Investigation of the GATA Repetitive DNA Sequence of the Domestic Horse (*Equus caballus*)

A thesis presented in partial fulfilment
of the requirements for the degree of
Master of Science in Genetics at
Massey University

Andrea J. Ede

June 1990

ABSTRACT

The variation in copy number and organisation of the simple quadruplet repeat $(GATA)_n$ in the genome of most animals has made it a potential tool for DNA fingerprinting. This study was undertaken to explore this application and to investigate its abundance and organisation in the horse genome.

Using the synthetic oligomer $(GATA)_5$ end-labelled with ^{32}P as a probe, the copy number of $(GATA)_n$ in genomic DNA from leukocytes of male and female horses was determined, and the extent of its polymorphism investigated on Southern blots of DNA digested with various restriction enzymes. To investigate its organisation, a genomic clone containing $(GATA)_n$ was isolated, characterized by restriction mapping and sequenced.

$(GATA)_n$ constituted 1% of the horse genome. Like the mouse, there was no quantitative sex variation. *Mbo* I digestion generated a large number of horse DNA fragments of various sizes up to 5kb which hybridized to the $(GATA)_5$ probe. Simpler profiles were produced by digestion with *Taq* I, *Alu* I, *Hae* III and *Hinf* I. The profiles were highly conserved between individuals and between family members indicating the $(GATA)_5$ is unlikely to be informative as a DNA fingerprinting probe.

Some intensely hybridizing DNA fragments appeared to be maternally transmitted. This seems to be a novel observation.

A 3.6kb fragment which hybridized to the $(GATA)_5$ probe was cloned from horse genomic DNA. It was restriction mapped, the GATA-containing region identified and sequenced. Only about 150bp contained tandemly repeated GATA motifs in strings of about 3-6 repeats interspersed with $(GAT)_{1-2}$ regions.

The lack of quantitative sex variation suggests that $(GATA)_n$ may not have a role in sex determination in horses. Also, its lack of polymorphism makes it unlikely to be informative as a DNA fingerprinting probe.

ACKNOWLEDGEMENTS

I wish to thank the following for assistance during the various stages of this project:

My chief supervisor, Dr Tom Broad, DSIR who made this research possible and showed enthusiastic interest throughout. Co-supervisors Dr Ian Anderson of the Equine Blood Typing and Research Centre, Massey University who organised funding and blood samples; and Professor Barry Scott, Department of Microbiology and Genetics, Massey University for advice on the presentation of the thesis.

I gratefully acknowledge the generosity of the New Zealand Racing Conference for funding, enabling this research to be carried out. Thanks also to Mr Jack O'Brien of the Sovereign Lodge Stud for supplying blood samples.

Many thanks to Pauline Lewis, DSIR, for helpful advice, techniques and answering numerous questions. Thanks to Dr John Forrest who gave much assistance in the early stages of this project. Thanks also to all those involved in the Animal Gene Mapping programme who have given support and advice.

For help in tracing pedigrees thanks go to Bruce Farndale, EBRTC and Wrightsons Bloodstock, Palmerston North.

Thanks to Lee McNabb and DIT staff, DSIR for computing assistance.

For often useful advice thanks to Margaret Carpenter, Drs Murray Grant and Sin Phua, all of DSIR, Palmerston North.

Many thanks to all the great people I met at DSIR, Palmerston North who made working on this project enjoyable (most of the time).

Table of Contents

Table of Figures

1.0 Introduction

2.0 Material and Methods

3.0 Results

4.0 Discussion

Table of Tables

1.0 Introduction

2.0 Materials and Methods

3.0 Results

1.0 INTRODUCTION

1.1 REASONS FOR THIS STUDY

Shakespeares' King Richard III expressed the extreme value of horses to man in his urgent cry: "A horse, a horse, my kingdom for a horse!".

The horse was domesticated thousands of years ago. It has played a major part in the history of man, in fields as diverse as commerce, war and sport. The horse has had a profound affect on Man's perception of himself and the world. A symbol of power was recognized in the figure of a man on horseback by early civilizations.

Today the horse is still a visible member of society even though it has been ousted from many of its former fields of operation. A good example is the Thoroughbred. A multimillion dollar industry has evolved around this breed. It has taken 300 years of conventional breeding techniques to develop the Thoroughbreds' characteristics of speed and stamina to their current level (Wagoner, 1978). It could take just a few generations to develop these characteristics to the same level and beyond using modern molecular biology techniques.

A better understanding of the genetics of the horse is required before molecular biology can be utilized for breed improvement. The organization of the genes, how they are regulated, and their location within the genome needs to be known. As a small step along this pathway I have chosen to study the occurrence of the GATA repetitive sequence in the horse. This sequence has been present in nearly all those species so far studied and as such may prove to be a useful genetic marker.

1.2 CHARACTERISTICS, EVOLUTION AND SIGNIFICANCE OF REPETITIVE DNA

1.2.1 Introduction

A repetitive sequence is defined as a sequence of nucleotides which is reiterated in the genome. These sequences may be clustered in particular areas or interspersed throughout the genome amongst unique sequence DNA. The number of nucleotides which make up the sequence can vary from just a couple of base pairs (as in simple sequences) to many hundreds (as in some satellites). The number of reiterations of these sequences can range from a few tens to many thousands.

Repetitive DNA occupies a large proportion (about 30-40%) of the eukaryotic genome (Hardman, 1986). This contrasts with prokaryotes whose relatively small genomes consist predominantly of low copy-number DNA sequences.

In mammals, studies of repetitive DNA have focussed on human and mouse genomes. Relatively little is known about its organization in domestic animals, particularly the horse.

Present definitions of repetitive sequences are based on data obtained from sequencing, restriction enzyme cutting, Southern blotting and hybridization studies. These definitions may be overlapping and ambiguous for any particular sequence. Confusion also arises from the many terms used to describe these sequences, including: "elements", "repeats", "repeated DNA" and "repetitive DNA". These terms are used synonymously in this study.

For the purposes of this study, repetitive DNA in eukaryotes has been grouped as shown in Table 1.1. These groupings are by no means rigid as much overlap exists between the various classes.

REPETITIVE DNA				
Interspersed elements			Tandemly repeated sequences	
SINES	LINES	Classical satellites	Long tandem repeats	Short tandem repeats
Alu	L1	Bkm Sat I	midisatelites multigene families	simple sequences minisatellites sqr (GATA)

Table 1.1 Classes of Repetitive DNA

Classical satellites were the first type of repetitive DNA to be recognized. They had a buoyant density in cesium chloride which was different from the majority of an organisms' DNA (Lewin, 1986). Later studies involving DNA-DNA hybridization led to an increase in the numbers and types of repeated sequences reported.

Interspersed tandemly repetitive DNA sequences found in the mammalian genome were first described by Wyman and White (1980). These hypervariable (or minisatellite) regions of reiterated sequences showed multi-allelic variation and correspondingly high heterozygosity (Wyman and White, 1980). These minisatellites now have a wide range of applications (Jeffreys *et al*, 1985c).

1.2.2 Interspersed Elements

These single units are scattered throughout the genome. Two classes are recognised: short interspersed elements (SINES) less than 500 bp; and long interspersed elements (LINES) more than 500bp (Fowler *et al*, 1987).

1.2.2.1 SINES

The Alu family of primates is the most well characterized SINE. This consists of 300 bp repetitive DNA that can be cleaved at a common flanking site by the restriction endonuclease Alu 1 (Hardman ,1986). The Alu family accounts for a minimum of 3-6% of the human genome. It is a major fraction of SINEs in other mammals (Jelinek and Schmid, 1982). Other SINES include the NTS family originally found in the nontranscribed spacer (hence, NTS) region of ribosomal DNA. It contains oligonucleotide stretches homologous to Alu, but their significance is not clear (Singer, 1982).

The Alu element shows strong sequence conservation. In human DNA it usually consists of a head-to-tail tandem arrangement of two related sequences, each about 130 bp long terminating with an A-rich segment (Jelinek and Schmid,1982).

Some Alu-like SINEs may be representatives of a new class of eukaryotic mobile element. Alu-like sequences carry variable (A)-rich 3' tails and are flanked by terminal repeats, except where there is clear evidence of deletion. These terminal repeats are presumed to be analogous to those generated by target site duplication of eukaryote and prokaryote transposable element insertions. However, SINEs are unlike the better known transposable elements in that they are shorter, lack internal terminal repeat sequences and the length of the direct flanking repeats vary from one family member to another. Several SINE family members are transcribed *in vivo* (Singer, 1982).

Retroposon is a term used to refer to sequences which have RNA origins and dispersed positions. Some repetitive sequences such as the Alu family are thought to be generated from RNA intermediates by a mechanism involving reverse transcription (Rogers, 1984). Common properties include, sequence boundaries exactly corresponding to RNA species; a repetitive (A)-rich tail at the 3' end; and direct terminal repeats of 8-19 bp of the flanking sequences at the 5' and 3' ends. In several instances Alu-like sequences have inserted into a known target sequence and the terminal repeat is demonstrable as a duplication of the target sequence.

Alu-like sequences and retroposons in general, have a strong tendency to insert into each others (A)-rich tails generating composites, which are themselves propagated as single retroposons. The primate Alu is the classic example, being a dimer of homologous sequences. The first (Alu.A) carries the functional RNA polymerase III promoter and the second (Alu.B) has a substantial "insert". These sequences are homologous to the 7SL RNA gene (Ulle *et al*,1984) where the "insert" is larger. 7SL is an essential functional RNA, involved in the synthesis of secreted proteins, and is a polymerase III transcript. Thus Alu may be a dimer of 7SL pseudogenes, which have arisen by internal deletions.

1.2.2.2 LINES

There are only a few known families of LINEs. L1 is the single major LINE family in primates. It constitutes 1-2% of the human genome. Homologous L1 sequences are found in other mammals, for example, the MIF-1 family in mice. L1 probes hybridize to many scattered chromosomal locations. They flank genes and are found in introns and within centromeric satellite DNA.

L1 elements are highly variable in length, from approximately 500bp to 7 kb. Most have common 3' ends with variable A-rich tails. They are heterogeneously 5' truncated. Most of the 5' truncated specimens which have been totally sequenced are flanked by short direct repeats typical of transposons. L1 elements are also often rearranged through inversions, internal deletions, and other permutations typical of linear sequences (Singer and Skowronski, 1985). Within the L1 family, most have common 3' ends with variable (A)-rich tails. Most of the 5' truncated specimens which have been totally sequenced are flanked by typical short direct repeats.

L1 elements may foster rearrangements both within L1 elements and in neighbouring genomic regions. L1 elements possess an open reading frame which is conserved in rodents and primates (Rogers, 1984). Variants may therefore arise via foldback of the nascent cDNA strand during reverse transcription. This sometimes occurs during in vitro cDNA cloning (Rogers, 1984).

L1 is transcribed at low levels by polymerase III but is not polyadenylated and is confined to the nucleus. No full length LINEs have been completely sequenced, nor have the transcription unit(s) been mapped. The origin of full length L1 LINEs is therefore unknown (Rogers, 1984). L1 is probably a multigene family composed of some functional genes and a large number of pseudogenes, many of which are truncated. This family is different from other multigene families, firstly, in that the copy number is very high compared to even the largest family described thus far - the U1-RNA gene family of humans, which has fewer than 100 genes and about

1000 pseudogenes. Secondly, the pattern of truncation is unique (Singer and Skowronski, 1985). It is not known how many functional genes are included in the L1 family or when and where the putative genes are expressed.

Some of the many non-coding family members may influence the expression of neighbouring genes in significant ways. One truncated mouse segment was shown to enhance transcription of an expression vector construct in monkey Cos1 cells. If they are important modulators of the expression of neighbouring genes, L1 segments might associate with actual genes. This association might be expected to be conserved among mammalian genes. The presence, therefore, of L1 units in similar relative positions downstream of the beta globin genes in mice and humans is potentially interesting (Singer and Skowronski, 1985).

1.2.3 Tandemly Repeated Sequences

These sequences comprise 5-10% of mammalian genomes. They are characterized by the head-to-tail repetition of lengths of DNA, generally of some common sequence.

1.2.3.1 Satellite DNA

These were isolated on CsCl density gradients. They are normally specific for a given taxonomic family, or in some cases genus or species (Fanning, 1987). In general, these simple, tandemly repeated sequence arrays are present in centromeric and telomeric heterochromatin. Normally they are transcriptionally quiescent.

The length of the simplest repeating unit in each class is generally constant, but sequence divergence within these units is possible giving a "family" of sequences within each class. The repeat units may be as small as 4-5bp (eg snake Bkm satellite, human Sat II and Sat III) but are more typically 170-250bp long. The

sequences of individual satellite family members in any one class are chromosome specific or nearly specific in origin (Fowler *et al*, 1987a).

Many satellite DNAs are believed to be the product of duplication-amplification events. For example, a short monomer sequence of 5-50 bp may duplicate to form a dimer. Over time, the dimer accumulates random base substitutions and at some point, duplicates to form a tetramer. Superimposed over this small process is a second, larger process whereby sections of the repeat structure are amplified, often giving rise to hundreds or thousands of tandemly linked copies. Many examples of satellites that have arisen by such process are known in rodents, primates, artiodactyls and insects (Fanning, 1987). The exact biochemical mechanisms giving rise to these sequences is unknown. Initially, it is thought, some type of slippage during DNA replication is involved followed by unequal crossover between the tandem arrays. An exception to the dimer formation is satellite II of the domestic goat (*Capra hiraus*). It has 700bp repeat units present in the genome primarily in the form of 2100bp trimers. This particular satellite DNA may represent one of the few cases when the unequal crossover mechanism does not give rise to a dimeric structure (Buckland and Elder, 1985).

No entirely convincing evidence exists for a function of satellite DNA sequences in somatic tissues. They may have functional roles in the germ line, for example, in the regulation of recombination at meiosis. These are undermethylated in the germ line. This is opposite to the situation with specific gene sequences that are methylated and inactive in the germ line, and undermethylated when actively transcribed in somatic tissues. This may point to a germ line function for some satellites, correlating with selective hypomethylation of their sequences. The true significance of the observation, however, is not yet understood (Hardman, 1986).

Satellite sequences may have a structural role in chromosome centromeres or telomeres. Telomeres often contain repeated but quite complex DNA sequences

which may extend for many kilobases from the molecular end of the chromosomal DNA. These "telomere-associated" sequences may mediate many of the telomere-specific interactions that occur both among telomeres and between telomeres and the nuclear envelope. Sequences at, or very close to, the extreme ends of the chromosomal DNA molecules consist of simple, satellite-like, tandemly repeated DNA sequences. It is likely that these "simple telomeric" sequences are essential functional components of telomeric regions. These are needed to supply a chromosomal end with both stability and the ability to be completely replicated (Blackburn and Szostak, 1984).

1.2.4 Long Tandem Repeats

1.2.4.1 Midisatellites

These consist of long tandem repeats of simple sequences. One has been found in the human genome. It consists of some 250-500kb of repetitive DNA that is clustered at a single locus near the telomere the short arm of chromosome 1 (Nakamura *et al*, 1987). It contains a core sequence which bears some homology to the repetitive sequence of the insulin gene and the zeta-globin pseudogene. It is suspected that the sequence GTGGG, which is common within at least four different kinds of repeating units and is similar to the lambda chi sequence, may have a role in recombination (Nakamura *et al*, 1987).

The genomic organisation of the "midisatellite" differs from the other "minisatellite" loci reported, with respect to copy number, the size of the locus, and its extremely polymorphic pattern (Nakamura *et al*, 1987).

1.2.4.2 Multi-gene Families

Ribosomal 5S RNA genes and histone genes in some but not all organisms are examples of long tandem arrays of complex repeated sequences. Some portions

are transcriptionally active and represent multigene families in which the copy number per haploid genome varies between a few hundred to many thousands (Hardman, 1986).

1.2.5 Short Tandem Repeats

1.2.5.1 Simple sequences

Simple sequences are mostly less than 100 bp long. They consist of only one, or a few tandemly repeated nucleotides. They are interspersed in many eukaryotic genomes near genes, in some introns and in DNA regions between immunoglobulin genes. They have also been found within variants of the repetitive Alu-elements, within satellite sequences, as well as in other regions of the genome that can not be related to any function (Tautz and Renz, 1984). All types of simple repetitive sequences probably exist.

Simple sequences may have arisen by slippage or unequal crossover which took place at randomly occurring short runs of the sequences. Both mechanisms would lead to constant formation and deletion of simple sequences. They would be expected to be found in all regions of the genome which do not undergo selection. Hence, the occurrence of simple sequences in eukaryotes is not a matter of evolutionary conservation, but instead depends on a number of factors, including: (i) the frequency of accidental amplifications and deletions; (ii) the extent to which the mechanisms spread the sequences between homologous chromosomes; (iii) the degree to which the sequences are tolerated in the genome; and (iv) on the number of possible formation sites for simple sequences, ie, redundant DNA. The absence of large amounts of simple sequences in prokaryotes could be due to any one of these factors, singly or in combination.

It is possible that some simple sequences might have been formed and distributed in the genome by additional mechanisms. For example, AA/TT may equally well

arise by reverse transcription of poly A tails of mRNA and integrated into the genome. However, subsequent slippage and unequal crossover must be expected to occur in all simple sequence regions regardless of their actual mode of origin.

Simple sequences are distinctly different from simple satellite sequences in that they are interspersed in the genome and are usually transcribed into RNA. Different types of simple sequences can be clustered within a small region of DNA, eg, CpG islands which differ from bulk DNA by being non-methylated at CpG dinucleotides. These sequences occur as discrete islands usually 1-2 kb long and are dispersed in the genome. There are approximately 30 000 islands in the haploid genome of mammals (Bird, 1987). GpC dinucleotides are rare in eukaryotic DNA but where they occur, they are often found clustered near the 5' ends of certain genes, where they presumably fulfill a functional role and are maintained by selection (Fanning, 1987). The proportion of islands in the genome that mark genes is likely to be large (Bird, 1987). The CpG dinucleotides found in interisland DNA are methylated at the 5' cytosine residue. As a consequence of methylation, cytosine is prone to deamination giving rise to thymidine. This could account for the low number of CpG dinucleotides in interisland DNA.

Several suggestions have been made concerning a possible function of simple sequences: in chromatin folding; homogenization of repetitive gene arrays; as "hot-spots" for recombination; in the evolution of new genes; in telomere formation; and, in gene regulation (Tautz and Renz, 1984). All these proposals are concerned only with certain types of simple sequence. In general, however, the predominant role of simple sequence repeats may be for recombination. This is supported by the fact that simple sequences may easily form single stranded regions, which are due to slippage. These single stranded regions might serve as "hot-spots" for strand invasion during initiation of the recombination event. They might also be able to combine different chromosome regions which otherwise share no homology, a mechanism which has been proposed for the switching region of immunoglobulin genes. Simple sequences should, therefore, be

regarded as a source of naturally occurring rearrangement and variation (Tautz and Renz, 1984).

1.2.5.2 Minisatellites

Regions made up of short tandemly repeated sequences are known as minisatellites. Many have been found near or within genes often because the gene and its surrounds were being studied (Fowler *et al*,1987). There is estimated to be at least 1500 "minisatellites" in the human genome (Fowler *et al*, 1987).

Minisatellites do not constitute a true "family" of sequences: specifically, they are not directly derived from each other in the way a family of transposable sequences might be. They are, however, related in the sense that they are based on very similar "core" sequences. These are in the region of 15 bases long and constitute the basis of the repeat unit (Lewin, 1986). Some minisatellite core sequences show remarkable conservation throughout nature . A minisatellite-like sequence found in protein III of the wild type M13 phage has been the most interesting found so far. It has been used to locate variable number tandem repeats (VNTRs) in human, bovine, equine, murine and canine genomes (Vassart *et al*, 1987).

The core sequence may help generate minisatellites by promoting the initial tandem duplication of unique sequence DNA and/or by stimulating the subsequent unequal exchanges required to amplify the duplication into a minisatellite (Jeffreys *et al*, 1985a). Tandemly repetitive sequence arrays may be the normal, expected consequences of a situation where unequal crossovers are not actually prevented. This mechanism operates independently of selective pressure. It could, however, be adapted to amplify selected genes which may confer some phenotypic advantage. Tandemly repeated genes such as 5S RNA, histones and rRNA are commonly found in eukaryotic genomes. Amplification of dihydrofolate reductase genes in cells treated with methotrexate is an extreme case of the rapid amplification of a tandemly repeated eukaryotic gene family under conditions of strong selective pressure (Hardman, 1986).

Minisatellites can be polymorphic due to an insertion/deletion mutational event causing lengthening or shortening of the overall fragment. The length of the repeat unit inserted or deleted is typically between 10 and 64 bp. Tandem arrays of such units may exist at either unique or a number of dispersed genomic sites (Fowler *et al*, 1987).

Minisatellites have been found which are highly variable with respect to the number of repeat cores found at a locus in a population. These have been referred to as hypervariable minisatellites (Jeffreys *et al*, 1985a). Only a limited number of hypervariable loci have been discovered in human DNA; these include minisatellites 5' to the insulin gene, alpha globin gene, type II collagen gene, apolipoprotein B gene, and the D14S1 locus. These minisatellites differ substantially in their variability, ranging from only 6 different alleles detected at the collagen hypervariable region, to more than 80 at the D14S1 locus (Wong *et al*, 1987).

A number of hypervariable loci studied in mice showed that they were autosomal, dispersed, not preferentially associated with centromeres or telomeres (Jeffreys *et al*, 1987).

Hypervariable minisatellites may be recombination "hot-spots". The core sequence is similar in length and G content to the chi sequence, a signal for generalized recombination in E. coli . Hence similar sequences might be used for related mechanisms in eukaryotes (Jeffreys, 1987).

1.3 THE GATA SEQUENCES

These are a subfamily of the "simple quadruplet repeats" or "middle repetitive, dispersed DNA sequences" found initially in the banded krait minor (Bkm) DNA satellite, isolated in a CsCl density gradient. This satellite was visible in DNA from only females of the Indian snake *Bungarus fasciatus* (Singh *et al*, 1980). The satellite was conserved throughout the snake group and mainly concentrated on

the sex determining W chromosome (Singh *et al*,1980). Snakes lacking sex chromosomes possessed related sequences but these had no sex-associated differences (Singh *et al*, 1980). A major component of the Bkm satellite was the simple quadruplet repeat $(GATA)_n$ (Singh *et al*, 1981)

Sequences cross-hybridizing to Bkm have since been found in various eukaryotes, from slime-molds to man (Arnemann *et al*, 1986; Singh *et al*,1984). Cloned Bkm-positive genomic fragments of Drosophila and mouse contained long tracts of the tetranucleotide GATA (Singh *et al*, 1984). A probe of long repeated tracts of GATA gave hybridization patterns similar to the original Bkm probe (Schafer *et al* 1986a; Singh *et al*, 1984). However, *in situ* hybridization of a short repeat, $(GATA)_4$ did not (Schafer *et al*, 1986a). $(GATA)_n$ sequences have since been found in vertebrates, invertebrates and plants, but not in any significant length in the ovine or bovine genomes (Miklos *et al*, 1989; Weising *et al*, 1989).

In addition to GATA repeats, clusters of GACA were found in a genomic clone of a female specific satellite DNA from the snake, *Elphe radiata.* Genomic and cDNA clones from Drosophila and mouse singled out using this snake clone as a probe also contain GACA interspersion (Epplen *et al*, 1982; Schafer *et al*, 1986a).

1.3.1 Chromosome Locations of GATA sequences

In situ hybridization has shown $(GATA)_n$ sequences occupy a concentrated area on the Y chromosome in mice, but most of the grains hybridized to the autosomes (Schafer *et al*, 1986a). The Y chromosome of mice appeared to contain a disproportionately large amount of simple repetitious DNA. An attractive explanation for this is that long tandem arrays of simple repeated sequences are generated at high frequency throughout the genome. They are retained for longer on the Y chromosome due to the absence of homologous pairing at meiosis (Platt and Dewey, 1987).

A Bkm probe showed significant hybridization to the sex determining Y chromosome in both XY male and XY sex reversed female horses (Kent *et al*, 1988). They were also found on chromosomes 3, 4 and probably 30. The degree of Bkm hybridization on these autosomes was much less than that seen on the Y chromosome. This suggested that fewer GATA repeats were present in these autosomes than the Y chromosome (Kent *et al*, 1988).

GATA sequences have been predominantly found in heterochromatic regions (Nanda *et al*, 1988). These regions are generally transcriptionally silent.

1.3.2 Possible Functions of GATA Sequences

The presence of these sequences in a wide range of organisms does not necessarily imply that they must have some kind of conserved function.

It was postulated that GATA sequences were in some way associated with sex determination because of their strong association with the sex chromosomes (Kiel-Metzger and Erickson, 1984; Chandra, 1985). Most eukaryotes tested to date have varying tracts of GATA sequences which are positioned as to correlate with some aspect of sex determination (Epplen *et al*,1988), sexual differentiation, sex chromosome differentiation (Jones and Singh, 1985), dosage compensation or X inactivation (Miklos *et al*, 1989). However, Durbin *et al*, (1989) showed that although chromosome 17 of mice did possess Bkm-related sequences they could not be related to those regions on chromosome 17 involved with sex determination.

No sex linkage of Bkm has been detected in the moth *Ephestia kuehniella* (Traut, 1987). $(GATA)_n$ tracts of any significant length (as reflected by hybridization intensities) are absent from bovine, ovine and chicken genomes at standard hybridization stringencies (Miklos *et al*, 1989). Some middle repetitive DNA sequences are located exclusively on the sex determining W chromosome of some bird species. These same sequences, however, are totally absent from other bird

species. Hence, the very restricted and intriguing sex chromosomal pattern is not conserved even within birds (Tone *et al*, 1984).

Mice appeared to show male-specific transcription of GATA sequences in the liver (Schafer *et al*, 1986b). However, this sex-specific transcription was not conserved in other rodents such as rats (Miklos *et al*, 1989).

The abundant occurrence of these sequences may reflect their involvement in roles such as regulation of gene expression, especially at the transcriptional level; as "hot-spots" for gene recombination or rearrangement; or they could be especially reactive with mutagens and carcinogens (Hamada *et al*, 1984).

GATA sequences may have an unique DNA conformation *in vivo*. This may be similar to the (T-G) and (C-G) elements which have been shown to form Z-DNA *in vivo* (Hamada *et al*, 1984). Interconversion between the B and the Z forms may play a role in gene regulation. Reversible interconversion would change the distortion of DNA at a proximal or distal site resulting in activation or inactivation of associated genes (Hamada *et al*, 1984).

The available functional options that can be invoked for this family of sequences is seriously limited by the discovery of two mammalian genomes (bovine and ovine) which lack $(GATA)_n$ tracts of any reasonable length (Miklos *et al*, 1989).

1.3.3 Are GATA sequences transposable elements?

Hypervariable Bkm cross-hybridizing sequences were found on the autosomes of the moth, *Ephestia kuehniella* (Traut, 1987). They were unusual in two respects: firstly, changes of restriction fragment length polymorphisms appeared at a high rate in the offspring of some crosses but were not present in others. Secondly, homologous loci could be "empty" of Bkm cross-hybridizing components (Traut, 1989). The high rate of restriction fragment size changes as well as the loss of Bkm positive material in some hybrids and the stability of fragments in others is

reminiscent of the bursts of transposition (Traut, 1989). However, unlike transposable elements in other organisms, the putative transposable Bkm elements of *Ephestia* were concentrated on two or three autosome pairs, at least in those strains investigated (Traut, 1989).

Most of the dispersed, middle repetitive DNA sequences in Drosophila, also belong to the mobile element class (Finnegan and Fawcett, 1986).

Five cDNAs from mouse which contained $(GATA)_n$ sequences have been sequenced. Nearly all the cDNAs possessed octomeric inverted repeats which flanked the $(GATA)_n$ and/or $(GACA)_n$ tracts. Most octomers began with TG and ended with CA. Thus, they were similar to the TG....CA sequences of mobile elements, Mu bacteriophage and various retroviruses (Schafer *et al* 1986).

1.3.4 Are GATA Sequences Transcribed and Translated?

One of the mouse cDNAs referred to above had a long open reading frame which included $(GATA)_n$ and $(GACA)_n$ tracts, whereas the other four cDNAs had frequent stop codons distributed throughout the cloned inserts (Schafer *et al,* 1986). These mouse cDNA sequences may be transcribed in a developmentally specific manner as are some Drosophila mobile elements. Alternatively, some $(GATA)_n$-containing sequences may well be transcribed by default due to read-through from nearby genes (Stephenson *et al*, 1981).

Tissue specific transcription to poly$(A)^+$-RNA appeared to occur to numerous regions of GATA sequences in blowflies. GATA sequences appeared to be actively transcribed during all stages of development investigated. When genomic DNA of blastoderm embryos was compared with adult genomic DNA some loci hybridizing to GATA displayed a marked stage-specific variation in length

(Kirchhoff, 1988). Stage- and tissue-specific differences in GATA transcription may point to a "sequence dependent" function, but equally they may simply reflect the general differential gene activity of specialized tissues (Kirchhoff, 1988).

Repeats of the quadruplet GATA produce a hypothetical hydrophobic repeated sequence of the four amino acids Leu-Ser-Ile-Tyr after transcription and translation. It is not yet known, however, whether the RNAs are indeed translated.

