe YAC (AA0801) was hybridized to an array of Alu-PCR products from 1536 YAC clones from the cX library (see text), each gridded in duplicate. The YAC hybridizes to itself (P) and to nine other clones. The YAC probe contains the marker loci DXS67 and DXS68 from Xp21.3. The two clones marked M also contain the DXS67 marker locus. The two clones marked F were localized by FISH to Xp22.1-21.3.

genomic Imperial Cancer Research Fund (ICRF) and CEPH "mega" YAC libraries. Together these libraries contain a theoretical 15-fold coverage of the X chromosome. Each of the 50,000 clones was amplified using a microtitre plate PCR robot (Meier-Ewert et al. 1993) and the products were gridded onto nylon membranes in high-density arrays. The inter Alu-PCR products from the YAC probes were radiolabeled and hybridized to the gridded arrays. In a total of 543 successful hybridizations of 764 performed with HHMI YAC probes to these arrays, 3978 different clones were identified (1727 ICRF, 643 CEPH, 1608 HHMI).

The 2370 positive clones from the ICRF and CEPH libraries were repicked into microtitre dishes to create a collection of X chromosomal YAC clones (the cX library). Hybridization filters were generated from the cX and HHMI libraries as described above. Inter Alu-PCR products from 316 clones in the cX library and a further 124 clones in the HHMI library were hybridized to these filter arrays, thus generating additional clone overlap information (Fig. 3).

Anchoring of YACs on the X Chromosome Map

In parallel to the large-scale experimental strategy described above, we collected genetic and physical mapping information on a large number of YAC clones used in this project. We have relied predominantly on preexisting information from two main sources: the Reference Library Database (RLDB; Zehetner and Lehrach 1994) and the CEPH/Genethon data base (Chumakov et al. 1995). The RLDB is a repository of mapping information obtained by the distribution of many types of reference libraries (YAC, cosmid, PAC, cDNA, etc.), including the ICRF and CEPH YAC libraries. We queried the RLDB for all human YAC clones previously mapped to the X chromosome, and retrieved 1723 records, of which 10% had also been confirmed by secondary screening. In addition, 28 RLDB participants provided 42 contigs in candidate regions for disease genes. Although there was some overlap between these data sets, information was collected on the marker content of 1181 clones. From the CEPH-Genethon data base, 711 YAC clones associated with an X chromosome marker and derived from the whole genome map were retrieved.

In total, these two sources provided marker information on 3074 YAC clones. Of these, 1150 were also identified in our hybridization and were used to annotate our contig map with marker information. However, it was not possible to treat all the outside data with the same level of confidence. We based our decisions on two conventions. First, we assumed that contigs provided by RLDB participants were completely correct with regard to the marker content of the YACs, unless a conflict between two or more groups existed in which case the situation that agreed with our data was assumed to be correct. The same applied to confirmed results from the RLDB. Second, nonconfirmed results from the RLDB and marker assignments derived from the CEPH data base were considered as only indicative and never used as sole evidence for the positioning of a clone on the map.

Localization of Unanchored Clones

FISH mapping experiments were performed with YAC clones belonging to unanchored contigs or with clones for which confirmation was needed before placing them on the map. Of 301 clones selected, 212 were assigned to the X chromosome, of which 48 were X-autosome chimerae.

The 89 non-X clones were mainly in singleton contigs or in very short contigs that could not be linked to other contigs. A radiation hybrid map of the X chromosome constructed in our laboratory was also used for localizing contigs lacking markers relative to each other (Kumlien et al. 1996). The map, comprising 72 hybrids, was constructed using 50 STSs spread evenly along the chromosome. Inter Alu-PCR products from the hybrids were hybridized to filters of the cX library, and, conversely, 450 YACs were hybridized to the hybrid panel. This allowed 971 YACs to be placed with confidence in ~3-Mb intervals (average distance between the STSs used).

Overlap Refinement and Confirmation

As data analysis proceeded (see below) we selected 1149 clones for Alu-PCR-based gel fingerprinting (Coffey et al. 1996). YAC clones were amplified individually by Alu-PCR, and the products separated on polyacrylamide gels. Fingerprints were analyzed automatically by the program contigC (derived from Contig9; Sulston et al. 1988), which yields a probability of overlap based on the number of bands shared between clones. Subsequently, detailed manual comparison of fingerprints was used to confirm potential overlaps and to provide a suggested order of clones based on subsets of shared and nonshared bands. In a second method, Alu-PCR products from 340 YACs were hybridized to a library of cloned Alu-PCR products from the X chromosome (Fig. 4). Shared hybridization patterns between YAC clones suggested YAC overlaps. Results were analyzed by Probeorder, a software successfully applied when constructing a YAC map of the *Schizosaccharomyces pombe* genome (Maier et al. 1992; Mott et al. 1993). When applied to raw hybridization results, Probeorder uses the simulated annealing algorithm to find the optimal order of probes based on their hybridization pattern. In addition, clones are ordered according to the order of the markers they contain. Approximately 2000 cloned inter Alu-PCR products were identified and added to the map. These clones constitute a pool of potential single copy probes, and 350 of them were hybridized to the cX library. An additional set of 100 single-copy probes were developed from the ends of YAC inserts by the vectorette PCR method (Riley et al. 1990), and these were hybridized back to the cX library.

