A Study of the Organisation of the Chicken Genome

Charlotte Kate Bruley

A thesis submitted in partial fulfilment of the requirements of the University of Edinburgh for the degree of Doctor of Philosophy

This programme of research was carried out at the Roslin Institute and Centre for Genome Research, University of Edinburgh.

September 1999

Φορ ϑερεμψ, α τηεσισ νοτ ωριττεν ιν χραψον

Declaration

I declare that the work presented in this thesis is my own, except where otherwise stated. All experiments were designed by myself, in collaboration with my supervisors, Dr. David Burt and Dr. John Mullins. No part of this work has been or will be, submitted for any other degree, diploma or qualification.

Charlotte Kate Bruley

September 1999

Acknowledgements

I would like thank all the people in my lab for their help: Andy, Bob, David, Anne, Cheryl, Dawn, Frazier, Pam, Simon, Stuart and Linda. Special thanks should go to Jackie who not only had the honour/misfortune of sharing a bench with me but also read my thesis. Thanks also to Stuart, Lisa, Helen Sang and Alan Archibald for reading my thesis and making helpful comments. Help was also given at various times by many different people from the Roslin Institute library, computing department and the Helen Sang and Mike Clinton labs. Pat Hope did much of my PhD. administration stuff at the CGR and Sheila Strachan organised my travel arrangements.

I would like to thank Nat Bumstead who provided the DNAs analysed in Chapter 3. David Waddington advised me on my t-tests and calculations while overlooking my dislike of statistics. Aedin Culhane and Nadia Anwar's e-mails and phone calls were invaluable while getting to grips with CLUSTAL and PHYLIP.

Thanks for everything's go to Kate, Terry, Sharon, Paul, Dale, Zack, Johanne, Ashley, James, Sophie, Kay, Becks, Jonathan, Kai and Antonio. David F. and Kat supported me through the last few years by numerous visits to the bingo hall. I am grateful to Mari, who fed me and told me that I spent too much time in the lab. Also, thanks to Alison who kept me entertained with stories about P. and to Liz who managed to fill my flat full of Germans during the festival.

My supervisor John Mullins was very helpful and gave me the opportunity to give regular talks at the CGR. I would like to thank David Burt for all his help and support he has given me as my supervisor. Also for the opportunity to visit Washington. I enjoyed the meeting almost as much as the shopping I did afterwards.

Abstract

Avian genome mapping efforts have concentrated on the domestic chicken (*Gallus gallus*) due to its economic importance and as a model of vertebrate development. Its karyotype consists of 6 large macrochromosomes and 33 small microchromosomes and its genome is 1.2 x 10^9 bp in size. Previous work has shown that microchromosomes are GC rich and CpG island rich whilst macrochromosomes are AT rich and CpG island poor. This suggests that the microchromosomes, though small, are more gene dense than macrochromosomes. Indeed I found that microchromosomes were approximately twice as gene dense as macrochromosomes.

The distribution of the avian retrotransposon repeat family Chicken Repeat 1 (CR1) was used to test the hypothesis, which considers if microchromosomes are gene dense, they should have fewer mobile repeats because they contain less non-essential DNA. A two-fold difference in CR1 repeat density macrochromosomes Vs microchromosomes was observed. This, combined with the numbers of CR1s calculated, supported the hypothesis. In addition to this, new CR1 subfamilies were assigned.

Conservation of the linkage group 5' *TH-INS-IGF2* 3' in chickens was also studied. Work was carried out to determine the order of these genes. The conserved linkage of the genes *TH* and *INS* was established but their order could not be determined.

Table of Contents **Page No.**

List of Figures

List of Tables

Abbreviations

A	Adenine
AMP	Ampicillin
APS	Ammonium persulfate
BAC	Bacterial artificial chromosome
BDH	Hydroxybutyrate dehydrogenase gene
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BSA	Bovine serum albumin
C	Cytosine
cDNA	Complementary DNA
CIP	Calf intestine phosphatase
CLAS	Classical
cM	Centimorgan
CNS	Central nervous system
CR1	Chicken repeat 1
DDBJ	DNA Database of Japan
DNA	Deoxyribonucleic acid
EDTA	Ethylene diaminetetra acetic acid
EL	East Lansing
EMBL	European Molecular Biology Laboratory
EST	Expressed sequence tag
FISH	Fluorescence *in situ* hybridisation
g	Gram
G	Guanine
GPI	Glycosyl phosphatidylinositol
GRAIL	Gene Recognition and Analysis Internet Link
HCR	Highly Conserved Region

IPTG	Isopropylthio-β-Dgalactoside
kb	Kilobase
Kd	Kilodaltons
L	Litre
LINE	Long interspersed nuclear element
LTR	Long terminal Repeat
M	Molar
mA	Milliamps
MAC	Macrochromosome
Mb	Megabase
mg	Milligram
MIC	Microchromosome
MICR	Microsatellite
μg	Microgram
μl	Microlitre
μM	Micromolar
ml	Millilitre
mM	Millimolar
mRNA	Messenger RNA
Mya	Million years ago
NCBI	National Centre for Biotechnology Information
nm	Nanomole
O.D.	Optical density
ORF	Open reading frame
PAC	P1-derived artificial chromosome
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
pg	Picograms

PRINS	Primed *in situ* labelling technique
QTL	Quantitative trait loci
RAPD	Random amplified polymorphic DNA
RNA	Ribonucleic acid
rpm	Revolutions per minute
RFLP	Restriction fragment length polymorphism
SDS	Sodium dodecyl sulfate
SINES	Short interspersed nucleotide elements
SSC	Sodium chloride sodium citrate
SNP	Single nucleotide polymorphism
STE	Sodium chloride Tris EDTA
STS	Sequence tagged site
SSCP	Single-strand Conformation Polymorphism
T	Thymine
TAE	Tris-acetate/EDTA electrophoresis buffer
TBE	Tris-borate/EDTA electrophoresis buffer
TE	Tris-EDTA buffer
TEMED	N,N,N',N'-Tetramethylethylenediamine
URL	Universal resource locator
UV	Ultra violet light
V	Volts
WL	White Leghorn
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside
YAC	Yeast artificial chromosome

Chapter 1

Introduction

1.1 Introduction

Avian genome mapping efforts have concentrated on the domestic chicken (*Gallus gallus*) due to its importance in meat and egg production and because of its utility as a model of vertebrate development. The first genetic linkage map of the chicken was reported by (Hutt, 1936) and was the first map of its kind to be reported for a livestock species. This was followed by linkage maps for cattle, sheep and pigs (Beattie, 1994). Since this early map, the chicken map has been updated at various intervals. For example, the chicken linkage map by (Bumstead and Palyga, 1992) was the first molecular map of a livestock species that attempted to address the whole genome. An international collaboration is currently underway to establish a comprehensive molecular map of the chicken genome (Bitgood and Somes Jr, 1993), (Burt *et al.*, 1995), (Burt *et al.*, 1997) and (Burt and Cheng, 1998).

1.2 Trait Gene Identification

There are a number of reasons for carrying out genetic mapping of a livestock species. Firstly, to understand the genetic control of economically important, biologically significant traits. Animal welfare and health can be improved by using marker assisted selection to eliminate disease. In addition rare breeds can be preserved, aspects of speciation and evolution can be investigated and models for human disease developed. Economically important genes which control traits such as growth rate can also be isolated (Beattie, 1994) and (Edwards, 1994). These traits are quantitative in nature i.e. there is continual variation of discrete traits, sex or colour and are located at quantitative trait loci (QTL). QTL mapping locates the region of interest in order to act as a launch pad for a gene hunt (Figure 1.1), and can be done using a number of different approaches. Mapping is also carried out in order to study single gene defects, an example being the mapping of the autosomal dwarf locus in chicken (Ruyter-Spira *et al.*, 1998). Genetic, cytogenetic and physical maps

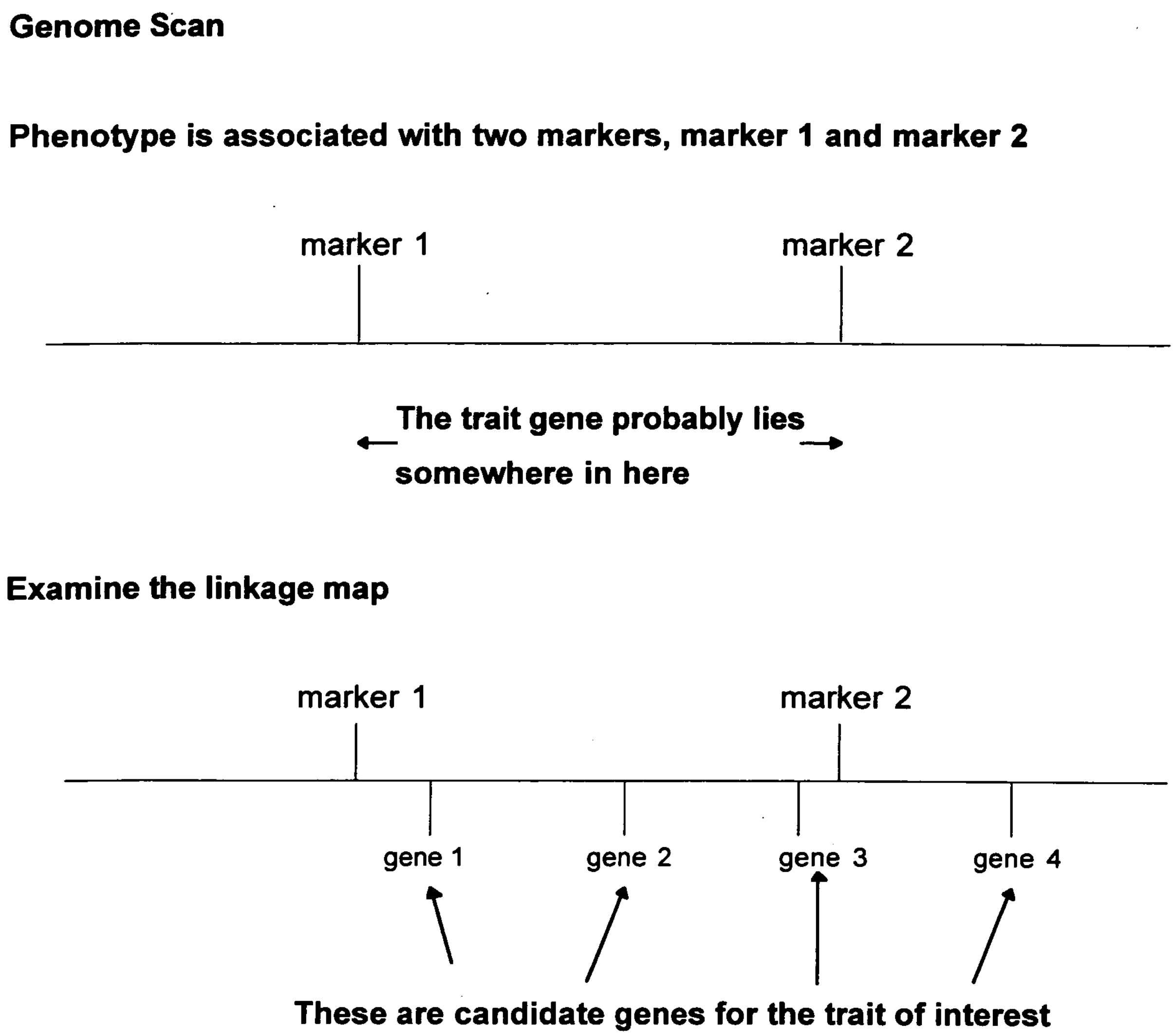

Genome Scan
Phenotype is associated with two markers, marker 1 and marker 2
marker 1
marker 2
The trait gene probably lies somewhere in here
Examine the linkage map
marker 1
marker 2
gene 1
gene 2
gene 3
gene 4
These are candidate genes for the trait of interest

Figure 1.1 Trait Gene Identification

A phenotype is associated with two markers and the area is examined for candidate genes.

are used to identify and isolate genes of interest. A brief description of each type of map follows:

Genetic maps study the linkage relationships of gene loci (linkage analysis) and are based on genetic markers. A cytogenetic map is based on the physical location of probes to chromosomes e.g. genes and is usually based on fluorescence *in situ* hybridisation (FISH). A more detailed physical map can be constructed by ordering overlapping cloned DNA fragments into contigs. The most detailed physical map is the complete nucleotide sequence (Fries, 1993) and (Smith *et al.*, 1994).

1.2.1 Gene Identification Without Map Information

Gene identification without prior map information relies on having information about previously isolated genes. For instance, when trying to isolate a disease gene, only a partial knowledge of the disease is enough to make an educated guess about which gene is the cause. An example of this was the work which showed that the familial cancer syndrome of Li and Fraumeni is caused by a germline missense mutation in the p53 genes (Ballabio, 1993; Collins, 1995). In chickens this approach has been used to isolate the *dw* gene which encodes a growth hormone receptor. A recessive mutation in this gene affects growth regulation and development and is responsible for sex-linked dwarf chickens (Burnside *et al.*, 1991); (Hull *et al.*, 1993); (Burnside *et al.*, 1992).

1.2.2 Functional Cloning

Functional cloning describes the identification of a trait gene without any knowledge of its chromosomal location but knowing its protein product, for example its amino acid sequence, or its function (Ballabio, 1993; Collins, 1995). This involves the detection and analysis of gene transcripts (Chen *et al.*, 1994) and to achieve this, cDNA libraries are screened using suitable probes. A second possible strategy is to use the polymerase chain reaction (PCR) to amplify cDNA with oligos from protein sequences. Genes isolated using this approach include the β-globin gene associated with Sickle Cell Anaemia (Ballabio, 1993).

1.2.3 Gene Identification With Map Information

To isolate and identify the gene or genes controlling a particular genetic trait the first step is to construct a low resolution map to estimate their number and location. Once this information has been gathered, a higher resolution map using DNA markers flanking the trait gene(s) can be developed. The gene(s) can then be identified by physically cloning the region and identifying genes within it. An alternative to this is to propose candidate genes based on genes previously mapped to this region or predictions based on comparative maps.

1.2.4 Positional Cloning

The isolation of a gene by its map position, without any knowledge of its function, is known as positional cloning (Collins, 1995) and (Parimoo, 1995). Methods used include the identification of expressed sequences and chromosome walking. Positional cloning is used when information about the biochemical or the molecular basis of the trait is not available. This approach has been effective in identifying a wide range of genes such as the cystic fibrosis gene and the genes involved with Beckwith-Weidemann syndrome. Notably, however, this approach is not as applicable to livestock species as the existing genetic maps are less dense (Mannens *et al.*, 1996), (Ballabio, 1993), (Parimoo, 1995) and (Womack, 1998). Also the cystic fibrosis gene was cloned from a region of 500,000 bp, whereas when searching for a particular QTL a region of 20-40 Mb may have to be searched.

1.2.5 Positional Candidate Gene

The positional candidate gene method of gene isolation is a combination of functional and positional cloning using knowledge of gene function and a genetic/physical map (Collins, 1995). This approach has been successful in isolating disease genes for familial hypertrophic cardiomyopathy (Ballabio, 1993). The positional candidate gene approach can be carried out in two possible ways.

The first approach is used when a new disease locus is assigned to a chromosomal region. This is analysed for candidate genes and their features are compared with the features of the disease. These include any biochemical defects, imprinted inheritance, anticipation, developmental defects and an animal model. The second approach uses information from a gene isolated in one species to isolate it in another species by sequence comparisons. This involves a new gene being assigned to a region of a chromosome by examining sequence domains, expression patterns, imprinted expression, sequence instability, developmental expression and conservation of synteny across species (Ballabio, 1993).

An example of this approach is the high-resolution mapping of sequence tagged sites (STS) and expressed sequence tags (EST) loci on human chromosome 2p13.3 and corresponding markers on mouse chromosome 11. These regions contain known genes and the mouse wobbler (*wr*) region. The detailed mapping of this region has narrowed down the *wr* region thus allowing the selection and exclusion of positional candidate genes for *wr* (Resch *et al.*, 1998).

1.2.6 Comparative Candidate Positional Gene

A positional candidate gene approach relies on having a detailed gene map. This can be a problem for livestock maps which have few loci associated with functional genes (Archibald, 1998). Functionally important sequences such as exons are more likely to be conserved across species than introns, interspersed repeat sequences or intergenic regions (Sedlacek *et al*., 1993). The comparative candidate positional gene approach uses comparative maps to predict candidate genes. This approach takes advantage of the evolutionary history of chromosomes and the detailed human and mouse maps available by transferring information from a 'map-rich' species to 'map-poor' species (Archibald, 1998) and (Womack, 1998). In livestock the comparative candidate positional gene strategy has proved effective in the study of malignant hyperthermia and the genetic control of colour in pigs (Archibald, 1998).

1.2.7 Genetic Mapping

Genetic maps summarise the linkage relationship of gene loci (linkage analysis) and are based on genetic markers. The markers are placed relative to each other and the order determined by recombination. The unit of recombination is known as the centimorgan (cM). Two markers are one cM apart if they recombine in meiosis once every 100 opportunities they have to do so. 1 cM is equivalent to approximately 1 Mb in human, 2 Mb in mouse and 0.3 Mb in chicken (Smith and Burt, 1998).

Genetic maps are used to locate genes or "trait-genes" which control traits of interest e.g. disease resistance and to provide information on genome structure. This information can be used as a resource for genetic analysis and to help study how genomes evolved (Patterson, 1995; Weeks, 1995). On these maps the markers are placed relative to each other, their order being determined by recombination.

Complex or polygenic traits are governed by multiple genes, a notable human example of a polygenic trait disease being Type 1 diabetes (Patterson, 1995). Polygenic traits are of major interest in livestock species as they govern economically important traits (Edwards, 1994). Genetic linkage maps can be used to locate regions of the genome that control these traits of economic importance, most of which are quantitative in nature. Genes controlling these traits are located at

QTLs and the mapping of them has been made easier by having dense maps with a high number of polymorphic markers (Beattie, 1994).

1.2.8 Genetic Markers

To create a comprehensive genetic linkage map, polymorphic marker loci which are evenly distributed throughout the genome are required. A marker must be polymorphic i.e. it must exist in different forms (alleles) enabling the chromosomes carrying the different alleles to be distinguished. Two classes of marker loci, Type I and Type II are defined (O'Brien, 1991). Table 1.1 outlines a number of DNA-based markers used in the construction of maps and their applications and Table 1.2 the genetic markers used on the chicken map.

1.2.9 Type I Markers

Type I markers are coding gene loci which tend to be conserved across species making them ideal markers for the construction of comparative maps (O'Brien *et al.*, 1993) and (O'Brien *et al.*, 1997). Type I loci are less useful than Type II markers when mapping trait genes because they are not, or are only slightly polymorphic. They are, however, useful when identifying the boundaries of conserved synteny and facilitate the identification of candidate genes when a particular trait gene locus has been assigned to a chromosome region.

1.2.10 Type II Markers

Type II markers are highly polymorphic DNA segments which are often species specific, and are useful when mapping trait genes and for the construction of

Marker Type	Acronym	Alias	Required	Major Use
Restriction length polymorphism	RFLP		Cloned DNA	Linkage mapping
Simple sequence repeats	SSR	Microsatellite	Cloned DNA and/or sequence	Linkage mapping
Sequence-tagged site	STS		DNA sequence	Physical mapping
Expressed sequence tag	EST		cDNA sequence	Physical and linkage mappin
Random amplified polymorphic DNA	RAPD		Short oligonucleotide primer	Fingerprinting
Variable number of tandem repeats	VNTR	Minisatellite	Repetitive sequence hybridisation probe	Fingerprinting
Chicken repeat 1-based marker	CR1		CR1 oligonucleotide primer set	Linkage mapping
Amplified fragment length polymorphism	AFLP		Designed oligonucleotide primer set	Fingerprinting, linkage mappi

Table 1.1 Types of DNA-Based Markers

(Dodgson *et al.*, 1997)

Marker Type	**Total**
SINES or CR1 repeats	45
Endogenous retroviruses	37
VNTRs or minisatellites	52
Random genomic clones	109
RAPDs	68
Classical genes	10
cDNAs of unknown function	37
cDNAs of known function	92
Microsatellites associated with genes	43
Total number of genes	201
Anonymous microsatellites	1020
Total number of microsatellites	1063
Total number of markers	1513

Table 1.2 Genetic Markers on the Chicken Map

CR1- Chicken Repeat 1; VNTR- Variable number of tandem repeat; RAPD-Random amplified polymorphic DNA; SINES- Short interspersed elements; (Burt and Cheng, 1998) linkage maps (Crooijmans *et al.*, 1993). Using the chicken linkage maps as an example (Table 1.2), common Type II markers include Chicken Repeat 1 (CR1), short interspersed repeats (SINES), restriction fragment length polymorphisms (RFLPs), minisatellites and simple sequence repeats such as microsatellite markers. Microsatellites are tandem repeats of one to six bp motifs with variations in the number of repeats of a simple sequence i.e. $CA_{(n)}$. They are a popular research tool due to being widely distributed in eukaryotic genomes, multi-allelic, co-dominant and assayable by PCR. The automated typing of large populations can be carried out using microsatellites (Burt and Cheng, 1998) and (Dodgson *et al.*, 1997).

Type II markers are of less use in comparative studies as they are poorly conserved between distantly related species. They can be used in linkage mapping between closely related species. For example, there is a close evolutionary relationship between sheep and cattle, which has led to cattle microsatellites being used in the construction of a sheep linkage map (Broad *et al.*, 1998) and (Marshall Graves, 1998).

A recent example of type II markers being used to characterise another avian genome is the development of a genetic map of the turkey genome. To characterise the genome and to aid in the construction of genetic maps, turkey genomic libraries enriched with the TG, GAT, and CCT simple repeats have been produced (Huang *et al.*, 1999).

1.2.11 Genetic Linkage Analysis

Recombination occurs during meiosis when pairs of homologous chromosomes come together and exchange segments. The further a marker is from a gene, the greater the chance of recombination between them. Linkage analysis compares within a family, the inheritance of a marker with the inheritance of DNA markers of a known chromosomal location. The co-inheritance of a gene and a marker suggests that they are physically close on the chromosome. Two probability calculations are made, the first to find the probability that the observed inheritance pattern has occurred by chance and is completely unlinked. The second probability calculation assumes there is linkage between the two. The ratio of the two probabilities is calculated, and expresses the odds for and against that degree of linkage. The logarithm of the ratio is called the logarithm of the odds or the lod score. A lod score equal to, or greater than, 3 strongly suggests that two markers are linked and programs such as Map Manager (Manly, 1993) can be used to carry out these calculations.

1.2.12 Mapping the Chicken Genome

The chicken genome is being mapped to identify genetic markers linked to traits of economic value, to find animal models of human disease, and to aid the study of the evolution of the vertebrate genome. To achieve this, large pedigrees and crosses between lines which show extreme differences in phenotypes are required. When linkage is established candidate genes can then be identified.

1.2.13 International Reference Crosses

1.2.14 The Compton Cross

With the aim of producing a linkage map to help identify and locate genes affecting salmonella disease resistance a reference mapping population, called the Compton Cross, was developed by Nat Bumstead and Jan Palyga at the Institute for Animal Health in the UK. This was a backcross between two inbred White leghorn lines which differed in disease resistance. A single F_1 female was backcrossed to a line 15I male. The Z chromosome could not be mapped with this reference family due to a female being used in the backcross. The mapping panel was formed with

DNA from 56 individuals and the map constructed using RFLP and microsatellite markers (Bumstead and Palyga, 1992).

1.2.15 The East Lansing Cross

The East Lansing reference family was developed by Lymen Crittenden at Michigan State University in the USA and is a backcross between a single female from the highly inbred UCD-003 White Leghorn (WL) line and a male from the partially inbred UCD-001 Red Jungle Fowl line. Two of the F_1 males were backcrossed with 10 and 8 UCD-003 WL females to generate 208 and 192 progeny, respectively. A mapping panel of 52 backcross progeny (1 F_1 male x 4 WL females) has been produced. The initial map consisted of RFLP, random amplified polymorphic DNA markers and chicken repeat element 1 markers (Crittenden *et al.*, 1993). Microsatellite markers have subsequently been mapped.

1.2.16 The Wageningen Cross

With approximately fifty offspring each, the East Lansing and Compton reference mapping populations are small and microsatellite coverage of the maps is not complete. The Wageningen reference family, has been produced by Martien Groenen (University of Wageningen, The Netherlands) and the poultry breeding company Euribrid. It comprises of a cross of two commercial broiler lines. A total of 430 microsatellite markers were analysed and a linkage map developed. The Wageningen map has markers in common with the maps based on the Compton and East Lansing reference populations, which is of use when drawing data together to generating a consensus map (Groenen *et al.*, 1998).

1.3 Databases

Molecular biology and genome research has benefited from the collection of data in public domain databases. Large scale sequencing efforts have generated a large amount of data which is analysed to identify coding sequences and to search for sequence similarities (Parimoo, 1995). Databases can be accessed by the Internet and contain data from large and small genome projects (Gelbart, 1998). Often a database is a simple way of finding an up to date copy of a map of a particular species or to isolate a potential candidate gene.

Genome databases can be divided into two groups, generalised and specialised. Generalised databases archive nucleic acid sequences such as the GenBank/EMBL/DDBJ or polypeptide sequences such as PIR and SwissProt (Gelbart, 1998). Gene homologies from sequence data are found using sequence alignment tools, such as the Basic Local Alignment Search Tool (BLAST) program (Altschul *et al.*, 1990). Two types of searches can be carried out using the general databases outlined in Table 1.3. The first, BLASTN compares a nucleotide query sequence against nucleotide sequence databases. The second, BLASTX, examines a sequence for the presence of protein coding regions against databases of protein sequences (Gish and States, 1993).

Specialised genome databases, Table 1.4, are available for a variety of mapping projects including human, mouse and livestock species. Each database often documents a specific model organism or a particular biological function, such as protein family databases (Gelbart, 1998). Table 1.4 also outlines comparative databases such as the Comparative Mapping Home Page, (La Trobe University, Australia) and the Livestock Animal Genome Databases, (Roslin Institute, Scotland). These document homologous genes and their chromosomal location among mammalian species (Nadeau *et al.*, 1995).

Program	Database	Searches
BLASTN Compares a nucleotide query sequence against a nucleotide sequence database	nr	All non-redundant GenBank +DDBJ sequences (but no EST's or STS's)
	gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences
	dbEST	Non-redundant database of GenBank+EMBL+DDBJ EST Divisions
	dbSTS	Non-redundant database of GenBank+EMBL+DDBJ STS divisions
	TDB	Provides access to cDNA/EST sequence and related data from The Institute for Genomic Research (TIGR), Human Genome Sciences (HGS) and world wide EST projects.
BLASTX Compares a nucleotide query sequence translated in all frames against a protein sequence database	nr	All non-redundant GenBank+EMBL+DDBJ sequences (but no EST's or STS's)

Table 1.3 Sequence Databases Used During BLAST Searches
EST-Expressed sequence tagged site; STS-Sequence tagged site; nr-Non-redundant; EMBL- European Molecular Biology Laboratory; DDBJ- DNA database of Japan ;TDB-TIGR database ;db- database

Species	Database Name and Institution	Universal Resource Locator (URL)
Chicken	Chicken genome database, Arkdb-CHICK	http://www.ri.bbsrc.ac.uk/chickmap/
Human, mouse	Genome Database (GDB)	http://gdbwww.gdb.org/gdb
Human	UniGene and Gene Map '98	http//www.ncbi.nlm.nih.gov
Human	Online Medline Inheritance in Man (OMIM)	http://gdbwww.gdb.org/omim/docs/omimtop.html
Mouse	Mouse Genome Database (MGD)	http://www.informatics.jax.org
Fugu	Fugu genome Project, Welcome Trust, Genome Campus, Cambridge, UK	http://fugu.hgmp.mrc.ac.uk
Comparative	Comparative Mapping Home Page, La Trobe University, Australia	http://www.latrobe.edu.au/www/genetics/compmap.html
Comparative, cat, cattle, chicken, pig, sheep, horse deer, turkey	Livestock Animal Genome Databases, Roslin Institute, Scotland, UK	http://www.ri.bbsrc.ac.uk/genome_mapping.html
C.elegans	ACeDB C.elegans database, The Sanger Centre, Hinxton Hall, Cambridge, UK	http://www.sanger.ac.uk/Projects/C_elegans

Table 1.4 Genome Databases on the Internet
(Wakefield, *et al.*, 1998)

continued

Table 1.4 continued

Species	Database Name and Institution	Universal Resource Locator (URL)
Cattle	Institut National de Recherche Agronomique, jouy-en-Josaas, Laboratoire de Genetique Biochimique, France	http://locus.jouy.inra.fr/cgi-bin/bovmap/intro.pl
Cattle, pig	USDA Meat Animal Research Centre Clay Centre, Nebraska, USA	http://sol.marc.usda.gov/marc/html/gene1.html
Dog	Dog Genome Project, University of Michigan, USA	http://www.msu.edu/`K9genome-index.html
Drosophila	FlyBase (Drosophila Mapping database) Harvard University Cambridge, USA	http://flybase.bio.indina.edu/
Marsupials Monotremes	Roobase (Marsupial Genome Database) La Trobe University, Australia	http://www.latrabe.edu.au/www/genetics/roobase.html
Rat	RATMAP database, Goteborg University, Sweden	http://ratmap.gen.gu.se
Shrew	International Sorex araneus Cytogenetics Committee	http://meiosis.bionet.nsc.ru/isacc.html
Sheep	Sheepbase, Roslin Institute, Scotland, UK	http://www.ri.bbsrc.ac.uk/sheepmap/
Vertebrate animal species other than	OMIA (Mendialian Inheritance in Animals) University of Sydney, Australia	http://www.angis.su.oz.au/Databasers/BIRX/omia/
Zebrafish	Institute of Neuroscience, University of Oregon, USA	http://zfish.uoregon.edu/

1.4 Model Species For Genome Studies

Model species are used as a tool to isolate genes and determine their function when direct study in a particular species (e.g. human) is difficult because of complexity and ethical considerations. The functional analysis of gene loci is greatly simplified by comparative studies in other species. In addition, model species can be used as a means of developing techniques which are required to isolate genes in other species. General characteristics of a model species include being well investigated, possessing a reasonably sized, compact genome, plus knowledge of genetic and comparative maps to facilitate positional cloning. Other requirements include having similar methods of gene replication, recombination and control of gene expression, plus showing structural and functional homologies with other genomes (Levy, 1994). Several model species for genome studies include mouse, the puffer fish *Fugu rubripes*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans* and livestock species such as the chicken.

The mouse has become an important species for the modelling of human disease and studying gene function (Brown, 1994) and its genetic map is complete (Dietrich *et al.*, 1996) i.e. a new genetic marker has a greater than 95% chance of linkage to another-previously-mapped marker. Advantages of this species include being able to control breeding, the number of different coat colours, anatomic and behavioural variants and the availability of inbred strains (O'Brien *et al.*, 1993) and (Dietrich *et al.*, 1995). It is also possible to generate gene knockouts in the mouse (Dietrich *et al.*, 1995).

Comparisons of human and mouse maps have shown areas of conserved synteny and gene order (DeBry and Seldin, 1996). This information plus the large number of genes mapped in mouse make it an ideal species for comparative studies. New human genes can be isolated by locating them in the mouse first (Brown, 1994). An example of this is the discovery of several loci predisposing to diabetes which were first mapped in mice and then in humans (Edwards, 1994).

The puffer fish *Fugu rubripes* has a similar gene collection to the human genome (Elgar *et al.*, 1996). As there is a lack of traditional linkage studies with interspecific backcross analysis, use of *Fugu* as a model species may appear limited. Studies have concentrated on DNA and RNA levels, and examing comparative gene structure and organisation (Elgar and Clark, 1998). The haploid genome size of *Fugu* is approximately 400 Mb and contains little repetitive DNA, smaller introns, a high gene density, and no pseudogenes or dispersed repeats. These features allow cross species hybridisation between *Fugu* and more complex genomes, making it an ideal system for the discovery of genes. However, *Fugu* and mammalian species are evolutionary some 420 million years apart, which can cause problems with this method of gene isolation. Unless a sequence is highly conserved, it might not always be possible to isolate a gene by sequence hybridisation (Elgar, 1996) and (Elgar and Clark, 1998). Also, many fish are polyploid or are derived from polyploid ancestors.

The bakers yeast *Saccharomyces cerevisiae* has been used by geneticists as a model for higher eukaryotes and basic biological research. It easy to manipulate, and has an efficient, compact genome of 12052 kb with its genes arranged on 16 chromosomes. A great effort was made to sequence its genome which is now complete (Butler, 1996) and the experience gained can be used in other sequencing projects. The analysis of the complete sequence is underway and will serve as a model for post-sequencing studies for other organisms (Dujon, 1996).

From the sequence, the ultimate anatomy of the genome can be examined. In yeast, features such as average G+C content, gene density, pseudogenes, open reading frames (ORFs), length of intergenic regions, number/length of introns and conservation of intron-exon-junction sequences have all been studied. The yeast genome was found to contain genes which have no clear function or sequence homologies with any other organism (including yeast). These genes are known as orphans and make up approximately 30-35% of ORFs in the yeast genomes. To determine if these are real genes, a combination of prediction by computer and gene knockout experiments need to be carried out (Dujon, 1996).

An example of analysis of the *S. cerevisiae* sequence is a sequence similarity search between yeast and human sequence. This was carried out to identify yeast homologs of human disease-associated genes and to understand their function. A total of 31% of the human disease-associated genes analysed had homology to yeast genes and it was suggested that the catalytic domains of the yeast and human proteins have identical functions (Foury, 1997).

The free-living nematode *Caenorhabditis elegans* acts as a model organism in the study of higher eukaryotes and in cell biology, cell death, development, neurobiology, sex determination and genetics (Chalfie and Jorgensen, 1998), (Metzstein *et al.*, 1998) and (Wilson, 1999). Advantages of *C. elegans* as a model include being well characterised, its small size, rapid generation time and the generation of gene knockouts (Chalfie and Jorgensen, 1998). *C.elegans* has a small, compact genome of approximately 97 Mb which has been completely sequenced. This is the first multicellular organism to be sequenced in its entirety, and this work has served as a pilot study for the human genome project. For example, software tools developed for the *C. elegans* project have been applied to human genome sequencing. The completed *C. elegans* sequence has revealed 18, 000 predicted protein coding genes and approximately 1000 RNA genes. The identification and analysis of their roles is now underway (Hodgkin and Herman, 1998) and (Wilson, 1999).

1.4.1 Chicken as a Model Vertebrate

The chicken is a good model species for a number of reasons. Advantages of the chicken for gene mapping include their nucleated red blood cells which enable large amounts of DNA to be extracted from small samples. They possess a relatively small genome of approximately 1.2 x 10^9 bp, which is one third the size of the human genome. This suggests that the chicken genome is three times more compact and could be easily sequence sampled (Burt *et al.*, 1995), (Bumstead and Palyga, 1992) and (Holden, 1996).

Sequence sampling (see also section 2.17, Materials and Methods) is a simple method of constructing high resolution physical maps from regions of genomic DNA (Smith *et al.*, 1994). Clones covering the region of interest are isolated and partially sequenced. The DNA sequence is analysed to identify gene homologies, repetitive elements and putative intron-exon boundaries (Smith *et al.*, 1994).

Using the genetic linkage map, conserved syntenic groups and conserved gene order between avian and mammalian species have been identified, allowing comparative mapping experiments to be carried out. In addition the chicken can also act as a model for the molecular basis of vertebrate limb development. Mutations such as *wingless, limbless* and *talpid 3* (*ta^3*) have been studied (Burt *et al.*, 1995) and (Johnson and Tabin, 1997).

1.5 Avian Genome Size

Avian species tend to have smaller genomes than those of amphibians, reptiles or mammals and it has been suggested that avian genome size reduced over a long period of time due to a series of deletions in the introns, see Figure 1.2. It has been hypothesised that the metabolic demands flight puts on a bird has placed a restraint on genome size (Hughes and Hughes, 1995). Bats also have a small genome size compared to other non-flying mammals. Both birds and bats have a high metabolic rate and a small genome (Burton and Bickham, 1989) and (Van Den Bussche *et al.*, 1995). It is interesting to note that among flightless birds genome size tends to be larger than in flying birds (Hughes and Hughes, 1995). For example the Turtle Dove has a mean DNA content of 2.46 picograms, the domestic chicken one of 2.47 and Jackass Penguin 3.26 picograms (Tiersch and Wachtel, 1991).

The genome size (nuclear DNA content) of 135 bird species was analysed. The genome size was found to be the most conservative and uniform of any vertebrate class. This suggests that bird genomes have evolved from a small ancestral

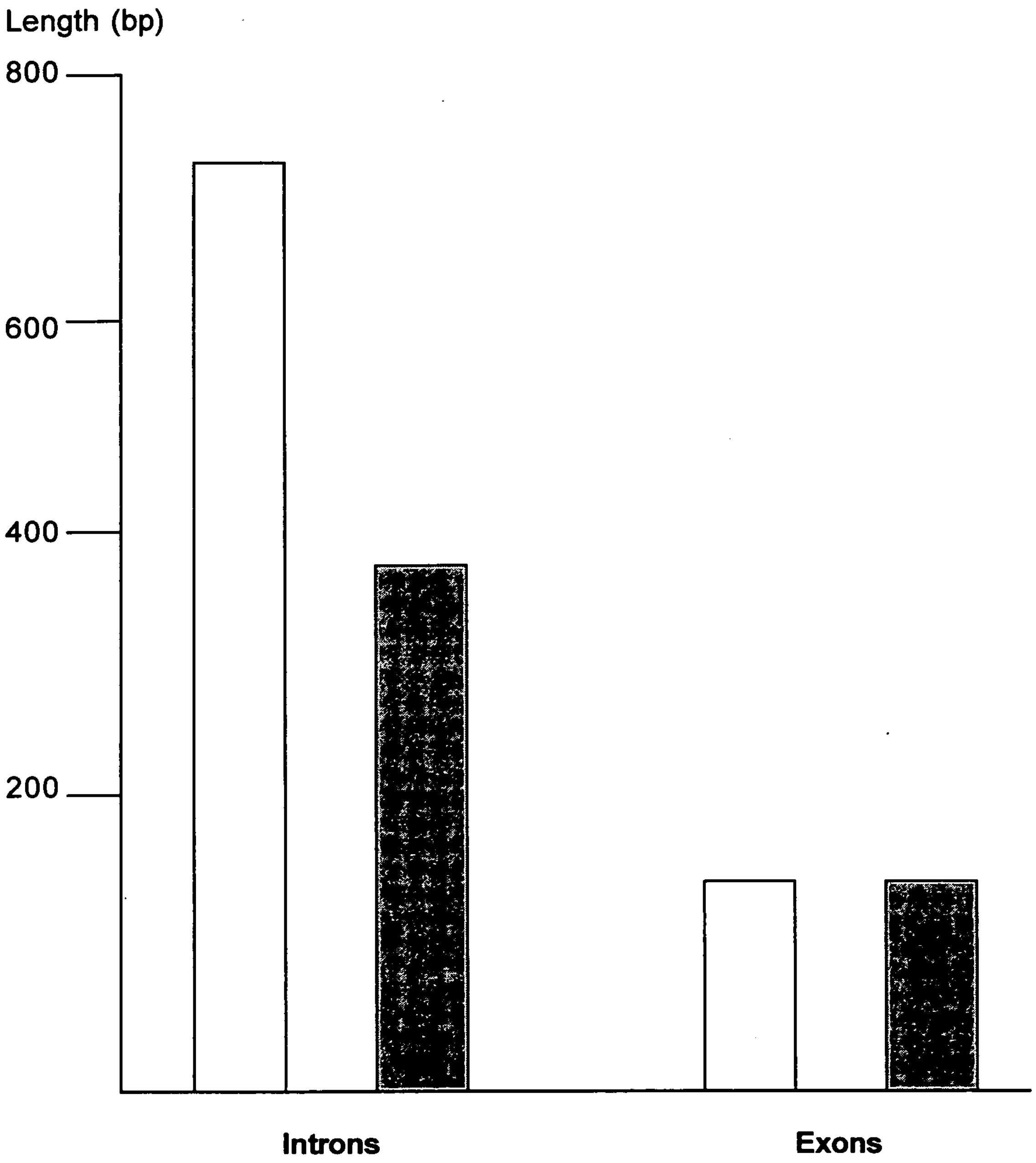


Figure 1.2 The Lengths of Human and Chicken Exons and Introns

The mean lengths of 111 homologous introns and 141 homologous exons from 31 genes of human (open columns) and chicken (grey columns) are described. Human introns are longer than their chicken homologues but exons are only marginally larger than their chicken counterparts. The difference is not statistically significant (Hughes and Hughes, 1995).

genome that was reduced before the emergence of the primitive bird ancestor (Tiersch and Wachtel, 1991).

1.6 Avian Karyotypes

Avian chromosomes are divided, by size, into two groups called macrochromosomes and microchromosomes. The macrochromosomes are similar in size to human chromosomes and are easily distinguishable from each other by their morphology and specific banding patterns. The microchromosomes are smaller and are maintained at a constant number, with telomeric sequences. They are cytologically indistinguishable from each other and are barely visible under a light microscope, making classification impossible (Carlenius, 1981), (Fritschi and Stranzinger, 1985) and (Hinegardner, 1976). Microchromosomes are observed in lower numbers than avians in vertebrates such as fish but disappear in mammals and may be evidence of them being ancestral chromosomes (Fillon, 1998) and (Rodionov, 1996).

In birds the sex chromosomes, Z and W, are in reverse with respect to mammals with the male being the homogametic sex, ZZ, and the female heterogametic, ZW (Crawford, 1990) and (Fritschi and Stranzinger, 1985).

The karyotypes of 234 bird species were compared and the average number of diploid chromosomes was found to be 80, comprising of 8 pairs of macrochromosomes and 32 pairs of microchromosomes (Tegelstrom and Ryttman, 1981). Table 1.5 summarises the diploid number of chromosomes for different species of birds. Exceptions to this are the Falconiformes, especially the Accipitridae family, which have three to six microchromosome pairs (De Boer and Sinoo, 1984).

1.7 Repetitive DNA Sequences in Avian Genomes

Repetitive sequences are dispersed throughout the avian genome, but in general, the percentage of repeats in birds is low compared to mammals. One repeat found

Common Name	Diploid Number of Chromosomes	Sex Chromosomes
Chicken	78	Z,W
Turkey	80	Z,W
Japanese Quail	78	Z,W
Mallard Duck	78	Z,W
Pheasant	82	Z,W
Falconiformes:		
Lanner Falcon	52	Unknown
Kestrel	52	ZW
Buzzard	68	ZW
Parakeet	58	ZW
Ratitae:		
Ostrich	80	No Heteromorphism
Cassowary	80	No Heteromorphism
Emu	80	No Heteromorphism
Rhea	82	ZW

Table 1.5 The Diploid Number of Chromosomes For Different Bird Species

throughout the avian genome, are short interspersed DNA elements called chicken repeat 1 (CR1). First isolated in chickens, they are widely distributed in vastly differing avian genomes such as emu, crane and chicken (Chen *et al.*, 1991). CR1 repetitive elements belong to the class of non-long terminal repeat retrotransposons, which also includes L1 elements in mammals (Vandergon and Reitman, 1994). In addition to avian CR1 repeats, reptilian homologs have been discovered, indicating they occurred some time before the divergence of reptile and bird lineage's.

Minisatellites in the chicken genome, numbering 50-100,000, are not evenly distributed and are generally unlinked because of high recombination rates in and around them (Bruford and Burke, 1994). The most common marker on the East Lansing map is the microsatellite. Clustering of microsatellites has been observed and they appear to be more frequent on the microchromosomes and may be a reflection of this type of repetitive element having a tendency to locate near the centromeres and telomeres which make up a larger percentage of microchromosomal DNA or an artefact. CA microsatellites have been found to be infrequent on the Z chromosome and moderately infrequent on the microchromosomes (Primmer *et al.*, 1997). The number of CA microsatellites is 10-fold lower than that found in mammals (Crooijmans *et al.*, 1993) and (Primmer *et al.*, 1997).

1.8 Comparative Mapping

Information for comparative mapping is generated from the genetic maps of many different species. Many chromosomal sequences are highly conserved over a wide range of species and comparative maps between species are informative in various ways. They can be used to understand chromosomal rearrangements that have occurred during the divergence of mammalian lineage's (Eppig, 1996). Conserved chromosomal segments can be used to predict linkages, identify candidate genes and to study vertebrate genome organisation and evolution (Nadeau and Sankoff, 1998).

1.8.1 Comparative Mapping in Chickens

It is now possible to compare the chicken genome maps with the more complete maps of mouse and human (Andersson *et al.*, 1996), (Burt *et al.*, 1995; Burt *et al.*, 1997). Comparisons can be used to find candidate genes. For example, the *NRAMP1* gene forms part of a conserved syntenic group containing at least 24 expressed loci and is defined by the genes *Col3a1* and *Col6a3*. This syntenic group has been found in mouse (chromosome 1), human (chromosome 2q35), sheep (chromosome 2q41), cattle (chromosome 2) and rat (chromosome 9). The chicken homolog of the *NRAMP1* gene has been assigned to a linkage group found on chromosome 7, which also contains at least four other genes belonging to the syntenic group from human and mouse. In the conserved region in chicken, the order of the genes was not completely conserved. This suggests that some kind of rearrangement has occurred (Girard-Santosuosso *et al.*, 1997). Further examples of conservation observed between chicken, mouse and humans are described in Table 1.6 and describes regions of conserved synteny and gene order between human, mouse and chicken. Conserved gene order can be observed with the genes *TCP1, IGF2R, VIP, ESR, MYB, PLN* and *FYN* which are all located on chromosome 3 in chickens and chromosome 6 in humans.

1.8.2 Conservation of Synteny

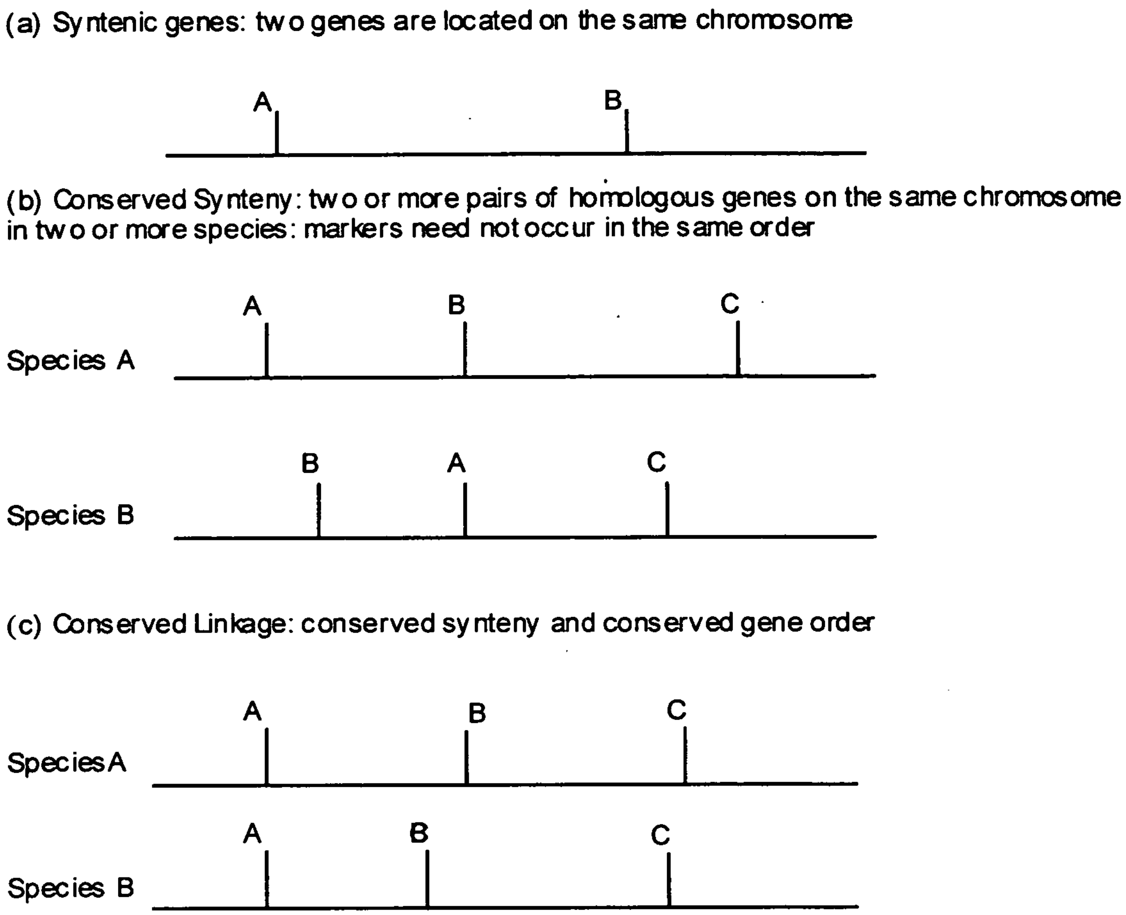
Two genes are said to be syntenic if they are located on the same chromosome (see Figure 1.3 a). When the same set of genes are grouped together on one chromosome in one species and the same group are found together on a chromosome in another species this is known as conserved synteny (see Figure 1.3 b and Table 1.6). Conserved linkage, Figure 1.3 c, occurs when the order of genes or markers is the same in two or more species (also see Table 1.6).

Though birds diverged from mammals 300-350 million years ago (Mya) (Nanda *et al.*, 1999), conserved blocks of synteny between human and chicken chromosomes

Locus	Chicken	Human	Mouse
BMP2	3	20p12	2 76.0
TGFB2	3 q22-q23	1q41	1 101.5
ACTN2	3	1q42-q43	13 7.0
HMX1	3	4p16.1	5 13.1
T	3	6q27	17 4.0
TCP1	3	6q25-q27	17 7.5
IGF2R	3	6q25.3	17 7.4
VIP	3	6q24-q27	10 S[a]
ESR	3	6q25.1	10 12.0
MYB	3 q24-26	6q23.3-q24	10 16.0
PLN	3	6q22.1	10 S[a]
FYN	3	6q21	10 25.0
CCNC	3 q27-q29	6q21	10 S[a]
ME1	3	6q12	9 48.0
EEF1A1	3	6q14	4 S[a]
BMP5	3	6q12-q13	9 42.0
GSTA2	3	6p12	9 43.0
ODC1	3	2p25	12 6.0
MYCN	3	2p24.3	12 4.0

Table 1.6 Examples of Conserved Synteny, Segments and Gene Order Between Chicken, Human and Mouse

S[a] mapped using synteny data



(a) Syntenic genes: two genes are located on the same chromosome

(b) Conserved Synteny: two or more pairs of homologous genes on the same chromosome in two or more species: markers need not occur in the same order

(c) Conserved Linkage: conserved synteny and conserved gene order

Figure 1.3 Synteny and Linkage Conservation

have been identified. Regions of human chromosomes 6, 4 and 9 were found to be homologous to regions on chicken chromosomes 3, 4 and Z, respectively (Chowdhary *et al.*, 1998). The largest region of conserved synteny observed is between the chicken Z chromosome and human chromosome 9, with 11 out of 18 genes on the Z having orthologs on chromosome 9. The order of these genes has been altered by an inversion (Nanda *et al.*, 1999).To aid the chicken comparative mapping effort thirty new type I markers have been assigned by FISH (Sazanov *et al.*, 1998). A total of 16 mapped to a macrochromosome and 28 to a microchromosome. Their assignment extended known syntenic groups on four of the macrochromosomes.