1.3.5 Possible Origin of GATA sequences

These sequences may have arisen independently in several taxa by a process involving slipped-strand mispairing of the two strands of DNA and/or unequal recombination (Levinson *et al*, 1985). Nevertheless, they may have specialized functions, such as the modification of nearby gene expression, as has been shown for genetically engineered constructs containing simple repeats (Hamada *et al*, 1984). Therefore their accumulation on sex chromosomes might be favoured by natural selection.

1.4 APPLICATIONS OF REPETITIVE DNA SEQUENCES

1.4.1 DNA Polymorphisms Arising from Repetitive Sequences

1.4.1.1 VNTRs

DNA polymorphisms can be due to the number of tandem repeats present in a sequence. The sequence may be present at a number of loci in the genome. A restriction enzyme which cuts outside the tandemly repeated sequence is used to demonstrate this type of polymorphism (fig 1.1a). Hence, this type of polymorphism is called a variable number tandem repeat (VNTR) or minisatellite. Detection of VNTRs is not dependent on the restriction enzyme used, provided it does not cleave the repeat unit. VNTR loci provide ideal genetic markers.

1.4.1.2 RFLPs

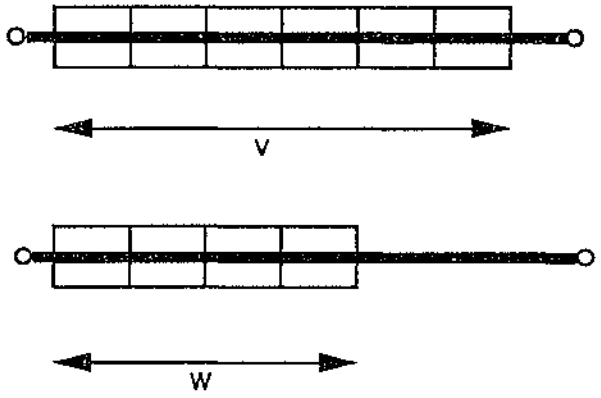
Length variation can be due to the formation or deletion of a restriction site. The restriction site can be located within or outside the repeat sequence. Digestion with the particular restriction enzyme can therefore demonstrate this type of polymorphism (fig 1.1b). This type of polymorphism is called a restriction fragment length polymorphism (RFLP). A specific base mutation generates RFLPs.

Detection of RFLPs and VNTRs forms the basis of the following applications:

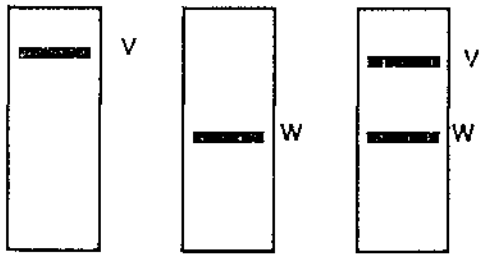
i) Pedigree identification

An individual-specific "fingerprint" of DNA bands can be produced using a specific probe. When this probe is based on the core tandem repeat sequence of a hypervariable minisatellite, it detects many highly variable loci simultaneously. The resulting "DNA fingerprint" somewhat resembles the bar codes commonly found on retail goods.

Fig 1a **VNTR**

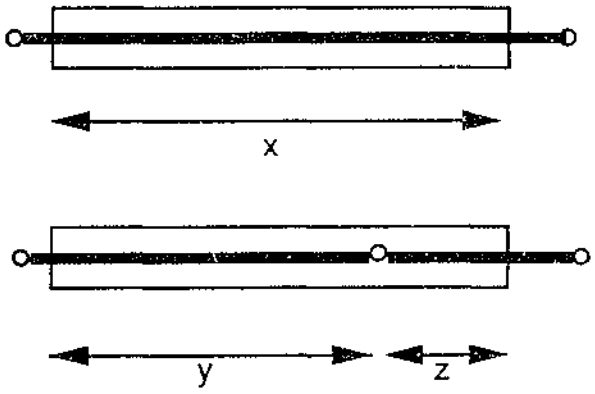


High variation in the number of repeat units between individuals

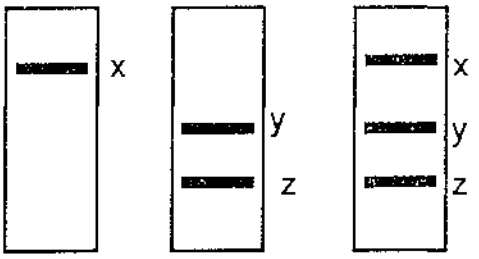


Homozyogous Homozygous Heterozygous

Fig 1b **RFLP**



new restriction site arises by mutation or conversion



Homozyogous Homozygous Heterozygous

Key

o Restriction endonuclease site
□ Repeat unit (in VNTR)

Fig 1.1 VNTR's and RFLPs

In humans, the technique can be applied to DNA obtained from samples of blood, semen (Morton *et al*, 1987) and body tissue (eg hair roots, Higuchi *et al*, 1988). The techniques strength is the possibility of positive identification of an individual through genetic tests, not just exclusion of identity (Lewin, 1986).

New length alleles of hypervariable human minisatellites arise from mutations. Mutations are sporadic, occurring with similar frequencies in sperm and oocytes. They can involve the gain or loss of substantial numbers of repeat units, consistent with length changes arising primarily by unequal exchange at meiosis. The mutation rate is sufficiently high to be directly measurable in human pedigrees. Germline stability must therefore be taken into account when using hypervariable loci as genetic markers, particularly in pedigree analysis and parenthood testing (Jeffreys *et al*, 1988)

DNA fingerprinting was first described in humans but has since been applied to a wide variety of other animals including: birds, cats, dogs, horses, mice, pigs, sheep, house sparrows and yaezes (goat x ibex)(Morton *et al*,1987).

There are many instances in veterinary work where this technique could be of value. For example, confirmation of identity of thoroughbred horses, inbred strains of laboratory animals, genetic identity of cell lines as well as many research interests such as chimaeras, cloning, etc (Morton *et al*, 1987)

ii) General linkage analysis and gene mapping

VNTRs are often located at unique loci near genes. When restriction enzymes cleave VNTRs, part of the flanking DNA may be cleaved. The flanking DNA can be cloned. It can then be used as a probe for gene mapping and to investigate (by cross hybridization) the distribution of similar sequences located elsewhere in the genome.

iii) Marker-assisted selection

Genetic improvement of animal populations is limited by the fact that most traits of economic importance are polygenic in nature and are influenced by a variety of environmental and developmental factors. Therefore it is generally not possible to determine the genotype of any particular individual by examination of phenotype alone. Traits of this nature are termed "quantitative traits" and the polygenic loci involved in their expression are termed "quantitative trait loci" (QTL) (Beckman and Stoller, 1987).

There is a problem in identifying QTL and manipulating them in breeding programs. RFLPs in agricultural populations can be examined for direct effects on traits of economic value, while linkage relationships between RFLPs and QTL can enable RFLPs to be used as genetic markers to monitor the transmission of useful QTL alleles from parent to offspring in the course of breeding populations (Beckman and Stoller, 1987).

Numerous polymorphic markers could make the accurate identification of breeding stock and their derivatives possible so that patenting of improved stocks could be feasible. Unique or rare alleles or combinations of alleles at several marker loci might be used to allow accurate genotyping and discrimination among stocks. Another use could be to monitor the introgression of a gene, or genes from one stock into another, by selective backcrossing or crosses with a rare or unique marker haplotype (or marker "bracket") which includes the gene (Smith and Simpson, 1986).

iv) Linkage analysis of disease susceptibility

Genetic disease loci can often be mapped by correlating the inheritance (or segregation) of a disease trait with the inheritance of a specific chromosomal

region. This involves studies of genetic linkage in families (Nakamura *et al*, 1987b).

Locating defective genes associated with inherited diseases requires good genetic markers. The fragments which make up DNA fingerprints may be the best markers so far characterized (Lewin, 1986). They are only useful within very large families, not between families. However, within families they can point to an association between a fingerprint band and a disease locus. Then subsequent cloning of the band is required to generate RFLPs to be screened and selected by population studies. The usefulness of probes for genetic analysis is affected by the frequency of RFLPs as well as their recombinational distance from the disease gene (Caskey, 1987).

Linkage analysis has localized the genes responsible for several major genetic diseases, including Huntington's chorea, Duchenne muscular dystrophy, adult polycystic kidney disease, and cystic fibrosis (Nakamura *et al*, 1987b).

DNA fingerprinting with synthetic GATA/GACA oligonucleotide probes has revealed a high level of RFLPs in the sex reversing (Sxr) region in mice (McLaren, 1988). This is an example of GATA sequences acting as genetic markers.

1.4.2 Other potential applications using repetitive DNA

Unlike the above, these applications are base on invariant characteristics of particular repetitive sequences:

i) Chromosome-specific identification

Specific chromosomes could be identified using probes based on chromosome-specific satellite DNA sequences. *In situ* hybridization techniques could be used

with a high degree of accuracy with such probes. This sort of accuracy is important, for example, in identifying chromosomes in somatic cell hybrids.

ii) Species-specific identification

Closely related species could be differentiated using sequences which are unique to particular species (such as the minisatellite types). This sort of application may be useful in conservation work where hybrids of closely related species are occurring and monitoring the individual species based on phenotypic character becomes difficult.

iii) Other uses

These include determination of engraftment or rejection of donor cells following tissue transplantation and studies on tumour clonality.

1.5 GENETICS OF THE HORSE

The horse family Equidae, consists of a single genus *Equus* with seven generally recognised species. This genus is particularly well represented in paleontologic records and is believed to have diversified 4-5 million years ago into the lines leading to present day forms (Ryder *et al*, 1978). Prezwalski's horse, *E prezwalskii*, is the only true wild (nonferal) horse and is thought to be the ancestor of the domestic horse, *E caballus*. Horses with features apparently identical to those of Prezwalski's horse are vividly depicted in the cave paintings of southern France and northern Spain.

The domestic horse has a diploid chromosome number of 64, including the X and Y chromosomes (Ryder *et al*, 1978). Very little has been published about the genomic organisation of the horse at the DNA level.

1.6 AIMS OF THIS INVESTIGATION

To undertake an investigation of GATA repetitive sequences in the domestic horse (*Equus caballus*). These sequences belong to the class of simple quadruplet repeats (sqr). They have been shown to be present in a number of eukaryotes.

Key questions to be answered are:

(a) are these repeats present in the horse genome?

(b) if they are, at what frequency (high, medium or low)?

(c) how are they organized (tandem arrays vs single interspersed repeats)?

(d) do they contribute to DNA polymorphisms in the horse, of the RFLP or VNTR types?

(c) does the level of polymorphism present make them suitable as DNA fingerprinting probes (ie sufficient to distinguish individuals within family groups)?

(f) do they cross hybridize with other tandem repeats in the horse?

MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Solutions

Commonly used stock solutions were made according to the procedures of Maniatis *et al* (1982). Abbreviations of stock solutions used throughout this project are listed in table 2.1.

2.1.2 Horses Used in this Investigation

Table 2.2 a-j shows the pedigrees of the horses used. Data was obtained from the New Zealand Thoroughbred Stud book and from Wrightson Bloodstock Ltd, Palmerston North, New Zealand.

2.2 THE GATA OLIGOMER

The oligonucleotide JF-1, prepared by Dr J Cutfield at Otago University on an ABI DNA synthesiser, was a 20mer consisting of the nucleotides G-A-T-A repeated five times, ie $(GATA)_5$. A second oligonucleotide, JF-2, was made which was complementary to JF-1 but out of phase by one nucleotide. It was a 20mer consisiting of the nucleotides $TAT(CTAT)_4C$.

2.2.1 Oligonucleotide end-labelling

This method utilizes the ability of the enzyme, T4 polynucleotide kinase to catalyse the transfer of the gamma phosphate of ATP to a free 5'OH terminus of double or single stranded DNA or RNA. JF-1 was radioactively end labelled with [γ-^{32}P] ATP (3000Ci/mmol, NEN) following the procedure of Forrest (1988).

EDTA	disodium ethylenediaminetetraacetate
IPTG	isopropyl thiogalactoside
SDS	sodium dodecyl sulphate
SSC	saline sodium citrate
SSPE	sodium chloride, sodium dihydrogen phosphate, EDTA
TBE	tris, borate, EDTA
TE	tris, EDTA
TEMED	N,N,N',N',-tetramethylethylenediamine
TES	tris, EDTA, NaCl
Tris.Cl	tris(hydroxymethyl)aminomethane, at various pH's
X-gal	5-bromo,4-chloro,3-indolyl, D galactopyranoside

Table 2.1 Abbreviations of Stock Solutions

Table 2.2 Horses used in the Investigation

The following pedigrees are displayed in the standard manner: the paternal parent is on the upper branch and the maternal parent is on the lower. For example, Kindom Bay's sire was Otehi Bay and his dam was Golden Praise. His paternal Grandsire was Biscay and maternal Grandsire was Golden Plume (Table 2.2a)

Table 2.2a Pedigree of Kingdom Bay (stallion)

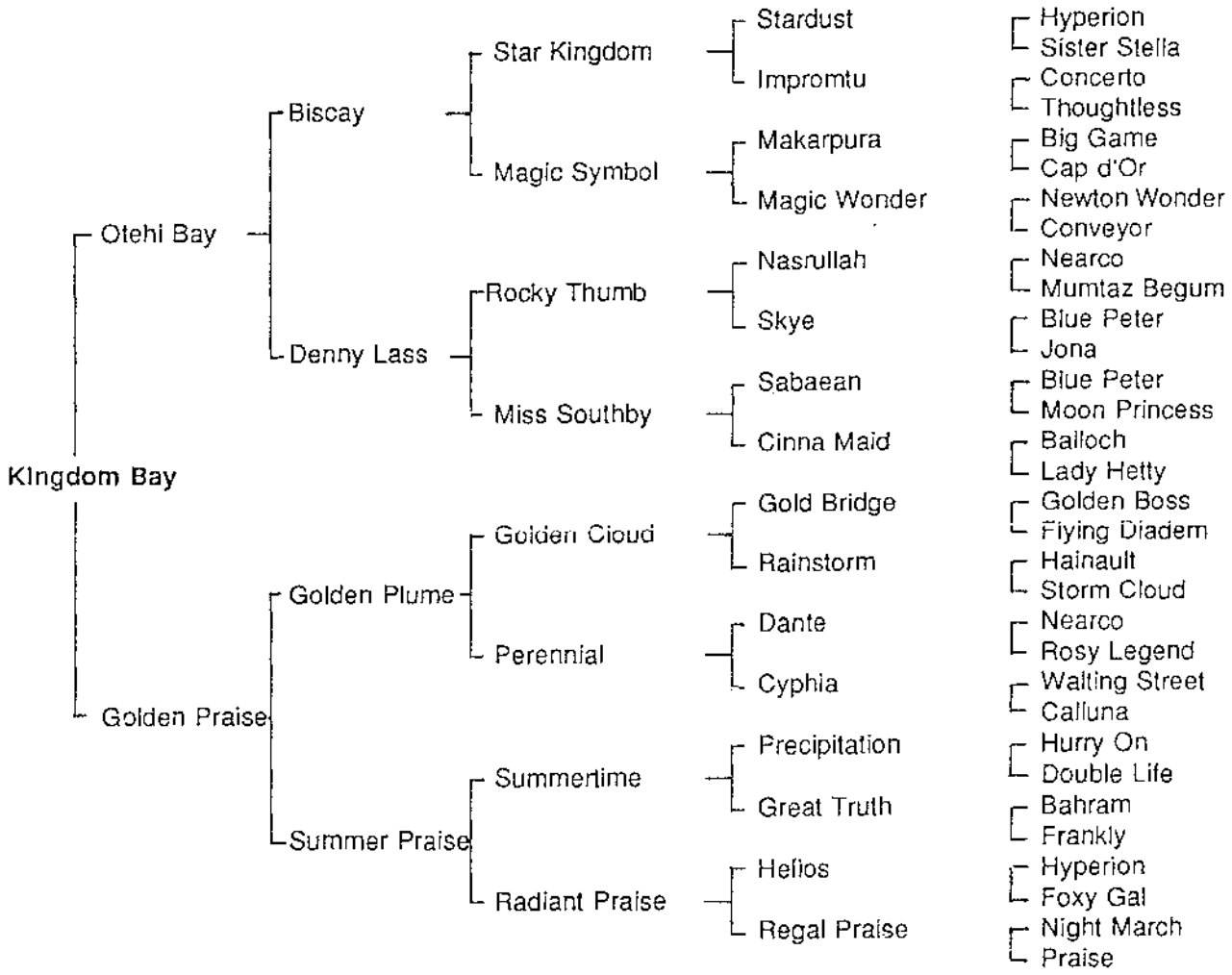


Table 2.2b Pedigree of Western Bay (stallion)

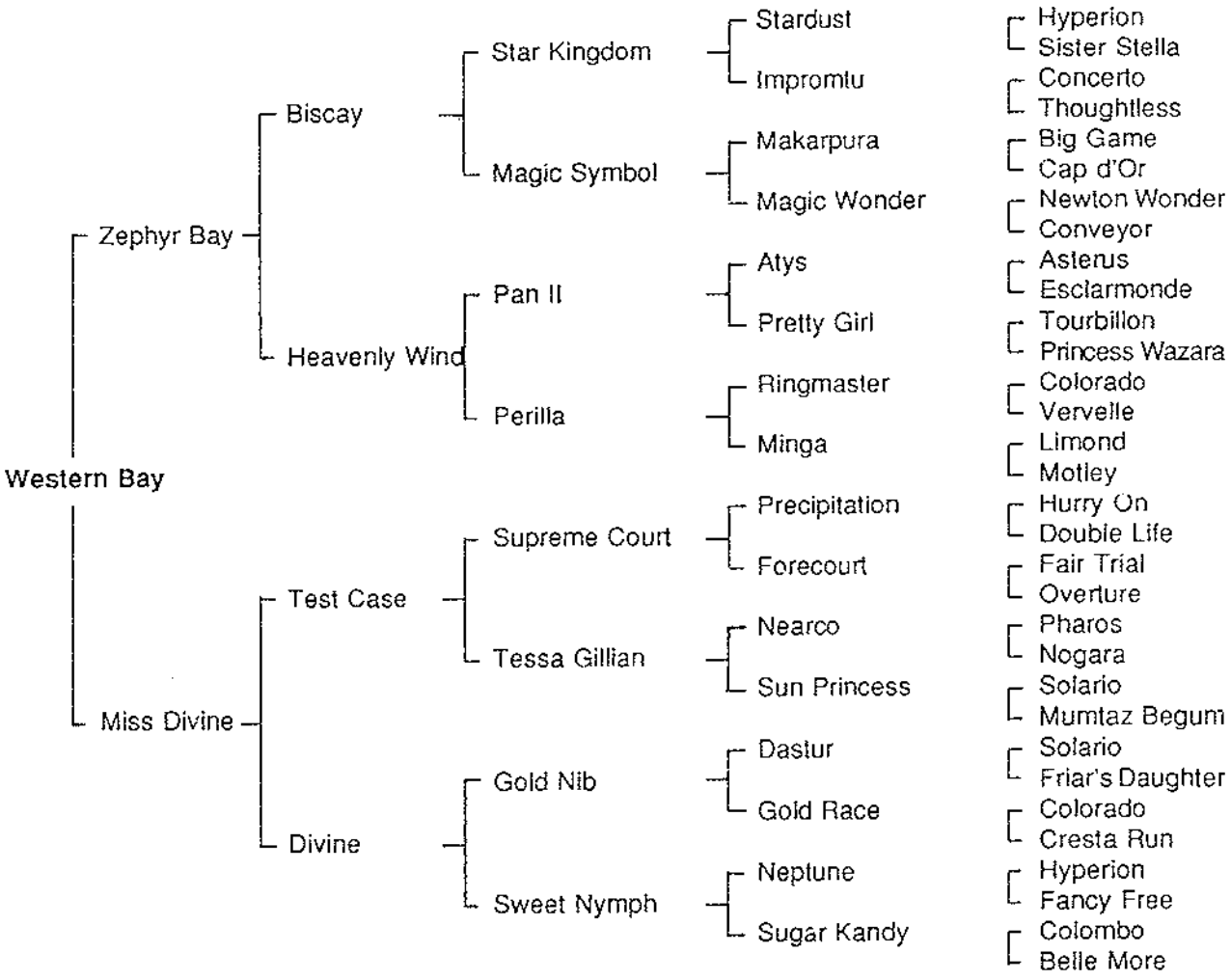


Table 2.2c Pedigree of Darling Daughter (mare)

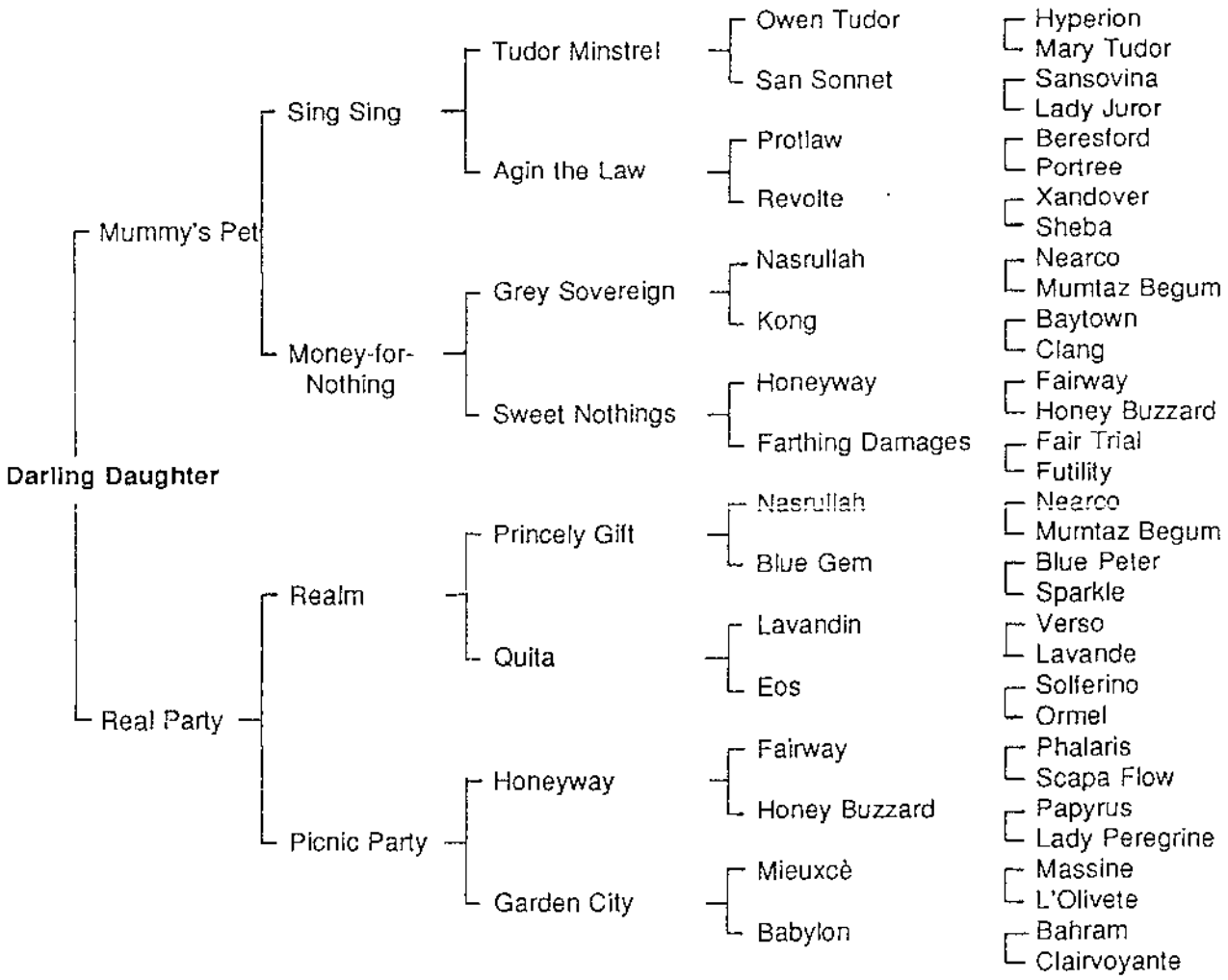


Table 2.2d Pedigree of Latchmi (mare)

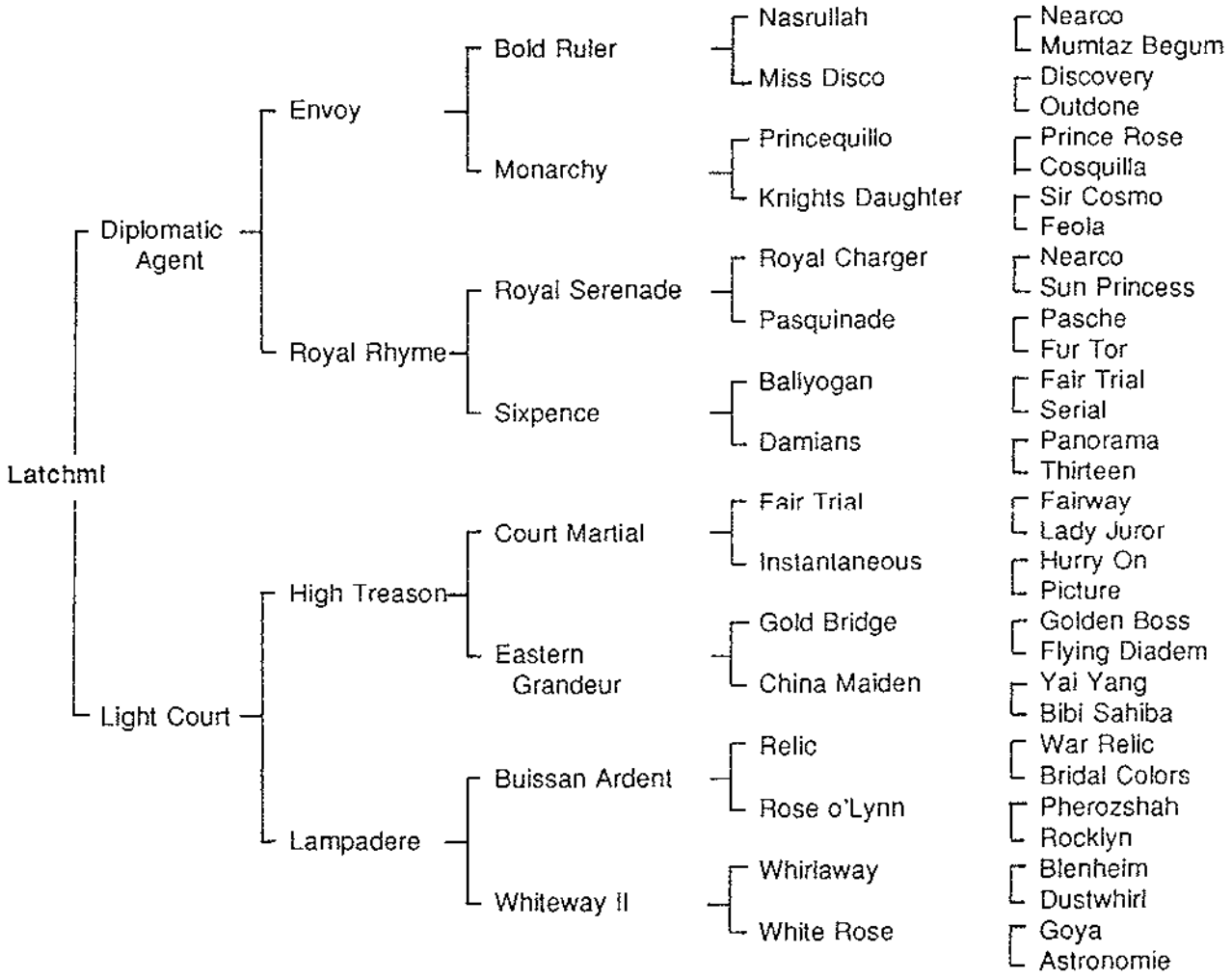


Table 2.2e Pedigree of Lovely Habit (mare)