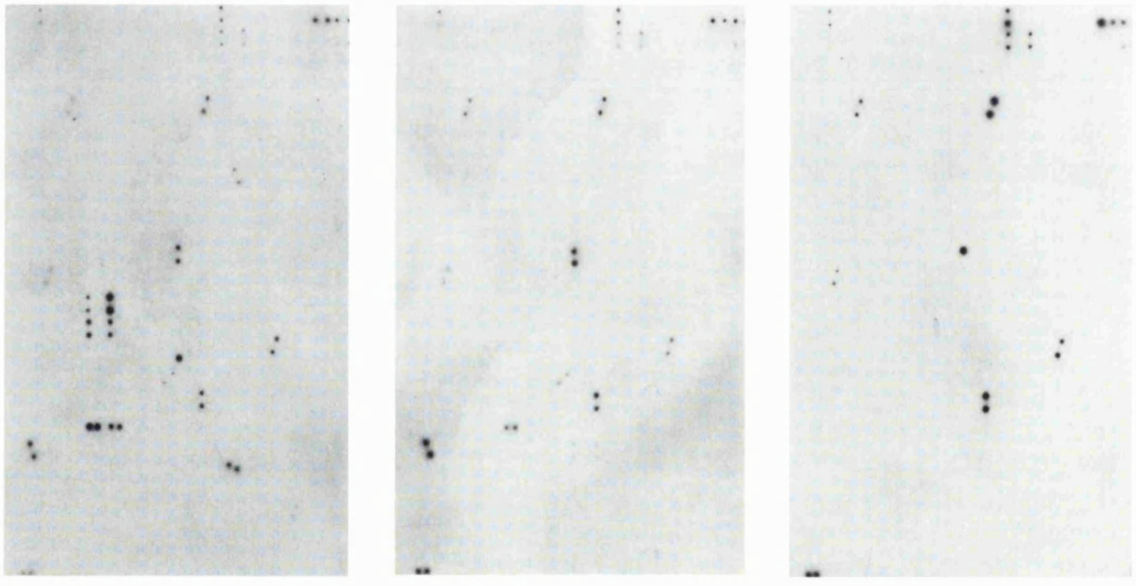


Figure 4 Hybridization results of three overlapping YAC clones on the library of cloned Alu-PRC products of the X chromosome. Each filter measures 7 × 11 cm and contains 9600 DNA spots, each in duplicate. Each clone has been amplified by Alu-PCR before spotting. Clones shared between the three YAC clones are clearly visible, and indicate that the YACs overlap.

Automatic Data Integration and Map Construction

The YAC to YAC hybridization data and the positional information were integrated using Probeorder. Positions derived from FISH experiments and from the radiation hybrid map were also taken into account at this stage. We used Probeorder to analyze the entire data set (2700 hybridization results) and constructed 113 clusters comprising 4087 clones. Clusters were ordered along the chromosome according to positional information. In 38 cases, clusters contained a single probe (singletons) and, therefore, were not useful for contig construction. The remaining 75 clusters containing 3973 clones were used for constructing contigs manually. This was done by considering the different types of mapping information in order of importance; FISH and marker content information were considered together first, and then fingerprints (gel- and hybridization-based) were compared to confirm the overlaps and determine a relative order of clones within a cluster. In this procedure, reliable contigs were progressively extracted from the Alu-PCR hybridization data set generated by Probeorder. Links between clusters were identified in the output of Probeorder by seeking YAC probes hybridizing consistently to clones present in two different clusters. Identification of such links reduced the total number of contigs from 75 to 25. When available, results generated by RLDB collaborators with the same clones were compared and used to orientate contigs and confirm orders and overlaps between clones. Nonetheless, each single overlap presented on the map can be deduced directly from the experimental data described above, with the exception of 19 clones in the two most telomeric bands (Xp22.33 and Xq28), which were contributed directly by outside groups (Ried et al. 1995 and Rogner et al. 1995, respectively).

Finally, marker/YAC association derived from RLDB supported by FISH results was used to place the contigs on the consensus map of markers, constructed jointly by the X chromosome community. Because all of the RLDB collaborators that indirectly contributed to our project also participated in the establishment of the consensus map, the integration of our contigs with the consensus map was greatly facilitated. The result of this strategy is a YAC contig map of the X chromosome integrating 655 genetic markers with 906 YAC clones, organized into 25 contigs (see Fig. 1). The total coverage is estimated to be 80% of the length of the chromosome, or 125 Mb of DNA. Based on the hybridization fingerprints, 79 intervals could be defined in the YAC contigs, in which 1420 cloned PCR products were placed.

Public Availability

At an early stage in the course of this study, we opted for ACEDB as a graphical data base system, first to store the collection of mapping information derived from their various sources and, subsequently, as a software tool to construct the map in its graphical representation. In the data base called IXDB (Integrated X DataBase), the map presented here is combined with information obtained from the Integrated Genome Database, which uses ACeDB to assemble data from the major genome-related data bases (GDB, OMIM, GenBank, RLDB, etc.). IXDB is available on the World Wide Web at http://www.mpimg-berlin-dahlem.mpg.de/~xteam. In this repositiory, all the experimental data supporting the map is presented in a user friendly environment. The YAC clusters constructed automatically by Probeorder are also available at the same address.