1.8.3 Homologs, Paralogs and Orthologs

Different types of molecular homology have been described. The first are homologs, which are genes descended from a common ancestor; for example all the globin genes. When deciding if two genes are homologous, features such as shared functions, sequence, similar expression, conserved map position, subcellular location and substrate specificity should be considered (Andersson *et al.*, 1996). These are important when considering ancestral genome duplications, transpositions and subsequent functional divergence. Sequence similarity, though important, can be misleading because of shared motifs, pseudogenes and related gene family members (Eppig, 1996).

Orthologous genes, such as the human β- and chimp β-globin genes, have diverged from a common ancestral gene after speciation events. Homologous genes which undergo gene duplication events such as tetrapliodization and regional duplications prior to divergence give rise to paralogous genes. Paralogous genes, an example of which are the β- and γ-globulin genes, make up many gene families and superfamilies (Eisen, 1998) and (Lundin, 1993).

1.9 The Chicken Genome and Karyotype

Table 1.7 outlines the general characteristics of the chicken genome. Its karyotype consists of macrochromosomes (chromosomes 1-5 and sex chromosome Z) and microchromosomes (chromosomes 6-8, the remaining chromosomes and the W sex chromosome). Figure 1.4 describes the karyotype (Bloom *et al.*, 1993). Chromosomes 1-8 and the sex chromosomes, Z and W have been distinguished by GTG- and RBG-banding studies carried out by the International Committee for the Standardisation of the Avian Karyotype (Ladjali *et al.*, 1999). The remaining chromosomes remain indistinguishable by this method. The physical lengths and DNA content of the eight largest chromosomes are summarised in Table 1.8 (Smith and Burt, 1998). An approach using FISH with cDNA, yeast artificial chromosome (YAC) or cosmid probes specific to microchromosomes could solve the problem of microchromosome identification. Sixteen microchromosomes have been identified with two-colour FISH experiments (Fillon *et al.*, 1998).

1.10 Recombination on Macrochromosomes and Microchromosomes

In chicken oocytes, the frequency of chiasmata on the macrochromosomes is linearly dependent on chromosome length. This was not observed for the microchromosomes, with one to two chiasmata on the larger microchromosomes (nos. 6-10) and one chiasma in the remaining chromosomes. The ZW sex chromosomes were also found to have one chiasma. Compared to human spermatocyte chromosomes, macrochromosome crossing over was found to occur twice as often and was even higher in the microchromosomes. A high frequency of exchanges in the microchromosomes has been observed in plants and animals and is related to each chromosome requiring at least one chiasma for homologous chromosomes to separate correctly in anaphase. For crossing-over in microchromosomes, there must be an increase in the frequency of

Microchromosomes	Macrochromosomes
31 microchromosome	8 macrochromosomes
Size: 1/10 the size of macrochromosomes	Size: Close to human
30% of genome[a]	70% of genome[a]
Hard to distinguish by morphology and banding patterns	Easy to distinguish by morphology and banding patterns
Have telomeres and centromeres[b]	Have telomeres and centromeres[b]
Maintained at a constant number	Maintained at a constant number
GC rich?	AT rich?
CpG island rich?	CpG island poor?
Gene rich?	Gene poor?

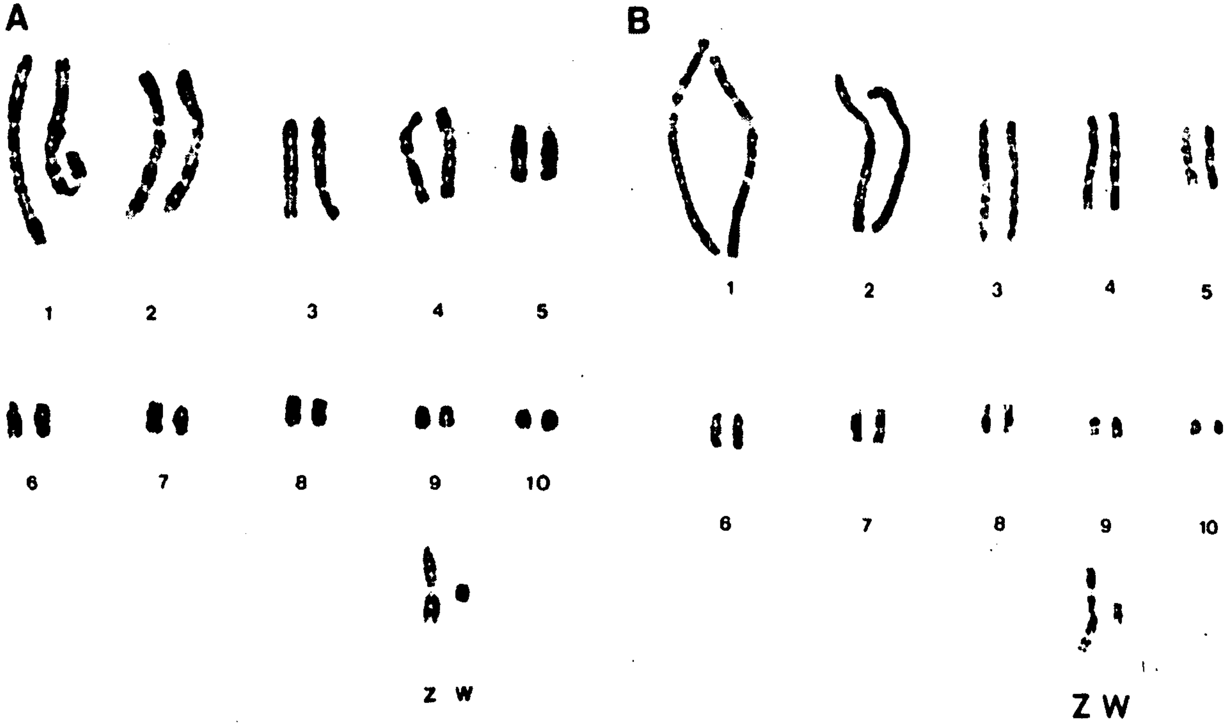
Table 1.7 Characteristics of Chicken Macrochromosomes and Microchromosomes

[a] (Smith and Burt, 1998); [b](Rodionov, 1996); The chicken genome is 1.2 billion bp in size, comprising of an estimated 60-100,000 genes (Burt *et al.*, 1995). The GC content is 42% and the O/E CpG ratio 0.25. The major repeats are the CR1 elements (Vandergon and Reitman, 1994) and the repeat fraction is 12% (Bloom *et al.*, 1993).

Figure 1.4 The Chicken Karyotype

Clear R- and G- banding patterns can be observed in all macrochromosomes and microchromosomes to size number 10. (A) RBG-banded karyotype for pairs 1 to 10 and sex chromosomes; (B) GTG-banded karyotype for pairs 1 to 10 and sex chromsomes.

Figure 1.4



Chromosome	Area %[a]	DNA content[b] ± se (%)	Physical length (Mb)	Measured length ± se (arbitrary units)
1	20.5	20.8 ± 0.8	250	1.14 ± 0.07
2	12.8	15.1 ± 0.3	181	0.84 ± 0.04
3	9.0	11.5 ± 0.3	138	0.63 ± 0.02
4	7.1	9.1 ± 0.3	109	0.48 ± 0.02
5	5.8	5.3 ± 0.2	64	0.34 ± 0.02
6	3.2	3.5 ± 0.1	42	0.24 ± 0.01
7	3.2	3.4 ± 0.2	41	0.22 ± 0.01
8	1.9	2.5 ± 0.1	30	0.20 ± 0.01
Z	7.1	8.4 ± 0.4	101	0.49 ± 0.03
W	1.9	2.8 ± 0.2	34	0.20 ± 0.01

Table 1.8 Physical Lengths of Chicken Chromosomes

[a] - (Bloom *et al.*, 1993); [b] - (Smith and Burt, 1998)

reciprocal recombination in all areas of the chromosome or a “hot spot” of recombination (Rodionov *et al.*, 1992).

1.11 CpG Islands, Gene Density/Distribution and Chicken Microchromosomes

1.11.1 Cytology

Chromosome identification and karyotyping can be carried out using banding patterns of metaphase chromosomes. Such studies have also shown chromosomes to be made up of segments with different structural, replication and staining properties. Compositional and functional differences have been observed between mammalian G- and R-bands (see Table 1.9). It has been observed that genes mapped accurately enough lie in R- bands and their GC-rich subclass T-bands rather than G-bands (Holmquist, 1992) and (Sumner *et al.*, 1993).

Giemsa staining of vertebrate chromosomes has highlighted GC and AT-rich regions. The Giemsa light R bands replicate early during S phase and are GC-rich with the majority of house-keeping genes associated with them. In contrast, G band DNA replicates later in S phase and is more AT-rich than R band DNA (Holmquist *et al.*, 1982) The majority of tissue specific genes are located in G bands. Similar banding patterns have been observed on both macrochromosomes and microchromosomes, which complements these earlier studies (McQueen *et al.*, 1996). From replication timing experiments, microchromosomes were found to replicate early during the first half of S phase and the macrochromosomes late during the mid to late S phase. Staining of chicken and other avian chromosomes suggested that certain areas such as the microchromosomes have a high proportion of early replicating DNA and were GC-rich (Ponce de Leon *et al.,* 1992) and (McQueen *et al.,* 1998). CpG islands were associated with early replicating, GC-rich, R band DNA. On macrochromosomes, however, the DNA tended to be late-replicating and AT-rich (Craig and Bickmore, 1994).

Banding	Method	Relationship to G Band
G	Giemsa staining after proteolytic digestion	
R	Giemsa staining after heat denaturation in saline. Or staining with chromomycin antibiotics which have a bias for GC-rich DNA	Reverse to G bands i.e. G positive bands are R-negative and vice versa
T	R banding at elevated temperatures	A subset of R bands
Q	Staining with AT-specific fluorescent dyes e.g. quinacrine	G positive bands are also Q positive

Table 1.9 Chromosome Banding Techniques

(Craig and Bickmore, 1994)

1.11.2 CpG Islands

CpG islands are regions of non-methylated DNA which have a high G + C content of 60-70% compared to the 40% G + C of bulk DNA and are generally between 0.5 and 2 kb in length. CpG islands are often located within the promoter region of genes and have been found at all house keeping genes plus around 40% of tissue specific genes (Antequera and Bird, 1993), (Cross, 1995), (Delgado *et al.*, 1998). They are known to contain multiple binding sites for transcription factors and can act as initiation sites for transcription and DNA replication (Delgado *et al.*, 1998) and (Jones, 1999). Genes can be isolated by locating CpG islands and in certain cases the methylation of CpG islands is associated with gene inactivation.

Common in mammals and higher vertebrates, CpG islands are less abundant in lower vertebrates and are found in some plant species (Sumner *et al.*, 1993), (Cross, 1995) and (Parimoo, 1995). The definition of CpG islands used in this thesis is a region of DNA 200 bp or more in length, an observed over expected (O/E) CpG content of greater than 0.6 and a %GC of more than 50% (Antequera and Bird, 1993) and (Jones, 1999).

To examine their distribution in various genomes, CpG island libraries for human (Cross *et al.*, 1994), mouse (Cross *et al.*, 1997) and pig (McQueen *et al.*, 1997) were constructed. From these studies, CpG islands were found to have a non-random distribution, with clustering of CpGs being observed.

Originally, it was assumed that chicken microchromosomes carried very few genes, the bulk being found on the macrochromosomes. A CpG island library was constructed to determine the distribution of genes in the chicken genome (McQueen *et al.*, 1996). To the library was prepared with *Mse* I-digested chicken liver DNA using differential binding to a methyl-CpG binding column before and after *de novo* methylation. Studies with this library showed that on the microchromosomes, CpG islands appear to be very concentrated and therefore potentially gene rich. The macrochromosomes had very few CpG islands however and are therefore possibly gene poor. Microchromosome euchromatin therefore appears to be gene rich and replicates earlier than macrochromosomes. This matches human CpG islands which are concentrated in early replicating domains (McQueen *et al.*, 1996). This corresponds with banding studies which also place CpG islands within the early replicating portions of the genome.

Out of 20 chicken genes chosen at random from the sequence database, 14 had CpG islands (McQueen *et al.*, 1996). There are in the region of 45,000 CpG islands per haploid genome in humans and approximately 37, 000 found in mouse (Antequera and Bird, 1993) and (Cross, 1995). By assuming that chicken has a similar number of CpG islands to mammals, it was estimated to be a gene with a CpG island at approximately 1 gene per 10 kb of microchromosomal DNA (McQueen *et al.*, 1996). If true, this would mean that the gene density on microchromosomes would be close to the maximum value known for vertebrates and it has been suggested that there are some similarities between the puffer fish, *Fugu*, which has a compact genome and the microchromosome portion of the chicken genome (McQueen *et al.*, 1996). It is interesting to note that the majority of *Fugu* genes, including the housekeeping genes, lack CpG islands at their 5' ends (Elgar, 1996).

1.12 Aims and Objectives of the Thesis

The macrochromosomes and microchromosomes appear to differ in regards to repeat distribution, G + C content and gene density. Based on this, this thesis asks and attempts to answer a number of questions.

1) Are microchromosomes gene dense and macrochromosomes, by comparison, less gene dense?

2) Are CpG islands concentrated on the microchromosomes as suggested by prior studies?

3) If microchromosomes are gene dense, will they have fewer repeats?

Other features of the chicken genome include the mobile CR1 repeats.

4) If microchromosomes are gene dense, then they should have fewer repeats because they contain less "junk" DNA and therefore have less room for integration. Is the distribution of CR1 repeats affected by gene density?

5) There is evidence to suggest that CR1s are an ancient class of repeat, existing before the divergence of avians and reptiles. How has this repeat has evolved?

6) Is there any evidence for the types of selective forces that constrain the expansion of the chicken genome and do these forces act equally on macrochromosomes and microchromosomes?

In addition, regions of conserved genes (coding and non-coding regulatory sequences), conserved syntenic groups and conserved gene order have been found between chicken and other vertebrate species.

7) In Humans the genes Tyrosine Hydroxylase (TH), Insulin (INS) and Insulin-Like Growth Factor-II (IGF-II) are contiguous and map to chromosome 11p15.5, forming the linkage group 5' TH-INS-IGF-II-3'. In the chicken these genes have been mapped to chromosome 5 but their order is unknown. What is the order of these genes in chicken?

The main approach used in this study was DNA sequence sampling. With this technique, cosmids from different portions of the genome are partially sequenced and the data analysed by database searches for gene homologies.

8) Is sequence sampling and database searching an efficient method of finding genes?

Chapter 2

Materials and Methods

2.1 Molecular Biology Techniques

All molecular biology techniques were carried out as described by (Sambrook *et al.*, 1989) unless outlined below. All modifying enzyme reactions were carried out according to the manufacturers instructions and used with buffers supplied. Any commercial kits were used following the given protocols. All glassware and plastics were presterilized by autoclaving.

2.2 Centrifugation

Centrifugation of eppendorf tubes (3,000 to 13, 000 rpm) was performed using a MSE Micro Centaur (Scotlab, UK) bench top microfuge. Centrifugation of corex tubes (4, 000 rpm to 12, 000 rpm) was performed using a JA20 rotor in a J2-21M/E centrifuge (Beckman, UK).

2.3 Buffers and Solutions

All chemicals were of analytical grade or equivalent and were obtained from Sigma, UK, BDH Laboratory Supplies, UK and Difco Laboratories, USA unless otherwise stated. Restriction enzymes were obtained from Promega, Boehringer Mannheim, New England Biolabs and Cambio. To make the solutions Millipore deionised water, autoclaved where appropriate, was used. All buffers and solutions were sterilised by autoclaving or filtration where appropriate.

2.3.1 General Stock Solutions

Tris-borate/EDTA Electrophoresis Buffer (TBE) 10x stock

0.89 M Tris base

0.89 M Boric acid

20 mM EDTA pH 8.0

Tris-acetate/EDTA Electrophoresis Buffer (TAE) 50x stock

40 mM Tris Acetate

0.5M EDTA pH 8.0

Tris-EDTA buffer (TE) pH 8.0 10x stock

10 mM Tris-HCl

1 mM EDTA pH 8.0

SSC 20x stock

3 M Sodium chloride

300 mM Sodium citrate

Adjusted to pH with 10M NaOH

Phenol:Chloroform

Equal amounts of phenol and chloroform were mixed and equilibrated by extracting several times with 0.1 M Tris.HCl (pH 7.6). The mixture was stored under an equal volume of 0.01M Tris.HCl (pH 7.6) at 4 °C in a dark glass bottle

5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) 2% (20 mg/ml) Stock

X-gal was dissolved in dimethylformamide to make a 20 mg/ml solution.

2.3.2 Loading Dyes

Loading Dye for Sequencing

95% Deionised formamide

0.5 M EDTA (pH8)

0.25% Dextran blue dye

Loading Dye for Gel Electrophoresis

40% (w/v) sucrose in water

0.5 M EDTA (pH 8)

0.25% Bromophenol blue

Loading Dye For SSCP Gels

95% Deionised formamide

0.25% BPB xylene cyanol

0.5M EDTA (pH 8)

2.4 Bacterial Strains

The *Escherichia coli* strain, DH5α (*sup*E44, Δ*Lac* U169(ϕ80 *lac* ZΔM15) *hsd*R17 *rec*A1 *gyr*A96 *thi-1 rel*A1 *sup*E44 *hsd*R17 *rec*A1 *end*A1 *gyr*A46 *thi rel*A1 *lac*- F' [*pro*AB+*lacIqlac* ZΔM15 tn (tet^r)]) was used in transformation experiments.

2.4.1 Culture Conditions

Bacterial strains were cultured on Luria agar plates with appropriate antibiotics and stored, short term, at 4°C. For longer term storage, they were cultured overnight in Luria broth with the appropriate antibiotics. The following day the cells were centrifuged for 10 minutes at 4,000 rpm, the supernatant was removed and the pellet re-suspended in one part 10 mM $MgSO_4$ to three parts 70% glycerol. This was frozen in liquid nitrogen and stored at -70°C.

2.4.2 Media

Luria Broth	per litre
Bacto-yeast extract	5 g
Bacto-tryptone	10 g
Sodium Chloride	10 g

Adjusted to pH 7.6 with 10M KOH

Luria Agar	per litre
Sodium chloride	10 g
Bacto-tryptone	10 g
Bacto-yeast extract	5 g

2.4.3 Preparation of Electrocompetent *Escherichia coli* Cells

The *E.coli* strain DH5α was streaked out on a luria agar plate and incubated over night at 37 °C. Single colonies were picked and used to inoculate two 10 ml overnight cultures of Luria broth which were grown up overnight at 37 °C. The following day the overnight was used to inoculate two 500 ml cultures of Luria, which were grown, at 37°C, to an A_{600} nm of 0.5-0.8. The cells were prepared as described by (Dower *et al.*, 1988).

2.4.4 Transformation of *Escherichia coli* Cells by Electroporation.

Introduction of a vector into a bacterial strain was carried out by electroporation. To carry out the transformations, 0.1 µg of DNA was mixed with 40 µl of competent *E.coli* cells. This was electroporated (2500V) in 2 mm cuvettes (Equibio) using a Celliect electroporator. The cells were allowed to recover at 37 °C for 15 min by shaking at 225 rpm in 500 µl of LB. 50 µl, 200 µl or all of the cells were plated out on to Luria agar plates containing 50 µg/ml ampicillin, 100 µg/ml IPTG and 50 µg/ml X-gal. The plates were inverted and incubated overnight at 37°C.

2.5 Polymerase Chain Reaction (PCR)

2.5.1 PCR Reactions

PCR primers were designed with the Primer3 program available from the primer design menu at HGMP (http://www-genome.wi.mit.edu/). Primer3 selects primers for PCR reactions according to GC content, size, PCR product size, primer-dimer possibilities and oligonucleotide melting temperature. The primers were synthesised by Genosys UK Ltd. A typical PCR reaction mix of 25-100 µl comprised of:

Primers 0.5 pM/µl each

Template DNA 1ng/µl

Mg^{2++} 1.5mM

dNTPs 200 µM each

Taq DNA polymerase 0.5 U

10x Buffer

PCR cycling conditions used throughout this study were:

Denature	95°C	3 min	
Anneal	60°C	2 min	
Extension	72°C	3 min	x 25
Denature	95°C	1 min	

Reactions using amplitaq gold (Perkin Elmer) had an initial 15 min denaturation step at 95°C and a final 10 min extension at 72 °C.

2.6 Cloning Vectors

2.6.1 Plasmids

For general subcloning purposes the pBluescript II SK$^+$ plasmid (Figure 2.1) was used 2.12. In Chapter 3, 93 random genomic clones from were analysed and the vectors used to create these clones were plasmids pBluescript II SK$^+$, pUC 18 (cloning site *Bam*HI) and pT7T3.

2.6.2 Cosmids

Two chicken cosmid libraries were used in this study. The first was the commercially available chicken cosmid library, made by CLONTECH (California) from adult Leghorn male liver DNA. The clones are constructed in pWE15, Figure 2.2.

The second chicken cosmid library available from (Buitkamp *et al.*, 1998) was constructed using the SuperCos 1 (sCos 1) cloning vector (Figure 2.3) and

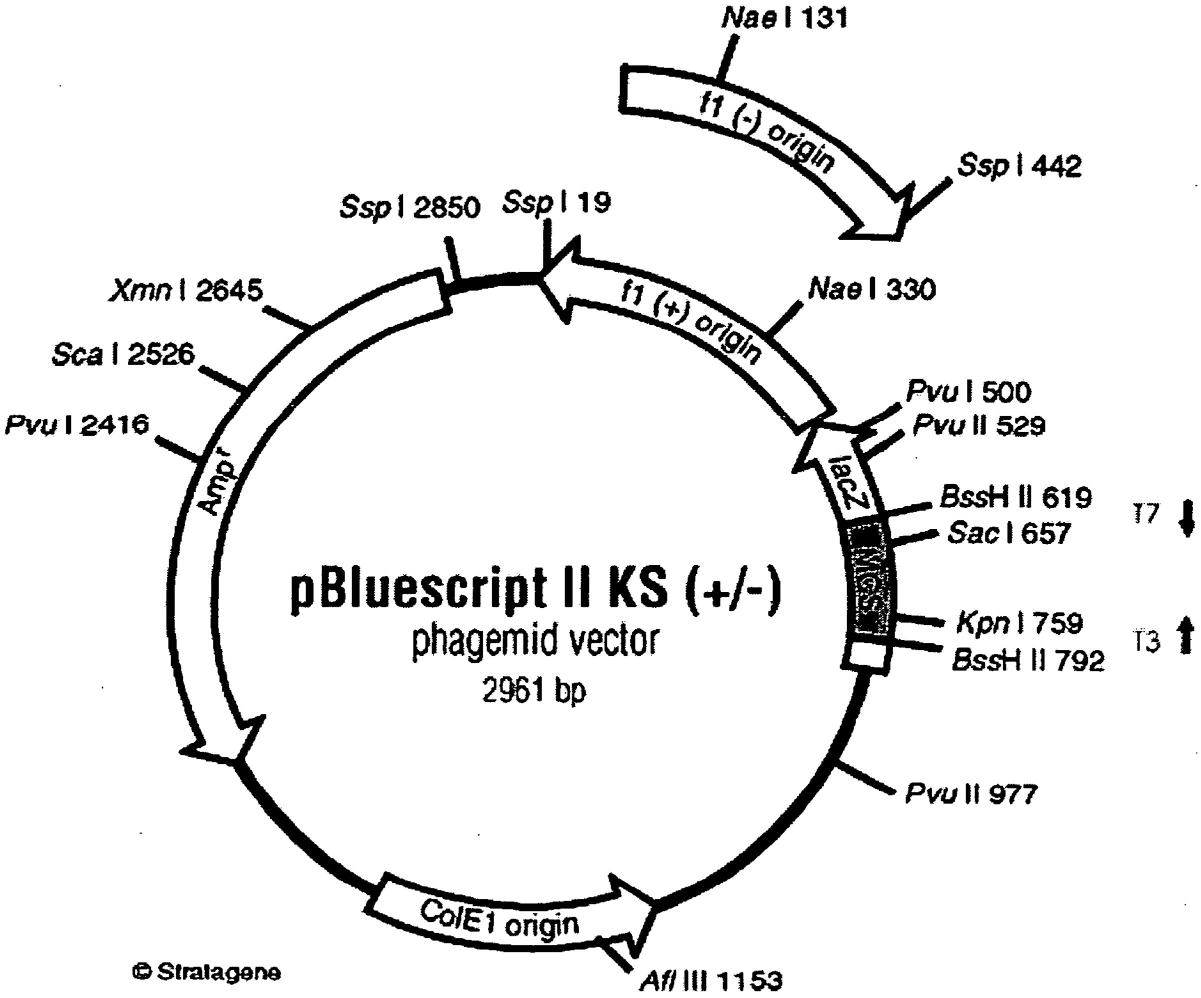


Figure 2.1 pBluescript SK (+/-) Phagemid Vector

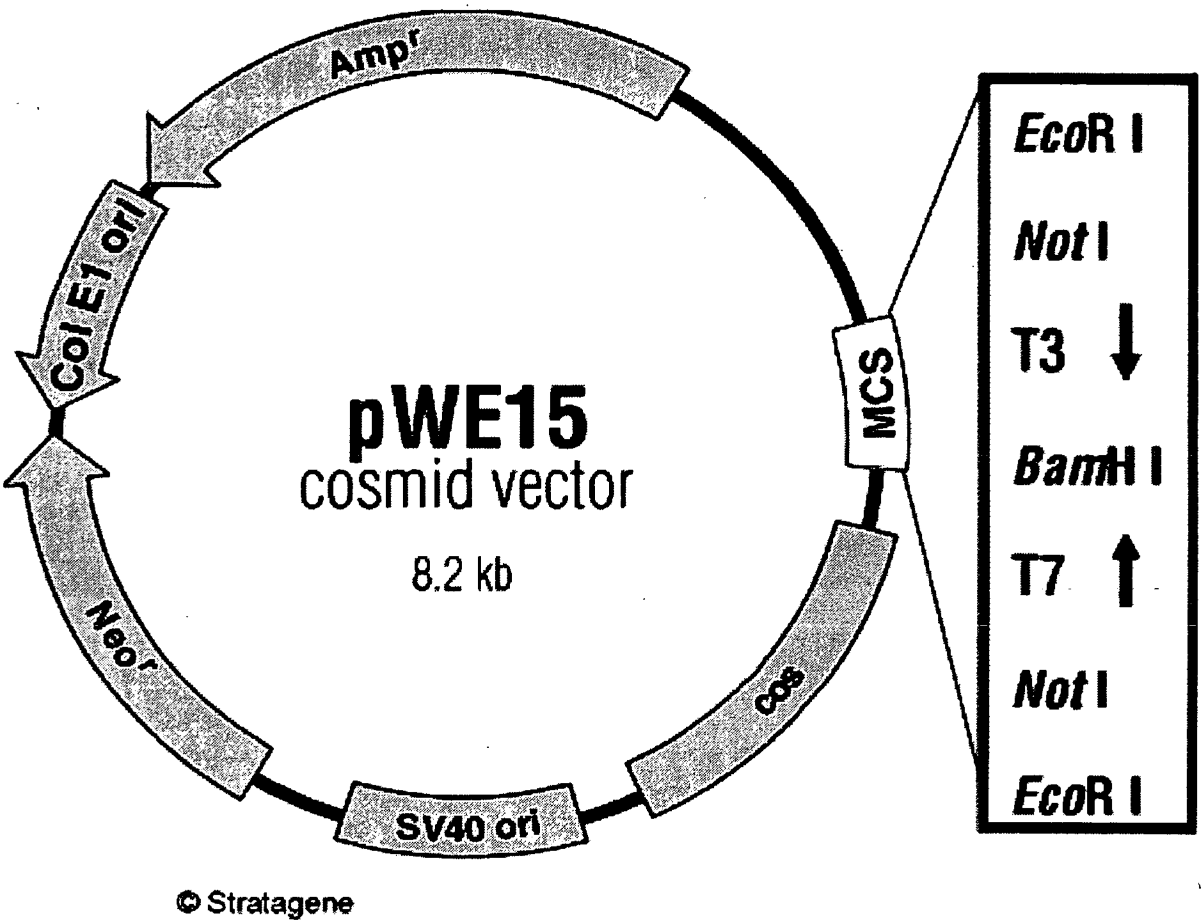


Figure 2.2 pWE15 Cosmid Vector

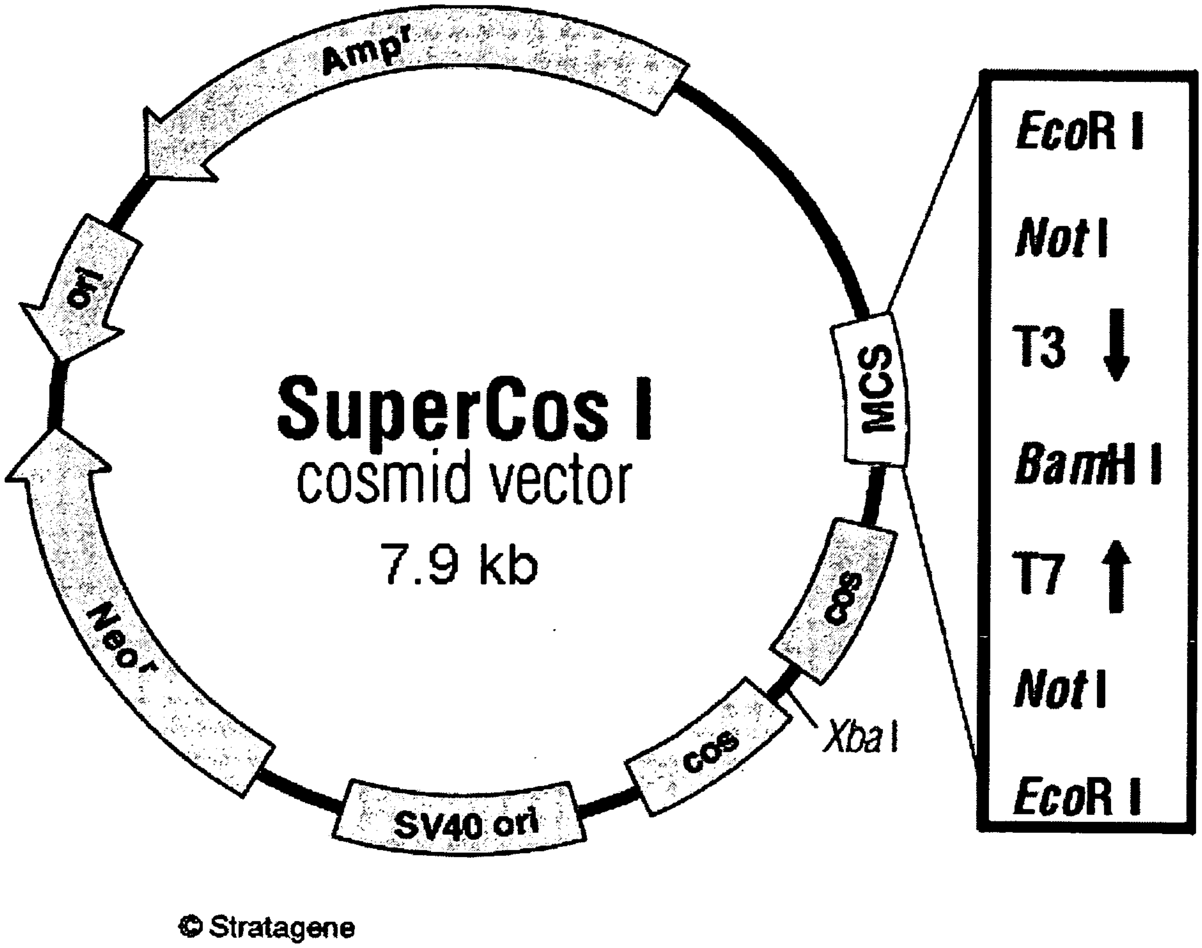


Figure 2.3 SuperCos I Cosmid Vector

female genomic DNA. The library had been gridded by (Buitkamp *et al.*, 1998) onto four nylon filters for the identification of clones by colony hybridisation.

2.6.3 Preparation of Plasmid and Cosmid DNA

Plasmid DNA was prepared using either the QIAprep Spin Plasmid Kit or the QIAwell Plasmid Kits (Qiagen). Cosmid DNA was prepared using the Qiagen Plasmid Maxi kits.

2.7 Phenol/Chloroform Extraction and Ethanol Precipitation of DNA

Equal volumes of phenol/chloroform:iso-amyl alcohol is added to the DNA, vortexed to produce an emulsion and centrifuged at 14,000 rpm for 5 minutes. The upper phase containing DNA was recovered and transferred to a clean microfuge tube. A further volume of phenol:chloroform was added and the process was repeated. The second upper phases was combined with the first and an equal amount of chloroform:isoamyl alcohol was added. This was vortexed and microfuged at 14,000 rpm. The upper layer was removed and transferred to a clean microfuge tube. To precipitate the DNA 0.1 volume of 3M sodium acetate and 2 volumes of cold 96% ethanol was added. This was placed on ice for 15 minutes, then centrifuged at 14,000 rpm for 15 minutes. The pellet was washed in 70% ethanol, dried and resuspended in a suitable amount of TE.

2.8 Phosphatase Treatment of Vector

10x Calf Intestine Phosphatase (CIP) Buffer

0.5 M Tris HCl (pH 9)

10 mM $MgCl_2$

1 mM Zn Cl_2

10 mM Spermidine

10x Sodium chloride Tris EDTA (STE)

100 mM Tris HCl pH 8

1 M NaCl

10 mM EDTA

Calf Intestine Phosphatase (CIP) Dilution Buffer

0.5 M Tris HCl pH8

1 mM EDTA

5 µg of pBluescript II SK$^+$ (Stratagene) was digested with EcoRV restriction at 37 °C for one hour. This was phenol/chloroform extracted and the pellet re-suspended in 50 µl of TE. To check the digest, 2 µl was electrophoresed on a 0.8% agarose gel. The linerised vector was treated with 1 µl of calf intestine phosphatase (CIP) at 37 °C for thirty minutes. Another 1 µl of CIP was added and the reaction was heated at 37 °C for a further thirty minutes, after which, 160 µl distilled water, 10x STE and 10% SDS was added. This was heated at 65 °C for 15 minutes, phenol/chloroform extracted and re-suspended in 100 µl TE to give a final concentration of 50 ng/µl.

2.9 DNA Electrophoresis

2.9.1 Agarose Gel Electrophoresis

Agarose gels were prepared with appropriate amounts of 1x TAE buffer and agarose. To every 100 ml of molten agarose, 10 µl of 10 mg/ml ethidium bromide was added. DNA samples and appropriate size markers were loaded on to the gels using a small amount of loading dye. Electrophoresis was carried out at 80-100 volts for 2-3 hours or at 10-15 volts overnight in 1X TAE buffer.

2.9.2 Acrylamide Gel Electrophoresis

5% Acrylamide Gel Mix

40% Acrylamide solution

2% NN'-Methylene-bis-acrylamide solution

10 x TBE

DNA fragments of less than 500 bp in size were resolved on 5% acrylamide gels. For each gel 5% acrylamide gel, 100 µl of 10% ammonium persulfate and 100 µl of 99% N,N,N',N'-Tetramethylethylenediamine (TEMED) was used. The DNA samples were loaded using appropriate markers and electrophoresed at 100 volts.

2.10 Preparation of DNA from Agarose Gels

Bands were excised from agarose and the DNA was extracted from the gel slice using the QIAEX II Agarose Gel Extraction kit (Qiagen) and re-suspended in 20 µl of water. To check that the extraction had been a success, 2 µl of resuspended DNA was electrophoresed on a 0.8% agarose gel.

2.11 Calculation of DNA Concentration

The concentration and purity of DNA samples was determined spectrophotometrically by optical density measurements at 260 and 280 nm. $OD_{260} = 1 = 50$ µg/ml DNA.

2.12 Isolation of Cosmid Clones and Preparation of Random Subclones for DNA Sequencing

Cosmid clones from the Clontech chicken genomic library were picked at random and the DNA prepared using Qiagen purification columns (Qiagen). The restriction enzyme *Cvi*JI (Cambio) was used to partially digest 5 μg of cosmid DNA. The *Cvi*JI enzyme was used as it cuts at the site PuG/CPy to produce random fragments (Fitzgerald *et al.*, 1992; Gingrich *et al.*, 1996). Digests were carried out with 0.2 U enzyme/μg DNA at 37°C for one hour followed by 15 minutes at 65 °C and electrophoresed on a 1% agarose gel. Fragments between 500 bp and 2 kb in size were excised and DNA extracted with a Qiaex II DNA extraction kit (Qiagen GmbH, Germany). The fragments were subcloned into *Eco*RV digested pBluescript II SK$^+$ (Stratagene) and transformed by electroporation. The plasmid DNA prepared with a QIAwell plasmid prep. kit (Qiagen).

2.13 Ligation Reactions

Ligation reactions between linerised and dephosphorylated pBluescript II SK$^+$ (Stratagene) and insert DNA were carried out in the presence of T4 DNA ligase at 16°C overnight. A ratio of 1:1 or 1:3 vector to insert DNA was used in the ligation reaction.

2.14 Examination of Plasmid DNA Inserts

To confirm the presence of insert DNA in the plasmid and to ensure there were no multiple inserts or religated vector, double digests with the restriction enzymes *Eco*RI and *Hin*dIII were carried out and analysed on 0.8% agarose gels.. The sites of these enzymes flank the *Eco*RV cloning site.

2.15 SuperCos 1 (sCos 1) Gridded Cosmid Library Screen

The sCos 1 gridded cosmid library was screened for clones containing the genes encoding insulin (INS), tyrosine hydroxylase (TH) and insulin like growth factor II (IGF2). PCR products from each of the genes were used as probes in the library screen.

2.15.1 sCos 1 Gridded Cosmid Library Screen Solutions

Hybridisation Solution

10% PEG 8000

7% SDS

1.5X SSC

Wash Buffer 1

2X SSC; 0.1% SDS

Wash Buffer 2

0.5X SSC; 0.1% SDS

2.15.2 Radiolabelling of Probe DNA

The probes were labelled using the "Prime-a-gene" kit (Promega) and labelled with $P^{32}\alpha$dCTP as per manufacturers instructions. The probes were denatured at 95 °C for two minutes and placed directly onto ice and 20 mM EDTA was added. Probes were precipitated by the addition of 1/10 volume NaAc (3M) and 3X vol. 96% ethanol and resuspended in 50 µl Tris HCl pH 8.

2.15.3 Hybridisations

The four SuperCos 1 nylon filters were prehybridised in 100 ml of the hybridisation solution at 65°C for at least 30 min. The denatured probe was added and hybridised overnight at 65°C. The filters were washed at 65°C with preheated wash solutions 2XSSC; 0.1%SDS, and 0.5XSSC; 0.1%SDS.

2.16 Autoradiography

Filters containing $P^{32}\alpha$dCTP were exposed to Dupont Cronex film or Kodak Biomax MS film (Sigma-Aldrich Techware, UK). Exposure time was from 1 hr at room temperature to two weeks at -70°C. All films were developed using a X-OGraph Compact x2 (X-OGraph Ltd, Malmesbury, Wiltshire).

2.17 Sequence Sampling

Sequence sampling is a simple method of constructing high resolution physical maps from regions of genomic DNA (Smith *et al.*, 1994). Figure 2.4 describes the isolation of clones covering the region of interest, breaking the clones down into fragments and sequencing. The clones are not sequenced to completion, with the resulting map covering around 40% of the complete DNA sequence but this has proved enough to identify such features as gene homologies, repetitive elements and putative intron-exon boundaries (Smith *et al.*, 1994).

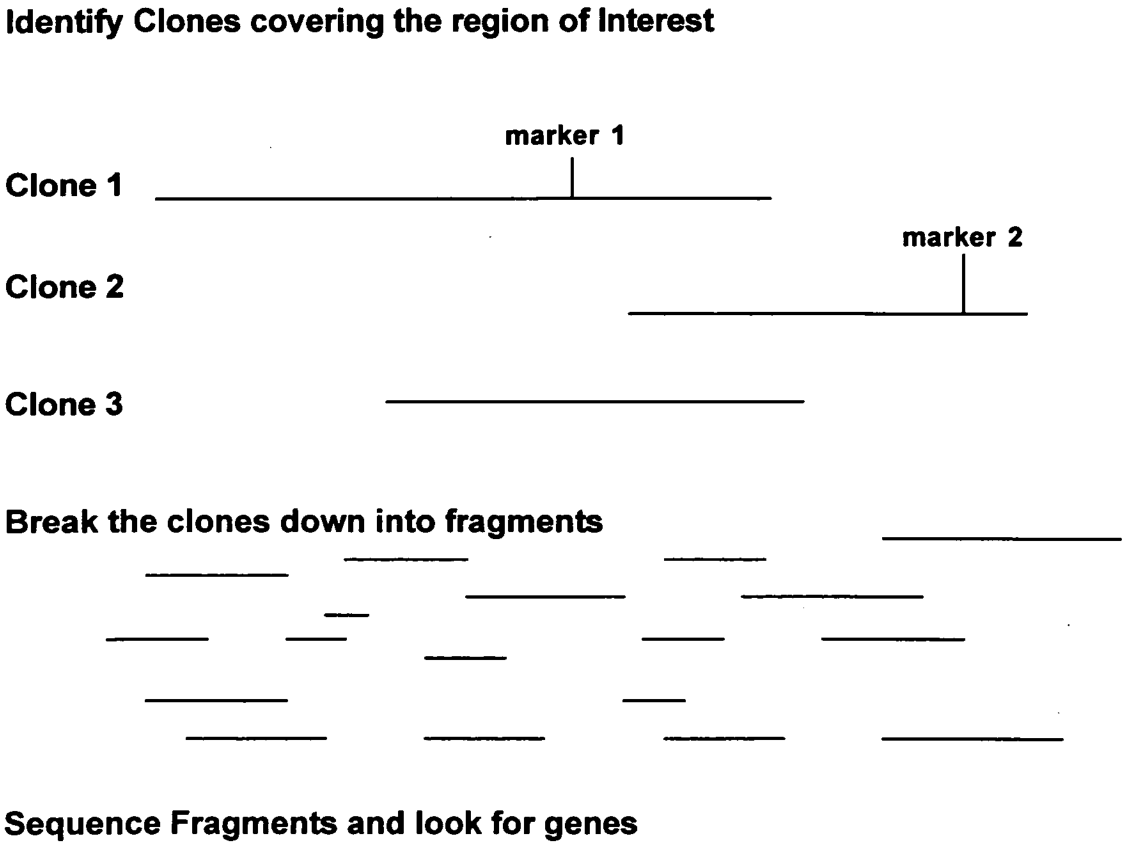

Identify Clones covering the region of Interest
marker 1
Clone 1
marker 2
Clone 2
Clone 3
Break the clones down into fragments
Sequence Fragments and look for genes

Figure 2.4 Sequence Sampling

Clones covering the region of interest are isolated and partially sequenced. The DNA sequence is analysed to identify gene homologies, repetitive elements and putative intron-exon boundaries (Smith *et al.*, 1994).

2.18 Sequencing

2.18.1 Estimation of Number of Clones to Sequence

To estimate the number of clones required for sequencing, the following equation was used: $p = 1 - [1 - x/y]n$, where “p” is the probability of cloning the entire cosmid, “x” the length of an average sequence read, “y” the cosmid insert size (35, 000 bp) and “n” the number of sequences. Enough subclones were sequenced to give at least 40% coverage of the cosmid. The equation assumes complete randomness.

2.18.2 Cosmid Clone and Sequence Naming Strategy

Each cosmid sequenced was given an internal laboratory number. Subclones were named according to a standard convention which identifies the cosmid they came from, the enzyme used to create them and an individual number. For example a series of 48 subclones from cosmid 001 using the enzyme *Cvi*JI would be named “Cos 001 Cvi 01 → Cos 001 Cvi 48”.

The sequences were named in such a way as to show which subclone they came from and which primer e.g. T3 or T7 was used. A number was also given to distinguish it from other sequences from that subclone and primer in the past. For example the first sequence from subclone Cos 001 Cvi 01 using a T3 primer would be called “Cos 001 Cvi 01.T3.1”.

2.18.3 DNA Sequencing of Subclones

Subclones were sequenced from both DNA strands with T3 and T7 primers. Each sequencing reaction was carried out using the dye terminator kit (Applied Biosystems, Perkin Elmer, UK). 200 ng DNA and 3.2 pM of each primer were used in the reactions. Sequencing reactions were carried out on a Hybaid PCR machine. The conditions were 94 °C for 3 minutes, followed by 25 cycles of 94 °C for 30 seconds, 55 °C for 15 seconds and 60 °C for 4 minutes.

Unincorporated dye terminators were removed by ethanol precipitation, as described in the protocol accompanying the dye terminator cycle sequencing ready reaction kit. The pellets were re-suspended in 2 µl loading dye, vortexed, centrifuged, denatured at 95 °C for 2 minutes and placed on ice. The entire sample was loaded onto the sequencing gel.

2.18.4 Electrophoresis of Sequencing Products

4% Acrylamide Sequencing Gel Mix

Urea	18 g
40% acrylamide solution	5 ml
Water	25 ml
Deionizing beads	1 g

Prior to filtering the gel mix, 5 ml of 10x TBE was filtered (to avoid deionisation). The mixture was then de-gassed for 2 min.

The sequencing plates were washed thoroughly, rinsed with distilled water and air dried. The plates were assembled and filled by capillary action with the gel mix. A 36 well sharks tooth comb (teeth facing away from the gel) was inserted into the top of the plates to form a well and secured. The gel was allowed to polymerise for at least two hours before use.

The sharks tooth comb was removed from the gel and the well flushed out with distilled water to remove any unpolymerised acrylamide. The comb, teeth facing into the gel, was replaced. The gel was placed into a frame and secured onto an ABI 377 automatic sequencer.

2.18.5 Analysis of Sequence Data

Raw sequence data was analysed using the Staden program (Bonfield *et al.*, 1995; Bonfield and Staden, 1996; Staden *et al.*, 1996). This is a general sequence assembly package which runs on UNIX machines. It has two main parts, called Pregap and Gap. Pregap assesses sequence quality and looks for cloning vector. A modified version of the Pregap Staden program (Bonfield *et al.*, 1995; Bonfield and Staden, 1996) was used to remove any poor quality sequence from the raw sequence. Poor DNA sequence and cloning vector sequence (cosmid or plasmid) is removed by Pregap. This is followed by post-processing which takes the remaining good sequence and reformats it for database searches. The alignment program Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) was used to carry out overnight BLASTN (DNA vs. DNA) and BLASTX (protein vs. protein) searches are carried out against a number of databases (local *Fugu* and chicken microsatellite databases[1], non-redundant nucleotide and protein databases, EST and STS databases) to find gene homologies. High scores and probability were used as the basis of deciding if a gene homology from the database searches was significant or not. For a gene homology to be considered significant for a BLASTN search, a high score of greater than 150 and a probability of 10^{-9} was necessary. For a protein-protein BLASTX search (Gish and States, 1993) a score of greater than 75 and a probability of 10^{-6} indicated a significant match. Common repeats such as microsatellites are screened out by the BLAST program but not species specific repeats such as the Chicken Repeat 1 elements. To isolate common repeats, the report repeats program was used. This program masks any sequence which is not repeat-like (Claverie and States, 1993) and (Law, Personal Communication). The GC content of the sequences was also assessed. Sequences were assembled into contigs with the Seqman program (DNAStar).

[1] The local *Fugu* database comprised of sequences from MRC HGMP Research Centre Fugu Project. The chicken microsatellite database comprises of all known chicken microsatellites.

2.19 Alignments and Phylogenetic Analysis of CR1 Repeats

2.19.1 Sequence Alignments

CR1 repeat sequences were re-formatted to GCG format and alignments carried out using the CLUSTALW program (Thompson *et al.*, 1994). The multi-sequence alignment editor Seqlab was used to fine tune the alignments and to incorporate any additional information about the sequences e.g. conserved domains.

2.19.2 Construction of Phylogenetic Trees[2]

Programs from PHYLIP were used to construct phylogentic trees from CR1 repeat sequences. The sequences were re-formatted to PHYLIP format with the tophylip program (Wright, 1999). The Ts/Tv ratios were calculated using PUZZLE (Strimmer *et al.,* 1996);(Strimmer *et al.*, 1997). The DNADIST program was used to calculate the pairwise distances between the DNA sequences and, for DNA, it takes into account transitions and transversions. The Ts/Tv ratio calculated from PUZZLE was used in these calculations. DNADIST has four different methods for calculating pairwise distances between DNA sequences. These are Jules Cantor (J-C), Kimura 2-parameters, DNAML and Jin & Nei. The Kimura method, which takes into account different Ts/Tv ratios, was used as it is suited to a large dataset (Wright, 1999).

The NEIGHBOUR program was used to draw a tree from DNADIST data sets. This is a distance based method of analysing distances from nucleic acids and protein sequence data. The Neigbor-Joining (NJ) method was used as it is fast and does not assume that the evolutionary rate is constant across lineage's. The NJ method finds the "best" or nearest "best" tree (Wright, 1999).

[2] The molecular biology help pages at http://www.bbsrc.ac.uk/molbiol/information.html were referred to for information on the GCG programs used.

2.19.3 Statistical Tests on Phylogenetic Trees

Bootstrapping was used to test tree topology. The SEQBOOT program was used to shuffle the dataset a 100 times, to generate a large set of datasets. The DNADIST program was used to create a large set of distance matrices. To create a large number of trees from the distance matrices NJ was used. To form a consensus tree from this large set of trees, the CONSENSE program was used and DRAWTREE and DRAWGRAM programs were used to visualise the final tree (Wright, 1999).

2.20 Genetic and Physical Mapping of Random Cosmids

The random cosmid clones were mapped by physical and genetic means. The genetic mapping was carried out using PCR length variants and single-strand conformation polymorphisms (SSCPs). The physical mapping was carried out by FISH.

2.20.1 Genetic Mapping by PCR Length Variants

Sequence data from the cosmids was analysed for repeats using the program 'report repeats' (Law, Personal Communication). PCR primers (Table 2.1) were designed around suitable repeats and PCR reactions carried out using parental DNA from the East Lansing cross. The PCR products were electrophoresed on acrylamide gels and examined for size differences.

