THE UNIVERSITY
of EDINBURGH

The abundance and diversity of endogenous retroviruses in the chicken genome

Andrew Stephen Mason

Thesis submitted for the degree of Doctor of Philosophy

College of Medicine and Veterinary Medicine

The Roslin Institute and Royal (Dick) School of Veterinary Studies

University of Edinburgh

2017

Acknowledgements

The completion of this PhD project and thesis has been the most academically and mentally challenging undertaking I have faced. It is without a doubt that I would have been unable to do so without the encouragement and support of a great many people, who I would like to take this opportunity to thank.

Firstly, to my academic supervisors: Professor Dave Burt, Dr Paul Hocking and Dr Sam Lycett. Dave and Paul, thank you for devising this project, taking on a 'fresh faced' graduate, and continually supporting and driving my academic development. You've both been incredibly supportive of my desire to take on extra activities, whether that be a training course, public engagement or undergraduate teaching, all of which have enriched my experience. You've shared a wealth of knowledge and life experience with me, and for that I am grateful. Thank you also to Sam for taking me on during my final year and critically reviewing my work with fresh eyes. You've been a great help.

Secondly, to my industrial supervisor at Hy-Line International, Dr Janet Fulton. You've been a great source of inspiration, opened my eyes to new ideas, but you kept me grounded and focused on the project. Thank you also for so generously hosting me during my month's stay in Des Moines, and opening your home for games, whisky and sage advice. Thank you also to the entire Molecular Genetics Laboratory group at Hy-Line, particularly Ashlee Lund for her tireless work on the KASP assays.

Thirdly, thank you to the collaborators who so generously shared data during my identification of ALVEs: Professors Chris Ashwell, Bernhard Benkel, Marc Eloit, Olivier Hanotte, Susan Lamont, Rudolf Preisinger and Douglas Rhoads. Special thanks are also due to Dr Scott Tyack from the EW group who shared ideas and troubleshooting methods during the development of the ALVE identification pipeline, and to Dr Gregor Gorjanc at Roslin for helping with modelling ALVE detection rates.

I have been lucky to part of the Burt lab at Roslin, full of clever and incredibly helpful people. Thank you specifically to Richard Kuo and Dr Lel Eory for dealing with many bioinformatic questions, to Bob Paton and Dr Kasia Miedzinska-Bielecka for help and patience in the lab, and to Dr Jacqueline Smith for good advice and a dose of reality. Beyond great academic support at Roslin, I was fortunate to find a group of people in

the department (and beyond) with whom I shared laughs, cakes and beers during my project. Special mentions to Serap, Graham, Kate, Brendan, Fiona, Carys and John: you guys kept Roslin fun. Thank you for the good memories, and the best of luck to you all.

I owe my Mum and Dad a huge thank you for all the love and support over the years. They always drive me to do better, support me through any decision, and even cope with my excited scientific monologuing. You're both incredible and it's my constant goal to make you proud. As well as the love and support of my sisters and their families, I have been so lucky during this project to see my family extend to the Hawleys, my brilliant in-laws. Thank you for welcoming me so completely into your lives.

Finally, and above all others, I need to acknowledge the love and brilliance of my incredible wife, Cathy. Completing PhDs simultaneously has been a challenge, but we've supported each other throughout, and could put a smile on the other's face after any long day or inconclusive result. You are absolutely everything to me, and I am so fortunate to be spending the rest of my life with you. Here's to a life-long, post-PhD adventure, and all the amazing things to come!

Chicken Lookin' by Alex Hawley
My very own motivational poster.

Declaration

I declare that this thesis has been composed solely by myself, and that it has not been submitted, in whole or in part, in any previous application for a degree or other professional qualification. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Andrew Mason

Abstract

Long terminal repeat (LTR) retrotransposons are autonomous eukaryotic repetitive elements which may elicit prolonged genomic and immunological stress on their host organism. LTR retrotransposons comprise approximately 10 % of the mammalian genome, but previous work identified only 1.35 % of the chicken genome as LTR retrotransposon sequence. This deficit appears inconsistent across birds, as studied Neoaves have contents comparable with mammals, although all birds contain only one LTR retrotransposon class: endogenous retroviruses (ERVs). One group of chicken-specific ERVs (Avian Leukosis Virus subgroup E; ALVEs) remain active and have been linked to commercially detrimental phenotypes, such as reduced lifetime egg count, but their full diversity and range of phenotypic effects are poorly understood.

A novel identification pipeline, LocaTR, was developed to identify LTR retrotransposon sequences in the chicken genome. This enabled the annotation of 3.01 % of the genome, including 1,073 structurally intact elements with replicative potential. Elements were depleted within coding regions, and over 40 % of intact elements were found in clusters in gene sparse, poorly recombining regions. RNAseq analysis showed that elements were generally not expressed, but intact transcripts were identified in four cases, supporting the potential for viral recombination and retrotransposition of non-autonomous repeats. LocaTR analysis of seventy-two additional sauropsid genomes revealed highly lineage-specific repeat content, and did not support the proposed deficit in Galliformes.

A second, novel bioinformatic pipeline was constructed to identify ALVE insertions in whole genome resequencing data and was applied to eight elite layer lines from Hy-Line International. Twenty ALVEs were identified and diagnostic assays were developed to validate the bioinformatic approach. Each ALVE was sequenced and characterised, with many exhibiting high structural intactness. In addition, a K locus revertant line was identified due to the unexpected presence of ALVE21, confirmed using BioNano optic maps. The ALVE identification pipeline was then applied to ninety chicken lines and 322 different ALVEs were identified, 81 % of which were novel. Overall, broilers and non-commercial chickens had a greater number of ALVEs than were found in layers.

Taken together, these two analyses have enabled a thorough characterisation of both the abundance and diversity of chicken ERVs.

Lay Summary

Retroviruses are a group of viruses which are unable to replicate themselves. Instead, during infection, these viruses insert their own genetic material into host cell DNA (the genome) where it is then replicated, producing more virus. Depending on the insertion site, these viruses can disrupt host gene function, which can lead to cancer. Furthermore, these insertions can elicit generational effects. If a retrovirus inserts within the DNA of a sperm or egg cell, the retrovirus will be passed on to the next generation as an endogenous retrovirus (ERV), and will be present in every cell of the offspring.

In mammals, approximately 10 % of the genome consists of ERVs. However, in the chicken only 1.35 % of genome is derived from ERVs, even though other birds have levels equivalent to mammals when differences in genome size are considered. Despite this apparent deficit of ERVs, a chicken-specific ERV group, the ALVEs, is known to induce tumours and inhibit productivity in commercial flocks. This PhD project sought to perform a better annotation of ERVs in the chicken and other birds, characterise ALVEs in commercial flocks, and more fully identify ALVE diversity across chickens.

A new annotation pipeline, LocaTR, was developed to identify ERVs in genome sequence. This pipeline was used to annotate ERVs in sixty-seven bird species, including chicken, and six reptile outgroups. The previously identified ERV content in chickens was almost doubled and the lineage analysis did not support the previously proposed deficit of ERVs in the chicken compared to other birds. Chicken ERV distribution and intactness was assessed to predict their effects on the host. Most ERVs were ancient and highly degraded, but 1,073 structurally intact ERVs were identified.

A further ALVE identification pipeline was developed to identify novel ALVE insertions in eight elite layer lines from Hy-Line International. Twenty different ALVEs were identified, diagnostic tests were developed, and each insert was sequenced and characterised. Many of the inserts remain highly intact supporting a continued influence on chicken biology. Analysis of ninety additional datasets identified over three hundred ALVEs, 81 % of which were novel to this study.

This PhD project has enabled a thorough characterisation of chicken ERVs, and the two pipelines are applicable to other research, such as viral-induced cancers in humans.

List of Abbreviations

ADOL	Avian Disease and Oncology Laboratory
ALV	avian leukosis virus
ALVE	avian leukosis virus subgroup E
ART-CH	avian retrotransposon in chicken
ASLV	avian sarcoma leukosis virus
BAM	binary alignment map
BASH	Bourne again shell
BEL	brown egg layers
BL	Brown Leghorn
BLASR	basic local alignment with successive refinement
BLAST	basic local alignment search tool
BLV	bovine Leukaemia virus
bp	base pairs
BWA	Burrows-Wheeler aligner
CBCS	Cot-based cloning and sequencing
cDNA	complementary deoxyribonucleic acid
ChIP	chromatin immunoprecipitation
CIGAR	compact idiosyncratic gapped alignment report
cM	centimorgans
CMAP	consensus map
CNV	copy number variant
CR1	chicken repeat 1
CRISPR	clustered regularly interspaced short palindromic repeats
CSV	chicken syncytial virus
DDBJ	DNA Data Bank of Japan
DEPS	Directional Evolution in Protein Sequences
DFE	distribution of fitness effects
DIRS	Dictyostelium intermediate repeat sequence
dN	number of non-synonymous substitutions per non-synonymous site
DNA	deoxyribonucleic acid
dS	number of synonymous substitutions per synonymous site
dsDNA	double stranded deoxyribonucleic acid
EAV	endogenous avian virus
ELISA	enzyme-linked immunosorbent assay
EMBOSS	European Molecular Biology Open Software Suite
ENA	European Nucleotide Archive
env	envelope
ERV	endogenous retrovirus
EVE	endogenous viral element
EW	Erich Wesjohann

FAO	Food and Agriculture Organisaton
FFV	feline foamy virus
gag	group-specific antigen
GB	gigabytes
Gbp	gigabase-pairs
GGERV	*Gallus gallus* endogenous retrovirus
GLIMMER	Gene Locator and Interpolated Markov Modeler
GLM	general linear model
GM	genetically modified
GTR	generalised time reversible
GyDB	Gypsy Database
HERV	human endogenous retrovirus
HGT	horizontal gene transfer
HL	Hy-Line
HMMER	Hidden Markov Modeler
HTLV	Human T-Lymphotropic Virus
ICR	internal complementary region
IGV	Interactive Genome Viewer
INT	integrase
ITR	inverted tandem repeats
JL	J-Line
K/T	Cretaceous-Tertiary
KASP	Kompetitive Allele-Specific Polymerase Chain Reaction
kb	kilobase pairs
KJF	Kees-Jan Francoijs
KoRV	Koala retrovirus
Ks	number of synonymous substitutions per synonymous site
LB	lysogeny broth
LB-amp	lysogeny broth supplemented with ampicillin
LD	LTR Digest
LH	LTR Harvest
LINE	long interspersed element
lncRNA	long non-coding ribonucleic acid
LS	LTR_STRUC
LTR	long terminal repeat
MATLAB	Matrix Laboratory
Mbp	megabase pair
MdEV	*Mus dunni* endogenous virus
MDV	Marek's disease virus
MGS	MGEScan_LTR
MMTV	mouse mammary tumour virus
mRNA	messenger ribonucleic acid

MT	methyl transferase
MUSCLE	Multiple Sequene Comparson by Log-Expectation
MYA	million years ago
NCBI	National Center for Biotechnology Information
ND	non-defined
Ne	effective population size
ng	nanogram
NGS	next generation sequencing
NLS	nuclear localisation signal
NSAC	Nova Scotia Agricultural College
ORF	open reading frame
PacBio	Pacific Biosciences
PAR	pseudoautosomal region
PBS	primer binding site
PCA	principal component analysis
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
pHMMs	profile hidden Markov models
piRNA	piwi-interacting ribonucleic acid
pol	polymerase
PPT	polypurine tract
PRO	protease
RAM	random access memory
RAxML	Randomized Axelerated Maximum Likelihood
ReTe	RetroTector
REV	reticuloendotheliosis virus
RFLP	restriction fragment length polymorphisms
RH	RNaseH
RIR	Rhode Island Red
RIW	Rhode Island White
RJF	red junglefowl
RM	RepeatMasker
RNA	ribonucelic acid
RNAi	ribonucleic acid interference
rRNA	ribosomal ribonucleic acid
RSV	Rous sarcoma virus
RT	reverse transcriptase
SAM	sequence alignment map
SIE	structurally intact elements
SINE	short interspersed element
SNP	single nucleotide polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection

SPF	specific pathogen-free
ssDNA	single stranded deoxyribonucleic acid
ssRNA	single stranded ribonucleic acid
SU	surface
SV	structural variant
TF	transcription factor
TM	transmembrane
Tm	Melting temperature
tRNA	transfer ribonucleic acid
TSD	target site duplications
TSS	transcription start site
TU	transcriptional units
TV	tumour virus
U3-R-U5	unique 3' - repeated - unique 5'
µl	microlitre
UTR	untranslated region
VLDL	very low density lipoprotein
WdSV	Walleye dermal sarcoma virus
WEL	white egg layers
WGS	whole genome (re)sequencing
WL	White Leghorn
WPR	White Plymouth Rock
YR	tyrosine recombinase

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction

Repetitive sequences in eukaryotic genomes have been widely dismissed as 'junk' DNA, however recent functional genome annotation has revealed the great abundance, diversity, and evolutionary significance of these elements. This PhD project will more completely characterise repetitive elements with a retrovirus-like structure in the chicken genome, including an assessment of recurrent retroviral integrations which detrimentally affect commercial poultry production. This chapter introduces these repetitive elements, and the chicken as a study species.

1.1 LTR retrotransposons

Eukaryotic repetitive DNA ranges from low-complexity, short nucleotide repeats, to large, replication-competent transposable elements which can move around the genome. Transposable elements are divided into two major groups depending on whether they replicate via an RNA intermediate (class I; retrotransposons) or a DNA intermediate (class II; DNA transposons). Most class II elements jump around the genome by excising and then inserting in a new location ('cut and paste'), whereas class I elements retrotranspose, producing a new element copy which then inserts at a different location ('copy and paste'), propagating the genomic content of these elements and resulting in high copy numbers. There are no cellular mechanisms for the clean excision of retrotransposons from the genome (Havecker et al. 2004; Stoye 2012).

Class I retrotransposons are divided into long terminal repeat (LTR) retrotransposons and the potentially more ancestral non-LTR retrotransposons (Flavell et al. 1997; Kordis 2005). LTR retrotransposons are a diverse group of autonomous elements (sequences which can move by themselves) abundant in eukaryotic genomes. Approximately 10 % of the mammalian genome consists of LTR retrotransposon-derived elements, but in some lineages, such as many plants (*e.g.* conifers) and amphibians (*e.g.* plethodontid salamanders), these can account for 80 – 90 % of the genomic DNA (Roth et al. 1997; Bromham 2002; McCarthy et al. 2002; McCarthy & McDonald 2004; Havecker et al. 2004; Vitte & Panaud 2005; Chaparro et al. 2007; Sun et al. 2012; Nystedt et al 2013).

LTR retrotransposons have a wide range of genomic and physiological effects, especially due to their role in insertional mutagenesis, their propensity for recombination, and the ability of some groups (such as vertebrate retroviruses) to become extracellular (Mattick et al. 2010; Aswad & Katzourakis 2012; Stoye 2012). The structure, evolutionary history, and biological effects of these elements are described below.