DISCUSSION

We describe a predominantly hybridization-based experimental approach that has been applied to establish YAC clone contigs covering ~80% (125 Mb) of the human X chromosome in 25 contigs. The map comprises some 750 discrete markers of all types (genetic, vectorette, inter Alu-PCR products). We generated a large experimental data set that was first processed with computer programs to lower its complexity. A stringent manual analysis was then performed on each YAC cluster, using all available information.

We observed that in the Alu-PCR hybridization data, 22% of the probes did not hybridize to themselves. The hybridization data generated by these probes was still considered in the analysis, for the following reasons. The most frequent source of false negative can be ascribed to the absence of the probe DNA on the filter, either because the robot pins did not transfer liquid on this particular spot, or because the YAC did not amplify properly in the waterbath PCR robot. In these cases connections between the probe and the clones it identifies are still correct. Alternatively, it is possible that probes were accidentally

mixed up, leading to apparent false negatives. These results could be detected easily in the redundant data set, as they present a clear aberrant hybridization pattern in the later analysis. We removed 51 probes from the data set based on this observation (7% of all probes). False-positive hybridization results can introduce false connections between clones. It is impossible to measure accurately the rate of false positives, but based on the number of links between probes that had to be ignored in the manual analysis, we estimate it at 10%–15%. The reason for their presence can be attributable either to human error (typing, scoring, sample handling), which is particularly acute in a large-scale project, or to nonspecific sequence similarities between clones. Mis-scoring was limited by the fact that each X-ray film was scored by two different persons independently. Human error was corrected further by scanning the data set with programs to detect specific patterns, for example, cross-contamination (probes in adjacent wells with identical hybridization patterns) or nonremoval of a probe from a filter (same clones positive in two successive hybridizations). Nonspecific sequence similarities are a well-known problem in mapping large regions of the human genome, and in addition come to the high level of chimerism observed in YAC libraries (30%). Whether attributable to repeat sequences or gene families, this problem can be avoided only by using complementary techniques to help make decisions. We have addressed this problem by complementing the Alu-PCR data with a battery of different types of data (fingerprints, marker content, FISH, end mapping, radiation hybrids) in a stringent manual analysis.

The map covers 80% of the chromosome, with 25 gaps. The depth of coverage is uneven (Fig. 1) with up to a 20-fold difference within 2Mb around the Menkes syndrome locus, for example. Long-range coverage is balanced, however, with 38% of the YACs localized on the short arm (36% of chromosomal length), and the remaining 62% on the long arm. The largest gap in the YAC contig map is in Xq27, where almost the complete band, which measures ~11 megabases, is not represented. Because a complete lack of Alu sequences over such a large region can be excluded, the reason for this underrepresentation must be ascribed to an unfortunate absence of probes mapping to this region in our random selection of HHMI clones. The region is represented at least partially in the target libraries as YAC contig construction has been reported in this cytogenetic band (Zucchi et al. 1996). Nevertheless in some cases we do observe a correlation between large gaps and the presence of a G-dark band. This is consistent with studies showing that these regions are relatively poor in Alu sequences (Korenberg and Rykowski 1988). However, not all G-dark bands are poorly represented (for example, Xp22.2, Xq13.1, and Xq23).

We have compared our map with X chromosome YAC contig maps built as part of whole genome efforts by CEPH (Chumakov et al. 1995) and by the Whitehead/MIT (Hudson et al. 1995) groups. In both cases, the X chromosome stands out because of its poor coverage compared with the average of the autosomes. This is principally attributable to the low representation of the X chromosome in the CEPH library made from a male cell line, which was the only substrate for the construction of the physical maps. Also the human and mouse X chromosomes contain either fewer CA repeats or fewer polymorphic CA repeats (Dietrich et al. 1996). This leads to a lower density of genetic markers available for whole-genome physical maps based on this type of STS. Therefore, an independently derived map of the X chromosome in YAC clones using libraries enriched for X chromosome DNA and independent from CA repeat content is particularly complementary. The approximate coverage of the X chromosome in the CEPH map, calculated with markers in common with the workshop consensus map, is 52 Mb (32%). The marker order between our map and the CEPH map agrees well except in one instance, where a group of markers is clearly misplaced in the CEPH map (XIST is placed in Xp11). Comparison with the Whitehead/MIT map of the X chromosome is more difficult, as the majority of markers have been developed very recently by this group and, therefore, are not placed on our map. It was possible to find 35 DXS markers common to both maps, for which the order broadly agrees, except for the first half of the short arm. In that region, the order of the nine common DXS markers strongly disagrees with the X chromosome community consensus map and with our map, over a 30-Mb region. The Whitehead/MIT order used for comparison is extracted from the radiation hybrid/STS content map. Again, on the basis of the physical distances between common markers in our map and the Whitehead/MIT map, we estimate the coverage of this map at ~50 Mb. This is also sustained by the maximal length of the contigs presented, based on the average length of a

YAC clone. The consensus map established at the sixth X chromosome workshop (Nelson et al. 1995) reported an 80% coverage of the chromosome in YAC contigs and the presence of 24 gaps in the map. This consensus was derived by collating the contigs from >50 different groups, and concentrates on marker order rather than attempting to present YAC clone organization. Therefore, the estimation of the size and number of gaps and the YAC coverage has to be taken with caution.