Clone	Primers	Expected PCR Product Length (bp)	Annealing temp (°C)
Cos28	A: CACAGATTCTCAGCTTCCAG B: TACGGGCTTCTGAATCTTAT	272	60
Cos35	A: GTTAGTCTGCATTACTGGCC B: CTGTATGCCTCAGTGGCAAA	342	60
Cos34	A:TCCTGTGGGTCTTTTCCAAC B: AACAACAAGAAATGGGGTGG	250	60
Cos30	A: TTGAGTTTGTGCTGTGAGTC B: TTAAAAAGCGATGCGAAGCT	288	60
Cos33	A: TCTTGCCTTGATGCACTTTG B:CACAGTAAGACAGCGGAAT	225	60
Cos27	A: AATTGCAGATGTGTCCTCAG B: TTTCAGAGATGGTTATCTCC	261	60
Cos32	A: CAAGTGGAAAATGTGATGCG B: TGATCCGATTTACAGCCTCC	314	60

Table 2.1 Primers Used in SSCP and PCR Analysis

2.20.2 Genetic Mapping by Single-strand Conformation Polymorphisms (SSCPs)

Preparation of a 15% Acrylamide:Bis Acrylamide Gel for SSCP Analysis

Acrylamide (37:1, 20%)

10X TBE

Water

SSCPs are sequences which differ by as little as one base and can be detected by the migration of short single stranded fragments down nondenaturing gels (Beier, 1993). SSCPs of PCR products can be used as markers for linkage analysis. To 15 µl of PCR product, 15 µl of formamide loading dye was added. The samples were placed on at heating block at 99 °C for three minutes, then immediately placed on ice to cool for at least five minutes. The samples were then loaded onto the gel. A Hoffer Gel rig was used. The gels were electrophoresed in pre-cooled 1XTBE buffer overnight at 12 mA/gel. The voltage limit was set at 500 V.

2.20.3 Silver Staining of SSCP Gels

The fragments were detected by silver staining. The staining solutions were aspirated off between each wash. To the gel, 10% ethanol was added, left for 5 minutes, then oxidised in 1% nitric acid for three minutes. A brief rinse in distilled water was carried out prior to the addition of the silver nitrate solution. The gel was left in the silver nitrate for twenty minutes, then rinsed in distilled water. The gel was reduced in a sodium carbonate/formaldehyde solution until the DNA bands become visible on the gel. The reducing process stopped by the addition of 10% glacial acetic acid to the gel for two minutes. The gel can be stored in distilled water prior to drying down.

2.20.4 Physical Mapping of Cosmids by Fluorescence *In situ* Hybridisation

Chicken metaphase spreads were prepared from chicken embryo fibroblasts after treatment with 0.6 µg/ml colcemid solution and 0.56% KCl hypotonic treatment.

2.20.5 Fluorescence *in situ* Hybridisation (FISH) Solutions

Blocking Buffer

4X SSC

0.1% Tween

3% Bovine Serum Albumin

Fix

3:1 ethanol:Glacial acetic acid

Hybridisation Mix

Deionized formamide

50% dextran sulphate

20X SSC (filtered)

0.5M sodium phosphate buffer (pH 7)

50X Denhardt's solution

10X Nick Translation Buffer

0.5 M Tris 7.4

100 mM $MgSO_4$

1 mM DTT

10 mg BSA

2.20.6 Nick Translation of DNA

DNA was labelled with biotin-16-d-UTP (Boehringer Mannheim) by nick translation according to the following reaction:

Bio-16-dUTP
DNA (1 μg)
dNTP's (0.5 mM A, C, G + 0.25 mM T)
Water (up to 25 μl)
DNA Polymerase I (10 Units/ml)
DNAse (1 μg/ml)
10X nick buffer

The reaction was mixed well and pulse microfuged prior to incubating at 16 °C for forty minutes. The reaction was then incubated at 65 °C for ten minutes. To check the size of the fragment (optimum size 200-500 bp), 2 μl was electrophoresed on a 0.8% agarose gel. The reaction was stopped with 1 μl of 0.5 M EDTA and stored at -20 °C.

2.20.7 Slide Preparation

Cells were dropped onto ethanol-cleaned slides, fixed with 3:1 ethanol:glacial acetic acid and allowed to air dry. The slides were dehydrated in a series of five minute ethanol washes, 70%, 90% and 100%. Slides were air dried and incubated at 65 °C for 1 hour. The slides were then denatured in 70% formamide/2X SSC at 65 °C for one minute and immediately immersed in cold 70% ethanol for five minutes. The slides were put through a second series of 70%, 90% and 100% ethanol washes, five minutes for each and air dried.

2.20.8 Preparation of Probe DNA

Probe DNA was prepared as follows:

Probe mixture	1μl
Salmon sperm DNA (10 mg/ml)	0.5 μl
Hybridisation mix	14 μl

This was incubated at 65 °C for 10 minutes followed by a 37 °C incubation for around twenty minutes. Both slides and DNA were kept at 37 °C until they were ready to use. The probe was added to the slides and the cover slips sealed with cowgum. This was incubated overnight at 37 °C in a moist chamber.

2.20.9 Post-Hybridisation Detection

The coverslips were rinsed off in 2X SSC and the slides were put through a series of four 42 °C , five minute washes, the first two being 50% formamide/2X SSC. The second two washes were in 0.1 X SSC. Following this the slide was briefly washed in 4X SSC/0.1% Tween at room temperature. The slide was then incubated with blocking buffer, under parafilm, at 37 °C in a moist chamber for twenty minutes.

During the incubation, the avidin-FITC solution was prepared (1:200 dilution) and centrifuged at 7000 rpm for ten minutes then left for a further ten minutes in the dark. The avidin-FITC solution was added to the slide, under parafilm and incubated in a moist chamber for 1 hour minutes. The slides were then washed three times at room temperature in 4X SSC/0.1% Tween. The slides were then mounted with Vectashield containing propidium iodide (Vector Labs).

Chapter 3

Sequence Sampling in the Chicken: A Test Case Using Random Clones From the Compton Reference Backcross

3.1 Introduction

The aim of this section of the thesis is to determine whether DNA sequencing can be used to find genes by a sequence sampling approach. This demands a method of processing random sequences by stripping out vector sequence, and by rapidly searching for gene homologies. Repeats, such as simple microsatellites and CR1s, were also identified.

A simple test system was established by examining 93 random genomic clones, which were originally used as probes for RFLP mapping on the Compton cross to produce a linkage map to assist in the location of genes affecting salmonella disease resistance (Bumstead and Palyga, 1992; Bumstead *et al.*, 1994). These clones were readily available in plasmid form (no subcloning required), and had been mapped onto the chicken genetic map. To facilitate comparisons with other studies, macrochromosomes were defined as chromosomes 1-5 and Z, whereas the microchromosomes were chromosomes 6-8 and the remaining chromosomes including the W sex chromosome.

The results from this analysis were used give preliminary estimates of gene homologies, chicken DNA %GC content, O/E CpG, CR1 and simple repeat content. This was then used to address questions of are microchromosomes more gene dense than macrochromosomes and are CpG islands concentrated on the microchromosomes? The distribution of different types of repeats was also investigated.

3.2 Results

3.2.1 Sequencing the Compton Clones

Of the 93 clones sequenced, four contained no chicken insert DNA. The remaining 89 clones produced unique chicken sequence and the genetic mapping data are described in Table 3.1, which also shows the lengths of sequences and the observed over expected (O/E) CpG content for each clone.

3.2.2 Test of the Randomness of the Clones

It was important to test that random DNA could be cloned and to this end the division of clones between macrochromosomes and microchromosomes was examined. Of the 89 genomic clones analysed, 60 mapped to macrochromosomes, 21 to microchromosomes and 8 has as yet not been mapped as they show no linkage to other markers. This is a 75/25 division of clones between the two chromosome types, which is close to 70/30 the division of macrochromosomes and microchromosomes in the genome, and helps confirm the randomness of the clones used in this study.

3.2.3 Criterion for a Gene Homology in Database Searches

Manual BLAST searches using the NCBI server (http://www.ncbi.nlm.nih.gov/BLAST/) against a number of databases were carried out to identify potential coding sequences. High scores and probability, as described in Chapter 2, were used as the basis of deciding if a gene homology from the database searches was significant. As the BLAST program filters out non-species specific repeats such as microsatellites, those studied were identified by unfiltered BLAST searches.

Table 3.1 Random Clone Genetic Mapping Information

Mic- microchromosome; Mac-Macrochromosome; UN-unlinked; LG-linkage group; COM LG Size-Compton linkage group size; Seq Len-length of sequence determined; CpG O/E-observed over expected based on G+C content; 'R' *Eco*RI fragments in Bluescript KS; 'T' *Eco*RI fragments in pT3T7 (Pharmacia); 'S' *Bam*HI fragments in pUC18 (Pharmacia); 'U' partial *Eco*RI in pT3T7 (Pharmacia).

Locus	Plasmid	Seq	Seq	% GC	CpG O/E	Chr.	COM	COM	Position
	Name	Primer	Len			No.	LG	LG	Compton
			(bp)					size	(kosambi)
								(cM)	
COM072	U3/10	T3	332	37.34	0.30	1	1	703	14.15
COM072	U3/10	T7	431	33.41	0.03	1	1	703	14.15
COM060	T2/72A	T7	429	34.26	0.03	1	1	703	128.63
COM060	T2/72A	T3	570	35.61	0.34	1	1	703	128.63
COM061	R2	ALL	552	39.85	0.17	1	1	703	155.80
COM036	T4/12	T7	141	31.91	0.00	1	1	703	308.56
COM036	T4/12	T3	295	35.59	0.11	1	1	703	308.56
COM037	T2/22	T3	184	27.17	0.00	1	1	703	330.36
COM037	T2/22	T7	318	35.53	0.15	1	1	703	330.36
COM038	T2/12	T3	379	39.31	0.13	1	1	703	357.53
COM038	T2/12	T7	421	41.56	0.07	1	1	703	357.53
COM040	T3/4	T3	410	34.63	0.08	1	1	703	465.24
COM040	T3/4	T7	429	40.09	0.07	1	1	703	465.24
COM041	T5/24	T3	160	33.12	0.10	1	1	703	468.94
COM042	T5/12	T3	237	38.39	0.14	1	1	703	489.60
COM042	T5/12	T7	421	27.07	0.03	1	1	703	489.60
COM104	T4/20	T3	389	39.07	0.04	1	1	703	496.89
COM104	T4/20	T7	471	38.64	0.10	1	1	703	496.89
COM105	R55	T3	422	40.75	0.00	1	1	703	505.83
COM105	R55	T7	231	32.46	0,14	1	1	703	505.83
COM043	T2/20	T3	316	36.70	0.26	1	1	703	512.98
COM043	T2/20	T7	371	32.07	0.26	1	1	703	512.98
COM098	T2/82	T7	471	35.24	0.03	2	2	456	3.65
COM098	T2/82	T3	430	34.88	0.03	2	2	456	3.65
COM035	T28	T3	262	31.67	0.00	2	2	456	3.65
COM035	T28	T7	421	35.86	0.00	2	2	468	3.65
COM034	T2/71	T7	281	31.31	0.17	2	2	456	5.98
COM034	T2/71	T3	440	31.59	0.00	2	2	456	5.98
COM032	T12/48	T3	282	37.94	0.05	2	2	456	54.04
COM032	T12/48	T7	328	39.64	0.05	2	2	456	54.04
COM067	T12/91	T7	451	29.49	0.03	2	2	456	89.84
COM067	T12/91	T3	601	36.77	0.05	2	2	456	89.94
COM080	T2/38B	T3	429	41.95	0.23	2	2	456	176.95
COM080	T2/38B	T7	429	34.96	0.07	2	2	456	176.95

continued

Table 3.1 continued

Locus	Plasmid Name	Seq Primer	Seq Len (bp)	%GC	CpG O/E	Chr. No.	COM LG	COM LG size (cM)	Position Compton (kosambi)
COM024	T7NB	T7	216	36.11	0.00	2	2	456	267.17
COM024	T7NB	T3	449	43.43	0.00	2	2	456	267.17
COM023	T12/108	T7	351	43.02	0.09	2	2	456	273.01
COM023	T12/108	T3	470	41.06	0.06	2	2	456	273.01
COM022	T2/11	T3	240	35.00	0.06	2	2	456	274.80
COM022	T2/11	T7	321	34.26	0.00	2	2	456	274.80
COM021	T13/40	T3	262	42.74	0.06	2	2	456	292.27
COM021	T13/40	T7	311	47.26	0.10	2	2	456	292.27
COM020	T7/40	T7	441	38.77	0.22	2	2	456	307.11
COM020	T7/40	T3	422	32.22	0.15	2	2	456	307.11
COM046	T12/53A	ALL	484	48.14	0.93	2	2	456	404.85
COM045	T2/73B	T7	291	32.30	0.33	2	2	456	432.31
COM045	T2/73B	T3	316	28.16	0.05	2	2	456	431.31
COM001	S50	T7	308	27.27	0.00	3	3	498	0.00
COM001	S50	T3	385	29.87	0.20	3	3	498	0.00
COM002	T2/41	T3	332	37.34	0.10	3	3	498	14.70
COM002	T2/41	T7	431	33.17	0.00	3	3	498	14.70
COM003	T12/88	T3	361	43.76	0.22	3	3	498	50.86
COM003	T12/88	T7	391	35.03	0.12	3	3	498	50.86
COM004	T2/47	T3	254	41.73	0.06	3	3	498	74.29
COM004	T2/47	T7	401	32.91	0.12	3	3	498	74.29
COM005	T12	T3	343	42.56	0.00	3	3	498	78.21
COM005	T12	T7	351	35.89	0.04	3	3	498	78.21
COM095	T2/58	T7	310	38.71	0.05	3	3	498	87.62
COM095	T2/58	T3	421	36.10	0.19	3	3	498	87.62
COM096	T2/46	ALL	626	40.25	0.15	3	3	498	91.11
COM100	T4/18	ALL	726	35.26	0.13	3	3	498	113.41
COM006	T4/7	T7	481	35.75	0.06	3	3	498	113.41
COM006	T4/7	T3	266	41.72	0.06	3	3	498	113.41
COM007	T5/51	T3	302	41.39	0.16	3	3	498	133.94
COM007	T5/51	T7	351	35.61	0.04	3	3	498	133.94
COM008	T3/2	T3	238	28.15	0.00	3	3	498	145.55
COM008	T3/2	T7	391	34.78	0.04	3	3	498	145.55

continued

Table 3.1 continued

Locus	Plasmid Name	Seq Primer	Seq Len (bp)	%GC	CpG O/E	Chr. No.	COM LG	COM LG size (cM)	Position Compton (kosambi)
COM009	T12/33	T3	336	45.53	0.19	3	3	498	163.95
COM009	T12/33	T7	593	37.77	0.21	3	3	498	163.95
COM029	T3/5	T7	201	35.82	0.33	3	3	498	218.11
COM029	T3/5	T3	308	28.89	0.05	3	3	498	218.11
COM028	T12/67	T3	428	30.37	0.07	3	3	498	225.59
COM028	T12/67	T7	411	40.87	0.12	3	3	498	225.59
COM027	T2/53	T7	327	37.92	0.15	3	3	498	232.72
COM027	T2/53	T3	340	37.64	0.04	3	3	498	232.72
COM030	T3/7	T3	210	29.52	0.07	3	3	498	238.05
COM030	T3/7	T7	431	36.42	0.00	3	3	498	238.05
COM026	R2/7	T7	423	39.71	0.15	3	3	498	303.82
COM026	R2/7	T3	429	39.62	0.07	3	3	498	303.82
COM064	T12/104	T3	428	36.91	0.03	4	4	342	30.30
COM064	T12/104	T7	441	41.72	0.11	4	4	342	30.30
COM063	R2/14	T7	431	32.94	0.07	4	4	342	36.14
COM063	R2/14	T3	421	36.10	0.07	4	4	342	36.14
COM053	T2/70AB	T3	423	36.40	0.03	4	4	342	84.00
COM053	T2/70AB	T7	451	43.23	0.00	4	4	342	84.00
COM058	T2/68	T7	441	38.77	0.00	4	4	342	135.68
COM058	T2/68	T3	452	38.05	0.10	4	4	342	135.68
COM057	T4	T7	451	37.02	0.07	4	4	342	202.83
COM057	T4	T3	647	38.48	0.10	4	4	342	202.83
COM054	T2/38	T3	423	39.48	0.15	4	4	342	208.54
COM054	T2/38	T7	431	33.41	0.11	4	4	342	208.54
COM052	T2/73A	T3	440	35.68	0.07	4	4	342	295.24
COM052	T2/73A	T7	440	42.50	0.40	4	4	342	295.24
COM069	T2/29	T3	429	36.83	0.03	4	4	342	405.39
COM069	T2/29	T7	431	31.32	0.03	4	4	342	405.39
COM017	T2/59	T3	393	31.29	0.16	5	6	193	0.00
COM017	T2/59	T7	410	35.12	0.08	5	6	193	0.00
COM016	T5/1AB	T7	371	33.96	0.08	5	6	193	10.35
COM016	T5/1AB	T3	389	38.30	0.16	5	6	193	10.35

continued

Table 3.1 continued

Locus	Plasmid	Seq	Seq	% GC	CpG	Chr.	COM	COM	Position
	Name	Primer	Len		O/E	No.	LG	LG	Compton
			(bp)					size	(kosambi)
								(cM)	
COM015	R11	T3	401	34.16	0.00	5	6	193	18.37
COM015	R11	T7	424	40.80	0.19	5	6	193	18.37
COM014	R2/1	T3	337	35.90	0.00	5	6	193	26.09
COM014	R2/1	T7	337	37.09	0.00	5	6	193	26.09
COM013	T4/2	T7	311	45.65	0.21	5	6	193	40.42
COM013	T4/2	T3	389	39.07	0.04	5	6	193	40.42
COM012	T5/21	T3	258	43.02	0.25	5	6	193	50.76
COM012	T5/21	T7	391	39.64	0.12	5	6	193	50.76
COM011	T9	T3	257	35.79	0.00	5	6	193	65.99
COM011	T9	T7	351	32.76	0.04	5	6	193	65.99
COM010	T2/72B	T3	570	35.61	0.34	5	6	193	81.58
COM010	T2/72B	T7	429	34.26	0.03	5	6	193	81.58
COM070	R34	T7	402	39.80	0.04	5	6	193	112.49
COM070	R34	T3	449	32.07	0.07	5	6	193	112.49
COM056	T2/55B	T3	453	34.21	0.07	6	10	183	0.00
COM056	T2/55B	T7	452	30.97	0.07	6	10	183	0.00
COM055	T2/55A	T3	453	34.21	0.07	6	10	183	28.71
COM055	T2/55A	T7	452	30.97	0.07	6	10	183	28.71
COM049	T2/28	T3	381	39.63	0.21	7	7	147	0.00
COM049	T2/28	T7	381	41.47	0.08	7	7	147	0.00
COM047	T5/39	T7	451	35.25	0.14	7	7	147	16.98
COM048	S52	T7	241	46.05	0.20	7	7	147	18.81
COM048	S52	T3	431	44.08	0.30	7	7	147	18.81
COM076	T2/66	T3	339	39.82	0.09	7	7	147	108.15
COM076	T2/66	T7	481	31.42	0.04	7	7	147	108.15
COM059	T12/46	ALL	601	40.43	0.16	8	14	149	67.00
COM085	T12/74	T7	361	39.88	0.13	8	14	149	97.00
COM085	T12/74	T3	479	44.25	0.17	8	14	149	97.00
COM082	T4/19	T3	410	69.51	1.72	8	12	102	107.70
COM082	T4/19	T7	431	42.92	0.15	8	12	102	107.70
COM050	T12/106	ALL	755	37.08	0.08	9	6	140	103.55
COM073	T4/13	T7	328	56.70	1.45	9	6	140	155.01
COM073	T4/13	T3	479	32.77	0.06	9	6	140	155.01

continued

Table 3.1 continued

Locus	Plasmid	Seq	Seq	% GC	CpG	Chr.	COM	COM	Position
	Name	Primer	Len		O/E	No.	LG	LG	Compton
			(bp)					size	(kosambi)
								(cM)	
COM075	U1/26	T7	401	29.92	0.12	C21	21	34	11.04
COM075	U1/26	T3	429	35.19	0.03	C21	21	34	11.04
COM107	T5/1AB	T7	371	33.96	0.08	C22	UN	?	0.00
COM107	T5/1AB	T3	389	38.30	0.16	C22	UN	?	0.00
COM083	R3/1	T7	409	28.85	0.00	C17	17	46	10.88
COM083	R3/1	T3	460	40.43	0.14	C17	17	46	10.88
COM066	T12/28	T7	421	33.01	0.07	C17	17	?	59.87
COM065	R33	T3	426	34.50	0.07	C17	17	?	61.03
COM065	R33	T7	441	37.86	0.00	C17	17	?	61.03
COM062	T4/6	T7	259	42.47	0.18	mic	15	42	8.23
COM062	T4/6	T3	452	41.37	0.21	mic	15	42	8.23
COM079	T12/61	ALL	484	42.14	0.20	mic	?	?	9.03
COM077	T12/81	T3	364	43.95	0.27	mic	?	0	66.75
COM077	T12/81	T7	431	32.94	0.03	mic	?	0	66.75
COM081	T2/69	T7	401	42.19	0.40	mic	?	?	50.00
COM074	R54	T3	186	25.26	0.09	M?	28	84	0.00
COM074	R54	T7	424	37.73	0.19	M?	28	84	0.00
COM099	T13/2	T3	428	31.54	0.00	UN	UN	?	0.00
COM102	T12/11	T3	381	36.48	0.08	UN	UN	?	0.00
COM102	T12/11	T7	394	28.42	0.08	UN	UN	?	0.00
COM031	R8	T7	234	35.89	0.00	UN	UN	?	0.00
COM031	R8	T3	306	34.31	0.10	UN	UN	?	0.00
COM071	S4	T3	431	43.38	0.26	UN	UN	?	0.00
COM071	S4	T7	451	36.36	0.14	UN	UN	?	0.00
COM108	T12/53AB	ALL	484	48.14	0.93	UN	UN	456	0.00
COM109	T12/53AB	ALL	484	48.14	0.93	UN	UN	456	0.00
COM099	T13/2	T3	428	31.54	0.00	UN	UN	?	0.00
COM099	T13/2	T7	448	34.82	0.10	UN	UN	?	0.00
COM084	T2/70AB	T3	423	36.40	0.03	UN	UN	342	0.00
COM084	T2/70AB	T7	451	43.23	0.00	UN	UN	342	0.00
TOTAL:			63,913 bp						
		Mac	43,463 bp						
		Mic	15,107 bp						
		UN	5,343 bp						

3.2.4 Database Search Results

The database search results, summarised in Table 3.2 show three significant hits. Two were against human ESTs (Figures 3.1 and 3.2a), both of which mapped to a macrochromosome. The database hit for T2/55T7 (*COM55*) was against the 3' end of a human cDNA clone (Figure 3.1). This is likely to be a 3'UTR, a conserved regulatory sequence; the cDNA clone has yet to be mapped in humans. The EST database match was against clone T2/73B (Figure 3.2a). Further study of the sequence revealed a potential open reading frame/exon region (Figure 3.2b). The third significant hit was against the human gene encoding for the transforming growth factor-beta II receptor (TGFβR2) which mapped to chromosome 6 (a microchromosome).

3.2.5 Gene Homology: Transforming Growth Factor-Beta Type II Receptor

Transforming growth factor-beta (TGF-β) is a multifunctional protein, which is secreted by many different cell types. It has roles in mammalian development and cellular processes such as proliferation and differentiation, extracellular matrix synthesis and the regulation of the synthesis of other growth factors and their receptors (Bae *et al.*, 1995), (Lawler *et al.*, 1994) and (Moustakas *et al.*, 1993). There are three known mammalian, closely-related TGF-β isoforms, TGF-β1, TGF-β2 and TGF-β3, each with a distinct role in foetal development (Moustakas *et al.*, 1993) and (Lawler *et al.*, 1994).

There are also three types of cell surface receptors, types I, II and III, associated with TGF-β. Types I and II receptors are transmembrane serine/threonine kinases which form a complex essential for signalling responses; TGF-β has an inhibitory growth effect on many different cell types through binding these receptors (Hougaard *et al.*, 1999), (Moustakas *et al.*, 1993) and (Lawler *et al.*, 1994).

Clone	**Chromosome**	**Locus**	**Database Homology**	**Accession Number**
T2/73B	2	*COM045*	Human cDNA clone	T84468
T2/73AT7	4	*COM052*	TGF-beta II Receptor	D50683
T2/55AT7	6	*COM055*	Human cDNA clone	AI148083

Table 3.2 Database Homologies from Compton Cross Clones

```
                                                                 Smallest
                                                                   Sum
                                                          High  Probability
                                                         Score  P(N)       N

gb|AI148083|AI148083 qb39a10.x1 Soares_pregnant_uterus_Nb...232  2.0e-26   4

gb|AI148083|AI148083 qb39a10.x1 Soares_pregnant_uterus_NbHPU Homo sapiens
cDNA clone IMAGE:1698618 3', mRNA sequence [Homo sapiens] Length = 517

Minus Strand HSPs:

Score = 134 (37.2 bits), Expect = 2.0e-26, Sum P(4) = 2.0e-26
Identities = 34/45 (75%), Positives = 34/45 (75%), Strand = Minus / Plus

Query: 441 AAATTCTTTAGTKRAATATATACACATTACACAGTATCTTAAAAA 397
           ||  ||||||||  ||||| |||||||||||  |||| |  | ||
Sbjct:  76 AATCTCTTTAGTTAAATATGTACACATTACATGGTATTTGTATAA 120

Score = 82 (22.8 bits), Expect = 2.0e-26, Sum P(4) = 2.0e-26
Identities = 18/20 (90%), Positives = 18/20 (90%), Strand = Minus / Plus

Query: 338 TTAGACTGTAATGTCAGTTT 319
           ||||| | ||||||||||||
Sbjct: 177 TTAGAGTATAATGTCAGTTT 196

Score = 189 (52.5 bits), Expect = 2.0e-26, Sum P(4) = 2.0e-26
Identities = 49/63 (77%), Positives = 49/63 (77%), Strand = Minus / Plus

Query: 175 AACTTTGATTGCATACAGAGGTTAGAATGACCAGCACTAAAATAACCCCCATCTCTACAA 116
           ||  ||||||| ||||  ||| || ||||||||||||| |  |||||||||| ||| | |
Sbjct: 344 AAGCTTGATTGTATACTCAGGCTACAATGACCAGCACTGATGTAACCCCCATATCTGCTA 403

Query: 115 AAT 113
           | |
Sbjct: 404 AGT 406

Score = 232 (64.4 bits), Expect = 2.0e-26, Sum P(4) = 2.0e-26
Identities = 56/69 (81%), Positives = 56/69 (81%), Strand = Minus / Plus

Query:  69 GTGSTAGCAAATTGTGCGTTTAGGCAGCCACGTGGCCAATAAGGTAGATGTTGTCATAAA 10
           ||  |||||| |      |||| |||||||| ||||||||||||||||| ||||||||||
Sbjct: 441 GTCCTAGCAATTGTGCTTTTTATGCAGCCACATGGCCAATAAGGTAGATATTGTCATAAA 500

Query:   9 GGTGCCCTA 1
           | |||||||
Sbjct: 501 GATGCCCTA 509
```

Figure 3.1 T2/55T7 (COM055) BLASTN Results from the EST Database

```
                                                                   Smallest
                                                                     Sum
                                                            High  Probability
                                                           Score     P(N)      N

gb|T84468|T84468      yd47d09.r1 Homo sapiens cDNA clone...204  3.0e-07   1

gb|T84468|T84468 yd47d09.r1 Homo sapiens cDNA clone 111377 5'.
            Length = 495

Plus Strand HSPs:

Score = 204 (56.4 bits), Expect = 3.0e-07, P = 3.0e-07
Identities = 48/57 (84%), Positives = 48/57 (84%), Strand = Plus / Plus

Query:  94 GATCAAGATCTTGTTAAAACCTTGAGTTGTTTAACTATGATAATCACTCCTGTCTTT 150
           || |||||||||||  ||||| ||||||||||  ||||||||||||| |||| | ||
Sbjct: 234 GAGCAAGATCTTGTGCAAACCCTGAGTTGTTTGTCTATGATAATCACACCTGGCATT 290
```

Figure 3.2a T2/73B (COM045) BLASTN Results

```
GAATTCGTAG GAAGATCTAA AATCAGATGG TGTGTTAAAT GCAGCTTATA
ATATCCTTTT TTGTTGATTT TAGGATCAAG ATCTTGTTAA AACCTTGAGT
TCCTGTCTTT GCTGAGGTAA GTTTTCCTTT CAAATTAGTT TTTAAATTAT
TCATTTTGTA TCATAATCCC TAGCAGATAA AAATTATAGA ACTTAAAAGA
ATGCCXTTTT TTTTTTCAAA
```

🟩 Splice acceptor

🟨 Conserved region, probably an exon

🟥 GTAAG Splice donor

Fiqure 3.2b T2/73B (COM45) Potential Exon

Figure 3.2 T2/73B (COM045) BLASTN Results and Potential Exon

The Type III receptor is a transmembrane proteoglycan which may facilitate the binding of TGF-β to the type II receptor. Certain tumour cell lines, which are resistant to the growth inhibitory effects of TGF-β, lack the expression of the type II receptor; thus abnormalities in TGF-β and its receptors may be involved in the loss of growth control often observed during carcinogenesis (Bae *et al.*, 1995) and (Moustakas *et al.*, 1993).

TGF-β I inhibits the proliferation of normal epithelial cells and diminishes the responsiveness to TGF-β I is a feature of most carcinoma cells. It binds to TGF-β RII, forming a duplex which recruits the type I receptor to create an active complex. Mutations in TGF-β RII have been found in human cancers of the colon, stomach, head, lung and neck (Takenoshita *et al.*, 1996).

Sequence data from T2/73AT7 (COM052) scored a significant match to human mRNA for TGF-beta type II receptor alpha (TGF-betaIIR) (Ogasa *et al.*, 1996). Figure 3.3 presents the BLASTN and, unusually, there is 100% sequence identity between the human and chicken sequence. A closer examination of the sequences shows that the alignment is in 3' UTR. This may explain why there is no BLASTX results, as the database match is in the untranslated region. The T2/73AT3 sequence showed no homology to the TGF-βIIR gene and probably represents an intron. A BLAST search was carried out with the entire human mRNA TGF-betaIIR sequence. This significant hits to human TGF-β receptors rat, mouse and chicken TGF-beta type II receptor mRNAs were observed. Its is therefore likely that T2/73AT7 contains the 3' UTR of the chicken transforming growth factor-beta type II receptor. As figure 3.3 shows, bases 4208-4507 of 3'UTR of TGF-βIIR aligned with the chicken sequence. A BLASTN search was carried out to see if this region has been sequenced in other species. The single significant BLASTN results was against itself, therefore this region has not been found in other species. This is not surprising given the number of cDNAs and ESTs that are in the databases.

```
                                                              Smallest
                                                                Sum
                                                       High  Probability
                                                      Score  P(N)       N

dbj|D50683|D50683    Homo sapiens mRNA for TGF-betaII... 531  1.1e-93    4

dbj|D50683|D50683 Homo sapiens mRNA for TGF-betaIIR alpha, complete cds
Length = 5759

Minus Strand HSPs:

Score = 220 (60.8 bits), Expect = 1.1e-93, Sum P(4) = 1.1e-93
Identities = 44/44 (100%), Positives = 44/44 (100%), Strand = Minus / Plus

Query: 297 GCTCCCAGCCTTCATCCTTTTCTAAAAAGGAGCAAATTCTCACT 254
            ||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4208 GCTCCCAGCCTTCATCCTTTTCTAAAAAGGAGCAAATTCTCACT 4251

Score = 265 (73.2 bits), Expect = 1.1e-93, Sum P(4) = 1.1e-93
Identities = 53/53 (100%), Positives = 53/53 (100%), Strand = Minus / Plus

Query: 253 TAGGCTTTATCGTGTTTACTTTTTCATTACACTTGACTTGATTTTCTAGTTTT 201
            |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4253 TAGGCTTTATCGTGTTTACTTTTTCATTACACTTGACTTGATTTTCTAGTTTT 4305

Score = 531 (146.7 bits), Expect = 1.1e-93, Sum P(4) = 1.1e-93
Identities =107/108 (99%), Positives = 107/108 (99%), Strand = Minus / Plus

Query: 204 TTTCTATACAAACACCAATGGGTTCCATCTTTCTGGCTCCTGATTGCTCAAGCACAGT 146
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4303 TTTCTATACAAACACCAATGGGTTCCATCTTTCTGGCTCCTGATTGCTCAAGCACAGT 4361

Query: 145 TTGGCCTGATGAAGAGGATTTCAACTACACAATACTATCATTGTCAGGA 97
            |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4362 TTGGCCTGATGAAGAGGATTTCAACTACACAATACTATCATTGTCAGGA 4410

Score = 480 (132.6 bits), Expect = 1.1e-93, Sum P(4) = 1.1e-93
Identities = 96/96 (100%), Positives = 96/96 (100%), Strand = Minus / Plus

Query:  96 TATGACCTCAGGCACTCTAAACATATGTTTTGTTTGGTCAGCACAGCGTTTCAAAAAGTG 37
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4412 TATGACCTCAGGCACTCTAAACATATGTTTTGTTTGGTCAGCACAGCGTTTCAAAAAGTG 4471

Query:  36 AAGCCACTTTATAAATATTTGGAGATTTTGCAGGAA 1
            ||||||||||||||||||||||||||||||||||||
Sbjct:4472 AAGCCACTTTATAAATATTTGGAGATTTTGCAGGAA 4507
```

Figure 3.3 T2/73AT7 (COM052) BLASTN Results Showing the TGFβRII Gene

3.2.6 C + G and CpG Island Content

Previous studies, suggest that microchromosomes may be more GC/CpG rich than macrochromosomes (McQueen *et al.*, 1996). A CpG island was defined as having a O/E CpG content of greater than 0.6 and a %GC of more than 50% (Antequera and Bird, 1993). The sample studied here enabled routine searches of unique sequence for %GC and CpG island content to be set up. The GC content was examined by constructing a plot of % GC against size of linkage group. The GC content appears to be uniform over the whole genome, regardless of linkage group size (Figure 3.4) except for two examples on the microchromosomes.

A plot of O/E CpG content vs. % GC was carried out to test CpG island characteristics (Figure 3.5). Two potential CpG islands were isolated from microchromosomal DNA (Chromosomes 8 and 9). These 'islands' were sequences which stood out from the bulk DNA and matched the criteria (Antequera and Bird, 1993). Figure 3.5 also shows two clones which had a high observed over expected CpG content, but did not have a %GC of greater than 50% and are therefore not potential CpG islands.

The bulk DNA has an average O/E CpG content of 0.14 and an average %GC of 41.16 (Table 3.1) and this is similar to that published for whole DNA. This was confirmed by plotting the observed over expected CpG ratio against linkage group size (Figure 3.6). As with GC content, CpG ratios appear to be fairly uniform regardless of linkage group size. The two GC rich, potential CpG islands are separate from these. As only two GC rich microchromosomes clones were isolated, it is not known if this a true reflection of all microchromosomes.

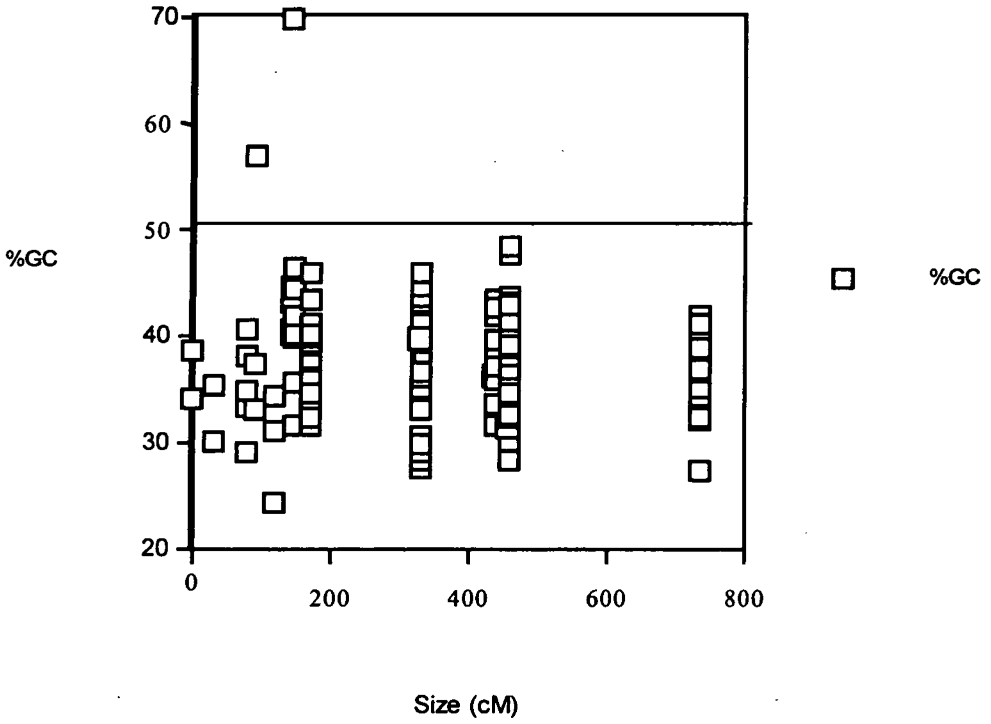


Figure 3.4 % GC vs. Size of the Linkage Group

The GC content appears to be uniform regardless of linkage group size. Macrochromosomes 1 to 5 + Z are shown, with 1 being the largest linkage group. The smaller linkage groups are either on microchromosomes or they have not yet been linked to the larger linkage groups. Two clones stand out as having a high %GC, both of which map to a microchromosome.

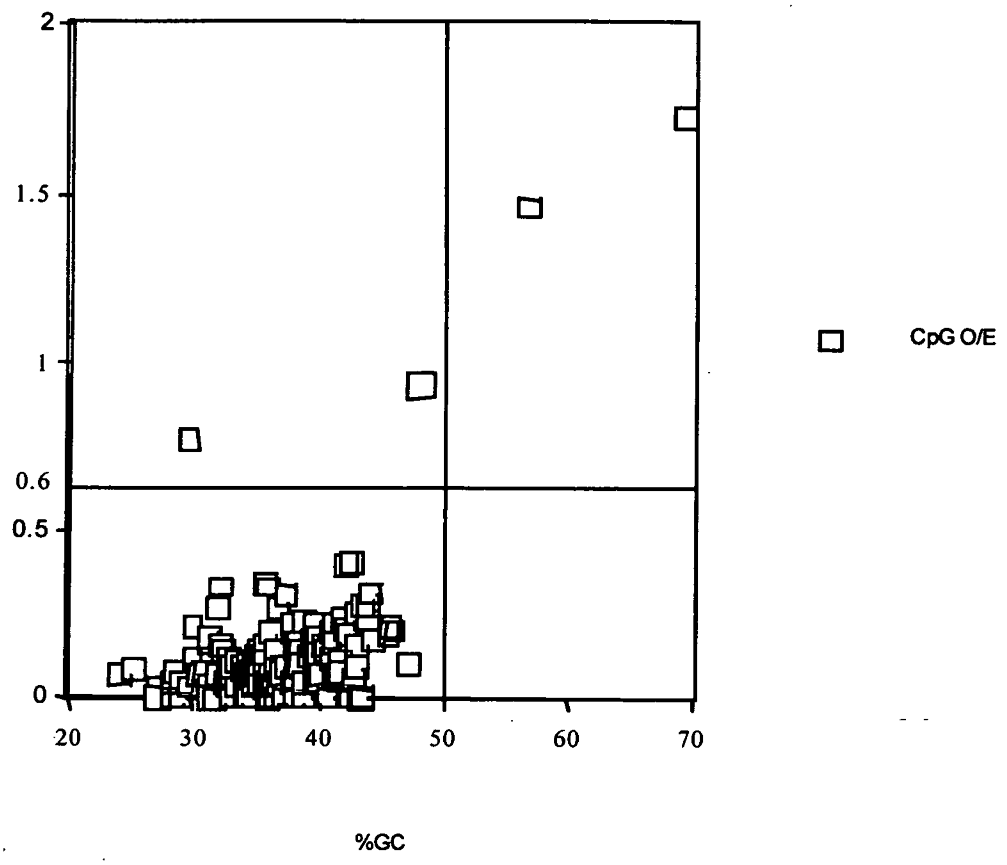


Figure 3.5 Observed Over Expected CpG Content vs. % GC

Two distinct groups are shown. The largest is the bulk DNA, which is not overly GC rich and therefore has no potential CpG islands. The second group shows two CpG islands, which map to a microchromosomes 8 and 9. Two clones have a high CpG O/E but a %GC of below 50%.

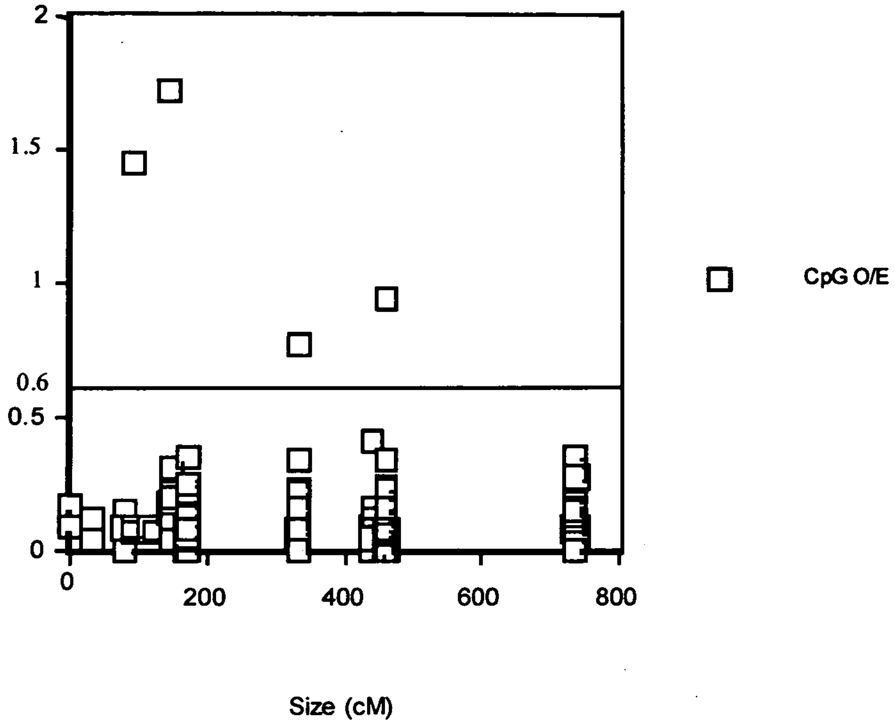


Figure 3.6 Linkage Group vs. Observed Over Expected CpG Content

The GC content is seen to be to be fairly uniform, regardless the size of the linkage group. Therefore, GC rich clones are separate from the rest of the DNA.

3.2.7 Distribution of Repeats

Database searches were carried out to determine which types of repeats are found in the chicken genome, and assess their distribution and their relation to gene density. Repeats appear to be common in this sample. Of the 89 sequences analysed, four were found to have CR1 elements and thirteen were simple repeats such as microsatellites.

3.2.8 Simple Repeats

Thirteen simple repeats were found (Table 3.3) and represent potential new markers. Eight of these repeats were on clones mapping to a macrochromosome, three to a microchromosome, one to linkage group 21, and one clone was unlinked.

In the 89 clones analysed, an AT repeat was the most common repeat type. A total of eight of these repeats were found, five of which mapped to a macrochromosome, two to a microchromosome and one to linkage group 21. Other repeat types were also found but at lower frequencies. Two T repeats were found on macrochromosomes 2 and 5, a single CT repeat mapped to macrochromosome 1 and the clone containing the poly A repeat had an unlinked map position. The clone mapping to microchromosome 9 carried a CA repeat.

3.2.9 CR1 Elements

Of the eighty nine clones analysed, four were found to have a CR1 element (Table 3.4) all of which mapped to a macrochromosome. With the small sample size it is impossible to draw any conclusions about the distribution of CR1 repeats between macrochromosomes and microchromosomes. Chapter 6 analyses CR1 repeats in more detail with a larger dataset.

Clone	Chromosome	Locus	Repeat Type
T2/22T3	1	*COM037*	AT
T2/72ABT7	1	*COM060*	CT
T2/73BT7	2	*COM045*	T
T2/11T3	2	*COM022*	AT
T12/91T7	2	*COM067*	AT
T2/58T7	3	*COM095*	AT
T3/5T7	3	*COM029*	AT
T9T3	5	*COM011*	T
T2/28T3	7	*COM049*	AT
T12/74T7	8	*COM085*	AT
T4/13T7	9	*COM073*	CA
U/26T7	LG21	*COM075*	AT
R8T7	UN	*COM031*	A

Table 3.3 Simple-, di- and mono-nucleotide Repeats

LG-Linkage group; UN-Unlinked

Clone	Chromosome	Locus
T5/24T3	1	*COM041*
T2/82T3	2	*COM098*
T12/108T7	2	*COM023*
T2/70ABT7	4	*COM053*

Table 3.4 CR1 Elements

3.3 Discussion

3.3.1 Sequencing Experiments

One of the aims of sequencing the Compton clones was to establish procedures, such as the Staden program and database searches. It was also used to test the hypothesis that microchromosomes are gene dense, that CpG islands are concentrated on microchromosomes and about the distribution and types of repeat found throughout the genome. Due to the small sample size, however only suggestions as to genome content can be made.

3.3.2 Database Homologies

The criteria used when deciding whether a sequence match from the database searches was significant and indicated homology, proved effective, as a total of three significant hits against two human ESTs and the human TGF-betaIIR gene, were recorded. False matches would have had scores of lower than 150 and a probability of 10^{-9} for a BLASTN search and 75 and a probability of 10^{-6} for a protein-protein BLASTX search. These three database homologies can be used to update the chicken/human comparative map. The sequence sampling did find gene homologies, but is not effective with such small fragments and more DNA is needed. Sampling cosmid clones should prove more successful as is demonstrated in chapter 5.

A total of 64,948 bp of unique DNA was sequenced, with 44,322 bp on the macrochromosomes and 12,953 bp on the microchromosomes; this was the equivalent of sequencing approximately two cosmids. Previous work predicts a gene every 30 kb on the macrochromosomes and a gene every 10 kb on the microchromosomes (McQueen *et al.*, 1996). The work described in this chapter predicted one gene every 22 kb on the macrochromosomes and one gene every 12 kb on the microchromosomes. The discrepancy could reflect the size of the dataset, or a gene density which is lower than was previously predicted. Sequence sampling of random cosmids (Chapter 5) should resolve this.

3.3.3 Conservation of Gene Structure

Examples of conserved gene structure were found from the database searches with hits against non-coding regions of the genome. The database hits against the 3' end of a human cDNA and likely to be a conserved regulatory sequence. Bacterial endoribonucleases such as RNase E play a role in mRNA degradation, and human mRNAs from short-lived oncogenes and growth factors have 3' UTRs which play a role in the control of mRNA stability. These regions contain the nucleotide motif AUUUA which is important for mRNA stability. This motif is recognised by *E.coli* RNase E and its human counterpart and cleavages are introduced in the 3'UTR, leading to mRNA decay (Wennborg *et al.*, 1995). The T2/55T7 sequence is A-rich indicating that this region may contain the AUUUA motif and be involved in mRNA stability.

The alignment between T2/73AT7 sequence and the 3'UTR region of the human TGF-β RII gene had 100% sequence identity. This may be an example of a highly conserved region (HCR). HCRs are found in the non-coding parts of genes and have been observed between species that have diverged over 300-350 Mya (Hedges, 1994); (Nanda *et al.*, 1999). HCRs between birds and mammals include the dystrophin gene 3' UTR which has three highly conserved regions. A long conserved region in the 3'UTR of the BTG1 antiproliferative genes has also been observed between human and chicken (Duret *et al.*, 1993). HCRs are frequent in genes essential to cell life and are frequent in the 3' non-coding regions. The 3'-HCRs are AT-rich and lie in the 3' transcribed region of the gene, suggesting that they have a role in post-transcriptional processes. They are found less frequently in promoters and introns.

3.3.4 GC Content

Analysis of bulk DNA gave an average GC content of 38%. Previous work has indicated that microchromosomes appear to be more GC-rich than macrochromosomes (McQueen *et al.*, 1996) and so it was expected that as chromosome size decreased, the %GC would increase; both GC rich clones mapped to a microchromosome. So few GC rich clones may be due to the GC-rich portions of the genome not being cloned and therefore not sequenced. The DNA samples were prepared using a variety of restriction digests which could discriminate against GC-rich region being cloned. A second possibility is that microchromosomes are not as GC rich as is traditionally held. The randomness of the clones can answer this. The Compton clones appear to be a random sample, with a distribution of 75:25 macrochromosomes:microchromosomes respectively. Based on this, the observed GC content is most likely an accurate reflection of the total genome. Therefore, the second explanation suggested above, that microchromosomes are less GC-rich than previous estimates may be true.

3.3.5 CpG Island and Gene Content

The GC and CpG island content of the sequences was determined to examine the questions of whether CpG islands are concentrated on microchromosomes, and whether microchromosomes are gene rich and macrochromosomes gene poor. Potential islands were isolated using plots of O/E CpG content Vs % GC and plotting O/E CpG ratio against linkage group size. The CpG islands identified satisfy the criteria of O/E 0.6 and GC% of 50% and are distinct from bulk DNA (Antequera and Bird, 1993). As both CpG islands map to a microchromosome, this suggests that microchromosomes are gene dense but, as only a relatively small sample was analysed, it is not known if this is significant or not. A larger dataset would resolve the question of whether microchromosomes are more CpG island rich and have a higher gene density than macrochromosomes.

3.3.6 Distribution of Repeats

Database searches were carried out to see whether repeats could be found by sequence sampling and, if so, how they are distributed; also, if microchromosomes are indeed more gene dense than macrochromosomes, they will have fewer repeats. This may be due to the microchromosomes containing less 'junk' DNA.

Repeats were located by sequence sampling, and both, simple and species-specific repeats, were found and represent potential new markers. In Chapters 5 and 7, sequence sampling data will be analysed with the Staden program but at this stage of the study the Staden package was not available within the laboratory, and BLAST searches were carried out manually with the NCBI server. A feature of BLAST searches is that simple repeats are filtered out, therefore additional unfiltered BLAST searches were carried out. This was time consuming, although successful at finding repeats. In chapter 5 the program Report Repeat, which masks non-repeat like sequence, was used to identify simple repeats (Law, Personal Communication).

3.3.7 Distribution of CR1 Repeats

Four CR1 repeats were identified which all mapped to macrochromosomes. CR1 repeats are members of the non-LTR class of retrotransposable elements which retrotranspose by a 'nick and prime' mechanism similar to other families of non-LTR elements (Haas *et al.*, 1997). CR1 repeats may be more likely to transpose themselves into regions rich in 'junk' DNA as insertion into an exon would have a deleterious effect. CR1 repeats are therefore expected to be more frequent on the macrochromosomes. CR1 data from this chapter have been combined with random cosmid CR1 sequences to address the question of CR1 bias on macrochromosomes, and is discussed in more detail in Chapter 6.

3.3.8 Distribution of Simple Repeats

A total of thirteen simple repeats were found, of which eight mapped to a macrochromosome and five to a microchromosome. This bias towards macrochromosomes may have arisen because simple repeats evolve by expansion. The greater the junk DNA content on the macrochromosomes would make expansion easier, and as a result simple repeats may be more likely.

