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	midisatellites 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.

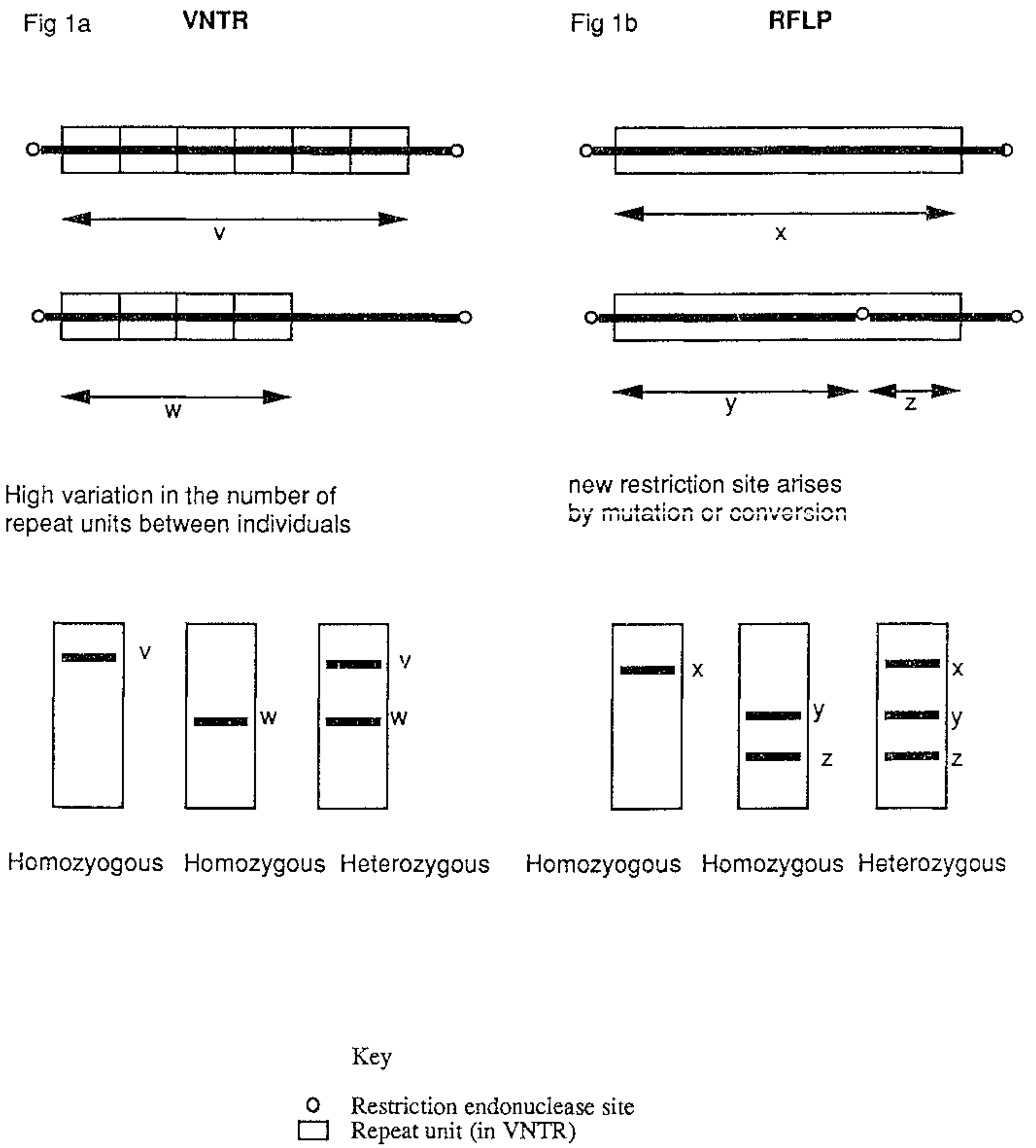
Fig 1a
VNTR
v
w
High variation in the number of repeat units between individuals
v
w
v
w
Homozyogous
Homozygous
Heterozygous
Fig 1b
RFLP
x
y
z
new restriction site arises by mutation or conversion
x
y
z
x
y
z
Homozyogous
Homozygous
Heterozygous
Key
O Restriction endonuclease site
Repeat unit (in VNTR)

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?