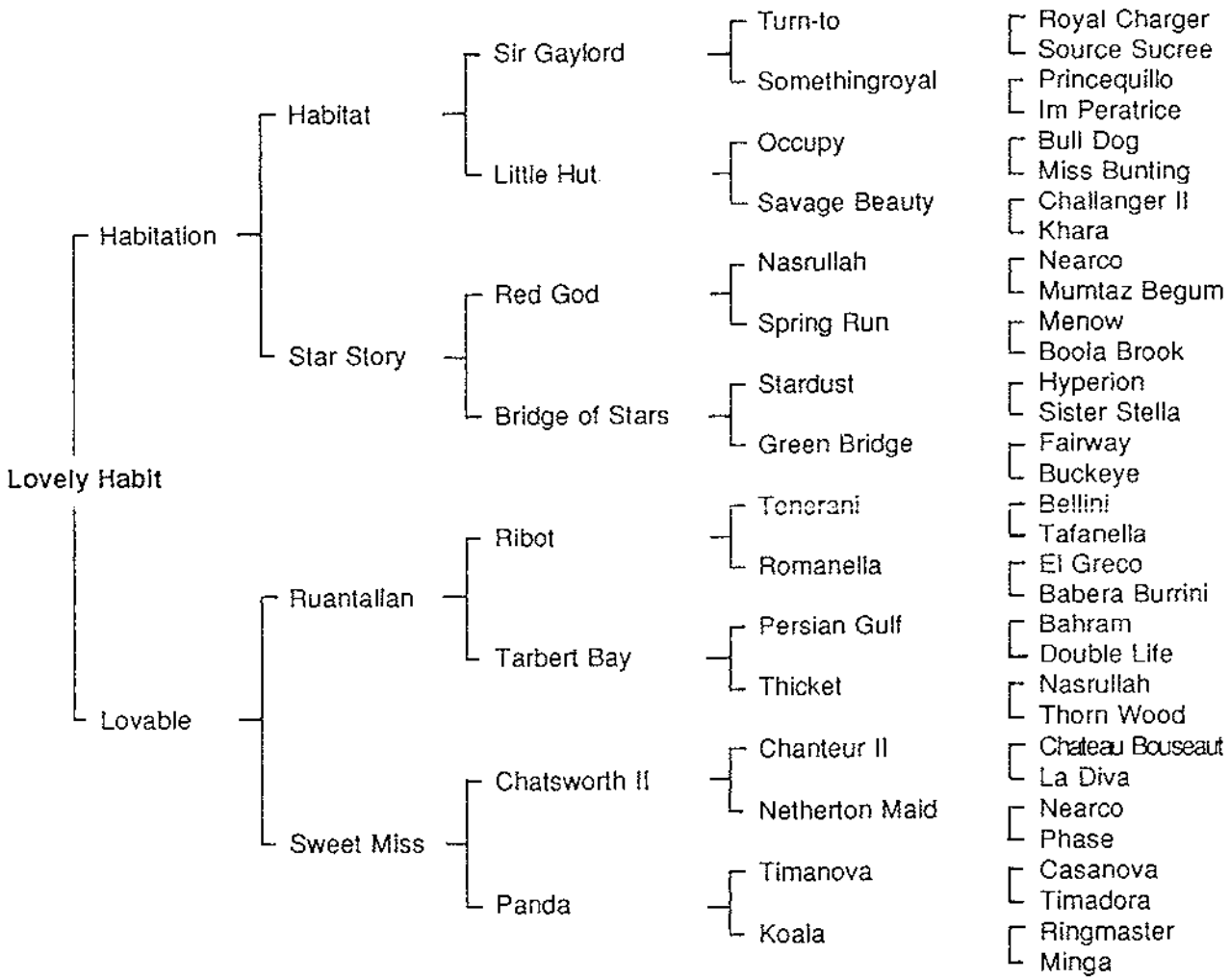
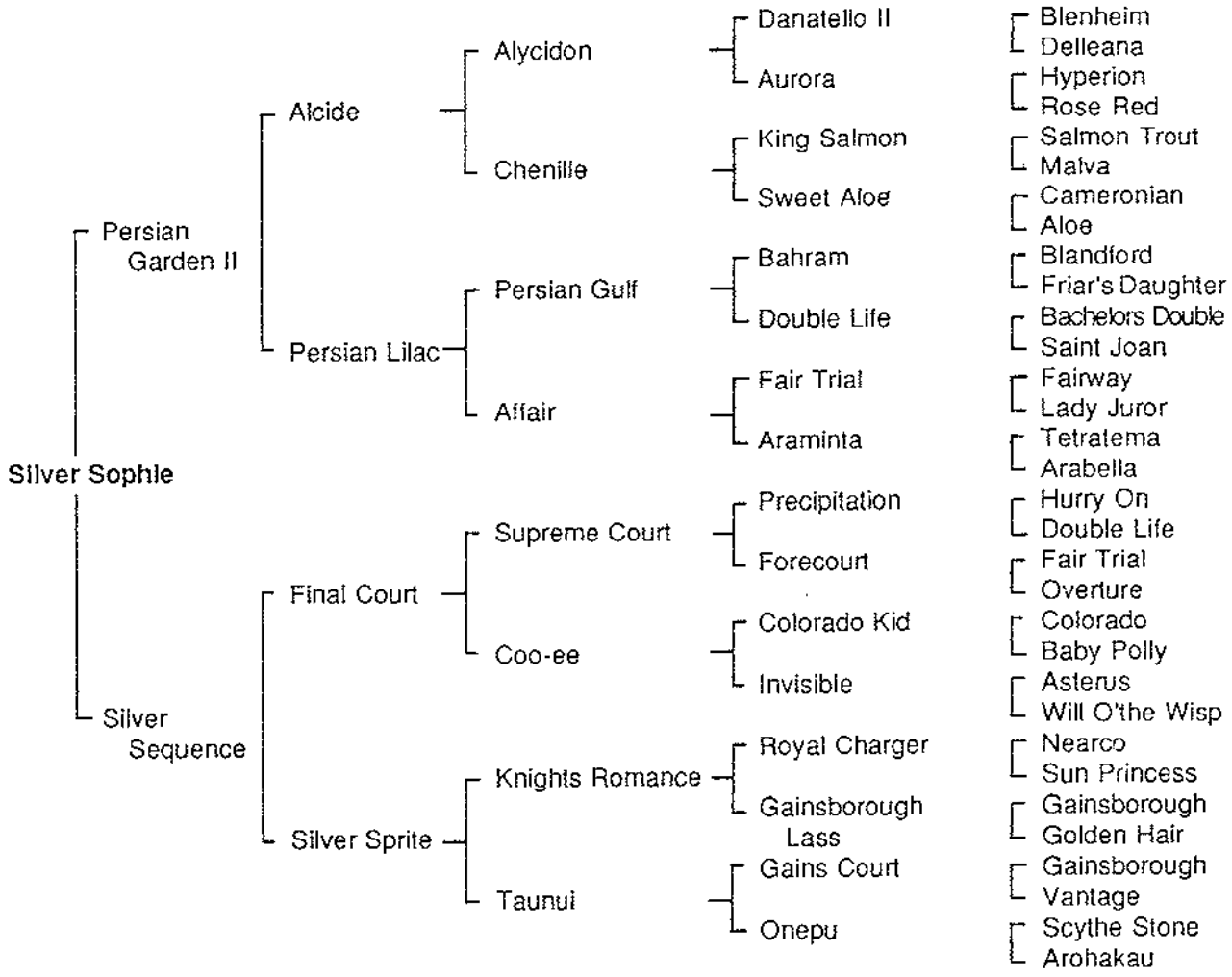


Table 2.2f Pedigree of Silver Sophie (mare)



All solutions were stored and added at 0°C. Reactions contained 4 pmol JF-1 (100ng/µl H_2O), 1x kinase buffer, 1µl0.1M dithiothreitol, 1µl 0.033M spermidine, 3µl 1mg/ml bovine serum albumin, 1µl 10mCi/ml [γ-32P]ATP. This was mixed, centrifuged briefly and 0.5µl removed for thin layer chromatography (TLC). Seven units of T4 polynucleotide kinase were added, mixed and centrifuged briefly and incubated for one hour at 37°C. The reaction mixture was transferred to ice and a further sample removed for TLC. If labelling was complete the reaction was stopped with 1x stop solution (125 mM EDTA, 5% SDS (w/v)). Probe could be stored for up to two week at -20°C.

2.2.2 PEI Cellulose Thin Layer Chromatography (TLC)

This was used to check the end-labelling reaction. Samples removed from the labelling reaction were spotted about 5mm above the lower edge onto a strip (about 40 x 80mm) of PEI cellulose (Schleicher and Schull) adjacent to each other. The samples were air dried and then chromatographed for 10-15 minutes in a foil-covered beaker containing approximately 2mm of 0.75M KH_2PO_4 adjusted to pH 3.5 with orthophosphoric acid. The chromatogram was wrapped in glad wrap and autoradiographed for 1-2 minutes.

Fig 2.1a shows a typical TLC result, with complete labelling occurring after 60 minutes incubation. Very little unincorporated gamma ^{32}P-ATP or ^{32}P-phosphate was detected.

2.2.3 Checking the probe size

Fig 2.1b shows JF-1 had retarded mobility compared to JF-8 (an 18-mer) verifying it was slightly larger and was of the size stated by the manufacturer (Otago University, section 2.2). A discrete band was present for each probe indicating neither had degraded.

Fig 2.1a **Testing P^{32} End-Labelling of JF-1**

The PEI cellulose was chromatographed for 15 min and autoradiogrpahed for 2 minutes.

t=60 minutes, complete labelling, spot represents labelled JF-1.
t=0 minutes, initial labelling, the large spot represents unincorporated ^{32}P-ATP

Fig 2.2b **Confirmation of JF-1's size by polyacrylamide minigel electrophoresis**

Approximately 3ng of end-labelled probe was electrophoresed for 1 hour along a 12% gel. The gel was autoradiographed overnight.

JF-1, 20mer
JF-2, 18mer


A
1
2
B
1
2

2.3 MANIPULATION OF GENOMIC DNA

2.3.1 DNA Extraction

Blood samples were collected by venipuncture into heparinised tubes and stored at 4^{o}C until extracted. Leucocytes were harvested after lysis of erythrocytes in 0.16M ammonium chloride 0.16M Tris.Cl buffer, pH7.5 (90:10, v/v) according to standard procedures (Mishell and Shiigi, 1980). DNA was extracted by the phenol-RNase-dialysis procedure of Maniatis *et al* (1982). The concentration of the DNA was determined spectrophotometrically and by short duration gel electrophoresis with lambda DNA standards in a 1% agarose TBE minigel as outlined by Maniatis *et al* (1982). Quality of the extracted DNA was assessed by measuring the OD 260/280 ratio (to check for phenol or protein contamination) following Maniatis *et al* (1982), and electrophoresis alongside high molecular weight lambda DNA (to check for DNA degradation). The DNA was stored at 4^{o}C dissolved in TE buffer.

Table 2.3 shows the spectrophotometric results of the DNA extracted from the Sovereign Lodge Stud horses. The OD 260/280 ratio was lower than the desired 1.8-2.0; dialysis a second time had no effect on this ratio. Electrophoresis showed the DNA to be high molecular weight, with little evidence of degradation and was therefore used in this investigation. Subsequently, it showed unsatisfactory digestion with restriction enzymes, but after repeated proteinase K, phenol and phenol-chloroform extraction the DNA gave similar results as the control DNA. The control DNA had been prepared from blood samples of horses in the Massey University herd, by Dr J W Forrest.

2.3.2 Concentration of DNA

DNA was concentrated by ethanol precipitation and butanol extraction according to Maniatis *et al* (1982).

Sample	A_{260}	A_{280}	[DNA]ng/μl	260/280
TE (blank)	0	0	0	-
Kingdom Bay	0.196	0.239	9.8	1.51
Latchmi	0.222	0.255	11.1	1.53
Filly (KB-L)	0.118	0.202	5.9	1.34
Darling Daughter	0.356	0.348	17.8	1.51
Filly (KB-DD)	0.311	0.316	15.5	1.51
Western Bay	0.271	0.297	13.5	1.46
Lovely Habit	0.251	0.280	12.6	1.49
Colt (WB-LH)	0.453	0.414	22.7	1.50
Silver Sophie	0.362	0.348	18.1	1.50
Colt (WB-SS)	0.303	0.323	15.2	1.40

Table 2.3 Horse DNA Extraction Results

2.3.3 Restriction Endonuclease Digestion of Genomic DNA

A variety of enzymes (see table 2.4) were used to digest genomic DNA. These were chosen based upon reports by other researchers (eg Ali *et al*, 1986) as well as on the collective experience of co-workers in our laboratory. Usually 5µg of genomic DNA was digested with about 20 units of enzyme in the appropriate restriction endonuclease reaction buffer and at the appropriate temperature overnight. For clearly distinguishable polymorphic bands, genomic DNA had to be digested to completion. Digestion was checked by electrophoresis along a minigel using about 200ng DNA. The presence of high molecular weight DNA indicated incomplete digestion, but complete digestion could not always be ascertained on a minigel. Partially digested DNA often spooled out of its own accord from the agarose gel wells, making quantitative comparison of adjacent lanes difficult.

2.3.3.1 Restriction endonuclease reaction buffers were made following the protocol of Maniatis *et al*, 1982. Low, medium and high salt stock solutions were stored at -20^{o}C in 100µl aliquots. Aliquots were thawed to room temperature before use, and then kept at 4^{o}C for one week before they were discarded and a new aliquot started.

2.4 ELECTROPHORESIS, BLOTTING AND HYBRIDIZATION

2.4.1 Horizontal Agarose Gel Electrophoresis

Two gel sizes were used throughout this project: minigels (5 x 80 x 120 mm) and large gels (5 x 200 x 220 mm). The gels were made at various concentrations of agarose (BRL ultrapure and Sigma low EEO agarose) ranging from 0.5-2% (w/v). For both gels 1x TBE electrophoresis buffer was used. Minigels were electrophoresed at 100V for 1-2h with 1µg/ml of ethidium bromide (10mg/ml) included in the buffer.

Enzyme	Restriction Site	Source
Acc I	AG/CT	BRL
Alu I	AG/CT	BRL
Bam HI	G/GATCC	Amersham
Eco RI	G/AAATTC	Boehringer
Hae III	GG/CC	Amersham
Hinc II	GT(C,T)/(G,A)AC	Pharmacia
Hind III	/AGCTT	BRL
Hinf I	G/ANTC	Biolabs
Kpn I	GGTAC/C	Biolabs
Mbo I	/GATC	Pharmacia
Pst I	CTGCA/G	Boehringer
Sal I	G/TCGAC	Boehringer
Sau 3A	/GATC	Amersham
Sam I	CCC/GGG	BRL
Taq I	T/CGA	BRL
Xba I	T/CTAGA	Biolabs

Table 2.4 Restriction Enzymes used in this Project

Large gels electrophoresed at 80V for 16h and stained afterwards in 1L 1x TBE containing 0.5µg/ml (v/v) ethidium bromide (10mg/ml). In all gels, BRL low molecular weight DNA standards or lambda-*Hind* III DNA were used as size markers.

Bromophenol Blue Loading Dye (10x) was made as according to Maniatis *et al* (1982). This was added at 1x to the DNA samples to be electrophoresed just before they were loaded. In the case of genomic DNA samples, the DNA-loading dye mixture was heated for 5 minutes at 70^{o}C then chilled on ice before loading. This encouraged the DNA to remain at the bottom of the well and not float away.

2.4.2 Polyacrylamide Gels

2.4.2.1 Minigels

A 12% polyacrylamide gel was made with 3ml 20x stock solution Bis acrylamide (as according to Maniatis *et al*, 1982), 2ml of 5x TBE and 5ml water. To the gel solution 100µl of freshly prepared 10% (w/v) ammonium persulphate and 6µl of TEMED was added. The gel was poured between two thin glass plates of the Biorad minigel apparatus.

2.4.2.2 Polyacrylamide/Urea Sequencing Gels

A 6% polyacrylamide-6M urea gel was prepared from 15ml of a 40% stock acrylamide solution (made as according to Maniatis *et al*, 1982), 42g urea, 10ml 10x TBE and the volume made up to 100ml with water. The urea was dissolved by gentle stirring and the solution filtered (0.4µm) and degassed on a vacuum pump. The solution was placed on ice for several minutes.

Clean glass plates were siliconised to facilitate the removal of the gel after electrophoresis and prevent unwanted bubble formation during pouring.

To 100ml of gel solution, 50µl of TEMED and 1ml of freshly prepared 10% (w/v) ammonium persulphate was added. The plates were poured on a 30^{o} angle and allowed to set for at least one hour (preferably overnight) before use.

2.4.2.3 Electrophoresis and autoradiography of sequencing gels

The microfuge tubes containing the DNA were heated at 75-80^{o}C for 2 min. Gel wells (created by "sharks tooth" combs) were flushed out with 1x TBE electrophoresis buffer immediately before loading. This removes any urea that may have leached out of the gel and disrupt the thin layering of the sample in the well. Three microlitre samples were loaded, at timed intervals such that the samples ran for a total of 9-, 7-, 4-, and 2 hours respectively. Not all these time intervals were used on every gel. Electrophoresis was begun at 50mA with voltage limited to 1500V. The current dropped to about 20mA towards the completion of the run.

At the end of the electrophoresis the siliconised plate was removed and the gel transferred to Whatman #1 paper, covered with Gladwrap, and dried on a vacuum gel drier at 80^{o}C for 45 minutes. The gel was not fixed. After drying, the Gladwrap was removed and X-ray film placed directly onto the the gel. Film was exposed for 36 hours at -20^{o}C and an intensifying screen was required to enhance the signal.

2.4.3 Southern Blotting

Capillary blotting was used to transfer DNA from agarose gels to nylon membrane (BioTrace RP, Gelman Sciences Inc, 0.45µm pore size), following the methods of Reed and Mann (1985) and Rigaud *et al* (1987).

The DNA was denatured to its single stranded form by gently agitating the gel in a shallow tray containing 1L of 0.4M NaOH/1.5M NaCl for 20 minutes. The gel was rinsed in deionised water before being placed in 1L 0.02M NaOH/1.5M NaCl (transfer solution) for 20 minutes with shaking to equilibrate.

The transfer apparatus was set up according to fig 2.2.

The membrane, previously wetted with boiling hot deionised water, was placed on top of the gel, rolled gently with a glass rod to remove air bubbles, the three pieces of 3 MM paper were "just-wet" with transfer solution (0.02M NaOH/1.5M NaCl) and applied in the same manner.

The paper towels were changed twice, after about one-half and two hours, transfer solution added to the reservoir as required. After overnight transfer, the DNA was fixed to the nylon membrane by placing it DNA side up on a pad saturated with 0.4M NaOH for 10 minutes. The blot was rinsed in 2x SSC for 10 minutes, lightly blotted dry then stored between two pieces of blotting paper at room temperature. The gel was stained in 1L 1x TBE containing 50µl ethidium bromide (10mg/ml) overnight and then examined using the UV transilluminator for evidence of any non transferred DNA.

Fig 2.3 demonstrates the efficiency of Southern blotting by following the transfer of radioactively labelled DNA. The blot emitted a strong signal and the gel emitted none, indicating efficient transfer of DNA from the gel to the membrane.

2.4.4 Alkaline DNA Dot Blotting

This was used for quantitative measurements of DNA hybridization and for colony screening. The BioRad Dot Blot apparatus was used according to manufacturers instructions using BioTrace nylon membrane.

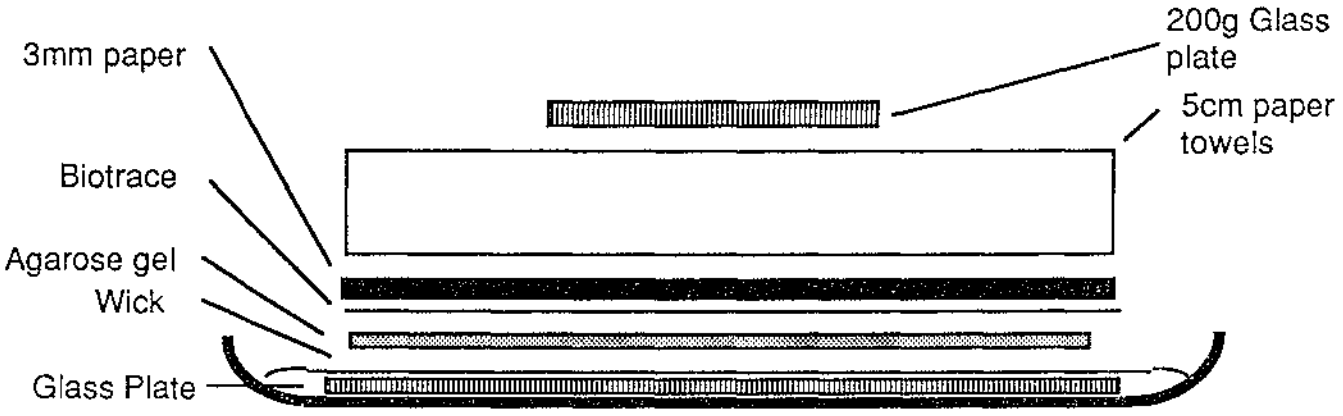


Fig 2.2 Southern Blot Apparatus

Fig 2.3 **Efficiency of Southern Blotting**

Approximately 10ng of ^{32}P labellled pUC, digested with *Taq* I was electrophoresed in a large gel then blotted overnight. Both gel and membrane were autoradiographed 1 day.

1 Membrane after DNA transfer

2 Gel after DNA transfer

1 2

2.4.5 Hybridization of JF-1 to genomic DNA

2.4.5.1 Calculation of Hybridization Temperature

In 1M Na^+ solution, the temperature at which hybridization occurs when there is 100% homology between the oligonucleotide probe and genomic DNA can be calculated approximately by the following equation from (Berent *et al*, 1985):

$$T_h = 2 \times (\text{no. of A.T bp}) + 4 \times (\text{no. of G.C bp}) - 5$$

Therefore, for JF-1 (ie $(GATA)_5$)

$$T_h = 2 \times (15) + 4 \times (5) - 5$$
$$= 45$$

2.4.5.2 Hybridization of Probe to Membrane Bound DNA

The hybridization solution consisted of 5xSSPE, 7% (w/v) SDS, 0.5% (w/v) Blotto and 1% (w/v) PEG 6000 based on that of Reed and Mann (1985).

The blot was prehybridized at the hybridization temperature for 2 hours in hybridization solution (10ml/10cm^2 membrane) before hybridization was carried out at 45^{0}C overnight in the same hybridization solution containing 100ng ^{32}P end-labelled JF-1. At this temperature it was expected that hybridization would occur only if there was 100% homology between the JF-1 probe and the genomic DNA.

On completion of hybridization the blot was washed in 6x SSC which has a 1M Na^+ concentration, at 45^{0}C for 15 minutes with vigorous agitation, lightly blotted dry, wrapped in glad wrap before being autoradiographed 3-7 days on Fuji RX double sided X-Ray film, with intensifying screens (DuPont Cronex "Lightning Plus") at -20^{0}C.

2.5 GROWTH OF *E. coli*

DH5α (BRL) and JM109 strains of *E. coli* were grown at 37°C in liquid or solid media and for DH5α, ampicillin (50μg/ml) was included in the media.

2.5.1 Maintenance of cultures

DH5α was maintained on LB plates containing 50μg/ml ampicillin, JM109 was maintained on minimal media plates. Both strains were kept at 4°C and were subcultured every 4-6 weeks. For long term preservation the bacteria were kept in a liquid media of LB containing 50% glycerol and stored at -20°C.

2.5.2 Media and Solutions

Ampicillin (Sigma) was made to 10mg/ml in 50% ethanol (v/v), stored at -20°C.

Luria Broth and agar were made as according to Maniatis *et al*, 1982.

Minimal media agar was made in two stock solutions: solution A contained 1.7g K_2HPO_4, 0.9g KH_2PO_4, 0.2g $(NH_4)_2SO_4$, 0.1g *tri* sodium citrate and deionised water to a total volume of 38ml, then autoclaved. Solution B contained 3g of agar dissolved in 160ml H_2O, autoclaved then cooled to 45°C and 0.2ml 20% $MgSO_4.7H_2O$ (w/v), 0.1ml 1% thiamine (w/v) and 2ml 20% glucose (w/v) added. Solution A was added to solution B, mixed and poured into ten 8.5cm petri dishes.

Terrific Broth was made in two stock solution: K stock contained 4.6g KH_2PO_4, 25.1g K_2HPO_4 made up to 200ml with deionised water and autoclaved in 25ml aliquots. T stock contained 12g bactotryptone, 24g yeast extract, 4g glycerol, made up to 900ml with deionised water and sterilized by autoclaving. Before use T stock and K stock were mixed in a ratio of 9:1 (v/v).

2.6 CLONING A GENOMIC GATA FRAGMENT

Two genomic horse DNA fractions (section 2.6.1) were blunt-end ligated in pUC (section 2.6.2), which had been digested with *Sma* I and *Hinc* II in the multiple cloning site. The plasmid was transformed into DH5α (sections 2.6.3 and 2.6.4) and single colony transformants were picked into both liquid media (in 96 well microtitre plates) and onto solid LB amp plates, for future reference.

Colonies from the microtitre plate were dot blotted onto nylon membrane (section 2.4.4). Control colonies included DH5α (without plasmid), DH5α with pUC (no insert), DH5α with ligated pUC but a blue colony indicating unsuccessful ligation. Also included on the blot was 100ng of horse DNA and 100ng of human DNA. The blot was hybridized with JF-1 to detect GATA positive clones.

2.6.1 Isolation of genomic DNA fragments

Genomic horse DNA was digested with *Hae* III and fractions were isolated along a salt gradient (by Dr J Forrest, DSIR, Biotechnology). Two fractions, containing approximately 3kb and 1kb fragments respectively, were electrophoresed in a minigel, blotted and hybridized with JF-1. Both fractions showed the presence of GATA (fig 2.4).

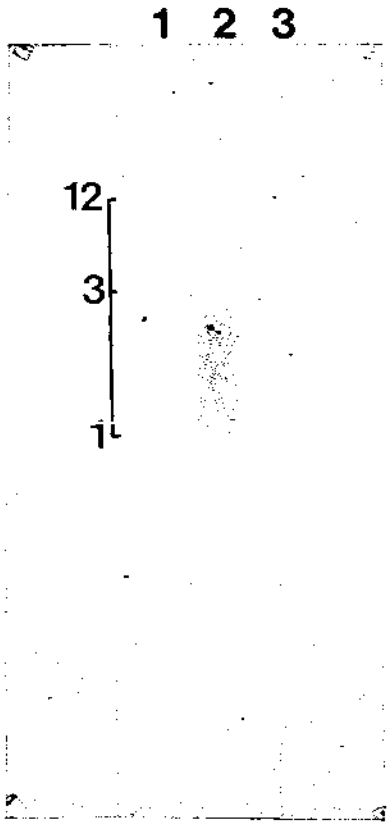
2.6.2 Ligation of insert DNA into vector DNA

The vectors used were the plasmid pUC 18 (Yanish-Perron *et al*, 1985) and the bacteriophages M13 *mp18* and *mp19* (Messing, 1983). From the method of Maniatis *et al* (1982) ligations had a 3:1 molar ratio of insert to vector in a total volume of 20µl. The vector was treated with 0.01U calf intestinal phosphatase (CIP, BRL) to dephosphorylate the 5' end to prevent self ligation of the vector. Ligations were carried out overnight at room temperature in ligation buffer, 1mM rATP, 1U T4 DNA ligase (Boehringer).

Fig 2.4 **Screening Genomic Horse Fractions 31 and 37 for the Presence of $(GATA)_5$ Sequences**

These fractions were electrophoresed on a 0.8% TBE agarose minigel, Southern blotted and hybridized with JF-1 at 45°C and washed at 42°C in 6x SSC.

1 Fraction 31 (containing approx. 3kb fragments)
2 Fraction 37 (containing approx. 1kb fragments)
3 BRL low molecular weight DNA standard


1 2 3
12
3
1

2.6.3 Competent Cell Preparation of DH5α

From an overnight "starter" culture of DH5α a 1:100 dilution was made in 50ml LB, and incubated at 37°C until an OD_{600} of about 0.5 was reached (this usually took just over 2 hours). The culture was placed on ice for 30 minutes then centrifuged (2000*g*, 2 minutes). All subsequent steps were performed at 4°C. The supernatant was discarded and the pellet resuspended in about 2ml of 0.1M $MgCl_2$: further $MgCl_2$ was added to make up a final volume of 50ml. The cells were then centrifuged (3000*g* for 2 minutes), the supernatant discarded and the pellet resuspended in 2.5ml 50mM $CaCl_2$. The cells were placed on ice 1-2 hours prior to using for transformation. Increased efficiency of transformation was found by leaving the cells overnight on ice before use.

2.6.4 Transformation of Competent Cells

To 10ng of plasmid DNA 100µl of competent cells (see the above method) were added, mixed gently then chilled on ice for 30 minutes. The cells were heat-shocked for 1 minute at 42°C, chilled on ice for 2 minutes, 0.5ml of LB added and the cells incubated at 37°C for 1 hour. For blue/white selection, 20µl of IPTG (25mg/ml) and 20µl of X-gal (25mg/ml) was added to the transformed cells. A dilution series could be made in LB before spread plating the transformants onto Luria agar plates containing 50µg/ml ampicillin. The plates were incubated overnight at 37°C, examined and the transformation efficiency determined. The plates could be stored at 4°C.

2.6.5 Preparation of Plasmid DNA

2.6.5.1 Minipreparation

The method of Ish-Horowicz and Burke (1981) was modified as follows:

(i) a 10ml culture was grown to late log phase and harvested by

centrifugation (1000*g*, 7 min, 4°C), the supernatant was removed by aspiration leaving the bacterial pellet dry as possible,
(ii) 10µl of freshly prepared lysozyme solution (50mg/ml solution I) was added after the addition of solution I,
(iii) the pellet was washed with 1ml 70% ethanol (v/v) and centrifuged 5 minutes at 12000*g*. The pellet dried *in vacuo* for 10-20 minutes then resuspended in 30µl TE (pH 8.0).

2.6.5.2 Large-scale Plasmid Preparation

This method based on that of Maniatis *et al* (1982) is similar to the minipreparation except that the volumes used and the incubation times have been increased.

A 500ml culture was harvested by centrifugation (6000*g*, 10min at 4°C), the pellet resuspended in 15ml solution I and 1.5ml lysozyme solution. The mixture was incubated at room temperature for 10 minutes, 30ml solution II added, incubated on ice 10 minutes, then 30ml solution III added, mixed by swirling, and incubated on ice for 10 minutes followed by centrifugation (8000*g*, 20 minutes, 4°C).