The most common repeat motifs were AT repeats, which were found on both chromosome types. A single microchromosomal CA repeat was found. It has been reported, that repeats are evenly distributed over the macrochromosomes and intermediate chromosomes, with low concentrations on the microchromosomes (Primmer *et al.*, 1997), and so a greater number of CA repeats would have been expected. This difference in numbers of CA repeats may reflect the primed *in situ* labelling technique (PRINS) approach used (Volpi and Baldini, 1993). Distribution of microsatellite sequences is discussed in Chapter 5.

3.4 Conclusion

The work carried out in this chapter highlights the effectiveness of a sequence sampling approach. Gene homologies and repeats were identified, suggesting it is an effective means of gene discovery and method of characterising genome organisation. In future studies a larger DNA sample, such as a cosmid, would be better suited as it is easier to map than a plasmid. This chapter has provided preliminary estimates of gene homologies, chicken DNA %GC content, O/E CpG, CR1 and simple repeat content. Procedures which were used in later work were established, and the need for faster and simpler analysis of raw DNA sequence was highlighted. Significant data on the differences between macrochromosomes and microchromosomes required a larger data set. Therefore, randomly selected cosmids were sequence sampled. The mapping of these cosmids is described in Chapter 4, the question of gene density is discussed in Chapter 5 and the evolution and distribution of CR1 elements in Chapter 6.

Chapter 4

Integration of Genetic and Physical Maps

4.1 Introduction

Randomly selected cosmids were mapped to either a macrochromosome or a microchromosome (see 4.1.1 for definition) physically by FISH and genetically by either PCR length variants or SSCP analysis. This was carried out as part of the gene density study (Chapter 5) and to contribute to the integration of the physical and genetic linkage maps of the chicken.

4.1.1 The Chicken Karyotype

The standard chicken karyotype distinguishes chromosomes 1-8 and the sex chromosomes, Z and W, by GTG- and RBG-banding studies carried out by the International Committee for the Standardisation of the Avian Karyotype (Ladjali *et al.*, 1999). This numbering system only indicates which chromosomes can be identified cytologically.

The definition of macrochromosomes and microchromosomes used in this chapter classifies the macrochromosomes as chromosomes 1-8 and the Z chromosome. The thirty microchromosomes are defined as the remaining chromosomes and the W chromosome. As they are hard to distinguish individually, microchromosomes were initially ordered by decreasing size. Recently, however, it has become possible to identify sixteen pairs of chicken microchromosomes by two-colour FISH (Fillon *et al.*, 1998). This allows genetic mapping data to be related to physical maps of the macrochromosomes.

4.1.2 Genetic Mapping in the Chicken Genome

Efforts to map the chicken genome have produced large amounts of genetic mapping information, with the genetic map close to completion[3] (Smith and Burt, 1998). The three main genetic maps are based upon the Compton (Bumstead and Palyga, 1992), East Lansing (Crittenden *et al.*, 1993) or Wageningen (Groenen *et al.*, 1998) reference populations. There is, however, little physical mapping data as few genomic clones have been mapped in this manner. Data from this investigation and from other laboratories involved in the EC-Chickmap project will rectify this. This is currently being addressed by the FISH mapping of cosmid (Buitkamp *et al.*, 1998) and bacterial artificial chromosome (BAC) clones (Crooijman, 1998).

4.1.3 Integration of the Physical and Genetic Maps

To allow integration of the chicken physical and genetic maps, linkage groups must be assigned to both the macrochromosomes and microchromosomes: for example, the chicken gene for insulin-like growth factor 1 was physically assigned to the short arm of chromosome 1, close to the centromere and genetically to position 149.9 on linkage group E01C01C11W01 (Klein *et al.*, 1996). A number of other expressed loci have also been mapped to this region of chromosome 1 such as the genes for the avian leukosis viral proteins ALVE1 and ALVE6A and the histone *H5* and *LYZ* genes. In addition, the *PGR*, *GAPD* and *OTC* genes have been mapped to the long arm of chromosome 1. Genetic linkage mapping with the East Lansing reference cross has produced markers for these genes on linkage group E01C01C11W01. This work has localised the *IGF1* and *GAPD* genes to the short and long arms respectively of the chromosome, on either side of the centromere. This information was used to position the chromosome 1 centromere between these two genes.

[3] Complete genetic map i.e. a new genetic marker has a greater than 95% chance of linkage to another previously mapped marker

Table 4.1 summarises loci which have been both genetically and physically mapped to chromosomes 1-8 and Z. The integration of the two kinds of map has allowed comparisons to be made between the chicken map and other vertebrate maps (Klein *et al.*, 1996). The subject of comparative mapping is discussed in Chapter 7.

This work has also aided the orientation of linkage groups. The physical and genetic mapping of chicken genomic clones is underway to achieve this, mostly as part of EC Chickmap project (Burt and Cheng, 1998); eight new polymorphic markers, isolated from BAC and P1-derived artificial chromosome (PAC) clones, were localised to the macrochromosomes by genetic mapping on the East Lansing and Compton reference families (Morisson *et al.*, 1998) and the clones physically mapped by FISH. It is now possible to align physical and genetic maps for chromosomes 1-8 and the Z chromosome, in conjunction with the cosmid mapping data presented here. The orientation of linkage groups for chromosomes 3 and 4 has also been established. Only the linkage group from chromosome 8 amongst the macrochromosomes remains to be orientated.

4.1.4 Genetic and Physical Mapping of Random Cosmid Genomic Clones

Sixteen random cosmids were genetically mapped onto the East Lansing map by either SSCP analysis (Beier, 1993) or by analysing DNA length variants from PCR products and physically mapped by FISH (Smith *et al.*, 1999). For this thesis, seven of these cosmids have been genetically mapped and were localised to a specific macrochromosome. This was done to allow comparisons of features such as gene density, CpG content and repeat distribution between macrochromosomes and microchromosomes to be made. This will also provide markers for two-colour FISH experiments, which will help to identify individual microchromosomes (Fillon *et al.*, 1998). Physical and genetic mapping data were submitted to the chicken genome database at http://www.ri.bbsrc.ac.uk/chickmap/, which is maintained at the Roslin Institute.

Table 4.1 Loci Which Have Been Genetically and Physically Mapped in Chicken

* - FLpters estimated from published cytogenetic band positions

Locus	Physical	FLpter	Genetic Position		References
	Position		EastLansing	Compton	
Chromosome 1					
ALVE6A	1p26	0.02*	0.0		(Levin *et al.*, 1994)
GCT0006	1p24-p22	0.13	64.5	10.1	(Morisson *et al.*, 1998)
LYZ	1p22-p15	0.20*	106.3	61.6	(Bumstead and Palyga, 1992);(Sang, Personal communication)
GCT0015	1p22-p21	0.20*	108.6	67.0	(Morisson *et al.*, 1998)
ROS0147	1p21-p15	0.22*	124.6		(Smith *et al.*, 1999)
H5	1p14-p13	0.25*	140.3		(Crittenden *et al.*, 1993);(Levin *et al.*, 1994);(Cheng *et al.*, 1995)
IGF1	1p12-p11	0.31	149.9		(Klein *et al.*, 1996)
ALVE1	1p12-p11	0.36*	189.2		(Ponce de Leon *et al.*, 1991);(Bumstead and Palyga, 1992)
GAPD	1q11-p12	0.47*	214.7	260.5	(Cheng and Crittenden, 1994); (Burt, 1994)
GCT0013	1q14	0.57	322.7		(Morisson *et al.*, 1998)
OTC	1q13-q14	0.57*	325.0		(Shimogiri *et al.*, 1993)
GCT0007	1q31-q35	0.79	398.3		(Morisson *et al.*, 1998)
PGR	1q42-q44	0.95*	464.5	578.2	(Dominguez-Steglich *et al.*, 1992); (Toye *et al.*, 1997)
Chromosome 2					
OVY	2q11-q12	0.49*	219.3	221.3	(Dominguez-Steglich *et al.*, 1992)
GCT0023	2q11-q12	0.53	249.4	250.4	(Morisson *et al.*, 1998)
Chromosome 3					
GCT0011	3p11-q11	0.16*	72.4		(Smith *et al.*, 1999)
TGFB2	3q22-q23	0.23	81.6	14.7	(Morrice and Burt, 1995)
MYB	3q24-q26	0.53*	196.4		(Symonds *et al.*, 1984); (Soret *et al.*, 1990)
GCT0008	3q27-q28	0.62	195.0	149.1	(Morisson *et al.*, 1998)
ROS0119	3q27-q29	0.63	218.0		(Smith *et al.*, 1999)
GCT0019	3q28q2.10	0.68	225.6	187.7	(Morisson *et al.*, 1998)
Chromosome 4					
PGK1	4p14-p11	0.11*	49.10	71.01	(Rauen *et al.*, 1994); (Spike *et al.*, 1996)
Chromosome 5					
TH	5q12	0.34*	42.8		(Dominguez-Steglich *et al.*, 1992)
ROS0099	5q12	0.34	42.8	22.2	(Smith *et al.*, 1999)
TGFB3	5q21-q22	0.60*	113.1		(Burt *et al.*, 1995); (Burke *et al.*, 1994)
Chromosome 6					
ROS0160	6q12	0.41	52.4		(Smith *et al.*, 1999)
SCD1	6q14	0.58	58.5		(Fillon *et al.*, 1997); (Pitel *et al.*, 1998)

continued

Table 4.1 continued

Locus	**Physical**	**FLpter**	**Genetic Position**		**References**
	Position		**EastLansing**	**Compton**	
Chromosome 7					
ROS0121	7p11	0.26	E27		(Smith *et al.*, 1999)
Chromsome 7					
RPL37A	7q12-q14	0.57*	62.3		(Nanda *et al.*, 1996); (Girard-Santosuosso *et al.*, 1997)
ROS0128	7q13	0.58	68.1		(Smith *et al.*, 1999)
NRAMP	7q13	0.60	76.8	62.5	(Hu *et al.*, 1995); (Girard-Santosuosso *et al.*, 1997)
Chromosome 8					
RPL5	cen		51.4	80.2	(Nanda *et al.*, 1996)
Z Chromosome					
ATPAL1	Zp24-p23	0.06*	20.0		(Fridolfsson *et al.*, 1998)
CHDIZ	Zp12-p13	0.60	143.4		(Griffiths and Korn, 1997)
ACO1	Zq21	0.77	198.7		(Saitoh *et al.*, 1993); (Smith and Cheng, 1998)

4.2 Results

This chapter describes the genetic mapping of random cosmid clones and how these data have been used as part of the effort to integrate the physical and genetic maps of the chicken genome.

4.2.1 Genetic Mapping of the Random Cosmids

Cosmid clones numbered 27, 28, 30, 32, 33, 34 and 35 were randomly picked from a commercially available chicken cosmid library (CLONTECH, California). Seven new markers were developed and genetically mapped. For Cosmid 28 (*ROS0105*) the size differences of PCR products were used in the analysis (see Figure 4.1). The other cosmids, 27 (*ROS0110*), 30 (*ROS0108*), 32 (*ROS0149*), 33 (*ROS0107*), 34 (*ROS0120*) and 35 (*ROS0150*) no size differences with the PCR products were observed, therefore SSCP analysis of PCR products was carried out. Linkage analysis was carried out using the Map Manager program (Manly, 1993) and their positions, with reference to the East Lansing map, are shown in Table 4.2, which also contains marker information.

4.2.2 Physical Mapping of the Random Cosmids

The cosmids were physically mapped by FISH to a chromosome as shown in Table 4.3 (Smith *et al.*, 1999). The FISH mapping of cosmid 29 was carried out for this thesis, and localisation of it to the Z chromosome is shown in Figure 4.2.

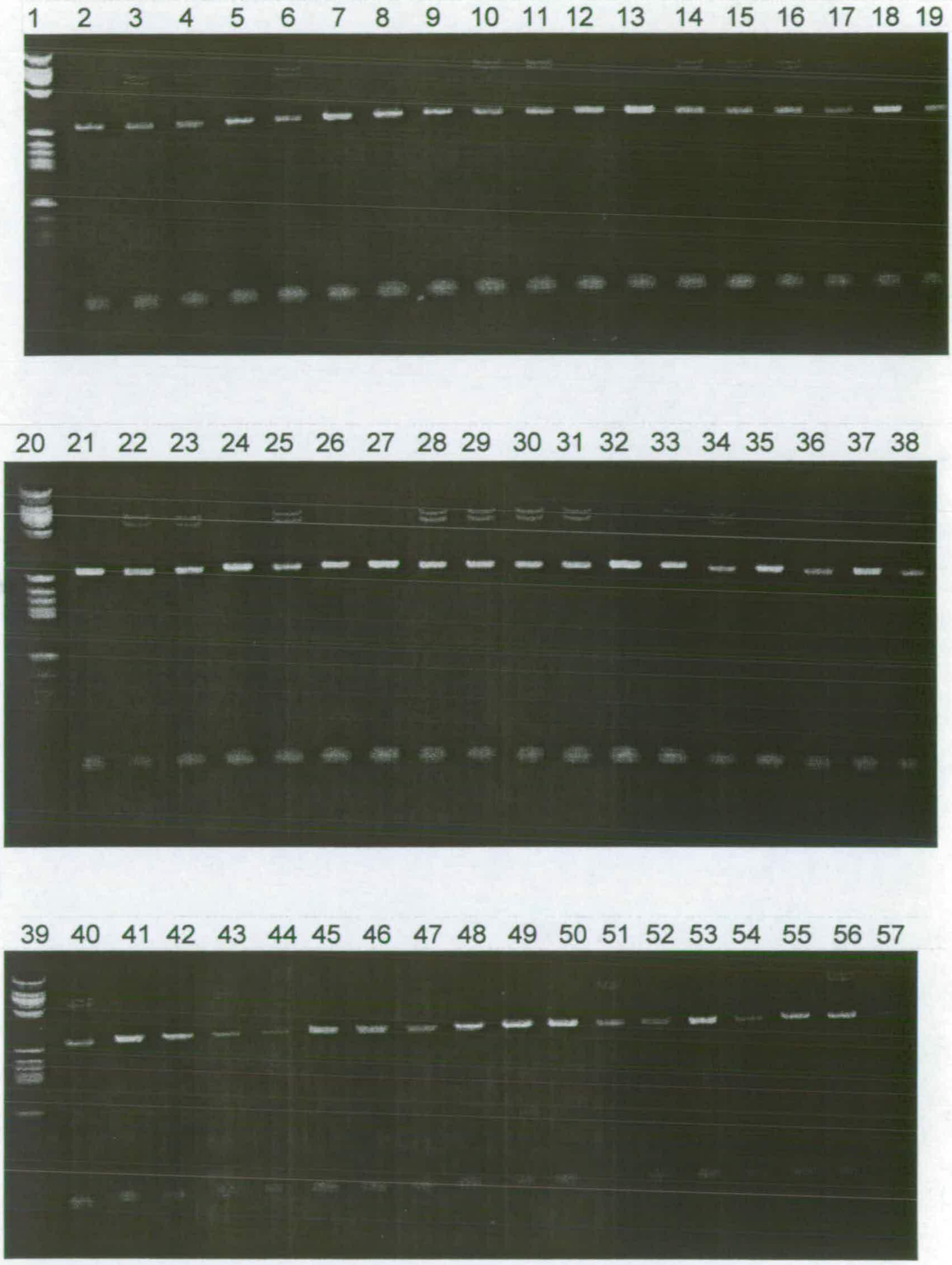


Figure 4.1 Cosmid 28 PCR Products Size Differences

Lanes: 1, 20 and 39 molecular weight marker X (Boehringer Mannheim), 2 Male 3 Female, 4-19, 21-38, 40-57 offspring

Marker	Cosmid	Chromosome	Genetic Map	GenBank
	Clone		Position	Accession
			(cM)	Number
ROS0105	Cos28	2	180.70	AJ231935
ROS0150	Cos35	2	379.23	AJ232107
ROS0120	Cos34	2	464.08	AJ232087
ROS0108	Cos30	3	217.95	AJ231966
ROS0107	Cos33	4	4.01	AJ232060
ROS0110	Cos27	5	113.11	AJ231874
ROS0149	Cos32	8	29.35	AJ232024

Table 4.2 Genetic Markers Used in SSCP and PCR Analysis

Locus	Physical Position	FLpter
Chromosome 2		
ROS0105	2p12-p11	0.26
ROS0150	2q26-p32	0.75
ROS0120	2q32-q35	0.89
Chromosome 3		
ROS0108	3q29-q33	0.63*
Chromosome 4		
ROS0107	4p14-p13	0.04*
Chromosome 5		
ROS0110	5q21-q22	0.60
Chromosome 8		
ROS0149	cen	
Z Chromosome		
ROS0249	Zq11-q12	0.52
Microchromosome		
ROS0250	Micro	-
ROS0251	Micro	-

Table 4.3 Physically Mapped Loci in the Chicken Based on Randomly Selected Cosmid Clones

* - FLpters estimated from given cytogenetic band positions; Micro-Microchromosome

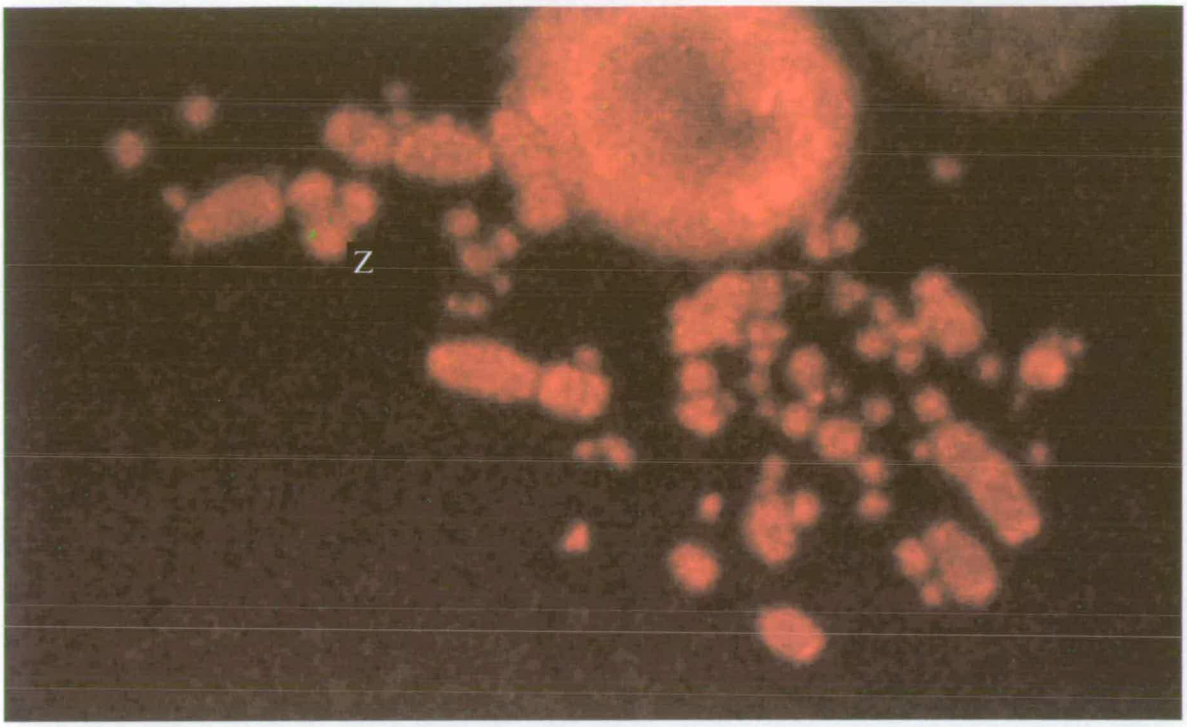


Figure 4.2 FISH Mapping of Cosmid 29 to the Z Chromosome

Fluorescence *in situ* hybridisation of cosmid 29 to metaphase chromosomes of a female (ZW) chicken. Cosmid 29 was labelled with biotin-16-dUTP and detected with avidin-FITC. The DNA was stained with propidium iodide.

4.2.3 Assignment of New Markers and the Physical Orientation of Linkage Groups

This work has assigned seven new genetic markers on the East Lansing map to six chicken chromosomes. On chromosome 2 the three new markers (*ROS0105*, *ROS0150* and *ROS0120*) were used to orientate the linkage group E06C02W02 (Figure 4.3). The orientation of the chromosome 3 linkage group E02C03W0 was established (Figure 4.4) and I added a new marker, *ROS0108*, to the maps. Relative to the genetic map compiled by Bumstead *et al.,* 1996, the linkage group for chromosome 3 should be inverted. The reason for this was that the genetic map was not based on any physical markers.

The orientation of the chromosome four linkage group E05C04W04 was also established (Figure 4.5) and the new marker *ROS0107* contributed to this. Orientation of the chromosome 5 linkage group E07E34C05W05 was established (Figure 4.6) with a new marker, *ROS0110*, being assigned. A new marker, *ROS0149*, was also assigned to chromosome 8 (Figure 4.7). As chromosome 8 is metacentric and so difficult to orientate cytogenetically therefore its orientation with respect to the genetic map is difficult to determine. The clones ROS0149 and RPL5 were mapped to this chromosome, both genetically and physically; this will allow future clones to be mapped with respect to these markers by two-colour FISH. It will be important to obtain high resolution banding alongside FISH signals, when determining orientation of this chromosome in the future.

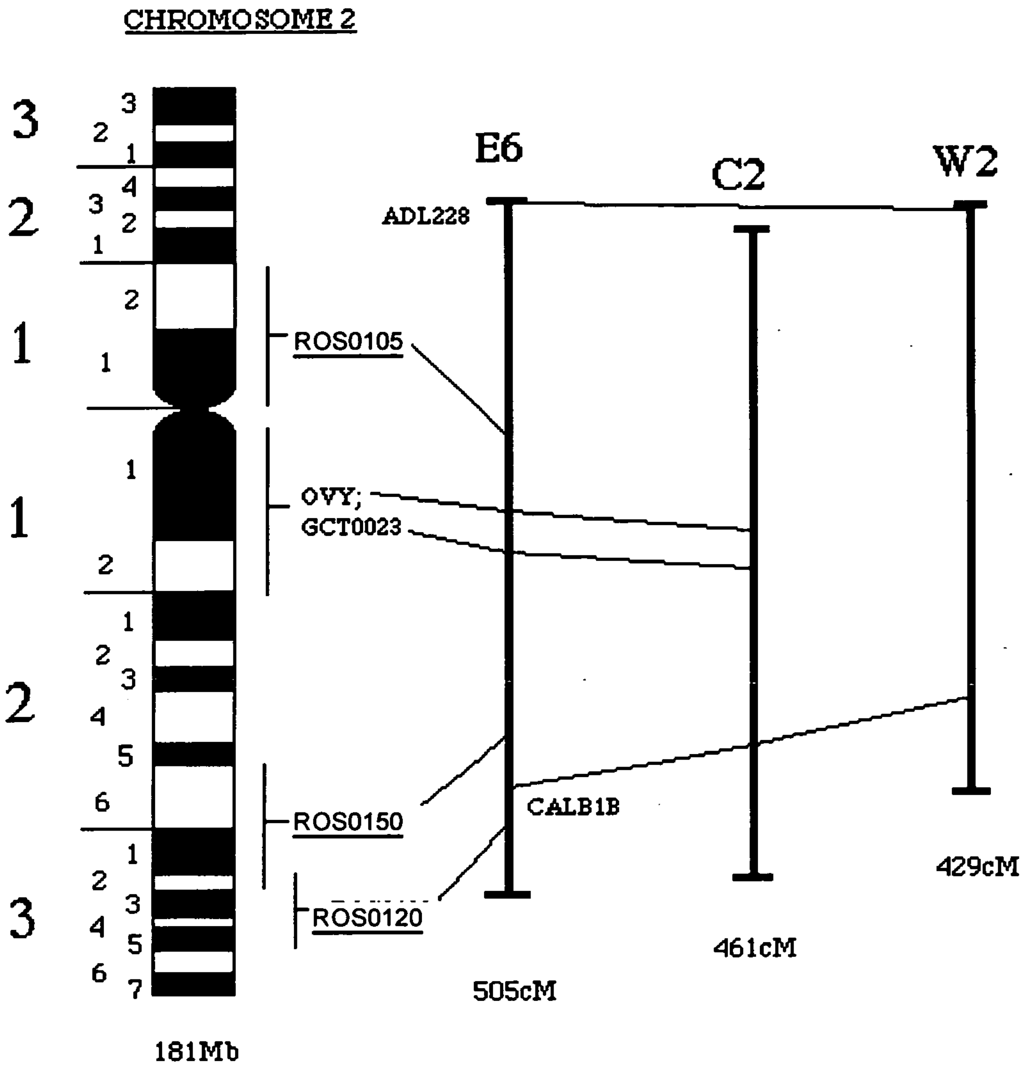


Figure 4.3 Integration of the Physical and Genetic Maps of Chromosome 2

Physical ideograms represent the RBG banding pattern; E- East Lansing genetic linkage group; C- Compton genetic linkage group; W-Wageningen genetic linkage group.

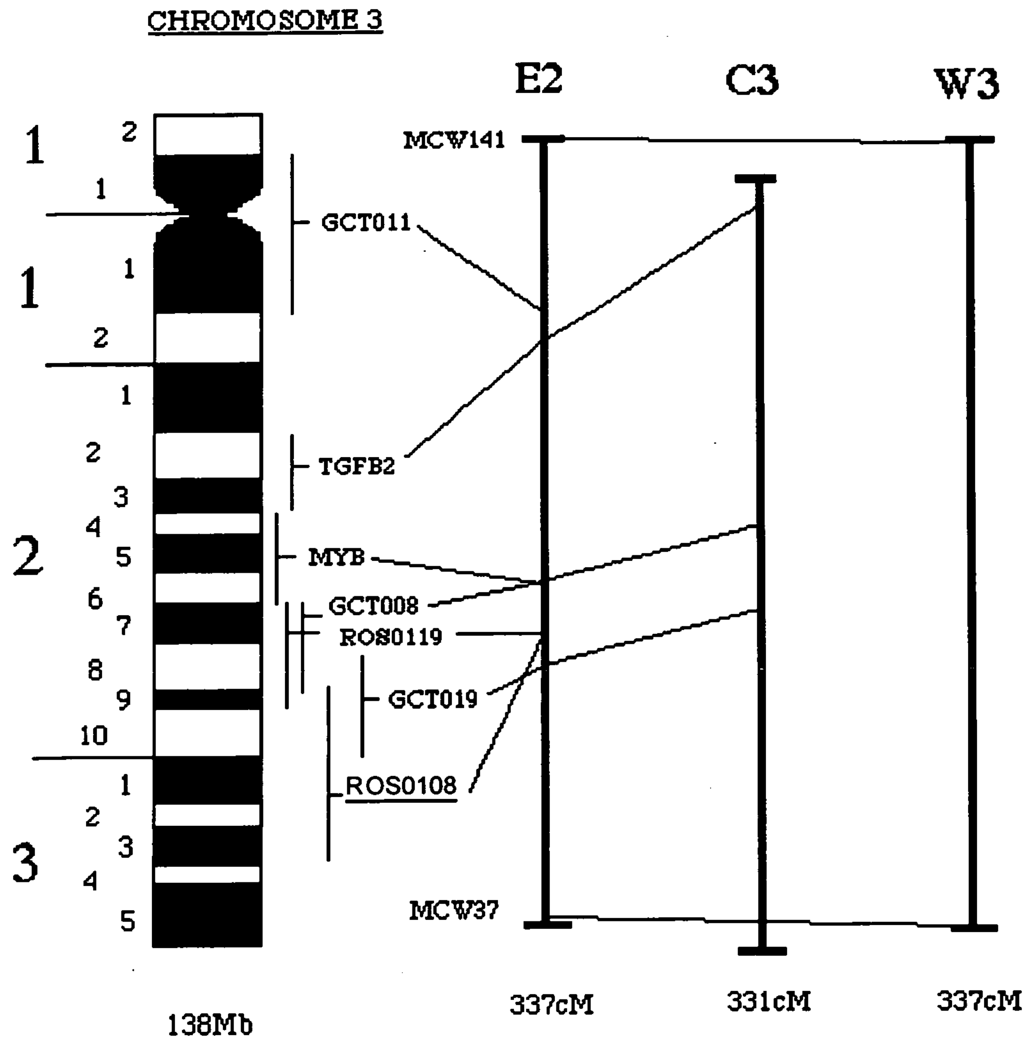


Figure 4.4 Integration of the Physical and Genetic Maps of Chromosome 3

Physical ideograms represent the RBG banding pattern; E- East Lansing genetic linkage group; C- Compton genetic linkage group; W-Wageningen genetic linkage group

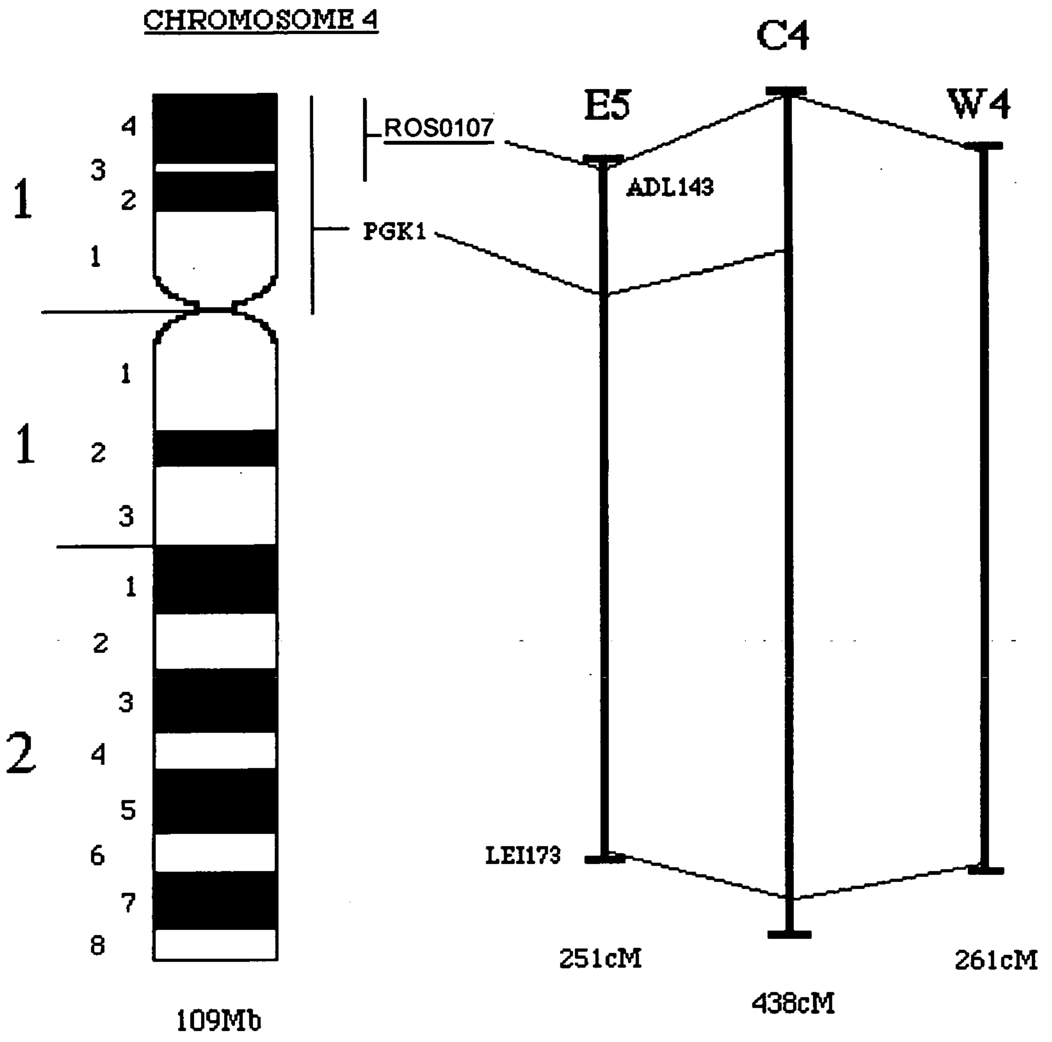


Figure 4.5 Integration of the Physical and Genetic Maps of Chromosome 4

Physical ideograms represent the RBG banding pattern; E- East Lansing genetic linkage group; C- Compton genetic linkage group; W-Wageningen genetic linkage group

CHROMOSOME 5

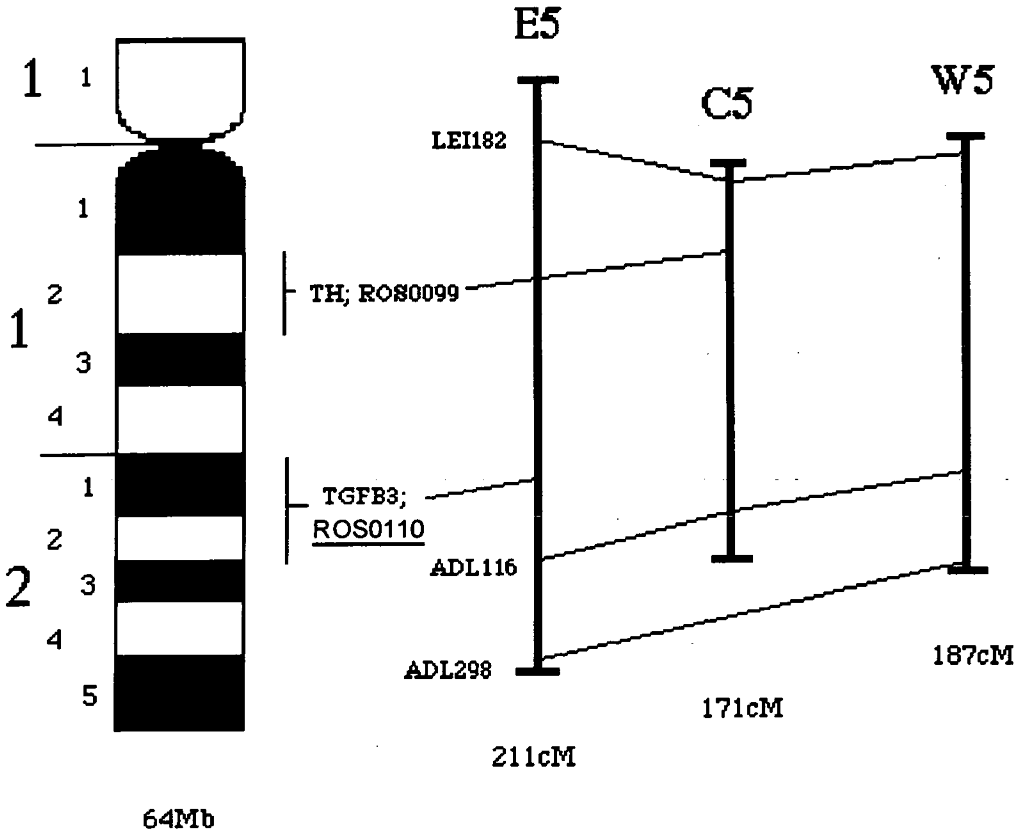


Figure 4.6 Integration of the Physical and Genetic Maps of Chromosome 5

Physical ideograms represent the RBG banding pattern; E- East Lansing genetic linkage group; C- Compton genetic linkage group; W-Wageningen genetic linkage group

CHROMOSOME 8

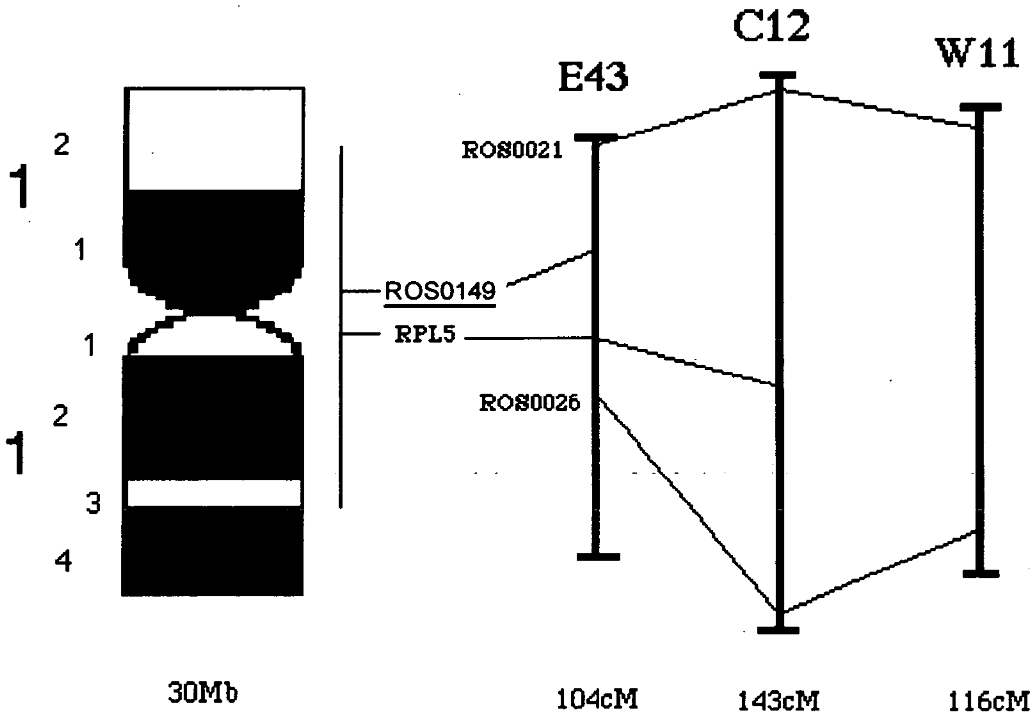


Figure 4.7 Integration of the Physical and Genetic Maps of Chromosome 8

Physical ideograms represent the RBG banding pattern; E- East Lansing genetic linkage group; C- Compton genetic linkage group; W-Wageningen genetic linkage group. This chromosome is metacentric and therefore difficult to orientate. The two markers shown can be used to orientate other markers relative to them.

4.2.4 Integration of the Physical and Genetic Maps

Table 4.3 outlines the physical mapping data generated during this study. Information from this table and Table 4.1 was used to integrate the genetic and physical maps as shown in Figures 4.3-4.7 and discussed below. In these figures, data for the East Lansing, Compton and Wageningen maps are included. Although the data presented here were used to orientate East Lansing linkage groups, common markers on the other two maps allows them to be aligned.

4.2.5 Genetic Coverage

It was possible to correlate genetic (cM) to physical distances (FLpter) in the regions for which data was available. Graphs were constructed for chromosomes 1, 2, 3 and Z (Figure 4.8) and from these, estimations of genetic position from physical data and *vice versa* could be made (Smith *et al.,* 1999). Where there was a lack of direct FLpter data, fractions were estimated from cytogenetic band positions. The slope of the graph was used to estimate coverage of chromosome 2 and, as shown in Table 4.4, it is complete. Although data from this study did not contribute to the coverage of chromosome 1, Table 4.4 describes its genetic coverage and it is complete. The estimated coverage of chromosome 3, is also complete. The Z chromosome linkage group is close to completion with an estimated 86% coverage; the remaining 14% of the chromosome is a heterochromatin block at the telomeric end of the long arm of the chromosome. Prior to this study, it was estimated that this region made up around 20% of the Z chromosome (Saitoh *et al.*, 1993). At the present time there are not enough data to estimate the coverage of chromosomes 4, 5, 6, 7 and 8, highlighting areas for more mapping work.

Figure 4.8 Correlation of Fractional Length Measurement With Genetic Location of Loci Which Have Been Physically and Genetically Mapped

The approximate position of the centromere is marked for each chromosome where vertical and horizontal lines bisect the axes. FLpter- the ratio of the distance of the FISH signal to the telomere of the p-arm divided by the length of the whole chromosome. The equation of the slope from each graph was used to determine the genetic length of each chromosome. Substituting physical values of 0 and 1 into the equation gave the resultant estimated size.

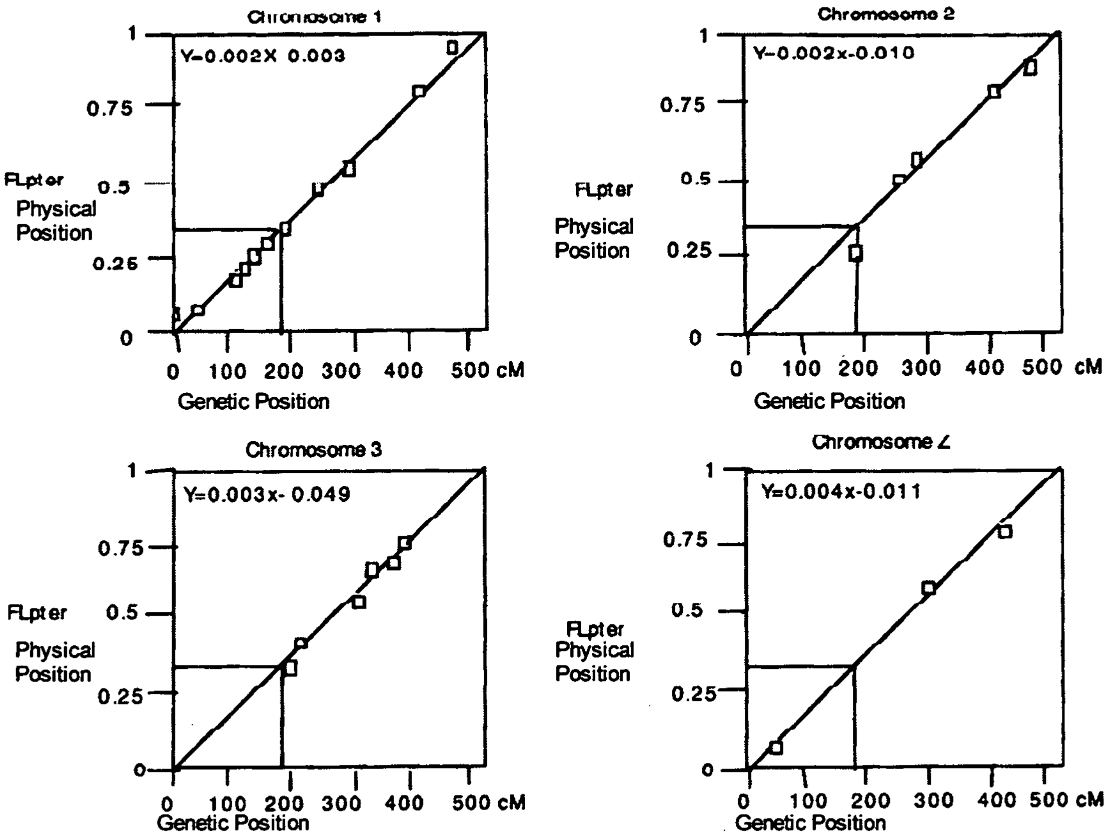


Figure 4.8

		Linkage Groups		
Chromosome	Estimated size (cM)^b		Actual Size (cM)^b	% Coverage
	y=0	y=1		
1	1.5	501.5	515.23	103%
2	5.0	505.0	495.83	99%
3	16.3	349.6	329.03	99%
Z	2.8	252.7	214.23	86%

Table 4.4 Estimated Coverage of East Lansing Linkage Groups

(Smith *et al.*, 1999); ^{b}-(Smith and Burt, 1998)

4.3 Discussion

The overall aim of this thesis is to examine aspects of chicken genome organisation. As part of this, randomly selected cosmids have been sequence sampled and the data used to answer questions such the gene density on macrochromosomes versus microchromosomes and the distribution of repeats. To answer any of these questions, the cosmids required mapping to a chromosome. By employing both physical and genetic methods, this work has contributed to the integration of the two types of maps.

4.3.1 Assignment and Orientation of Linkage Groups

I have assigned seven new markers to the East Lansing map which has contributed to the orientation of linkage groups E06C02W02 (Chromosome 2), E02C03W03 (Chromosome 3), E05C04W04 (Chromosome 4) and E07E34C05W05 (Chromosome 5).

Other problems, such as the orientation of chromosome 8 have been addressed. The positioning of the markers *ROS0149* (from this study) and *RPL5* on the physical and genetic maps enables its orientation by using these two markers as probes in two-colour FISH experiments. Future clones can be mapped with respect to these.

This work has also been used to update present maps. An example of this is the inversion of the linkage groups for chromosomes 3 and 4 on the genetic map compiled by Bumstead *et al.,* 1996. They have now been orientated correctly, with respect to the physical maps of these chromosomes.

4.3.2 Genetic Coverage

Genetic coverage was estimated for the three largest macrochromosomes and the Z chromosome. The degree of coverage of these chromosomes by the genetic linkage group is being confirmed by FISH mapping markers from linkage group ends (Vignal, Personal Communication). These chromosomes are either complete or close to completion, and so future mapping efforts should be focused on those chromosomes with relatively poor coverage.

4.4 Conclusion

A contribution has been made to the integration of the physical and genetic chicken maps by mapping random cosmids for this sequence sampling study. A means of establishing the orientation of chromosome 8 (see section 4.2.3 and Figure 4.7) has been established and a number of linkage groups have been placed in the correct orientations.

Chapter 5

Direct Assessment of Gene Density Differences Between Chicken Macrochromosomes and Microchromosomes

5.1 Introduction

One of the primary aims of this study was to investigate the hypothesis that chicken microchromosomes are more gene dense than macrochromosomes. Therefore a direct assessment of gene density, on both types of chromosome, was carried out by sequence sampling random chicken cosmids. Comparisons of the number of genes mapped to each set of chromosomes from previous genetic and physical mapping data were also made.

5.1.1 The Chicken Genome

The chicken karyotype comprises of thirty nine chromosome pairs, divided into six large macrochromosomes and thirty-three small microchromosomes. To facilitate comparisons with other studies, macrochromosomes were defined as chromosomes 1-5 and Z, whereas the microchromosomes were chromosomes 6-8, the remaining chromosomes and the W chromosome. Microchromosomes make up around 30% of the chicken genome (Smith and Burt, 1998) and appear to have a higher CpG content than macrochromosomes (McQueen *et al.*, 1996). Given that CpG islands are associated with genes, these results suggest a higher gene density on the microchromosomes.

However, CpG islands may not always be associated with genes; chicken chromosomal regions associated with the nuclear periphery often contain CpG-rich repetitive DNAs. The chicken nuclear membrane repeat, for example, is comprised of different multiples of a 41-42 bp tandemly repeated sequence, organised into large blocks (Matzke *et al.*, 1990). These sequences are located almost solely on the microchromosomes and are organised into large tandem arrays, making up 10% of the chicken genome. This repeat has not been found in either turkey or quail DNA, suggesting that the amplification of this repeat occurred recently in avian evolution (Matzke *et al.*, 1990).

5.1.2 Gene Density Studies

Chapter 1 described how microchromosomes appear to have a higher gene density than macrochromosomes. The prediction was made that a gene could be expected at the rate of 1 gene per 10 kb of microchromosomal DNA. If true, then 75% of chicken genes would be found on the microchromosomes and would have important practical implications e.g. gene discovery. From EST work (Burt and Bumstead-unpublished), 40% of chicken cDNAs were found to have a database hit. A gene homology can therefore be expected every 10-20 kb of DNA. In the cosmid clones, which are analysed in this chapter, between 2-4 gene homologies can therefore be expected per cosmid.

5.2 Results

5.2.1 Assessment of Gene Density Differences Between Chicken Macrochromosomes and Microchromosomes

The hypothesis that chicken microchromosomes are more gene dense than macrochromosomes was investigated by carrying out a direct assessment of gene density by sequence sampling randomly selected chicken cosmids mapping to both microchromosomes and macrochromosomes. Random cosmids were used to avoid the bias towards genes being present in the cosmid.

Putative genes were identified by comparison with known expressed sequences in public databases and features such as CpG content (discussed in this chapter) and repetitive elements were analysed (Chapter 6). The distribution of repeats in the genome was used to test the hypothesis that if microchromosomes are gene dense, they should have fewer repeats, as they should contain less 'junk' DNA.

The gene density differences between macrochromosomes and microchromosomes was small. The sample size used to determine differences in gene density by sequence sampling was too small to be statistically significant. Therefore, a comparison was made of the number of genes mapped to either a macrochromosome or to a microchromosome from previous genetic and physical mapping data. It was then possible to estimate the difference in gene density more precisely. These data were then compared with the gene density estimate from sequence sampling.

5.2.2 Calculation of the Number of Cosmid Sub-Clones for Sequencing

For sequence sampling, the cosmid clones were partially sequenced, and random subclones of the cosmids were generated with the restriction enzyme *Cvi*JI (Cambio UK). At least 40% of each cosmid clone was sequenced. To calculate the number of clones required for this, if the sequencing was completely random, the equation $p=1-[1-x/y]^n$ was used where 'x' is the average sequence length, 'y' is the cosmid insert size (35 kb) and 'n' is the number of sequences. Table 5.1 describes this information.

40% coverage of each cosmid was used in the *Fugu* Landmark Mapping project, where a sequence scanning approach was employed, which is very similar to the sequence sampling method used here (Nurminsky and Hartl, 1996). This level of coverage was found to be sufficient when searching for genes. Over 1000 *Fugu* cosmids have been scanned and now specific regions of the genome can be compared with mammals (Elgar, 1996);(Elgar *et al.*, 1998).

Both contig assembly and estimated sequence lengths are similar, confirming the randomness of the sequence sampling approach. Randomness was essential for good coverage of the cosmid.

5.2.3 Gene Homologies from Database Searches

In total seven gene homologies, as outlined in Table 5.2, were discovered: three mapped to macrochromosomal cosmids and four to microchromosomal cosmids.

5.2.4 Gene Homologies on the Macrochromosomes

Three gene homologies were discovered from the eight macrochromosomal cosmids: genes homologous to the human 5-Hydroxytryptamine 1D receptor (5-HT_{1D}) (Hamblin and Metcalf, 1991) and a platelet-activating receptor protein (PTAFR) (Jacobs *et al.*, 1997) were found on cosmid 27. The chicken transforming growth factor beta receptor gene (TGFβR1) was present on cosmid 28.