We are working on gap closure using two strategies, bypassing the use of inter Alu-PCR and establishing useful landmarks for a cosmid/P1/P1 artificial chromosome (PAC) map of the chromosome. First, we are identifying these *Escherichia coli*-based clones using genetic and physical STS markers developed in the CEPH and Whitehead/MIT mapping efforts, which are likely to be positioned in our gaps. The STSs are amplified from total human DNA and used as hybridization probes. The positive clones in turn are used to screen the genomic YAC libraries to identify clones missed by Alu-PCR YAC probes. Second, we are using a combination of L1 (Line repeat) and Alu primers to amplify YAC clones from the ends of our contigs. These are used to screen the same cosmid and PAC libraries, therefore identifying cosmids and PAC clones at the edges of the gaps. When used to screen against the YAC genomic libraries, these probes can identify new clones extending from the original contig.

Arising out of the X chromosome workshop was a common accord that a repository of all YAC clones known or supposed to map to the X chromosome must be established, in combination with a dedicated data base that would make available all the published mapping information. We have taken on this project and have distributed 15 copies of a 9000-clone collection to genome centers worldwide. Also, we make available high-density gridded YAC colony filters of the collection, and DNA pools for PCR screening. This will increase the value of the X chromosome YAC resources available worldwide, and will allow verification and completion of the existing consensus YAC map. The clone collection includes the cX library reported here and clone sets from groups based at the Sanger Centre, the Baylor College of Medicine, the Washington University School of Medicine, and many others.

Clearly, the mapping of the X chromosome is reaching a stage where increasing efforts will be put into the construction of higher resolution maps in bacterial cloning systems [cosmids, P1, PACs, bacterial artificial chromosomes (BACs)], in which the YAC clone resources and maps will play a central role. These bacterial clones will be essential for the large-scale genomic sequencing and transcriptional mapping of the chromosome. Using the YAC contig map presented here, we have started a systematic identification of PAC, BAC, and X chromosome-specific cosmid clones. This is the next logical step toward a high resolution ("sequence ready") clone and transcript map of the chromosome, itself the consummate template for large-scale sequencing projects.

METHODS

YAC Libraries

The human YAC libraries used were the ICRF whole genomic library (Larin et al. 1991; M.T. Ross, S. Meier-Ewert, and A. Monaco, unpubl.), the CEPH "mega" YAC library (Chumakov et al. 1995) (plates 737–984), and the University of Pennsylvania X chromosome-specific library (Lee et al. 1992; HHMI thereafter). The HHMI library was made from a hybrid cell line (Micro-21D) carrying Xpter-Xq27.3 as its human component. The ICRF library comprises clones derived from the DNA of the cell lines GN1416B (48, XXXX National Institute of General Medical Sciences, Human Genetic Cell Repository), OXEN [49, XYYYY (Bishop et al. 1983) and HD1 (46, XX, homozygous for Huntington disease; Wexler et al. 1987)]. The CEPH library was made from a male lymphoblastoid cell line DNA source. The total coverage of the three libraries combined is estimated to be 14.5 X chromosome equivalents.

Large-scale Inter Alu-PCR of YAC Clones

Whole yeast DNA was extracted by a modification of the procedure of Chumakov et al. (1992) and used as template for large-scale inter Alu-PCR of YAC clones. The three libraries were replicated into 96-well microtitre plates containing 100μl selective medium (Anand et al. 1989) (SD ura, -trp), and cultures were grown for 3 days at 30°C. Cells were pelleted for 10 min at 2000 rpm (Beckman J6-MC), and supernatants were removed by inversion. Cell pellets were washed in 50 μl SCE buffer [1M sorbitol, 0.1M sodium citrate (pH 5.8), 10 mM EDTA], then harvested as before. Yeast cells were converted to spheroplasts by incubation for 1 hr at 37°C in 25 μl SCE containing 4 mg/ml novozym (NovoBiolabs) and 10 mM dithiothreitol. Then, 60 μl 0.14 N NaOH were added to each well and plates were incubated for 7 min at room temperature. DNA extracts were neutralized with 60 μl of 1M Tris-HCl (pH 8.0) and stored at −20°C.

Inter Alu-PCR was carried out in 67 mM Tris-HCl (pH 8.8), 16.7 mM $(NH_4)_2SO_4$, 6.7 mM $MgCl_2$, 0.5 mM each dNTP, 170 μg/ml BSA, 10 mM 2-mercaptoethanol, 1.3 μM primers ALE1 and ALE3 (Cole et al. 1991), and 0.6 units of *Taq* polymerase. A mixture sufficient for ≤10,000 reac-

tions was dispensed in 50-μl aliquots into the wells of 384-well polypropylene microtitre dishes (Genetix). DNA from individual clones was transferred to the reaction plate with a 96-pin plastic device (Genetix), and plates were heat-sealed with a plastic film. PCR was carried out by use of a large capacity waterbath robot (Meier-Ewert et al. 1993) for 30 cycles of 3 min at 94°C followed by 6 min at 65°C, with an initial denaturation of 10 min and final extension of 10 min.