The DNA was precipitated by isopropanol, placed at -20°C for 1 hour then centrifuged (12000*g*, 20 minutes at 4°C). The pellet was washed twice with about 40ml 70% ethanol and centrifuged (12000*g*, 10 minutes, at 4°C). The supernatant was removed by aspiration and the pellet dried *in vacuo* for about 20 minutes. It was important not to over-dry the pellet otherwise it became difficult to redissolve in 5ml of TE. The solution was centrifuged (7000*g*, 10min) and the supernatant transferred to a fresh tube. To check the yield of DNA about 0.5µl was electrophoresed in a minigel before progressing further.

Plasmid DNA was separated from bacterial chromosome DNA in a CsCl gradient. The CsCl and ethidium bromide were added, the DNA centrifuged (8000*g*, 20 min, 4) and the supernatant transferred. If the supernatant was not clear, then it was

centrifuged again. The density was checked by weighing a 100µl aliquot: a density of 1.55g/ml was required. The DNA was transferred to a 5ml Beckman ultracentrifuge tube and centrifuged (16 hours at 140000 *g*, 15°C) and then visualized by fluorescence with UV light (366nm) and the plasmid DNA removed using a 18 gauge needle.

Ethidium bromide was extracted from the DNA by inversion with an equal volume of salt buffered isopropanol (to 100ml isopropanol 100ml TE and 2.5 M NaCl was added, stirred overnight).

CsCl was removed from the DNA by dialysis against 2L TES buffer with 4 changes over several hours. The DNA was precipitated in 0.5 volumes ammonium acetate (7.5M) and 2 volumes absolute ethanol at -20°C overnight. The DNA was centrifuged (12000*g*, 30 min, 4°C), the pellet dried *in vacuo* and resuspended in 500µl TE. The DNA concentration was measured spectrophotometrically and on a minigel.

2.7 RESTRICTION MAPPING, SEQUENCING AND ANALYSIS

2.7.1 Preparation of DNA by Electroelution

This method from Maniatis *et al* (1982) was used to isolate a DNA insert from a recombinant vector. About 20µg of the plasmid containing the required insert was digested such that two discrete pieces of DNA resulted representing the vector and insert. An aliquot was first checked on minigel before the total digest was electrophoresed on a 0.8% SeaPlaque or SeaKem GTG agarose large gel at 50V overnight in 1x TBE. The gel was stained in ethidium bromide and the DNA visualized by fluorescence using the UV transilluminator. Using a sterile scalpel blade, a slice of the gel was cut out containing the appropriate piece of DNA. This was placed in a dialysis bag containing 0.5x TBE and then submerged in an electrophoresis tank containing 0.5x TBE and electroeluted following the method of Maniatis *et al* (1982). The DNA was ethanol precipitated and redissolved in 50µl TE.

2.7.2 Restriction Mapping of a GATA Fragment

A variety of enzymes were used to digest insert DNA. Usually 200ng of DNA was digested with 5-10U of enzyme in the appropriate restriction enzyme buffer (subsection 2.3.3.1) for 1-2 hours. Digestion was examined by electrophoresis in a minigel. BRL low molecular weight standard DNA was used as a standard to determine the molecular weight of DNA fragments. The relative mobilities of DNA fragments were measured and molecular weights determined graphically from a plot of relative mobility against $\log_{10}$ molecular weight (Sanger *et al*, 1982).

2.7.3 Subcloning

The DNA to be subcloned was obtained by digestion with appropriate restriction enzymes (section 2.3) followed by electroelution (section 2.7.1). The DNA was ligated into the M13 vector following standard procedures (section 2.6.2).

2.7.4 Competent Cell Preparation for Sequencing

This method was used for bacterial cells which were to be transformed with M13. These cells had to be initially cultured on minimal media plates to select for the F' plasmid. This is necessary for the production of the F pilus through which the M13 bacteriophage enters and subsequently is released from the bacterial cells.

An overnight starter culture of JM109 was made in LB. A 1:100 dilution of this was made in 20ml LB which was incubated for 2-3 hours until the OD_{600} was between 0.4 and 0.6. The culture was dispensed as 1.5ml aliquots into precooled eppendorf tubes and chilled on ice for 5 minutes. These were centrifuged for 1 minute, the supernatant discarded and the pellet resuspended in 1ml chilled 0.1M $MgCl_2$. The cells were chilled on ice 10 minutes before being centrifuged for 1 minute. The supernatant was discarded and the pellet resuspended in 1ml chilled 0.1M $CaCl_2$ and returned to ice for 30 minutes. The cells were centrifuged for 1

minute, the supernatant discarded and the pellet resuspended in 200µl chilled 0.1M $CaCl_2$. The cells could be kept for 24 hours when stored on ice.

2.7.5 Transformation of competent cells for sequencing

Lawn cells were prepared by inoculating 5ml LB with a single colony and grown with shaking at 37°C until they reached late log phase (3-4 hours).

Competent cells (prepared by method 2.7.4) were transformed with 10µl of ligated M13 (section 2.6.2), mixed briefly then placed on ice for 30 minutes. The transformation mix was heat-shocked at 42°C for 2 minutes then placed on ice. For each transformation 3 tubes were set up each containing 100µl fresh JM109 lawn cells, 10µl 0.1M IPTG, 20µl X-Gal (25mg/ml), 2.5ml LB top agar and 2, 20 or 200µl of the transformation mix. These were mixed briefly and spread quickly onto LB plates and allowed to set before being incubated at 37°C overnight.

2.7.6 Preparation of Single Stranded M13 DNA

Clear plaques were picked using a Gilson pipette tip and inoculated into 1.3ml Terrific broth (section 2.5.2) containing 20µl of an overnight culture of JM109. The cultures were grown at 37°C with vigorous shaking for 4-6 hours (maximum) and then centrifuged for 5 minutes. About 1ml of the supernatant containing the bacteriophage was transferred to eppendorf tubes, 200µl 20% PEG 8000-2.5M NaCl was added and the bacteriophage was precipitated at room temperature for 20 minutes or overnight at 4°C. The cells were resuspended in 100µl supernatant, 500µl 50% glycerol added and then stored at -20°C for future reference.

The bacteriophage was pelleted at 12000*g* for 10 minutes, the supernatant discarded. It was centrifuged a further 2 minutes and the remaining supernatant

removed with a drawn-out pasteur pipette. The pellet was resuspended in 100µl TE, then extracted with 50µl phenol-chloroform to removed the protein coat from the bacteriophage. The mixture was vortexed for 15 seconds, the phases allowed to separate for 15 minutes and then vortexed for a further 15 seconds before being centrifuged for 2 minutes and the aqueous phase transferred to a fresh eppendorf tube. Another 50µl phenol-chloroform was added and the DNA re-extracted. Remaining phenol was removed by the addition of 50µl chloroform-isoamyl alcohol 24:1 (v/v), vortexed for 10 seconds and centrifuged for 2 minutes, the aqueous phase transferred to a new tube. The DNA was ethanol precipitated at -20°C overnight, pelleted at 12000*g* for 20 minutes and the supernatant discarded. The pellet washed with 200µl 70% ethanol, vortexed briefly then centrifuged at 12000*g* for 20 minutes. The supernatant was discarded and the DNA resuspended in 20µl TE. To determine the yield of DNA, a 2µl aliquot was electrophoresed in a minigel.

2.7.7 Sequencing and Computer-Based Sequence analysis

Sequencing was carried out using the Sanger dideoxy sequencing method, following the protocol in the Sequenase kit (USB) with no modifications. Polyacrylamide/urea sequencing gels were prepared as according to section 2.4.2.2. Gels were electrophoresed and autoradiographed as according to section 2.4.2.3.

The sequence was analysed using the following programs (supplied by Dr P Stockwell, Biochemistry Department, University of Otago, New Zealand, 1985):

NBATIN, a sequence gel entry system, version 3.0
VTUTIN, a gel management editing package, version 1.0
RANS, a rapid nucleotide sequencing program
FASTN, an EMBL database search and homology identification program.

The sequencing programs were run on a microvax 3400. Hardcopy of printouts were obtained from a dot-matrix printer queued to the VAX.

3.0 RESULTS

3.1 QUANTIFICATION OF GATA IN THE HORSE GENOME

3.1.1 Are GATA sequences present in domestic animal genomes?

The amount of probe which binds to DNA fixed onto a membrane depends on the temperatures at which hybridization occurs and that which the blot is washed at, and the salt concentration of the hybridization and wash solutions, as well as the number of sequences complementary to the probe and the extent of their complementarity or homology. These variables had to be tested before the number of GATA repeats in the genome could be determined. This testing was initially performed on dot blots.

A set of four dot blots (section 2.4.4) containing known amounts of genomic DNA from various species including horse, were hybridized at 35°C. At this temperature, hybridization theoretically occurred if there was 80% or more homology between the probe and the DNA (section 2.4.5). The dot blots were washed sequentially in decreasing salt concentration and increasing temperature which resulted in an overall increase in stringency. Fig 3.1 shows that as the stringency increased, the intensity of the hybridization signal decreased. However, even at high stringency there was significant association between the probe and genomic DNA reflecting a strong level of homology and/or a relatively high number of GATA repeats in the horse genome. This showed that GATA sequences were present in the horse genome, at about the same level as in the human genome.

A slightly different set of dot blots containing horse, human, cow, goat and sheep DNA were used as positive controls in all subsequent hybridizations involving JF-1. When the probe was hybridized to these dot blots and washed at high stringency

(45^{o}C, 6x SSC), there was a noticeable difference in the amount of hybridization between the various species (fig 3.2). The horse, human and goat DNA were strongly hybridizing, but relatively little hybridization occurred to sheep and cow DNA. Negligible hybridization occurred to the negative control of pUC , M13 and lambda DNA.

3.1.2 Is there a sex-specific quantitative difference in horse?

A dot blot series was created which contained 100ng amounts of genomic DNA from male and female horses, as well as from the offspring of some of these horses. The quantity of the DNA was first verified by electrophoresing 100ng samples along a TBE agarose minigel. Fig 3.3 shows that the quantity of DNA for each sample was the same. The probe was hybridized to duplicate dot blots using standard conditions (section 2.4.5). Fig 3.4 shows there was no noticeable difference in hybridization intensity between the sexes suggesting that there was not quantitative sex-difference in GATA repeats in the horse.

3.1.3 How much GATA present in horse?

The above dot blot (fig 3.4) also included a known amount of an oligomer complementary to JF-1. By comparing the intensity of signal from this dot to those of the horse DNA dots, the amount of GATA present in horse could be estimated. The assumptions made were firstly, a similar signal of complementary oligomer and genomic DNA meant the same number of repeats were present in each. Secondly, JF-1 was present in excess and finally, that hybridization occurred only with sequences that were 100% homologous.

A similar signal appeared for 1ng of complementary JF-1 as for 100ng of genomic horse DNA, therefore the number of $(GATA)_5$ repeats in the horse can be calculated as follows:

Fig 3.1 **Determination of Optimal Hybridization conditions for JF-1.**

Dot blots of genomic DNA were hybridized with JF-1 at 35°C, prewashed in 6x SSC at room temperature before being washed at increasing stringency.

Dots represent hybridization to DNA of:

1 Blank (negative control)
2 pUC 18, M13 mp18, λ, 100ng each
3 Horse 100ng
4 Horse 10ng
5 Human 100ng
6 Human 10ng
7 Sheep 100ng
8 Cattle 100ng

Dot blots were washed at increasing salt concentrations and temperatures which gave an overall increase in stringency:

(a) 3x SSC, 0.1% SDS at room temperature
(b) 3x SSC, 0.1% SDS at 35°C
(c) 1x SSC, 0.1% SDS at 35°C
(d) 1x SSC, 0.1% SDS at 45°C

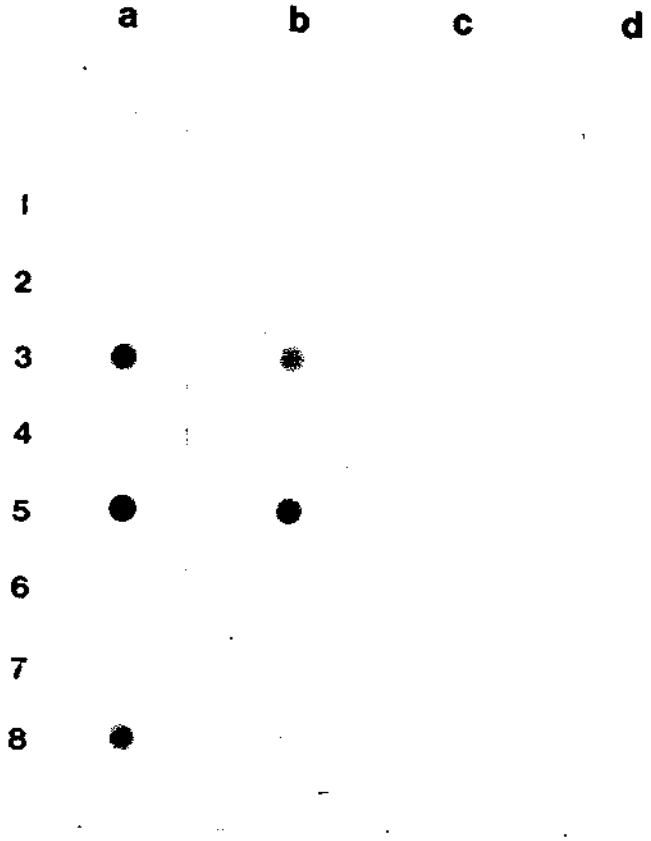

a
b
c
d
1
2
3
4
5
6
7
8

Fig 3.2 **Hybridization positive control.**

DNA dot blot hybridized and washed at high stringency conditions of 45°C hybridization and washed at 45°C in 6x SSC for 15 minutes. An identical set of dot blots was included in all subsequent hybridization as a positive control.

1 pUC, M13 mp18, λ, 100ng each
2 Horse 100ng
3 Horse 10ng
4 Human 100ng
5 Human 10ng
6 Cattle 100ng
7 Sheep 100ng
8 Goat 100ng


1
2
3
4
5
6
7
8

Fig 3.3 Confirmation of DNA Quantity by gel electrophoresis

A 0.8% TBE agarose minigel showing 100ng samples of genomic horse DNA. Gel was electrophoresed for 10 minutes.

1 Low molecular weight marker DNA
2 Kingdom Bay
3 Latchmi
4 Filly (Kingdom Bay x Latchmi)
5 Filly (Kingdom Bay x Darling Daughter)
6 Western Bay
7 Lovely Habit
8 Colt (Western Bay x Lovely Habit)
9 Colt (Western Bay x Silver Sophie)

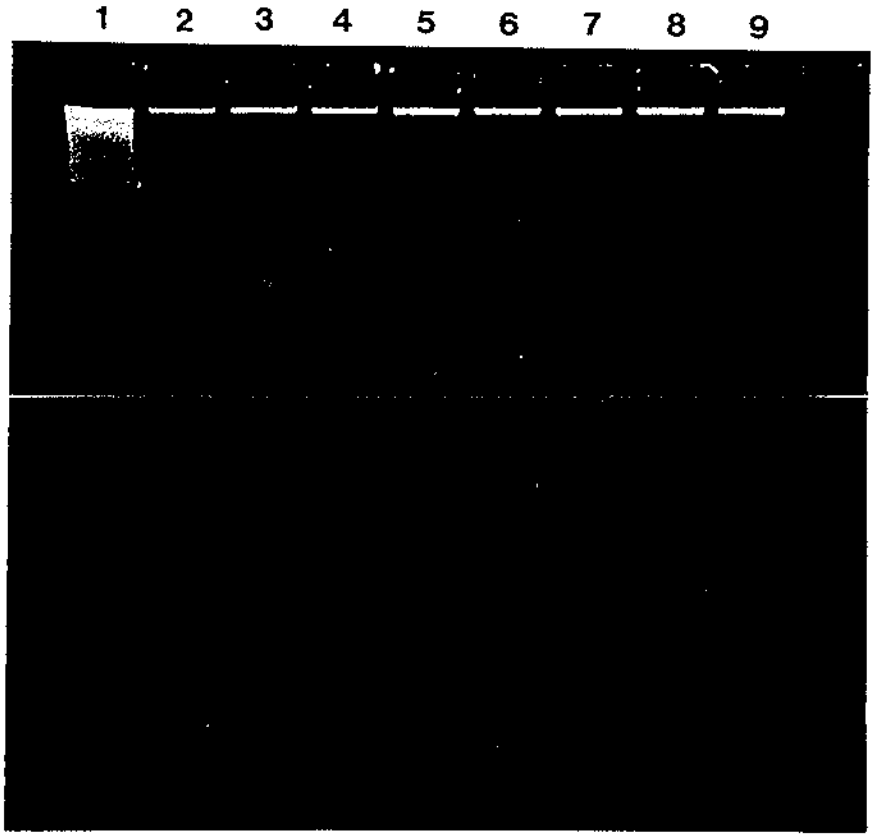

1
2
3
4
5
6
7
8
9

Fig 3.4 **Quantitative comparison of male and female horse DNA with respect of GATA sequence content**

Dot blot was hybridized and washed at high stringency conditions of 45^{o}C for hybridization and washed at 45^{o}C in 6x SSC for 15 minutes.

Rows A-D are duplicated on E-H respectively.
Columns 1-3 contain known amounts of JF-2 (a complementary oligo of JF-1), 10pg, 100pg and 1ng amounts respectively.

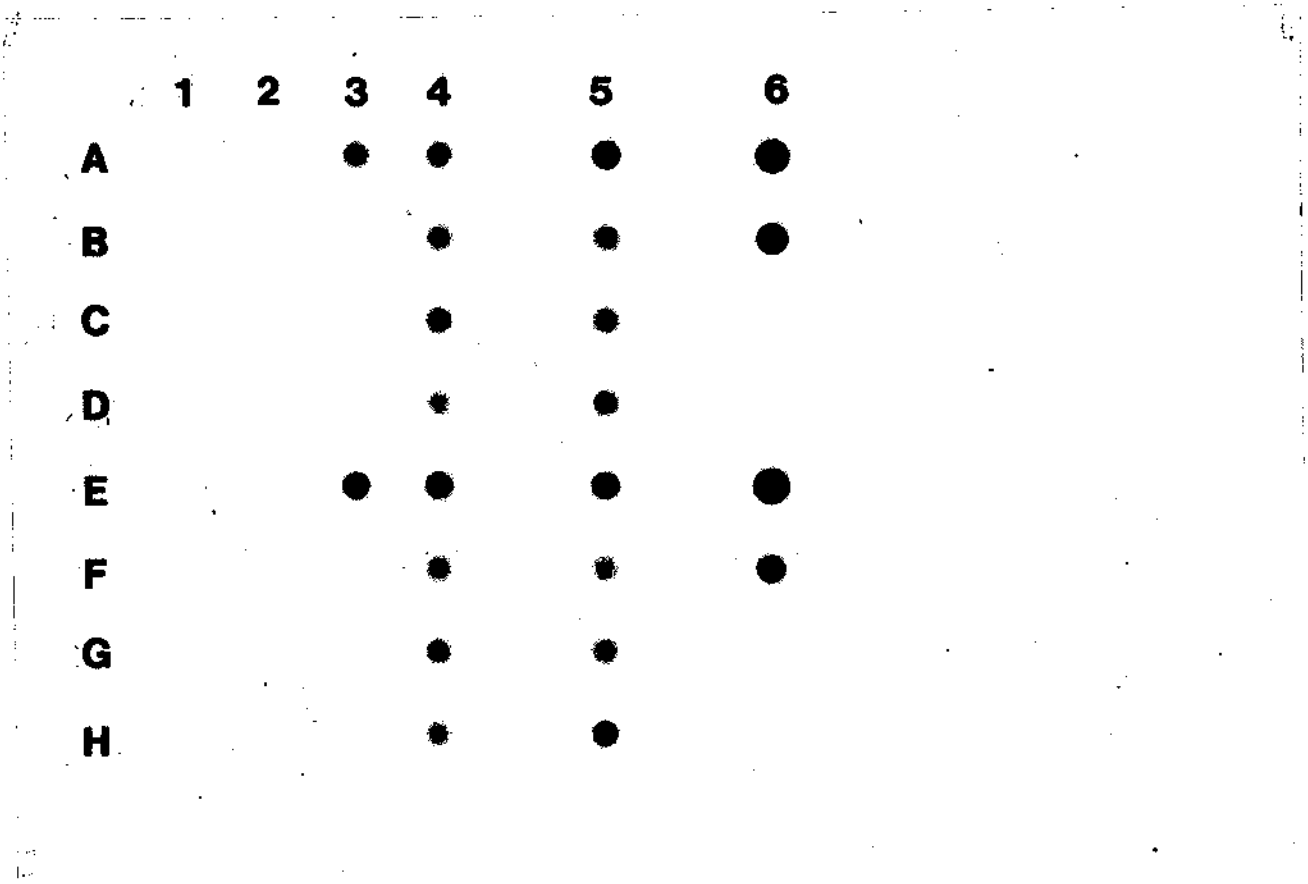
Column 4 contains 100ng of DNA from male horses

(A) Kingdom Bay
(B) Western Bay
(C) Colt (Western Bay x Lovely Habit)
(E) Colt (Western Bay x Silver Sophie)

Column 5 contains 100ng of DNA from female horses

(A) Latchmi
(B) Filly (Kingdom Bay x Latchmi)
(C) Lovely Habit
(D) Filly (Kingdom Bay x Darling Daughter)

Column 6A contains 100ng zebra DNA
Column 6B contains 100ng donkey DNA


1 2 3 4 5 6
A
B
C
D
E
F
G
H

The molecular weights of the nucleotides A, G, C, and T are:

Adenylic acid (AMP)	365.24
Guanylic acid (GMP)	361.21
Cytidylic acid (CMP)	321.18
Thymidylic acid (TMP)	320.19

The molecular weight of $(GATA)_5$

$$= (G \times 5) + (A \times 10) + (C \times 5)$$
$$= 1806.05 + 3652.4 + 1600.95$$
$$= 7059.4 \qquad \text{(i)}$$

The number of moles in 1 ng $(GATA)_5$

$$= m / Mr$$
$$= 1 \times 10^{-9} / 7059.4$$
$$= 1.42 \times 10^{-13} \qquad \text{(ii)}$$

where m refers to the mass in grams and Mr the molecular mass.

The number of molecules of $(GATA)_5$ in 1ng

$$= n \times N_o$$
$$= 1.42 \times 10^{-13} \times 6.022 \times 10^{23}$$
$$= 8.53 \times 10^{10} \qquad \text{(iii)}$$

where n refers to the number of moles and N_o as Avagadro's number (the number of molecules present in one mole).

The molecular weight of the horse genome

= the average base Mr x no. bases in the genome
= 341.955 x 3×10^9
= 1.03×10^{12} (iv)

The number of moles in 100ng genomic horse DNA (calculated as in (ii))

= $1 \times 10^{-7} / 1.03 \times 10^{12}$
= 9.75×10^{-20} (v)

The number of "genome equivalents" in 100ng (calculated as in (iii))

= $9.75 \times 10^{-20} \times 6.022 \times 10^{23}$
= 5.87×10^4 (vi)

Therefore, if there are 8.52×10^{10} $(GATA)_5$ repeats in 5.87×10^4 genomes, in a single genome there would be (iii)/(vi) $(GATA)_5$ repeats, ie

= $8.53 \times 10^{10} / 5.87 \times 10^4$
= 1.45×10^6 $(GATA)_{5+n}$ repeats (vii)

where $0 < n < 5$, as a single JF-1 probe can hybridize to a sequence containing more than 5 GATA repeats, but any more than 10 and a second JF-1 probe can hybridize adjacent to the first.

The proportion of these sequence in the genome can be calculated. The genome size is stated in bases, so, the number of $(GATA)_5$ sequences also needs to be converted to the number of bases. Each $(GATA)_5$ sequence contains 20 bases, therefore, the proportion of sequences within the genome can be calculated as follows:

$$P = (\text{no. of } (GATA)_5 \text{ sequences} \times 20 / \text{size of the genome}) \times 100$$
$$= (1.45 \times 10^6 \times 20 / 3 \times 10^9) \times 100$$
$$= 1.0\% \quad \text{(viii)}$$

So, $(GATA)_5$ sequences take up approximately 1.0% of the horse genome.

The probability of single GATA sequences occurring at random in the genome would be 1:4 x 4 x 4 x 4, ie 1:256 bases. The number of repeats this represents in the genome can be calculated as:

$$= \text{no. bases in the genome / probability of GATA sequences}$$
$$= 3 \times 10^9 / 256$$
$$= 1.17 \times 10^7 \quad \text{(ix)}$$

The expected proportion of this in the genome can be calculated as in (viii),

$$= ((1.17 \times 10^7) \times 4 / 3 \times 10^9) \times 100$$
$$= 1.6\% \quad \text{(x)}$$

The expected amount of single GATA sequences present in the horse genome, which occurred by random would be 1.6%.

But, the probability of $(GATA)_5$ occurring is $1:4^{20}$, which means the probability of $(GATA)_5$ sequences having occurred at random would be exceedingly small!

A stronger signal was produced from an equivalent amount of donkey and zebra DNA. This indicates that there are more GATA sequences in these species than is present in the horse.

3.2 DO GATA SEQUENCES SHOW LENGTH POLYMORPHISMS IN HORSE?

GATA polymorphisms could be detected in human DNA that had been digested with the restriction enzymes *Alu* I, *Hae* III, *Hinf* I, or *Mbo* I, and subsequently hybridized with a GATA oligomer (Ali *et al* 1986). This paper was used as a basis for determining if polymorphic GATA sequences occurred in the horse.

3.2.1 Detection of length polymorphisms in horse DNA

Genomic horse DNA was digested (section 2.3.2), concentrated to approximately 30µl (section 2.3.4), electrophoresed (section 2.4.1) then Southern blotted (section 2.4.3). Included in each blot was a human DNA sample as a positive control.

On a large gel a number of clearly visible satellite bands were seen after ethidium bromide staining. These generally occurred only when there was complete digestion of the DNA. The pattern of the bands varied depending on the restriction enzymes used, but for any one enzyme the banding pattern was the same for all individuals within a species. These bands represent highly repetitive DNA sequences. This type of banding can be clearly seen in fig 3.5a, especially around the 3 kb region in the horse samples.

In all cases the blots were hybridized at 45^{o}C, and washed at 40-42^{o}C. This was 3-5^{o}C below T_h at which hybridization theoretically occurred with 100% homology (section 2.4.5). These temperatures were based on results from dot blot experiments, section 3.1. The maximum number of bands were produced at these temperatures. These included DNA fragments which did not have strong homology to the probe, such as the bands in the lanes containing standard low molecular weight DNA (eg fig 3.5b). These had no homology to the probe (confirmed by a computor-based homology search, data not shown) but were still visible in the autoradiographs.

The effect of a few degrees in washing temperature was best demonstrated in fig 3.5b, where the blot was first washed at 40°C. This resulted in nonspecific hybridization of the probe to low molecular marker DNA as well to a number of genomic DNA fragments. At high stringency, hybridization did not occur to these specific fragments (fig 3.5c).

Mbo I was the most promising enzyme, producing the largest number of bands with considerable variation in band size (figs 3.5 and 3.6) indicating some GATA sequences were polymorphic. *Alu I* produced a number of bands with some variation between individual horses (fig 3.7). The enzymes *Hae* III and *Hinf* I produced few bands (figs 3.8 and 3.9). The number of bands and amount of variation produced by digestion with the enzymes *Alu* I, *Hae* III and *Hinf* I meant that they were not suitable for subsequent DNA fingerprinting studies. No significant banding patterns resulted with *Taq* I (fig 3.10). This was probably because of incomplete digestion, even after many attempts the gels did not run satisfactorily.

Enzymes which cut infrequently in the horse genome were also tried but with poor results (data not shown). These enzymes included *Kpn* I, *Eco* RI, *Hind* III, and *Hpa* I which either did not cut to completion and/or the resulting bands were blurred.

3.2.2 Family studies

DNA from three families of horses, each consisting of sire, dam and offspring were digested with *Mbo* I. Initally, there was a noticeable difference between the Sovereign Lodge DNA samples and those previously prepared from the Massey herd (fig 3.11). However, after the Sovereign Lodge samples had been cleaned up (section 2.3.1) they gave similar results to the control samples (fig 3.5c). A large number of bands were visible. The curving pattern of adjacent lanes is a result of the electrophoresis box used.