Table 5.1 Estimated and Actual Sequence Lengths From Sequence Sampled Random Cosmids

The equation $p=1-[1-x/y]^n$ was used to calculate the estimated sequences lengths with 'p' is the probability of cloning the entire cosmid, 'x' as the length of the average sequence read, 'y' the cosmid insert size (35,000 bp) and 'n' the number of sequences; Mic- microchromosome

Cosmid	Physical Assignment	% Coverage of Cosmid	Assembled unique sequence (bp)	No. of sequences	Average sequence read (bp)	Total sequence read (bp)
28	2p12-p11	59	13537	106	322	34132
35	2q26-q32	54	12199	90	333	29970
34	2q32-q35	54	15069	77	389	29953
08	3q11	47	16423	52	463	24076
30	3q23-q33	44	14785	61	369	22509
16	4p13-p12	64	26788	72	539	38808
33	4p14-p13	45	14226	66	344	22704
27	5q21-q22	51	18262	80	345	27600
		Average: 52.25	**Total: 131289**	**Total: 604**	**Average: 380**	**Total: 229752**

continued

Table 5.1 continued

Cosmid	Physical Assignment	% Coverage of Cosmid	Assembled unique sequence (bp)	No. of Sequences	Average sequence read (bp)	Total Sequence Read (bp)
01	Mic	66	21480	80	320	41600
07	Mic	37	13570	49	362	17738
14	Mic	44	12248	46	490	22540
20	Mic	72	24010	114	427	48678
21	Mic	62	19354	72	521	37512
31	Mic	46	11710	68	352	23936
32	Mic	48	15577	88	287	25256
36	Mic	43	13266	64	339	21696
		Average: 52.25	**Total: 131215**	**Total: 581**	**Average: 411**	**Total: 238956**

Table 5.2 Gene Homologies Found by Sequence Sampling

Mic-microchromosome

Cosmid	Chicken Cosmid GenBank Accession Numbers	Physical Assignment	Gene Homologies
28	AJ231919-231950	2p12-p11	(Chicken) Transforming Growth Factor Beta Receptor (TGFBR1)
35	AJ232107-232137	2q26-q32	None
34	AJ232079-232106	2q32-q35	None
08	AJ231737-231764	3q11	None
30	AJ231951-231979	3q23-q33	None
16	AJ231786-231815	4p13-p12	None
33	AJ232048-232078	4p14-p13	None
27	AJ231872-231918	5q21-q22	(Human) 5-Hydroxytryptamine 1D receptor (5-HT$_{1D}$); (Human) Platelet Activating Factor Receptor (PTAFR)

continued

Table 5.2 continued

Cosmid	Chicken Cosmid GenBank Accession Numbers	Physical Assignment	Gene Homologies
07	AJ231709-231736	Mic	(human) Hydroxybutyrate Dehydrogenase
14	AJ231765-231785	Mic	(human) M130 Antigen
20	AJ231816-231844	Mic	Chicken CEPU-1 Gene
21	AJ231845-231871	Mic	None
31	AJ231980-232004	Mic	None
32	AJ232005-232047	Mic	(human) KIAA0677 Protein
36	AJ232138-232169	Mic	None

5.2.5 5-Hydroxytryptamine 1D Receptor

The neurotransmitter and local hormone 5-Hydroxytryptamine (5-HT), or serotonin, is widely distributed in animals. It plays a role in many physiological and pathophysiological pathways centred around the central nervous system (CNS) and intestine. 5-HT stimulates the contraction of smooth muscles such as blood vessels, the intestine (to mediate peristalsis) and the uterus, and also stimulates sensory nerve endings. It has a potential role in platelet aggregation and microvascular control. In the CNS, 5-HT is thought to be involved in functions such as the control of mood, anxiety, hallucinations, sleep, vomiting, thermoregulation and pain perception (Saudou and Hen, 1994) and (Watson and Arkinstall, 1994).

There is a wide range of 5-HT receptors which can be divided into seven subfamilies, 5-HT$_1$-5-HT$_7$, based on amino acid sequence homology and coupling to second messengers. Each subfamily is broken down further into receptor subtypes which have individual brain distributions (Saudou and Hen, 1994);(Clement *et al.*, 1996). The 5-HT$_1$ subfamily, to which the 5-HT$_{1D}$ (formally known as 5-HT$_{1D\alpha}$) receptor belongs, are G-protein coupled receptors and comprise of five subtypes which all share a degree of sequence homology (~50%) and have overlapping pharmacological specifities. The receptors are located within the lipid bilayer; they comprise of seven transmembrane proteins and are linked to the inhibition of adenylate cyclase activity (Shih *et al.*, 1991);(Watson and Arkinstall, 1994). The receptor is located in neurons in the CNS and in vascular smooth muscles such as the coronary artery, and may have a role, alongside receptors 5-HT$_{1A}$ and 5-HT$_{1B}$, in the control of the release of 5-HT as well as controlling neurotransmitter release. It is also thought to have a role in feeding behaviour, anxiety, depression, cardiac function and movement (Weinshank *et al.*, 1992), (Watson and Arkinstall, 1994) and (Wurch *et al.*, 1997). The 5-HT$_{1D}$ receptor has been identified in dog, guinea pig and human (Saudou and Hen, 1994) and (Wurch *et al.*, 1997). The receptor 5-HT$_{1B}$, (formally known as 5- HT$_{1D\beta}$) is the rodent homolog of the 5-HT$_{1D}$ receptor and is found in opossum, rat, hamster and mouse. It has a different pharmacological profile due to a single amino acid change. Stimulation of the 5-HT$_{1B}$ receptor is thought to lead to an increase of anxiety and locomotion but a decrease in food intake and aggressive behaviour (Clement *et al.*, 1996).

The 5-HT$_{1D}$ gene, which had not previously been isolated in chicken, was found on cosmid 27, which maps to macrochromosome 5 at position q21-q22. Figures 5.1 and 5.2 present the BLASTN and BLASTX results from the database search. Significant hits with chicken sequence were observed both in the coding sequence of the gene and the 5' glycosylation site. Significant hits to a number of 5-HT$_{1D}$ receptors from a wide range of species were also observed, the highest scores and probabilities for the human 5-HT$_{1D}$ receptor. Other species with high scores and probabilities include pig, *Fugu*, plus the mouse and rat 5-HT$_{1B}$ receptor. The chicken 5-HT$_{1D}$ receptor also shares homology with other members of the 5-HT$_1$ subfamily such as the mouse 5-HT$_{1E}$ and human 5-HT$_{1a}$ receptors. The chicken 5-HT$_{1D}$ receptor appears to be orthologous to the human 5-HT$_{1D}$ in that they have diverged from each other after speciation events. The chicken 5-HT$_{1D}$ is paralogous to mouse 5-HT$_{1B}$ receptors as it has diverged after a gene duplication event (Eisen, 1998).

5.2.6 Platelet-Activating Receptor Protein

The phospholipid platelet-activating factor (PAF) acts as an intercellular messenger. It is involved in a range of activities including platelet activation, allergic response, asthma, septic shock, arterial thrombosis and other inflammatory process. It also has a role in the activation of polymorphonuclear leukocytes, monocytes and macrophages, decreasing cardiac output, the stimulation of uterine contraction and glycogenolysis in the liver (Seyfried *et al.*, 1992) and (Prescott *et al.*, 1990). The effects of PAF are mediated by specific cell surface receptors such as the platelet-

```
                                                                  Smallest
                                                                    Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

gb|M81589|HUMSER1DRA Homo sapiens serotonin 1D receptor (..972  3.1e-73   1
gb|M89955|HUM5HT1DA  Human 5-HT1D-type serotonin receptor..972  3.6e-73   1
emb|Y11868|SS5HTSR1D S.scrofa mRNA for serotonin 1D recep..900  2.4e-67   1
gb|L20335|MUSGPCR14  Mouse serotonin-1D receptor homologu..873  8.7e-66   1
gb|U60825|OCU60825   Oryctolagus cuniculus 5-HT1D alpha r..882  9.9e-66   1
emb|Z50162|OC5HT1DAR Oryctolagus cuniculus gene for 5HT1D..873  6.5e-65   1
emb|X94908|MM5HT1D   M.musculus 5-HT1D gene for seroprote..855  1.8e-63   1
gb|S74770|S74770     serotonin 5HT1D alpha receptor homol..721  1.3e-53   1
emb|X83865|FR5HT1D   F.rubripes 5HT1D gene                 714  1.9e-50   1
gb|U82175|CPU82175   Cavia porcellus 5-HT1B receptor gene..438  1.9e-26   1
gb|S45398|S45398     serotonin 1D receptor {clone S8-beta..379  3.7e-22   1
gb|L04962|HUMSRCPT1F Homo sapiens serotonin receptor (HTR..230  4.4e-16   2
emb|Z14224|MMSR5HT1E M.musculus mRNA for 5HT1E beta serot..212  6.5e-16   2
gb|L05597|HUMSEROTON Human serotonin receptor gene, compl..230  6.5e-16   2
gb|U80852|CPU80852   Cavia cobaya 5-hydroxytryptamine 1F ..230  7.1e-09   1

gb|M81589|HUMSER1DRA Homo sapiens serotonin 1D receptor (5-HT1D`) mRNA,
complete cds. Length = 1200

Score = 972 (268.6 bits), Expect = 3.1e-73, P = 3.1e-73
Identities = 286/403 (70%), Positives = 286/403 (70%), Strand = Plus / Plus

Query:   1 CTATCACAGATGCTTTGGAATATGCCAAACGCCGGACTGCTGGCCGAGCAATGCTCATGA 60
           | |||||||||||  |||||||    |||||| |||| |||||||  ||      |||||
Sbjct: 459 CAATCACAGATGCCCTGGAATACAGTAAACGCAGGACGGCTGGCCACGCGGCCACCATGA 518

Query:  61 TCGCTGTGGTTTGGATGATCTCCATTAGTATTTCTGTGCCACCATTTTTCTGGAGGCAAG 120
           ||||  | || |||   ||||||||  | || ||  | || ||  | |||||| |||| |
Sbjct: 519 TCGCCATTGTCTGGGCCATCTCCATCTGCATCTCCATCCCCCCGCTCTTCTGGCGGCAGG 578

Query: 121 TGAAAGCTCATGAAGAAATCNCGAANTGTAATGTGAACACAGATCAGATTTCCTACACAA 180
             || || || || || ||  || | |||   ||||||||   |||||| |||||||| |
Sbjct: 579 CCAAGGCCCAGGAGGAGATGTCGGACTGTCTGGTGAACACCTCTCAGATCTCCTACACCA 638

Query: 181 TTTATTCCACCTGTGGAGCTTTCTACATTCCAACTGTGCTCCTCCTAATATTATACNGTA 240
           | || ||||||||||| || |||||||||||  | ||| | ||| | ||  ||||  |
Sbjct: 639 TCTACTCCACCTGTGGGGCCTTCTACATTCCCTCGGTGTTGCTCATCATCCTATATGGCC 698

Query: 241 GGATTTATGTANCANCTCGATCCAAGATCCTGAAGCCACCCTCACTGTATGGGAAACGAT 300
           |||| ||     |  | ||   |   |||||||| ||||||||||| |||||||| || |
Sbjct: 699 GGATCTACCGGGCTGCCCGGAACCGCATCCTGAATCCACCCTCACTCTATGGGAAGCGCT 758

Query: 301 TCACTACTGCACACCTGATAACTGGCTCTGCTGGGTCTTCCCTCTGCTCCATTAACGCAA 360
           |||| || || ||||| || || |||||||| ||||| || ||||||||  | ||| | |
Sbjct: 759 TCACCACGGCCCACCTCATCACAGGCTCTGCCGGGTCCTCGCTCTGCTCGCTCAACTCCA 818

Query: 361 GCCTTCATGAAGGGCATTCCCATTCCGGTGGATCCCCGATATT 403
           |||| ||||| ||||| || || || | ||| |||||  | ||
Sbjct: 819 GCCTCCATGAGGGGCACTCGCACTCGGCTGGCTCCCCTCTCTT 861
```

Figure 5.1 Cosmid 27 BLASTN Results: 5-Hydroxytryptamine 1D Receptor

```
                                                                Smallest
                                                                  Sum
                                                Reading  High  Probability
Sequences producing High-scoring Segment Pairs:   Frame Score  P(N)      N

TREMBL:P79400 P79400 SEROTONIN RECEPTOR 1D (FRAGMENT)..+3   560  1.0e-72  1
SWISSPROT:5H1D_HUMAN P28221 homo sapiens (human). 5-h..+3   557  2.1e-72  1
SWISSPROT:5H1D_RABIT P49145 oryctolagus cuniculus (ra..+3   554  5.4e-72  1
TREMBL:O02823 O02823 5-HT1D ALPHA RECEPTOR. 7/97       +3   554  5.4e-72  1
SWISSNEW:5H1D_MOUSE ID 5H1D_MOUSE STANDARD; PRT; 374 ..+3   542  2.5e-70  1
SWISSPROT:5H1D_RAT P28565 rattus norvegicus (rat). 5-..+3   540  4.7e-70  1
TREMBL:Q61615 Q61615 SEROTONIN 1D RECEPTOR (FRAGMENT)..+3   542  5.4e-70  1
SWISSPROT:5H1D_CANFA P11614 canis familiaris (dog). 5..+3   523  1.1e-67  1
TREMBL:O08891 O08891 5-HT1D RECEPTOR. 7/97             +3   523  1.1e-67  1
SWISSNEW:5H1D_CAVPO ID 5H1D_CAVPO STANDARD; PRT; 376 ..+3   520  2.7e-67  1
SWISSNEW:5H1D_FUGRU ID 5H1D_FUGRU STANDARD; PRT; 379 ..+3   460  2.5e-60  2
TREMBL:Q64054 Q64054 SEROTONIN 5HT1D ALPHA RECEPTOR H..+3   449  5.9e-57  1
SWISSNEW:5H1B_RAT ID 5H1B_RAT STANDARD; PRT; 386 AA.   +3   320  2.2e-45  3
SWISSPROT:5H1B_RAT P28564 rattus norvegicus (rat). 5-..+3   320  2.2e-45  3
SWISSPROT:5H1B_DIDMA P35404 didelphis marsupialis vir..+3   173  1.5e-44  4
SWISSNEW:5H1B_CRIGR ID 5H1B_CRIGR STANDARD; PRT; 386 ..+3   329  1.3e-43  2
SWISSPROT:5H1B_CRIGR P46636 cricetulus griseus (chine..+3   329  1.3e-43  2
SWISSNEW:5H1B_HUMAN ID 5H1B_HUMAN STANDARD; PRT; 390 ..+3   326  3.7e-43  2
SWISSPROT:5H1B_HUMAN P28222 homo sapiens (human). 5-h..+3   326  3.7e-43  2
SWISSNEW:5H1B_MOUSE ID 5H1B_MOUSE STANDARD; PRT; 386 ..+3   321  1.8e-42  2

TREMBL:P79400 P79400 SEROTONIN RECEPTOR 1D (FRAGMENT). 5/97
Length = 291

Plus Strand HSPs:

Score = 560 (257.6 bits), Expect = 1.0e-72, P = 1.0e-72
Identities = 104/133 (78%), Positives = 115/133 (86%), Frame = +3

Query:   3 ITDALEYAKRRTAGRAMLMIAVVWMISISISVPPFFWRQVKAHEEIXXCNVNTDQISYTI 182
           ITDALEY+KRRTAG A  MIA+VW ISI IS+PP FWRQ +AHEEI  C VNT QISYTI
Sbjct:  60 ITDALEYSKRRTAGHAAAMIAIVWAISICISIPPLFWRQARAHEEISDCLVNTSQISYTI 119

Query: 183 YSTCGAFYIPTVLLLILYXRIYVXXRSKILKPPSLYGKRFTTAHLITGSAGSSLCSINAS 362
           YSTCGAFYIP++LL+ILY RIY   R++IL PPSLYGKRFTTAHLITGSAGSSLCS+N S
Sbjct: 120 YSTCGAFYIPSLLLIILYGRIYRAARNRILNPPSLYGKRFTTAHLITGSAGSSLCSLNPS 179

Query: 363 LHEGHSHSGGSPI 401
           LHEGHSHS GSP+
Sbjct: 180 LHEGHSHSAGSPL 192
```

Figure 5.2 Cosmid 27 BLASTX Results: 5-Hydroxytryptamine 1D Receptor

activating receptor (PTAFR), which is a part of the G- protein coupled receptor family; the binding of PAF stimulates GTPase activity. The receptors are found in cells and tissues where PAF has effects (Seyfried *et al.*, 1992) and PTAFR is involved in the pathogenesis of many diseases and also has a role in normal physiological processes such as homeostasis and reproduction. It can activate tyrosine kinase pathways which leads to the phosphorylation of Src proteins (Chase *et al.*, 1996).

DNA sequence from cosmid 27 produced a significant database hit against the *PTAFR* gene. The match was observed in the coding region of the gene and this appears to be the first time this gene has been found in a non-mammalian species. The chicken *PTAFR* gene has not previously been isolated (Figures 5.3 shows the BLASTN results) and was also found on cosmid 27, which maps to chicken macrochromosome 5q21-q22. A BLASTX search proved unsuccessful for this gene.

5.2.7 Conservation of Synteny

The *PTAFR* and *5-HT*$_{1D}$ genes are syntenic in chicken as they are located on the same chromosome. The sequencing data from this cosmid is incomplete, and so one cannot say how close these two genes are to each other. However, as the cosmid insert size is 38 kb, the genes are at the very most this distance apart. The genes are also syntenic in human and mouse, forming a small conserved segment, but appear to be further apart. In humans, both the *PTAFR* and *5-HT*$_{1D}$ genes map to 1p35-p34.3 and are estimated to be 100 kb apart. In mouse, *Ptafr* maps to chromosome 4 at position 62.4 and *5-HT*$_{1B}$ to chromosome 4 at position 66.0. The *PTAFR* gene has not been found in other species and without this information it cannot be determined whether this conserved segment is present in other species.

```
                                                              Smallest
                                                                Sum
                                                       High  Probability
                                                       Score  P(N)      N

NEWEMBL:AF002986 ! Af002986 Homo sapiens platelet activat..490  7.2e-33   1

NEWEMBL:AF002986 Af002986 Homo sapiens platelet activating receptor homolog
(H963) mRNA, complete cds. 11/97
Length = 1272

Plus Strand HSPs:

Score = 490 (135.4 bits), Expect = 7.2e-33, P = 7.2e-33
Identities = 158/233 (67%), Positives = 158/233 (67%), Strand = Plus / Plus

Query:    3 ATTCCAATAAAAACCATTGAAGAAAGACCTAACGCAAGGTGCATCGATTTCAAAACAAAA 62
            ||||| || |||  |||  | ||||   | || |   | || || || || ||||   ||
Sbjct:  673 ATTCCCATCAAAGACATCAAGGAAAAGTCAAATGTGGGTTGTATGGAGTTTAAAAAGGAA 732

Query:   63 TTTGGGAGAGACTGGCACGTGTTCACTAACTTTGTGTGCACAGCAATATTCCTGAATTTT 122
            ||||| ||| | |||||  || | || || ||  | ||   |||||||||  | |||||
Sbjct:  733 TTTGGAAGAAATTGGCATTTGCTGACAAATTTCATATGTGTAGCAATATTTTTAAATTTC 792

Query:  123 TCAGCTGTGATACTCATTTCCAATTTCCTTGTTGTCAGACAGCTCTACCAGAACAAATAC 182
            |||||  | ||  | || ||||||| ||||||  |  |||||||||||   |||||| |
Sbjct:  793 TCAGCCATCATTTTAATATCCAATTGCCTTGTAATTCGACAGCTCTACAGAAACAAAGAT 852

Query:  183 AGCGAGAGTTACACAAATGTGAAGAAAGCCCTGGTGAGCATACTGCTGCTGAC 235
            |  || | |||| |||||||||| || || ||  | | ||||||  |  ||||
Sbjct:  853 AATGAAAATTACCCAAATGTGAAAAAGGCTCTCATCAACATACTTTTAGTGAC 905
```

Figure 5.3 Cosmid 27 BLASTN Results: Platelet Activating Receptor Homolog

5.2.8 Transforming Growth Factor Receptor Gene

The TGF-β family has effects on cell proliferation, differentiation and organisation. The transforming growth factor TGFβR1, or receptor protein kinase 2 (RPK-2), is a member of this family as it has a kinase domain which is related to the type II receptor for TGF-β. The TGFβR1 sequence has the most similarity with chick receptor protein kinase 1 (RPK-1), which is another member of this family. Potential ligands for TGFβR1 include growth factors such as the bone morphogenetic proteins, Vg1-related protein and inhibins (Nohno *et al.*, 1993).

Three regions of the TGFβR1 receptor were isolated, two in the exon and one in the 3' end, prior to the poly A site. Figure 5.4a presents the BLASTN and figure 5.4b the BLASTX results. Although the TGFβR1 cDNA was isolated from a chick embryonic DNA library prior to this study (Nohno *et al.*, 1993), it had not been mapped. Genetic and physical mapping data indicate that the gene maps to the short arm of macrochromosome 2.

5.2.9 Gene Homologies on the Microchromosomes

Four gene homologies were found across the eight microchromosomal cosmids. The chicken CEPU-1 gene on cosmid 20; the gene for a chicken scavenger receptor-like protein homologous to a human M130 antigen protein on cosmid 14; a homolog of the human (R)-3-hydroxybutyrate dehydrogenase (BDH) gene (Marks *et al.*, 1992) on cosmid 7; and cosmid 32 produced a significant hit to human sequence for KIAA0667, which appears to be a zinc finger protein (Ishikawa *et al.*, 1998).

```
                                                                     Smallest
                                                                       Sum
                                                              High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

dbj|D14460|CHKRPK2     Chicken RPK-2 mRNA for receptor pro..524  7.2e-36   1
gb|U37065|MSU37065     Mustela sp. TGF-b type I receptor m..380  4.4e-22   1
gb|L11695|HUMALK5A     Human activin receptor-like kinase ..380  4.7e-22   1
gb|U21860|XLU21860     Xenopus laevis type I serine/threon..375  1.2e-21   1
dbj|D28526|MUSTGFBIR   Mouse mRNA for TGF-beta type I rece..357  4.4e-20   1

dbj|D14460|CHKRPK2 Chicken RPK-2 mRNA for receptor protein kinase, complete
cds.
Length = 2186 Plus Strand HSPs:

Score = 524 (144.8 bits), Expect = 7.2e-36, P = 7.2e-36
Identities = 106/109 (97%), Positives = 106/109 (97%), Strand = Plus / Plus

Query: 165 GGCAAACCAGCAATTGCCCACAGAGATTTGAAATCAAAAAACATATTGGTAAAGAAGAA 222
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:1074 GGCAAACCAGCAATTGCCCACAGAGATTTGAAATCAAAAAACATATTGGTAAAGAAGAA 1131

Query: 223 TGGAACATGCTGCATTGCAGACCTGGGGTTGGCAGTTAGGCATGATTCA 271
            |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:1132 TGGAACATGCTGCATTGCAGACCTGGGGTTGGCAGTTAGGCATGATTCA 1180
```

Figure 5.4a Cosmid 28 BLASTN Results

```
                                                                     Smallest
                                                                       Sum
                                                    Reading  High  Probability
Sequences producing High-scoring Segment Pairs:       Frame  Score  P(N)      N

pir||A56693          receptor protein kinase RPK-2 -..+3   180   9.1e-18    1
gi|841310            (U21860) type I serine/threonin..+3   179   1.3e-17    1
gi|1045610           (U37065) TGF-b type I receptor ..+3   177   2.4e-17    1
pir||JC2062          transforming growth factor-beta..+3   177   2.5e-17    1

pir||A56693 receptor protein kinase RPK-2 - chicken >gi|285700 (D14460)
receptor protein kinase [Gallus gallus]
Length = 440 Plus Strand HSPs:

Score = 180 (82.8 bits), Expect = 9.1e-18, P = 9.1e-18
Identities = 34/35 (97%), Positives = 34/35 (97%), Frame = +3

Query:    213 IGGIHEDYQLPYYDLVPSDPSVEEMKKVVCEQKLR 317
              IGGIHEDYQLPYYDLVPSDPSVEEMKKVVCEQKLR
Sbjct:    354 IGGIHEDYQLPYYDLVPSDPSVEEMKKVVCEQKLR 388
```

Figure 5.4b Cosmid 28 BLASTX Results

Figure 5.4 Cosmid 28 BLASTN and BLASTX Results: Transforming Growth Factor Receptor Gene

5.2.10 The Chicken CEPU-1/Opioid-Binding Cell Adhesion Molecule Gene

Members of the immunoglobulin superfamily mediate cell-cell interactions in the immune system and the developing nervous system. The gene *CEPU-1* is a member of this superfamily. In chicken it has been found to be expressed in the developing Purkinje neurons in the cerebellum. (Spaltmann and Brummendorf, 1996). The CEPU-1 protein shows high sequence similarity with other members of this superfamily, such as the opioid-binding cell adhesion molecule (OBCAM); this has a potential role in cell adhesion and as a neuropeptide receptor and, like CEPU-1, is expressed in the nervous system (in the striatum, cerebral cortex and cerebellum regions of the brain) (Wu *et al.*, 1990) and (Shark and Lee, 1995). Both CEPU-1 and OBCAM are glycosylphosphatidylinositol (GPI)-anchored proteins (Hachisuka *et al.*, 1996) and (Spaltmann and Brummendorf, 1996). The CEPU-1 gene had been isolated from chicken prior to this study and has now been mapped to a microchromosome. Figures 5.5 and 5.6 highlight the BLASTN and BLASTX results showing hits to the chicken *CEPU-1* and human *OBCAM* genes. A chicken *OBCAM* gene does not appear to be present in any database and there is no literature documenting it. The BLAST search results suggest that the chicken *CEPU-1* gene may be the chicken *OBCAM* gene and is therefore the ortholog of human *OBCAM*. If they were different genes this would be seen clearly in the BLASTN results, with separate database matches to *OBCAM* and *CEPU-1* genes. This is confirmed by the BLASTX results, which are identical for both *CEPU-1* and *OBCAM* genes.

5.2.11 Human M130/Bovine WC1 Antigens

Monoclonal antibodies were used to define the human monocyte/macrophage-associated antigen M130. It is a transmembrane glycoprotein located both on the cell

```
                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

emb|Z72497|GGCEPU1 G.gallus mRNA for CEPU-1                   431  1.2e-25   1
gb|L34774|HUMOBCAM Human (clone pHOM) opioid-binding cell..386  7.6e-22   1
gb|M88710|RATCALMB Rattus norvegicus cell adhesion-like m..359  1.5e-19   1
gb|U16845|RNU16845 Rattus norvegicus neurotrimin mRNA, co..350  8.3e-19   1
emb|X12672|BTOBCAM Bovine mRNA for opioid binding protein..350  8.5e-19   1

emb|Z72497|GGCEPU1 G.gallus mRNA for CEPU-1 Length = 1257

Minus Strand HSPs:

Score = 431 (119.1 bits), Expect = 1.2e-25, P = 1.2e-25
Identities = 87/89 (97%), Positives = 87/89 (97%), Strand = Minus / Plus

Query: 599 AGGAGTGCCCGTGCGCAGCGGAGATGCCACCTTCCCCAAAGCTATGGACAACGTGACTGT 540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 162 AGGAGTGCCCGTGCGCAGCGGAGATGCCACCTTCCCCAAAGCTATGGACAACGTGACTGT 221

Query: 539 GCGGCAAGGGGAGAGTGCCACGCTCAGGT 511
            |||||||||||||||||||||||||||||
Sbjct: 222 GCGGCAAGGGGAGAGTGCCACGCTCAGGT 250


gb|L34774|HUMOBCAM Human (clone pHOM) opioid-binding cell adhesion molecule
mRNA, complete cds. Length = 1478

Minus Strand HSPs:

Score = 386 (106.7 bits), Expect = 7.6e-22, P = 7.6e-22
Identities = 82/89 (92%), Positives = 82/89 (92%), Strand = Minus / Plus

Query: 599 AGGAGTGCCCGTGCGCAGCGGAGATGCCACCTTCCCCAAAGCTATGGACAACGTGACTGT 540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct: 131 AGGAGTGCCCGTGCGCAGCGGAGATGCCACCTTCCCCAAAGCTATGGACAACGTGACGGT 190

Query: 539 GCGGCAAGGGGAGAGTGCCACGCTCAGGT 511
             ||||| |||||||| ||||| |||||||
Sbjct: 191 CCGGCAGGGGGAGAGCGCCACCCTCAGGT 219
```

Figure 5.5 Cosmid 20 BLASTN Results: CEPU-1 and OBCAM Genes

```
                                                                      Smallest
                                                                         Sum
                                                     Reading High  Probability
Sequences producing High-scoring Segment Pairs:        Frame Score P(N)      N

sp|Q90773|CEPU_CHICK CEPU-1 PROTEIN PRECURSOR /gnl|P..+2 141  1.6e-11    1
pir||JC1238          opioid-binding protein (clone D..+2 141  1.6e-11    1
gnl|PID|d1032473     (AB011810) CEPU-1 [Gallus gallu  +2 141  1.6e-11    1
sp|Q62718|NTRI_RAT   NEUROTRIMIN PRECURSOR (GP65) /p..+2 141  1.6e-11    1
sp|P32736|OPCM_RAT   OPIOID BINDING PROTEIN/CELL ADH..+2 141  1.6e-11    1
sp|Q14982|OPCM_HUMAN OPIOID BINDING PROTEIN/CELL ADH..+2 141  1.6e-11    1
sp|P11834|OPCM_BOVIN OPIOID BINDING PROTEIN/CELL ADH..+2 141  1.6e-11    1


sp|Q90773|CEPU_CHICK CEPU-1 PROTEIN PRECURSOR gnl|PID|e244389 (Z72497)
CEPU-1 [Gallus gallus] prf||2207311A CEPU-1 [Gallus gallus]
Length = 353

Plus Strand HSPs:

Score = 141 (64.6 bits), Expect = 1.6e-11, P = 1.6e-11
Identities = 28/29 (96%), Positives = 28/29 (96%), Frame = +2

Query:     2 GVPVRSGDATFPKAMDNVTVRQGESATLR 88
             GVPVRSGDATFPKAMDNVTVRQGESATLR
Sbjct:   26 GVPVRSGDATFPKAMDNVTVRQGESATLR 54


pir||JC1238 opioid-binding protein (clone DUZ1) - rat gi|203246
(M88709) cell adhesion-like molecule [Rattus norvegicus]
Length = 338

Plus Strand HSPs:

Score = 141 (64.6 bits), Expect = 1.6e-11, P = 1.6e-11
Identities = 28/29 (96%), Positives = 28/29 (96%), Frame = +2

Query:     2 GVPVRSGDATFPKAMDNVTVRQGESATLR 88
             GVPVRSGDATFPKAMDNVTVRQGESATLR
Sbjct:   21 GVPVRSGDATFPKAMDNVTVRQGESATLR 49
```

Figure 5.6 Cosmid 20 BLASTX Results: Chicken CEPU-1 and OBCAM Genes

surface and within the cell and is found on all circulating monocytes and most tissue macrophages. The cDNA contains an extracellular domain with nine repeating elements similar to scavenger receptor domains. This type of domain is also found on other antigens such as CD5 and CD6, the WC1 cattle antigen, in the long form of the scavenger receptor and in complement factor I (Law *et al.*, 1993).

Scavenger receptors are made up of six domains: cytoplasmic, membrane-spanning, alpha-helical coiled-coil, collagen like (which has a role in ligand binding) and a type-specific C-terminal. Scavenger receptor proteins have been detected in macrophages of various tissues and organs and have a potential role in defence against pathogenic agents (Itakura *et al.*, 1993). The scavenger receptor domains on M130 and WC1 are regular in structure (Law *et al.*, 1993). In bovine species, the WC1 antigen is only expressed on the surface of $CD4^-$ $CD8^-$ $\gamma\delta$ T lymphocytes (Wijngaard *et al.*, 1992). The extracellular portion of the WC1 antigen is made up of 11 cysteine-rich scavenger receptor protein domains. *WC1* gene families have been found in sheep, goats, pigs and horses. The *WC1* gene family appears to be less complex in human and mouse genomes, with fewer genes (Wijngaard *et al.*, 1994).

Figures 5.7 and 5.8 show the BLASTN and BLASTX alignments respectively between chicken DNA and the M130 and WC1 antigen coding sequences. The alignments clearly show that a chicken member of the WC1 gene family has been found and it is orthologous to the human M130 gene.

5.2.12 Homologs of the Human Hydroxybutyrate Dehydrogenase Gene

BDH has been previously cloned and characterised from a human heart cDNA library (Marks *et al.*, 1992); it is a mitochondrial membrane enzyme requiring phosphatidylcholine for activity (Green *et al.*, 1996; Marks *et al.*, 1992). Its amino acid sequence shows homology with a superfamily of short-chain alcohol dehydrogenases (Churchill *et al.*, 1992). BDH is widely distributed in different

```
                                                                  Smallest
                                                                    Sum
                                                           High  Probability
Sequences producing High-scoring Segment Pairs:           Score  P(N)      N

EM_HU1:HSM130A Z22968 H.sapiens mRNA for M130 antigen. 1..468  2.3e-38   2
EM_HU1:HSM130AC1 Z22969 H.sapiens mRNA for M130 antigen ..468  2.3e-38   2
EM_HU1:HSM130AE Z22971 H.sapiens mRNA for M130 antigen e..468  2.4e-38   2
EM_HU1:HSM130AC2 Z22970 H.sapiens mRNA for M130 antigen ..468  3.2e-38   2
EM_RO:MM12434 U12434 Mus musculus mscd6 precursor (Cd6) ..446  6.1e-35   2
EM_OM:SSSRP2 X99333 S.scrofa mRNA for scavenger-receptor..435  1.2e-34   2
EM_RO:MM35370 U35370 Mus musculus T cell accessory signa..437  3.9e-34   2
EM_OM:BBWC11MR X63723 B.bovis WC1.1 mRNA. 5/94            429  1.7e-33   2
EM_RO:MM37544 U37544 Mus musculus T cell surface glycopr..428  1.9e-33   2
EM_RO:MM37543 U37543 Mus musculus T cell surface glycopr..428  2.4e-33   2
EM_OM:SSSRP3 X99334 S.scrofa mRNA for scavenger-receptor..415  7.8e-33   2
EM_OM:S76311 S76311 T19=180-200 kda membrane protein sca..420  1.8e-31   2
EM_HU1:HS346251 U34625 Human T cell surface glycoprotein..379  2.6e-29   2
EM_HU1:HS346241 U34624 Human T cell surface glycoprotein..379  2.7e-29   2
EM_HU1:HS346231 U34623 Human T cell surface glycoprotein..379  2.8e-29   2
EM_HU1:HSCD6 X60992 H.sapiens CD6 mRNA for T cell glycop..379  3.1e-29   2
EM_OV:PM20652 U20652 Petromyzon marinus scavenger recept..445  9.0e-27   1
EM_RO:MM37438 U37438 Mus musculus CRP-ductin-alpha mRNA,..424  5.5e-25   1
EM_RO:RN32681 U32681 Rattus norvegicus ebnerin mRNA, com..372  1.2e-20   1
EM_OM:SSSRP4 X99335 S.scrofa mRNA for scavenger-receptor..339  6.7e-18   1


EM_HU1:HSM130A Z22968 H.sapiens mRNA for M130 antigen. 11/96
Length = 3703

Plus Strand HSPs:

Score = 468 (129.3 bits), Expect = 2.3e-38, Sum P(2) = 2.3e-38
Identities = 152/226 (67%), Positives = 152/226 (67%), Strand = Plus / Plus

Query:  43 AGCTGGGCGCTGTGCAGGGAGAGTGGAGATCTACTACCAGGGCCAATGGGGCACGGTCTG 102
           ||  || |||||||| |||||||| ||||||||  |  |||||   ||||||||  ||||
Sbjct:2258 AGGAGGTCGCTGTGCTGGGAGAGTAGAGATCTATCATGAGGGCTCCTGGGGCACCATCTG 2317

Query: 103 CGACGACGCCTGGGACACGGCCGACGCTGATGTTGTTTGCCGCCAGCTGANCTGCGGGTG 162
            || |||  |||||||  |   || ||  | || |||||| | ||||||  ||| ||
Sbjct:2318 TGATGACAGCTGGGACCTGAGTGATGCCCACGTGGTTTGCAGACAGCTGGGCTGTGGAGA 2377

Query: 163 GGCTGTGGAGGCGGCCGGCTCCGCTCGGTTTGGCGAGGGCTCCGGGCANATCTGGCTGGA 222
           |||  |  | ||  | || || ||||  ||||| || ||  | ||||  |||||||||||
Sbjct:2378 GGCCATTAATGCCACTGGTTCTGCTCATTTTGGGGAAGGAACAGGGCCCATCTGGCTGGA 2437

Query: 223 TGGTGTGAACTGCTCTGGGACTGAAGCTGCTCTCTGGGACTGTCAT 268
           ||   |||| |||  ||| |  ||| |     | ||| | || |||
Sbjct:2438 TGAGATGAAATGCAATGGAAAAGAATCCCGCATTTGGCAGTGCCAT 2483
```

Figure 5.7 Cosmid 14 BLASTN Results: M130/WC1 Antigens

```
                                                                 Smallest
                                                                   Sum
                                                   Reading High  Probability
Sequences producing High-scoring Segment Pairs:     Frame Score  P(N)      N

pir||I38005         M130 antigen (cytosolic variant ..+2   219  5.7e-28   2
pir||I38004         M130 antigen (cytosolic variant ..+2   219  5.7e-28   2
pir||I38006         M130 antigen (extracellular vari..+2   219  5.7e-28   2
pir||S36077         M130 antigen - human /gi|312142 ..+2   219  5.7e-28   2
gi|4105084          (AF043112) hensin [Oryctolagus c..+2   218  5.1e-27   2
pir||A57190         ebnerin precursor - rat /gi|9753..+2   204  7.6e-27   3
gnl|PID|e328724     (AJ000342) DMBT1 protein, 5.8 kb..+2   204  8.5e-26   2
pir||S56744         mucin (clone pGM7-1) - bovine /b..+2   204  1.3e-25   3
gnl|PID|e254903     (X99334) scavenger-receptor prot..+2   189  4.1e-25   2
gnl|PID|e254902     (X99333) scavenger-receptor prot..+2   187  6.9e-25   3
sp|P30205|WC11_BOV  ANTIGEN WC1.1 /pir||A46496 antig..+2   195  7.6e-25   2
bbs|117475          WC1 antigen [cattle, CD4-CD8- ga..+2   195  7.6e-25   2
pir||S56745         mucin (clone pGM31-1) - bovine /..+2   196  2.7e-24   2
pir||JC4361         scavenger receptor Cys-rich epid..+2   188  6.9e-24   2
gnl|PID|e254811     (X99335) scavenger-receptor prot..+2   189  6.9e-24   2


pir||I38005 M130 antigen (cytosolic variant 2) - human gi|312146
(Z22970) M130 antigen cytoplasmic variant 2 [Homo sapiens]
Length = 1156

Plus Strand HSPs:

Score = 219 (100.3 bits), Expect = 5.7e-28, Sum P(2) = 5.7e-28
Identities = 38/74 (51%), Positives = 46/74 (62%), Frame = +2

Query: 47 GRCAGRVEIYYQGQWGXXXXXXXXXXXXXXXXRQLXCGWAVEAAGSARFGEGSGXIWLDG 226
           GRCAGRVEIY++G WG                RQL CG A+ A GSA FGEG+G IWLD
Sbjct:721 GRCAGRVEIYHEGSWGTICDDSWDLSDAHVVCRQLGCGEAINATGSAHFGEGTGPIWLDE 780

Query:227 VNCSGTEAALWDCH 268
           + C+G E+ +W CH
Sbjct:781 MKCNGKESRIWQCH 794

Score = 84 (38.5 bits), Expect = 5.7e-28, Sum P(2) = 5.7e-28
Identities = 13/17 (76%), Positives = 15/17 (88%), Frame = +1

Query:277 WGQHDCGHKEDAGVVCS 327
           WGQ +C HKEDAGV+CS
Sbjct:798 WGQQNCRHKEDAGVICS 814
```

Figure 5.8 Cosmid 14 BLASTX Results: M130/WC1 Antigens

tissues and has been found in mitochondria from bovine heart, rat brain and liver, and in smooth, fast and slow twitch and cardiac muscle (Marks *et al.*, 1992).

The BLASTN and BLASTX results of chicken DNA aligned with BDH coding DNA are in Figures 5.9a and 5.9b. The *BDH* gene has not been previously detected in chicken and has now been mapped to a microchromosome. It is orthologous to the human and rat *BDH* genes.

5.2.13 Human mRNA for KIAA0677

KIAA0677 was isolated as part of 100 new cDNA clones from human brain cDNA libraries (Ishikawa *et al.*, 1998) and maps to human chromosome 1. KIAA0677 expression was detected in human heart, brain, lung, liver, skeletal muscle, kidney, pancreas, spleen, testis and ovary. This EST contains a potential zinc finger and shares homology with a putative 90.2 kDa zinc finger protein (Ishikawa *et al.*, 1998). Sequence from cosmid 32 produced a significant hit to the human sequence for the protein KIAA0667. The match is part of the coding sequence, and Figure 5.10 a and b show the BLAST results.

5.2.14 Calculation of Gene Density on the Macrochromosomes and Microchromosomes

5.2.15 Relative Gene Density

In this study, three genes per 131 kb of macrochromosomal DNA and four genes per 131 kb of microchromosomal DNA were found. The relative gene density was estimated by comparing the total amount of DNA sequenced for each chromosome type with the number of genes isolated on the macrochromosomal and microchromosomal cosmids. From these estimates it appears that the microchromosomes are 1.3 times as gene dense as the macrochromosomes. As the

```
                                                                  Smallest
                                                                     Sum
                                                         High  Probability
Sequences producing High-scoring Segment Pairs:          Score  P(N)      N

gb|M93107|HUM3HBDH   Homo sapiens heart (R)-3-hydroxybuty..385  7.7e-22   1
gb|M89902|RATDBHYDEH Sprague-Dawley D-beta-hydroxybutyrat..315  6.2e-16   1

gb|M93107|HUM3HBDH Homo sapiens heart (R)-3-hydroxybutyrate dehydrogenase
mRNA, 3' end. Length = 1357 Minus Strand HSPs:

Score = 385 (106.4 bits), Expect = 7.7e-22, P = 7.7e-22
Identities = 91/109 (83%), Positives = 91/109 (83%), Strand = Minus / Plus

Query: 196 CTCACNTCCACGGCCCCCTACACTCGCTACCACCCCATGGATTACTACTGGTGGCTGCGC 137
           || ||  ||||  |||||||||| ||||||||||||||||| |||||||||||||||||
Sbjct: 916 CTGACCGCCACCACCCCCTACACCCGCTACCACCCCATGGACTACTACTGGTGGCTGCGA 975

Query: 136 ATGCAGATCATGACGCACATGCCTGCAGCCATTTCAGACCGGCTCAAAA 88
           |||||||||||||| ||| |||||| |||||| || |||  | || | |
Sbjct: 976 ATGCAGATCATGACCCACTTGCCTGGAGCCATCTCCGACATGATCTACA 1024
```

Figure 5.9a Cosmid 7 BLASTN Results

```
                                                                  Smallest
                                                                     Sum
                                                Reading  High  Probability
Sequences producing High-scoring Segment Pairs:   Frame  Score  P(N)      N

sp|Q02338|BDH_HUMAN  D-BETA-HYDROXYBUTYRATE DEHYDROG..-3   185  6.1e-17   1
sp|P29147|BDH_RAT    D-BETA-HYDROXYBUTYRATE DEHYDROG..-3   168  1.3e-14   1
pir||B42845          3-hydroxybutyrate dehydrogenase..-3    95  7.1e-06   2

sp|Q02338|BDH_HUMAN D-BETA-HYDROXYBUTYRATE DEHYDROGENASE PRECURSOR
(BDH) (3-HYDROXYBUTYRATE DEHYDROGENASE) pir||A42845
3-hydroxybutyrate dehydrogenase (EC 1.1.1.30) - human (fragment)
gi|177198 (M93107) (R)-3-hydroxybutyrate dehydrogenase [Homo sapiens]
Length = 343 Minus Strand HSPs:

Score = 185 (84.7 bits), Expect = 6.1e-17, P = 6.1e-17
Identities = 32/54 (59%), Positives = 38/54 (70%), Frame = -3

Query:    234 ETFQTEWMLKSKPRVCVPVRVLTSTAPYTRYHPMDYYWWLRMQIMTHMPAAISD 73
              ET+ +     + P +      LT+T PYTRYHPMDYYWWLRMQIMTH+P AISD
Sbjct:    285 ETYCSSGSTDTSPVIDAVTHALTATTPYTRYHPMDYYWWLRMQIMTHLPGAISD 338
```

Figure 5.9b Cosmid 7 BLASTX Results

Figure 5.9 Cosmid 7 BLASTN and BLASTX Results: Hydroxybutyrate Dehydrogenase Gene

```
                                                                    Smallest
                                                                      Sum
                                                              High  Probability
                                                             Score  P(N)       N

dbj|AB014577|AB014577    Homo sapiens mRNA for KIAA0677 p..415  3.1e-24    1

dbj|AB014577|AB014577 Homo sapiens mRNA for KIAA0677 protein, complete cds
Length = 4417

Minus Strand HSPs:

Score = 415 (114.7 bits), Expect = 3.1e-24, P = 3.1e-24
Identities = 103/128 (80%), Positives = 103/128 (80%), Strand = Minus / Plus

Query: 128 TCAATAGCTTCAGATATGCGCTTCACAGAGATCTTCGCAGAGAAGGAGGTCAGGCAAGA 68
           ||| |||| ||||| ||||||||||  ||||| ||| ||||||| ||||| | ||||||
Sbjct:3178 TCAGTAGCCTCAGACATGCGCTTCAATGAGATTTTCACAGAGAAAGAGGTTAAGCAAGA 3237

Query:  69 GAGGAAGAGACAAAGAGTGATCAATTCACGCTACCGGGAAGATTACATTGAACCTGCCT 8
            |  |||  | ||| |||| ||||| ||| | |||||||||||||| ||||| |||||
Sbjct:3238 AAAGAAACGGCAACGAGTTATCAACTCAAGATACCGGGAAGATTATATTGAGCCTGCAC 3297

Query:   7 GTACCGG 1
            ||||||
Sbjct:3298 ATACCGG 3304
```

Figure 5.10a Cosmid 32 BLASTN Results

```
                                                                  Smallest
                                                                    Sum
                                                    Reading  High  Probability
                                                     Frame  Score  P(N)       N

dbj|BAA31652|     (AB014577) KIAA0677 protein [Homo...-2    124   1.3e-08    1

dbj|BAA31652| (AB014577) KIAA0677 protein [Homo sapiens]
Length = 1064

Minus Strand HSPs:

Score = 124 (56.8 bits), Expect = 1.3e-08, P = 1.3e-08
Identities = 26/42 (61%), Positives = 27/42 (64%), Frame = -2

Query:    127 SIASDMRFTEIFAXXXXXXXXXXXXXXXINSRYREDYIEPALYR 2
              S+ASDMRF EIF                INSRYREDYIEPALYR
Sbjct:  1019 SVASDMRFNEIFTEKEVKQEKKRQRVINSRYREDYIEPALYR 1060
```

Figure 5.10b Cosmid 32 BLASTX Results

Figure 5.10 Cosmid 32 BLASTN and BLASTX Results: KIAA0667

numbers of genes found by this approach was small, the relative gene density was also calculated from genetic and physical mapping data (see sections 5.2.19, 5.2.20 and 5.2.21).

5.2.16 Absolute Gene Density

On the macrochromosomes, one gene homology per 44 kb of DNA was found, whereas on the microchromosomes, 1 gene homology per 33 kb of DNA was recorded. Sequencing projects involving *C.elegans, E.coli* and *S.cerevisiae* have shown that approximately 50% of all genes are found by sequence homology with the current databases (Jones, 1995), due to the limited number of conserved sequences amongst eukaryotes which diverged 540-580 million years ago. Approximately 40% of eukaryotic genes contain these conserved sequences and 85% of these regions have been previously characterised within the sequence databases (Green *et al.*, 1993). A sequencing project involving 1000 chicken cDNAs also confirmed this observation, with 45% of genes found by database hits (Bumstead, Personal communication). These gene density predictions are an underestimate of the number of genes and the true gene density is likely to be closer to one gene every 22 kb on the macrochromosomes and a gene every 17 kb on the microchromosomes.

5.2.17 Estimation of Gene Number in the Chicken Genome

The gene density estimates of 1 gene every 22 kb on the macrochromosomes and 1 gene every 17 kb on the microchromosomes, and the sizes of the chromosomes (Smith and Burt, 1998), were used to calculate the number of genes in the chicken genome (excluding the Z and W chromosomes). The total number was estimated to be 59,000, with 38,000 estimated to be macrochromosomal and 21,000 microchromosomal.

5.2.18 CpG Content

The GC content of the raw sequencing data was analysed. Potential CpG islands were identified as having a GC% greater than 50% and an observed over expected figure of 0.6 (Antequera and Bird, 1993). From this data it was impossible to tell if there were true CpG islands due to the incomplete sequencing data and lack of knowledge of the methylation status of the DNA *in vivo*.

As an alternative, the average CpG content of each cosmid was analysed. This is an indirect measure of CpG islands and the results are summarised in Table 5.3. The average number of CpGs/kb of sequence for each cosmid and overall average CpG contents for macrochromosomes and microchromosomes were calculated. On average, the macrochromosomes were found to contain 13.63 CpGs/kb and the microchromosomes 21.69 CpGs/kb. The CpG content of microchromosomal DNA was therefore estimated to be 1.6 times higher than on the macrochromosomes ($p< 0.0001$ and 95% confidence intervals of 1.2, 2.1).

5.2.19 Relative Gene Density from Genetic and Physical Mapping Data

A difficulty with this data is determining how significant the differences in gene density between the two chromosome types as the sample size is small. To combat this, the distribution of genes from physical and genetic mapping data was used to calculate relative gene density.

As the likelihood of mapping a gene to a chromosome by physical or genetic methods relies on the underlying gene density, previous genetic and physical mapping data was used to directly estimate the relative gene density. This was achieved by a comparison of the number of genes mapped by either method to both macrochromosomes and microchromosomes. Only markers which were mapped randomly with no *a priori* knowledge of their position were used, as some of the

Cosmid	Physical Assignment	CpGs/kb	Average CpGs/kb
28	2p12-p11	9.21	
35	2q26-q32	13.97	
34	2q32-q35	11.88	
08	3q11	15.17	
30	3q23-q33	15.09	
16	4p13-p12	15.13	
33	4p14-p13	9.85	
27	5q21-q22	18.73	**13.63**
01	Mic	23.84	
07	Mic	16.67	
14	Mic	20.84	
20	Mic	27.32	
21	Mic	20.62	
31	Mic	29.23	
32	Mic	14.33	
36	Mic	20.66	**21.69**

Table 5.3 Cosmid Sequencing Sampling Data: CpGs/kb and Average CpGs/kb

Mic-Microchromosome

markers which have been mapped were chosen with a predetermined knowledge of their position in the genome.

5.2.20 Physical Mapping Gene Density Data

A total of 42 anonymous clones and 32 genes have been mapped physically and are listed in Table 5.4. The mapping distribution of these clones is approximately equal between the two chromosome types. If the distribution of the anonymous clones was according to genome size (as represented by the macrochromosomes and microchromosomes), then a 70:30 split would be expected. This discrepancy would indicate that there may be some kind of cloning bias inherent within the genomic libraries or a gene density difference. To allow for the differences in the length of genome being sampled by this method, a number of anonymous physical markers mapped with no *a priori* knowledge of gene content were used. This served as an indirect measure of the physical size of genome sampled.

A total of 20 anonymous clones and 14 genes have been assigned to the physical maps of macrochromosomes. This is a ratio of 0.70 genes/anonymous loci. For the microchromosomes, 20 anonymous and 18 gene markers have been mapped. This translates to a ratio of 0.90 genes/anonymous loci. The physical mapping data presented here can be used to estimate the microchromosomes to be 1.3 times as dense as the macrochromosomes (with a 95% confidence interval of 0.5, 3.3). In total, 108 genes have been mapped by FISH but many of these target a specific chromosome or were selected based on comparative maps with human (unpublished data). The small sample of randomly mapped genes and the skewed distribution of cloned genomic DNA is a limitation of this approach. To combat this, genetic linkage data was analysed with a larger sample size.