1.1.1 LTR retrotransposon structure

Superficially, LTR retrotransposons and non-LTR retrotransposons share a similar, virus-like architecture (Figure 1.1), containing structural and enzymatic genes, terminated by the target site duplications (TSDs) formed during genomic integration. LTR retrotransposons are clearly differentiated by the presence of the eponymous long terminal repeats. These domains, which range in length from one hundred base pairs (bp) to two kilobase pairs (kb), exhibit a shared U3-R-U5 structure (domain names explained below; page 4), consisting of conserved blocks interspersed with hypervariable regions (Conklin 1991; Benachenhou, Jern, et al. 2009).

The LTRs drive internal protein expression, even bidirectionally (Dunn et al. 2006), due to their modular arrangement of polyadenylated regions arranged into promoters and enhancers, particularly in the U3 domain. The U3-R boundary marks the transcription start site (TSS) and is preceded by a highly conserved TATA box (Benachenhou, Blikstad, et al. 2009). LTR hypervariable regions consist of transcription factor (TF) binding sites which confer host- and tissue-specific expression. In addition, the LTRs give a clear, repeated demarcation to the retroelement, which aids retention of structural integrity after retrotransposition. Non-LTR retrotransposons are commonly truncated at the 5' end, resulting in a loss of promoter activity, due to failed replication of the 5' sequence (Cordaux & Batzer 2009).

LTR retrotransposons typically contain two protein coding genes (Figure 1.1), each of which codes for multiple constituent proteins. The *gag* (group-specific antigen; core structural proteins) gene codes for the matrix, capsid and nucleocapsid proteins required for forming a retrotransposon virion. The *pol* (polymerase; enzymic proteins) gene codes for the protease (cleaves the translated retrotransposon proteins), reverse

transcriptase (produces double stranded DNA from the retrotransposon RNA intermediate), RNaseH (non-sequence-specific ribonuclease) and integrase (integrates retrotransposon DNA into the host genome) proteins required for retrotransposition.

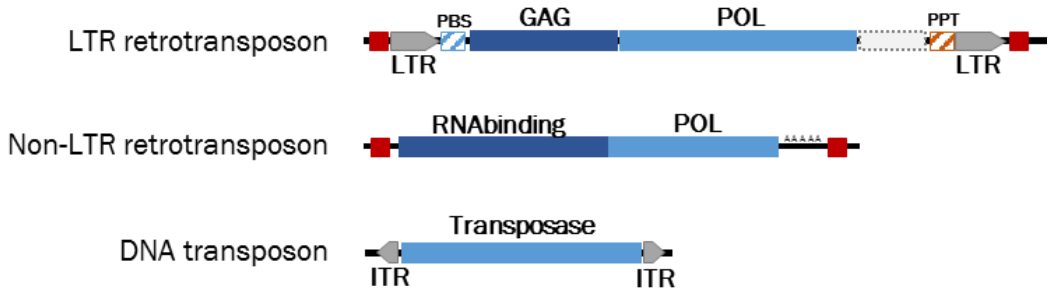


Figure 1.1 Transposon archetypal structures. LTR retrotransposons and Non-LTR retrotransposons are class one transposable elements which replicate via an RNA intermediate which is reverse transcribed to cDNA, enabling replication in a 'copy and paste' manner. These elements are demarcated by target site duplications (red boxes) and share a generally homologous enzymic *polymerase* (POL) gene. The first LTR retrotransposon gene (GAG) encodes group-specific antigens which are not homologous with the Non-LTR retrotransposon RNA binding domain. LTR retrotransposons sometimes also contain accessory genes (grey dotted box) such as the *envelope* gene found in the Retroviridae. LTR retrotransposons begin and end with the eponymous LTRs. The 5' LTR is immediately followed by the primer binding region (PBS) and the 3' LTR is immediately preceded by the polypurine tract (PPT). Non-LTR retrotransposons have a 5' promoter and 3' A-rich regions. DNA transposons are class two transposable elements which jump around the genome in a 'cut and paste' manner. They are demarcated by short, genera-specific inverted tandem repeats (ITRs).

Some LTR retrotransposon groups contain accessory genes acquired from other infectious agents or the host genome. These often provide additional functionality, and can lead to host genic dysregulation, retroelement niche expansion, or the ability of LTR retrotransposons to become extracellular through acquisition of an *env* (envelope) gene, which mediates host cell surface binding and viral entry (Malik et al. 2000; Llorens et al. 2011). Additionally, LTR retrotransposons contain a primer binding site (PBS), where a host tRNA primes reverse transcriptase, and a polypurine tract (PPT), which primes the second stage of reverse transcription (Arkhipova et al. 1986; Zhang et al. 2014).

Retrotransposition and expression

LTR retrotransposons are transcribed by host RNA polymerase starting at the TSS at the 5' LTR U3-R boundary, and terminating at the 3' LTR R-U5 boundary. This produces a transcript which begins R-U5 (repeated – unique 5' sequence) and ends U3-R (unique 3' sequence – repeated), hence the LTR domain nomenclature (Arkhipova et al. 1986; Katz & Skalka 1990). Transcripts of the negative 'template' strand will be translated at the host ribosome into retroviral proteins, and positive strand transcripts form the templates for retrotransposition. Retrotransposition results in identical LTR sequences at the point of integration, enabling the estimation of element age from LTR nucleotide divergence, based on host rates of neutral evolution (Kijima & Innan 2010).

The mechanism for retrotransposition is summarised in Figure 1.2.

Integration sites and genomic fates

The integration site greatly impacts the evolutionary success of an LTR retrotransposon, as natural selection will act to remove insertions which are detrimental to the host. This is often difficult to study accurately as the observable distribution is a result of the effects of selection (Bushman 2003). The LTR retrotransposon integrase is non-sequence-specific, but the accessory genes or domains in some groups enable specific targeting. For example, chromodomain-mediated cDNA tethering to RNA polymerase III facilitates the integration of chromovirus LTR retrotransposons (part of the Gypsy/Ty3 supergroup) near RNA genes transcribed by this polymerase (Malik & Eickbush 1999; Bushman 2003). Furthermore, integration bias largely depends on when existing LTR retrotransposons can retrotranspose (due to epigenetic silencing), and whether its cDNA and integrase (produced in the cytoplasm) can access the genome whilst the nuclear membrane is present (Desfarges & Ciuffi 2010).

Most LTR retrotransposon insertions have limited impact on the host, so are retained in the genome but degrade over many host generations. Degradation occurs through polymerase slippage (causing frameshifts or single nucleotide polymorphisms; SNPs), the random endonucleolytic action of integrase, and homologous and non-homologous recombination events (Katz & Skalka 1990; Bromham 2002). These processes erode

the archetypal LTR retrotransposon structure leading to truncations or the excision of internal domains (Figure 1.3). Commonly, the identical LTR pairs will recombine leaving 'solo LTRs' which are sixty to a hundred times more abundant in eukaryotic genomes than intact elements (Stoye 2001).

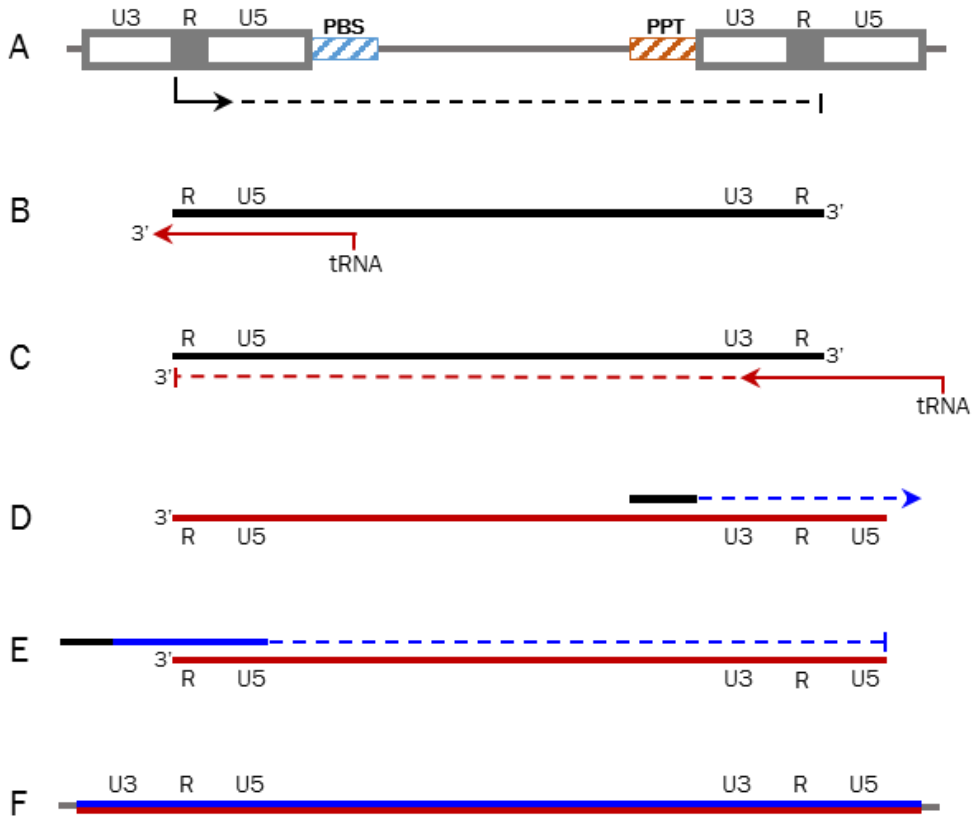


Figure 1.2 Mechanism for LTR retrotransposon expression and retrotransposition. A). An intact, replication competent LTR retrotransposon hijacks host cell replicative machinery to express a positive single strand RNA (ssRNA) copy from the transcription start site (TSS) located at the U3-R boundary in the 5' LTR. Transcription extends to the 3' LTR R-U5 boundary. Similar transcription to produce a negative ssRNA enables expression and translation of the internal coding regions. B). A host tRNA binds to the primer binding region (PBS) of the positive ssRNA and reverse transcriptase (RT) produces complementary single strand DNA (ssDNA). C). After reaching the end of the ssRNA template, the ssDNA and RT lifts off and binds to the 3' end due to the shared LTR R region homology, and RT continues. D) RNaseH activity breaks down the ssRNA template to leave the polypurine tract (PPT), which acts as a primer for RT synthesis of the ssDNA complementary strand. E). In the same manner as part C, the RT lifts off to synthesise the remainder of the complementary strand, creating the double strand DNA ready for genomic insertion via the action of integrase (F). Figure adapted from Arkhipova et al. 1986 and Zhang et al. 2014.

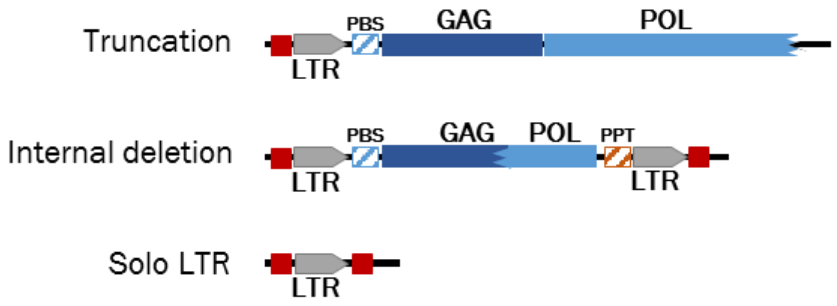


Figure 1.3 LTR retrotransposon degradation. Degradation can be a gradual process with coding domains gradually acquiring frameshift and nonsense mutations, or degradation can be accelerated due to recombination. This can lead to truncation, removing the ends of elements, or internal deletion due to non-homologous recombination. In many cases the paired LTRs recombine, excising the internal coding regions leaving a 'solo LTR'.

Cellular regulation of LTR retrotransposons

Most LTR retrotransposons degrade relatively rapidly in evolutionary timescales, but their effects immediately after insertion can be varied and unpredictable (section 1.1.3). In addition, over short timescales uncontrolled retrotransposition and insertional mutagenesis creates persistent but varied challenges for the host, so it is advantageous to control retrotransposon activity. Eukaryotic cells do this through two core processes: epigenetic silencing by sequence methylation and histone modifications, and innate antiviral defence (Stoye 2012).

DNA base methylation in eukaryotes occurs almost exclusively at cytosine residues and is used in transcriptional regulation through the silencing of promoter and enhancer regions. In a similar manner, cells target transposable elements to be methylated to silence their activity (Weber & Schübeler 2007). Methylation is observed across LTR retrotransposon sequences, but particularly in the LTRs themselves due to the presence of promoters and enhancers. The LTR U3 domain is commonly the most methylated region as this is where cellular machinery and transcription factors bind (Cam et al. 2008; Reiss et al. 2010). The hypervariable nature of the U3 domain across LTR retrotransposons likely represents an evolutionary response to epigenetic targeting (Benachenhou et al. 2013; Goodier 2016).

In addition to the action of host cell methyl transferases, targeted methylation of retrotransposons is facilitated by co-evolved mechanisms. Cam and colleagues (2008) identified that complementary binding of expressed retrotransposon-derived sequences physically recruited methylation machinery to integration sites, and a family of zinc-finger proteins with LTR retrotransposon-sequence-specificity were required for identifying and methylating non-promoter-like regions (Rowe et al. 2013). Cytosine methylation is a heritable control, and is proactively regulated in animal genomes as methylation marks are reset after any cell division. The high copy number and recurrent LTR retrotransposon activity in many plants, may be due, at least in part, to the absence of these post-replication checks, as 'reactivated' elements may be able to retrotranspose for multiple generations until their epigenetic marks are reset (Rigal & Mathieu 2011).

Sequence-level silencing can account for control in adult cells, but not during early development when methylation is stripped. Most early development epigenetic modification studies have only been completed in mammalian model systems, but recent work has shown the importance of histone marks and modifications in limiting transposable element activity (Rowe & Trono 2011; Leung & Lorincz 2012). The heritable relevance of histone modifications has also been shown in plants, where deacetylase and deubiquitination mutants in *Arabidopsis* were shown to exhibit higher incidence of transposable element insertions (Rigal & Mathieu 2011). Despite histone-induced regulation, transposable element activity is still significantly elevated during early development (Bromham 2002; Cohen et al. 2009; Faulkner et al. 2009; Reiss et al. 2010; Feschotte & Gilbert 2012).

Epigenetic marks are the major, heritable mechanism for transposable element control. However, these marks are easily modified and will often change during times of cellular stress or dysregulation, resulting in increased transposable element activity. Additionally, recent insertions will likely be incompletely methylated (if at all), leading to partial element expression (Reiss & Mager 2007; Varriale 2014). Innate antiviral defence mechanisms provide secondary, cumulative control of these elements. This can include a range of host viral inhibitors (often based on detecting RNA or DNA intermediaries), restriction enzymes, or more complex pathways such as RNA interference (RNAi) (Rigal & Mathieu 2011; Chung et al. 2014). RNAi can specifically target retrotransposon transcripts which have been previously expressed, as short sequences (approximately 20

to 40 bp) are retained as references which can be used to bind new transcripts and mark them for degradation. Sequences with close homology will also be targeted as exact base pair complementarity is not required (Rigal & Mathieu 2011; Chung et al. 2014). Whilst these mechanisms target expressed retrotransposons, enzymes such as cytidine deaminases target existing genomic insertions and mutate the polyadenylated regions, disrupting promoter and enhancer activity (Chiu & Greene 2008). In addition, LTR retrotransposons themselves are known to influence the expression, replication and transmission of other LTR retrotransposons (introduced further in section 1.1.3).