Fig 3.5 **GATA hybridization to a *Mbo* I digest of Sovereign Lodge DNA family samples after the DNA had been "cleaned-up" by proteinase K digestion and phenol-chloroform extraction**

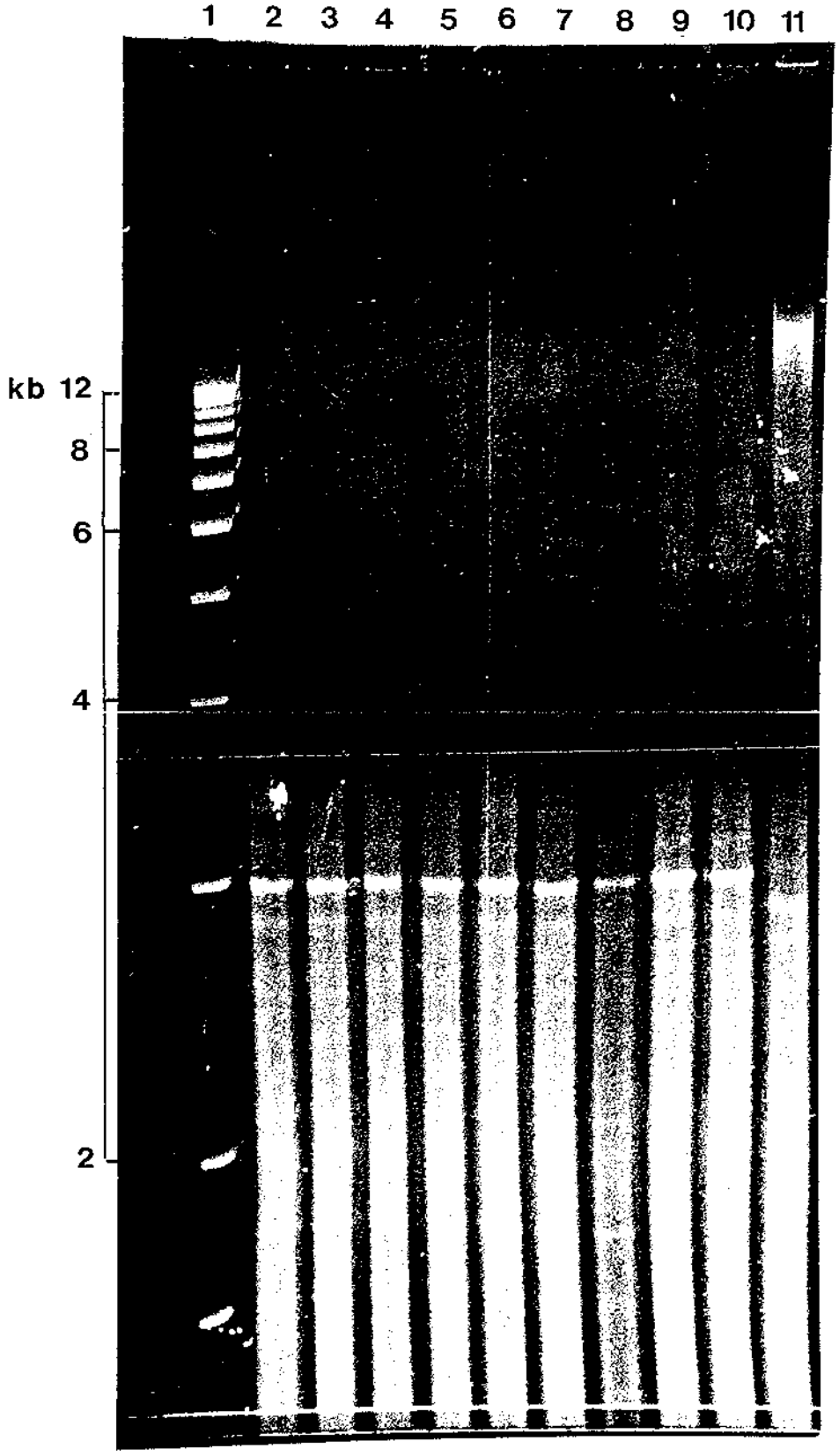
(a) Agarose gel. Digested DNA was electrophoresed in 1.0% agarose gel for 17 hours at 80V.

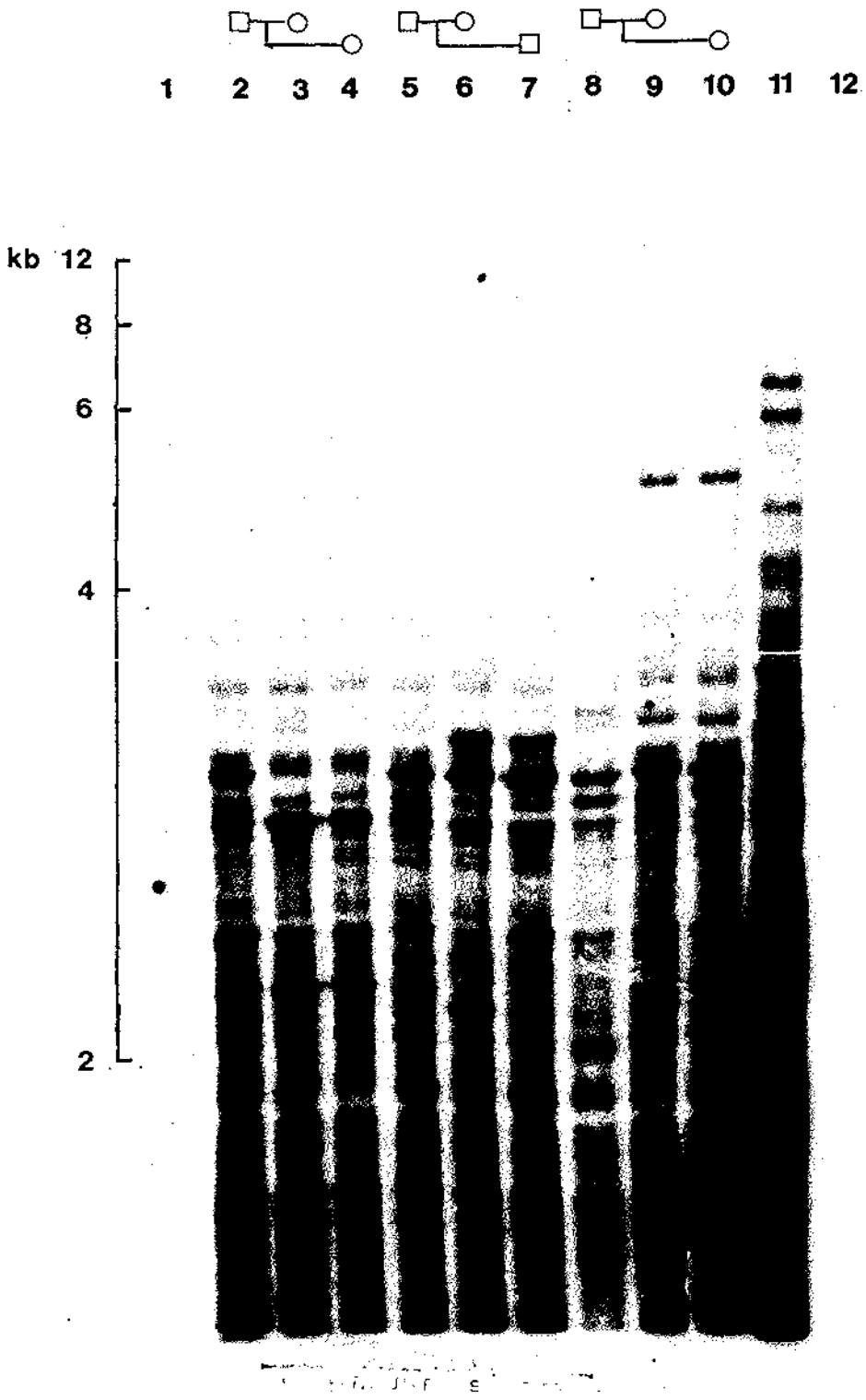
(b) Autoradiograph. Following Southern blotting, the membrane was hybridized at standard conditions of 45^{o}C and washed at 40-42^{o}C in 6x SSC for 15 minutes (section 3.2.1) then autoradiographed for three days at -20^{o}C with two screens.

(c) The above blot was rewashed at 45^{o}C. It was then autoradiographed for three days at -20^{o}C with two screens.

Lanes contain DNA from:

1 BRL low molecular weight standard
2 Kingdom Bay
3 Latchmi
4 Filly (Kingdom Bay x Latchmi)
5 Western Bay
6 Lovely Habit
7 Colt (Western Bay x Lovely Habit)
8 Massey herd horse 5
9 Massey herd horse 6
10 Massey herd horse 8
11 Human


1 2 3 4 5 6 7 8 9 10 11
kb 12
8
6
4
2


1 2 3 4 5 6 7 8 9 10 11 12
kb 12
8
6
4
2

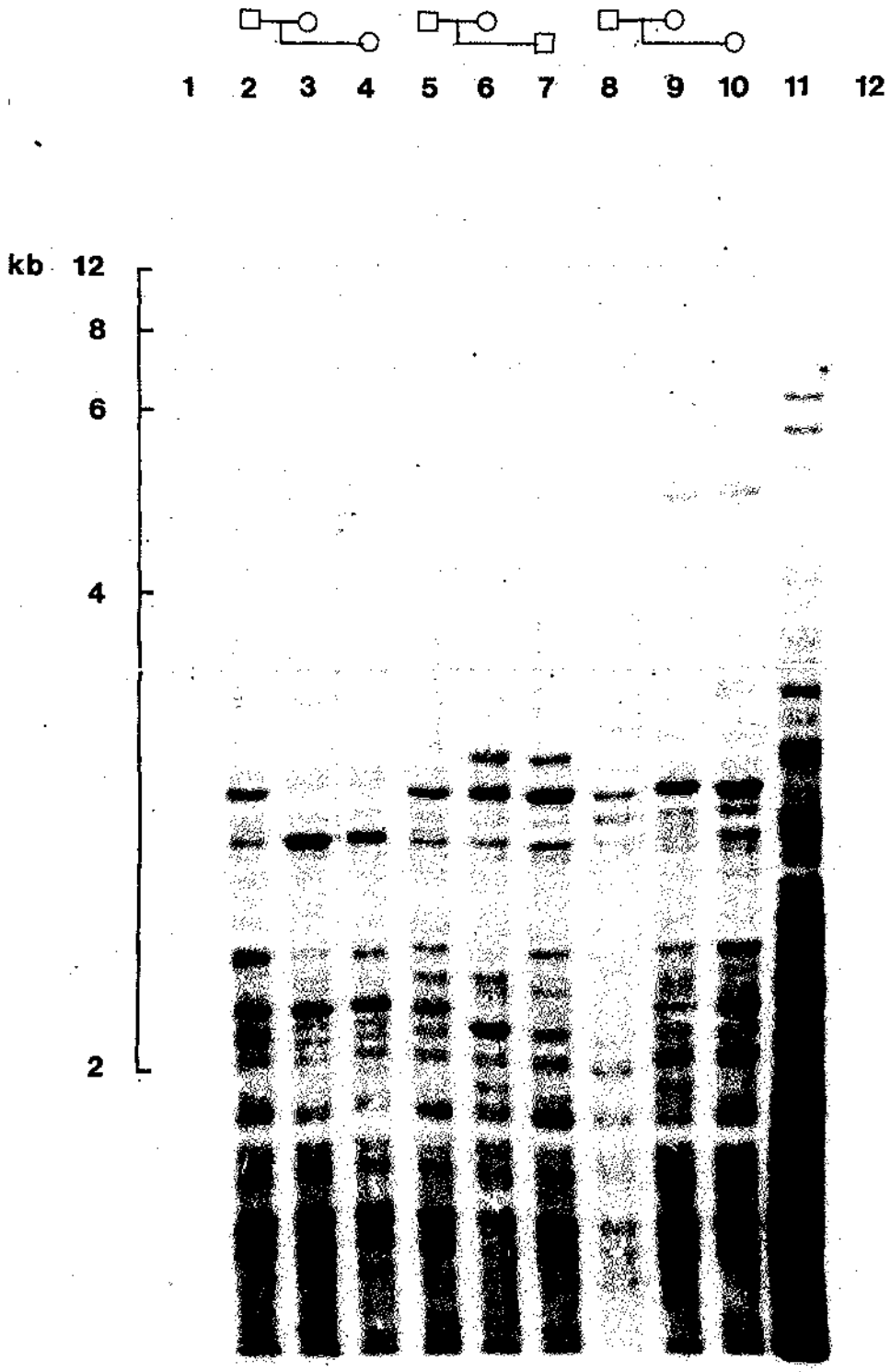

1 2 3 4 5 6 7 8 9 10 11 12
kb
12
8
6
4
2

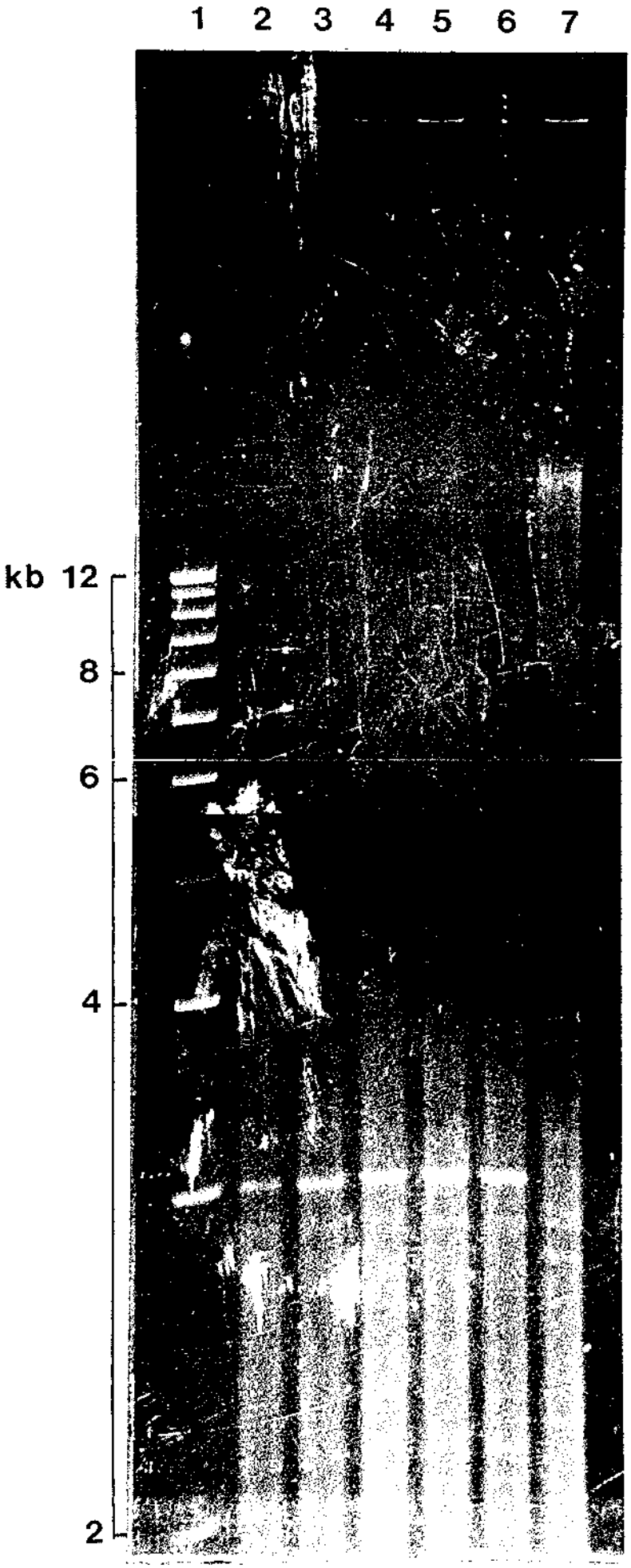
Fig 3.6 **GATA hybridization to a *Mbo* I digest of genomic DNA**

(a) Agarose gel. Digested DNA was electrophoresed along a 1.0% agarose gel for 17 hours at 80V.

(b) Autoradiograph. Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) then autoradiographed for five days at -20^{o}C with two screens.

Lanes contain 5µg genomic DNA, digested with 20U enzyme

1 BRL low molecular weight standard DNA
2 Massey herd horse 5 (male)
3 Massey herd horse 6 (female)
4 Massey herd horse 8 (")
5 Massey herd horse 9 (")
6 Massey herd horse 11 (")
7 Human
8 BRL low molecular weight standard DNA


1
2
3
4
5
6
7
kb
12
8
6
4
2

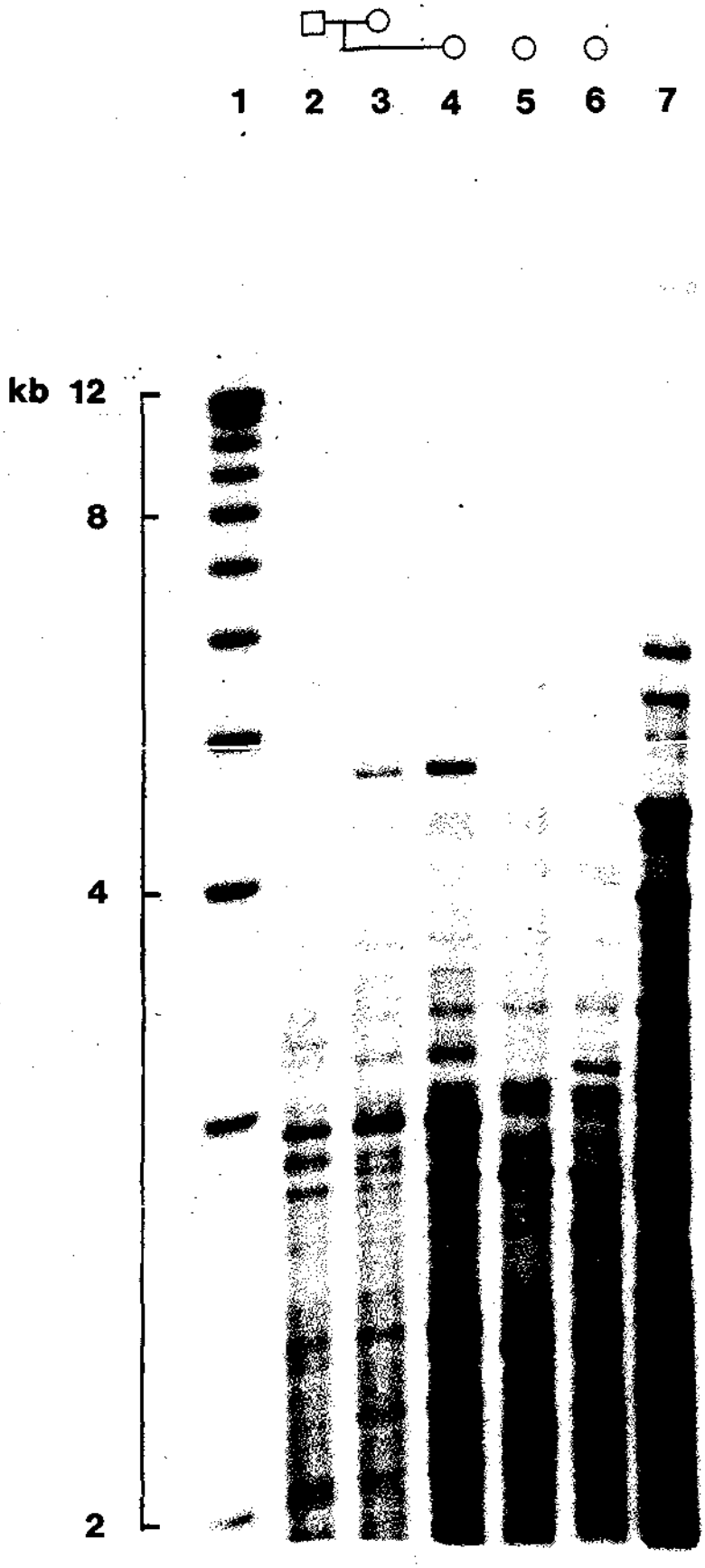

1 2 3 4 5 6 7
kb 12
8
4
2

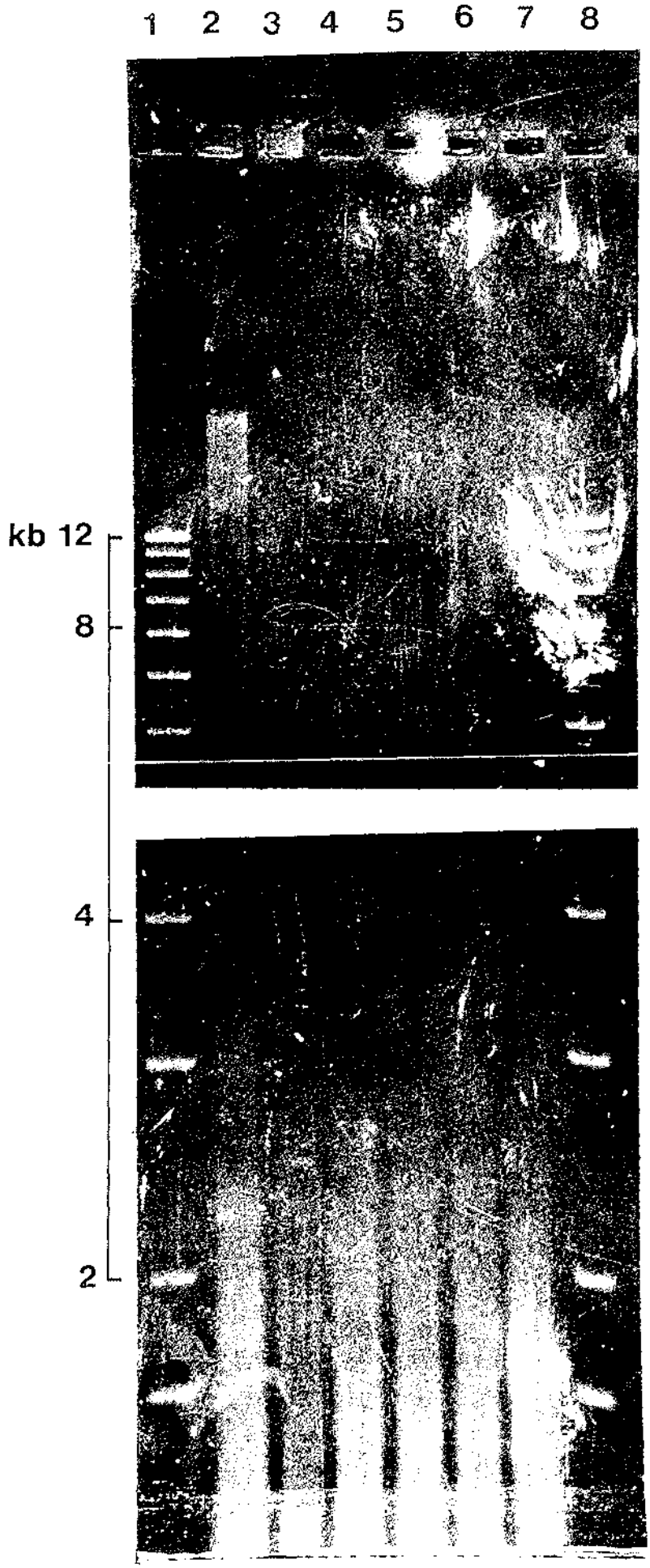
Fig 3.7 **GATA hybridization to an *Alu* I digest of genomic DNA**

(a) Agarose gel. Digested DNA was electrophoresed along a 0.8% agarose gel for 16 hours at 60V.

(b) Autoradiograph. Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) then autoradiographed for five days at -20^{o}C with two screens.

Lanes contain 5µg genomic DNA, digested with 20U enzyme

1 BRL low molecular weight standard DNA
2 Human
3 Massey herd horse 5 (male)
4 Massey herd horse 6 (female)
5 Massey herd horse 8 (")
6 Massey herd horse 9 (")
7 Massey herd horse 11 (")
8 BRL low molecular weight standard DNA


1 2 3 4 5 6 7 8
kb 12
8
4
2

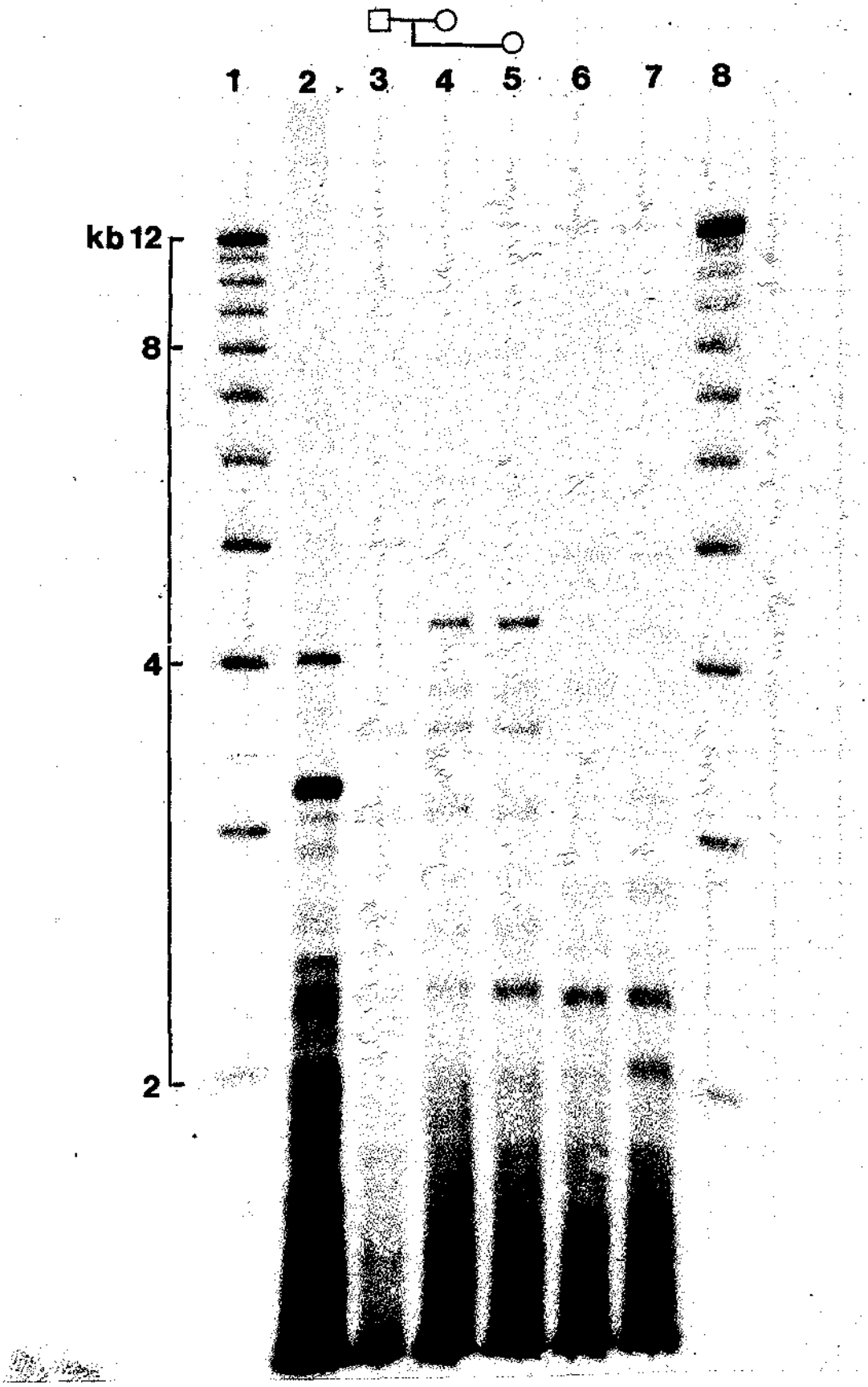

1
2
3
4
5
6
7
8
kb 12
8
4
2

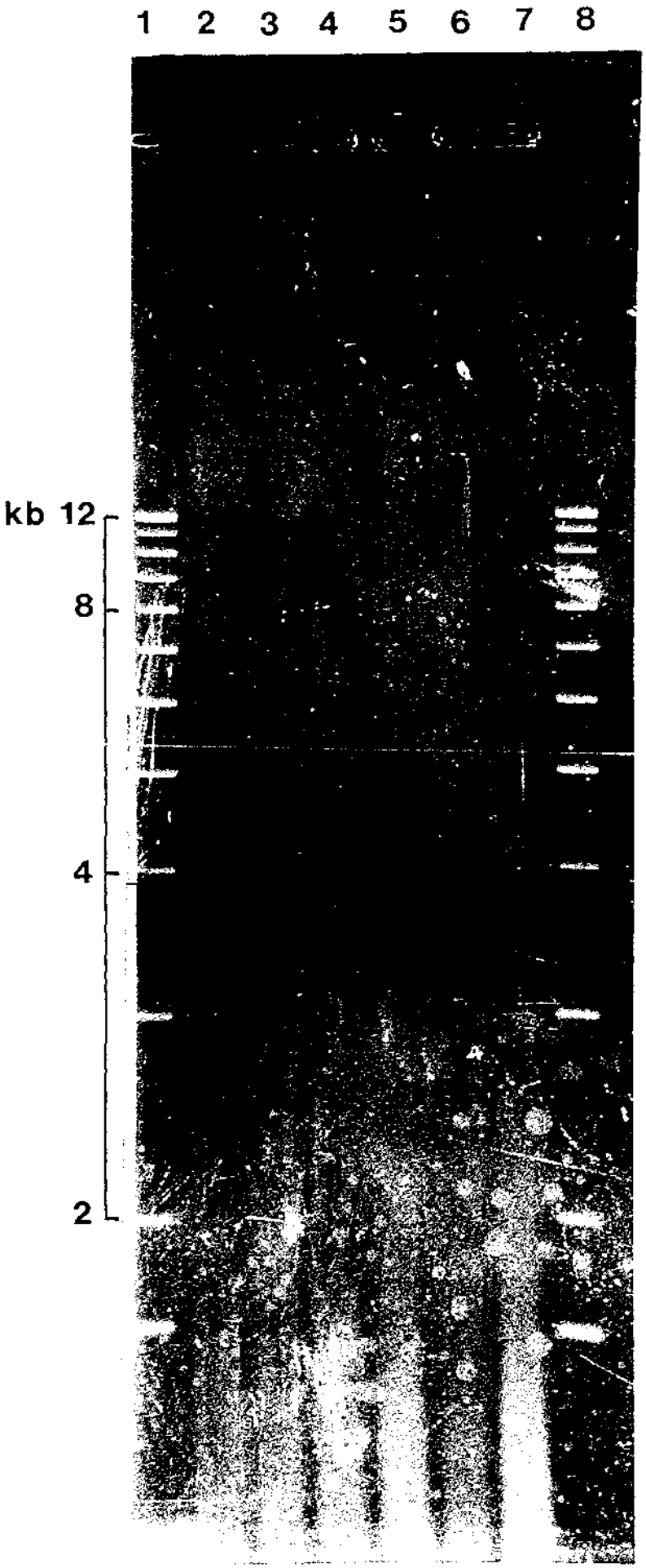
Fig 3.8 **GATA hybridization to a *Hae* III digest of genomic horse DNA**

(a) Agarose gel. Digested DNA was electrophoresed along a 0.8% agarose gel for 17 hours at 60V.