Locus	Chr No	R	Locus	Chr No	R	Locus	Chr No	R
ASCL	MAC	0	AVDL@	MAC	1	IGF1R	MIC	0
CCND2*	MAC	1	IFN1*	MAC	1	MC1R	MIC	0
CCNDP2*	MAC	0	TRKB*	MAC	1	NCAM1*	MIC	1
DCN	MAC	0	ALDH	MIC	0	NGFB*	MIC	1
GAPD*	MAC	0	NRAMP1*	MIC	1	OPCML	MIC	0
H5*	MAC	0	RPL37A*	MIC	1	OVM*	MIC	1
HISA@*	MAC	0	RPL5*	MIC	1	PGA@	MIC	0
IGF1*	MAC	1	ABL1	MIC	0	PPY	MIC	0
PGR*	MAC	1	ACACA*	MIC	1	PRNP	MIC	0
UCP2	MAC	0	ADORA1	MIC	0	RAF1*	MIC	0
TGFBR1	MAC	0	ADORA3	MIC	0	RARB	MIC	0
ACTB*	MAC	0	AK1*	MIC	0	RNR*	MIC	1
CCNC*	MAC	1	ANX2	MIC	0	RPL7A*	MIC	1
MYB*	MAC	0	B2M*	MIC	0	SLC6A4	MIC	0
HMG14*	MAC	1	BBC1	MIC	1	SUV3	MIC	1
IL8	MAC	0	BMP7	MIC	0	TAX1	MIC	0
IFR2*	MAC	1	CAMLG	MIC	0	TF*	MIC	1
KIT	MAC	0	CCNE*	MIC	1	TRAF1	MIC	0
PGK1*	MAC	1	CD3E	MIC	0	H3F3B*	MIC	0
CCND1*	MAC	1	CDC2L1*	MIC	1	DMD*	MIC	1
HTR1D	MAC	0	COS0032*	MIC	0	FASN*	MIC	1
MAX*	MAC	1	CRABP1	MIC	0	FES*	MIC	0
TGFB3*	MAC	1	DCM11*	MIC	1	FLN2	MIC	0
TH*	MAC	1	SCD*	MIC	1	FMOD	MIC	0

Table 5.4 Genes Which Have Been Mapped Physically by FISH

* References in Chicken genome database (http://www.ri.bbsrc.ac.uk) R-random, 1 = yes 0 = no; MIC-Microchromosome; MAC-Macrochromosome

5.2.21 Relative Gene Density from Genetic Mapping Data

The numbers of genetically mapped genes were compared with anonymous loci associated with genetic markers to calculate gene density. The data in Table 5.5 shows the total number of randomly mapped genes to be 147. To construct the genetic map, a range of genetic markers such as microsatellites, RFLPs, RAPDs, SSCPs and CR1 repeats have been employed (Burt and Cheng, 1998). The genetic map is now close to completion, as a new genetic marker has a greater than 95% chance of linkage to another previously mapped marker (Smith and Burt, 1998).

5.2.22 Distribution of Microsatellite Sequences and Other Genetic Markers

It has been hypothesised that microsatellite sequences are not randomly distributed across the chicken genome and are present on the microchromosomes at a lower density (Primmer *et al.*, 1997). This was tested by comparing the distribution on the genetic map of microsatellites and other genetic markers for anonymous and gene sequences. The results are shown in Table 5.6. It was expected, based on their physical size (Smith and Burt, 1998), that in the case of anonymous markers, 70% would map to the macrochromosomes, and 30% to the microchromosomes. However, more microsatellite markers than expected were found on the microchromosomes. Out of a total of 368, 157 belong to a microchromosome, (43%). The discrepancy between the results presented here and that of Primmer may be due to the PRINS approach used by (Primmer *et al.*, 1997) on the chicken microchromosomes.

The distribution of other genetic markers on the macrochromosomes appears to be random and proportional to the physical size of each chromosome (70%:30%). This indicates that the distribution of genes based on these kinds of genetic markers are more likely to mirror to the true distribution of genes over the chicken genome.

Table 5.5 Genes Which Have Been Mapped by Genetic Linkage Analysis

* References in the Chicken genome database (http://www.ri.bbsrc.ac.uk/) R-Random, 1= yes 2 = no; SSCP-single strand conformation polymorphism; SNP-single nucleotide polymorphism; RFLP-restriction fragment length polymorphism; MICR-microsatellite; CLAS-classical

Locus	Chr. No.	Type	R	Locus	Chr. No.	Type	R	Locus	Chr. No.	Type	R
SMOH	1	SSCP	1	BMP2	3	RFLP	1	LPL*	Z	SSCP	1
NRCAM*	1	SNP	1	TGFB2*	3	RFLP	1	ALDOB*	Z	SNP	1
GNRH	1	SSCP	1	ACTN2*	3	SNP	1	GGTB2*	Z	SNP	1
NAGA*	1	SNP	1	HMX1*	3	RFLP	1	ROS0028E	6	MICR	0
LGALS4*	1	SNP	1	T	3	RFLP	1	PDE6C*	6	SNP	1
COM0119E*	1	RFLP	1	IGF2R*	3	SNP	1	ACTA2*	6	MICR	0
IGF1*	1	RFLP	1	VIP	3	SSCP	1	SCD*	6	SSCP	1
LDHB*	1	RFLP	1	ESR*	3	RFLP	1	PSAP*	6	RFLP	1
GAPD*	1	SNP	1	PLN*	3	MICR	0	CPII*	7	CLAS	1
HSD3B*	1	SNP	1	BMP5	3	RFLP	1	ROS0019E	7	MICR	0
COM0155E*	1	RFLP	1	GSTA2*	3	SNP	1	CD28*	7	RFLP	1
ROS0044E	1	MICR	0	ODC1	3	RFLP	1	EEF1B2*	7	RFLP	1
ROS0081E	1	MICR	0	CPPP*	3	CLAS	1	ROS0021E	8	MICR	0
LAMP1*	1	MICR	0	COM0094E*	4	RFLP	1	GGTB1*	8	SNP	1
COM0092E*	1	RFLP	1	HMG14A*	4	RFLP	1	VTG2*	8	SNP	1
RB1*	1	SNP	1	FMR1	4	MICR	0	ROS0026E	8	MICR	0
FUCT4*	1	SNP	1	SPP1*	4	SNP	1	PLA2G2A*	8	SNP	1
ROS0025E	1	MICR	0	COM0117E*	4	RFLP	1	B@*	16	RFLP	1
ROS0055E	1	MICR	0	MSX1	4	RFLP	1	RFP-Y@*	16	RFLP	1
WNT11*	1	RFLP	1	CD8A*	4	RFLP	1	CPAA*	E04	CLAS	1
SHH	2	SSCP	1	CAPN1*	5	SNP	1	CPEE*	E04	CLAS	1
PENK*	2	SNP	1	RYR3*	5	SNP	1	PGA@*	E04	SNP	1
CA2*	2	SNP	1	HTR1D	5	SSCP	1	ARF4*	E16C17W22	RFLP	1
CALB1*	2	RFLP	1	TGFB3*	5	RFLP	1	MIF*	E18C15W15	RFLP	1
NPY	2	SSCP	1	COM0089E*	5	RFLP	1	IGL@*	E18C15W15	MICR	0
ROS0018E	2	MICR	0	DNCL*	5	RFLP	1	CRYBB1*	E18C15W15	MICR	0
CP49*	2	RFLP	1	CKB*	5	SNP	1	I*	E22C19W28	CLAS	1
TGFBR1	2	RFLP	1	BMP4	5	RFLP	1	ROS0054E	E22C19W28	MICR	0
PRL	2	SSCP	1	PRLR*	Z	SSCP	1	GLI	E22C19W28	SSCP	1
BMP6	2	RFLP	1	MSU0068E*	Z	RFLP	1	TGFB1	E25C31	RFLP	1
BCL2*	2	MICR	0	ROS0072E	Z	MICR	0	RYR1	E25C31	SSCP	1
ZNF5*	2	SNP	1	PTCH*	Z	SSCP	1	GNRHR	E29C09W09	SSCP	1
ROS0023E	2	MICR	0	ROS0017E	Z	MICR	0	IGF1R*	E29C09W09	RFLP	1
ROS0074E	2	MICR	0	CHRNB3*	Z	SNP	1	AGC1*	E29C09W09	SSCP	1

continued

Table 5.5 continued

Locus	Chr. No.	Type	R	Locus	Chr. No.	Type	R	Locus	Chr. No.	Type	R
HMX3*	E29C09W09	RFLP	1	CPRR*	E38	CLAS	1	POU2F3*	E49C20W21	RFLP	1
CYP19	E29C09W09	SSCP	1	COM0152E*	E38	RFLP	1	APOA1*	E49C20W21	SNP	1
CCNE	E30C14W10	SSCP	1	ROS0020E	E41W17	MICR	0	W*	E49C20W21	CLAS	1
MYH@*	E31E21C25W12	MICR	0	RPL7A*	E41W17	RFLP	1	ACACA*	E52W19	SSCP	1
ROS0022E	E31E21C25W12	MICR	0	ABL1*	E41W17	SNP	1	CRK*	E52W19	SSCP	1
H3F3B*	E31E21C25W12	RFLP	1	CD39L1*	E41W17	SNP	1	AMH	E53C34W16	MICR	0
FASN*	E31E21C25W12	SSCP	1	AMBP*	E41W17	SNP	1	TVA*	E53C34W16	RFLP	1
COM0093E*	E31E21C25W12	RFLP	1	HSF1*	E46C08W18	MICR	0	CDC2L1*	E54	SNP	1
HLF	E31E21C25W12	MICR	0	ITGAM*	E46C08W18	MICR	0	AGRN*	E54	SNP	1
BMP7*	E32	RFLP	1	POU4F3*	E48C28W13	RFLP	1	ENO1*	E54	SNP	1
FZF*	E32	SNP	1	CAMLG*	E48C28W13	SNP	1	PLOD*	E54	SNP	1
ROS0078E	E36C06W08	MICR	0	CDX1*	E48C28W13	SNP	1	SLC2A1*	E54	RFLP	1
EIF4A2*	E36C06W08	RFLP	1	ROS0083E	E48C28W13	MICR	0	TP53*	E57	SN P	1
SNON*	E36C06W08	SNP	1	MSX2	E48C28W13	RFLP	1	COL1A1*	E59C35W20	MICR	0
ROS0073E	E38	MICR	0	OPCML	E49C20W21	SSCP	1	ROS0071E	E59C35W20	MICR	0

	Anonymous Genetic Markers				Genetic Markers of Genes			
Chromosomes	Microsatellites	%	Other	%	Microsatellites	%	Other	%
MAC(5A+Z)	211	57.3	193	69.4	13	39.4	58	50.9
MIC(33A+W)	157	42.7	85	30.6	20	60.6	56	49.1
Total	368	100.0	278	100.0	33	100.0	114	100

Table 5.6 Distribution of Genetic Markers on Macrochromosomes and Microchromosomes

The genetic markers are based on the East Lansing genetic map (Burt *et al*., 1995)

5.2.23 Gene Density

A total of 58 gene markers, excluding microsatellite-based markers, have been randomly assigned to the genetic maps of the macrochromosomes (Table 5.7). For the microchromosomes, a total of 56 gene markers have been mapped. By comparing the genetic mapping data to the physical size of the genome, it can be estimated that the microchromosomes are 2.3 times as gene dense as the macrochromosomes (with a 95% confidence interval of 1.6, 3.3).

Chromosome	Size (Mb)[a]	Genetic Linkage[b]		Relative	Physical Linkage[d]		Relative
Number		Anonymous	Gene	Density[c]	Anonymous	Gene	Density
MAC(5A+Z)	843	193	58		20	14	
MIC(33A+W)	357	85	56	2.3	20	18	1.3

Table 5.7 Gene Density on Macrochromosomes and Microchromosomes Based on Genetic and Physical Mapping Data

[b]See Table ; [d]See Table ; [a](Smith and Burt, 1998); [c]Using physical sizes (Mb) as shown in this table; MAC-Macrochromosome; MIC-Microchromosome

5.3 Discussion

5.3.1 Sequence Sampling Experiments

The results presented in this chapter indicate that sequence sampling is an effective method of gene discovery, with genes being found in cosmids with a 50% success rate. This approach has led to the identification of previously unknown chicken genes such as *5-HT$_{1D}$* and *OBCAM.*

The random cosmids required subcloning prior to sequencing. Sonication and restriction enzyme digestion were considered as ways of fragmenting the DNA into suitable sizes for sequencing. Sonication is advantageous in that it produces random overlapping fragments, but after initial experiments this method was rejected as large concentrations of DNA were required for subcloning and optimal conditions varied widely between cosmids. Other drawbacks to this method included a low transformation efficiency and a requirement for the additional step of end-repairing the fragments prior to cloning.

Fragmentation by restriction digests was evaluated initially using *Sau*3A and *Eco*RI. After initial experiments, this approach was also rejected as the cosmid had to be subcloned several times with different enzymes for sufficient cosmid coverage. The restriction enzyme *Cvi*JI was then employed as it cuts at a sufficient frequency to produce fragments with a randomness similar to that gained with sonication (Xia *et al.*, 1987) and unlike sonication, it required no blunt ending prior to ligation into a plasmid. A drawback with using this enzyme was that, like sonication, the conditions varied from cosmid to cosmid, demanding testing and high concentrations of DNA. A potential solution to the demand for large amounts of DNA is a transposon method, which is currently being used within this laboratory (Biolabs, 1998).

The Staden package was used to analyse the sequencing data and was modified to assess GC/CG composition (Staden *et al.,* 1996). It was effective at screening out poor DNA sequence and vector. The criterion for a gene homology using a BLASTN search (DNA Vs DNA), was a high score of greater than 150 and a probability of less than 10^{-9}. For a protein-protein BLASTX search (Gish and States, 1993) a score of greater than 75 and a probability of less than 10^{-6} was taken to indicate a significant match. These conditions were more stringent than the criteria used to define a significant gene homology in a similar study which "sequence scanned" nineteen chicken cosmids (Nurminsky *et al.,* 1996);(Clark *et al.*, 1999). Prior to this study cosmid sequence scanning was used on *Fugu* cosmids to study gene content, genome organisation and synteny (Elgar, 1996);(Elgar, *et al.*, 1998). For the chicken cosmids a probability score of less than 10^{-4} for protein database homologies yielded twelve significant database matches, with 3 to a macrochromosome and 9 to a microchromosome. These results are in agreement with the data presented here as both studies found more significant hits on the microchromosomes than on the macrochromosomes. However, the sequence scanning study analysed more microchromosomal (10) than macrochromosomal cosmids (8). An and intermediate microchromosome cosmid was also analysed but it did not have any gene homologies. This may, in part, account for the higher number of database matches. Also, a single, highly gene rich, microchromsomal cosmid was isolated. This contained four significant database matches, accounting for half of the database hits for the microchromosomal cosmids, this one cosmid thus biasing the number of genes. No such cosmid was isolated during the work carried out for this thesis.

The database searches described in this chapter, with a 50% hit rate, were effective at finding gene homologies. A drawback was that repeats, other than the species-specific CR1 elements, were filtered out, but a BLAST search without the filter would have been too inefficient. In addition, such a search would detect 'false' matches due to simple repeats such as all DNAs with an AC_n repeat. However, this problem of false matches would only be encountered if searching for repeats by database searches. For simple repeats it is advantageous to use a program that detects them such as report repeats (Law, Personal Communication). This is an adaptation of the XBLAST and XNU programs which mask repetitive sequences (Claverie and States, 1993);(Law, Personal Communication).

Few significant hits to human ESTs were found, most likely reflecting the 350 million years of evolutionary divergence between birds and mammals and the bias of ESTs for 3' ends. With such a distance between species, homologies may not be detected. More chicken ESTs are required and work is currently underway to rectify this (Burt, Personal Communication).

Gene finder programs such as GRAIL (Gene Recognition and Analysis Internet Link) were not used for a number of reasons. First, the average sequence read of 400 bp meant the sequences were too small for an accurate isolation of potential exons and introns. The programs are modelled on human sequences and do not take into account such features as the GC content of chicken DNA. In addition, it is also difficult to determine if the potential exon found is real without a database match. It has been reported that the GeneMark program has been used to analyse chicken sequences for open reading frames. However when compared with known gene content it gave a 17% correct ORF prediction. Therefore no predictions concerning novel genes or potential coding motifs were made (Clark *et al.*, 1999).

5.3.2 Conservation of Synteny

Cosmid 27 contains two G-protein receptor genes, *PTAFR* and *5-HT$_{1D}$* and maps to the chicken macrochromosome 5 at position q21-q22 and is an example of conserved synteny. As they are both members of a G-protein receptor family they may be the products of an ancient tandem gene duplication. As mapping information on the region in chicken is limited, it is not known if this suggested duplication would have generated more members of the G-protein receptor family in the region. A survey of the corresponding chromosome region in humans has shown no other G-protein receptors map in this region, so this may not be the case.

As the sequencing data on this cosmid is incomplete, the orientation of *5-HT_{1D}* and *PTAFR* cannot be derived. The proximity of the two genes to each other cannot be determined without further work.

5.3.3 Relative Gene Density on Macrochromosomes and Microchromosomes

Analysis of chicken cosmids using CpG islands as a measure of gene content was carried out (McQueen *et al.*, 1998). The results showed CpG island-like fragments to be approximately six times denser on the microchromosomes and to account for 75% of all chicken genes. If true, this should have been reflected in a higher number of significant hits from the sequence sampling work presented in this study. For the sixteen cosmids used, up to sixteen gene homologies would be expected from the microchromosomal cosmids and two or three from the macrochromosomal cosmids. From this study it is estimated that the microchromosomes account for only 50% of all chicken genes, accounting for the lower number of homologies from the sequence sampling work.

This study found the relative gene density, from sequence sampling data, on microchromosomes to be only 1.3 times as gene dense as macrochromosomes. Genetic and physical mapping data gives a gene density difference of between 1.5-fold to 3.5-fold. These estimate is lower than the six-fold estimate suggested by a prior study based upon CpG islands and acetylation studies (McQueen *et al.*, 1998). In this study sequence sampling was used. This is a very direct means of detecting genes by sequence similarity with known genes available in databases. However, only a small data set was examined, which may account for the lower gene density estimates. Despite this, the sequence sampling results, in conjunction with genetic and physical mapping data, do indicate a definite two-fold difference in gene density between the two chromosome types. Therefore, both studies are in qualitative agreement in that microchromosomes are more gene dense.

5.3.4 Estimation of Total Gene Number in the Chicken Genome

An estimate of 59,000 genes (38,000 macrochromosome, 21,000 microchromosome) in the chicken genome was made, with a 21:38 split of genes between microchromosomes and macrochromosomes. An equal gene density between microchromosomes and macrochromosomes would give a split of genes of 25:75. This total gene number of 59,000 is in the region of previous estimates made for vertebrates of 80,000, (Antequera and Bird, 1993), 60,000 (Elgar, 1996), and 64,000 (Fields *et al.*, 1994). The estimation of gene numbers calculated in this study is close to an earlier report which based its estimate of 55,000 genes in the chicken genome on CpG island data, of which 42,000 genes were predicted to be microchromosomal portion, and 13,000 in the macrochromosomal (McQueen *et al.*, 1998).

5.3.5 CpG Content

The CpG content of microchromosomal DNA was 1.6 times higher than that of the macrochromosomes, and similar to the two-fold differences found by gene mapping. This is in contrast to the earlier studies which suggest that microchromosomes are CpG island-rich and macrochromosomes are CpG island-poor (McQueen *et al.*, 1996; McQueen *et al.*, 1998). It was suggested that microchromosomes may have a six-fold greater CpG content than the macrochromosomes (McQueen *et al.*, 1998), which is much higher than the estimate in this study and may reflect the different approaches used to assess CpG content as well as differences in sample size. Whether this high CpG content is related to the actual number of CpG islands or to some other feature of microchromosomal DNA remains to be seen. So far, no bias for house-keeping genes on the microchromosomes and for tissue-specific genes on the macrochromosomes has been found (McQueen *et al.*, 1998), implying other reasons for the CpG content differences between the chromosome types. However, these results do suggest a correlation between CpG content and gene content.

5.4 Conclusion

This study has shown that sequence sampling is a reasonable means of finding gene homologies, with a 50% success rate being observed. Previously un-categorised chicken genes have been cloned and a new example of conserved synteny has been identified.

The results suggest that microchromosomes are approximately twice as gene dense as macrochromosomes, which is not as high as other studies have indicated. Similarly, there is a 1.6 fold difference in CpG content between macrochromosomes and microchromosomes, which is lower than previous predictions.

Finally, these data allowed an estimate to be made of the number of genes in the chicken genome. The estimate of 59,000 (on the autosomes) is close to estimates for other vertebrates.

Chapter 6

Distribution of CR1 Repeats

6.1 Introduction

In comparison to other vertebrate species, the chicken has fewer repeats comprising of an estimated 17% of the genome and may be a reflection of the small genome size (Eden and Hendrick, 1978). A similar observation has been made for bats which also have a small genome size and fewer microsatellite repeats than other mammals (Van Den Bussche *et al.*, 1995). However, this apparent lack of repeats has not hampered mapping efforts as the chicken genetic map is close to completion (Smith and Burt, 1998).

Short, simple repetitive elements and species-specific, CR1 repeats are retrotransposons found throughout the chicken genome. How these repeats are distributed can be used to test the hypothesis that if microchromosomes are gene dense, then they should have fewer repeats because they contain less 'junk' DNA and therefore a limited potential target area for insertion events. This is assuming the insertion of a CR1 repeat into a gene would have deleterious consequences. This chapter discusses the distribution and evolution of avian CR1s.

6.1.2 Interspersed Repetitive Elements

Interspersed repetitive elements are divided into two categories. The first, such as Short Interspersed Nucleotide Elements (SINES) are a class of retroposons which do not encode their means of retrotransposition. They are derived from transcripts of RNA polymerase III. An example of a SINE repeat are *Alu* elements which are found in the human genome.

The second category of interspersed repetitive elements which do encode their own means of retrotransposition and have retrovirus-like elements flanked by long terminal repeats (LTRs). Elements without these repeats are called non-LTR retrotransposons, and in vertebrates non-LTR retrotransposon families include the mammalian L1 (or LINE-1, long interspersed nucleotide element-1), the *Xenopus* Tx1 repeat and the CR1 element in birds (Vandergon and Reitman, 1994).

6.1.3 Chicken Repeat 1 Elements

CR1 elements are short interspersed DNA elements which were first discovered in chicken. First thought to be homologous to human *Alu*I and mouse B1-B2 repeats, they are now known to belong to the non-long terminal repeat transposons repeat family (Olofsson and Bernardi, 1983). The CR1 repeats have retroviral LTR features such as terminal inverted repeats, primer binding sites and terminal sequence homology (Shapira *et al.*, 1991) and retrotranspose by a 'nick and prime' mechanism similar to other families of non-LTR elements (Haas *et al.*, 1997). Also, CR1 elements often contain ORFs which are able to code for proteins (Silva and Burch, 1989) and (Vandergon and Reitman, 1994). They have conserved 3' ends with a truncated 5' end, with their lengths ranging from 160 to 850 bp. Most CR1s are 400 bp or less with only 3-5% close to 800 bp. Only 0.1% are over 2 kb in length and have a full length parental element (Hache and Deeley, 1988), (Chen *et al.*, 1991) and (Burch *et al.*, 1993). From the phylogeny of chicken CR1s, they have been subdivided into at least six subfamilies, named A-F. It has been estimated that there is approximately one CR1 element per 10,000 bp in the chicken genome (Dodgson *et al.*, 1997).

Over time, CR1 repeats have become polymorphic by a loss/gain of DNA sequences and have become an important part of chicken genetic maps. A number of markers are associated with a CR1 repeat (Cheng *et al.*, 1995) and (Okimoto *et al.*, 1997). For example, 24 CR1 based markers were placed on the East Lansing map using PCR.

CR1 elements are not only confined to chicken but have been observed in other avian species as diverse as emu, cassowary, pelican, stork, condor, quail, crane, owl, magpie, peacock, songbird, duck and turkey (Chen *et al.*, 1991), (Okimoto *et al.*, 1997), (Shapira *et al.*, 1991) and (Silva and Burch, 1989).

6.1.4 Distribution and Number of CR1 Repeats in the Chicken Genome

The CR1 repeats could be distributed in a uniform or non-uniform manner across the genome. Early data suggest the latter, with the highest concentration of repeats found in the G+C rich (48%) portion of chicken DNA (Olofsson and Bernardi, 1983). The reason for this is not known but may lie in how the CR1 subfamilies are dispersed. There may, for example, be a clustering of certain types of subfamily on the microchromosomes. In addition, how the subfamilies came to be distributed would have been governed by their ability to transpose themselves into different regions of the genome.

Hot spots for CR1 insertion have been found. For example the chicken β-globin cluster contains a high number of CR1 repeats, sizes ranging from 38-938 bp. There appears to be no pattern to their distribution within the β-globin cluster as they were found both close to and distant from genes and hypersensitive sites (Reitman *et al.*, 1993).

Estimates of the number of CR1s in the chicken genome vary. Early hybridisation data estimates varied between 7,000 and 30,000 (Burch *et al.*, 1993; Stumph *et al.*, 1984). From sequence analysis, 100,000 CR1 repeats were predicted and it was also estimated that they make up 2% of the chicken genome (Vandergon and Reitman, 1994). Little is known about numbers of CR1 repeats in other birds but hybridisation studies indicate there may be similar numbers in the duck genome (Li *et al.*, 1995).

6.1.5 Potential Roles of CR1 Repetitive Elements

Various studies have suggested that some CR1 elements may play a role in gene expression. A hypothesis as to how this came about is that the CR1 repeat was acquired, then some kind of regulatory role would have evolved. For instance, CR1 repeats in the region of a gene could indicate a role in the control of gene expression. DNA sequence comparisons between sarus crane (*Grus antigone*) and emu (*Dromaius novaehollandiae*) CR1 elements have shown two regions, a transcriptional silencer and a nuclear protein binding site, to be highly conserved (Chen *et al.*, 1991). Such features suggest a functional role.

CR1 repeats can be associated with genes such as the one located in the 5' flanking end of the chicken α-skeletal actin gene, *asa,* with its 3' end closest to the start of the gene. It was initially thought to act as a transcriptional silencer but it was later shown to have little effect on the *asa* gene (French *et al.*, 1990). Other examples of CR1 associated genes include the avian very low density apolipoprotein II (apoVLDII) gene which encodes a small phospholipid binding protein and has a CR1 repeat in the 5' flanking region which also has three DNAse hypersensitive sites (Hache and Deeley, 1988).

The 5' flanking region of two avidin-related genes *Avr4* and *Avr5* both have CR1 repeats which point towards the genes. Both repeats contain a nuclear protein binding consensus sequence and have a 191 bp deletion in a region corresponding to the functional silencer regions. These were found within the CR1 elements upstream of the chicken lysozyme and apoVLDLII genes (Wallen *et al.*, 1996).

CR1s have been found in the 3'-flanking region of the chicken vitellogenin gene. This gene produces a yolk precursor protein found solely in the liver of laying hens. The CR1 element lies 2.2 kb downstream of the gene, pointing away from it in a region showing changes in chromatin structure. This indicates a possible role in the determination of the structural state of the surrounding chromatin [Schip, 1987 #76].

6.1.6 CR1-Like Elements in Other Species

The genomes of fish, amphibian and reptilian species such as Turtle (Kajikawa *et al.*, 1997), frog and two species of torpedo ray (Burch *et al.*, 1993) have non-longterminal-repeat (non-LTR) retrotransposons which display sequence similarity with CR1 repeats. This suggests CR1s may have arisen before the divergence of birds and reptiles (300-400 Mya). Due to this, they have been named CR1-like elements. (Drew and Brindley, 1997) and (Okimoto *et al.*, 1997).

Non-LTR retrotransposons sequences called the SR1 family have been isolated from the human blood fluke *Schistosoma mansoni.* These repeats share amino acid and structural similarities with CR1-like elements, placing them in the CR1-like group of retrotransposons. This is the first time this type of repeat has been discovered in a non-vertebrate (Drew and Brindley, 1997). It is possible that CR1-like elements are not only found in vertebrates and may indicate that the original CR1 repeat is older than was previously thought, possibly > 500 million years old.

6.1.7 Evolution of the CR1 Repeat

Chicken CR1s have been subdivided into at least six subfamilies, named A-F and Figure 6.1 outlines how they may have evolved. Within subfamilies B, C, D and F highly similar elements were observed, indicating that a distinct progenitor spawned each of these subfamilies. In subfamily C a nucleotide divergence of only 5-8% was observed, suggesting a recent occurrence of retrotransposition. The A and E subfamilies could have come from ancestors of these four progenitors or from another, distinct progenitor. The ancient nature of CR1 repeats was demonstrated by the divergence of the consensus sequences from each of the subfamilies. In each subfamily they have truncated 5' ends and a 3' end made up of ≥2 repeats of an 8-bp sequence (Vandergon and Reitman, 1994). After a repeat has been truncated and

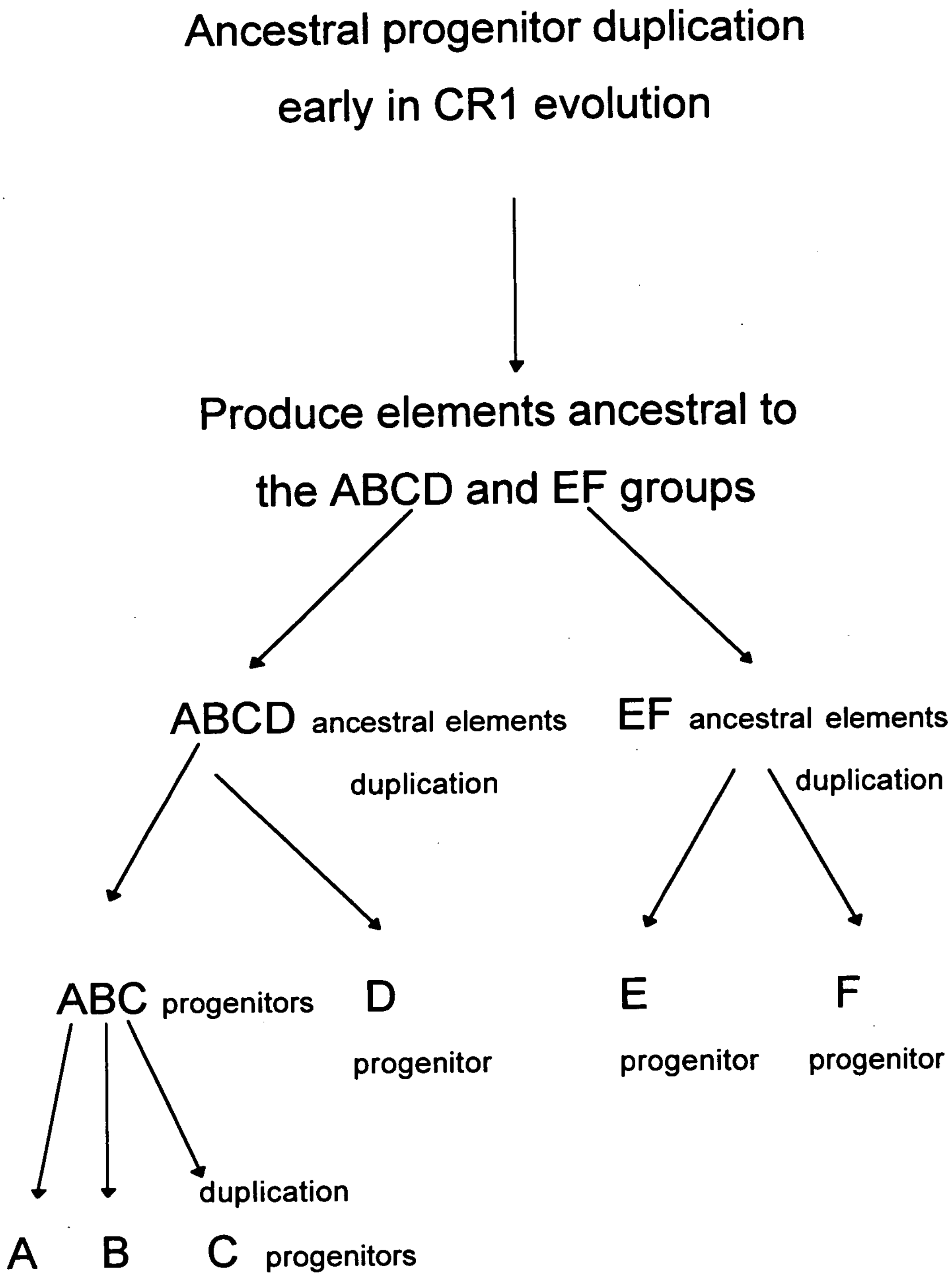


Figure 6.1 Model for the evolution of the CR1 family

The model is based on a 1994 paper by (Vandergon and Reitman, 1994) using 95 CR1 sequences transposed it has been predicted that it would lose any detectable similarity with other CR1 repeats after approximately 40 million years. A portion of the chicken genome may therefore comprise of CR1 repeats degenerated as far as to be undetectable (Okimoto *et al.*, 1997).

6.1.8 Analysis of the Distribution and Evolution of the CR1 Repeat

In Chapter 5, it was shown that microchromosomes are more gene dense than macrochromosomes and presumably more compact with less 'junk DNA'. It is possible that microchromosomes, which should have less 'junk' DNA, may have fewer CR1 repeats than the macrochromosomes. To test this hypothesis, a survey of the distribution of CR1 repeats across the genome was carried out using gene density and genetic mapping data. How the repeats are distributed can also give clues to the nature of macrochromosomes and microchromosomes.

There is evidence to suggest that CR1s are an ancient class of repeat, existing prior to the divergence of avians and reptiles. To examine how this repeat has evolved, a phylogenetic analysis was carried out on the CR1 data from this thesis and other studies. CR1 repeats were assigned to subfamilies and the distribution of subfamilies was examined with respect to chromosome and chromosome type (macrochromosome vs. microchromosomes).

6.2 Results

Database search results from the sequencing of clones from the Compton Cross (Chapter 3) and random cosmid clones (Chapter 5), generated a number of significant hits against CR1 repeats. Here I have collated the data and analysed their number, distribution and evolution.

6.2.1 Distribution of CR1 Repeats from Sequence Sampling Data

As Table 6.1 shows, a total of thirty three new CR1 repeats were found in the sixteen random cosmids studied. Twenty two mapped to a macrochromosome (including two CR1s on the Z chromosome) and eleven mapped to a microchromosome. This demonstrates that CR1 repeats are found on all chromosome types in the chicken, with more CR1 repeats mapping to a macrochromosome than to a microchromosome.

6.2.2 Number of CR1 Repeats on the Macrochromosomes and Microchromosomes

Based on the mapping data (Table 6.2) the number of CR1 repeats in the chicken genome has been estimated to be 172,000, with 142,000 on the macrochromosomes (1-5, Z chromosome) and 30, 000 on the microchromosomes (6-38, W chromosome).

Table 6.1 CR1 Repeats from the Gene Density Study

Mic- Microchromosome; Mac-Macrochromosome; Macrochromosomes are defined as chromosomes 1-5 and the Z chromosome. Microchromosomes are defined as chromosomes 6-8, the remaining chromosomes and the W chromosome; The CR1/Seqs column gives the number of CR1 repeat database hits which were found; The maximum number of CR1 repeats from contigs is the number of CR1 repeats found after alignments have been carried out. For example, for cosmid 28, three CR1 database matches were found out of 106 sequences analysed. From the alignments, two of the CR1 sequences overlapped, therefore the maximum number of CR1s found were 2.

Cosmid	Physical Assignment	CR1/Seqs	% of cosmid with a CR1 sequence	Maximum Number of CR1 Repeats from Contigs
28	2p12-p11	3/106	2.8	1
35	2q26-q32	3/90	3.3	2
34	2q32-q35	7/77	9.0	6
08	3q11	4/52	7.7	3
30	3q23-q33	6/61	9.8	5
16	4p13-p12	0/72	0.0	0
33	4p14-p13	2/66	3.0	1
27	5q21-q22	2/80	2.5	2
29	Z	3/54	5.5	2
01	Mic	0/80	0.0	0
07	Mic	3/49	6.1	1
14	Mic	8/46	17.4	6
20	Mic	2/114	1.8	1
21	Mic	0/72	0.0	0
31	Mic	0/68	0.0	0
32	Mic	1/88	1.1	1
36	Mic	2/64	3.1	2
			Total number	**of CR1s: 33**
				Mac: 22
				Mic: 11

	Number of CR1 repeats from database searches	Mb[a]	Estimate number of CR1 repeats in the genome	Relative density
Macrochromosome (5A+Z)	22 (131 kb)	843x 10^6	142,000	**2**
Microchromosomes (33A+W)	11 (131 kb)	357x10^6	30,000	**1**
Total	**33**	**1200x10^6**	**172,000**	

Table 6.2 Number and Relative Density of CR1 Repeats

[a] (Smith and Burt, 1998); CR1 data from sequencing experiments detailed in Chapter 5; A- autosomes

6.2.3 Density of CR1 Repeats on the Macrochromosomes and Microchromosomes

By comparing the macrochromosome and microchromosome data (Table 6.2), a relative CR1 repeat density of 2:1 (95% confidence interval 0.94 4.12), microchromosome vs. macrochromosome was calculated. From these data the macrochromosomes appear to be more CR1 dense than the microchromosomes. This difference in density is a trend, as the confidence interval suggests that this difference may not be significant and both chromosome types may have an equal density of CR1s.

6.2.4 Length of CR1 Repeats on Macrochromosomes and Microchromosomes

The physical position and length of the CR1 repeats found in experiments carried out in chapters 3 and 5 are presented in Table 6.3. Out of a total of 39 CR1 repeats, 28 were found to map to a macrochromosome and 11 to a microchromosome. Overall, the average length of the CR1 repeats was 150 bp. The average length on the macrochromosomes 159 bp and the average length on the microchromosomes was 124 bp. A plot of CR1 repeat length against linkage group size (Figure 6.2) was carried out and shows the lengths of microchromosomal CR1s tend to be shorter than the lengths of macrochromosomal CR1s, with the majority below 150 bp in length. On the macrochromosomes, the length of CR1s is evenly spread, with more repeats of 150 bp or above observed. This observation could be real or due to fewer CR1 database hits being found on the microchromosome cosmids sampled. To determine if this trend of difference in CR1 lengths between macrochromosomes and microchromosomes was significant, a t-test was carried out. A t value of 0.138 suggests that the differences in the lengths of CR1s between the two chromosome types is not significant. A larger data set would be required to see if lengths of repeats on the two chromosome types is different or not.

Table 6.3 Length, Map Position and Predicted Subfamily Assignments of CR1 Elements

UN- Unknown subfamily; Total Number of Repeats: 39; Macrochromosome 28, Microchromosome 11; Average length of CR1 repeats: 150 bp; Average length of CR1 on Macrochromosomes: 159 bp; Average length of CR1 on Microchromosomes: 124 bp

Clone Name	Physical assignment	Length of Repeat (bp)	Predicted CR1 Sub Family Assignment
T5/24T7	1	163	C
T2/82T3	2	208	UN
T12/108T7	2	73	UN
028-C42-T3	2p11-p12	73	F
035-C04	2q26-q32	70	D
035-CR1-01	2q26-q32	207	D
034-C16-T3	2q32-q35	201	F
034-C19-T7	2q32-q35	135	F
034-C67-T7	2q32-q35	209	F
034-CR1-01	2q32-q35	193	UN
034-CR1-04	2q32-q35	331	E
034-CR1-05	2q32-q35	227	F
034-C21-T3	2q32-q35	114	E
008-C17-T7	3q11	216	B
008-CR1-01	3q11	146	B
008-CR1-02	3q11	203	F
030-CR1-01	3q23-q33	209	F
030-CR1-02	3q23-q33	182	F
030-C02-T7	3q23-q33	123	E
030-C20-T3	3q23-q33	78	F
030-C21-T3	3q23-q33	78	E
030-C39-T3	3q23-q33	127	B
T2/70ABT7	4	249	F
033-C13-T7	4p14	68	F
037-C18-T7	5	170	E

continued

Table 6.3 continued

Clone Name	Physical assignment	Length of Repeat (bp)	Predicted CR1 Subfamily Assignment
027-C06-T3	5q21-q22	161	UN
029-C31-T7	Z	138	F
029-C32-T3	Z	138	F
007-CR1-01	Mic	198	UN
014-C08-T7	Mic	99	F
014-C02-T3	Mic	123	UN
014-C14-T7	Mic	83	F
014-C28-T7	Mic	73	UN
014-C29-T7	Mic	98	F
014-C32-T7	Mic	137	D
036-C45-T7	Mic	280	UN
036-C30-T3T7	Mic	52	F
020-CR1-01	Mic	148	F
032-C26-T3	Mic	78	E

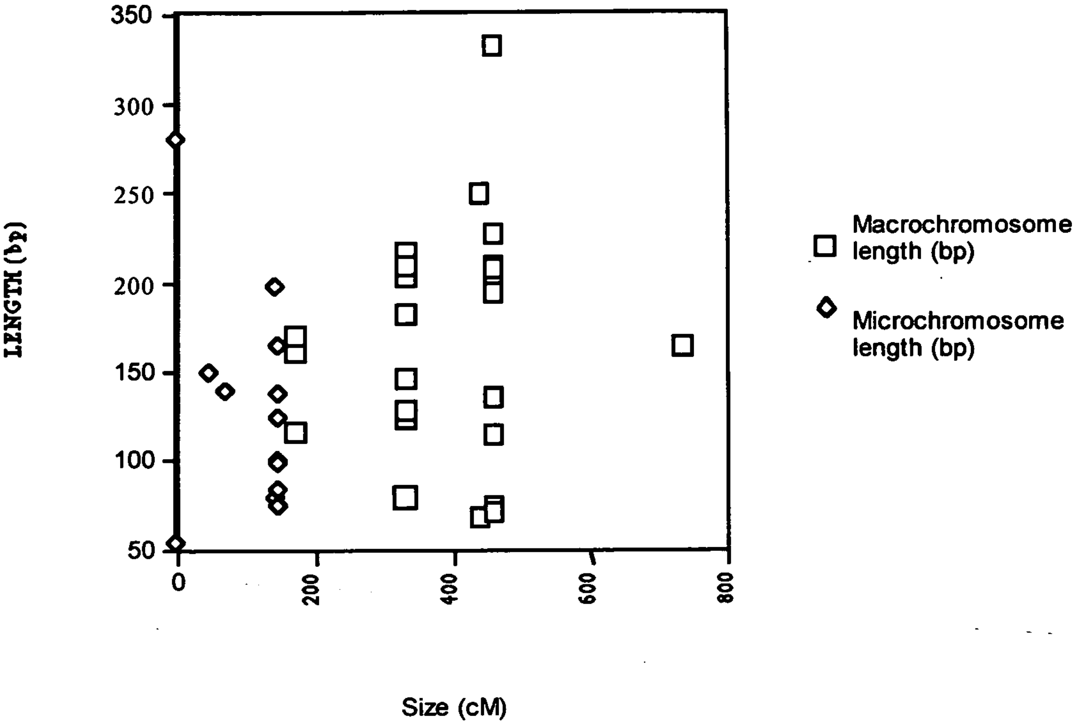


Figure 6.2 Length of CR1 Repeats vs. Linkage Group Size

6.2.5 Assignment and Distribution of CR1 Sub-Families on Macrochromosomes and Microchromosomes

6.2.5.1 Subfamily assignment by CLUSTALW Alignments

To predict which sub-families the new CR1 sequences belonged to, alignments with the CLUSTALW program (Thompson *et al.*, 1994), were carried out. For this, 52 CR1 sequences of 150 bp or more with known subfamilies (Vandergon and Reitman, 1994) were aligned with 39 new CR1 repetitive sequences of a similar length, and unknown subfamilies from this study. As Table 6.3 shows, it was possible to predict subfamilies for 30 new CR1 sequences. Out of these, 18 were found to belong to subfamily F, which is also one of the most ancient and was found on both chromosome types. As this was the largest subfamily group in the 52 sequences of known subfamilies, this is not surprising. An alignment with a new CR1 sequence (T52/4T7) and subfamily C sequences was also assigned. Fewer members of Subfamily C are expected to be found as it is a product of a recent transposition (Vandergon and Reitman, 1994). Other alignments suggest predicted assignments to CR1 subfamilies B, D and E. Several sequences, referred to as unknown (UN) in table 6.3, did not show any similarity to any of the subfamily sequences or to each other and may represent new subfamilies.

Table 6.3 has evidence of clustering of CR1 elements in certain regions of the genome by the number of CR1s found on each cosmid. For example, four out of the six CR1s found on cosmid 34 belong to subfamily F. Possible clustering of subfamily B is observed on cosmid 08. Cosmid 30 showed the greatest number of different subfamilies (3 F, 2 E and 1 B).

These results do show, however, that all CR1 subfamilies appeared to be distributed across the macrochromosomes and microchromosomes with a possible bias of subfamilies towards chromosome type.

6.2.5.2 Subfamily Assignment by Phylogenetic Tree Studies

Phylogenetic trees were constructed to confirm subfamily predictions made by the CLUSTALW alignments and the results are presented in Table 6.4. Seven sequences (described as UN in table 6.3) did not align with subfamilies A-F and were therefore assigned to new subfamilies, G-N. Four of the new subfamilies were found on the microchromosomes and three on the macrochromosomes.

The eight new subfamilies may be ancient CR1 repeats which share characteristics with CR1-like elements. To find any similarities, alignments and tree analysis were carried out with subfamilies G-N and CR1-like elements from Frog, Lizard, Snake, Ray and Turtle. The results showed subfamily G was related to the lizard CR1-like repeats and subfamily I to the ray CR1-like element. The other subfamilies showed little similarities to these repeats.

Table 6.4 Assignment of Subfamilies to New CR1 Repetitive Sequences from Phylogenetic Tree Studies

Mic-Microchromosome; New subfamilies, G-N, were assigned to seven new CR1 sequences.

Clone Name	Physical assignment	Subfamily assignment from phylogenetic trees
T5/24T3	1	C
T2/82T3	2	G
T12/108T7	2p11-p12	L
035-CR1-01	2q26-q32	D
034-C16-T3	2q32-q35	F
034-C19-T7	2q32-q35	F
034-C67-T7	2q32-q35	F
034-CR1-01	2q32-q35	H
034-CR1-04	2q32-q35	E
034-CR1-05	2q32-q35	F
008-C17-T7	3q11	B
008-CR1-02	3q11	F
030-CR1-01	3q23-q33	F
030-CR1-02	3q23-q33	F
T2-70/ABT7	4	F
037-C18-T7	5	E
027-C06-T3	5q21-q22	K
014-C08-T7	Mic	F
014-C02-T3	Mic	M
014-C28-T7	Mic	N
036-C45-T7	Mic	I
007-CR1-01	Mic	J

6.3 Discussion

The results from Chapter 3 on the distribution of CR1 repeats was inconclusive as only four repeats were found and the sample was too small to be significant. A larger data set has allowed the distribution of these repeats to be investigated more fully and has found CR1 elements throughout the chicken genome, with more being found on the macrochromosomes than on the microchromosomes. Clustering of some CR1s was observed and this could be due to a lateral spread of the repeats.

6.3.1 Density of CR1 Repeats (Macrochromosomes vs. Microchromosomes)

The relative density of the CR1 repeats may not be significant. Therefore a larger data set is required. The 2:1 density (macrochromosomes vs. microchromosomes) of the CR1s is interesting as the macrochromosomes are twice as CR1 dense as the microchromosomes but with gene density, the opposite is observed (Chapter 5). From size alone, 70% of the CR1s would be expected to map to a macrochromosome but genetic linkage data shows that 78% of CR1s are located on macrochromosomes 1-5.

6.3.2 Number of CR1 Repeats in the Chicken Genome

Based on the sequencing data, the number of CR1 repeats was calculated as 172,000. This is close to the estimate of 100,000 CR1s per haploid genome made using nonredundant chicken GenBank DNA files longer of than 10,000 bp. Fewer CR1s (11) were used in these calculations, therefore this could be an underestimate of the true numbers of CR1s (Vandergon and Reitman, 1994).

The estimates of the number of CR1s using hybridisation data are therefore under representations of the true numbers of CR1s (Burch *et al.*, 1993; Stumph *et al.*, 1984). As there are likely to be a number of different CR1 progenitor elements, a single probe would not detect all CR1 subfamilies. In addition, older, more divergent CR1s have undergone mutations, becoming undetectable by hybridisation but still recognisable by sequence analysis.

6.3.3 Macrochromosomes, Microchromosomes, CR1 elements and Junk DNA

The distribution, density and number of CR1 elements all suggest that there are more of these repeats on the macrochromosomes than on the microchromosomes. These observations could be evidence of the CR1 elements being able to insert themselves into the macrochromosomes more easily because there is more room, in the form of junk DNA, for them to do so. The microchromosomes, being more gene dense would have less junk DNA and therefore less room for the CR1s. In addition, is there more selective pressure on the microchromosomes to reduce in size?

6.3.4 Distribution of CR1 Repeat Subfamilies

CR1 subfamilies arose at different times in evolution and their distribution could provide clues as to how and when the macrochromosomes and microchromosomes evolved. Out of the 39 CR1s isolated, the F subfamily was the most common on both chromosome types. The microchromosomes analysed have older CR1s (subfamilies D, E and F) whereas the macrochromosomes have a mix of newer subfamilies (BC) and older CR1s. Both chromosome types have the new subfamilies, G-N. If macrochromosomes are the result of the fusion of microchromosomes, then microchromosomes existed before macrochromosomes and would have older CR1 elements. Alternatively, if microchromosomes arose from the fission of macrochromosomes then they would have had the older CR1s. As macrochromosomes still had more room on them, new CR1s could transpose themselves. On the microchromosomes there would be pressure to lose non-essential

DNA. There would be less room for new CR1 repeats to transpose into. Mapping and analysing more CR1s would help elucidate which hypothesis is true.

It has been suggested that more than six CR1 subfamilies exist (Vandergon and Reitman, 1994). New subfamilies have been isolated during this study and there is the possibility that subfamilies G-N are older than subfamilies A-F. The G-N subfamilies will have undergone more deletions and lost the ability to transpose themselves. As they would have altered so much over time they would be difficult to recognise by hybridisation methods and would have only been isolated now through sequencing. If these are ancient CR1 repeats, when did they first appear? Subfamilies G and I share similarities with lizard and ray CR1-like elements, respectively. This suggests that these repeats existed before birds diverged from other vertebrates. After birds diverged from other vertebrates these repeats would have experiences a gain/loss of DNA.

6.4 Conclusion

There appears to be a two-fold difference in CR1 repeat density when comparing macrochromosomes with microchromosomes. This, along with the number and distribution of the repeat, go some way to support the hypothesis that if microchromosomes are gene dense or more compact (i.e. less junk DNA), then they will have fewer repeats. Leading on from this data the question of why genome size is constrained and what mechanisms control it can be asked.

Chapter 7

Comparative Mapping in the Chicken Genome

7.1 Introduction

Chickens are “known’ to have diverged from mammals approximately 300-350 Mya (Hedges, 1994); (Nanda *et al.*, 1999). Despite this lengthy period of divergence, conservation of genes (coding and non-coding regulatory sequences), conserved syntenic groups and conserved gene order have been identified between chicken and other vertebrate species.