1.1.2 Evolutionary origins

LTR retrotransposons have been identified throughout the eukaryotic lineage, supporting their presence in the last eukaryotic common ancestor (Llorens et al. 2008; Llorens et al. 2011). These elements are absent in studied Prokaryotes, and no LTR retrotransposon precursors have been identified. However, both DNA transposons and non-LTR retrotransposons are found in prokaryotes, and it is likely that LTR retrotransposons are chimeric elements formed from recombination between elements of these two ancestral classes (Malik & Eickbush 2001; Boeke 2003). There has been some suggestion that the uncoupling of transcription and translation in eukaryotic cells, and the presence of a nuclear membrane barrier, made the two stage retrotransposition of LTR retrotransposons more successful than other transposon classes (Malik & Eickbush 2001).

In addition to their chimeric origin, the 'superficially conserved' internal coding sequences of LTR retrotransposons also appear to have complex evolutionary histories, with multiple independent acquisitions from various viral sources in different lineages (Malik et al. 2000; Malik & Eickbush 2001; Peterson-Burch & Voytas 2002; Havecker et al. 2004; Capy 2005; Benachenhou, Blikstad, et al. 2009). Consequently, lineages constructed from separate domains produce different group relatedness, but the use of the highly conserved *reverse transcriptase* (RT) gene (required at all times for retrotransposition) has become the field standard for retrotransposon classification and phylogeny construction (Jern et al. 2005; Llorens et al. 2008).

Phylogenies constructed from RT divide up the LTR retrotransposons into five distinct groups (Figure 1.4): Bel/Pao (Semotiviridae), Copia/Ty1 (Pseudoviridae), DIRS (*Dictyostelium* intermediate repeat sequence), Gypsy/Ty3 (Metaviridae) and the retroviruses (Retroviridae). LTR retrotransposons across the five groups are typically 4 – 10 kb in length, but can elongate considerably depending on the presence of accessory genes. The Bel/Pao, Copia, Gypsy and retroviruses all share a similar version of the archetypal LTR retrotransposon structure (Figure 1.5). In contrast, the DIRS elements exhibit a highly divergent structure lacking, most strikingly, the eponymous LTRs. These elements phylogenetically group within LTR retrotransposons due to RT homology, but have inverted terminal repeats (ITRs) and have acquired a *methyl transferase* and *tyrosine recombinase* rather than the typical *protease* and *integrase* (Goodwin & Poulter 2001; Poulter & Goodwin 2005; Piednoël et al. 2011).

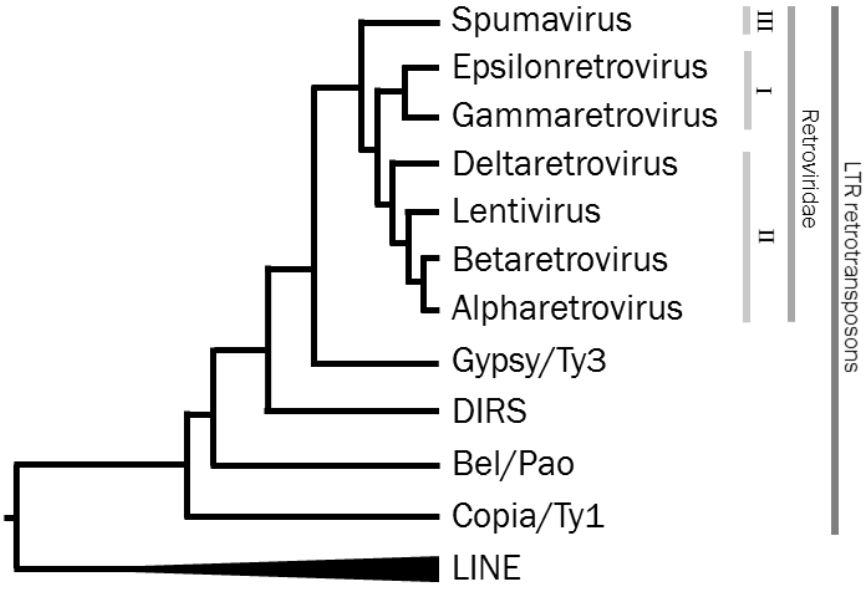


Figure 1.4 LTR retrotransposon cladogram based on *reverse transcriptase* (RT) with the non-LTR retrotransposon LINE family as outgroup. The Retroviridae are divided into seven genera, and have historically been grouped into three classes labelled with Roman numerals based on human retrovirus classification (I, II and III). Despite their divergent structure (Figure 1.5), DIRS elements sit between the Metaviridae (Gypsy/Ty3) and Semotiviridae (Bel/Pao). The Pseudoviridae (Copia/Ty1) are the most basal LTR retrotransposon lineage. The broad LINE branch represents extensive non-LTR retrotransposon diversity. Cladogram constructed based on Poulter & Goodwin 2005; Llorens et al. 2008; Piednoël et al. 2011; Llorens et al. 2011; Benachenhou et al. 2013.

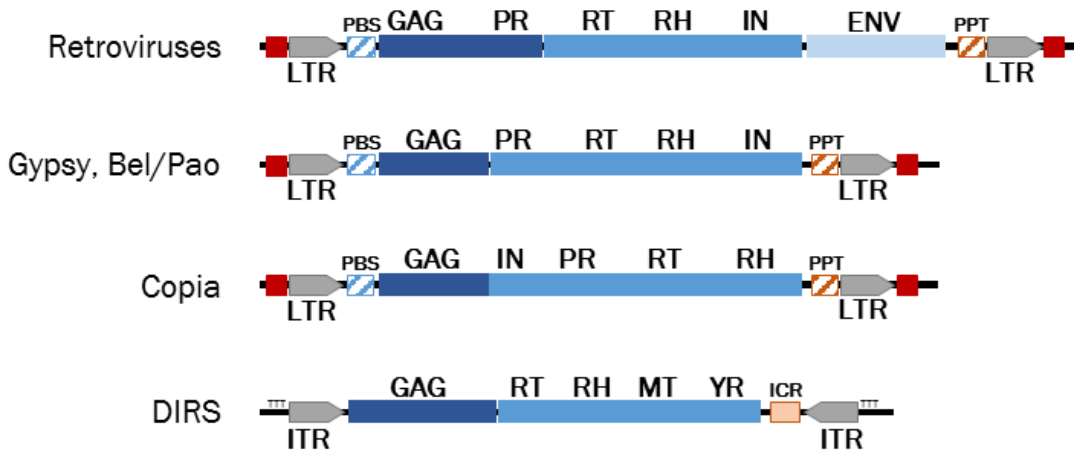


Figure 1.5 LTR retrotransposon archetypal structures. The Retroviruses, Gypsy, Bel/Pao and Copia groups have the standard LTR retrotransposon structure with a pair of LTRs terminally demarked by a target site duplication (red box). The 5' LTR finishes with the primer binding site (PBS) and the 3' LTR is preceded by the polypurine tract (PPT). The internal sequences are the *gag* (group specific antigens; dark blue), and *polymerase* (blue) including RT (*reverse transcriptase*), RH (*RNaseH*), IN (*integrase*) and PR (*protease*). In the retroviruses, Bel/Pao and Gypsy the *polymerase* order is PR-RT-RH-IN, but the more ancient Copia group has IN first. In retroviruses, the PR is encoded as part of the *gag* gene leading to higher *protease* expression. Retroviruses typically have an *envelope* gene (light blue) which encodes the virion proteins required for an extracellular lifecycle. Some Gypsy and Copia elements have independently acquired *envelope* genes, but these are exceptions. DIRS retrotransposons have a highly divergent structure with inverted tandem repeats (ITRs) rather than LTRs, although these are much longer and more complex than those in DNA transposons (Figure 1.1). DIRS elements are demarcated by the trinucleotide TTT and the ICR (internal complementary region) has homology with both ITRs, facilitating DIRS retrotransposition via a circular double stranded DNA (dsDNA) intermediate. DIRS elements have a typical *gag*, but use a methyl transferase (MT) and tyrosine recombinase (YR) rather than the typical protease and integrase.

Eukaryotic distribution

Copia and DIRS elements have the widest eukaryotic distribution, with representation in single-cell eukaryotes such as the Diatoms (Piednoël et al. 2011; Benachenhou et al. 2013). The diverse Gypsy group also diverged before the last common ancestor of plants, animals and fungi, but the Bel/Pao clade is limited to metazoans, and retroviruses are limited to vertebrates (Llorens et al. 2009; Llorens et al. 2011).

Despite these general distributions there is substantial lineage specificity, with the total loss of some LTR retrotransposon groups at various taxonomic levels (Bromham 2002; Peterson-Burch & Voytas 2002; Havecker et al. 2004; Kordis 2005; Volff 2009). For example, DIRS elements are absent in mammals and birds despite their presence in other vertebrate groups (Piednoël et al. 2011). The abundance and diversity of particular lineages is largely dependent on stochastic processes influenced by genomic expansion and contraction, changes in effective population size, inter-element recombination, and infections from exogenous viruses (Flavell et al. 1997; Katzourakis et al. 2005; Weiss 2006; Kanda et al. 2013).

Retroviruses – acquisition of an envelope

Technically, the term 'retrovirus' can be used to describe any retrotransposon which can escape a host cell and infect other cells with successful integrations (Llorens et al. 2011). Practically, when this term is used it refers to vertebrate retroviruses of the Retroviridae family. This distinction is important as multiple LTR retrotransposon lineages have acquired *env* genes enabling them to become exogenous viruses (Malik et al. 2000; Peterson-Burch & Voytas 2002). Interestingly, a number of these do not appear to have then become infectious agents, suggesting that they instead facilitate virion chaperoning during retrotransposition when single stranded viral RNA or DNA would be targeted for degradation by the host cell (Havecker et al. 2004). However, the true benefit of this to the retrotransposon is unknown, as recent work has suggested that elements without an *env* proliferate to a greater extent (Magiorkinis et al. 2012).

The Retroviridae (hereafter simply referred to as 'retroviruses') originated from within the Metaviridae (Gypsy/Ty3). There is conflicting evidence from the different coding regions (confounded by possible convergence and horizontal gene transfer; HGT) as to whether retroviruses are monophyletic or whether the three subclasses (class I, II and III; Figure 1.4) arose independently (Llorens et al. 2008; Benachenhou et al. 2013). The defining feature of all elements in this group is the presence of an *env* gene. This enables retroviruses to replicate in a host cell, exit the cell, and then fuse with a target cell and integrate into its genome, propagating horizontally. This has led to examples of cross-species niche expansion, such as the ongoing Koala retrovirus (KoRV) infection which

originated from a gibbon gammaretrovirus (Tarlinton et al. 2006), or the mammalian gammaretrovirus REV (reticuloendotheliosis virus) which now infects chickens and turkeys (Payne 1998).

Importantly, as retroviruses still integrate into the host genome, if integration occurs within the germline the retrovirus sequence will become part of the inherited genome, and have an intracellular lifecycle functionally indistinguishable from other LTR retrotransposons (Magiorkinis et al. 2012). These endogenous retroviruses (ERVs) are very common across vertebrates, although this 'fossil record' of retroviral integrations is very uneven. Endogenous lentiviruses were first identified in 2007 in the European rabbit (*Oryctolagus cuniculus*) and have since been identified in two lemur species (Katzourakis et al. 2007; Gilbert et al. 2009), and an endogenous deltaretrovirus was discovered for the first time earlier this year in the Natal long fingered bat (*Miniopterus natalensis*) (Farkašová et al. 2017). Limited identification of some groups is due to a combination of retrovirus propensity for infecting germline cells, the likelihood of recent infection (enabling identification), and the analysis of appropriate genomes. A limited number of different endogenous alpharetroviruses have been described as they are limited to avian genomes having evolved from avian betaretroviruses (Bolisetty et al. 2012). In addition, endogenization events are rare, with most ERV genomic copy number variants (CNVs) due to intracellular retrotransposition following the initial integration event (Bock & Stoye 2000; Katzourakis et al. 2005). Consequently, if novel ERVs are either degraded on arrival, or before retrotransposition occurs, they are unlikely to be well represented in a genome.

As most LTR retrotransposons are solely intracellular elements, ERVs pose unique challenges for host cells. Novel infections without vertical inheritance enable exogenous retroviruses to evolve at rates up to six orders of magnitude faster than in vertebrate genomes, as well as horizontally introduce novel accessory genes (including detrimental oncogenes or beneficial additions) from other vertebrate hosts and their parasites, or other viruses. Persistent infection also means that there are young, structurally intact genomic integrations, in locations potentially detrimental to the host, as there has been insufficient time for removal by selection, epigenetic silencing or natural decay (Doolittle & Feng 1989; Stoye 2001; Reiss & Mager 2007; Rigal & Mathieu 2011; Kanda et al. 2013).

In many cases, vertebrate ERVs do not correspond to current exogenous retroviral infections, but rather represent a record of historical infections and provide an answer to how viruses which evolve so quickly could have been around for so many millions of years (Doolittle & Feng 1989). Some ERVs re-emerge from the genome, likely through recombination with unrelated viruses, to pose novel exogenous threats, such as Avian Leukosis Virus subgroup J in chickens (Doolittle & Feng 1989; Venugopal 1999; Katzourakis et al. 2005; Kanda et al. 2013).

1.1.3 Genomic and physiological impacts of LTR retrotransposons

The co-option and maintenance of LTR retrotransposon-derived elements is rare, particularly over large evolutionary timescales. Highly deleterious integrations will either be removed from a population, or be rapidly targeted for epigenetic silencing and will eventually degrade. Any maintained insertion must either be selectively advantageous and its further replication controlled, or in a genomic location where maintenance is possible, such as in poorly recombining regions or regions in linkage with genes under selection (Gogvadze & Buzdin 2009; Stoye 2012; Magiorkinis et al. 2013). Novel LTR retrotransposon integrations can elicit a wide range of impacts, some of which are relevant for disease or productivity traits. Depending on the host species, the LTR retrotransposons which have most recurrent impact will differ due to lineage specificity and individual element intactness.

Impact on genome stability and host gene functionality

LTR retrotransposons can comprise large proportions of the genome, and in some species (such as plethodontid salamanders) individual elements exist in over a million copies (Chaparro et al. 2007; Sun et al. 2012). This can elicit metabolic stress on the host due to the large quantity of DNA which needs to be copied at each cell division, and physiological stress as cell size needs to increase to cope with a bloated nucleus (Roth et al. 1997; Cavalier-Smith 2005). Rates of retrotransposition and deletion are generally balanced, but periods of elevated or reduced activity can lead to lineage-specific element complements, and can affect the overall genome size (Promislow et al. 1999).