(b) Autoradiograph. Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) then autoradiographed for five days at -20°C with two screens.

Lanes contain 5µg genomic DNA, digested with 20U enzyme

1 BRL low molecular weight standard DNA
2 Massey herd horse 5 (male)
3 Massey herd horse 6 (female)
4 Massey herd horse 8 (")
5 Massey herd horse 9 (")
6 Massey herd horse 11 (")
7 Human
8 BRL low molecular weight standard DNA


1 2 3 4 5 6 7 8
kb 12
8
4
2

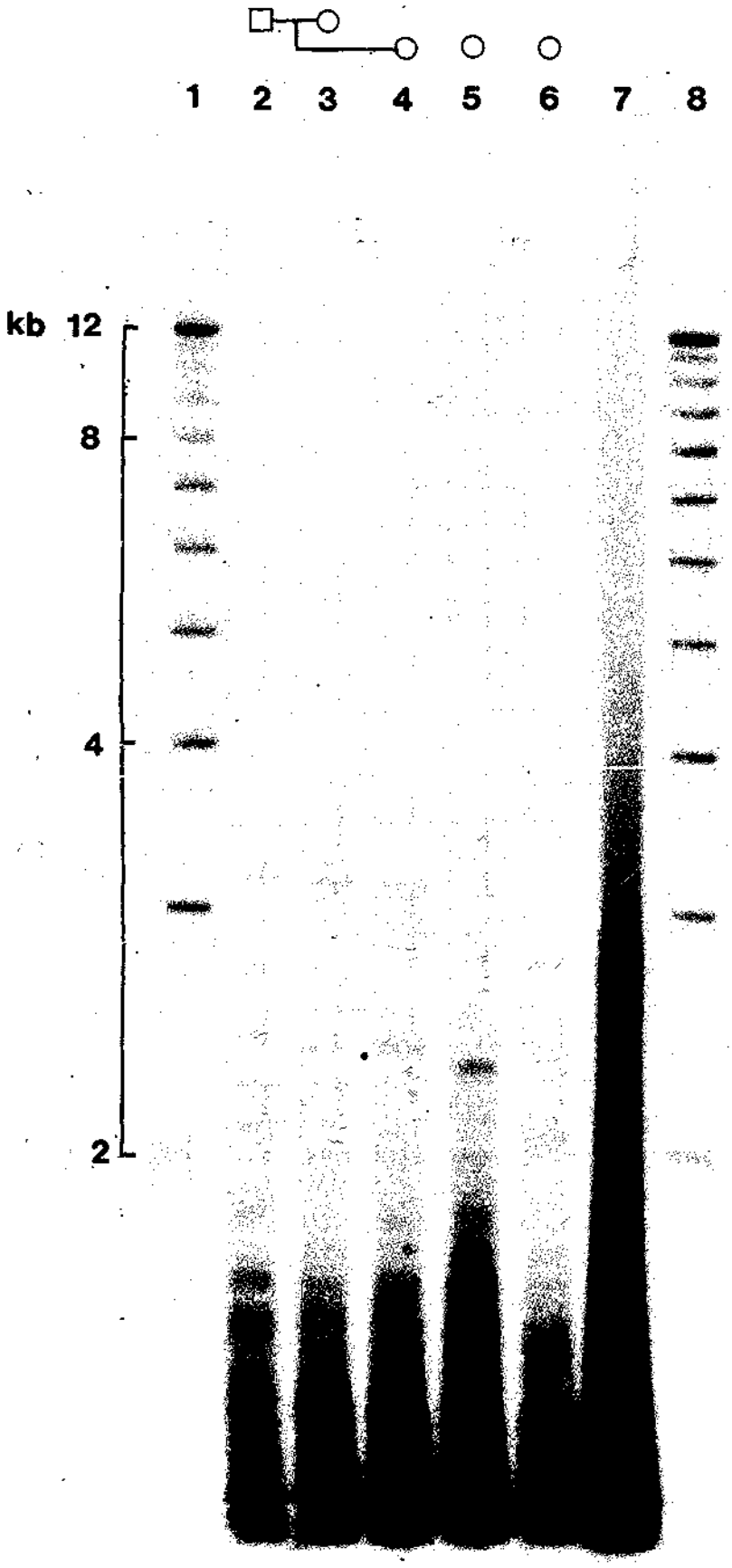

1 2 3 4 5 6 7 8
kb 12
8
4
2

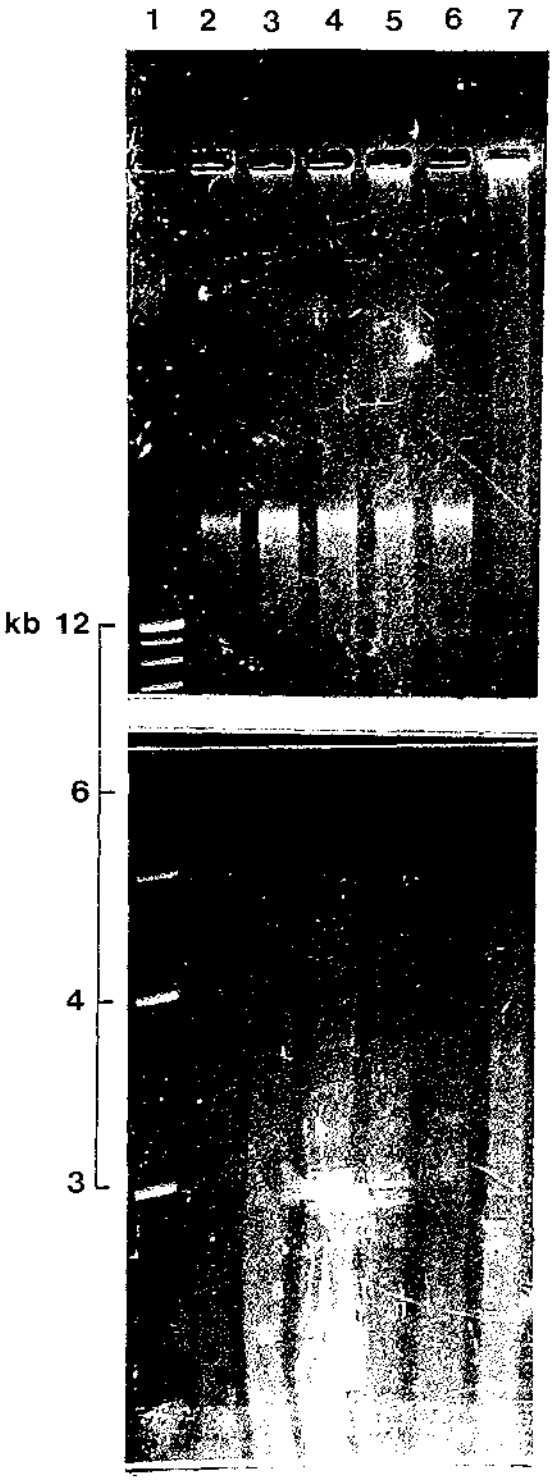
Fig 3.9 **GATA hybridization to a *Hinf* I digest of genomic horse DNA**

(a) Agarose gel. Digested DNA was electrophoresed along a 1.0% agarose gel for 17 hours at 80V.

(b) Autoradiograph. Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) autoradiographed for five days at -20^{o}C with two screens.

Lanes contain 5µg genomic DNA, digested with 20U enzyme

1 BRL low molecular weight standard DNA
2 Massey herd horse 5 (male)
3 Massey herd horse 6 (female)
4 Massey herd horse 8 (")
5 Massey herd horse 9 (")
6 Massey herd horse 11 (")
7 Human


1 2 3 4 5 6 7
kb 12
6
4
3

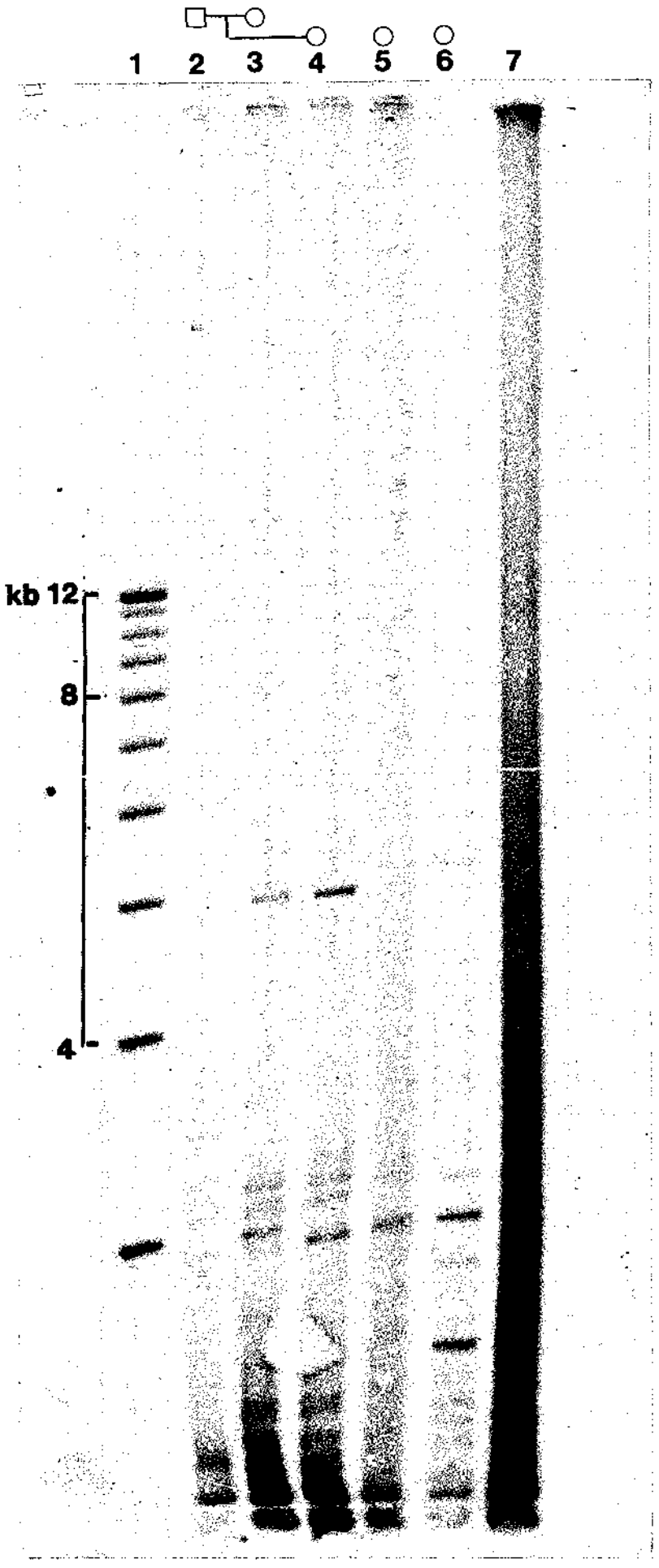

1
2
3
4
5
6
7
kb 12
8
4

Fig 3.10 GATA hybridization to a *Taq* I digest of genomic horse DNA.

(a) Agarose gel. Digested DNA was electrophoresed along a 0.8% agarose gel for 18 hours. The ethidium bromide staining at high molecular weights indicates incomplete digestion with *Taq* I. Although the same amounts of DNA were loaded in each lane, incomplete digestion meant that the DNA would spool out of the gel wells of its own accord.

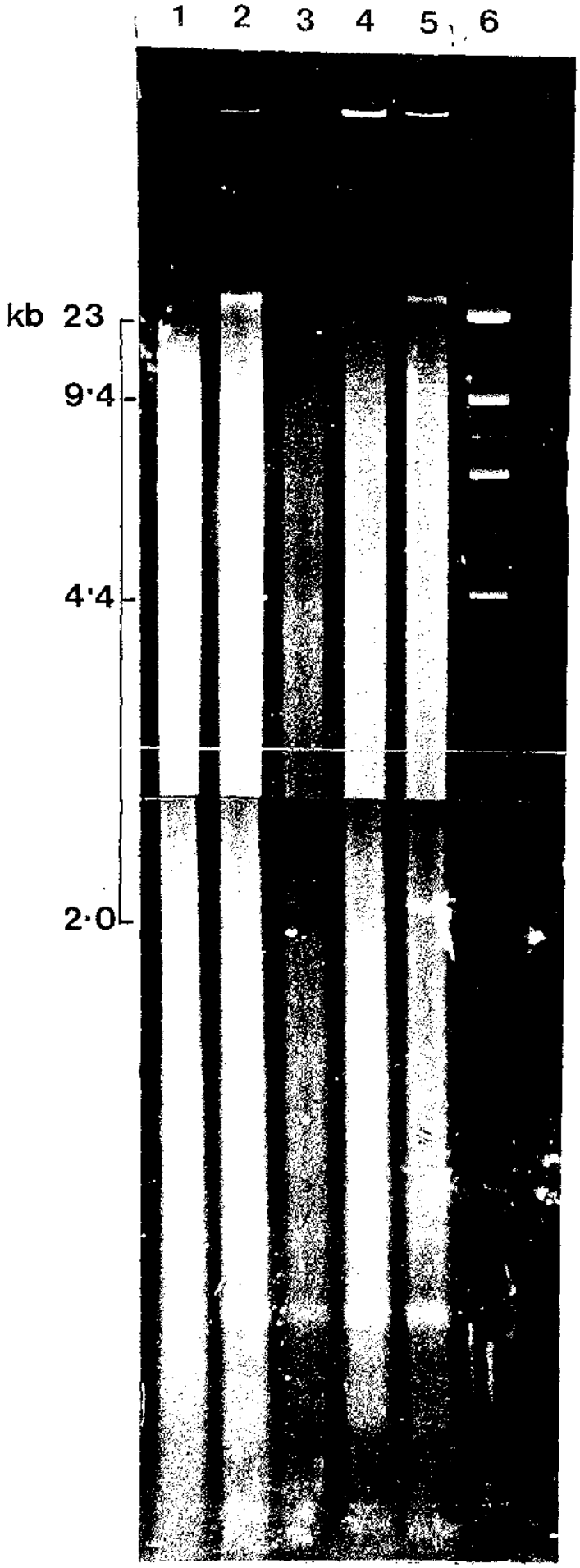
Lanes were loaded with 5µg genomic DNA

1 Human
2 Kingdom Bay
3 Filly (Kingdom Bay x Darling Daughter)
4 Darling Daughter
5 Silver Sophie
6 λ DNA digested with *Hind* III

(b) Autoradiograph of gel shown in (a). Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) then autoradiographed for two and a half days with two screens at -20^{o}C.

Lanes contain 5µg genomic DNA

1 λ DNA digested with *Hind* III
2 Human
3 Kingdom Bay
4 Filly (Kingdom Bay x Darling Daughter)
5 Darling Daughter
6 Silver Sophie


1
2
3
4
5
6
kb
23
9·4
4·4
2·0

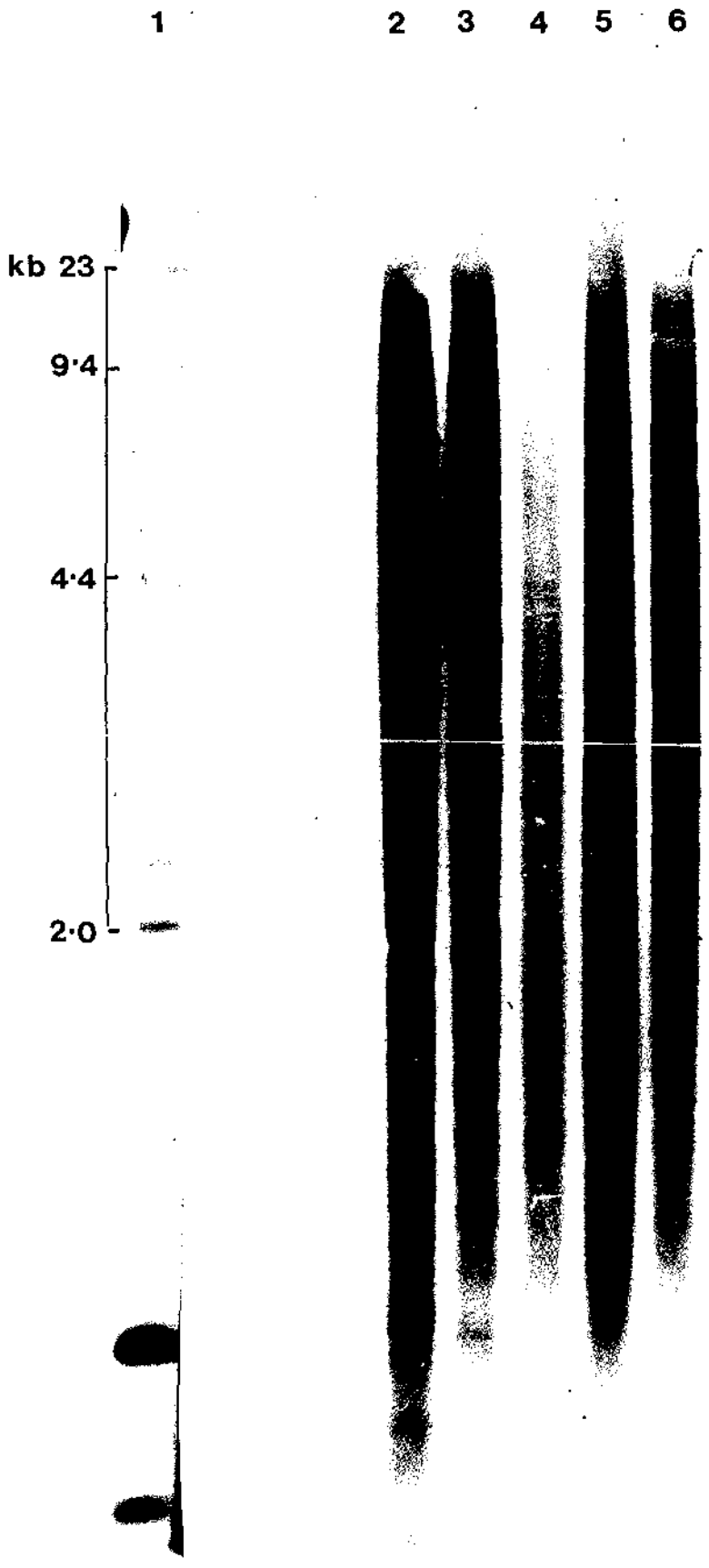

1
2
3
4
5
6
kb 23
9·4
4·4
2·0

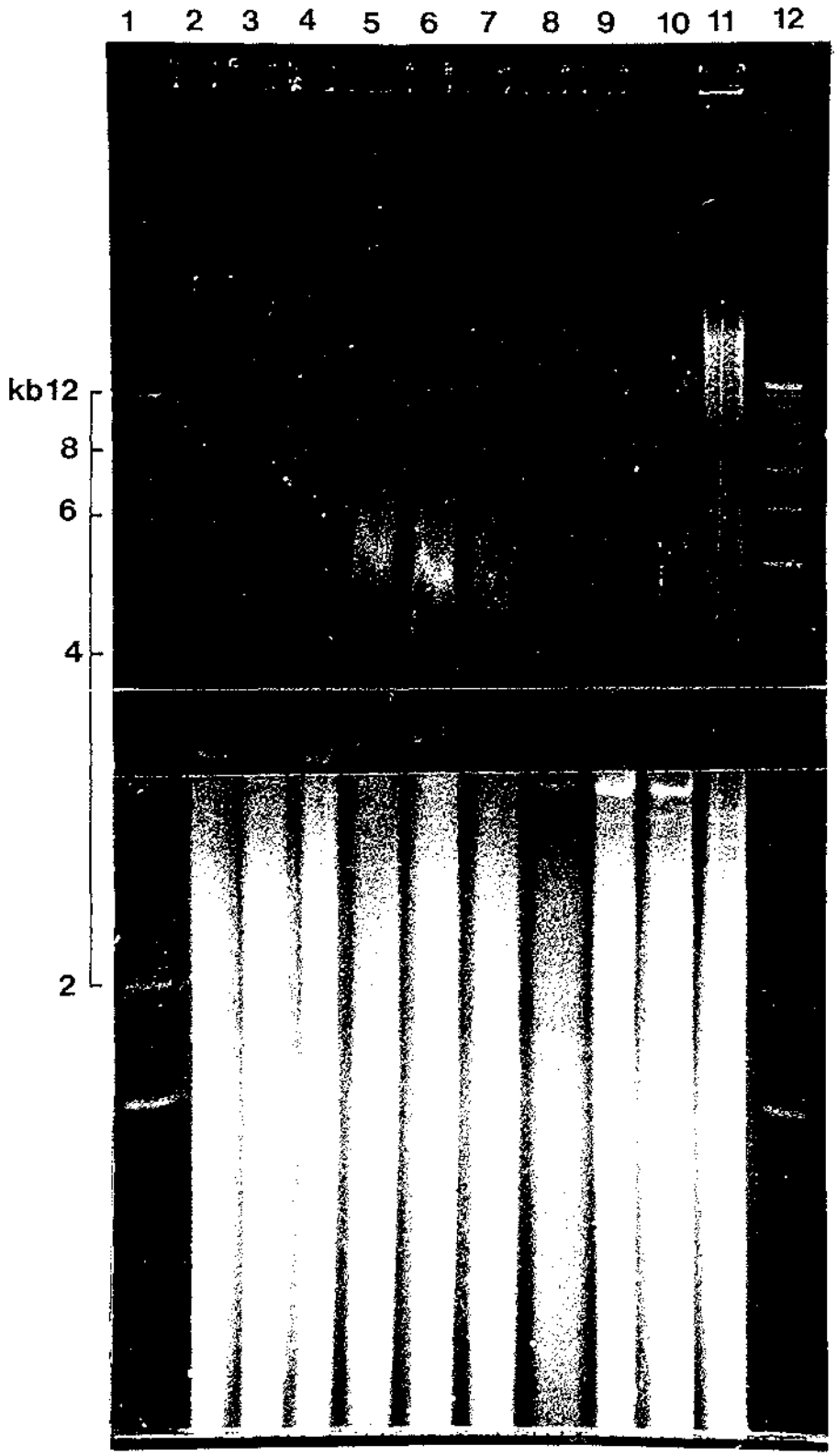
Fig 3.11 **GATA hybridization to a *Mbo* I digest of Sovereign Lodge DNA family samples.**

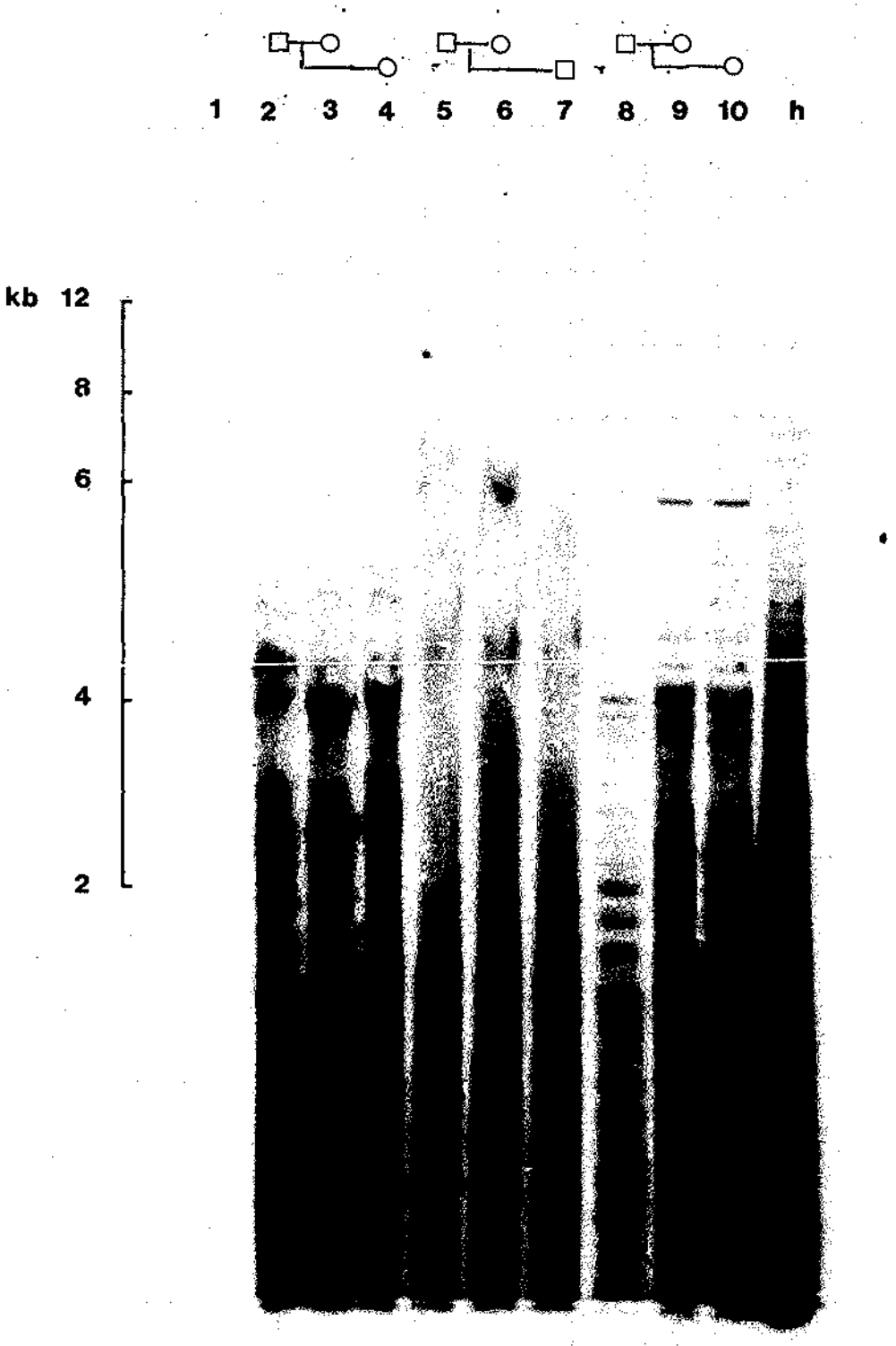
(a) Agarose gel. Digested DNA was electrophoresed along a 1.0% agarose gel for 17 hours at 80V.

(b) Autoradiograph. Following Southern blotting, the membrane was hybridized and washed at standard conditions (see fig 3.5b) then autoradiographed for three days at -20^{o}C with two screens.

Lanes contain DNA from:

1 BRL low molecular weight standard DNA
2 Kingdom Bay
3 Latchmi
4 Filly (Kingdom Bay x Latchmi)
5 Western Bay
6 Lovely Habit
7 Colt (Western Bay x Lovely Habit)
8 Massey herd horse 5
9 Massey herd horse 6
10 Massey herd horse 8
11 Human


1
2
3
4
5
6
7
8
9
10
11
12
kb12
8
6
4
2


1 2 3 4 5 6 7 8 9 10 h
kb 12
8
6
4
2

Although DNA digested with *Mbo* I produced 50-100 highly polymorphic bands, the polymorphisms were conserved, therefore, there was little variation between individuals and families. It is, however, interesting that some bands were very intense in females and appeared in the offspring, independent of their sex.

3.2.3 GATA Polymorphisms between species

JF-1 was hybridized under standard conditions to a zooblot (prepared by P Lewis, DSIR) containing horse, donkey, cow, goat, human and sheep DNA (fig 3.12). There were different levels and patterns of hybridization to the various species. As the same amount of DNA was in each lane, the difference appears to be due to the amount of GATA sequences in the species. The surprising result was the large difference in the amount of hybridization between closely related species - horse and donkey, sheep and goat.

3.3 GENOMIC ORGANIZATION OF A GATA SEQUENCE

3.3.1 Isolation of a genomic GATA DNA fragment

Only one fraction (containing 1kb fragments) was successfully ligated into pUC (section 2.6). This was transformed into DH5α and the resulting transformants screened (section 2.6). Of approximately 180 colonies, one positive clone was found (fig 3.13) which was named p37/68. This clone was cultivated and the plasmid DNA isolated initially by minipreparation (section 2.6.5.1). An *Eco* RI and *Pst* I digest showed the 37/68 insert to be 3.6kb. The DNA was blotted and hybridized to JF-1 which showed GATA sequence was present in the insert and not in pUC.

3.3.2 Restriction mapping of insert 37/68

Plasmid DNA was extracted from a large scale plasmid preparation (section 2.6.5.2) and the insert isolated by digestion with *Eco* RI and *Pst* I followed by electroelution (section 2.7.2).