7.1.1 Conservation of Genes and Synteny

Conservation of genes has been observed a number of times in this study while searching for gene homologies via database searches. For example the chicken *5-HT$_{1D}$* receptor gene was identified in Chicken, Human, Mouse and *Fugu* (see Chapter 5). A second example is the *TGF-β RII* gene, which is an example of the 3’ UTR of a gene that is highly conserved (see Chapter 3) (Duret *et al.*, 1993).

An example of conservation of synteny between chicken and other species shown in Chapter 5. The genes *PTAFR* and *5-HT$_{1D}$* were found on the same cosmid and are therefore syntenic. Chapter 1, Table 1.5 outlines other chicken genes *TCP1, IGF2R, VIP, ESR, MYB, PLN* and *FYN*, which all show conservation of synteny.

7.1.2 Conservation of Gene Order

Genetic segments can be conserved throughout evolution and can be used to identify candidates for disease genes and facilitate mapping in other species. (Eppig, 1996). Detailed maps and sequence information is available for human and mouse, making it easy to identify homologs between these and other, less detailed genome maps such as those for the chicken. This chapter discusses the order of three genes, tyrosine hydroxylase (*TH*), insulin (*INS*) and insulin-like growth factor-II (*IGF2*), and their conservation in mammals and chicken.

7.1.3 Tyrosine Hydroxylase and Phenylaline Hydroxylase

Tyrosine hydroxylase regulates catecholaminergic neuronal activity by acting as the rate limiting step in catecholamine synthesis (Tillet *et al.*, 1997). It is biochemically induced by environmental stress and drugs and is expressed in all catecholamine-synthesising neurons in the CNS of vertebrates (Xue *et al.*, 1988) and (Boularand *et al.*, 1998). *TH* has been cloned in chicken (Carrier *et al.*, 1993), quail (Fauquet *et al.*, 1988), human (Grima *et al.*, 1987), rat (Grima *et al.*, 1985), mouse (Brilliant *et al.*, 1987) and eel (Boularand *et al.*, 1998).

The disease Phenylketonuria (PKU) is the most common inborn error of amino acid metabolism among Caucasians, an incidence of 1 in 10,000 in the UK. Phenyaline is an amino acid that is essential for growth in infants and nitrogen equilibrium in adults (Start, 1998). PKU is caused by a phenylaline hydroxylase (PAH) deficiency which lowers the enzymatic conversion of dietary phenylaline to tyrosine, leading to hyperphenylalaninaemia (Guldberg *et al.*, 1998; Start, 1998; Tyfield, 1997). There have been more than 300 different mutations of the human PAH gene which is located on chromosome 12 band region q22-q24, and contains 13 exons spanning 90 kb of DNA (Tyfield, 1997).

TH and *PAH* belong to a family of aromatic amino acid hydroxylases which also includes tryptophan hydroxylase (TPH). All three share a degree of sequence homology, biochemical and immunological properties and similar functional characteristics (Ledley *et al.*, 1985). A comparative analysis of *TH*, *PAH* and *TPH* sequences, regulation mechanism and tissue distributions, suggest that the duplications of the common ancestor of these three genes occurred before the emergence of arthropods (Fauquet *et al.*, 1988) and [Boularand, 1998 #195

7.1.4 Insulin, Insulin-Like Growth Factor I and Insulin-Like Growth Factor II

The insulin related gene family is involved in growth, development and metabolism. It compromises of *INS*, insulin-like growth factor-I (IGF1), IGF2, relaxin and several invertebrate insulin-related peptides. (McRory and Sherwood, 1997). The growth factors *IGF1* and *IGF2* share approximately 70% amino acid identity and with proinsulin they share a 50% amino acid identity (Rotwein, 1991). Studies on their evolution has indicated that insulin and the insulin-like growth factors only became distinct molecules after vertebrates arose (McRory and Sherwood, 1997).

7.1.5 Insulin

Insulin is a polypeptide hormone that increases the rate of glycogen, fatty acids and protein synthesis and stimulates glycolysis. Secreted in the beta cells of the pancreas, it promotes the entry of glucose, some other sugars and amino acids into muscle and fat cells, which lowers blood glucose levels. Precursors of the active hormone are preproinsulin and proinsulin.

Genes which are homologous but differ in structure can give information about how they evolved. For example, mammals and birds have a single insulin gene whereas rodents and three fish species (tuna, bonito and toadfish) have two (Perier *et al.*, 1980) and (Davies *et al.*, 1994). Comparison of the organisation of this gene in chicken and rat has shown that the single chicken gene and one of the rat insulin genes (*Ins*-2) have a common structure of two introns. The second rat insulin gene has a single intron, suggesting that the ancestral insulin gene had two exons. This second rat gene may have evolved by gene duplication and then lost of one of the introns from one copy of the gene (Perier *et al.*, 1980) and (Soares *et al.*, 1985).

7.1.6 Insulin Like Growth Factor I

Insulin Like Growth Factor I plays a role maintaining and promoting postnatal growth and mediating most of the actions of growth hormone. Synthesised in the liver, as well as other tissues, it can also function as a locally regulated autocrine or paracrine growth stimulator (Kajimoto and Rotwein, 1991), (Kallincos *et al.*, 1990), (Taylor *et al.*, 1991) and (Upton *et al.*, 1995).

The organisation of the *IGF1* gene in mammals is complex. The single-copy gene is transcribed and processed into multiple mRNAs which encode at least two peptide precursors. The organisation of the chicken *IGF1* gene is simpler and more compact compared to its mammalian homologs. Comparative analyses of this gene in chicken and mammals has defined features of *IGF1* common to vertebrates. The chicken gene has four exons spread over 50 kb of chromosomal DNA and maps to chromosome 1p14-p13. These are transcribed and processed into two mRNAs of 1.9 and 2.6 kb in size (Kajimoto and Rotwein, 1991).

7.1.7 Insulin-Like Growth Factor II

Insulin-Like Growth Factor II is a polypeptide that is involved in tissue differentiation and the regulation of embryonic growth and development. It has structural similarity with insulin-like growth factor I and insulin. It is synthesised by a number of different cell types, with transcripts most abundant in foetal tissue and expression levels falling postnatally. Mammalian *IGF2* genes have complex transcription patterns with multiple tissue-specific promoters and polyadenylation sites that give rise to a range of different *IGF2* mRNAs (Boulle *et al.*, 1993), (Kallincos *et al.*, 1990), (Darling and Brickell, 1996) and (Upton *et al.*, 1995).

The chicken *IGF2* cDNA has been characterised and has three coding exons and is interrupted by introns at similar positions to those found in the human and rat genes (Darling and Brickell, 1996).

7.1.8 Imprinted Genes

In the mammalian genome, certain autosomal genes are inherited from one parent in a silent state and in an active form from the other parent. This is known as imprinting and it has a role in development. Loss of imprinting results in a number of cancers such as Wilms tumour, and diseases such as the Prader-Willi and Angelman syndromes. Imprinted genes tend to cluster together, an example of which is the gene clusters on mouse chromosome 7 and human chromosome 11p15.5. This ~1.5 megabase region contains the maternally expressed *$p57^{KIP2}$*, *KyLQT1*, and *Mash2* genes at one end, the paternally expressed *INS*/*Ins-2* and *IGF2* genes in the centre, with the maternally expressed *H19* genes approximately 90 kb downstream (Bartolomei and Tilghman, 1997) and (Lalande, 1997).

It is not known if IGF2 and INS are imprinted in chicken. There are no known examples of imprinting in birds but there is no evidence against it. Human and mouse species diverged some 70 million years ago, suggesting that the mechanism of imprinting arose before this time (Marshall Graves, 1998). If these genes are imprinted in the birds, which diverged from vertebrates some 300-350 Mya, this would suggest that imprinting may have evolved more than 70 Mya. As imprinted genes are found in clusters, the fact that these two genes are close together in chicken could suggest that some conservation of the imprinting mechanism may have occurred.

7.1.9 Conservation of TH-INS-IGF2 Gene Order

In humans, *TH*, *INS* and *IGF2* are contiguous and map to chromosome 11p15.5, forming the linkage group 5' *TH-INS-IGF2* 3'. They have the same transcriptional polarity with 2.7 kb separating the *TH* and *INS* genes and 1.4 kb separating the *INS* and *IGF2* genes (O'Malley and Rotwein, 1988). Similar conservation is observed in mouse with *Th*, *Ins 2* and *Igf2* (Brilliant *et al.*, 1987) and

(Rotwein, 1991). This conservation of *TH-INS-IGF2* could be by chance, or may have some selective advantage e.g. imprinting.

In the chicken these three genes have been mapped both genetically and physically to chromosome 5 and are therefore linked but their order and how close they are to each other is unknown.

7.1.10 Evolution of the TH-INS-IGF2 and PAH-IGF1 Paralogous Segments

The study of gene families gives the opportunity to examine the evolution of paralogous segments. A paralogous region, *PAH* and *IGF1* is found on human chromosome 12 (O'Malley and Rotwein, 1988). The genes *INS* and *IGF2* may be products of tandem duplication. The genes *TH* and *PAH* regions are paralogous and may have arisen by chromosome/genome duplication. To find out what has occurred in the evolution of these genes, divergent groups such as fish, birds and mammals, need to be examined.

In the telost fish species Barramundi, the genes *TH* and *IGF2* are adjacent, but *INS* is not. It has been proposed that the *INS* gene between *TH* and *IGF2* may have been silenced in bony fish, and an active *INS* gene may lie on a paralogous chromosome segment between the genes *PAH* and *IGF1* (Collet *et al.*, 1998).

How these genes may have evolved, as suggested by work carried out prior to this study, is described in Figure 7.1 (Collet *et al.*, 1998). This figure shows a vertebrate ancestor with two insulin genes organised as, *TH-INS-IGF2* and *PAH-INS-IGF1*. As species diverged, silencing of one of the *INS* genes has occurred. In the reptilian ancestor the insulin gene between *PAH* and *IGF1* was silenced. This was maintained through the divergence of reptile/bird and mammals. The reverse occurred in bony fish, with silencing of the insulin gene between *TH* and *IGF2*. In bony fish, insulin must map elsewhere and could potentially lie between *PAH* and *IGF1*.

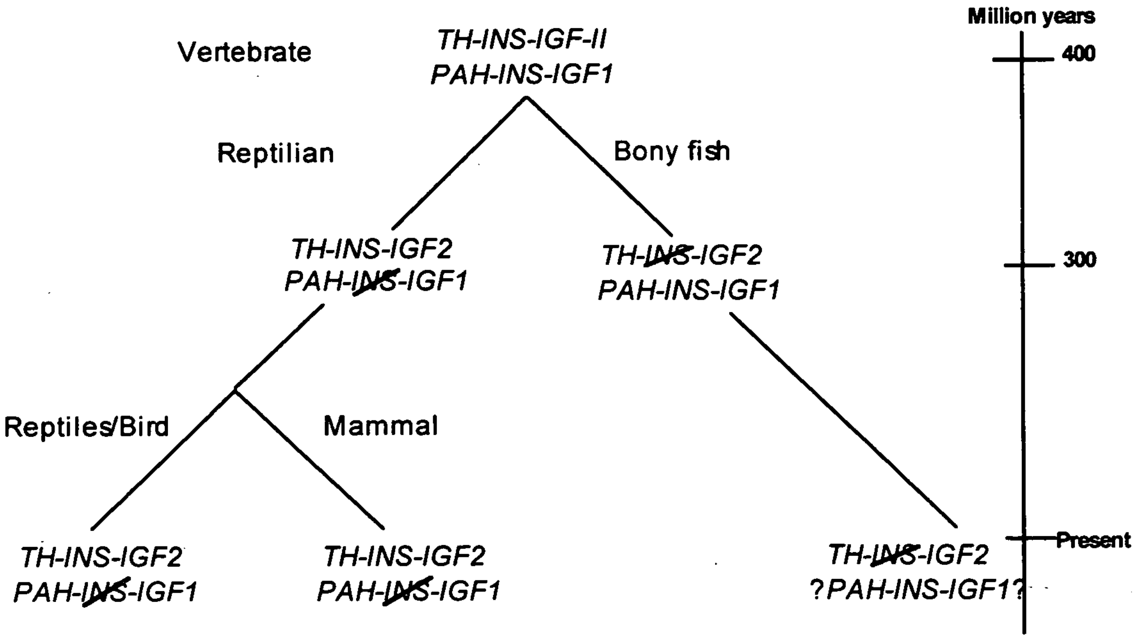


Figure 7.1 Model of the Evolution of the Genes Tyrosine Hydroxylase, Insulin and Insulin Like Growth Factor II

~~*INS*~~-silenced insulin gene; (Collet *et al.*, 1998); A vertebrate ancestor had two insulin genes and as species diverged one of the *INS* genes was silenced.

7.2 Results

7.2.1 Study of Conserved Linkage and Gene Order

To study the conservation of linkage and gene order between *TH*, *INS* and *IGF2*, cosmids containing the genes encoding tyrosine hydroxylase, insulin and insulin like growth factor II, were isolated and sequence sampled. Five cosmids, three *IGF2*, one *INS*, one *TH*, were isolated and digested with the restriction enzymes *Hin*dIII, *Bam*HI and *Pst*I (Figure 7.2) to find overlaps between the clones. Two of the three *IGF2* cosmids showed a great deal of overlap, as did the *TH* and *INS* cosmids. Based on the digest results, an *IGF2* clone (cosmid 40) and the INS cosmid (cosmid 41) were analysed. A total of 24,465 kb of unique DNA sequence was produced for the *IGF2* containing cosmid 40 and 27,596 kb of unique DNA sequence for *TH*/*INS* containing cosmid 41.

7.2.2 Gene Homologies

Database search results from cosmid 40 gave significant hits to the chicken *IGF2* gene, with matches to the three coding exons. Figure 7.3 and 7.4 show the BLASTN and BLASTX alignments with exon two of *IGF2*. Results from cosmid 41 database searches produced significant hits to both the chicken insulin gene (matches to exons 1 and 2). Figure 7.5 is the BLASTN alignment to exon one and figure 7.6 the BLASTX alignment to the chicken, turkey and ostrich genes. Alignments to the quail tyrosine hydroxylase gene are shown in Figures 7.7 (BLASTN) and 7.8 (BLASTX). A match to chicken tyrosine hydroxylase gene was not found as only the 5' end and the first exon of this gene has been characterised (Carrier *et al.*, 1993).

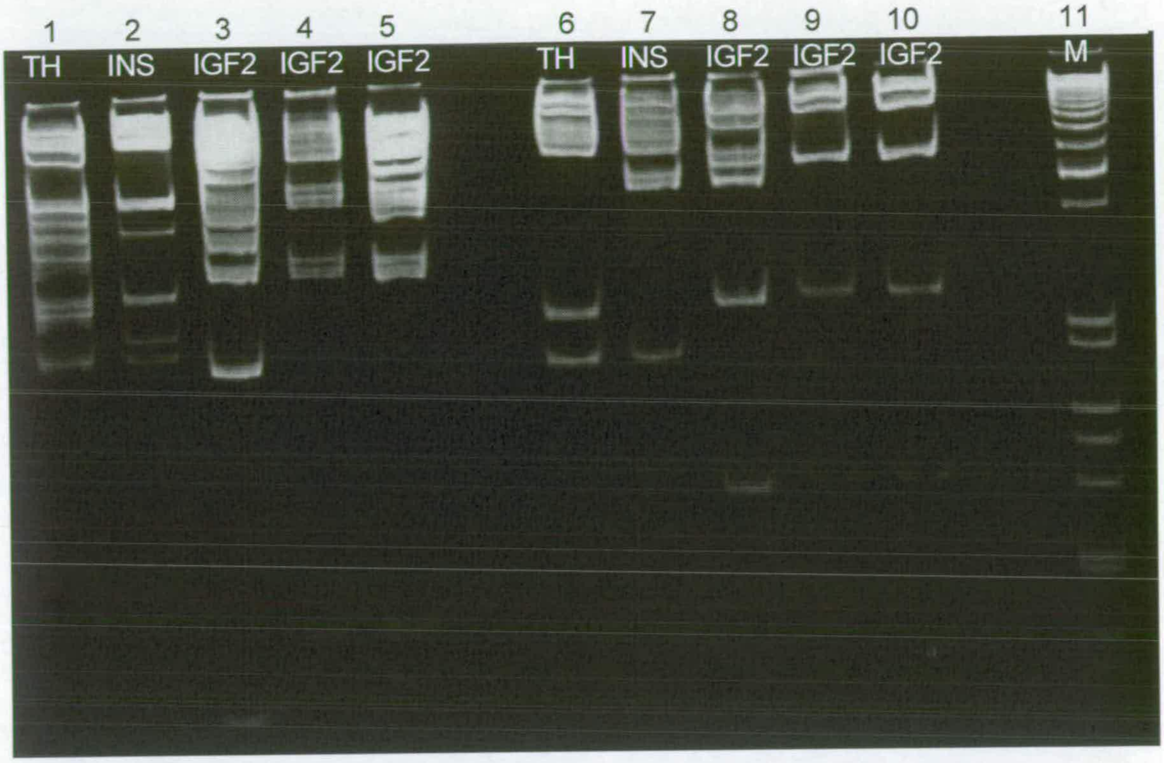


Figure 7.2 HindIII and BamHI Digests of INS, TH and IGF2 Cosmids

Lane: 1-5 HindIII digests, 6-10 BamHI, 11 Marker X (Boehringer Mannheim); Cosmids 040 (lanes 4 and 9) and 041 (lanes and 7) were analysed by sequence sampling.

```
                                                                 Smallest
                                                                   Sum
                                                           High  Probability
Sequences producing High-scoring Segment Pairs:           Score  P(N)      N

EM_OV:S82962 ! S82962 prepro-insulin-like growth factor-I...449  6.3e-56   2
EM_OV:ZFAJ3165 ! Aj223165 Zebra finch insulin-like growth...369  4.6e-39   2
EM_OM:SSIGF2 ! X56094 S.scrofa mRNA IGF2 for insulin-like...224  2.7e-08   1
EM_OM:CJIGF2PRT ! Aj001297 Callithrix jacchus mRNA fragme...215  7.9e-08   1
EM_HU1:HSGFI23 ! M14117 Human insulin-like growth factor ...210  1.5e-07   1
EM_HU1:HSIGFII4 ! X03426 Human IGF-II gene exon 4 for ins...210  1.8e-07   1
EM_RO:MMGFII ! M14951 Mouse insulin-like growth factor II...210  4.1e-07   1

EM_OV:S82962 S82962 prepro-insulin-like growth factor-II [chickens,
Genomic,1513 nt, segment 2 of 2]. 2/97
Length = 1513

Minus Strand HSPs:

Score = 449 (124.1 bits), Expect = 6.3e-56, Sum P(2) = 6.3e-56
Identities = 91/93 (97%), Positives = 91/93 (97%), Strand = Minus / Plus

Query: 176 GGTAGACCAGTGGGACGAAATAACAGGAGGATCAACCGTGGCATTGTGGAGGAGTGCTGC 117
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   8 GGTAGACCAGTGGGACGAAATAACAGGAGGATCAACCGTGGCATTGTGGAGGAGTGCTGC 67

Query: 116 TTTCGGAGCTGTGACCTGGCTCTGCTGGAAACC 84
            |||||||||||||||||||||||||||||||||
Sbjct:  68 TTTCGGAGCTGTGACCTGGCTCTGCTGGAAACC 100

 Score = 369 (102.0 bits), Expect = 6.3e-56, Sum P(2) = 6.3e-56
 Identities = 75/77 (97%), Positives = 75/77 (97%), Strand = Minus / Plus

Query:  77 GTGCCAAGTCCGTCAAGTCAGAGCGTGACCTCTCCGCCACCTCCCTCGCGGGCCTCCCAG 18
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 105 GTGCCAAGTCCGTCAAGTCAGAGCGTGACCTCTCCGCCACCTCCCTCGCGGGCCTCCCAG 164

Query:   17 CCCTCAACAAGGTAGGG 1
             |||||||||||||||||
Sbjct:  165 CCCTCAACAAGGTAGGG 181
```

Figure 7.3 Cosmid 40 BLASTN Results: The Chicken *IGF2* Gene, Exon 2

```
                                                                   Smallest
                                                                     Sum
                                                  Reading High  Probability
Sequences producing High-scoring Segment Pairs:     Frame Score P(N)      N

TREMBL:P79890 ! P79890 PREPRO-INSULIN-LIKE GROWTH FA..-1 164    8.0e-16  1
TREMBL:057687 ! 057687 IGF-II PRECURSOR. 6/98         -1 160    3.0e-15  1
SWISSPROT:IGF2_MUSVI ! P41694 mustela vison (america..-1 139    1.0e-12  1
SWISSPROT:IGF2_HORSE ! P51459 equus caballus (horse)..-1 118    3.4e-10  1
TREMBL:P78449 ! P78449 INSULIN-LIKE GROWTH FACTOR II..-1 118    3.0e-09  1
TREMBL:Q91443 ! Q91443 INSULIN-LIKE GROWTH FACTOR II..-1 120    6.0e-09  1
SWISSNEW:IGF2_HUMAN ! P01344 INSULIN-LIKE GROWTH FAC..-1 118    1.5e-08  1
SWISSPROT:IGF2_HUMAN ! P01344 homo sapiens (human). ..-1 118    1.5e-08  1
SWISSNEW:IGF2_HORSE ! P51459 INSULIN-LIKE GROWTH FAC..-1 118    1.5e-08  1
SWISSPROT:IGF2_PIG ! P23695 sus scrofa (pig). insuli..-1 118    1.5e-08  1
TREMBL:Q14299 ! Q14299 PREPROINSULIN-LIKE GROWTH FAC..-1 118    1.5e-08  1
TREMBL:042429 ! 042429 INSULIN-LIKE GROWTH FACTOR II..-1 110    5.5e-07  1
SWISSPROT:IGF2_CAVPO ! Q08279 cavia porcellus (guine..-1 105    2.1e-06  1
TREMBL:Q63265 ! Q63265 RAT INSULIN-LIKE GROWTH FACTO..-1 105    2.8e-06  1
SWISSPROT:IGF2_BOVIN ! P07456 bos taurus (bovine). i..-1 105    2.8e-06  1
SWISSPROT:IGF2_SHEEP ! P10764 ovis aries (sheep). in..-1 105    3.3e-06  1
SWISSPROT:IGF2_MOUSE ! P09535 mus musculus (mouse). ..-1 105    3.3e-06  1
SWISSPROT:IGF2_RAT ! P01346 rattus norvegicus (rat)...-1 105    3.3e-06  1
TREMBL_NEW:G2769668 ! G2769668 INSULIN-LIKE GROWTH F..-1 105    3.3e-06  1
TREMBL_NEW:G3158363 ! G3158363 INSULIN-LIKE GROWTH F..-1 105    3.8e-06  1


TREMBL:P79890 P79890 PREPRO-INSULIN-LIKE GROWTH FACTOR-II. 6/98
Length = 187

Minus Strand HSPs:

Score = 164 (75.4 bits), Expect = 8.0e-16, P = 8.0e-16
Identities = 32/43 (74%), Positives = 33/43 (76%), Frame = -1

Query:   173 RPVGRNNRRINRGIVEECCFRSCDLALLETYCAKSVKSERDLS 45
             RPVGRNNRRINRGIVEECCFRSCDLALLETYCAKSVKSERDLS
Sbjct:    53 RPVGRNNRRINRGIVEECCFRSCDLALLETYCAKSVKSERDLS 95

Score = 104 (47.8 bits), Expect = 5.9e-06, P = 5.9e-06
Identities = 22/23 (95%), Positives = 23/23 (100%), Frame = -3

Query:    75 AKSVKSERDLSATSLAGLPALNK 7
             AKSVKSERDLSATSLAGLPALNK
Sbjct:    85 AKSVKSERDLSATSLAGLPALNK 107
```

Figure 7.4 Cosmid 040 BLASTX Results: Exon 2 of the Chicken *IGF2* Gene

```
                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

EM_OV:GGINS1 ! V00416 Part of the chicken insulin gene (e..2232  3.3e-180   1
EM_OV:GGJINS1 ! J00872 Chicken preproinsulin gene, from 5..2232  3.3e-180   1
EM_OV:GGINSMRNA ! X58993 G.gallus mRNA for preproinsulin...1025  9.7e-80    1
EM_OV:S66611 ! S66611 preproinsulin [Selaphorus rufus=hum...722  1.9e-52    1
EM_RO:MMINSIIG ! X04724 Mouse preproinsulin gene II. 4/93   531  7.4e-34    1
EM_RO:RNINS21 ! M25583 Rat insulin 2 gene, exons 1 (parti...521  1.9e-33    1
EM_RO:RNINS2 ! V01243 Rat gene for insulin 2. 4/93          521  4.5e-33    1
EM_RO:RNINSII ! J00748 Rat insulin II gene (ins-2) with t...521  5.3e-33    1
EM_OM:CLINSU ! V00179 Dog gene encoding insulin. 3/91       508  5.8e-32    1
EM_OM:OANIGFII1 ! U00659 Ovis aries insulin gene, complet...499  3.8e-31    1
EM_RO:MMINSIG ! X04725 Mouse preproinsulin gene I. 9/93     483  7.8e-30    1
EM_OM:CEPPINS ! X61092 C.aethiops gene for preproinsulin....480  1.5e-29    1

EM_OV:GGINS1 V00416 Part of the chicken insulin gene (exon 1). 10/96
Length = 497

Plus Strand HSPs:

Score = 2232 (616.7 bits), Expect = 3.3e-180, P = 3.3e-180
Identities = 448/450 (99%), Positives = 448/450 (99%), Strand = Plus / Plus

Query:    1 CTGATGAATAAAATATTCCTTTCCTCTTCAGAAGGTCCATTTGCTTCTGTAGTCTTGTTT 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   41 CTGATGAATAAAATATTCCTTTCCTCTTCAGAAGGTCCATTTGCTTCTGTAGTCTTGTTT 100

Query:   61 TCACGTCAAAGGAGCTGAGGGACATAAGATGCCTGATGATAGCTTATTCCTCCCTTGCAA 120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  101 TCACGTCAAAGGAGCTGAGGGACATAAGATGCCTGATGATAGCTTATTCCTCCCTTGCAA 160

Query:  121 CCCCCCCGTGTCTCCTTTGCTTCCTACCTCTAGGCCTCCCCCAGCTCATCATGGCTCTCT 180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  161 CCCCCCCGTGTCTCCTTTGCTTCCTACCTCTAGGCCTCCCCCAGCTCATCATGGCTCTCT 220

Query:  181 GGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATG 240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  221 GGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATG 280

Query:  241 CAGCTGCCAACCAGCACCTCTGTGGCTCCCACTTGGTGGAGGCTCTCTACCTGGTGTGTG 300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  281 CAGCTGCCAACCAGCACCTCTGTGGCTCCCACTTGGTGGAGGCTCTCTACCTGGTGTGTG 340

Query:  301 GAGAGCGTGGCTTCTTCTACTCCCCCAAAGCCCGACGGATGTCGAGCAGCCCCTAGGTA 360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  341 GAGAGCGTGGCTTCTTCTACTCCCCCAAAGCCCGACGGATGTCGAGCAGCCCCTAGGTA 400

Query:  361 AGTCAGTTTGACCATGACTACATTCATATGCTATATGATGCAAAAAGCAACTGTCTATCT 420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  401 AGTCAGTTTGACCATGACTACATTCATATGCTATATGATGCAAAAAGCAACTGTCTATCT 460

Query:  421 TTGATGGTGACACAAGGAATGTCCTTGGTG 450
            ||||||||||||||||||||||||||||||
Sbjct:  461 TTGATGGTGACACAAGGAATGTCCTTGGTG 490
```

Figure 7.5 Cosmid 41 BLASTN: Chicken Insulin Gene, Exon 1

```
                                                                    Smallest
                                                                      Sum
                                                     Reading  High  Probability
Sequences producing High-scoring Segment Pairs:        Frame Score  P(N)      N
..
SWISSPROT:INS_CHICK ! P01332 gallus gallus (chicken)..+3   327  7.0e-40   1
SWISSPROT:INS_SELRF ! P51463 selasphorus rufus (humm..+3   270  5.7e-32   1
SWISSPROT:INS_SHEEP ! P01318 ovis aries (sheep). ins..+3   241  5.7e-28   1
SWISSPROT:INS_MACFA ! P30406 macaca fascicularis (cr..+3   240  7.5e-28   1
SWISSPROT:INS_CANFA ! P01321 canis familiaris (dog)...+3   237  2.0e-27   1
SWISSPROT:INS_CERAE ! P30407 cercopithecus aethiops ..+3   237  2.0e-27   1
SWISSNEW:INS_HUMAN ! P01308 INSULIN PRECURSOR. 11/98   +3   234  5.1e-27   1
SWISSPROT:INS_HUMAN ! P01308 homo sapiens (human). i..+3   234  5.1e-27   1
SWISSPROT:INS_CRILO ! P01313 cricetulus longicaudatu..+3   233  7.0e-27   1
SWISSPROT:INS_PSAOB ! Q62587 psammomys obesus. insul..+3   233  7.0e-27   1
SWISSPROT:INS2_RAT ! P01323 rattus norvegicus (rat)...+3   232  9.7e-27   1
SWISSPROT:INS1_RAT ! P01322 rattus norvegicus (rat)...+3   231  1.3e-26   1
SWISSPROT:INS_PANTR ! P30410 pan troglodytes (chimpa..+3   230  1.8e-26   1
SWISSPROT:INS_BOVIN ! P01317 bos taurus (bovine). in..+3   230  1.9e-26   1
SWISSPROT:INS_AOTTR ! P10604 aotus trivirgatus (nigh..+3   222  2.4e-25   1
SWISSPROT:INS2_MOUSE ! P01326 mus musculus (mouse). ..+3   217  1.2e-24   1
SWISSPROT:INS_RODSP ! P21563 rodentia sp. insulin pr..+3   125  3.1e-24   2
SWISSPROT:INS1_MOUSE ! P01325 mus musculus (mouse). ..+3   211  8.0e-24   1
SWISSPROT:INS_RABIT ! P01311 oryctolagus cuniculus (..+3   203  1.0e-22   1
SWISSNEW:INS_LOPPI ! P01341 INSULIN PRECURSOR. 11/98   +3   196  8.9e-22   1


SWISSPROT:INS_CHICK P01332 gallus gallus (chicken), meleagris gallopavo
(common turkey), and struthio camelus (ostrich). insulin precursor. 10/96
Length = 107

Plus Strand HSPs:

Score = 327 (150.4 bits), Expect = 7.0e-40, P = 7.0e-40
Identities = 63/65 (96%), Positives = 63/65 (96%), Frame = +3

Query:  171 MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ 350
            MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
Sbjct:    1 MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ 60

Query:  351 PLVSS 365
            PLVSS
Sbjct:   61 PLVSS 65
```

Figure 7.6 Cosmid 41 BLASTX Results: Chicken, Turkey and Ostrich Insulin Gene

```
                                                                Smallest
                                                                  Sum
                                                         High  Probability
Sequences producing High-scoring Segment Pairs:          Score  P(N)      N
..
EM_OV:CCTYRHA ! M24778 Quail tyrosine hydroxylase mRNA, c..612  3.0e-42   1
EM_RO:PSTYRHYDR ! Y09294 P.sungorus mRNA for tyrosine hyd..439  4.5e-27   1
EM_RO:MMTHRA ! M69200 Mouse tyrosine hydroxylase, complet..431  8.1e-26   1
EM_RO:RNTOHA ! M10244 Rat tyrosine hydroxylase mRNA, comp..431  8.1e-26   1
EM_RO:RNTOHAB ! L22651 Rat tyrosine hydroxylase (TH) mRNA..431  8.3e-26   1
EM_OM:BTTHA ! M36794 Bovine tyrosine hydroxylase mRNA, co..430  9.8e-26   1
EM_OM:BTTYRHAA ! M36705 Bovine tyrosine hydroxylase mRNA,..430  9.8e-26   1
EM_HU1:HSINSTHIG ! L15440 Homo sapiens tyrosine hydroxyla..429  1.4e-25   1
EM_HU1:HSHTH1R ! X05290 Human mRNA for tyrosine hydroxyla..416  1.6e-24   1
EM_HU2:HSTHR ! Y00414 Human mRNA for tyrosine hydroxylase..416  1.6e-24   1
EM_HU2:HSTHX ! M17589 Human tyrosine hydroxylase type 4 m..416  1.6e-24   1
EM_OV:AF033802 ! Af033802 Tilapia mossambica insulin-like..389  3.4e-22   1
EM_OV:LCAF7942 ! Af007942 Lates calcarifer tyrosine hydro..371  1.1e-20   1
EM_OV:AAAJ731 ! Aj000731 Anguilla anguilla mRNA for tyros..344  2.2e-18   1
EM_IN:DMRNATH ! X76209 D.melanogaster mRNA for tyrosine h..271  3.2e-12   1

EM_OV:CCTYRHA M24778 Quail tyrosine hydroxylase mRNA, complete cds. 8/90
Length = 2077

Minus Strand HSPs:

Score = 612 (169.1 bits), Expect = 3.0e-42, P = 3.0e-42
Identities = 132/144 (91%), Positives = 132/144 (91%), Strand = Minus / Plus

Query: 172 CCTTGTCAGATGAACCAGAGGTACGAGACTTTGATCCTGATGCTGCTGCCGTTCAGCCCT 113
           ||||||||||||||||||||||||| |||||||||||||| |||||||||||||||||||
Sbjct:1205 CCTTGTCAGATGAACCAGAGGTACGGGACTTTGATCCTGACGCTGCTGCCGTTCAGCCCT1264

Query: 112 ACCAGGACCAGAACTACCAGCCTGTGTATTTTGTGTCTGAGAGCTTCAGTGATGCCAAAA 53
            ||||||||||  |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:1265 GCCAGGACCAGCCCTACCAGCCTGTGTATTTTGTGTCTGAGAGCTTCAGTGATGCCAAAA1324

Query:  52 ACAAGCTGAGGTAGGACTGGGCAC 29
           ||||||||||| |  |    ||||
Sbjct:1325 ACAAGCTGAGGAACTATGCAGCAC 1348
```

Figure 7.7 Cosmid 41 BLASTN Results: The Quail Tyrosine Hydroxylase Gene

```
                                                                 Smallest
                                                                   Sum
                                                       Reading High  Probability
Sequences producing High-scoring Segment Pairs:          Frame Score P(N)      N

TREMBL:O42428 ! O42428 TYROSINE HYDROXYLASE (FRAGMEN..-1 198   4.1e-22   1
TREMBL:P70468 ! P70468 TYROSINE HYDROXYLASE (FRAGMEN..-1 196   9.7e-22   1
TREMBL:P97517 ! P97517 TYROSINE HYDROXYLASE (FRAGMEN..-1 196   9.7e-22   1
SWISSPROT:TY3H_PHASP ! P11982 phasianidae sp. (quail..-1 209   1.0e-21   1
SWISSPROT:TY3H_BOVIN ! P17289 bos taurus (bovine). t..-1 202   9.8e-21   1
SWISSPROT:TY3H_MOUSE ! P24529 mus musculus (mouse). ..-1 199   2.6e-20   1
SWISSPROT:TY3H_RAT ! P04177 rattus norvegicus (rat)...-1 196   6.8e-20   1
SWISSPROT:TY3H_HUMAN ! P07101 homo sapiens (human). ..-1 190   4.8e-19   1
SWISSPROT:TY3H_ANGAN ! O42091 anguilla anguilla (eur..-1 187   1.2e-18   1
SWISSPROT:TY3H_DROME ! P18459 drosophila melanogaste..-1 150   1.8e-13   1
TREMBL:Q24000 ! Q24000 TYROSINE HYDROXYLASE TYPE 2, ..-1 150   1.8e-13   1
SWISSPROT:TY3H_CAEEL ! P90986 caenorhabditis elegans..-1 134   2.8e-11   1
SWISSPROT:PH4H_CAEEL ! P90925 caenorhabditis elegans..-1 127   1.0e-09   1
TREMBL:O17498 ! O17498 PHENYLALANINE HYDROXYLASE (EC..-1 126   1.7e-09   1
TREMBL:O46110 ! O46110 PHENYLALANINE HYDROXYLASE (FR..-1 107   1.1e-06   1
TREMBL:Q27600 ! Q27600 PHENYLALANINE HYDROXYLASE (EC..-1 107   3.4e-06   1
TREMBL:Q27599 ! Q27599 PHENYLALANINE HYDROXYLASE (EC..-1 107   3.4e-06   1
SWISSPROT:PH4H_DROME ! P17276 drosophila melanogaste..-1 107   3.4e-06   1
SWISSPROT:PH4H_RAT ! P04176 rattus norvegicus (rat)...-1 107   3.4e-06   1
SWISSPROT:TR5H_RAT ! P09810 rattus norvegicus (rat)...-1 103   1.4e-05   1


TREMBL:O42428 O42428 TYROSINE HYDROXYLASE (FRAGMENT). 6/98
Length = 129

Minus Strand HSPs:

Score = 198 (91.1 bits), Expect = 4.1e-22, P = 4.1e-22
Identities = 37/43 (86%), Positives = 39/43 (90%), Frame = -1

Query:   170 LSDEPEVRDFDPDAAAVQPYQDQNYQPVYFVSESFSDAKNKLR 42
             LSDEPE R+FDP+AAAVQPYQDQ YQPVYFVSESFSDAK K R
Sbjct:    35 LSDEPETREFDPEAAAVQPYQDQTYQPVYFVSESFSDAKEKFR 77
```

Figure 7.8 Cosmid 41 BLASTX Results: Tyrosine Hydroxylase Gene Fragment

7.2.5 Conservation of Linkage and Gene Order

To determine the order of the three genes, contig assembly for both cosmids was carried out. Figure 7.9 describes the two contigs generated for cosmids 40 and 41 sequences and outlines which parts of the genes were isolated. Linkage of *TH* and *INS* could be established as they were found on the same cosmid. This is an example of conserved linkage as they are also linked in man and rodents. As the two genes were found on the same cosmid, they must be close together but where *IGF2* lies in relation to *INS*/*TH* is unknown. The Seqman contigs were merged to resolve this but no overlap between the contigs was found. The region expected to link the two contigs together, assuming there is conservation of avian and mammalian regions, would contain the final *INS* exon and the estimated 1.4 kb (found between the human genes) separating *INS* from *IGF2*. The order of the three genes cannot be determined as these data cannot tell us where *IGF2* is in relation to *TH* and *INS*.

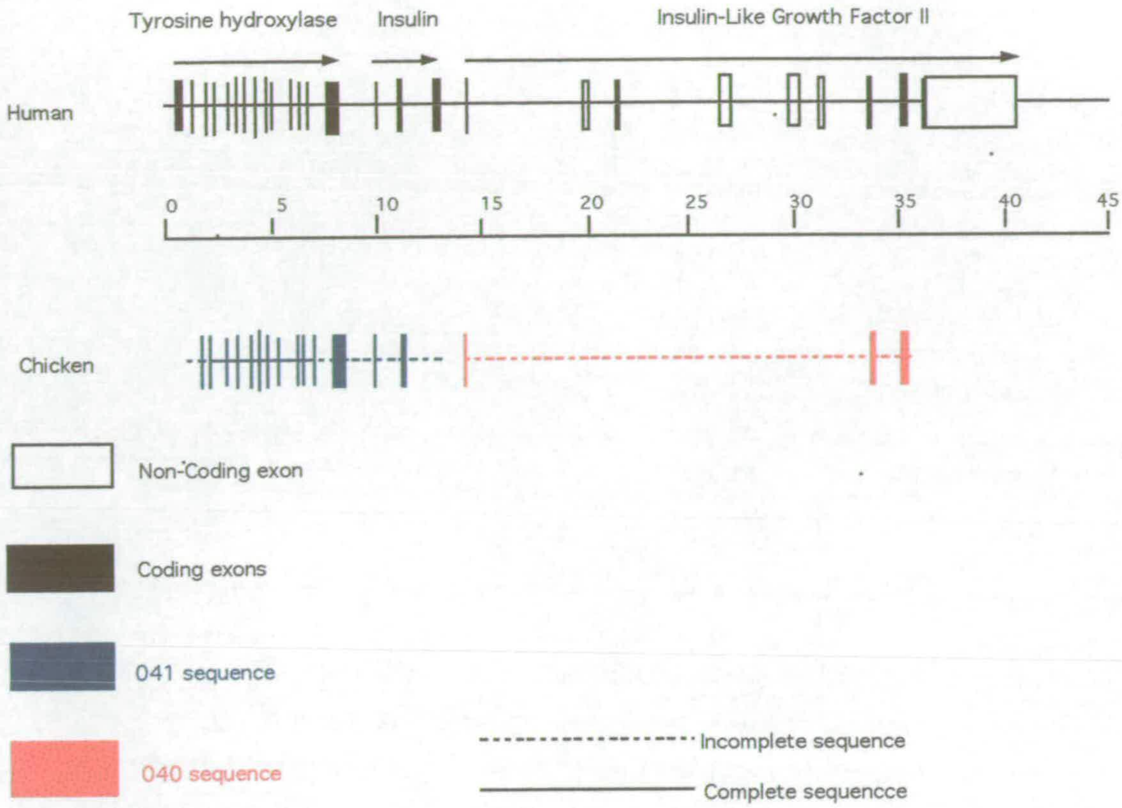


Figure 7.9 Sequence Sampling Results

7.3 Discussion

This work has established the linkage of the genes *TH* and *INS* by sequence sampling. The probable orientation of the two genes is:

$$\mathbf{5'\ \overrightarrow{\text{TH}} - \overrightarrow{\text{INS}}\ 3'}$$

The above orientation has been observed in both humans and in mouse (Brilliant *et al.*, 1987), (O'Malley and Rotwein, 1988) and (Rotwein, 1991) and is likely to be conserved in chicken. Where *IGF2* is in relation to *TH* and *INS* could not be determined by sequence sampling alone but it is expected that *IGF2* will be upstream of insulin. Therefore, the expected order of the three genes is expected to be:

$$\mathbf{5'\ \overrightarrow{\text{TH}} - \overrightarrow{\text{INS}} - \overrightarrow{\text{IGF2}}\ 3'}$$

7.4 Conclusion

Sequence sampling has proven to be an effective means of gene discovery, as shown by the results presented in chapters 3 and 5. As a method of establishing the order of genes, it is less effective. Further work to establish the order of these genes will involve aligning the *TH*, *INS* and *IGF2* cosmids along a BAC clone spanning the 5q11-q12 region. Alternatively, a fibreFISH approach could be used (Mann *et al.*, 1997a; Mann *et al.*, 1997b). In addition, as there is no evidence of imprinting in birds a future study examining the expression of the maternal and paternal *INS/IGF2* alleles can be carried out.

Chapter 8

General Discussion

8.1 Introduction

This study has highlighted some interesting features of the chicken genome:

(i) Differences in gene density between chromosome types (chapters 3 and 5), with microchromosomes twice as gene dense as macrochromosomes.

(ii) The number and distribution of CR1 elements between macrochromosomes and microchromosomes (chapter 6). The density of CR1 repeats is may be two times greater on the macrochromosomes than on the microchromosomes.

(iii) The conservation of genome organisation between chicken and other vertebrates (chapter 7).

In this chapter, these data are discussed with reference theories as to how the chicken genome has evolved.

8.2 Gene Density Differences Between Macrochromosomes and Microchromosomes

The genome sizes of birds, as represented by nuclear DNA content, varies little (range of 2.0-3.8 pg) compared to other vertebrates (range from 1-280 pg), suggesting some kind of restraint on genome size (Tiersch and Wachtel, 1991). Selective pressure placed on the genome to maintain a small size in birds may have been an adaptation for flight. Genome and cell size are correlated in vertebrates and avian cells are generally smaller than those of mammals. The smaller cell size allows for a greater rate of gas exchange necessary for flight (Hughes and Hughes, 1995). This work also suggests that the avian genome was derived from a larger genome, with reduction in size mostly occurring in the non-coding regions. Such compaction of the avian genome would have occurred gradually, over a long period of time (Hughes and Hughes, 1995). In this thesis I have estimated the number of genes in the chicken genome to be 59,000, which is similar to numbers estimated for mammals (see chapter 5). Therefore this trend in the reduction of genome size in birds was associated with an increase in gene density. I have also shown that there is difference in gene density within the genome, with a two-fold difference between macrochromosomes and microchromosomes. Is this because of differences in gene number or a more rapid DNA loss on the microchromosomes? The observed trend of shorter CR1 repeats on the microchromosomes (chapter 6) suggests a more rapid DNA loss. This difference was not statistically significant but the discovery of more CR1 elements may change this. For any comparison of CR1 repeat lengths to have a statistical significance of 95%, 100 new CR1 sequences are required.

What selective pressures is the chicken genome under? G-band DNA, which has been observed in birds and mammals, may be non-essential and hence not under such strong selection pressure (Craig and Bickmore, 1994). As the macrochromosomes have more AT-rich, gene poor, G-band DNA this could be evidence of the macrochromosomes being under less selective pressure. This could result in a higher number of repeats and more non-essential DNA. The microchromosomes, however, tend to be GC rich, gene rich and have more R-band DNA and therefore may be under greater selective pressure to maintain gene numbers and carry less non-essential DNA.

In chapter 6, I examined the idea that restraints on genome size could be reflected by the distribution of mobile CR1 repeats. The distribution of CR1 repeats is likely to be restricted to insertions within non-essential regions. This predicts that the density of CR1 repeats will be higher in less gene dense regions. Mapping studies do show a two fold difference in CR1 density between the gene poor macrochromosomes and gene rich microchromosomes. If the ancestral genome was less gene dense then we would predict that older CR1 repeats would be more widely dispersed. This is supported by the CR1 studies in chapter 6 where older CR1 subfamilies such as F are common, being widely distributed across both macrochromosomes and microchromosomes. In contrast, younger CR1 subfamilies such as B and C, are mostly found on the gene poor macrochromosomes. In the future more younger CR1 repeats need to be mapped for this difference to be statistically significant. One approach would be to isolate more CR1 repeats by sequencing and mapping specific BACs.

These observations ask the questions of how these differences in gene density evolved and how it is maintained? In mammals the gene content may be associated with DNA replication and recombination (Craig and Bickmore, 1994). Can a similar mechanism be true in chicken?

8.3 Correlation of Gene Density, Replication and Recombination

The gene dense microchromosomes replicate early during the first half of S phase (McQueen *et al.*, 1998). In mammals the high number of genes in the R and T bands may be linked to replication timing. Early replication may direct DNA into open transcriptionally competent chromatin, which also replicates early. Recombination may tend to initiate in these accessible regions (Craig and Bickmore, 1994). This may explain why there is a higher rate of recombination in the chicken gene dense microchromosomes (Rodionov *et al.*, 1992). Other factors may also be involved in maintaining this difference in recombination rate. In higher eukaryotes, chromosomal GC content is related to recombination frequency (Rodionov, 1996), and the high GC content in chicken microchromosomal DNA may ensure higher recombination (Rodionov, 1996). A higher than expected number of microsatellites has been found on the microchromosomes (chapter 5). The expected distribution of these markers is 70% on the macrochromosomes and 30% on the microchromosomes, but the observed distribution is 56% on the macrochromosomes and 44% on the microchromosomes. Microsatellites may be involved in recombination rate differences as they have been found in recombination hot spots (Brahmachari *et al.*, 1995). Due to the large number of small microchromosomes, to maintain at least one crossing over event on each, requires a higher recombination rate than found on the larger macrochromosomes (Rodionov *et al.*, 1992) and (Rodionov, 1996).

8.4 Evolution of the Avian Genome: Why Have Macrochromosomes and Microchromosomes?

Why are macrochromosomes and microchromosomes a unique and stable feature of birds? Prior to 350 million years ago, the ancestral vertebrate may have had 24 chromosomes (Morizot, 1994). Some 200 million years ago the common ancestor of birds/reptiles diverged and macrochromosomes and microchromosomes appeared in the bird lineage (Rodionov, 1997). Why did this happen in birds and not in reptiles? Why or how this event occurred is difficult to explain. However, clues to this may come from a discussion on the consequences of having macrochromosomes and microchromosomes. Was it by chance or did this have some adaptive value?

As bird and reptile lineage's diverged, the reptilian lineage may have undergone a series of fission's/fusion's which did not drastically change the number of chromosomes. In contrast as the bird lineage diverged, ancestral chromosomes underwent a number of fission events, increasing the number of chromosomes from 24 to approximately 40. Therefore, in birds the number of fission's out numbered the fusion events. This would imply that once a fission had occurred, a fusion was less likely to occur in birds than in mammals. Also, birds appear to have fewer chromosomal rearrangements, suggesting that microchromosomes may have ancestral gene arrangements (Fillon, 1998).

What are the reason for more fission events occurring in the chicken genome? One suggestion is that when fission occurred in mammalian chromosomes, the fragments, were unstable as they lacked telomeres and a centromere. This would favour the fusion of chromosomes to create a more stable structure. In contrast, chromosome fragments resulting from fission in the chicken have had telomeres and a centromere and were therefore more stable. Alternatively, as the chicken chromosomes have fewer repeats through DNA loss there would be less homologous sequence between the chromosomes. Chromosome rearrangements such as translocations would be less likely.

In birds, the increase in chromosome number would lead to instability in mitosis and meiosis with the need for more chromosome pairing (Rodionov, *et al.*, 1992) and (Rodionov, 1996), More crossing over was required and therefore the recombination rate was increased. This may have occurred by an increase in the gene density (see section 8.3) by DNA loss. Alternatively, the recombination rate could be maintained by chromosomal fusion. In birds it is the recombination rate that has been increased rather than the rate of chromosomal fusion's.

This process of DNA loss and chromosome fusion would have occurred until a state of equilibrium of 8 macrochromosomes and 32 microchromosomes (average number of avian chromosomes) was reached. This asks the question of which came first, flight or DNA loss? One hypothesis is that DNA loss occurred first, which led to smaller cells which were more energy efficient and enabling flight to be possible.

Bibliography

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 4013-4410.

Andersson, L.A., Ashburner, M., Audun, S., Barendse, W., Bitgood, J., Bottema, C., Broad, T., Brown, S., Burt, D.W., Charlier, C., Copeland, N., Davis, S., Davisson, M., Edwards, J., Eggene, A., Elegar, G., Eppig, J.T., Franklin, I., Grewe, P., Gill, T., Graves, J.A.M., Hawken, R., Hetzel, J., Hillyard, A., Jacob, H., Jaswinska, L., Jenkins, N., Kunz, H., Leven, G., Lie, O., Lyons, L., Maccarone, P., Mellersh, C., Montgomery, G., Moore, S., Moran, C., Morizot, D., Neff, M., Nicholas, F., O'Brien, S., Parsons, Y., Peters, J., Postlethwait, J., Raymond, M., Rothschild, M., Schook, L., Sugimoto, Y., Szpirer, C., Tate, M., Taylor, J., Vanderberg, J., Wakefield, J., Weinberg, J. and Womack, J. (1996) Comparative Genome Organization of Vertebrates. *Mammalian Genome*, **7**, 717-734.

Antequera, F. and Bird, A. (1993) Number of CpG Islands and Genes in Human and Mouse. *Proc. Natl. Acad Sci. USA*, **90**, 11995-11999.