High-density Gridded Filter Arrays of YAC Clones

A custom-built robotic device (Lehrach et al. 1990) was used to grid either YAC inter Alu-PCR products or live YAC cultures onto nylon membranes for hybridization screening.

For PCR products, a 384-pin plastic gridding tool (Genetix) was used to transfer a small amount (<0.5 μl) of liquid from each well of the reaction plate onto a 22 × 22 cm nylon membrane (Hybond N+, Amersham). By interleaving the gridding patterns, PCR prodcuts of 9000–18,000 clones were gridded in duplicate on a single filter. Therefore, the three libraries were accommodated on three to five filters.

Gridded filters were transferred onto Whatman 3MM paper soaked with 0.5 N NaOH, 1.5 M NaCl for 2 min, then neutralized in 1M Tris-HCl (pH 7.2), 1.5 M NaCl. Filters were air-dried before use in hybridization experiments.

For screening whole YAC clone DNA, high-density colony grids were produced. The primary library plates in 96-well dishes were condensed into 384-well plates containing SD medium. After 2 days at 30°C, these cultures were used to grid onto nylon membranes, which were processed as described previously (Ross et al. 1992)

Hybridization of Inter Alu-PCR Products of Individual YAC Clones to High-density Filter Arrays of Inter Alu-PCR Products

Inter Alu PCR of individual clones to be used as probes was carried out using the reaction mixture described above. Reactions were carried out in 0.5-ml tubes in a MJ-PTC100 PCR machine using the following cycling conditions: 94°C for 5 min, then 30 cycles of 93°C for 1 min, 65°C for 1 min, 72°C for 4 min, then a final extension of 72°C for 5 min. Reaction products were precipitated by the addition of ammonium acetate to 2.5 M and two volumes of absolute ethanol. For 22 × 22 cm filters, 10–20 ng of DNA were labeled by random priming (Feinberg and Vogelstein 1983) in a 40-μl reaction using 5 μCi alpha [^{32}P]dATP. For 7.5 × 11 cm filters, only 1 μCi dATP was used. Probes were pre-reassociated for 1 hr at 65°C in 125 mM sodium phosphate buffer (pH 7.2) containing 0.75 mg/ml human placental DNA. Hybridization occurred overnight at 65°C in Church hybridization buffer [0.5 M sodium phosphate (pH 7.2), 7% SDS, 1% BSA (fraction V), 1 mM EDTA]. Filters were rinsed in 40 mM sodium phosphate buffer (pH 7.2), 0.1% SDS twice at room temperature, then washed twice in the same buffer at 65°C. Autoradiography was carried out using blue-sensitive film (Genetic Research Instrumentation) at −70°C with a single intensifying screen.

Hybridization Fingerprinting of YAC Clones

A library of cloned inter Alu-PCR products of the X chromosome was constructed in the plasmid vector pAMP1 (GIBCO-BRL). Approximately 100 ng of DNA from the hybrid cell line 578 (Wieacker et al. 1984), which contains a single human X chromosome on a hamster background, was used as a template in a PCR reaction using the same conditions as for the YAC amplification above, with 1.5 mM primer ALE3CA (CAUCAUCAUCAUCCACTGCACTCCAGCCTGGG) and 1.5 mM primer ALE3CU (CUACUACUACUACCACTGCACTCCAGCCTGGG). Cycling was as follows: 94°C for 4 min, 30 cycles for 94°C for 30 sec, 68°C for 2 min, and a final extension at 72°C for 4 min. One to 2 μl of the PCR was used directly for the UDG annealing reaction according to the manufacturer's instructions. Electrocompetent DH5a were electroporated with 1–2 μl of the annealing reaction, and the resulting clones were picked, using a robotic device developed by us (Meier-Ewert et al. 1993), into 384-well microtitre plates.

Amplification of plamid inserts was carried out in 75 mM Tris-HCl (pH 9), 20 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.1% (wt/vol) Tween, 0.2 mM each dNTP, 1.5 μM primer ALE3, and 0.5 units of *Taq* polymerase. Reactions were set up as for YAC inter Alu-PCR. Template DNA was added directly from thawed glycerol stocks of the plasmid library in 384-well plates, with a 384-pin device. PCR was carried out as for large-scale YAC amplification. The PCR products were gridded robotically by use of the system described for YAC inter Alu-PCR products. A higher gridding density allowed the complete library of 4600 clones to be spotted in duplicated on a 7 × 11 cm filter. A regular array of India ink dots was also spotted to facilitate positive identification when using fluorescent detection.

A hybridization fingerprint was generated by hybridizing the inter Alu-PCR products of a YAC to the gridded filters of the cloned Alu-PCR library. Generation of the probes was as described for the YAC to YAC hybridization except that only the primer ALE3 was used, and the precipitation of the final products was omitted. Approximately 100 ng of probe were competed and hybridized as for the YAC to YAC hybridization. Washing and autoradiography of radioactive filters were also identical to the method above. X-ray films were scored by use of semi-automated methods and purpose-built software.