In general, LTR retrotransposons (and other transposable elements) follow the Pareto principle, where 20 % of elements account for 80 % of the total abundance (Magiorkinis et al. 2012). Consequently, there are many sites which could facilitate inter- and intra-chromosomal recombination, leading to chromosome fission and fusion events, and sequence deletion. Such events are likely to be highly deleterious for the host, so observed recombination between LTR retrotransposon loci is rare (Hughes & Coffin 2001). However, elevated transposable element activity in cancer genomes has been identified as a cause of genomic instability (Romanish et al. 2010). Recombination events can also facilitate sequence duplication (including entire gene blocks) (Cusack & Wolfe 2007; Langille & Clark 2007; Dorus et al. 2008; Han et al. 2009), and the action of LTR retrotransposon reverse transcriptase and integrase on host gene mRNA can cause the formation of retrogenes, which present opportunities for rapid host evolution (Bromham 2002; Kaessmann et al. 2009; Mattick et al. 2010).

LTRs themselves have been well documented as providing alternative, or tissue-specific promoters to host genes (Meisler & Ting 1993; Dunn et al. 2005; Dunn et al. 2006; Romanish et al. 2007; Cohen et al. 2009; Jacques et al. 2013; Z. Wang et al. 2013; Wragg et al. 2013). Interestingly, the influence of these mobile promoters can be felt from large up- and down-stream distances, making the bioinformatic prediction of LTR impacts on gene expression difficult (Li et al. 2012). Integrations within gene introns are much more common than those near (within 10 kb) upstream promoters and the transcription start site, and are known to disrupt splice sites which can cause or limit alternative transcripts, sometimes in a tissue-specific manner (Chang et al. 2006; Mattick et al. 2010; Isbel & Whitelaw 2012; Stoye 2012). Whilst the likely effects of such integrations can be predicted bioinformatically, the extent and range of the effects are yet to be fully studied due to the difficulty in assessing alternative splicing from short read next generation sequencing (NGS) data (Martin & Wang 2011; Ozsolak & Milos 2011).

The diverse roles of retroviral proteins

In some rare cases, LTR retrotransposon insertions have been co-opted by the genome and are now host genes with a range of cellular functions (Lynch & Tristem 2003; Youngson et al. 2005; Ono et al. 2006; Volff 2006; Sekita et al. 2008; Carré-Eusèbe et

al. 2009; Marco & Marín 2009; Volff 2009). The best studied of these are the mammalian *syncytin* genes, which are derived from an ERV *envelope* gene. *Syncytins* are crucial for the formation and maintenance of the placenta during pregnancy, as the envelope protein is expressed on the surface of cells enabling cell-cell interaction and fusion, potentially with a secondary immunosuppressive role which prevents foetal rejection (Mi et al. 2000; Gong et al. 2005; Dupressoir et al. 2009). Interestingly, *syncytin* genes arose independently in multiple mammalian lineages from distinct, but related, ERVs (Heidmann et al. 2009; Chuong et al. 2013; Lavialle et al. 2013).

For an LTR retrotransposon insertion to be truly co-opted, its replication needs to have been controlled and its detrimental effects reduced or eradicated (Stoye 2012). The continued expression of evolutionarily recent ERV insertions in vertebrate genomes has been shown to elicit a prolonged immunological burden on the host, mainly due to the production of replication competent virus which induces persistent viremia (Aswad & Katzourakis 2012). However, expression of endogenous retroviral proteins has also been shown to confer resistance to exogenous retroviruses (part of the wider endogenous viral element derived immunity), particularly if they are closely related. Consequently, such elements provide a selective advantage during recurrent retroviral infections, but are strongly selected against once that retrovirus is no longer a threat (Katzourakis & Gifford 2010; Aswad & Katzourakis 2012; Patel et al. 2012; Hurst & Magiorkinis 2014).

ERVs can mediate exogenous retroviral infection in two main ways. Firstly, production of envelope protein can inhibit exogenous retrovirus infection through receptor interference. The endogenous envelope acts as a competitive inhibitor by physically blocking the cellular receptors hijacked by infecting viruses. Examples include the mammalian gammaretroviral *Fv4* and *Rcmf*, and chicken alpharetroviral ERVs such as ALVE6 (Robinson et al. 1981; Smith et al. 1991; Varela et al. 2009; Ito et al. 2013; Kozak 2014). In some cases the cellular receptors have become mutated to block viral entry, but this can elicit a broad range of negative phenotypic effects on the host (Lepperdinger et al. 2001; Jadin et al. 2008). Secondly, whilst production of gag proteins is generally detrimental to the host (Astrin & Robinson 1979; Robinson et al. 1981), these proteins can also have an inhibitory effect on retroviral uncoating and reassembly, as well as halting nuclear transport and targeting viral RNA for degradation by forming double stranded hybrid sequences (Aswad & Katzourakis 2012). In hosts with multiple,

related ERVs which exhibit different levels of intactness, the cumulative effect of these elements on retroviral defence is complex and often finely balanced. In addition, whilst some expressed ERVs can protect host cells, the presence of ERV transcripts increases the likelihood of forming recombinant viruses (Venugopal 1999; Liu et al. 2011; Henzy et al. 2014).

1.2 The study of LTR retrotransposons in the chicken

Exogenous and endogenous retroviruses have been intensively studied in the chicken, particularly Avian Leukosis Viruses (ALVs; also known as Avian Sarcoma Leukosis Viruses, ASLVs) due to their diverse effects, including the modulation of retroviral infection dynamics, depression of commercial performance traits, and formation of tumours (Gavora et al. 1991; Ka et al. 2009; Payne & Nair 2012). However, the wider LTR retrotransposon content of the genome has yet to be completely described.

A comprehensive evaluation is required as the chicken is of great agricultural and economic importance, as well as being an important developmental, experimental and disease model. The Food and Agriculture Organisation (FAO) reports that over fifty billion broiler chickens are raised annually, producing over one hundred million tonnes of meat (2015 data). In addition, at least six billion layer chickens produce over one trillion eggs every year. Since 2000 consumption of chicken meat has increased by 35 %. However, this hides the increasing demand of developing countries, which by 2030 will consume more than double the quantity in 2000. Greater understanding of LTR retrotransposons, as genomic elements which limit productivity and detrimentally affect animal welfare, is therefore of vital importance to global food security.

Chicken domestication

Domestication of the red junglefowl (RJF; *Gallus gallus*) began six to eight thousand years ago, with multiple origins, back crosses and at least one hybridisation event with the grey junglefowl (*G. sonneratii*) (Rubin et al. 2010). Following this complex domestication process, the domestic chicken (*G. gallus domesticus*) and RJF have

retained high genetic identity and remain classified as the same species. Furthermore, it was an RJF individual which was sequenced for the chicken reference genome, in an attempt to avoid any genetic changes from the domestication process (Hillier et al. 2004). This was the first bird and first agricultural animal to be sequenced, and its phylogenetic location within the Galloanserae has enabled it to be particularly informative in the study of other commercially relevant bird species (Figure 1.6). However, the RJF individual sequenced was from an American zoo in Hawaii (originally from Malaysian stock), where there had potentially been crosses with domesticated chickens. Whilst Hillier and colleagues (2004) acknowledged this risk, the chosen bird was phenotypically and physiologically a typical RJF, including the seasonal laying of small, brown eggs. However, recent work has revealed significant genetic introgression in this individual from the domesticated White Leghorn (WL) breed (Ulfah et al. 2016), which likely means the reference genome does not truly represent the chicken's wild ancestor.

Intensive commercialisation

Traditionally, a diverse variety of chicken pure breeds and hybrids were kept in small flocks for their meat and eggs, with some breeds developed for cock-fighting or exhibition. Following the Second World War, large scale commercialisation focused on breeding specialised broiler (meat) and layer (egg) lines to overcome the observed genetic conflicts between production (growth rate) and reproduction (egg production) (Muir et al. 2008).

Despite the hundreds of well-characterised chicken breeds (Ekarius 2007; Roberts 2009), commercial stock has been derived from just a handful. Commercial layer lines are divided by egg colour. White egg layers (WELs) were largely derived from WLs, and brown egg layers (BELs) were developed from North American and European dual purpose breeds such as Plymouth Rock, Rhode Island Red and New Hampshire. For broilers, distinct lines were created for the sire (paternal) and dam (maternal) birds, again to avoid the genetic conflict between production and reproduction. The broiler dam line has an origin similar to the BELs, but the sire line was derived almost completely from the Cornish breed (British Cornish Indian Game breed) due to its compact body size and high proportion of breast muscle (Muir et al. 2008).

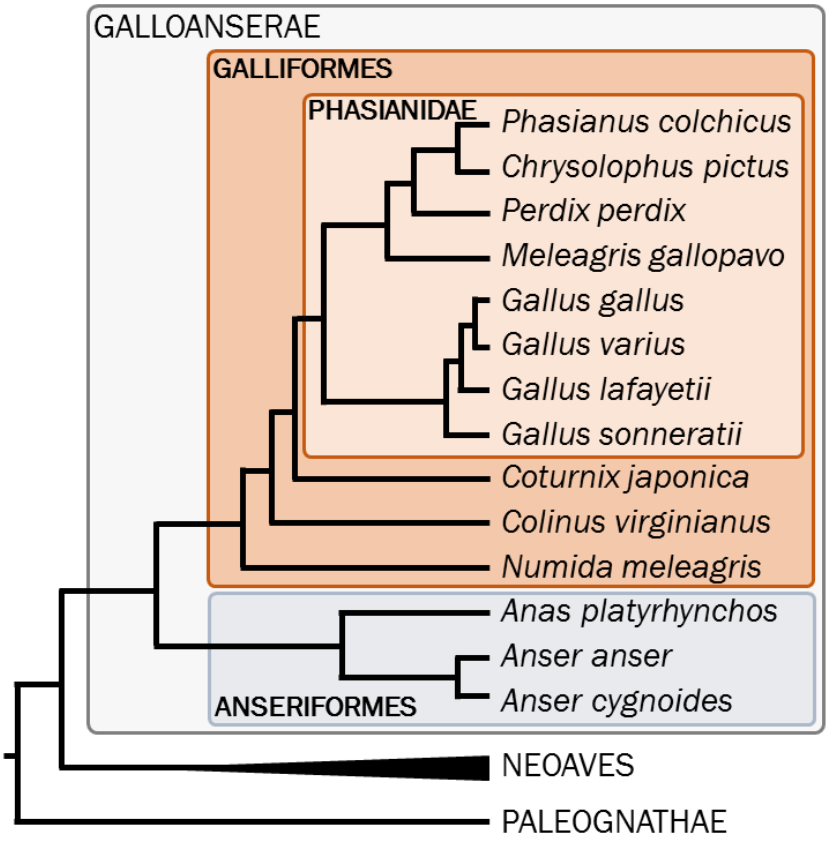


Figure 1.6 Galloanserae cladogram showing commercially relevant species and their general relationship to the chicken, including those within the Phasianidae pheasant family. The two Galloanserae orders are the land (Galliformes) and water (Anseriformes) fowl. The common names for the represented Galloanserae species are, from the top: common pheasant, golden pheasant, grey partridge, turkey, red junglefowl (incl. domestic chicken), green junglefowl, Sri Lankan junglefowl, grey junglefowl, Japanese quail (Old World quail), Northern bobwhite (New World quail), Guinea fowl, duck, greylag goose, and swan goose (also known as the Chinese goose). The broad branch leading to the Neoaves represents this group's large phylogenetic diversity. Cladogram adapted from Kan *et al.* (2010) and Meiklejohn *et al.* (2014).

These commercial programmes have proven particularly effective in broilers, increasing growth rate while decreasing food conversion rates, so that meat can be produced today for a lower price (in real terms) than it was in the 1950s (Havenstein et al. 2003). Havenstein and colleagues identified that 85 – 90 % of this improvement was due to selectively bred genetic modifications, with improvements in husbandry accounting for the remainder. Similarly, selective breeding in layers has augmented lifetime egg production, continues to increase total egg counts, and has reduced required feed intake

by 30 %, all whilst breeding for commercially desirable uniformity in egg size, shape and colour (McKay 2009).

Such intensive phenotypic manipulation has had genomic impacts, with decreased productivity-gene allelic diversity following hard selective sweeps (Rubin et al. 2010), and variation in DNA methylation patterning across coding and regulatory regions. Studies of the latter (Nätt et al. 2012; Hu et al. 2013) suggest that rapid responses to breeding programmes may be plastic epigenetic effects rather than 'permanent' nucleotide changes, a notion supported by the paucity of coding region deletions in intra-line comparisons, and the maintenance of high nucleotide diversity (Rubin et al. 2010).

These findings were welcome news to breeders, as continued commercial improvement requires genetic diversity, and traditional selective breeding methods often result in high inbreeding coefficients within lines (Bacon et al. 2000; Charlesworth 2009). In chickens, these effects are due to the limited numbers of breeds used in the initial development of commercial lines, use of within-line selection (where a limited number of individuals derive a large proportion of the next generation) and the knock-on effect of intense market competition, where commercial stock has been reduced to a limited number of lines in a few multinational companies (Muir et al. 2008). Maintenance of high nucleotide diversity is likely testament to the recent, complex domestication and the large ancestral effective population size (Ellegren 2005).

1.2.1 Chicken genome structure and sequence

Compared to mammalian genomes, the 1.2 gigabase-pair (Gbp) chicken genome is small but compact: approximately one third the length of the human genome but with a similar number of annotated protein-coding genes (Hillier et al. 2004). Chicken genomic DNA is arranged over thirty-eight autosomes and a pair of sex chromosomes, giving a diploid number of seventy-eight. The autosomes exhibit length variation across two orders of magnitude, with chromosomes 1 to 10 classified as macrochromosomes (although the literature sometimes refers to chromosomes 6 to 10 as 'intermediate' chromosomes), and 11 to 38 as microchromosomes. Microchromosomes exhibit higher GC content, a higher density of genes and CpG islands, elevated rates of synonymous

substitutions (Ks), lower repetitive DNA content, and reduced intronic and intergenic distances compared with the macrochromosomes (Ellegren 2005). Additionally, microchromosomes exhibit elevated recombination rates due to the obligate meiotic chromosomal crossovers which occur between all homologous chromosome pairs irrespective of their short lengths. This has the effect of elongating the chicken linkage map to 4,000 centimorgans (cM), 300 cM longer than the human linkage map, despite the three-fold difference in genome length (Dawson et al. 2007). Across the genome, the neutral nucleotide substitution rates are very similar between chicken (and birds more widely) and mammals (Helm-Bychowski & Wilson 1986; Hedges et al. 1996).

The avian sex chromosomes (W and Z) began to develop from an autosomal pair approximately 150 million years ago (MYA), and have an origin independent from the sex chromosomes of the mammalian or other reptilian lineages (Fridolfsson et al. 1998; Matsubara et al. 2006; Nam & Ellegren 2008). Unlike mammals, in birds the female is the heterogametic sex (WZ) and the male is homogametic (ZZ). The Z chromosome evolves faster than the W (ratio estimates vary between 1.7 to 6.5 times faster), exhibiting the "Faster-Z effect", which has been documented across a wide range of species (Meisel & Connallon 2013). In chickens, this effect is not due to a gender-specific mutation bias, but rather the ineffectiveness of selection on the Z chromosome, enabling mildly deleterious mutations to become fixed through genetic drift (Axelsson et al. 2004; Bergero & Charlesworth 2009; Wright et al. 2015).