Archibald, A.L. (1998) Comparative Genome Mapping-the Livestock Perspective.

Bae, H.W., Geiser, A.G., Kim, D.H., Chung, M.T., Burmester, J.K., Sporn, M.B., Roberts, A.B. and Kim, S.-J. (1995) Characterization of the Promoter Region of the Human Transforming Growth Factor-b Type II Receptor Gene. *The Journal of Biological Chemistry*, **270**, 29460-29468.

Ballabio, A. (1993) The Rise and Fall of Positional Cloning? *Nature Genetics*, **3**, 227-279.

Bartolomei, M.S. and Tilghman, S.M. (1997) Genomic Imprinting in Mammals. *Annu. Rev. Genet.*, **31**, 493-525.

Beattie, C.W. (1994) Livestock Genome Maps. *Trends in Genetics*, **10**, 334-338.

Beier, D.R. (1993) Single-Strand Conformation Polymorphism (SSCP) Analysis as a Tool for Genetic Mapping. *Mammalian Genome*, **4**, 627-631.

Bitgood, J.J. and Somes Jr, R.G. (1993) *Gene Map of the Chicken (Gallus gallus or G.domesticus).* Cold Spring Harbour Laboratory Press, Cold Spring Harbour.

Bloom, S.E., Delany, M.E. and Muscarella, D.E. (1993) Constant and Variable Features of Avian Chromosomes. In Etches, R.J. and Gibbons, A.M.V. (eds.), *Manipulation of the Avian Genome*. CRC Press, Florida, pp. 39-59.

Bonfield, J.K., Smith, K.F. and Staden, R. (1995) A New DNA Sequence Assembly Program. *Nucleic Acids Research*, **23**, 4992-4999.

Bonfield, J.K. and Staden, R. (1996) Experiment Files and Their Application During Large-Scale Sequencing Projects. *DNA Sequence-The Journal of Sequencing and Mapping*, **6**, 109-117.

Boularand, S., Biguet, N.F., Vidal, B., Veron, M., Mallet, J., Vincent, J.-D., Dufour, S. and Vernier, P. (1998) Tyrosine Hydroxylase in the European Eel (*Anguilla anguilla*): cDNA Cloning, Brain Distribution, and phylogenetic Analysis. *Journal of Neurochemistry*, **71**, 460-469.

Boulle, N., Schneid, H., Listrat, A., Holthuizen, P., Binoux, M. and Groyer, A. (1993) Developmental Regulation of Bovine Insulin-Like Growth Factor-II (IGF-II) Gene Expression: Homology Between Bovine Transcripts and Human IGF-II Exons. *Journal of Molecular Endocrinology*, **11**, 117-128.

Brahmachari, S.K., Meera, G., Balagurumoorthy, P., Tripathi, J., Raghavan, S., Shaligram, U. and Pataskar, S. (1995) Simple Repetitive Sequences in the Genome: Structure and Functional Significance. *Electrophoresis*, **16**, 1705-1714.

Brilliant, M.H., Niemann, M.M. and Eicher, E.M. (1987) Murine Tyrosine Hydroxylase Maps to the Distal End of Chromosome 7 Within a Region Conserved in Mouse and Man. *Journal of Neurogenetics*, **4**, 259-266.

Broad, T.E., Hill, D.F., Maddox, J.F., Montgomery, G.W. and Nicholas, F.W. (1998) The Sheep Gene Map. *ILAR Journal*, **39**, 160-170.

Brown, S.D.M. (1994) Integrating Maps of the Mouse Genome. *Current Opinion in Genetics and Development*, **4**, 389-394.

Bruford, M.W. and Burke, T. (1994) Minisatellite DNA Markers in the Chicken Genome I. Distribution and Abundance of Minisatellites in Multilocus DNA Fingerprints. *Animal Genetics*, **25**, 381-389.

Buitkamp, J., Ewald, D., Schalkwyk, L., Weiher, M., Masabanda, J., Sazanov, A., Lehrach, H. and Fries, R. (1998) Construction and Characterisation of a Gridded Chicken Cosmid Library With Four-Fold Genomic Coverage. *Animal Genetics*, **29**, 295-301.

Bumstead, N. Personal Communication

Bumstead, N. and Palyga, J. (1992) A Preliminary linkage map of the Chicken Genome. *Genomics*, **13**, 690-697.

Bumstead, N., Young, J.R., Tregaskes, C., Palyga, J. and Dunn, P.P.J. (1994) Linkage Mapping and Partial Sequencing of 10 cDNA Loci in the Chicken. *Animal Genetics*, **25**, 337-341.

Burch, J.B.E., Davis, D.L. and Haas, N.B. (1993) Chicken Repeat 1 Elements Contain a *pol*-like Open reading Frame and Belong to the Non-Long Terminal Repeat Class. *Proc. Natl. Acad. Sci. USA*, **90**, 8199-8203.

Burke, D., Burt, D.W. and Ponce de Leon, F.A. (1994) Chromosomal Assignment of the Chicken Transforming Growth Factor β3 Gene by FISH. *11th European Colloquium on Cytogenetics of Domestic Animals*, Denmark.

Burnside, J., Liou, S.S. and Cogburn, L.A. (1991) Molecular Cloning of the Chicken Growth Hormone Receptor Complementary Deoxyribonucleic Acid: Mutation of the Gene in Sex-Linked Dwarf Chickens. *Endocrinology*, **128**, 3183-3192.

Burnside, J., Liou, S.S., Zhong, C. and Cogburn, L.A. (1992) Abnormal Growth Hormone Receptor Gene Expression in the Sex-Linked Dwarf Chicken. *General and Comparative Endocrinology*, **88**, 20-28.

Burt, D.W. (1994) Mapping the Chicken GAPD Locus Using Heteroduplex Polymorphisms. *Animal Genetics*, **25**, 207.

Burt, D.W. (1999) Personal Communication

Burt, D.W., Bumstead, N., Bitgood, J.J., Ponce De Leon, F.A. and Crittenden, L.B. (1995) Chicken Genome Mapping: A New Era in Avian Genetics. *Trends in Genetics*, **11**, 190-194.

Burt, D.W., Bumstead, N., Burke, T., Fries, R., Groenen, M.A.M., Tixer-Boichard, M. and Vignal, A. (1997) Current Status of Poultry Genome Mapping-June 1997. *Proceedings of the 12th AVIAGEN Symposium: Current Problems in Avian Genetics*, 33-45.

Burt, D.W. and Cheng, H.H. (1998) The Chicken Gene Map. *ILAR*, **39**, 229-236.

Burton, D.W. and Bickham, J.W. (1989) Flow-Cytometric Analyses of Nuclear DNA Content in Four Families of Neotropical Bats. *Evolution*, **43**, 756-765.

Butler, D. (1996) Interest Ferments in Yeast Genome Sequence. *Nature*, **380**, 660-661.

Carlenius, C. (1981) R-, G- and C-Banded Chromosomes in the Domestic Fowl (*Gallus domesticus*). *Hereditas*, **94**, 61-66.

Carrier, A., Devignes, M.-D., Renoir, D. and Auffray, C. (1993) Chicken Tyrosine Hydroxylase Gene: Isolation and Functional Characterization of the 5' Flanking Region. *Journal of Neurochemistry*, **61**, 2215-2224.

Chalfie, M. and Jorgensen, E.M. (1998) *C.elegans* Neuroscience: Genetics to Genome. *Trends in Genetics*, **14**, 506-512.

Chase, P.B., Yang, J.-M., Thompson, F.H., Halonen, M. and Regan, J.W. (1996) Regional Mapping of the Human Platelet-Activating Factor Receptor Gene (PTAFR) to 1p35-p34.3 by Fluorescence in situ Hybridization. *Cytogenet Cell Genet*, **72**, 205-207.

Chen, E., d'Urso, M. and Schlessinger, D. (1994) Functional Mapping of the Human Genome by cDNA Localisation Versus Sequencing. *BioEssays*, **16**, 693-698.

Chen, Z.-Q., Ritzel, R.G., Lin, C.C. and Hodgetts, R.B. (1991) Sequence Conservation in Avian CR1: An Interspersed Repetitive DNA Family Under Functional Constraints. *Proc. Natl. Acad. Sci. USA*, **88**, 5814-5818.

Cheng, H.H. and Crittenden, L.B. (1994) Microsatellite Markers for genetic Mapping in the Chicken. *Poultry Science*, **73**, 539-546.

Cheng, H.H., Levin, I., Vallejo, R.L., Khatib, H., Dodgson, J.B., Crittenden, L.B. and Hillel, J. (1995) Development of a Genetic Map of the Chicken With Markers of High Utility. *Poultry Science*, **74**, 1855-1874.

Chowdhary, B.P., Raudsepp, T., Fronicke, L. and Scherthan, H. (1998) Emerging Patterns of Comparative Genome Organisation in Some Mammalian Species as Revealed by Zoo-FISH. *Genome Res.*, **8**, 557-589.

Churchill, P., Hempel, J., Romovacek, H., Zhang, W.W., Brennan, M. and Churchill, S. (1992) Primary Structure of Rat Liver D-Beta-Hydroxybutyrate Dehydrogenase from cDNA and Protein Analyses: A Short-Chain Alcohol Dehydrogenase. *Biochemistry*, **31**, 3793-3799.

Clark, M.S., Edwards, Y.J.K., McQueen, H.A., Meek, S.E., Smith, S., Umrania, Y., Warner, S., Williams, G. and Elgar, G. (1999) Sequence Scanning Chicken Cosmids: A Methodology for Genome Screening. *Gene*, **227**, 223-230.

Claverie, J.-M. and States, D.J. (1993) Information Enhancement Methods for Large Scale Sequence Analysis. *Computers Chem.*, **17**, 191-201.

Clement, Y., Hossein Kia, K., Daval, G. and Verge, D. (1996) An Autoradiographic Study of Serotonergic Receptors in a Murine Genetic Model of Anxiety-Related Behaviours. *Brain Research*, **709**, 229-242.

Collet, C., Candy, J. and Sara, V. (1998) Tyrosine Hydroxylase and Insulin-Like Growth Factor II but not Insulin are Adjacent in the Telost Species *Barramundi, Lates calcarifer*. *Animal Genetics*, **29**, 30-32.

Collins, F.S. (1995) Positional Cloning Moves from Perditional to Traditional. *Nature Genetics*, **9**, 347-350.

Craig, J.M. and Bickmore, W.A. (1994) The Distribution of CpG Islands in Mammalian Chromosomes. *Nature Genetics*, **7**, 376-381.

Crawford, R.D. (1990) Poultry Breeding and Genetics. .

Crittenden, L.B., Provencher, L., Santangelo, L., Levin, L., Abplanalp, H., Briles, R.W., Briles, W.E. and Dodgson, J.B. (1993) Characterization of a Red Jungle Fowl by White Leghorn Backcross Reference population for Molecular Mapping of the Chicken Genome. *Poultry Science*, **72**, 334-348.

Crooijmans, R.P., van Kampen, A.J.A., van der Poel, J.J. and Groenen, M.A.M. (1993) Highly polymorphic Microsatellite Markers in Poultry. *Animal Genetics*, **24**, 441-443.

Cross, S.H. (1995) CpG Islands and Genes. *Current opinion in Genetics and Development*, **5**, 309-314.

Cross, S.H., Charlton, J.A., Nan, X. and Bird, A.P. (1994) Purification of CpG Islands Using a Methylated DNA Binding Column. *Nature Genetics*, **6**, 236-224.

Cross, S.H., Lee, M., Clark, V.H., Craig, J.M., Bird, A.P. and Bickmore, W.A. (1997) The Chromosomal Distribution of CpG Islands in the Mouse: Evidence for Genome Scrambling in the Rodent Lineage. *Genomics*, **40**, 454-461.

Darling, D.C. and Brickell, P.M. (1996) Nucleotide Sequence and Genomic Structure of the Chicken Insulin-like Growth Factor-II (IGF-II) Coding Region. *General and Comparative Endocrinology*, **102**, 283-287.

Davies, P.O., Poirier, C., Deltour, L. and Montagutelli, X. (1994) Genetic Reassignment of the Insulin-1 (*Ins1*) gene to Distal Mouse Chromosome 19. *Genomics*, **21**, 665-667.

De Boer, L.E.M. and Sinoo, R.P. (1984) A Karyological Study of Accipitridae (Aves: Falconiformes), With Karyotypic Descriptions of 16 Species New to Cytology. *Genetica*, **65**, 89-107.

DeBry, R.W. and Seldin, M.F. (1996) Human/Mouse Homology Relationships. *Genomics*, **33**, 337-351.

Delgado, S., Gomez, M., Bird, A. and Antequera, F. (1998) Initiation of DNA Replication at CpG Islands in Mammalian Chromosomes. *The EMBO Journal*, **17**, 2426-2435.

Dietrich, W.F., Copeland, N.G., Gilbert, D.J., Miller, J.C., Jenkins, N.A. and Lander, E.S. (1995) Mapping the Mouse Genome: Current Status and Future Prospects. *Proc. Natl. Acad. Sci. USA*, **92**, 10849-10853.

Dietrich, W.F., Miller, J., Steen, R., Merchant, M.A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M.J., Ingalls, C.A., DeAngelis, M.M., Levinson, D.M., Kruglyak, L., Goodman, N., Copeland, D.G., Jenkins, N.A., Hawkins, T.L., Stein, L. and Page, D.C. (1996) A Comprehensive Map of the Mouse Genome. *Nature*, **380**, 149-152.

Dodgson, J.B., Cheng, H.H. and Okimoto, R. (1997) DNA Marker Technology: A Revolution in Animal Genetics. *Poultry Science*, **76**, 1108-1114.

Dominguez-Steglich, M., Jeltsch, J.-M. and Schmid, M. (1992) *In situ* Mapping of the Chicken Progesterone Receptor Gene and the Ovalbumin Gene. *Genomics*, **13**, 1343-1344.

Dower, W.J., Miller, J.F. and Ragsdale, C.W. (1988) High efficiency transformation of *E.coli* by High Voltage Electroporation. *Nucleic Acids Research*, **13**, 6127-6145.

Drew, A.C. and Brindley, P.J. (1997) A Retrotransposon of the Non-Long Terminal Repeat Class from the Human Blood Fluke *Schistosoma mansoni*. Similarities to the Chicken-Repeat-1-like Elements of Vertebrates. *Mol. Biol. Evol.*, **14**, 602-610.

Dujon, B. (1996) The Yeast Genome Project: What did we Learn? *Trends in Genetics*, **12**, 263-270.

Duret, L., Dorkeld, F. and Gautier, C. (1993) Strong Conservation of Non-Coding Sequences During Vertebrates Evolution: Potential Involvement in Post-Transcriptional Regulation of Gene Expression. *Nucleic Acids Research*, **21**, 2315-2322.

Eden, F.C. and Hendrick, J.P. (1978) Unusual Organization of DNA sequences in the Chicken. *Biochemistry*, **17**, 5838-5844.

Edwards, J.H. (1994) Comparative Genome Mapping in Mammals. *Current Opinion in Genetics and Development*, **4**, 861-867.

Eisen, J.A. (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, **8**, 163-167.

Elgar, G. (1996) Quality not Quantity: The Pufferfish Genome. *Hum. Mol. Gen.*, **5**, 1437-1442.

Elgar, G. and Clark, M. (1998) The Puffer Fish Gene Map. *ILAR Journal*, **39**, 249-256.

Elgar, G., Clark, M., Meek, S., Smith, S., Warner, S., Edwards, Y., Umrania, Y. and Williams, G. (1998) Analysis of the Fugu Genome. *8th International Workshop. Beyond the Identification of Transcribed Sequences: Functional Expression and Analysis*, Vienna, USA, p. 59.

Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B. and Brenner, S. (1996) Small is Beautiful-Comparative Genomics With the Puffer Fish (*Fugu rubripes*). *Trends in Genetics*, **12**, 145-150.

Eppig, J.T. (1996) Comparative Maps: Adding Pieces to the Mammalian Jigsaw Puzzle. *Current Opinions in Genetics and Development*, **6**, 723-730.

Fauquet, M., Grima, B., Lamouroux, A. and Mallet, J. (1988) Cloning of Quail Tyrosine Hydroxylase: Amino Acid Homology with Other Hydroxylases Discloses Functional Domains. *Journal of Neurochemistry*, **50**, 142-148.

Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994) How Many Genes in the Human Genome? *Nature Genetics*, **7**, 345-346.

Fillon, V. (1998) The Chicken as a Model to Study Microchromosomes in Birds: A Review. *Genet. Sel. Evol.*, **30**, 209-219.

Fillon, V., Langlois, P., Douaire, M., Gellin, J. and Vignal, A. (1997) Assignment of Steaoyl Coenzyme A Desatyrase Gene (SCD1) to Chicken Chromosome R-band 6q14 by *in situ* Hybridisation. *Cytogenetics and Cell Genetics*, 229-230.

Fillon, V., Morisson, M., Zoorob, R., Auffray, C., Douaire, M., Gellin, J. and Vignal, A. (1998) Identification of 16 Chicken Microchromosomes by Molecular Markers Using Two-Colour Fluorescence *in situ* Hybridisation (FISH). *Chromosome Research*, **6**, 307-313.

Fitzgerald, M.C., Skowron, P., Van Etten, J.L., Smith, L.M. and Mead, D.A. (1992) Rapid Shotgun Cloning Utilizing the Two Base Recognition Endonuclease Cvi JI. *Nucleic Acids Research*, **20**, 3753-3762.

Foury, F. (1997) Human Genetic Diseases: A Cross-Talk Between Man and Yeast. *Gene*, **195**, 1-10.

French, B.A., Bergsma, D.J. and Schwartz, R.J. (1990) Analysis of a CR1 (Chicken Repeat) Sequence Flanking the 5' End of the Gene Encoding α-Skeletal Actin. *Gene*, **88**, 173-180.

Fridolfsson, A.-K., Cheng, H.H. and Copeland, N.G. (1998) Evolution of the Avian Sex Chromosomes from an Ancestral Pair of Autosomes. *Proceedings of the National Academy of Sciences (USA)*, **95**, 8147-8152.

Fries, R. (1993) Mapping the Bovine Genome: Methodological Aspects and Strategy. *Animal Genetics*, **24**, 111-116.

Fritschi, S. and Stranzinger, G. (1985) Florescent Chromosome Banding in Inbred Chicken: Quinacrine Bands, Sequential Chromomycin and DAPI Bands. *Theoretical and Applied Genetics*, **71**, 408-412.

Gelbart, W.M. (1998) Databases in Genomic Research. *Science*, **282**, 659-661.

Gingrich, J.C., Boehrer, D.M. and Basu, S.B. (1996) Partial *Cvi* JI Digestion as an Alternative Approach to Generate Cosmid Sublibraries for Large-Scale Sequencing Projects. *BioTechniques*, **21**, 99-104.

Girard-Santosuosso, O., Bumstead, N., Lantier, I., Protais, J., Colin, P., Guillot, J.-F., Beaumont, C., Malo, D. and Lantier, F. (1997) Partial Conservation of the Mammalian *NRAMP1* Syntenic Group on Chicken Chromosome 7. *Mammalian Genome*, **8**, 614-616.

Gish, W. and States, D.J. (1993) Identification of Protein coding Regions by Database Similarity Search. *Nature Genetics*, **3**, 266-272.

Green, D., Marks, A.R., Fleischer, S. and McIntyre, J.O. (1996) Wild type and mutant Heart (R)-3-Hydroxybutyrate Dehydrogenase Expressed in Insect Cells. *Biochemistry*, **35**, 8158-8165.

Green, P., Lipman, D., Hillier, L., Waterson, R., States, D. and Claverie, J.-M. (1993) Ancient Conserved Regions in New Gene Sequences and the Protein Databases. *Science*, **259**, 1711-1716.

Griffiths, R. and Korn, R.M. (1997) A CHD1 Gene is Z Chromosome Linked in the Chicken *Gallus domesticus*. *Gene*, **197**, 229.

Grima, B., Lamouroux, A., Boni, C., Julien, J.F., Javoy-Agid, F. and Mallet, J. (1987) A Single Human Gene Encodes Multiple Tyrosine Hydroxylases Predicted to Differ in Their Functional Characteristics. *Nature*, **326**, 227-237.

Grima, B., Lamoutoux, A., Blanot, F., Faucon Biguet, N. and Mallet, J. (1985) Complete Coding Sequence of Rat Tyrosine Hydroxylase mRNA. *Proc. Natl. Acad. Sci. USA*, **72**, 3961-3965.

Groenen, M.A.M., R.P.M.A., C., Veenendaal, A., Cheng, H.H., Siwek, M. and van der Poel, J.J. (1998) A Comprehensive Microsatellite Linkage Map of the Chicken Genome. *Genomics*, **49**, 265-274.

Guldberg, P., Henriksen, K.F., Lou, H.C. and Guttler, F. (1998) Aberrant Phenylalnine Metabolism in Phenylketonuria Heterozygotes. *J. Inher. Metab. Dis.*, **21**, 365-372.

Haas, N.B., Grabowski, J.M., Sivitz, A.B. and Burch, J.B.E. (1997) Chicken Repeat 1 (CR1) Elements, Which Define and Ancient Family of Vertebrate Non-LTR Retrotransposons, Contain Two Closely Spaced Open Reading Frames. *Gene*, **197**, 305-309.

Hache, R.J.G. and Deeley, R.G. (1988) Organization, Sequence and Nuclease Hypersensitivity of Repetitive Elements Flanking the Chicken apoVLDLII Gene: Extended Similarity to Elements Flanking the Chicken Vitellogenin Gene. *Nucleic Acids Research*, **16**, 97-113.

Hachisuka, A., Yamazaki, T., Sawada, J. and Terao, T. (1996) Characterization and Tissue Distribution of Opioid-Binding Cell Adhesion Molecule (OBCAM) Using Monoclonal Antibodies. *Neurochem Int*, **28**, 373-379.

Hamblin, M.W. and Metcalf, M.A. (1991) Primary Structure and Functional Characterization of a Human 5-HT1D-type Serotonin Receptor. *Mol. Pharmacol.*, **40**, 143-148.

Hedges, S. (1994) Molecular Evidence for the Origin of Birds. *PNAS*, **91**, 2621-2624.

Hinegardner, R. (1976) *The Evolution of Genome Size.* Sinaur Associates, Massachusetts.

Hodgkin, J. and Herman, R.K. (1998) Changing Styles in *C.elegans* Genetics. *Trends in Genetics*, **14**, 352-357.

Holden, C. (1996) Gene-Hunters Choice: Fish or Fowl? *Science*, **271**, 1369.

Holmquist, G., Gray, M., Porter, T. and Jordan, J. (1982) Characterization of Giemsa Dark- and Light-Band DNA. *Cell*, **31**, 121-129.

Holmquist, G.P. (1992) Chromosome Bands, their Chromatin Flavours, and their Functional Features. *Am J Hum Genet*, **51**, 17-37.

Hougaard, S., Abrahamsen, N., Moses, H.L., Spang-Thomsen, M., Skovgaard Poulsen, H. (1999) Inactivation of the Transforming Growth Factor Beya Type II Receptor in Human Cell Lines. *Br J Cancer*, **79**, 1005-1011.

Hu, J., Bumstead, N., Burke, D., Ponce de Leon, F.A., Skamene, E., Gros, P. and Malo, D. (1995) Genetic and Physical Mapping of the Natural Resistance-Associated Macrophage Protein (NRAMP1) in Chicken. *Mammalian Genome*, **6**, 809-815.

Huang, H.-B., Song, Y.-Q., Hsel, M., Zahorchak, R., Chiu, J., Teuscher, C. and Smith, E.J. (1999) Development and Characterization of Genetic Mapping Resources for the Turkey (*Meleagris gallopavo)*. *The Journal of Heredity*, **90**, 240-242.

Hughes, A.L. and Hughes, M.K. (1995) Small Genomes for Better Flyers. *Nature*, **377**, 391.

Hull, K.L., Fraser, R.A., Marsh, J.A. and Harvey, S. (1993) Growth Hormone Receptor Gene Expression in Sec-Linked Dwarf Leghorn Chickens: Evidence Against a Gene Deletion. *Journal of Endocrinology*, **137**, 91-98.

Hutt, F.B. (1936) Genetics of the Fowl. VI. A Tentative Chromosome Map. *Neue Forsch Tiersucht Abstam (Duerst Festschrift)*, 105-112.

Ishikawa, K., Nagase, T., Suyama, M., Miyajima, N., Tanaka, A., Kotani, H., Nomura, N. and Ohara, O. (1998) Prediction of the Coding Sequences of Unidentified Human Genes. X. The Complete Sequences of 100 New cDNA Clones from Brain Which Can Code for Large Proteins *in vitro. DNA Res.*, **5**, 169-176.

Itakura, H., Matsumoto, A., Asaoka, H. and Kodama, T. (1993) Structure and Function of the Scavenger Receptor. *Nippon Rinsho*, **51**, 1083-1091.

Jacobs, K.A., Collins-Racie, L.A., Colbert, M., Duckett, M., Goldenfleet, M., Kelleher, K., Kriz, R., LaVillie, E.R., Merberg, D., Spaulding, V., Stover, J., Williamson, M.J. and McCoy, J.M. (1997) A Genetic Selection for Isolating cDNAs Encoding Secreted Proteins. *Gene*, **198**, 289-296.

Johnson, R.L. and Tabin, C.J. (1997) Molecular Models for Vertebrate Limb Development. *Cell*, **90**, 979-990.

Jones, P.A. (1999) The DNA Methylation Paradox. *Trends in Genetics*, **15**, 34-37.

Jones, S.J.M. (1995) An Update and Lessons From Whole-Genome Sequencing Projects. *Curr. Opin. Genet. Dev.*, **5**, 349-353.

Kajikawa, M., Ohshimo, K. and Okada, N. (1997) Determination of the Entire Sequence of Turtle CR1: The First Open Reading Frame of the Turtle CR1 Element Encodes a Protein With a Novel Zinc Finger Motif. *Mol. Biol. Evol.*, **14**, 1206-1217.

Kajimoto, Y. and Rotwein, P. (1991) Structure of the Chicken Insulin-like Growth Factor I Gene Reveals Conserved Promoter Elements. *The Journal of Biological Chemistry*, **266**, 9724-9731.

Kallincos, N.C., Wallace, J.C., Francis, G.L. and Ballard, F.J. (1990) Chemical and Biological Characterization of Chicken Insulin-Like Growth Factor-II. *Journal of Endocrinology*, 89-97.

Klein, S., Morrice, D.R., Sang, H., Crittenden, L.B. and Burt, D.W. (1996) Genetic and Physical Mapping of the Chicken IGF1 Gene to Chromosome 1 and Conservation of Synteny With Other Vertebrate Genomes. *Journal of Heredity*, **87**, 10-14.

Ladjali, K., Bitgood, J.J., Tixier-Boichard, M. and Ponce de Leon, F.A. (1999) International System for Standardizeed Avian karyotypes (ISSAK): Standardized Banded Karyotypes of the Domestic Fowl (*Gallus domesticus*). *Cytogenetics and Cell Genetics.*, **In press.**

Lalande, M. (1997) Parental Imprinting and Human Disease. *Annu. Rev. Genet.*, **30**, 173-195.

Law, A. Report Repeats, Personal Communication.

Law, S.K.A., Micklem, K.J., Shaw, J.M., Zhang, X.-P., Dong, Y., Willis, A.C. and Mason, D.Y. (1993) A New Macrophage Differentiation Antigen Which is a Member of the Scavenger Receptor Superfamily. *European Journal of Immunology*, **23**, 2320-2325.

Lawler, S., Candia, A.F., Ebner, R., Shum, L., Lopez, A.R., Moses, H.L., Wright, C.V.E. and Derynck, R. (1994) The Murine Type II TGF-β Receptor has a Coincident Embryonic Exprression and Binding Preference for TGF-β1. *Development*, **120**, 165-175.

Ledley, F.D., DiLella, A.G., Kwok, S.C.M. and Woo, S.L.C. (1985) Homology between Phenylalanine and Tyrosine Hydroxylases Reveals Common Structural and Functional Domains. *Biochemistry*, **24**, 3389-3394.

Levin, I., Santangelo, L., Cheng, H.H., Crittenden, L.B. and Dodgson, J. (1994) An Autosomal Genetic Linkage Map of the Chicken. *Journal of Heredity*, **85**, 79-85.

Levy, J. (1994) Sequencing the Yeast Genome: An International Achievement. *Yeast*, **10**, 1689-1706.

Li, X., Wistow, G.J. and Piatigorsky, J. (1995) Linkage and Expression of the Argininosuccinate Lyase/δ–Crystallin Genes of the Duck: Inserion of a CR1 Element in the Intergenic Spacer. *Biochimica et Biophysica Acta*, **1261**, 25-34.

Lundin, L.G. (1993) Evolution of the Vertebrate Genome as Reflected in Paralogous Chromosomal Regions in Man and the House Mouse. *Genomics*, **16**, 1-19.

Manly, K.F. (1993) A Macintosh Program for Storage and Analysis of Experimental Genetic Mapping Data. *Mammalian Genome*, **4**, 303-313.

Mann, S.M., Burkin, D.J., Griffin, D.K. and Ferguson-Smith, M.A. (1997a) A Fast, Novel Approach for DNA Fibre-Fluorescence *in situ* Hybridization Analysis. *Chromosome Res*, **5**, 145-147.

Mann, S.M., Burkin, D.J., Griffin, D.K. and Ferguson-Smith, M.A. (1997b) A Fast, Novel Fibre-FISH Technique. *Cytogenetics and Cell Genetics*, Vol. 77, p. P34.

Mannens, M., Alders, M., Redeker, B., Bliek, J., Steenman, M., Wiesmeyer, C., de Feinberg, A., Little, P. and Westerveld, A. (1996) Positional Cloning of Genes Involved in the Beckwith-Wiedemann Syndrome, Hemihypertrophy, and Associated Childhood Tumors. *Med Pediatr Oncol*, **27**, 490-494.

Marks, A.R., McIntyre, J.O., Duncan, T.M., Erdjumentbromage, H., Tempest, P. and Fleischer, S. (1992) Molecular Cloning and Characterization of (R)-3-Hydroxybutyrate Dehydrogenase from Human Heart. *J. Biol. Chem.*, **22**, 15459-15463.

Marshall Graves, J.A. (1998) Background and Overview of Comparative Genomics. *ILAR Journal*, **39**, 48-65.

Matzke, Varga, F., Berger, H., Schernthaner, J., Schweizer, Mayr, B. and Matzke, A.J.M. (1990) 41-42 bp Tandemly Repeated Sequence Isolated from Nuclear Envelopes of Chicken Erythrocytes Located Predominantly on Microchromosomes. *Chromosoma*, **99**, 131-137.

McQueen, H.A., Clark, V.H., Bird, A.P., Yerle, M. and Archibald, A.L. (1997) CpG Islands of the Pig. *Genome Research*, **7**, 924-931.

McQueen, H.A., Fantes, J., Cross, S.A., Clark, V.H., Archibald, A.L. and Bird, A.P. (1996) CpG Islands of Chicken are Concentrated on Microchromosomes. *Nature Genetics*, **12**, 321-324.

McQueen, H.A., Siriaco, G. and Bird, A.P. (1998) Chicken Microchromosomes Are Hyperacetylated, Early Replicating, and Gene Rich. *Genome Research*, **8**, 621-630.

McRory, J.E. and Sherwood, N.M. (1997) Ancient Divergence of Insulin and Insulin-Like Growth Factor. *DNA and Cell Biology*, **16**, 939-949.

Metzstein, M.M., Stanfield, G.M. and Horvitz, R.H. (1998) Genetics of Programmed Cell Death in *C. elegans* : Past, Present and Future. *Trends in Genetics*, **14**, 410-416.

Morisson, M., Pitel, F., Fillon, V., Pouzadoux, A., Berge, R., Vit, J.P., Zoorob, R., Auffray, C., Gellin, J. and Vignal, A. (1998) Integration of Chicken Cytogenetic and Genetic Maps: 18 New Polymorphic Markers Isolated from BAC and PAC clones. *Animal Genetics*, **29**, 348-355.

Morizot, D.C. (1994) Reconstructing the Gene Map of the Vertebrate Ancestor. *Anim. Biotech.*, **5**, 113-122.

Morrice, D.R. and Burt, D.W. (1995) A *Sst* I RFLP at the Chicken Transforming Growth Factor Beta-2 Locus (TGFβ 2). *Animal Genetics*, **26**, 210.

Moustakas, A., Lin, Y., Henis, Y, Plamondon, J., O'Connor-McCourt, M.D. and Lodish, H.F. (1993) The Transforming Growth Factor β Receptors Types I, II and III Form Hetero-oligomeric Complexes in the Presence of Ligand. *The Journal of Biological Chemistry*, **268**, 22215-22218.

Nadeau, J.H., Grant, P.L., Mankala, S., Reiner, A.H., Richardson, J.E. and Eppig, J.T. (1995) A Rosetta Stone of Mammalian Genetics. *Nature*, **373**, 363-365.

Nadeau, J.H. and Sankoff, D. (1998) The Lengths of Undiscovered Conserved Seqments in Comparative Maps. *Mammalian Genome*, **9**, 491-495.

Nanda, I., Shan, Z., Scharti, M., Burt, D.W., Koehler, M., Nothwang, H.-G., Grutzner, F., Paton, I.R., Windsor, D., Dunn, I., Engel, W., Staeheli, P., Mizuno, S., Haaf, T. and Schmid, M. (1999) 300 Million Years of Conserved Synteny Between Chicken Z and Human Chromosome 9. , **21**, 258-259.

Nanda, I., Tanaka, T. and Schmid, M. (1996) The Intron-Containing Ribosomal Protein-Encoding Genes L5, L7a and L37a are Unlinked in Chicken. *Gene*, **170**, 159-164.

New England Biolabs (1998) *In vitro* Transposon-based Kit for Sequencing Projects. GPS^{TM}-1, A Faster Alternative to Primer Walking, Random Subcloning and Nested Deletions. *The NEB Transcript. A Scientific Newsletter from New England Bioloabs*, Vol. 9, p. 5.

Nohno, T., Sumitomo, S., Ishikawa, T., Ando, C., Nishida, S., Noji, S. and Saito, T. (1993) Nucleotide sequence of a cDNA Encoding the Chicken Receptor Protein Kinase of the TGF-Beta Receptor Family. *DNA Sequence-J.DNA Sequencing and Mapping*, **3**, 393-396.

Nurminsky, D.I. and Hartl, D.L. (1996) Sequence Scanning: A Method for Rapid Sequence Acquisition from Large-Fragment DNA Clones. *Proc. Natl. Acad. Sci. USA*, **93**, 1694-1698.

O'Brien, S.J. (1991) Molecular genome Mapping: Lessons and Prospects. *Curr. Opin. Genet. Dev.*, **1**, 105-111.

O'Brien, S.J., Weinberg, J. and Lyons, L.A. (1997) Comparative Genomics: Lessons from Cats. *Trend in Genetics*, **13**, 393-399.

O'Brien, S.J., Womack, J.E., Lyons, L.A., Moore, K.J., Jenkins, N.A. and Copeland, N.G. (1993) Anchored Reference Loci for Comparative Genome Mapping in Mammals. *Nature Genetics*, **3**, 103-112.

O'Malley, L.O. and Rotwein, P. (1988) Human Tyrosine Hydroxylase and Insulin Genes are Contiguous on Chromosome 11. *Nucleic Acids Research*, **16**, 4437-4446.

Ogasa, H., Noma, T., Murata, H., Kawai, S. and Nakazawa, A. (1996) Cloning of a cDNA Encoding the Human Transforming Growth Factor-Beta Type II Receptor: Heterogeneity of the mRNA. *Gene*, **181**, 185-190.

Okimoto, R., Cheng, H.H. and Dodgson, J.B. (1997) Characterization of CR1 Repeat Random PCR Markers for Mapping the Chicken Genome. *Animal Genetics*, **28**, 139-145.

Olofsson, B. and Bernardi, G. (1983) The Distribution of CR1, an *Alu*-LIKE Family of Interspersed Repeats, in the Chicken Genome. *Biochimica et Biophysica Acta*, **740**, 339-341.

Parimoo, S. (1995) cDNA Selection and Other Approaches in Positional Cloning. *Analytical Biochemistry*, **228**, 1-17.

Patterson, J.C. (1995) A Complex Issue. *Trends in Genetics*, **11**, 463.
Perier, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) The Evolution of Genes: The Chicken Preproinsulin Gene. *Cell*, **20**, 555-566.

Pitel, F., Fillon, V., Heimel, C., Fur, N.L., Khadir-Mounier, C.E., Douaire, M., Gellin, J. and Vignal, A. (1998) Mapping of *FASN* and *ACACA* on Two Chicken Microchromosomes Disrupts the Human 17q Syntenic Group Well Conserved in Mammals. *Mammalian Genome*, **9**, 297-300.

Ponce de Leon, F.A., Yukui, L. and Smith, E.J. (1991) Reassignment of the *ev* 1 Locus by High Resolution Chromosomal *in situ* Hybridisation. *Poultry Science*, **70 (supp. 1)**, 95.

Prescott, S.M., Zimmerman, G.A. and McIntyre, T.M. (1990) Platelet-Activating Factor. *The Journal of Biological Chemistry*, **265**, 17381-17384.

Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Moller, A.P. and Ellegren, H. (1997) Low Frequency of Microsatellites in the Avian Genome. *Genome Res*, **7**, 471-482.

Rauen, K.A., LeCiel, C.D.S., Abbott, U.K. and Hutchison, N.J. (1994) Localization of the Chicken PGK Gene to Chromosome 4p by Fluorescence *in situ* Hybridisation. *Journal of Heredity*, **85**, 147-150.

Reitman, M., Grasso, J.A., Blumenthal, R. and Lewit, P. (1993) Primary Sequence, Evolution, and Repetitive Elements of the *Gallus gallus* (Chicken) β-Globin Cluster. *Genomics*, **18**, 616-626.

Resch, K., Korthaus, D., Wedemeyer, N., Lengeling, A., Ronsiek, M., Thiel, C., Baer, K., Jockusch, H. and Schmitt-John, T. (1998) Homology Between Human Chromosome 2p13.3 and the Wobbler Critical Region on Mouse Chromosome 11: Comparative High-Resolution Mapping of STS and EST Loci on YAC/BAC Contigs. *Mammalian Genome*, **9**, 893-898.

Rodionov, A.V. (1996) Micro Versus Macro: A Review of Structure and Functions of Avian Micro- and Macrochromosomes. *Russian Journal of Genetetics*, **32**, 517-527.

Rodionov, A.V. (1997) Evolution of Avian Chromosomes and Linkage Groups. *Russian Journal of Genetics,* **33** (6), 605-617.

Rodionov, A.V., Myakoshina, Y.A., Chelysheva, L.A., Solovei, I.V. and Gaginskaya, E.R. (1992) Chiasmata on Lampbrush Chromosomes of *Gallus gallus domesticus*: A Cytogenetic Study of Recombination Frequency and Linkage Group Lengths. *Genetika*, **28**, 53-63.

Rotwein, P. (1991) Structure, Evolution, Expression and Regulation of Insulin-Like Growth Factors I and II. *Growth Factors*, **5**, 3-18.

Ruyter-Spira, C.P., de Groof, A.J.C., van der Poel, J.J., Herbergs, J., Masabanda, J., Fries, R. and Groenen, M.A.M. (1998) The HMGI-C Gene is a Likely Candidate for the Autosomal Dwarf Locus in the Chicken. *The Journal of Heredity*, **89**, 295-300.

Saitoh, Y., Ogawa, A., Hori, T., Kunita, R. and Mizuno, S. (1993) Identification and Localization of Two Genes on the Chicken Z Chromosome: Implication of Evolutionary Conservation of the Z Chromosome Among Avian Species. *Chromosome Research*, **1**, 239-251.

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

Sang, H. Personal Communication.

Saudou, F. and Hen, R. (1994) 5-Hydroxytryptamine Receptor Subtypes in Vertebrates and Invertebrates. *Neurochem. Int.*, **25**, 503-532.

Sazanov, A.A., Masabanda, J., Ewald, D., Burt, D.W., Smith, J., Bruley, C.K. and Fries, R. (1998) Cytogenetic Mapping of 30 Type I Markers in Chicken. .

Sedlacek, Z., Konecki, D.S., Siebenhaar, R., Kioschis, P. and Poustka, A. (1993) Direct Selection of DNA Sequences Conserved Between Species. *Nucleic Acids Research*, **21**, 3419-3425.

Seyfried, C.E., Schweickart, V.L., Godiska, R. and Gray, P.W. (1992) The Human Platelet-Activating Factor Receptor Gene (PTAFR) Contains No Introns and Maps to Chromosome 1. *Genomics*, **13**, 832-834.

Shapira, E., Yarus, S. and Fainsod, A. (1991) Genomic Organization and Expression During Embryogenesis of the Chicken CR1 Repeat. *Genomics*, **10**, 931-939.

Shark, K.B. and Lee, N.M. (1995) Cloning, Sequencing and Localization to Chromosome 11 of a cDNA Encoding a Human Opioid-Binding Cell Adhesion Molecule (OBCAM). *Gene*, **155**, 213-217.

Shih, J.C., Yang, W., Chen, K. and Gallaher, T. (1991) Molecular Biology of Serotonin (5-HT) Receptors. *Pharmacology Biochemistry & Behavior*, **40**, 1053-1058.

Shimogiri, T., Kono, M., Mannen, H. and Mizutano, S. (1993) Chicken Ornithine Transcarbamylase Gene, Structure, Regulation and Chromosomal Assignment: Repetitive Sequence Motif in Intron 3 Regulates This Enzyme Activity. *Journal of Biochemistry*, **124**, 962-971.

Silva, R. and Burch, J.B.E. (1989) Evidence that Chicken CR1 Elements Represent a Novel Family of Retroposons. *Molecular and Cellular Biology*, **9**, 3563-3566.

Smith, E.J. and Cheng, H.H. (1998) Mapping Chicken Genes Using Preferential Amplification of Specific Alleles. *Microb. Comp. Genomics*, **3**, 13-20.

Smith, J. and Burt, D.W. (1998) Parameters of the Chicken Genome (*Gallus gallus*). *Animal Genetics*, **29**, 290-294.

Smith, J., Paton, I.R., Bruley, C.K., Windsor, D., Burke, D., Ponce de Leon, F.A. and Burt, D.W. (1999) Integration of the Genetic and Physical Maps of the Chicken Macrochromosomes. *Animal Genetics*. **In press**

Smith, M.W., Holmsen, A.L., Wei, Y.H., Paterson, M. and Evans, G.A. (1994) Genomic Sequence Sampling: A Strategy for High Resolution Sequence-Based Physical Mapping of Complex Genomes. *Nature Genetics*, **7**, 40-47.

Soares, M.B., Schon, E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J. and Efstratiadis, A. (1985) RNA-Mediated Gene Duplication: The Rat

Preproinsulin I Gene is a Functional Retroposon. *Molecular and Cellular Biology*, **5**, 2090-2103.

Soret, J., Vellard, M., Viegas-Pequignot, E., Apiou, F., Dutrillaux, B. and Perbal, B. (1990) Chromosomal Reallocation of the Chicken *c-myb* Locus and Organization of 3'-proximal Coding Exons. *FEBS*, **263**.

Spaltmann, F. and Brummendorf, T. (1996) CEPU-1, a Novel Immunoglobulin Superfamily Molecule, is Expressed by Developing Cerebellar Purkinje Cells. *The Journal of Neuroscience*, **16**, 1770-1779.

Spike, C.A., Bumstead, N., Crittenden, L.B. and Lamont, S.J. (1996) RFLP Mapping of Expressed Sequence Tails in the Chicken. *Journal of Heredity*, **87**, 6-9.

Staden, R., Bonfield, J. and Beal, K. (1996) *The New Staden Package Manual - Part 1*. The Medical Research Council, Laboratory of Molecular Biology.

Start, K. (1998) Treating Phenylketonuria by a Phenylaline-Free Diet. *Professional Care of Mother and Child*, **8**, 109-110.

Strimmer, K. and von Haesler, A. (1996) Quartet Puzzling:A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution,* **13**, 964-969.

Strimmer, K., Goldman, N. and von Haesler, A. (1997) Bayesian Probabilities and Quartet Puzzling. *Molecular Biology and Evolution*, **14**, 210-211.

Stumph, W.E., Hodgson, C.P., Tsai, M.-J. and O'Malley, B.W. (1984) Genomic Structure and Possible Retroviral Origin of the Chicken CR1 Repetitive DNA Sequence Family. *Proc. Natl. Acad. Sci. USA*, **81**, 6667-6671.

Sumner, A.T., de la Torre, J. and Stuppia, L. (1993) The Distribution of Genes on Chromosomes: A Cytological Approach. *J Mol Evol*, **37**, 117-122.

Symonds, G., Stubblefield, E., Guyaux, M. and Bishop, J.M. (1984) Cellular Oncogenes (c-ERB-A and c-ERB-B) Located on Different Chicken Chromosomes can be Transduced into the Same Viral Genome. *Molecular Cell Biology*, **4**, 1627-1630.

Takenoshita, S., Hagiwara, K., Nagashima, M., Gemma, A., Bennett, W.P. and Harris, C.C. (1996) The Genomic Structure of the Gene Encoding the Human Transforming Growth Factor β Type II Receptor (TGF-β RII). *Genomics*, **36**, 341-344.

Taylor, E.R., Seleiro, E.A. and Brickell, P.M. (1991) Identification of Antisense Transcripts of the Chicken Insulin-Like Growth Factor-II Gene. *Journal of Molecular Endocrinology*, **7**, 145-154.

Tegelstrom, H. and Ryttman, H. (1981) Chromosomes in Birds (Aves): Evolutionary Implications of Macro- and Microchromosomes Numbers and Lengths. *Hereditas*, **94**, 225-233.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, **22**, 4673-4680.

Tiersch, T.R. and Wachtel, S.S. (1991) On the Evolution of Genome Size of Birds. *Journal of Heredity*, **82**, 363-368.

Tillet, Y., Krieger, M. and Thibault, J. (1997) Isolation and Characterization of Ovine Tyrosine Hydroxylase mRNA. *Journal of Neurochemistry*, **68**, 2161-2160.

Toye, A.A., Bumstead, N. and Moran, C. (1997) A Pentanucleotidde Repeat Polymorphism Maps Progesterone Receptor (PGR) to Chicken Chromosome 1. *Animal Genetics*, **28**, 317.

Tyfield, L.A. (1997) Phenylketonuria in Britain: Genetic Analysis Gives a Historical Perspective of the Disorder But Will it predict the Future for Affected Individuals? *J Clin Pathol: Mol Pathol*, **50**, 169-174.

Upton, Z., Francis, G.L., Kita, K., Wallace, J.C. and Ballard, F.J. (1995) Production and Characterization of Recombinant Chicken Insulin-Like Growth Factor-II from *Escherichia coli*. *Journal of Molecular Endocrinology*, **14**, 79-90.

Van Den Bussche, R.A., Longmire, J.L. and Baker, R.J. (1995) How Bats Achieve a Small C-Value: Frequency of Repetitive DNA in *Macrotus*. *Mammalian Genome*, **6**, 521-525.

Vandergon, T.L. and Reitman, M. (1994) Evolution of Chicken Repeat 1 (CR1) Elements: Evidence for Ancient Subfamilies and Multiple Progenitors. *Mol. Biol. Evol.*, **11**, 886-898.

Vignal, A. Personal Communication.

Volpi, E.V. and Baldini, A. (1993) MULTIPRINS: A Method for Multicolour Primed *in situ* Labelling. *Chromosome Research,* **1,** 257-260.

Wakefield, M.J. (1998) Internet Comparative Mapping Resources. *ILAR Journal,* **39,** 66-67.

Wallen, M.J., Keinanen, R.A. and Kulomaa, M.S. (1996) Two Chicken Repeat One (CR1) Elements Lacking a Silencer-Like Region Upstream of the Chicken Avidin-Related Genes *Avr4* and *Avr5*. *Biochimica et Biophysica Acta,* **1308,** 193-196.

Watson, S. and Arkinstall, S. (1994) *The G-Protein linked Receptor Facts Book.* Academic Press Limited, London.

Weeks, D.E. (1995) Polygenic Disease: Methods for Mapping Complex Disease Traits. *Trends in Genetics,* **11,** 513-519.

Weinshank, R.L., Zgombick, J.M., Macchi, M.J., Branchek, T.A. and Hartig, P.R. (1992) Human Serotonin 1D Receptor is encoded by a Subfamily of Two Distinct genes: 5-HT1D Alpha and 5-HT1D Beta. *Proc. Natl. Acad. Sci USA,* **89,** 3630-3634.

Wennborg, A., Sohlberg, B., Angerer, D., Klien, G. and von Gabin, A. (1995) A Human RNase E-Like Activity That Cleaves RNA Sequences Involved in mRNA Stability Control. *Proc. Natl. Acad. Sci. USA,* **92,** 7322-7326.

Wijngaard, P.L.J., MacHugh, N.D., Metzelaar, M.J., Romberg, S., Bensaid, A., Pepin, L., Davis, W.C. and Clevers, H.C. (1994) Members of the Novel WC1 Gene Family Are Differentially Expressed on Subsets of Bovine $CD4^-CD8^-$ $\gamma\delta$ T. *Journal of Immunology,* **152,** 3476-3482.

Wijngaard, P.L.J., Metzelaar, M.J., MacHugh, N.D., Morrison, W.I. and Clevers, H.C. (1992) Molecular Characterization of the WC1 Antigen Expressed Specifically on Bovine CD4$^-$CD8$^-$ $\gamma\delta$ T Lymphocytes. *Journal of Immunology*, **149**, 3273-3277.

Wilson, R.K. (1999) How the Worm was Won. *Trends in Genetics*, **15**, 51-58.

Womack, J.E. (1998) The Cattle Gene Map. *ILAR Journal*, **39**, 153-159.

Wright, F. (1999) HGMP Training Course Notes: Phylogenetic Trees from Molecular Sequences. Biomathematics and Statistics, Scotland.

Wu, C.-S.C., Hasegawa, J., Smith, A.P., Loh, H.H., Lee, N.M. and Yang, J.T. (1990) Opioid-Binding Protein (OBCAM) is Rich in Beta-sheets. *Journal of protein Chemistry*, **9**, 3-7.

Wurch, T., Palmier, C., Colpaert, F.C. and Pauwels, P.J. (1997) Sequence and Functional Analysis of Cloned Guinea Pig and Rat Serotonin 5-HT$_{1D}$ Receptors: Common Pharmacological Features Within the 5-HT$_{1D}$ Receptor Subfamily. *Journal of Neurochemistry*, **68**, 410-418.

Xia, Y.N., Burbank, D.E., Uher, L., Rabussay, D. and Van Etten, J.L. (1987) IL-3A Virus Infection of a *Chlorella* -like Green Alga Induces a DNA Restriction Endonuclease With Novel Sequence Specificity. *Nucl. Acids. Res.*, **15**, 6075-6090.

Xue, F., Kidd, J.R., Pakstis, A.J., Castiglione, C.M., Mallet, J. and Kidd, K.K. (1988) Tyrosine Hydroxylase Maps to the Short Arm of Chromosome 11 Proximal to the Insulin and HRAS1 Loci. *Genomics*, **2**, 288-293.