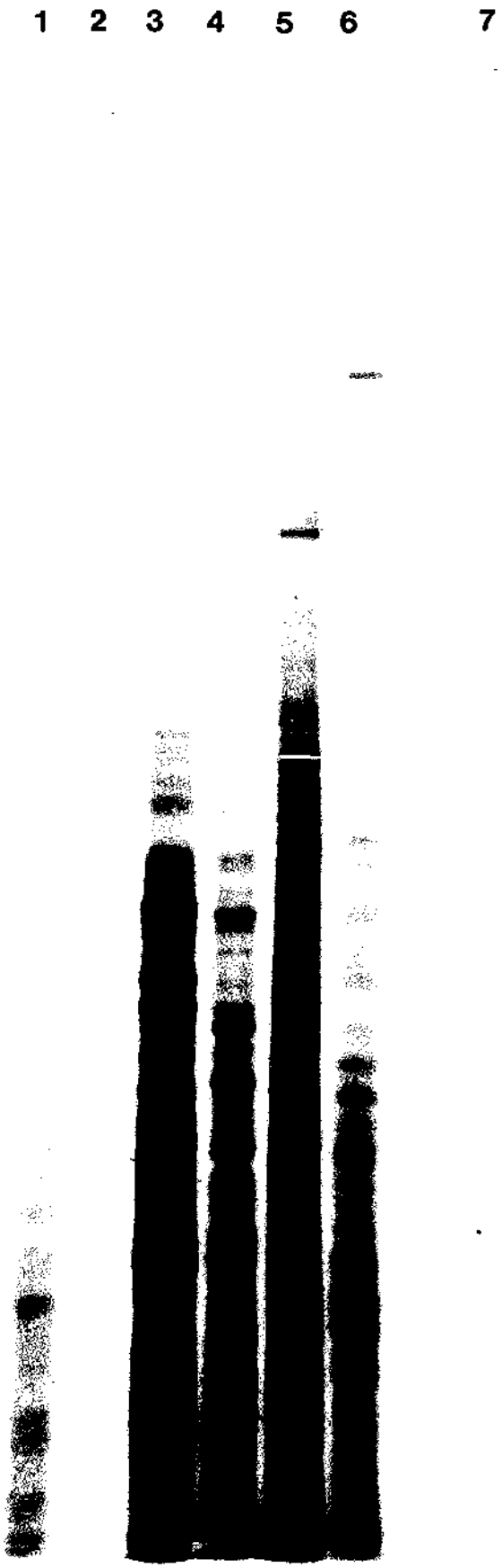
Fig 3.12 GATA hybridization to a multispecies blot

7µg genomic DNA was digested with *Sau* 3A, electrophoresed in a 1% agarose gel for 16 hours

JF-1 was hybridized to the blot at standard conditions (see fig 3.5b).

Lanes contain DNA from:

1 Sheep
2 Cattle
3 Donkey
4 Horse
5 Human
6 Goat
7 BRL low molecular weight DNA standard


1 2 3 4 5 6 7

Restriction mapping was performed on the insert only. Five of the fourteen enzymes used (table 2.3) had restriction sites within the insert. These were *Alu* I and *Sau* 3A, which had too many restriction sites to be useful for restriction mapping. *Hae* III and *Hind* III each had one restriction site respectively, and *Taq* I had two.

As the sizes of the 37/68 insert and pUC were known and single *Hind* III sites occurred in both, the relative orientation of the insert could be determined by digestion of p37/68 with *Hind* III. This resulted in a 4.2kb fragment, representing pUC and part of 37/68, and a 2.1kb fragment representing part of 37/68 only. The size of these fragments showed the *Hind* III site in 37/68 was closer to the *Eco* RI end than the *Pst* I end.

The restriction map was determined by performing double digests (fig 3.14) using combinations of the three enzymes as according to method 2.7.2. Fig 3.15 shows the final restriction map obtained.

To locate the GATA region within 37/68, the insert was digested with the three restriction mapping enzymes, electrophoresed, blotted, then hybridized with the probe. The fragments which the probe hybridized to were all from one end within 37/68 (fig 3.16), showing that there was only one GATA region.

3.3.3 Subcloning the GATA portion of 37/68

This was necessary for sequencing. The 3.6kb fragment was digested with *Hae* III and the 1kb Eco-Hae fragment purified by electroelution then directionally ligated into M13 *mp18* and *mp19* which had been previously digested with *Eco* RI and *Hinc* II as according to standard methods (section 2.7). This directional subcloning enabled sequencing to be initiated at both ends of the 1kb fragment.

Fig 3.13 **Screening of clones which contained DNA from fraction 37 with a GATA probe**

Blots were hybridized and washed at standard conditions (see fig 3.5b).

Upper blot

A1 blank - no bacteria, but undergone the same treatment as other dots.

A2 DH5α without plasmid.

A3 DH5α containing pUC, no insert.

A4 DH5α blue colony, ie, unsucessful transformation.

A7 approx. 100ng genomic human DNA.

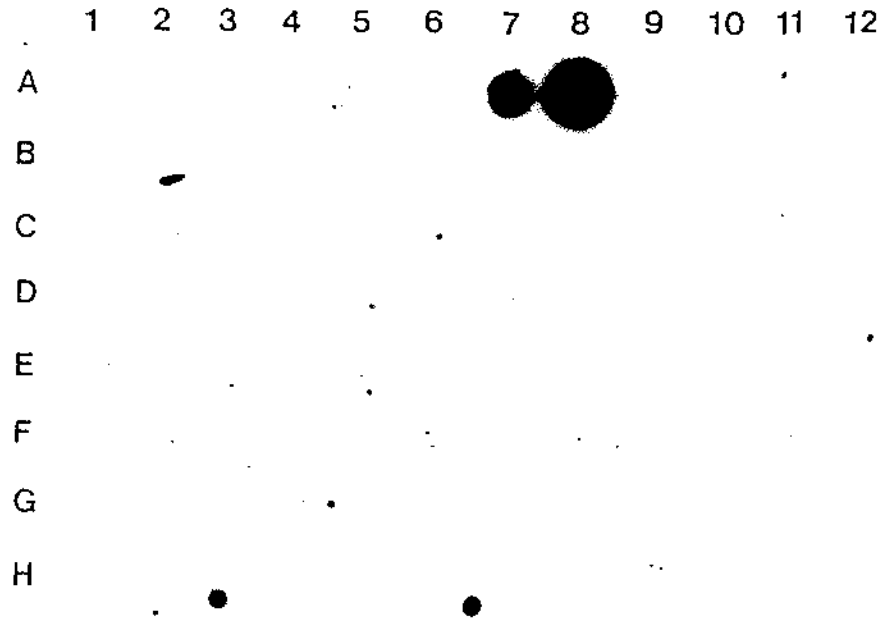
A8 approx. 100ng genomic horse DNA.

A1-A4 acted as negative controls.

A7-A8 acted as positive controls.

Lower blot

F8 is a positive clone, named 37/68.


1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H

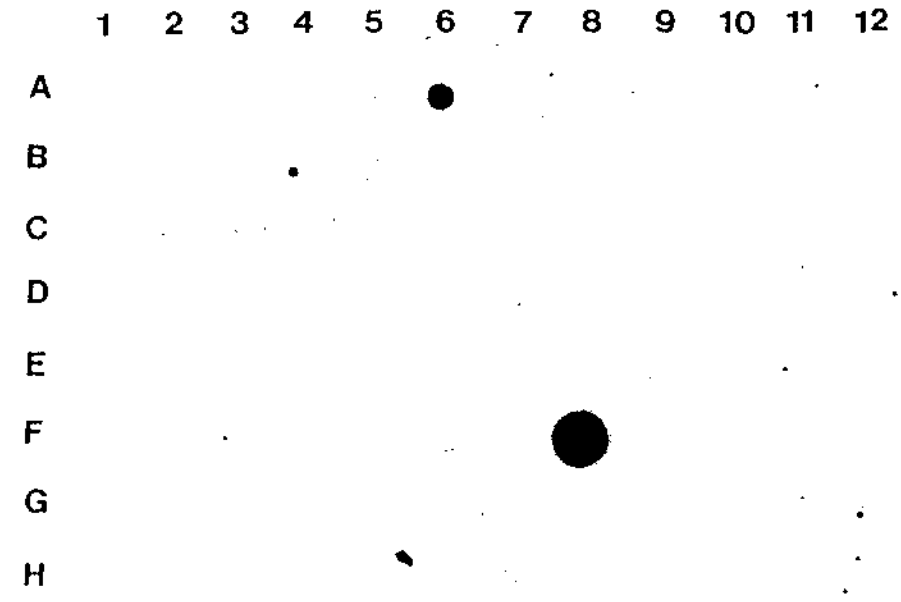

1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H

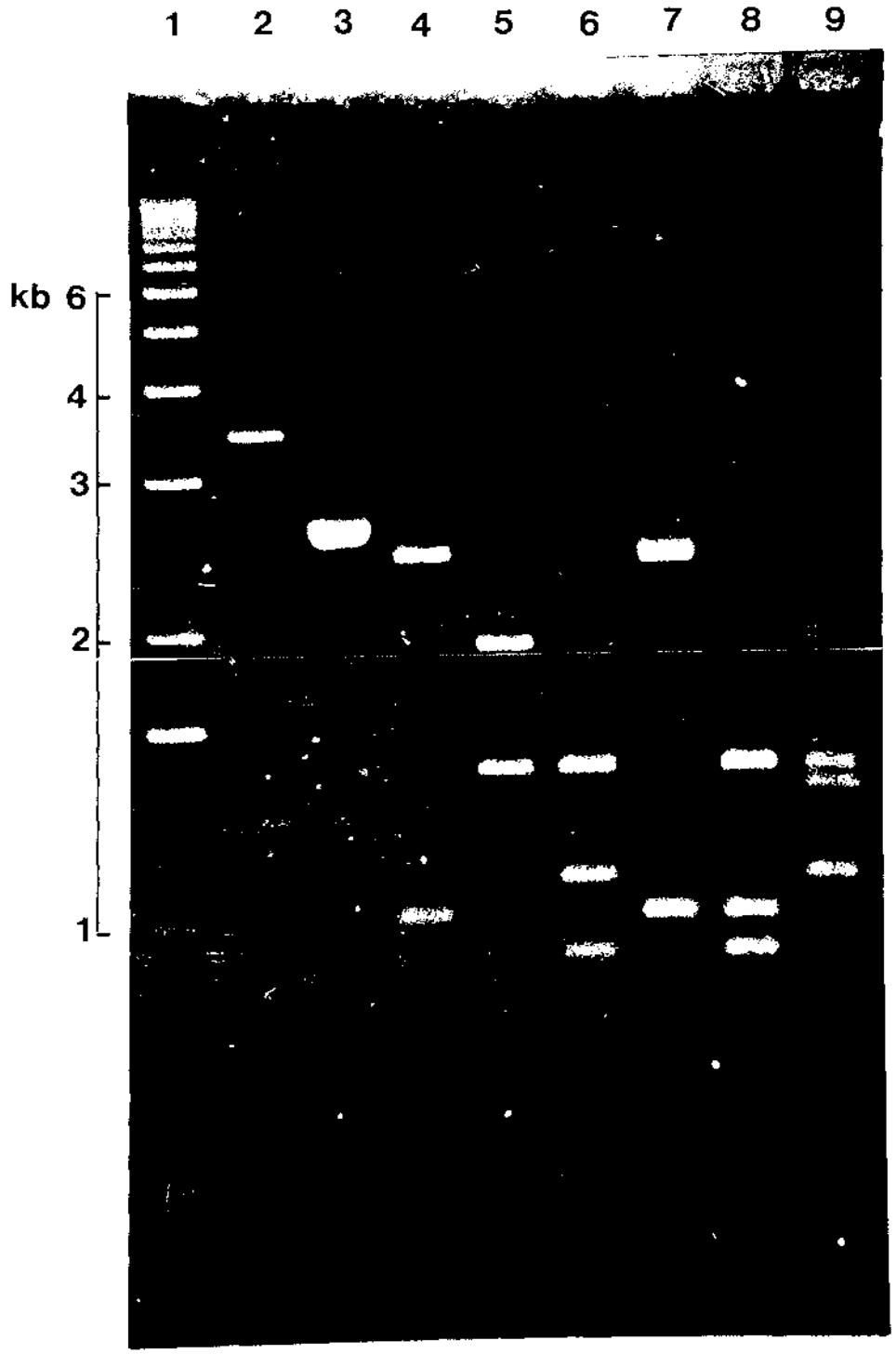
Fig 3.14 **Restriction digest of 37/68**

37/68 was digested with various restriction enzymes and the digested DNA was electrophoresed in a 0.8% TBE agarose gel at 100V for 16 hours.

Lanes contain

1. BRL Low molecular weight standard DNA
2. undigested 37/68 (positive control)
3. pUC 18 cloning vector digested with *Eco* RI (negative control)
4. *Hae* III digested 37/68
5. *Hind* III digested 37/68
6. *Taq* I digested 37/68
7. *Hind* III/*Hae* III digested 37/68
8. *Taq* I/*Hae* III digested 37/68
9. *Taq* I/*Hind* III digested 37/68

In Lane 9 the DNA was not completely digested. Repeated attempts led to no improvement.


1
2
3
4
5
6
7
8
9
kb 6
4
3
2
1

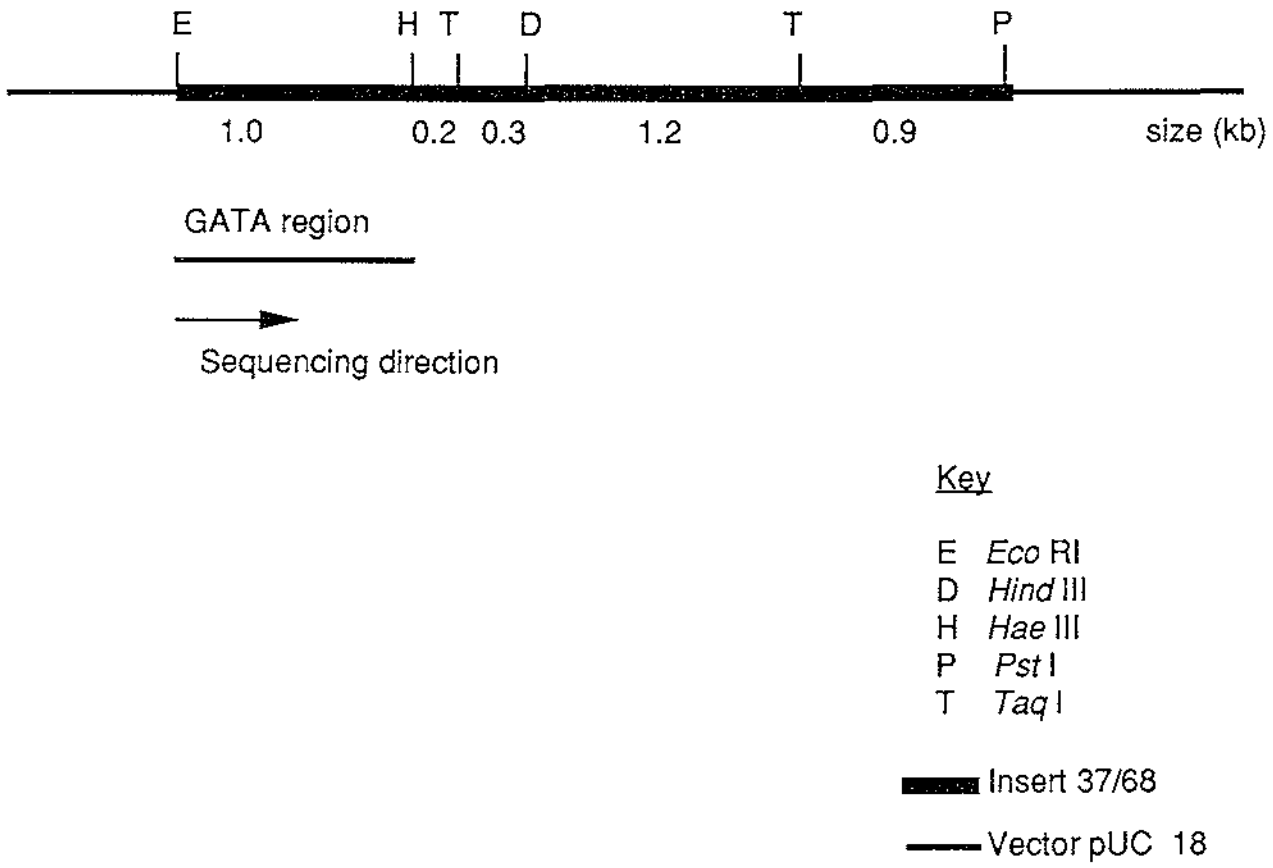


Fig 3.15 Restriction Map of 37/68

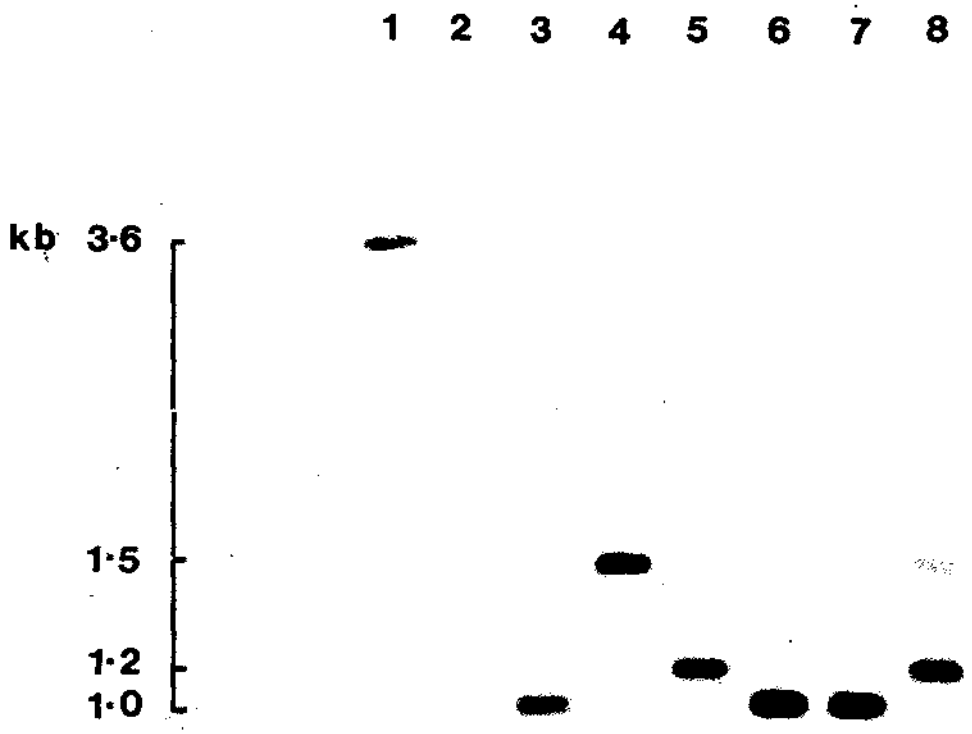
Fig 3.16 Localization of GATA sequences within 37/68

37/68 was digested with various restriction enzymes and the digested DNA was electrophoresed in a 0.8% TBE agarose gel at 100V for 16 hours. After Southern blotting and hybridization at standard hybridization and wash condition (see fig 3.5b), the blot was autoradiographed for 1 hour without intensifying screens.

Lanes contain

1 undigested 37/68 (positive control)
2 pUC 18 cloning vector (negative control)
3 *Hae* III digested 37/68
4 *Hind* III digested 37/68
5 *Taq* I digested 37/68
6 *Hind* III/*Hae* III digested 37/68
7 *Taq* I/*Hae* III digested 37/68
8 *Taq* I/*Hind* III digested 37/68

The faint 3.6kb bands visible in lanes 4, 6 and 8 indicates very slight incomplete digestion. These were not visible in the minigel (Fig 3.14).


1 2 3 4 5 6 7 8
kb 3·6
1·5
1·2
1·0

3.3.4 Sequencing the GATA portion of 37/68

Single stranded sequence data was obtained (section 2.7.7). There were 451 bases from the *Hind* III end (fig 3.19) and 506 bases from the *Eco* RI end sequenced (fig 3.21) which meant approximately 40 bases were required for the complete sequence of the 1kb fragment. Both sequences initiated from the multiple cloning site.

Upon performing a homology search of the EMBL database using the FASTN program (section 2.7.7), the 451 bases sequenced from the *Hind* III were found to have 99% identity to M13 DNA. It appears that during the ligation procedure (section 2.7.3) a fragment of M13 was ligated into the cloning cassette instead of part of 37/68. Personal enquiry has revealed cloning of sequencing vector is not uncommon when using the shot-gun sequencing approach. However, this data served as a useful control as to the accuracy of the sequencing performed and as a comparison for the level of identity with other homology searches.

A string of GATA sequences were found 347 bases from the *Eco* RI end. Within 190 bases, there were 34 strings of GATA sequences, the longest contained 6 repeats, and the shortest had just a single GATA sequence. There were an average of 3 GATA repeats per string. The GATA strings were nearly always separated by a G-A-T or G-A-T-G-A-T sequence, with only 5 exceptions. Only a single C base was present in the total 190 base sequence and this occurred near the 5' end of the GATA sequence.

The sequence preceding the GATA region was very A-T rich, with very few C bases.

1	GAATTGGAGC	TCGGCACCCG	GGGATCCTCT	AGAGTCCCCA	GATTTAATTT
51	TAATTATTTA	AAATTTAAAT	TAAAAAATGA	AAGCAGTATA	AAATATTTTT
101	CCACTAAACA	CAACTACTTG	TTTCAATAGG	ACTATATTTC	ATTTTAACCA
151	TTGAAAATTT	AATATATCTG	AATTGAGATG	GGATGGAAGT	GTAAAATACA
201	TGCTGGATTT	CAAAGACAGT	CTTTAAGGAA	AAAAATAAAT	GTAATATCTC
251	ATTAATAATT	TGTTTATATT	GATCACATGC	TGAATGATTA	TGTTTGATAT
301	ACCACATGGT	ATCTAATGTA	TTATTGGCTA	TCAGATATTT	AAGTTAGATA
351	**GATAGATAGA**	**TAGATACATA**	**GATAGATAGA**	**TGATGAGAGA**	**TAGATAGATA**
401	**GATAGATAGA**	**TGAATTAGAT**	**AGATGATGAG**	**AGATAGGATA**	**GATAGATAGA**
451	**TGATAGATAG**	**ATAGATGATG**	**ATAGATAGAT**	**AGATGATAGA**	**TGATAGATGA**
501	**TAGATAGATA**	**GATAGATAGA**	**TAGATGATAG**	**ATAGATAGG**	

Fig 3.21 Sequence Data Obtained From 37/68

Underlined sequence is the multiple cloning cassette sequence. Bold sequence is the highly repeated GATA sequence

Start	Stop	Phase	ORF Length (bases)	ORF Peptide length	ATG	length bases	peptide
326	538	2	213	71	365	178	58
340	537	1	198	66	-	-	-
417	539	3	123	41	-	-	-
134	250	2	117	39	197	39	18
175	267	1	93	31	-	-	-

Table 3.2a Open Reading frames of Transcribed 5'GATA3' Sequence

Start	Stop	Phase	ORF Length (bases)	ORF Peptide length	ATG	length bases	peptide
303	452	3	150	50	-	-	-
385	504	1	120	40	-	-	-
226	294	1	69	23	229	66	22
194	241	2	48	16	-	-	-
207	251	3	45	15	-	-	-

Table 3.2b Open Reading Frames of Transcribed 5'TATC3' Sequence

EMBL ref.	Homologous Sequence	% Identity	Overlap	OPT #	Region of Homology
HSTPA	Human tissue plasminogen activator (t-PA) gene	63	443	542	GATA
SKSEXSAT	Snake (*E. radiata*) w-chromosome sex-specific satellite DNA	54.6	542	472	GATA
RNRSBZ3	Rat repetitive sequence cluster, multiple "taga" copy region on 1.3kb *Eco* RI fragment	74	262	482	GATA
MMSQR3	Mouse simple repetitive DNA (sqr family) transcript (clone pmlc 31) with conserved GACA/GATA repeats	58.7	414	406	GATA
LGAB19	*Lemna gibba* chlorophyll a/b apoprotein gene	67.9	268	440	GATA
MMSQR4	Mouse simple repetitive DNA (sqr family) transcript	53.8	426	358	GATA
HSMGI1	Human myoglobin gene, exon 1	54.1	386	334	whole
MMPRPMPB	Mouse PRP gene encoding proline-rich protein MP-2	52.7	548	434	GATA
GMRDNA1	Tsetse fly rRNA with intergenic spacer and 183 rRNA gene 5'end	50.4	506	336	AT-rich
CHEGPSBA	*Euglena gracilis* chloroplast psbA locus with 32 kd protein gene (herbicide binding protein)	49.5	515	347	AT-rich
DDDISIC2	Slime mold (*D. discoideum*) discoidin-ic and id genes; 5' part id	52.2	460	258	AT-rich
GMGLY	Soybean glycinin subunit A-2B-1a gene	49.3	548	324	whole
MIDYRRN	*Drosophila yakuba* mitochondrial DNA molecule	53.3	518	412	AT-rich

Table 3.3 Results of EMBL database search using 539bp GATA sequence

EMBL ref.	Homologous Sequence	% Identity	Overlap	OPT #	Region of Homology
MMXX01	Mouse mRNA isolated by screening with snake satellite DNA	62.7	459	600	CTAT
XLB2VIT2	*Xenopus laevis* B2 vitellogenin gene (LF) intron 3	62.1	419	558	CTAT
DMRSXB	*D. melanogaster* Bkm-like DNA, proximal region X chromosome, 3' end, clone c314 2(8)	65.5	412	612	CTAT
MMRSY	Mouse sxr (Bkm-like DNA, sex-determining region of the Y chromosome	57.4	350	400	CTAT
HSIGK6	Human germline pseudogene for the leader peptide and variable region of a kappa immunoglobulin (subgroup V kappa I)	54.5	290	246	whole
HS7SKP41	Human small cellular 7SK pseudogene (clone 41)	60.4	273	322	CTAT
SCPMA1	Yeast PMA1 gene for plasma membrane ATPase	45.6	507	206	whole
RNMHCG	Rat embryonic skeletal muscle myosin heavy chain	49.8	414	250	CTAT
HSTGKOL	Human lys-tRNA, gln-tRNA, leu-tRNA genes and flanking	47.5	474	288	AT-rich
MISC37	Yeast mitochondrial genes for 15s rRNA and tRNA-Trp	50.2	506	340	whole
RNRRNA08	Rat rRNA spacer downstream of 28S rRNA gene 3'end	53.0	485	352	whole
MILTMXC1	*Leishmania tarentolae* maxicircle DNA fragment	48.7	505	314	whole
SVRATJ2	SV40/Fisher rat DNA junction, clone pEM5	50.0	504	316	whole

Table 3.4 Results of the EMBL database search using the complementary strand of 506bp sequence.

4.0 DISCUSSION

4.1 PRESENCE OF GATA SEQUENCES IN THE HORSE AND OTHER DOMESTIC ANIMAL GENOMES

By using a synthetic oligomer as a specific probe, the amount of GATA was calculated at approximately 1% of the genome. This figure is above that which would be expected if GATA sequences had occurred randomly (section 3.3.3).

There were large differences in the amount of GATA sequence between species including those which are closely related in evolutionary terms, such as horse and donkey, and sheep and goat. Very little GATA sequence was found in the sheep and cattle genomes, confirming the result of Miklos *et al* (1989).

Until recently the presence of these sequences in a wide range of organisms has been taken to indicate some evolutionarily conserved function. The absence of these sequences of any substantial length in the bovine and ovine genomes limits this theory. The presence of these sequences does not necessarily imply that they have a function.

The horse has been shown to contain large amounts of Bkm sequence on its Y chromosome, as well as chromosomes 3, 4 and 30 (Kent *et al*, 1988). The level of hybridization suggested, however, that there were fewer GATA repeats on the autosomes compared to the Y chromosome (Kent *et al*, 1988). But from my experiments there appears to be no noticeable difference in the level of hybridization between male and female horses (section 3.3.2). The probe I used was much shorter than Bkm and therefore probably hybridized to a greater number of GATA sequences over the whole genome complement. So like mice, the amount of GATA sequences on the autosomes may out number that which occurs on the Y chromosome (Schafer *et al*, 1986a).