Chicken genomic resources

The chicken reference genome has undergone several revisions since the publication of the first draft in 2004. At the beginning of this project version 4.0 (Gallus_gallus4.0; Galgal4) was majorly revised to correct errors on the Z chromosome, but still lacked chromosome-level assemblies for nine microchromosomes (29-31, 33-38) and the assemblies for chromosomes 16, 25, 32 and W were far shorter than their known lengths. The total assembly length was 1.05 Gbp, suggesting that at least 200 Mb of the genome was yet unsequenced (Gregory 2017). However, the chicken genome is constantly being improved and during this project the new Galgal5 assembly was produced (Warren et al. 2017), and the release of Galgal6 is predicted for next year.

Such active development is both a blessing and a hindrance for the researcher. The genome sequence is constantly improving and better reflecting true chicken biology, but the release of three assemblies in seven years means that annotated genomic coordinates and gene feature models are constantly changing. In addition, the knowledge that a new assembly is just around the corner has made genome database organisations, such as NCBI or Ensembl, loath to rerun their annotation pipelines. Despite this, the chicken genome annotation is one of the best outside mouse and human, due to considerable research effort and the availability of NGS data across a wide range of tissues and developmental stages. Improvements to both genome and annotation quality during the course of this project have been facilitated by the development of long read sequencing technologies and high resolution optic mapping (Kuo et al. 2017; Warren et al. 2017). In addition, study of chicken breed variation has been aided by the development of a high density 600K SNP chip (Kranis et al. 2013).

Genome stability

The 2n = 78 chicken karyotype is very representative of the entire avian lineage, where the karyotype is largely confined to a 2n = 76 to 80 range, with variation due to microchromosome fusion events. Similar genome organisation is also found in the distantly related Lepidosauromorpha reptilian lineage, suggesting that the avian karyotype highly resembles that of the amniote ancestor 310 MYA (reviewed by Ellegren 2010). Such chromosomal stability is mediated, at least in part, by the hexanucleotide tandem repeats near the telomeres, which are ten times more abundant in the chicken compared with mammals, and are present on all chromosomes in active configurations (Delany et al. 2003). Recombination rates are also elevated near the telomeres relative to the chromosomal average, potentially generating the high frequency of tandem repeats responsible for protecting the chromosome ends. Similar findings were later identified in the zebra finch (*Taeniopygia guttata*, 2n = 80) to a greater extent, supporting the role of telomeric DNA in avian chromosomal stability (Backström et al. 2010).

Chicken genome stability goes beyond the highly conserved karyotype, as avian genomes exhibit extended syntenic blocks, even between phylogenetically distant species (Ellegren 2010). Synteny is generally maintained by the lower rates of recombination

observed away from the telomeres, with internal macrochromosome regions exhibiting rates as low as 0.1 cM/Mb compared to greater than 2.5 cM/Mb near the telomeres (Delany et al. 2003). Elferink and colleagues (2010) further categorised the heterogeneity of recombination rates, identifying regions effectively devoid of recombination, even on the generally highly recombining microchromosomes. Concordantly, observed intrachromosomal rearrangements are also rare, occurring ten times less frequently than in mammals (Feuk et al. 2006; Pontius et al. 2007). Furthermore, repetitive elements are known to facilitate chromosomal rearrangements in many species by non-homologous recombination (Feuk et al. 2006; Aguilera & Gómez-González 2008; Carbone et al. 2009), but such effects are limited in the chicken and other galliform birds due to the paucity of annotated repetitive elements in the genome (Griffin et al. 2008).

Genomic repeat content

Annotation of the original chicken genome draft sequence identified a significantly lower proportion of the genome as interspersed repeats (approximately 10 %) compared with mammalian genomes (typically 40 – 50 %) (Hillier et al. 2004). This was later shown to be a more widespread avian phenomenon with similar levels identified in the genomes of other fowl such as the duck (*Anas platyrhynchos*) and turkey (*Meleagris gallopavo*), as well as in the higher order passerines such as the budgerigar (*Melopsittacus undulatus*), collared flycatcher (*Ficedula albicollis*) and zebra finch (Dalloul et al. 2010; Warren et al. 2010; Huang et al. 2013; Kawakami et al. 2014; Ganapathy et al. 2014). Additionally, the most abundant chicken repetitive element, Chicken Repeat 1 (CR1; a LINE family element), is present in approximately two hundred thousand copies, significantly fewer than the over one million L1 copies in mammalian genomes, even when scaled for the three-fold difference in genome size (Ellegren 2005). These initial draft genomes also began to show that whilst overall repeat content was similar between birds, the represented repeat families and their relative proportions were highly lineage specific. It was not until after the start of this project, with the release of further genomes, that this could be studied more completely, including through the analysis presented here in Chapter 4. Subsequent work on avian genomic repeat content has been reviewed in the introduction to that chapter (page 5).

This observed deficit of repetitive elements in avian genomes compared with mammals is at least partially responsible for their observed compactness. Such a deficit could be due to limited numbers of new retrotransposition events, and the commonly observed truncation of the 5' CR1 promoter region in avian genomes could have limited expansion of this group (Ellegren 2005). However, as avian genomes are generally very similar in size despite highly lineage-specific repeat content composition, it is likely that birds are particularly efficient at controlling and removing transposable elements.

There is a great deal of literature supporting the physiological costs of large genomes, including reduced organism-level size, slower cell division, reduced metabolic rates, and even issues with circulation and neurological capacity (Hanken & Wake 1993; Roth et al. 1997; Nuzhdin 1999; Cavalier-Smith 2005; Gregory 2005; Sun et al. 2012). For birds, flight requires a high metabolic rate, and a number of studies have linked this to the observed compact genome sizes (Hughes & Hughes 1995; Gregory et al. 2009; Wright et al. 2014). Similarly, compact genomes are also seen in bats compared with other mammalian groups, again due to the reduction of transposable element content (Van Den Bussche et al. 1995). A greater understanding of avian transposable element content is therefore necessary to understand not only the effects of individual elements on the host, but also what the content reveals about avian genome evolution.

1.2.2 Chicken LTR retrotransposon content

Until the release of the reference genome very little was known about the wider LTR retrotransposon content of the chicken. Most research had focused on individual ERVs which were known to have phenotypic and commercial relevance, particularly the alpharetroviral Avian Leukosis Virus subgroup E (ALVE) and Endogenous Avian Virus (EAV) insertions (Sacco & Venugopal 2001; Borisenko 2003; Payne & Nair 2012). These elements are relatively recent insertions and are therefore more likely to be structurally intact and able to elicit a modulatory phenotypic effect (introduced fully below). However, they are also at relatively low copy numbers, and therefore make up only a small fraction of the total LTR retrotransposon content.

In the 2004 first draft release of the chicken reference genome, Hillier and colleagues identified approximately 30,000 distinct LTR retrotransposon-derived elements, accounting for 1.3 % of the assembled genome. All identified sites were from ERVs, with no gypsy, copia or bel/pao homologues detected. Only one of the two known ALVEs in the reference genome bird were assembled and detectable, and it was only recently that the second ALVE was detected in one of the unassembled contigs localised to the very 5' end of chromosome 1 (Benkel & Rutherford 2014). ERVs were identified from each of the three retroviral clades, and many of the GGERVK10 (class II; betaretrovirus) and GGERVL (class III; spumavirus) family elements were found to be less than 3 % divergent, and therefore predicted to still be active within the genome. Gammaretroviral (class I) sequences identified in the chicken genome were heavily degraded. These elements are most closely related to exogenous Murine Leukosis Virus and the human endogenous retrovirus group HERV-I, and are likely the remnants of ancestral insertions common to all amniotes (Martin et al. 1997; Martin et al. 1999; Borysenko et al. 2008). Interestingly, no ERVs from current gammaretroviruses, such as chicken syncytial virus (CSV) or REV in turkeys, are detectable in bird genomes (Borysenko et al. 2008).

The following year, Wicker and colleagues (2005) performed Cot-based cloning and sequencing (CBCS) to identify repeats present at large copy numbers. This enabled better definition of some of the observed LTR retrotransposon element families, before the first full-genome bioinformatic analysis of the chicken in 2008 (Huda et al. 2008). Wicker identified four high copy number groups of LTR retrotransposons: *Birddawg* (7,404), *Kronos* (4,961), *Hitchcock* (3,324) and *Soprano* (1,362). Most identified elements were short, solo LTRs, but internal sequences were identified in *Birddawg* (894; 12.1 %), *Kronos* (1,517; 30.6 %) and *Soprano* (75; 5.5 %) elements, enabling family classification. All three were initially classified as previously unseen *gypsy* elements, but were later reclassified as endogenous spumaviruses (GGERVL-related, class III) when clustered with a larger number of reference LTR retrotransposon sequences (Bolisetty et al. 2012). The *Hitchcock* elements could be defined as solo LTRs due to the presence of TSDs at the termini and the absence of SINE-like features.

The development of a range of bioinformatic tools for the identification of LTR retrotransposons based on their canonical structure, rather than purely sequence

homology, facilitated the next stage of annotation of these elements in the chicken genome (all methods for detecting LTR retrotransposons have been reviewed in the introduction to chapter 3, page 5). Huda and colleagues (2008) used one structure-based method and identified fourteen distinct groups of chicken LTR retrotransposons, resolving across the three retroviral clades, based on the detection of eighty-nine full length sequences. Eleven of these groups were novel but generally filled gaps in the retroviral phylogeny, largely reflecting the detection of structurally intact (at least, LTR-RT-LTR) elements for the first time.

The most recent analysis, by Bolisetty and colleagues (2012), made use of multiple methodologies and performed an extensive analysis of LTR retrotransposon distribution and expression. The authors' approach enabled identification of 492 intact elements and approximately 30,000 solo LTR-like elements (proportionately over 60 % more solo LTRs than are found in mammals). This represented a large increase in total annotated sequence, but the increased length of the genome assembly (Galgal3) caused the proportion to remain at approximately 1.3 %. Intact elements were found to be significantly depleted within or nearby (< 10 kb) coding regions, and 40 % were found in clusters where the elements were unrelated by genera or insertion age. Together, these data support selection against deleterious LTR retrotransposon insertions, but the authors also concluded that the clusters may be functional, either by promoting recombination or acting as binding sites for cytoskeletal elements during cell division.

Bolisetty and colleagues also performed parallel annotations of the turkey and zebra finch genomes. Whilst the turkey was superficially like chicken (although with lower annotated content presumably due, in part, to its poorer genome assembly; scaffold N50 of 0.86 Mb compared to 11.06 Mb in chicken), the zebra finch had almost three times the LTR retrotransposon-derived sequence content, with 1,221 intact elements and approximately 78,000 solo LTR-like elements. As mentioned above, the overall zebra finch repeat content is very comparable to chicken, so this likely reflects lineage-specific LTR retrotransposon expansion. It is possible that such results were not biologically representative, as the authors' approach used a methodology potentially more biased towards the gammaretroviral sequences that are more common in the zebra finch, due to the program's development based on primate retroviral sequences. However, if these differences are representative it is then a question as to which taxonomic level the lineage

specificity extends. Bolisetty and colleagues suggested that this represented a deficit of LTR retrotransposons in Galliformes relative to the entire Neoaves superorder. But as this conclusion was drawn from only three species, a much wider and more comprehensive analysis of avian genomes is required to quantify this proposed deficit.

Summary of the current annotation of chicken LTR retrotransposon content

ERVs have been detected from all three retroviral clades, with full length examples from spumaviruses, alpha-, beta- and gammaretroviruses, but deltaretrovirus and lentivirus ERVs (both clade II) are absent. No gypsy, copia or bel/pao elements have been detected, and DIRS elements have also not been identified in any avian or mammalian genome, despite their presence in other sauropsids (Piednoël et al. 2011).

A total of 1.35 % (almost 15 Mbp) of the chicken genome (based on the Galgal3 assembly) has been annotated as LTR retrotransposon-derived elements, including approximately 500 intact elements. As expected, this value was far less than the 10 % typically observed in mammalian genomes, but it was also three times less than observed in the zebra finch, leading to the hypothesis that there is a deficit of LTR retrotransposons in Galliformes compared to higher order Neoaves.

Whilst endogenous alpharetroviruses make up only a small fraction of the total LTR retrotransposon content, these elements remain active and are commonly intact. In addition, ALV is the only known chicken retrovirus with both exogenous and endogenous current activity (Borysenko et al. 2008; Payne & Nair 2012).

1.2.3 The endogenous alpharetroviruses of the chicken genome

ALVE loci

Like other alpha- and betaretroviruses, ALV insertion is effectively random across the genome. However, the presence of a nuclear localisation signal (NLS) in the ALV *integrase* means that, like lentiviruses, ALV does not require the breakdown of the nuclear membrane to access genomic DNA. This likely explains the weak observable insertion site 'preference' of ALV for open chromatin, generally within or near regions

of the genome expressed by RNA polymerase II, such as protein coding genes (Narezkina et al. 2004; Justice & Beemon 2013). The impact of these insertions is modulated by host genetic susceptibility and the location of the insertion. Most ALVs are slowly transforming, producing lymphoid tumours over weeks or months via insertional mutagenesis, but the acquisition of accessory genes (such as *v-src* in Rous Sarcoma Virus; RSV) enables acute transformation and rapid tumour development (Payne 1998).

Purely exogenous ALVs (subgroups A-D and J) induce leukoid tumours across the Galliformes. Endogenous ALVs can also infect horizontally within a population, but have a species-specific range. Subgroup E (the ALVEs) are found in RJF and the domestic chicken but no other *Gallus* species, subgroups F and G are in pheasant, H in partridge and I in quail (Frisby et al. 1979; Venugopal 1999). ALV subgroups are defined by the virion surface protein constituent gp85 (encoded by the *envelope* gene), as this defines the specific TV (tumour virus) cell entry receptor. ALV-A enters by the TVA receptor, ALV-C by the TVC receptor, and subgroups B, D and E via the TVB receptor. Two TVB alleles have been identified which convey resistance to ALVE cell entry (*TVB*S3*, *TVB*R*), but these are both recessive to, and far less common than, the wildtype *TVB*S1* which is susceptible to all three subgroups (Hunt et al. 2008; Yu et al. 2008; Justice & Beemon 2013). There are no documented cases of ALVEs containing accessory oncogenes, and they are expressed at a level two to three orders of magnitude lower than exogenous ALV due to deletion of enhancers in the ALVE LTR U3 domain (Coffin et al. 1983; Norton & Coffin 1987; Conklin 1991). ALVE LTRs typically only have a single enhancer, rather than the tandem enhancer cassette present in exogenous ALV (Ruddell 1995). Consequently, ALVEs rarely induce tumour formation, but the presence of these loci in the host genome can modulate the infection dynamics of exogenous ALV (Benson et al. 1998; Payne 1998; Yu et al. 2008; Payne & Nair 2012; Kanda et al. 2013).