Hybridization of Cloned Inter Alu-PCR Products to High-density Filter Arrays of YAC Inter Alu-PCR Products

In all cases hybridizations were performed by use of digoxigenin labeling of the probes. DIG-11-dUTP was incorporated during PCR amplification of selected cloned inserts by use of primer ALE3 and 19:1 ratio of dTTP:dUTP. The PCR conditions were as described above for the amplification of the whole library before filter gridding, except that reactions were carried out in polycarbonate 96-well plates in an MJ-PTC100 thermocycler. Approximately 100 ng of amplified insert was subjected to competition, as were the YAC inter Alu-PCR probes, and hybridized to high-density arrays of YAC inter Alu-PCR products of the cX library. Washes and detection were carried out as recommeded by the manufacturer, using the Attophos substrate (JBL Scientific).

Gel Fingerprinting of YAC Clones

The method for gel fingerprinting of YAC clones by comparison of inter Alu-PCR products has been described elsewhere (Coffey et al 1996). Briefly, a single YAC colony was resuspended in 100 μl 10 mM Tris-HCl (pH 8.0), 0.1 mM EDTA. Five microliters of the suspension were used as template for the primers ALE1 and ALE3 in a 25-μl primary PCR. Reaction composition and cycling conditions were as described above for inter Alu-PCR of individual YAC probes. An aliquot of the primary PCR was used as template for a secondary PCR containing radiolabeled ALE1 and ALE3 primers. The products of the secondary PCR were electrophoresed on a 4% polyacrylamide, 7M urea gel containing Sau3AI digested and ^{35}S-labeled λ DNA markers. Dried gels were autoradiographed at room temperature for ~65 hr. Autoradiographs were then scanned (Amersham) and the image edited before analysis.

Generation of Vectorette End Probes from YACs

Protocols for total yeast DNA preparation was according to Riley et al. (1990), with essentially the following modifications. Novozym (8 mg/ml) was used instead of Lyticase, DTT (10 mM) was used instead of β-mercaptoethanol, and agarose was 2% before cell resuspension. The isolation of vectorette ends from YAC clones was performed according to Coffey et al. (1992). Prerassociation and hybridization of the probe and washing and autoradiography of the filters were performed as described above.

FISH Mapping of YAC Clones

Whole yeast DNA was used for FISH. Single YAC colonies were grown to saturation in 40 ml SD broth at 30°C. Cells were harvested at 1000*g* for 5 min, then resuspended in 3 ml of 0.9 M sorbitol, 0.1 M EDTA (pH 7.5) containing 50 μl zymolase 20T. Spheroplasting was carried out for 60 min at 37°C, then spheroplasts were pelleted at 200*g* for 5 min and resuspended in 5 ml of 50 mM Tris-HCl (pH 7.4), 20 mM EDTA. Five hundred microliters of 10% SDS were added, and samples were incubated at 65°C for 30 min. Potassium acetate was added to a concentration of 1 M and samples were placed on ice for 60 min. Debris was pelleted at 15,000*g* for 10 min, then DNA was precipitated with two volumes absolute ethanol. DNA was redissolved in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (TE) containing RNase and incubated for 30 min at 37°C. DNA was reprecipitated using 0.1 volumes of 3 M sodium acetate (pH 5.2) and two volumes absolute ethanol, washed in 70% ethanol and redissolved in 200 μl TE. The total yeast DNA was labeled by nick translation with biotin-16-dUTP and hybridized overnight onto metaphase spreads essentially as described (Lichter et al. 1988). Hybridization was visualized using a standard two-layer avidin-FITC protocol (Pinkel et al. 1988).

ACKNOWLEDGMENTS

We thank all members of the Reference Library System who provided information on their X chromosome YAC contigs; D. Cohen and D. Le Paslier for providing the CEPH mega YAC library; R. Nussbaum for the HHMI library; N. Carter and S. Povey for organizing FISH mapping of YAC clones; D. Nelson and D. Schlessinger for mapped YAC clones used as controls in our experiments; A. Ballabio and G. Ferrero for YAC maps across parts of Xp22 used as elements for comparison; F. Francis and M-L. Yaspo for critical reading of the manuscript. This work was supported by grants G9226552 from the Medical Research Council; CT910020 and CT930088 from the European Community. H.R.C. was supported by a bursary from the European Community (GT920336).

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

Albertsen, H.M., H. Abderrahim, H.M. Cann, J. Dausset, D. Lepaslier, and D. Cohen. 1990. Construction and characterization of a yeast artificial chromosome library containing 7 haploid human genome equivalents. *Proc. Natl. Acad. Sci.* **87:** 4256–4260.

Anand, R., A. Villasante, and C. Tylersmith. 1989. Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed-field gel-electrophoresis. *Nucleic Acids Res.* **17:** 3425–3433.

Bassi, M.T., M.V. Schiaffino, A. Renieri, F. De Nigris, L Galli, M. Bruttini, M. Gebbia, A.A. Bergen, R.A. Lewis, and A. Ballabio. 1995. Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome. *Nature Genet.* **10:** 13–19.