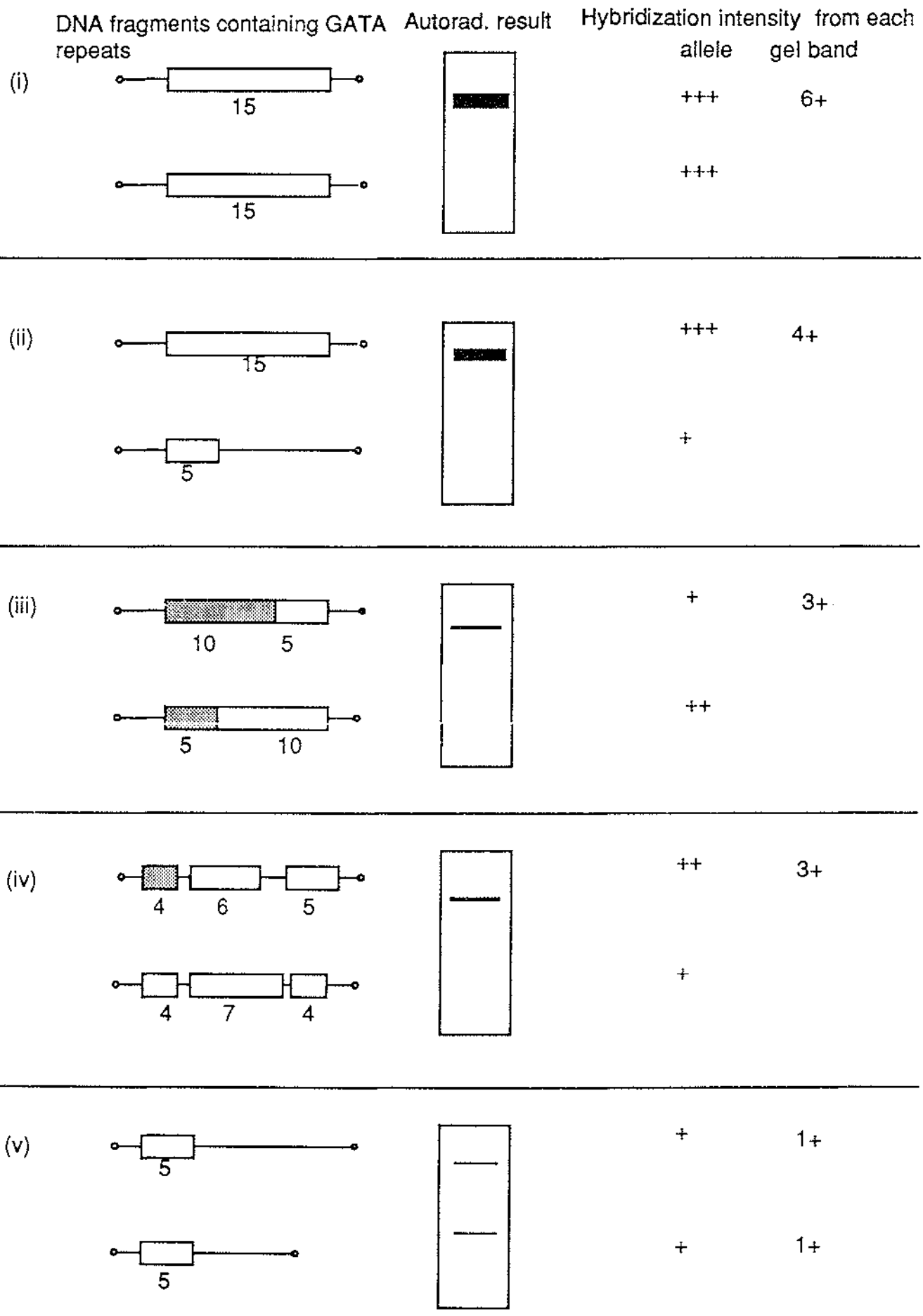


fig 4.1 Possible Explanations for Observed Autoradiograph Bands

4.2 GATA POLYMORPHISMS

A number of length polymorphisms were generated by digesting total genomic DNA to completion with restriction enzymes. These enzymes were chosen to cut outside the GATA sequence. The number and size of the length polymorphisms depended on the restriction enzyme used. The intensity of the polymorphisms also varied, reflecting the amount of probe binding to the DNA. For example digestion with *Mbo* I generated a large number of length polymorphisms (fig 3.5c). Nearly all these length polymorphisms were below 4 kb. Most likely there are a number of loci in the horse genome which contain GATA sequences.

Fig 4.1 summarizes the five possible explanations for the hybridization patterns of genomic digests. Two fragments of the same size may be derived from: homologous chromosomes; heterologous chromosomes; the same chromosome where the fragments may have been located next to each other as part of a larger tandem array or from dispersed loci within the chromosome. There is no way to determine from my results the possible chromosomal origin of the hybridizing fragments.

Both the length and the number of homologous sequences which hybridize to the GATA probe will determine the size and intensity of the band appearing on the autoradiograph. Fig 4.1, parts (i) and (ii) show how variation in the number of homologous sequences can affect the intensity of the autoradiograph result. The presence of more sequences will result in a more intense band on the autoradiograph. Part (iii) shows that although similar sequences may be present (represented by the shaded region) hybridization of the probe will not occur unless there is 100% (or very near) homology to the probe. Although the total number of homologous sequences present in a fragment may be substantial their distribution within the fragment will affect binding of the probe (part (iv)). The probe used in this investigation consisted of five repeats of the GATA sequence and would not bind to complementary sequences containing fewer repeats. Parts (i) to (iv) are examples of variable number tandem repeat (VNTR) polymorphisms.

Two homologous loci may differ with respect to the presence or absence of a restriction site and so the resulting fragment lengths differ (part (v)). As a result two different sized bands will appear in the autoradiograph. This is referred to as a restriction fragment length polymorphism (RFLP). An RFLP may also contain VNTR type variations.

In all cases in the present investigation the stringency of the hybridization and wash conditions affected the amount of probe which bound to the genomic DNA. At low stringency the probe bound to less homologous sequences, as comparison of figs 3.5b and 3.5c shows.

Although bands up to 4 kb were visible, it is unlikely that these DNA fragments were composed solely of GATA sequence, as the intensity of the band did not reflect intense levels of probe hybridization. To find out how GATA sequences were organized in the horse genome, a GATA positive piece of genomic DNA was isolated and analysed (see section 4.4).

4.3 APPARENT MATERNAL INHERITANCE OF CERTAIN "BANDS"

Several relatively dark bands compared to other bands in the same lane were present in individuals whose DNA had been digested with *Mbo* I. An unusual feature noted was that the dark bands which appeared in the dams examined had a corresponding band in their offspring. The dark bands of the sire did not appear in the offspring. This pattern occurred independent of the sex of the offspring.

Kent *et al* (1988) reported that the Y chromosome of horses contains long tracts of GATA; I therefore expected a strong hybridizing signal to have appeared from the male horses and this pattern to be repeated in their male offspring. This did not appear to occur. Perhaps these long Y chromosome-specific GATA sequences were interspersed with Mbo restriction sites (/GATC) and therefore no significantly long GATA sequences were present. However, dot blot evidence negates this possibility (fig 3.4, section 3.3.2), as this DNA was undigested.

The dark bands may have arisen by an imprinting (methylation) effect by the dam. A different banding pattern was observed with *Sau* 3A digested DNA (demonstrated in fig 3.12, lane 4). This enzyme is an isoschizomer of *Mbo* I, i.e. it recognizes the same sequence (GATC). However, with *Sau* 3A the sequence GmATC is cut, while GATmC is not. Mbo does not cleave at GmATC sites, while the effect of methylation of GATmC in not known (Anonymous, 1988).

Alternatively, these dark bands may have a mitochondrial origin. The method of DNA preparation involves the lysis of the cell membrane and collection of total genomic DNA and not only the isolation of nuclear DNA (section 2.3.1). Conceivably mitochondrial DNA could be purified along with nuclear DNA.

Mammals have small mitochondrial genomes. In humans, mice and cattle it is about 16.5kb. Each organelle appears to have between 1 to 10 genomes and each cell contains several hundred organelles (Lewin, 1986). Like bacterial genomes there is no non-coding DNA. Therefore, if the GATA probe is hybridizing to mitochondrial DNA it is probably also a functional gene. Analysis of five mouse cDNA sequences showed only one which had a long open reading frame which included $(GATA)_5$ tracts, whereas the other four cDNAs had frequent stop codons distributed throughout the cloned inserts (Miklos *et al*, 1989). Thus, the presence of GATA sequences in putative genes is not unknown.

4.4 GENOMIC ORGANISATION OF GATA SEQUENCES IN HORSE

Restriction mapping a positive genomic GATA sequence, (section 3.5.2), showed that GATA sequences occurred at a single region and were not distributed over the entire 3.6kb fragment (figs 3.15 and 3.16). This supports the hypothesis stated in section 4.2 that it was unlikely that the polymorphic bands were composed solely of GATA sequences. The concentration of GATA sequences resulted in a very strong hybridization signal appearing upon hybridization with the GATA probe.

Analysis of the 506 bases of 37/68 sequenced showed the organisation of the genomic GATA sequence appeared to be typical of dispersed, middle-repetitive sequences isolated from other organisms. This particular GATA genomic clone did not reveal GATA/GACA interspersion as has been reported by other researchers (eg Epplen, 1988). However, the presence of GAT interspersion with GATA sequences suggests that a slippage mechanism may have been involved in their generation (Jeffreys, 1987). The strong AT rich region upstream from the GATA sequence explains why the restriction enzyme *Alu* I was unsuitable for restriction mapping. This region contained a large number of Alu restriction sites.

Further analysis revealed the number of stop codons (ie TAA, TAG and TGA) present in the unsequenced complementary strand (containing 5'TATC3' sequences) was greater than that present in the sequenced strand. Open reading frames present in both the sequenced strand and its complement could be found using the program NLDNA. There were a greater number present in the sequenced 5'GATA3' strand. Tables 3.2a and 3.2b list the longest five ORFs found. Within the 5'GATA3' sequence two ORFs contained the initiation codon ATG. This is in agreement with the observation of Traut (1987) who found significant ORFs in cloned genomic Bkm sequences, all in the 5'GATA3' strand and repeated stop codons in the complementary 5'TATC3' strand.

Searching the EMBL database revealed a number of sequences which shared over 50% identity with the 506bp sequence. These data have been summarised in table 3.3. When the complementary strand was used a number of different homologous sequences appeared. These have been summarised in table 3.4. Both searches showed sequences homologous to the horse genomic clone present in a wide range of organisms such as yeast mitochondria, slime mold, Drosophila, amphibian, snake, mouse and human genomes. The convention now is for more weight to be given to the exon number and exon/intron boundaries when discussing the significance of homology searches as this is more in keeping with functional conservation of genes during evolution. Data from this study does

not allow any definitive comment to be made as to the possible function of the isolated GATA sequence.

The output from the homology search includes the percentage identity between the query sequence and homologous database sequence as well as the number of bases along which this homology occurs, and an optimum (OPT) number. This number indicates the significance of the homology between the query sequence and the database sequence in comparison to a match between two random sequences (Needleman and Wunsch, 1970). An optimum number above 150 indicates significant homology. Both searches had the "best" 13 homologous sequences with an OPT number above 150.

It is reassuring to note that high levels of identity were shown to occur between the horse genomic GATA sequence and the snake (*E. radiata*) sex-specific satellite DNA sequence isolated by Epplen *et al* (1982) and the mouse sequence isolated by Schaefer *et al* (1986). Both have conserved GACA/GATA repeats.

The presence of a yeast mitochondrial homologous sequence helps support the hypothesis that GATA sequences may occur in the horse mitochondrial genome and therefore could explain the apparent maternal inheritance pattern of some bands (section 4.3).

4.5 DNA FINGERPRINTING WITH GATA

A subsidiary aim of this study was to investigate whether GATA polymorphisms would be sufficient for Jeffreys-type DNA fingerprinting. I found that the $(GATA)_5$ probe did not show enough polymorphisms to be useful for a fingerprinting probe. Like all synthetic oligonucleotide probes JF-1 hybridizes to a very specific sequence. It may, in fact, be too specific. JF-1 could be used, however, to isolate a piece of genomic DNA, such as 37/68, which maybe a member of a polymorphic family. This piece of genomic DNA could act as a probe, and as such would be expected to hybridize to many related sequences which are interspersed in the

not allow any definitive comment to be made as to the possible function of the isolated GATA sequence.

The output from the homology search includes the percentage identity between the query sequence and homologous database sequence as well as the number of bases along which this homology occurs, and an optimum (OPT) number. This number indicates the significance of the homology between the query sequence and the database sequence in comparison to a match between two random sequences (Needleman and Wunsch, 1970). An optimum number above 150 indicates significant homology. Both searches had the "best" 13 homologous sequences with an OPT number above 150.

It is reassuring to note that high levels of identity were shown to occur between the horse genomic GATA sequence and the snake (*E. radiata*) sex-specific satellite DNA sequence isolated by Epplen *et al* (1982) and the mouse sequence isolated by Schaefer *et al* (1986). Both have conserved GACA/GATA repeats.

The presence of a yeast mitochondrial homologous sequence helps support the hypothesis that GATA sequences may occur in the horse mitochondrial genome and therefore explains the apparent maternal inheritance pattern of some bands (section 4.3).

4.5 DNA FINGERPRINTING WITH GATA

A subsidiary aim of this study was to investigate whether GATA polymorphisms would be sufficient for Jeffreys-type DNA fingerprinting. I found that the $(GATA)_5$ probe did not show enough polymorphisms to be useful for a fingerprinting probe. Like all synthetic oligonucleotide probes JF-1 hybridizes to a very specific sequence. It may, in fact, be too specific. JF-1 could be used, however, to isolate a piece of genomic DNA, such as 37/68, which maybe a member of a polymorphic family. This piece of genomic DNA could act as a probe, and as such would be expected to hybridize to many related sequences which are interspersed in the

genome producing a complex hybridization pattern (ie, a DNA fingerprint). This possibility remains to be tested.

If the GATA family of dispersed, middle-repetitive sequences is found to behave like mobile elements, their use for fingerprinting becomes questionable because of their lack of genomic stability.

It is debatable whether DNA fingerprinting of horses could be as successful as that in humans. Domestic horses have been selectively bred over the last 3000 years or more. The Thoroughbred originated from 3 stallions and approximately 100 mares (Wagoner, 1978) and have been selectively bred for the last 300 years. Compared to other domestic species, the thoroughbred is a closed population with a small genetic pool. Little variation has occurred at the DNA level and what there is has been homogenized via selective breeding. Well documented pedigrees show that the further back a pedigree is traced the greater the number of common ancestors appear between two apparently unrelated individuals. Inbreeding has occurred and continues to occur in the thoroughbred with the aim of reinforcing desired characteristics such as speed and endurance. The human population on the other hand has in the main actively outbred, creating much variation. This level of variation means that for fingerprinting purposes a single probe is more likely to show greater polymorphisms at loci in human populations than in horses.

5.0 CONCLUSION

GATA sequences were present in the horse genome, accounting for about 1% of the total genome. Multispecies studies revealed these sequences varied in amount between species, with much variation occurring between closely related species such as horse and donkey, sheep and goat. Very few GATA sequences occurred in the sheep and cattle genomes. The absence of substantial lengths of GATA sequences in these species limits the theory of an evolutionary conserved function for these sequences.

There were no quantitative sex-specific differences. Published results state there was a concentration of GATA sequences along the Y chromosome in horses. Perhaps the dispersed autosomal sequences far outnumber those which occur on the Y chromosome.

GATA sequences did contribute to DNA polymorphisms, especially when the restriction enzyme *Mbo* I was used, but the level of polymorphism was not enough for Jeffreys' type (multi-locus) DNA fingerprinting. A novel feature noted, however, was the apparent maternal inheritance pattern of some bands. A mitochondrial origin may explain this inheritance pattern. Homologous mitochondrial sequences were found following a database search.

A genomic fragment of DNA was isolated which contained GATA sequences. These were not distributed over the entire fragment but concentrated in one region, where they were arranged tandemly. This particular fragment showed GAT sequences interspersed amongst GATA. Computer-based searches of a nucleic acid sequence database revealed significantly homologous sequences from a wide range of organisms.

Further research of GATA sequence organisation in the horse genome is possible based on this investigation. The GATA positive clone, 37/68, needs further characterization - specifically sequencing past the GATA region into the 3' region. Analysis of the flanking regions may give some indication as to the possible function or origin of this DNA fragment. The presence of inverted repeats would add support to the hypothesis that GATA sequences belong to a group of mobile transposable elements. If the family of GATA sequences are mobile elements their use for fingerprinting would be severely limited due to their instability at genomic loci. But they could be useful for future manipulation of the horse genome by acting as vectors to introduce foreign DNA into the genome (eg. by homologous recombination) or act as markers for analysis of nearby loci.

Determination of the chromosomal location of GATA sequences in the horse genome would be interesting. Both the GATA oligomer and the isolated GATA genomic fragment could be used as probes. The problem with using the oligomer is that only a single radioactive molecule of ^{32}P can be attached to it. The exposure time for the hybridization therefore becomes extremely long. One way to get around this problem would be to ligate the oligomer to itself creating a much larger probe. This should be possible if a complementary oligomer was available, manufactured in such a way that its repeats were slightly out of phase with its complement. By allowing the oligomers to anneal together, double stranded oligomers would be created. These would possess sticky ends enabling them to be ligated into long tandem arrays. These could be radioactively labelled and be used as probes more efficiently than a single oligomer.

In situ hybridization with 37/68 would reveal whether this fragment is unique to a single locus, as such it would be a useful chromosome marker. Alternatively 37/68 may be a member of a multilocus family and may therefore prove useful as a fingerprinting probe if enough variation occurred amongst horses. Hybridizing 37/68 to a genomic DNA digest may reveal whether it is from a single locus or part of a multiloci family which may be polymorphic.

Northern blot analysis may show if this fragment is transcribed by hybridization to complementary RNA. Transcription may be the result of read-through from nearby genes, or the transcript may in fact be translated into some functional protein. The presence of ORFs, especially in the 5'GATA3' direction, may therefore be significant. The expected protein would be of simple structure due to the simple repeating sequence, and therefore may have more of a structural role than enzymatic.

GATA sequences may have developmental significance. During the horses lifecycle these sequences may be transcribed at specific times and/or in specific tissues. But, as Kirchhoff (1988) pointed out, stage- and tissue-specific differences in GATA transcription might also point to a "sequence dependent" function, or equally they might simply reflect the general differential gene activity of specialized tissues.

6.0 REFERENCES

ALI S, MULLER CR and EPPLEN JT (1986) DNA fingerprinting by oligonucleotide probes specific for simple repeats. Hum Genet 74:239-243

ALI S and WALLACE BR (1988) Intrinsic polymorphism of variable number tandem repeat loci in the human genome. Nuc Acids Res 16:8487-8496

ARNEMANN J, JAKUBICZKA S, SCHMIDTKE J, SCHAFER R and EPPLEN JT (1986) Clustered GATA repeats (BKm sequences) on the human Y chromosome. Hum Genet 73:301-303

BECKMANN JS and SOLLER (1987) Molecular markers in the genetic improvement of farm animals. Biotechnol 5:573-576

BIRD AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends In Genet 3:342-347

BLACKBURN EH and SZOSTAK JW (1984) The molecular structure of centromeres and telomeres. Ann Rev Biochem 53:163-194

BUCKLAND RA and ELDER JK (1986) On the mechanism of amplification of satellite II DNA sequences of the domestic goat (*Capra hiraus*). J Mol Biol 186:13-23

CASKEY CT (1987) Disease diagnosis by recombinant DNA methods. Science 326:1223-1229

CHANDRA HS (1985) Sex determination : A hypothesis based on noncoding DNA. Proc Natl Acad Sci USA 82:1165-1169

de la CHAPELLE A (1988) Invited editorial: The complicated issue of human sex determination. Am J Hum Genet 43:1-3

DURBIN EJ, ERICKSON RP and CRAIG A (1989) Characterization of GATA/GACA-related sequences on proximal chromosome 17 of the mouse. Chromosoma 97:301-306

EPPLEN JT, McCARREY JR, SUTOU S and OHNO S (1982) Base sequence of a cloned snake W-chromosome DNA fragment and identification of a male-specific putative mRNA in the mouse. Proc Natl Acad Sci USA 79:3798-3802

EPPLEN JT, STUDER R and McLAREN A (1988) Heterogeneity in the *Sxr* (sex-reversal) locus of the mouse as revealed by synthetic GATA/CA probes. Genet Res Camb 51:239-246

EPPLEN JT (1988) On simple repeated GAT/CA sequences in animal genomes: A critical reappraisal. J Hered 79:409-417

FANNING TG (1987) Origin and evolution of a major feline satellite DNA. J Mol Biol 197:627-634

FORREST JW, (1988) Oligonucleotide ^{32}P end-labelling. Personal communication.

FOWLER JCS, BURGOYNE LA, SCOTT AC and HARDING HWJ (1988) Repetitive DNA and Human Genome variation. In press (personal communication)

FOWLER C, DRINKWATER R, SKINNER J and BURGOYNE L (1988) Human satellite III DNA and example of a 'macrosatellite' polymorphism. In press (personal communication)

GADI IK and RYDER OA (1983) Molecular cytogenetics of the equidae. Cytogenet Cell Genet 35:124-130

HAMADA H, PETRINO MG, and KAKUNAGA T (1982) A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc Natl Acad Sci USA 79:6465-6469

HARDMAN N (1986) Structure and function of repetitive DNA in eukaryotes. Biochem J 234:1-11

HIGUCHI R, van BEROLDINGEN CH, SENSABAUGH GF and ERLICH HA (1988) DNA typing from single hairs. Nature 332:543-546

ISH-HOROWICZ D and BURKE JF (1981) Rapid and efficient cosmid cloning. Nuc Acids Res 9 (13):2989-2998

JEFFREYS AJ, WILSON V and THEIN SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67-73

JEFFREYS AJ, WILSON V and THEIN SL (1985) Individual specific fingerprints of human DNA. Nature 316:76-79

JEFFREYS AJ, BROOKFIELD JFY and SEMEONOFF R (1985b) Positive identification of an immigration test case using human DNA fingerprints. Nature 317:818-819

JEFFREYS AJ (1987) Highly variable minisatellites and DNA fingerprints. Biochem Soc Trans 15:309-317

JEFFREYS AJ, WILSON V, KELLY F TAYLOR BA and BULFIELD G (1987) Mouse DNA fingerprint analysis of chromosome localisation and germ line stablility of hypervariable loci in recombinant inbred strains. Nuc Acids Res 15:2823-2836

JEFFREYS AJ, ROYLE NJ, WILSON V and WONG Z (1988) Spontaneous mutation rates to new length alleles of VNTRs in human DNA. Nature 322:278-281

JELINEK WR and SCHMID CW (1982) Repetitive sequences in eukaryotic DNA and their expression. Ann Rev Biochem 15:813-844

JONES KW and SINGH L (1985) Snakes and the evolution of sex chromosomes. Trends in Genet 1:55-61

KENT MG, ELLISTON KO, SHROEDER W, GUISE KS and WACHTEL SS (1988) Conserved repetitive DNA sequences (Bkm) in normal equine males and sex-reversed females detected by *in situ* hybridization. Cytogenet. Cell Genet 48:99-102

KIEL METZGER K and ERICKSON RP (1984) Regional localization of sex-specific Bkm-related sequences on proximal chromosome 17 of mice. Nature 310:579-581

KIRCHHOFF C (1988) GATA tandem repeats detect minisatellite regions in blowfly DNA (Diptera: Calliphoridae). Chromosoma 96:107-111

KIYAMA R, MATSUI H and OISHI M (1986) A repetitive DNA family (Sau 3A) in human chromosomes - extrachromosomal DNA and DNA polymorphism. Proc Natl Acad Sci USA 83:4665-4669

KIYAMA R, OKUMURA K, MATSUI H, BRUNS GAP, KANDA N and OISHI M (1987) Nature of recombination involved in excision and rearrangement of human repetitive DNA. J Mol Biol 198:589-598

KODAMA H, SAITOH H, TONE M, KUHARA S, SAKAKI Y and MIZUNO S (1987) Nucleotide sequences and unusual electrophoretic behavior of the W chromosome-specific repeating DNA units of the domestic fowl, *Gallus gallus domesticus*. Chromosoma 96:18-25

LEWIN B (1985)Genes II. John Wiley & Sons New York.

LEWIN R (1986) DNA fingerprints in health and disease. Science 233:521-522

MANIATIS T, FRITSCH EF, and SAMBROOK J (1982) Molecular cloning: A laboratory manual, Cold Spring Harbor laboratory, Cold Spring Harbor, New York.

McLAREN A (1988) Sex determination in mammals. Trends in Genet 4:153-157

MIKLOS G.G, MATTHAEI KI and REED KC (1989) Occurrence of the $(GATA)_n$ sequence in vertebrate and invertebrate genomes. Chromosoma 98:194-200

MISHELL BB and SHIIGI SM (1980) Selected methods in cellular immunology. P23 Lysis of red blood cells with tris-buffered ammonium chloride. Published by W.H. Freeman and Company, New York.

MITCHELL AR, GOSDEN JR and MILLER DA (1985) A clones sequence p82H of the alphoid repeated DNA family found at the centromeres of all human chromosomes. Chromosoma 92:369-377

MORTON DB, YAXLEY RE, PATEL I JEFFREYS AF, HOWES SJ, and DEBENHAM PG (1987)Use of DNA fingerprint analysis in identification of the sire. Vet Rec 121:592-593

NAKAMURA Y, JULIER C, WOLFFE R, HOLM T, O'CONNEL P, LEPPERT M and WHITE R(1987) Characterization of a human 'minisatellite' sequence. Nuc Acids Res 15:2537-2547

NAKAMURA Y, LEPPERT M, O'CONNEL P, WOLFFE R, HOLM T, CULVER M, MARTIN C,FUJIMOTO E, HOFF M, KUMLIN E and WHITE R (1987b) Variable number tandem repeat markers for human gene mapping. Science 325:1616-1622

NANDA I, NEITZEL H, SPERLING, STUDER R and EPPLEN JT (1988) Simple GAT/CA Repeats characterize the X chromosomal heterochromatin of *Microtus agrestis*, European field vole (Rodentia, Cricetidae). Chromosoma 96:213-219

NEEDLEMAN SB and WUNSCH CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443-453

OHNO S and EPPLEN JT (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc Natl Acad Sci USA 80:3391-3395

PLATT THK and DEWEY MJ (1987) Multiple forms of male-specific simple repetitive sequences in the genus Mus. J Mol Evol 25:201-206

PLATT THK and DEWEY MJ (1989) Variable evolutionary stability of Y chromosomal repeated sequences in the genus Mus. Genet Res Camb 53:87-93

PROSSER J, FROMMER M, PAUL C ands VINCENT PC (1986) Sequence relationships of three human satellite DNAs. J Mol Biol 187:145-155

REED KC and MANN DA (1985) Rapid transfer of DNA from agarose gels to nylon membranes. Nuc Acids Res 13(20):7207-7221

ROGERS J (1984) Origin and evolution of retroposons. Int Rev Cytol 93:187-279

ROGERS JH (1985) Long interspersed sequences in mammalian DNA. Properties of newly identified specimens. Biochim Biophys Acta 824:113-120

RYDER OA, EPEL NC and BENIRSCHKE K (1978) Chromosome banding studies of the Equidae. Cytogenet Cell Genet 20:323-350

SANGER F, COULSON AR, HONG GF, HILL DF and PETERSEN GF (1982) Nucleotide sequence of bacteriophage λ DNA. J Mol Biol 162:729-773

SCHAFER R, ALI S and EPPLEN JT (1986a) The organization of the evolutionarily conserved GATA/GACA repeats in the mouse genome. Chromosoma 93:502-510

SCHAFER R, BOLTZ E, BECKER A, BARTELS F and EPPLEN JT (1986b) The expression of the evolutionarily conserved GATA/GACA repeats in mouse tissues. Chromosoma 93:496-501

SINGER MF (1982) SINES and LINES: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:233-234

SINGER MF and SKOWRONSKI J (1985) Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends Biochem Sci 10:119-122

SINGH L, PHILLIPS C and JONES KW (1984) The conserved nucleotide sequences of Bkm, which define Sxr in the mouse, are transcribed. Cell 36:111-120

SINGH L, PURDOM IF, and JONES KW (1981) Conserved sex-chromosome-associated nucleotide sequences in eukaryotes. Cold Spring Harbour Symp Quant Biol 45:805-814

SINGH L, PURDOM IF and JONES KW (1980) Sex chromosome associated satellite DNA: Evolution and conservation. Chromosoma 79:137-157

SMITH C and SIMPSON SP (1986) The use of genetic polymorphisms in livestock improvement. J Anim Breed Genet 103:205-217

SOUTHERN EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517

STEPHAN W (1987) Quantitative variation and chromosomal location of satellite DNAs. Genet Res Camb 50:41-52

STEPHENSON EC, ERBA HP and GALL JG (1981) Histone gene clusters of the Newt Notophthalmus are separated by long tracts of satellite DNA. Cell 24:639-647

TAUTZ D and RENZ M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nuc Acids Res 12:4127-4138

TONE M, SAKAKI Y, HASHIGUCHI T, and MIZUNO S (1984) Genus specificity and extensive methylation of the W chromosome-specific repetitive DNA sequences from the domestic fowl, *Gallus gallus domesticus*. Chromosoma 89:228-237

TRAUT W (1987) Hypervariable Bkm loci in a moth, *Ephestia kuehniella*: Does transposition cause restriction fragment length polymorphism? Genetics 115:493-498

TYLER-SMITH C and BROWN CRA (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. J Mol Biol 195:457-470

VASSART G, GEORGES M, MONSIEUR R, BROCAS H, LEQUARRE AS and CHRISTOPHE D (1987) A sequence in M13 phage detects hypervariable minisatellites in human and animal DNA. Science 235:683-684

WAGONER D (1978) Ed. Equine genetics and selection procedures. Published by Equine Research Publications, Texas.

WETHERALL JD, GROTH DM and CARRIK MJ (1988) Hypervariable markers associated with repetitive DNA sequences. Proceedings of the gene mapping workshop, Mt Victoria, Australia.

WEISING K, WEIGAND F, DRIESEL AJ, GUNTER K, ZISCHLER H and EPPLEN JT (1989) Polymorphic simple GATA/GACA repeats in plant genomes. Nuc Acids Res 17:10128

WONG Z, WILSON V, PATEL I, POVEY S and JEFFREYS AJ (1987) Characterisation of a panel of highly variable minisatellites clones from human DNA. Ann Hum Genet 51:269-288

WU JC and MANUELIDIS L (1980) Sequence definition and organisation of a human repeated DNA. J Mol Biol 142:363-386