ALVE insertions are recent and recurrent, as there is significant element variation between chicken populations (with approximately fifty ALVE loci identified to date), low element copy number, and generally high structural integrity. In fact, nearly half of the twenty-three ALVEs identified across various WL lines can produce replication competent virions (Benkel 1998; Borisenko 2003). The presence of replication

competent ALVEs, and ALVEs expressing the *gag* gene, has been associated with reductions in body weight, egg production, size and shell thickness, and increased retroviral shedding, which facilitates horizontal infection within a flock (Crittenden et al. 1984; Kuhnlein et al. 1989; Gavora et al. 1991; Ka et al. 2009). Conversely, expression of the ALVE *envelope* can mediate the effects of both exogenous and endogenous ALV through receptor interference, as the endogenous *env* glycoproteins physically block the TV receptors (Smith et al. 1990a; Smith et al. 1990b; Smith et al. 1991). As the number of ALVEs increases, the interplay of these effects becomes more complicated and less predictable, particularly when lines are interbred. The associated effects of ALVEs, and the methods for their detection, have been more extensively reviewed in the introduction to chapter 6 (page 5).

Given these detrimental productivity and complex immunological effects, it has become common practice to attempt to eradicate ALVEs from commercial lines. ALVEs are non-essential genetic components of the chicken genome, and 'ALVE-free' lines have been developed (Zhang et al. 2008). However, the commercial community has been unable to completely remove ALVEs from breeding stock, with WELs typically containing one to three ALVEs, BELs containing four to eight, and broilers containing six to ten (Benkel 1998). This has been a combination of the inability to detect all ALVE insertions, the association between some ALVEs and desirable traits (such as ALVE21 and slow feathering (Bacon et al. 1988; Tixier-Boichard et al. 1994; Tixier-Boichard & Boulliou-Robic 1997; Elferink et al. 2008; Bu et al. 2013), and ALVE-TYR and white plumage (Chang et al. 2006)), and the need to balance existing selective breeding programmes. A new methodology for detecting ALVE insertions and assessing the diversity between commercial lines is needed to better direct the elimination of these elements from the chicken genome.

EAVs

EAVs are a much older group of endogenous alpharetroviruses than ALVEs and exhibit a wider distribution among the Galliformes, with element divergence matching host co-speciation patterns. Despite a more ancient origin, in many species EAVs remain transcriptionally active and horizontal transmission is common in sympatric species

(Dimcheff & Drovetski 2000; Dimcheff et al. 2001). In the RJF and domestic chicken EAV sequences are generally inactive, even compared to other *Gallus* species (Sacco et al. 2001), but some elements have retained the ability to retrotranspose and express individual retroviral domains.

EAVs are divided into three main groups due to their phylogenetic clustering: EAV-0, EAV-HP and EAV-E51 (including EAV-E33 sequences which fall within the EAV-E51 cluster, but were originally identified from a different clone). EAV-0 sequences share the most recent common ancestor with ALVEs and are generally the most structurally intact, even retaining functional *polymerase*. EAV-E51 and EAV-HP elements all exhibit large deletions in their *polymerase* gene (rendering them non-autonomous) but retain intact *gag* and *envelope* domains in some instances (Boyce-Jacino et al. 1992; Bai et al. 1995; Sacco & Nair 2014). Despite this, group members retain high LTR identity, suggesting recent and recurrent retrotransposition, and there are independent, segregating EAV complements in various chicken lines (Dimcheff & Drovetski 2000; Sacco & Venugopal 2001; Wragg et al. 2015). Until recently, a fourth distinct EAV group was defined which shared 5' homology with EAV-HP and 3' homology with EAV-E51, but had a unique, yet truncated *polymerase* domain. These sequences, known as ART-CH (avian retrotransposon in chicken), are now known to be recombinant sequences between an EAV-HP (5' LTR and most of *gag*) and an EAV-E51 which had already had significant *polymerase* degradation (Sacco & Nair 2014).

Despite their general degradation, EAVs still retrotranspose and insertional mutagenesis remains a possibility. Individual insertions have been shown to elicit phenotypic effects, particularly due to the alternative promoter activity generated by a high number of solo LTRs (Sacco & Venugopal 2001). For example, independent EAV-HP LTR insertions in the promoter region of the solute carrier *SLCO1B3* (solute carrier organic anion transporter family member 1B3) were recently shown to upregulate gene expression causing the *oocyan* 'blue-green' egg shell phenotype in both Chinese (Z. Wang et al. 2013) and South American (Wragg et al. 2013) chicken breeds.

The movement of non-autonomous EAVs is facilitated by the presence of replication-competent retroviruses. This can either be other ERVs, such as intact ALVEs, or exogenous retroviruses during an infection. The emergent ALV-J is a result of the latter,

following recombination between an exogenous ALV-A and the intact *envelope* domain of an EAV-HP (Payne et al. 1991; Payne et al. 1992; Bai et al. 1995; Benson et al. 1998; Sacco & Flannery 2000). The altered ALV virion causes ALV-J to enter cells via the NHE1 receptor (originally named TVJ), changing the target cell from B cells to any myeloid cell, resulting in myelocytomas (rather than typical ALV-induced B cell lymphomas) in infected birds (Zhang et al. 2008; Payne & Nair 2012; Sacco & Nair 2014). Whilst this target cell change created a new immunological challenge for infected birds, expressed EAV-HP *envelope* has been shown to induce tolerance by receptor interference (Sacco et al. 2004).

The detrimental global impact of ALV-J has been well documented (Payne 1998; Fadly 2000; Payne & Nair 2012), and whilst its spread has largely been controlled in Europe and North America, it is still a recurrent issue in Asia where secondary ALV-J/ALVE recombinants (Liu et al. 2011) and acutely transforming strains (Chesters et al. 2001) have been identified. The potential for novel recombinant retroviruses, even derived from sequences which have been in the genome for millions of years, highlights the need for a full characterisation of ERVs in the chicken, including documenting the observed diversity between populations.

1.3 The scope of this PhD project

This project has two broad aims. Firstly, an updated characterisation of LTR retrotransposon-derived elements in the chicken genome, and an assessment of how this content compares with that observed in other avian genomes. This will involve the critical assessment of existing LTR retrotransposon detection methodologies and the creation of a new bioinformatic annotation pipeline, LocaTR, to identify these elements (Chapter 3). The intactness, distribution and expression potential of the chicken LTR retrotransposons will be assessed to give an accurate snapshot of the abundance of these elements. The LocaTR pipeline will then be applied to seventy-three sauropsid genomes, including six reptilian outgroups, to identify LTR retrotransposons across the lineage, assess any lineage-specific expansions or contractions, and evaluate the impact of genome assembly quality on LTR retrotransposon detection (Chapter 4).

Secondly, a new bioinformatic pipeline will be described to identify ALVE insertions from next generation sequencing data, to characterise the diversity of these youngest chicken ERVs and assess their potential impact on host genome stability and physiology (Chapter 6). The pipeline will be tested and validated by analysing eight elite layer lines from Hy-Line International, developing diagnostic assays for each of the identified ALVEs, and sequencing each ALVE insert. The potential phenotypic effects will be assessed, and potential methods identified for mediating the detrimental impacts of these insertions (Chapter 6). The wider ALVE diversity in chicken populations will be assessed by applying the pipeline to multiple datasets from commercial stocks, indigenous breeds, and wild RJF (Chapter 7).

Together, this will enable an updated annotation of transposable elements capable of insertional mutagenesis, recombination with exogenous retroviruses, and disruption of commercially relevant traits, with the aim of mitigating their detrimental effects on productivity and animal welfare.

Chapter 2: Materials and Methods (i)

This chapter outlines the methodology used for the next two chapters. These chapters concern the identification of all LTR retrotransposons within assembled avian genomes, beginning with the annotation of the chicken. This involves the development of a new identification pipeline to annotate these repetitive elements, and their full characterisation in terms of genomic distribution, age and intactness.

In this thesis 'element' refers to a single LTR retrotransposon. There may be multiple 'copies' of the same LTR retrotransposon in the genome (classified by LTR retrotransposon family or retroviral genera), but each integration is a separate 'element'.

2.1 Development of LocaTR – an integrated identification pipeline for LTR retrotransposons, using the Galgal4 chicken genome assembly

2.1.1 Genomic resources

The Galgal4 chicken genome assembly (GenBank: GCF_000002315.3) was used for development of the LocaTR identification pipeline and for the analysis of chicken LTR retrotransposon content.

2.1.2 Construction of the LocaTR LTR retrotransposon identification pipeline

LocaTR, a unified LTR retrotransposon identification pipeline, was created to connect the disparate LTR retrotransposon identification methodologies outlined below in section 2.1.3 into a clear, user friendly framework. The pipeline includes seventeen processing scripts directly run by the user, six accessory scripts containing standalone functions, a reference set of LTR retrotransposon sequences, nucleotide and protein domain profile Hidden Markov Models (pHMMs), and extensive documentation.

Most scripts were written in Python, with four of the position formatting scripts written in BASH. All scripts were written specifically for this pipeline, are extensively commented, and have built in help messages and error catches. All code can be found on the CD accompanying this thesis (Appendix 1) as well as online in the GitHub LocaTR repository: https://github.com/andrewstephenmason/LocaTR.

2.1.3 *In silico* methods for the identification of LTR retrotransposons

Three homology based identification programs and four structure-based identification programs were implemented as part of LocaTR to annotate LTR retrotransposons. Both branches of identification strategies have their strengths and weaknesses (outlined in section 3.1), and the individual programs also have their own biases. However, the combination of strategies and the use of multiple programs within each strategy was done to mitigate against those potential issues. Other structure-based approaches are available, but the four used here have a wide use in the literature and no complete identification redundancy between methods.

Homology-based identification

RepeatMasker v4.0.3 (Smit et al. 2013) analysis was performed on Galgal4 with default settings and also by specifying the '-species' flag as 'chicken' and then as 'vertebrates'. Under default settings, RepeatMasker uses the primate RepBase libraries (Jurka et al. 2005), but the '-species' flag directs RepeatMasker to annotated repeats in specific phylogenetic groups. In all cases the '-nolow' flag was used to reduce computational effort by ignoring low complexity DNA and simple repeats during masking. In addition, the three non-default sensitivity settings of RepeatMasker were tested to compare total annotated repeat content against processing time. Following testing of settings, LTR retrotransposon-annotated regions were extracted using a custom BASH script.

As well as using existing, generic RepBase libraries, a custom database of reference LTR retrotransposons was compiled from 717 single domain and full-length sequences (Appendix 2: AF01). These were sourced from the *Gallus*-specific RepBase libraries and sequences from the Gypsy Database (GyDB) of Mobile Genetic Elements (Llorens et al. 2011), selected for diverse LTR retrotransposon phylogenetic coverage, from avian host species where possible. These sequences were used as queries in local BLASTn and tBLASTx v2.2.28 (Altschul et al. 1990) searches of Galgal4, using an E-value threshold of 10^{-10}, with rejection of hits shorter than 100 bp (Katzourakis & Gifford 2010). These positions were combined with the locations identified by RepeatMasker, and annotated sites were merged if they overlapped or were fewer than 11 bp apart.

Each identified putative sequence was analysed again with RepeatMasker, specifying the '-species vertebrates' flag, and those locations with high homology to other repeat classes, such as Chicken Repeat 1 (CR1) LINE elements identified due to their *reverse transcriptase* domain, were removed. Putative LTR retrotransposon sites were further checked using a reciprocal tBLASTx search against the custom LTR retrotransposon database. This whole process was wrapped into Python scripts.

Additionally, ReDoSt v1.1 (Piednoël et al. 2011) was used with default settings for the identification of the structurally divergent DIRS elements. Putative DIRS elements had to have an identifiable *reverse transcriptase* domain, and either a recognisable *methyl transferase* or *tyrosine recombinase* domain. DIRS element locations were extracted using a custom Python script, including 350 bp of up- and down-stream sequence to capture the inverted terminal repeats (ITRs).

Structure-based identification

LTR_STRUC (LS) v1.1 (McCarthy & McDonald 2003) was the oldest structure-based program used for this study. Whilst it has continued to be used in LTR retrotransposon identification studies, the program itself is a black box: the source code is unavailable, there is very limited documentation, and it only runs as a foreground Windows executable. In the first instance, the ten poorly defined sensitivity settings were compared by analysing the Galgal4 chromosomes 1 and Z against the default analysis. Following this, LS was run on the full Galgal4 assembly with sensitivity 1, per contig, to address inherent memory issues, and then separately on reverse complemented sequence, as LS does not consider the reverse strand during identification. Sequences were then used as BLASTn queries against Galgal4 to obtain the putative element positions. Sequence preparation, batch running of LS on Windows, and identification of element locations was all managed with Python scripts.

LTR Harvest (LH) (Ellinghaus et al. 2008) was implemented as part of GenomeTools v1.5.1 (Gremme et al. 2013), compiled with HMMER v3.1b1 (Eddy 2009), with two custom parameter groupings in addition to the default settings (Table 2.1). Custom set 1 followed less stringent requirements for LTR length and sequence identity (enabling

detection of older, more divergent sequences). Custom set 2 went beyond this to also allow for nested LTR retrotransposons, where a more recent LTR retrotransposon insertion occurred within an existing LTR retrotransposon. Custom settings were chosen based suggestions in the LH user manual, and the annotated features of previously identified chicken LTR retrotransposon sequences.

Table 2.1 Default and chosen parameters for LTR Harvest optimisation.

LTR Harvest Parameter	Default	Set 1	Set 2
Min LTR length (bp)	100	80	80
Max LTR length (bp)	1,000	2,000	2,000
LTR homology cut-off (%)	85	75	75
Max element length (bp)	15,000	15,000	25,000
Report nested elements	no	best	all

Prior to running RetroTector (ReTe) v1.0.1 (Sperber et al. 2007), all contigs shorter than 30 kb were 'padded' at each end to aid the LTR retrotransposon identification algorithm, as suggested in the ReTe documentation. The padding was 15 kb of randomly generated sequence with equal base frequencies, and the sequence was checked prior to use with BLASTn and tBLASTx searches against the NCBI non-redundant database (Pruitt et al. 2002) to ensure it was devoid of gene or repeat identity. The primary 'SweepDNA' protocol masked undesirable repeat classes by converting identified sequences to ambiguous bases (Ns). This protocol masked Alu and L1 elements by default, but four additional 'brooms' (specific repeat-masking programs) provided by the ReTe authors were used to mask CR1 elements. Repeat-masking limits the search space and reduces false positives. The secondary 'SweepScripts' protocol was run with default settings and putative hits were extracted using custom Python scripting, rather than the 'CollectGenome' protocol which required management of multiple SQL databases.

The default configuration file for MGEScan_LTR (MGS; referred to as LTR_Rho in original literature) v1.3.1 (Rho et al. 2007) was altered to increase the diversity of detected elements. The minimum distance between LTRs was reduced from 2 kb to 1

kb, the maximum length for the LTR retrotransposon was increased from 20 kb to 25 kb, minimum LTR length was decreased to 80 bp, and maximum LTR length was increased from 1 kb to 2 kb. These changes were made to reflect the optimum parameters identified with the multiple tests of LH. MGS was run individually on each Galgal4 contig and the positions extracted with a custom BASH script.