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.

Burke, D.T., G.F. Carle, and M.V. Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806–812.

Chumakov, I., P. Rigault, S. Guillou, P. Ougen, A. Billaut, G. Guasconi, P. Gervy, I. Legall, P. Soularue, L. Grinas, et al. 1992. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380–387.

Chumakov, I.M., P. Rigault, I. Le Gall, C. Bellané-Chantelot, A. Billault, S. Guillou, P. Soularue, G. Guasconi, E. Poullier, I. Gros, et al. 1995. A YAC contig map of the human genome. *Nature* (*Suppl.*) **377:** 175–297.

Coffey, A.J., R.G. Roberts, E.D. Green, C.G. Cole, R. Butler, R. Anand, F. Giannelli, and D.R. Bentley. 1992. Construction of a 2.6-mb contig in yeast artificial chromosomes spanning the human dystrophin gene using an sts-based approach. *Genomics* **12:** 474–484.

Coffey, A., S. Gregory, and C.G. Cole. 1996. Alu-PCR

fingerprinting of YACs. In *Methods in molecular biology.* (ed. D. Markie), pp. 97–114. Human Press Inc., Totowa, NJ.

Cole, C.G., P.N. Goodfellow, M. Bobrow, and D.R. Bentley. 1991. Generation of novel sequence tagged sites (STSs) from discrete chromosomal regions using Alu-PCR. *Genomics* **10:** 816–826.

Collins, J., C. Cole, L.J. Smink, C.L. Garett, M.A. Leversha, C.A. Soderlund, G.L. Maslen, L.A. Everett, K.M. Rice, A.J. Coffey, S.G. Gregory, and R. Gwilliam. 1995. A high density YAC contig map of human chromosome 22. *Nature (Suppl.)* **377:** 367–379.

Dietrich, W.F., J. Miller, R. Steen, M.A. Merchant, D. Damron-Boles, Z. Husain, R. Dredge, M.J. Daly, K.A. Ingalls, T.J. O'Connor, C.A. Evans, M.M. DeAngelis, D.M. Levinson, L. Kruglyak, N. Goodman, N.G. Copeland, N.A. Jenkins, T.L. Hawkins, L. Stein, D.C. Page, and E. Lander. 1996. A comprehensive genetic map of the mouse genome. *Nature* **380:** 149–152.

Doggett, N.A., L.A. Goodwin, J.G. Tesmer, L.J. Meincke, D.C. Bruce, L.M. Clark, M.R. Altherr, A.A. Ford, H. Chi, B.L. Marrone, et al. 1995. An integrated physical map of human chromosome 16. *Nature (Suppl.)* **377:** 335–366.

Feinberg, A.P. and B. Vogelstein. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132:** 6–13.

Ferrero, G.B., B. Franco, E.J. Roth, B.A. Firulli, G. Borsani, J. Delmas-Mata, J. Weissenbach, G. Halley, D. Schlessinger, A.C. Chinault, H.Y. Zoghbi, D.L. Nelson, and A. Ballabio. 1995. An integrated physical and genetic map of a 35 Mb region on chromosome Xp22.3-Xp21.3. *Hum. Mol. Genet.* **4:** 1821–1827.

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

Gemmill, R.M., I. Chumakov, P. Scott, B. Waggoner, P. Rigault, J. Cypser, Q. Chen, J. Weissenbach, K. Gardiner, H. Wang, Y. Pekarsky, I. Le Gall, D. Le Paslier, S. Guillou, E. Li, L. Robinson, L. Hahner, S. Todd, D. Cohen, and H.A. Drabkin. 1995. A second-generation YAC contig map of human chromosome 3. *Nature (Suppl.)* **377:** 299–320.

Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S. Xu, X. Hu, A.M.E. Colbert, C. Rosenberg, et al. 1995. An STS-based map of the human genome. *Science* **270:** 1945–1954.

The HYP consortium. 1995. A gene (PEX) with homologies to endopeptidases is mutated in patients with X-linked hypophosphatemic rickets. *Nature Genet.* **11:** 130–136.

Korenberg, J.R. and M.C. Rykowski. 1988. Human genome organization—Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53:** 391–400.

Krauter, K. K. Montgomery, S. Yoon, J. LeBlanc-Straceski, B. Renault, I. Marondel, V. Herdman, L. Cupelli, A. Banks, J. Lieman, J. Menninger, P. Bray-Ward, P. Nadkarni, J. Weissenbach, D. Le Paslier, P. Rigault, I. Chumakov, D. Cohen, P. Miller, D. Ward, and R. Kucherlapati. 1995. A second-generation YAC contig map of human chromosome 12. *Nature (Suppl.)* **377:** 321–334.

Kumlien, J., A. Grigoriev, H. Roest Crollius, M.T. Ross, P.N. Goodfellow, and H. Lehrach. 1996. A radiation hybrid map spanning the entire human X chromosome integrating YACs, genes and STS markers. *Mamm. Genome* (in press).

Larin, Z., A.P. Monaco, and H. Lehrach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci.* **88:** 4123–4127.

Lee, J.T., A. Murgia, D.M. Sosnoski, I.M. Olivos, and R.L. Nussbaum. 1992. Construction and characterization of a yeast artificial chromosome library for Xpter-Xq27.3—A systematic determination of cocloning rate and X chromosome representation. *Genomics* **12:** 526–533.

Lehrach, H., R. Drmanac, J. Hoheisel, Z. Larin, G. Lennon, A.P. Monaco, D. Nizetic, G. Zehetner, and A. Poustka. 1990. Hybridization fingerprinting in genome mapping and sequencing. In *Genome analysis* (ed. K.E. Davies and S.M. Tilghman), Vol. 1, pp. 39–81. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Lichter, P., T. Cremer, J. Borden, L. Manuelidis, and D.C. Ward. 1988. Delineation of individual human chromosomes in metaphase and interphase cells by *in situ* suppression hybridization using recombinant DNA libraries. *Hum. Genet.* **80:** 224–234.

McKusick, V.A. 1994. *Mendelian inheritance in man. Catalogs of human genes and genetic disorders,* 11th ed. John Hopkins University Press, Baltimore, MD.

Maier, E., J. Hoheisel, L. McCarthy, R. Mott, A Grigoriev, A.P. Monaco, Z Larin, and H. Lehrach. 1992. Complete coverage of the Schizosaccharomyces pombe genome in yeast artificial chromosomes. *Nature Genet.* **1:** 273–277.

Meier-Ewert, S., E. Maier, A. Ahmadi, J. Curtis, and H. Lehrach. 1993. An automated approach to generating expressed sequence catalogs. *Nature* **361:** 375–376.

Mott, R., A. Grigoriev, E. Maier, J. Hoheisel, and H. Lehrach. 1993. Algorithms and software tools for ordering clone libraries: Application to the mapping of the genome of *Schizosaccharomyces pombe. Nucleic Acids Res.* **21:** 1965–1974.

Nelson, D.L., S.A. Ledbetter, L. Corbo, M.F. Victoria, R. Ramirezsolis, T.D. Webster, D.H. Ledbetter, and C.T. Caskey. 1989. Alu polymerase chain-reaction—A method

for rapid isolation of human-specific sequences from complex DNA sources. *Proc. Natl. Acad. Sci.* **86:** 6686–6690.

Pinkel, D., J. Landegent, C. Collins, J. Fuscoe, R. Segraves, J. Lucas, and J. Gray. 1988. Fluorescence *in situ* hybridization with human chromosome-specific libraries—Detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci.* **85:** 9138–9142.

Riley, J., R. Butler, D. Ogilvie, R. Finniear, D. Jenner, S. Powell, R. Anand, J.C. Smith, and A.F. Markham. 1990. A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res.* **18:** 2887–2890.

Ried, K., A. Mertz, R. Nagaraja, M. Trusgnich, J.H. Riley, R. Anand, H. Lehrach, D. Page, J.W. Ellison, and G. Rappold. 1995. Characterisation of a YAC contig spanning the pseudoautosomal region. *Genomics* **29:** 787–792.

Rogner, U.C., P. Kioschis, K. Wilke, W. Gong, E. Pick, A. Dietrich, U. Zechner, H. Hameister, A. Pragliola, G.E. Herman, et al. 1994. A YAC clone map spanning 7.5 megabases of human chromosome based Xq28. *Hum. Mol. Genet.* **3:** 2137–2146.

Ross, M., J.D. Hoeisel, A.P. Monaco, Z. Larin, G. Zehetner, and H. Lehrach. 1992. High-density gridded YAC filters: Their potential as genome mapping tools. In *Techniques for the analysis of complex genomes.* (ed. R. Anand), pp. 137–154. Academic Press, London, UK.

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

Wexler, N.S., A.B. Young, R.E. Tanzi, H. Travers, S. Starosta-Rubenstein, J.B. Penney, S.R. Snodgrass, I. Shoulson, F. Gomez, M.A. Ramos-Arroyo, G. Penchaszadeh, R. Moreno, K. Gibbons, A. Farinyarz, W. Hobbs, M.A. Anderson, E. Bonilla, P.M. Conneally, and J.F. Gusella. 1987. Homozygotes for Huntington's disease. *Nature* **326:** 194–197.

Wieacker, P., K.E. Davies, H.J. Cooke, P.L. Pearson, S.S. Bhattacharya, J. Zimmer, and H.H. Ropers. 1984. Towards a complete linkage map of the human X chromosome: Regional assignments of 16 cloned single copy DNA sequences employing a panel of somatic cell hybrids. *Am. J. Hum. Genet.* **36:** 265–276.

Zehetner, G. and H. Lehrach. 1994. The Reference Library System—Sharing biological material and experimental data. *Nature* **367:** 489–491.

Zucchi, I., S. Mumm, G. Pilia, S. MacMillan, R. Reinbold, L. Susani, J. Weissenbach, and D. Schlessinger. 1996. YAC/STS Map across 12 Mb of Xq27 and 25 kb resolution, merging Xq26-qter. *Genomics* **34:** 42–54.

Received March 29, 1996; accepted in revised form July 11, 1996.