Following identification of putative elements by these four programs, additional support was required before they were defined as LTR retrotransposons. Each element was analysed by RepeatMasker to check for other repeat classes, particularly CR1. LTR retrotransposon nucleotide motifs were identified using Dfam v1.2 (Wheeler et al. 2013) pHMMs with HMMER nhmmscan and an E-value threshold of 10^{-5}. Putative protein-coding regions within the elements were identified using GyDB *gag*, *pol*, *envelope* and accessory protein pHMMs with HMMER hmmscan and an E-value threshold of 10^{-10}, following translation of each element into all six reading frames with the EMBOSS v6.6.0 transeq tool (Rice et al. 2000). E-value thresholds were chosen to reflect recommendations in the HMMER user manual. pHMMs for the Galgal4 tRNA genes were also built to enable characterisation of the protein binding site (PBS) domain between the *5'* LTR and *gag* gene. tRNA genes were identified using tRNAscan-SE v1.3.1 (Lowe & Eddy 1997) with default parameters. tRNA sequences were then extracted using custom Python scripts and aligned by amino acid using MUSCLE v3.8.31 (Edgar 2004) with default settings, and pHMMs built using HMMER hmmbuild. HMMER hmmpress was used to create all pHMM flatfile databases.

Feature test results were used to generate composite confidence scores for each putative insertion. Elements were scored with a 'P' if any LTR retrotransposon coding domains were identified. Identification of *reverse transcriptase* would ideally be supported by identification of other domains, as *reverse transcriptase* alone could belong to a LINE. Elements were also scored with a 'D' if there was positive identification of LTR retrotransposon nucleotide motifs, with the 'D' followed by a number identifying where the first LTR retrotransposon-related motif positioned when ranked by E-value. If elements were annotated by RepeatMasker as LTR-related, they were also scored with an 'R', with the output showing how much of the element had matches to various repeat classes. Finally, elements were also scored with a 't' if the PBS site was identified.

Putative elements were ranked on assigned score, and those with no supporting evidence were automatically discarded. The result output for all other identified sites was checked manually. The highest confidence score for a putative element was 'D1+P+R+t', but lower confidence elements were also selected. In general, elements with just the 'R' classification were discarded unless most of the element had LTR-related annotation (and the annotation of other repeat classes was minimal), and elements with just the 't' classification were automatically discarded. User discretion was important with more ambiguous scores, such as 'D7+R', but putative elements were discarded if there was insufficient evidence that they were LTR retrotransposons.

Combining the homology and structure-based identification approaches

All elements identified by the structure-based identification programs were used in a secondary BLASTn/tBLASTx protocol, following the same approach as above with the custom LTR retrotransposon database. This was to identify degraded sequence or solo LTRs related to the newly identified, structurally intact elements (SIEs) which were not identified by the homology based identification. The final stage of the LocaTR identification pipeline was to combine and merge all LTR retrotransposon-related sequence identified in the homology, structure-based and secondary BLAST protocols using external element coordinates, resulting in a full LTR retrotransposon annotation.

2.2 Analysis of the LTR retrotransposons identified in the Galgal4 chicken genome assembly

2.2.1 Initial characterisation of structurally intact LTR retrotransposons

Structurally intact elements were aligned to the forty-five full length LTR retrotransposon reference sequences in the custom database using MUSCLE with default settings, and analysed for putative coding regions with GyDB protein pHMMs, using HMMER hmmscan and an E-value threshold of 10^{-10}. Domain alignments were inspected and SIEs were manually classified into either a retroviral genus or another LTR retrotransposon family. Protein assignments, particularly *reverse transcriptase*, took precedence over nucleotide alignments during classification.

LTR pair sequence identity was calculated from MUSCLE alignments with default settings. The insertion date for each element was calculated using this LTR pair identity and nucleotide divergence rates from the Galliformes (Helm-Bychowski & Wilson 1986). The potential impact of selection on element distribution over time was tested by randomly assigning intact element insertion ages 1,000 times, and averaging the proportions for age categories based on LTR identity.

Average GC content was calculated for the entire genome and per chromosome. Chromosomal GC content was correlated with $\log_{10}$ transformed values for chromosome length and gene density. GC content was calculated for each structurally intact element and compared to the genomic average as a function of their calculated insertion date.

2.2.2 Genomic distribution of LTR retrotransposons

LTR retrotransposon density and clustering

Element density (LTR retrotransposons per Mb) was calculated per chromosome and correlated with chromosome length, gene density and average chromosomal recombination rate (Elferink et al. 2010), all of which were $\log_{10}$ transformed to normality. Pairwise Pearson correlations were performed and a General Linear Model (GLM) fitted using element density as the response variable and chromosome length and recombination rate as covariates. Data from chromosomes 27 and Z were excluded from the GLM as outliers due to large residuals in the normality plots. Furthermore, data from chromosomes 16, 25, 32 and W could not be directly compared to the overall analysis due to substantial known sequence assembly gaps.

Density heterogeneity was considered through identification of structurally intact elements in clusters, with a cluster defined as at least five elements per Mb, compared to a calculated even distribution of one element per Mb. This level was chosen to match the cluster definition used in the previous work by Bolisetty and colleagues (2012) with the Galgal3 genome assembly, even though they identified half the number of structurally intact elements compared with the Galgal4 analysis presented in this thesis. Tests were completed using a stricter cluster cut-off of ten elements per Mb, to ensure the identified clusters were biologically representative rather than a statistical result of an

increased number of identified elements. Elements within each cluster were checked for their insertion age and LTR retrotransposon family classification.

The probability of clusters arising by chance was assessed by comparing the observed number of SIEs within clusters to 100,000 random integration distributions using equal numbers of modelled insertion sites. Differences between observed and simulated cluster numbers were quantified with exact binomial tests. Average recombination rates within clusters were obtained from the 500 kb-average-bin data from Elferink and colleagues (2010), following conversion of cluster positions to the Galgal3 assembly (WASHCU2; GenBank: GCF_000002315.2) using the 'Map to Reference' tool in Geneious v7.0.4 (Kearse et al. 2012). Galgal4 centromere locations were also mapped using the Galgal3 annotations (Table 2.2).

Table 2.2 Galgal4 centromeric locations identified using the flanking sequence to the 1.5 Mb of ambiguous bases used to mark the centromeres in the Galgal3 WASHCU2 annotation file. Whilst the Galgal4 centromeres often retain some ambiguous bases, their resolution was greatly improved.

Chromosome	Galgal3	Galgal4
1	76,857,403 - 78,357,402	74,621,949 - 74,673,419
2	52,291,242 - 53,791,241	52,740,394 - 52,743,010
4	19,307,569 - 20,807,568	19,201,396 - 19,215,816
5	6,508,834 - 8,008,833	5,826,741 - 5,827,880
8	10,229,801 - 11,729,800	9,981,100 - 9,992,722
Z	24,138,915 - 25,638,914	24,533,502 - 24,641,581

Analysis of LTR retrotransposon distribution on chromosomes 27 and Z

Chromosomes 27 and Z were identified as GLM outliers in terms of their LTR retrotransposon density. Density heterogeneity was inspected across each of these chromosomes individually, with particular regard to the recombination rate.

Chromosome 27 recombination rate and LTR retrotransposon density heterogeneity were correlated using the 500 kb-average-bin data from Elferink and colleagues (2010).

These densities were inspected manually through Ensembl to enable direct comparison with gene distribution. The Ensembl annotation file was used to calculate average gene exon number across the chromosome. In addition, the Ensembl Comparative Genomics Synteny tool was used to check chromosome 27 synteny with the genomes of the zebra finch (*Taeniopygia guttata*) and turkey (*Meleagris gallopavo*), as well as annotated gene paralogues also found on chromosome 27.

Element density in the highly recombining Z chromosome pseudoautosomal region (PAR; the 5' 630 kb of the chromosome; Smeds et al. 2014) was compared with the chromosomal average. In addition, the $\log_{10}$ element density vs $\log_{10}$ chromosome length correlation straight line equation was used to predict the expected LTR retrotransposon density of the Z, given its length. The observed density was compared to this expected value using the exact binomial test.

Distribution of LTR retrotransposons relative to known gene annotations

Structurally intact element locations were compared to the Ensembl Galgal4 version 79 annotation file using the BEDTools v2.23.0 intersectBed tool (Quinlan & Hall 2010). Elements overlapping with 'transcriptional units' (TU; regions including exons, introns, UTRs, and 5 kb up- and downstream regions, for protein and RNA genes) were annotated for strand and TU domain overlap. The shortest distance from each non-overlapping element to a TU was calculated and distances were binned in 10 kb ranges up to 100 kb.

Genome-wide and per chromosome analyses were completed and compared with randomly generated simulations of equal numbers of modelled insertion sites. Simulated insertion sites were modelled using a random number generator and repeated 100,000 times. Deviation of the observed distribution from the modelled data was quantified using individual category exact binomial tests and the Kolmogorov-Smirnov test for the overall distance distributions.

Similar distribution analyses were conducted relative to constrained genomic locations using two multiple sequence alignments from Ensembl; one consisting of twenty-three amniotes and another of seven sauropsids.

2.2.3 Analysis of the expression of structurally intact LTR retrotransposons

Available RNAseq data for quantification of expression

Twenty-three diverse chicken RNAseq datasets were used to provide evidence for any LTR retrotransposon expression. Of these, twenty were adult tissues (breast muscle, bursa, cerebellum, duodenum, gizzard fat, Harderian gland, heart muscle, ileum, kidney, left optic lobe, liver, lung, ovary, pancreas, proventriculus, skin, spleen, thymus, thyroid and trachea) from the Roslin Institute J-Line (EBI ENA: PRJEB12891), and three were White Leghorn embryonic stages HH4-5 (GenBank: SRX893876), HH14-15 (GenBank: SRX893868) and HH25-26 (GenBank: SRX893873).

All twenty-three datasets were quality checked with FASTQC v0.11.2 (Andrews 2012). Whereas the J-Line data were all high quality, the embryonic data exhibited low quality read ends and overrepresentation of adapter sequences. These were removed with Trim Galore v0.4.0 (Krueger 2013) using Cutadapt v1.4 (Martin 2011).

Mapping of RNAseq data and assessment of putative expression

Reads from each tissue dataset were mapped independently to the Galgal4 assembly using Bowtie2 v2.2.5 (Langmead & Salzberg 2012) and TopHat2 v2.0.14 (Kim et al. 2013). Inner insert size and strand orientation was identified during mapping, and transcripts assembled using Cufflinks v2.2.1 (Trapnell et al. 2010) without a reference annotation.

Locations for assembled transcripts were overlapped with the locations of structurally intact LTR retrotransposons using BEDTools intersectBED, requiring matched strand orientation. Overlaps were viewed using the Ensembl genome browser and Geneious, and the extent of expression for each LTR retrotransposon was assessed. In cases where expression appeared partial across the intact LTR retrotransposon, the relevant sequence was used as a BLASTn and BLASTx query against the NCBI non-redundant database to identify whether it matched intact domains or genes. All full-length or domain-matched transcripts were translated into the three forward reading frames and the number of stop codons assessed. Predicted proteins without a high frequency of interspersed stop codons were used as BLASTp (Altschul & Madden 1997) queries

against the NCBI non-redundant database and homologous results were aligned with MUSCLE and individually assessed. Putative conserved domains were identified using InterPro (Mitchell et al. 2015) and transmembrane topologies predicted using Phobius (Käll et al. 2004).

Further characterisation of Ovex1: a co-opted, gammaretrovirus-derived gene

Ovex1 is a chicken gene which had been previously characterised as a co-opted endogenous gammaretrovirus with expression limited to the ovaries (Carré-Eusèbe et al. 2009). However, RNAseq mapping in this project identified that *Ovex1* expression was more ubiquitous. Expression was quantified in all tissue datasets, including those where cufflinks failed to construct transcript models. Reads were extracted from the RNAseq BAM files for the *Ovex1* exon and whole gene region, and viewed in Geneious.

InterPro and Phobius results for the *Ovex1* protein (GenBank: NP001159385.1) were compared to similar analysis of *envelope* proteins from bovine leukaemia virus (BLV; deltaretrovirus; GenBank: AF033818.1), feline foamy virus (FFV; spumavirus; GenBank: NP056915.1), *Mus dunni* endogenous virus (MdEV; gammaretrovirus; GenBank: AAC318061.1), mouse mammary tumour virus (MMTV; betaretrovirus; GenBank: AAC82558.1), Rous sarcoma virus (RSV; alpharetrovirus; GenBank: BAD98245.1) and Walleye dermal sarcoma virus (WdSV; epsilonretrovirus; GenBank: AAC82608.1). Each of these retroviral, *envelope* proteins was also annotated for protein domain families and motifs using NCBI.

Ovex1 sauropsid homologues were identified using the chicken *Ovex1* protein sequence as a BLASTp query against the NCBI non-redundant database. These were MUSCLE aligned along with forty-seven GyDB retroviral *envelope* protein sequences, the alignment was trimmed, and a phylogenetic tree constructed using RAxML v8.1.15 (Stamatakis 2014) with the PROT-GAMMA-I-WAG protein evolutionary model and one hundred bootstraps. Patterns of selection in the sauropsid *Ovex1* protein homologue alignment were analysed using the DataMonkey (Pond & Frost 2005) hosted DEPS (Directional Evolution in Protein Sequences) program (Pond et al. 2008).

2.3 Comparative analyses of the LTR retrotransposon content of the Galgal4 and Galgal5 chicken genome assemblies

2.3.1 Analysis of the Galgal5 assembly with LocaTR

LTR retrotransposons were identified in the Galgal5 genome assembly (GenBank: GCF_000002315.4) using the LocaTR identification pipeline with default settings. The assembly statistics of Galgal4 and Galgal5 are compared in Table 2.3.

Table 2.3 Comparison of the Galgal4 and Galgal5 chicken genome assemblies.

Assembly features	Galgal4	Galgal5
Assembled length (bp)	1,046,932,099	1,230,258,557
Assembled chromosomes	1-28, 32, W, Z	1-28, 30-33, W, Z
Total contigs	27,041	24,693
Contig N50 (bp)	279,750	2,894,815
Placed contigs (bp)	1,014,655,963	1,091,712,069
Unplaced contigs (bp)	32,120,124	138,199,872
Total scaffolds	16,847	23,870
Scaffold N50 (bp)	12,877,381	6,379,610

2.3.2 Analysis of the LTR retrotransposons of the Galgal5 genome assembly

The distribution and location of LTR retrotransposons in the new, Galgal5 assembly was analysed in much the same manner as with Galgal4, described above (Section 2.2.2).

Correlations were made between element density, chromosomal recombination rate, and gene density, now using the updated Ensembl Galgal5 v86 feature annotation file. Data for chromosomes 27 and Z were excluded due to their far higher than expected density, which resulted in non-normal plots after $\log_{10}$ transformation. Whilst improvements have been made to chromosomes 16 and W they remain poor assemblies compared to their known lengths, so were again not considered in this analysis (Warren et al. 2017). Chromosomes 30, 31, 32 and 33 were also excluded for this reason. Structurally intact element distribution was inspected for evidence of

clusters, and the clustered elements were checked for insertion date or phylogenetic relatedness.

Element distribution was compared to the Ensembl annotation file using BEDTools intersectBED and closestBed in the same manner as with Galgal4. Random insertion distributions were modelled 100,000 times using BEDTools shuffle and compared using individual category exact binomial tests. Structurally intact element locations were also overlapped with a recently updated list of long non-coding RNA (lncRNA) genes (Kuo et al. 2017) using closestBed, not requiring matching strand orientation.

Throughout this annotation comparisons were made between the two assembly results. Genomic locations were converted between the assemblies using the NCBI Genome Remapping Service (www.ncbi.nlm.nih.gov/genome/tools/remap) where required.

2.4 Analysis of LTR retrotransposon content across the avian lineage

2.4.1 Genomic resources

Genome assemblies for sixty-seven bird species and six reptilian outgroup species were downloaded from the NCBI GenBank. RefSeq assembly versions were used where available to include the assembled mitochondrial chromosome. Assembly statistics and accession numbers for each of the seventy-three species used are shown in Table 2.4.

2.4.2 Homology approach using updated Galgal4 LTR retrotransposon content

Of the seventy-three genomes listed below in Table 2.4, twenty-one phylogenetically diverse species were initially analysed for their LTR retrotransposon content using a purely homology based approach. These twenty-one genomes, including two reptilian outgroups, have been identified by an asterisk in Table 2.4.

Table 2.4 Genome assembly statistics for the seventy-three species used in this study. Those species which were used in the initial, homology-only LTR retrotransposon identification strategy are indicated by an asterisk.

Species name	Common name	GenBank accession	Genome size (bp)	Scaff N50 (bp)	Contig N50 (bp)
Acanthisitta chloris	Rifleman	GCF_000695815.1	1,035,876,403	64,469	20,602
Alligator mississippiensis	American alligator	GCF_000281125.2	2,156,436,551	-	16,610
Amazona aestiva	Blue-fronted amazon parrot	GCA_001420675.1	1,129,535,839	-	26,171
Amazona vittata	Puerto Rican parrot	GCA_000332375.1	1,175,404,042	19,239	6,904
Anas platyrhynchos *	Mallard	GCF_000355885.1	1,105,052,351	1,233,631	26,114
Anolis carolinensis *	Green anole	GCF_000090745.1	1,799,143,587	4,033,265	79,867
Anser cygnoides domesticus	Swan goose	GCF_000971095.1	1,119,151,626	5,202,740	27,602
Antrostomus carolinensis	Chuck Will's Widow	GCF_000700745.1	1,119,683,066	46,345	22,156
Apaloderma vittatum *	Bar-tailed trogon	GCF_000703405.1	1,070,836,417	56,673	28,882
Aptenodytes forsteri *	Emperor penguin	GCF_000699145.1	1,254,347,440	5,071,598	31,730
Apteryx australis mantelli	Brown kiwi	GCF_001039765.1	1,523,972,539	-	17,252
Aquila chrysaetos canadensis	Golden eagle	GCF_000766835.1	1,192,743,076	9,230,743	172,329
Ara macao	Scarlet macaw	GCA_000400695.1	1,204,700,227	15,974	4,399
Balearica regulorum gibbericeps	East African crowned crane	GCF_000709895.1	1,127,622,302	52,178	23,331

Buceros rhinoceros silvestris	Rhinoceros hornbill	GCF_000710305.1	1,065,782,791	53,203	14,587
Calidris pugnax	Ruff	GCF_001431845.1	1,229,094,286	10,060,041	109,237
Calypte anna *	Anna's hummingbird	GCF_000699085.1	1,105,676,412	4,052,191	26,738
Cariama cristata	Red-legged seriema	GCF_000690535.1	1,132,245,425	55,197	24,645
Cathartes aura	Turkey vulture	GCA_000699945.1	1,152,571,117	36,359	15,248
Chaetura pelagica	Chimney swift	GCF_000747805.1	1,119,188,094	3,841,852	30,757
Charadrius vociferus	Killdeer	GCF_000708025.1	1,219,859,583	3,657,050	39,278
Chlamydotis macqueenii	Macqueen's bustard	GCF_000695195.1	1,086,566,339	45,221	21,641
Chrysemys picta belii *	Western painted turtle	GCF_000241765.3	2,365,766,571	6,605,846	21,349
Colinus virginianus	Northern bobwhite	GCA_000599465.1	1,171,855,925	-	6,061
Colius striatus	Speckled mousebird	GCF_000690715.1	1,075,931,597	46,063	25,860
Columba livia *	Rock pigeon	GCF_000337935.1	1,107,989,085	3,148,738	26,579
Corvus brachyrhynchos *	American crow	GCF_000691975.1	1,091,312,783	6,953,989	29,093
Corvus cornix cornix	Hooded crow	GCF_000738735.1	1,049,964,851	16,358,221	94,375
Coturnix japonica	Japanese quail	GCF_001577835.1	927,656,957	2,975,000	511,217

Crocodylus porosus	Australian saltwater crocodile	GCA_000768395.1	2,120,573,303	204,532	32,735
Cuculus canorus *	Common cuckoo	GCF_000709325.1	1,153,894,225	2,989,832	38,137
Egretta garzetta	Little egret	GCF_000687185.1	1,206,501,934	3,067,157	29,019
Eurypyga helias	Sunbittern	GCF_000690775.1	1,088,019,637	47,243	24,402
Falco cherrug	Saker falcon	GCF_000337975.1	1,174,811,715	4,154,532	31,327
Falco peregrinus *	Peregrine falcon	GCF_000337955.1	1,171,973,431	3,935,757	28,645
Ficedula albicollis	Collared flycatcher	GCF_000247815.1	1,118,343,587	6,542,656	410,964
Fulmarus glacialis	Northern fulmar	GCF_000690835.1	1,141,395,646	47,208	25,926
Gavia stellata	Red-throated loon	GCF_000690875.1	1,129,694,867	45,523	24,321
Geospiza fortis	Medium ground-finch	GCF_000277835.1	1,065,292,181	5,255,844	30,521
Haliaeetus albicilla	White-tailed eagle	GCF_000691405.1	1,133,549,865	57,319	25,143
Haliaeetus leucocephalus *	Bald eagle	GCF_000737465.1	1,178,409,481	9,145,499	105,493
Lepidothrix coronata	Blue-crowned manakin	GCA_001604755.1	1,079,577,334	4,972,619	141,849
Leptosomus discolor	Cuckoo roller	GCF_000691785.1	1,136,244,952	62,640	24,735
Manacus vitellinus	Golden-collared manakin	GCF_000692015.1	1,145,854,002	2,558,866	43,697

Meleagris gallopavo *	Turkey	GCF_000146605.2	1,128,339,136	3,801,642	26,671
Melopsittacus undulates *	Budgerigar	GCF_000238935.1	1,117,373,619	10,614,383	55,633
Merops nubicus	Carmine bee-eater	GCF_000691845.1	1,062,961,556	48,089	24,675
Mesitornis unicolor	Brown roatelo	GCF_000695765.1	1,087,290,853	47,102	22,740
Nestor notabilis	Kea	GCF_000696875.1	1,053,559,886	61,475	26,546
Nipponia nippon	Crested ibis	GCF_000708225.1	1,223,863,029	5,211,696	29,116
Opisthocomus hoazin	Hoatzin	GCF_000692075.1	1,203,712,246	2,937,227	28,179
Parus major	Great tit	GCF_001522545.1	1,020,309,133	71,365,269	148,693
Pelecanus crispus *	Dalmatian pelican	GCF_000687375.1	1,160,924,693	43,364	21,679
Phaethon lepturus	White-tailed tropicbird	GCF_000687285.1	1,152,958,507	47,896	22,941
Phalacrocorax carbo	Great cormorant	GCF_000708925.1	1,138,967,842	48,427	17,343
Phoenicopterus ruber	American flamingo	GCA_000687265.1	1,132,184,511	38,071	20,262
Picoides pubescens *	Downy woodpecker	GCF_000699005.1	1,167,323,935	2,093,929	24,809
Podiceps cristatus	Great crested grebe	GCA_000699545.1	1,134,922,578	30,087	17,412
Pseudopodoces humilis	Tibetan ground-tit	GCF_000331425.1	1,042,997,632	16,337,386	165,265

Pterocles gutturalis *	Yellow-throated sandgrouse	GCF_000699245.1	1,069,324,295	49,530	26,448
Pygoscelis adeliae *	Adelie penguin	GCF_000699105.1	1,216,617,519	5,118,896	22,195
Python bivittatus	Burmese python	GCF_000186305.1	1,435,052,152	213,970	10,658
Serinus canaria	Common canary	GCF_000534875.1	1,152,100,110	17,815,079	53,884
Struthio camelus australis *	African ostrich	GCF_000698965.1	1,225,041,896	3,593,425	34,997
Sturnus vulgaris	Common starling	GCF_001447265.1	1,036,755,994	3,416,708	151,865
Taeniopygia guttata *	Zebra finch	GCF_000151805.1	1,232,135,591	8,236,790	38,639
Tauraco erythrolophus	Red-crested turaco	GCF_000709365.1	1,155,540,733	56,334	22,885
Thamnophis sirtalis	Common garter snake	GCF_001077635.1	1,424,897,867	647,592	10,447
Tinamus guttatus *	White-throated tinamou	GCF_000705375.1	1,047,056,493	246,268	29,773
Tyto alba *	Barn owl	GCF_000687205.1	1,120,143,088	52,818	17,226
Zonotrichia albicollis	White-throated sparrow	GCF_000385455.1	1,052,600,561	4,866,725	112,748
Zosterops lateralis melanops	Silver-eye	GCA_001281735.1	1,036,003,386	-	34,514

Each genome was analysed with RepeatMasker specifying the "-species vertebrates -nolow" flags, and the LTR retrotransposon-annotated positions were extracted. This generic analysis was extended with a second RepeatMasker analysis which used a custom library built from the structurally intact Galgal4 LTR retrotransposons identified by the LocaTR pipeline. Annotated positions from these two methods were combined, and putative elements with high homology to other repeat classes were removed.

Annotated LTR retrotransposon content was correlated with genome length, scaffold N50 length and contig N50 length, and values were mapped to a cladogram based on the known phylogeny (Jarvis et al. 2015; Suh 2016).

2.4.3 LTR retrotransposon identification using LocaTR

LTR retrotransposons were annotated in each of the seventy-three sauropsid genomes using the LocaTR pipeline. The total LTR retrotransposon content values were correlated with genome length, scaffold N50 length and contig N50 length. The number of SIEs was scaled by genome size and correlated with contig and scaffold N50 values. All metrics were $\log_{10}$ transformed for normality.

Total content values were mapped to a cladogram based on the known phylogeny. A GLM was constructed to identify significant variables which explained the observed distribution of LTR retrotransposon content. Total content was fitted as the response variable, and the model consisted of taxonomic groupings, scaled SIE values, contig N50 length and genome length (the last three as covariates). Eight taxonomic groupings were fitted to best match the avian lineages at the K/T extinction event: Paleognathae, Galliformes, Columbea, Caprimulgiformes and Otidimorphae, Cursorimorphae and Opisthocomiformes, Aequornithia and Phaethantimorphae, Afroaves, and Australaves.

Chapter 3: A new look at the LTR retrotransposon content of the chicken genome

3.1 Introduction

Since the release of the first draft of the chicken genome, there have been four major studies which have improved the knowledge of its overall LTR retrotransposon content. Initial work by the International Chicken Genome Consortium (2004) used homology-based approaches to identify ERVs from all three retroviral classes, and Wicker and colleagues (2005) used Cot-based cloning and sequencing (CBCS) to identify the highest copy number LTR retrotransposon elements in the chicken genome. These were initially classified as gypsy elements, but were later reclassified as endogenous GGERVL spumavirus sequences (Bolisetty et al. 2012). Work by Huda and colleagues (2008) and, most recently, Bolisetty and colleagues (2012), expanded identification by including structure-based methodologies which take advantage of the archetypal LTR retrotransposon structure. Overall, this enabled the characterisation of 492 structurally intact elements and the annotation of 1.35 % of the genome as LTR retrotransposon-derived elements.

Despite these studies, this overall figure remains three times lower than the content of mammalian genomes, even when scaled for genome size. In addition, Bolisetty and colleagues (2012) found that the Neoaves had LTR retrotransposon content much more comparable to mammals, and suggested that this 'deficit' of LTR retrotransposons in chicken may be specific to the Galliformes. So, have all the chicken LTR retrotransposons been successfully identified, and is the apparent paucity of these elements biologically representative or simply a result of an incomplete annotation?

Since the work of Bolisetty and colleagues (2012), a new chicken genome assembly has been released (Galgal4), and previous work has shown that there has been a marked increase in repeat content annotation with each revised assembly. In addition, previous identification studies have been heavily homology-based or have only used one structure-based identification program, so it is possible that whole subsets of elements, and those that are lineage-specific, may have been missed completely. It is therefore necessary to undertake a review of the available methodologies and undertake a new, more comprehensive identification of LTR retrotransposon sequence in the chicken genome.

3.1.1 Review of the available identification methodologies

Homology-based approaches

LTR retrotransposons have largely been studied independently in a species-specific manner, but the increased availability of genomic sequence in the last fifteen years has led to the development of multiple repeat databases. It is common practice to begin any *de novo* repeat annotation with RepeatMasker (Smit et al. 2013), which identifies all repetitive DNA in the genome assembly and classifies it according to the RepBase libraries (Jurka et al. 2005). However, it is usual to then further extend the annotation with custom-built libraries or standalone BLAST searches (Altschul et al. 1990) using full or partial reference sequences from databases such as HERVd (Paces et al. 2004), RetOryza (Chaparro et al. 2007), and GyDB (Llorens et al. 2011), depending on the desired study species.

This homology-based approach profits from such an extensive knowledge base, enabling confident LTR retrotransposon annotation even when present in low copy number or truncated forms. However, identified elements are a product of the reference sequences used to find them, and there is significant bias towards well described repeat classes from extensively studied species and younger, more complete insertions (Bergman & Quesneville 2007). Additionally, heterogeneous conservation of repetitive sequence returns fragmented hits, rather than complete elements, and there is often a bias towards retroviral-like domains such as *reverse transcriptase*, leading to the detection of other retrotransposing elements such as LINEs. Crucially, homology-based identification alone is non-exhaustive and is unlikely to detect lineage-specific repeats.

Structure-based approaches

The conserved archetypal structure of LTR retrotransposons enables element identification independent of sequence homology, by instead modelling the element based on various distance and similarity constraints. Such approaches initially identify LTRs by their U3-R-U5 conserved structure, including polyadenylation signals, transcription factor binding sites and the transcription start site, and their demarcation by short inverted repeats. Candidate LTRs are then paired on distance and similarity

constraints with annotated pairs only classified as LTR retrotransposons when validated by the presence of further motifs which satisfy location, size and reading frame requirements. This stipulation for identification of candidate LTR pairs means that intact LTR retrotransposon discovery is heavily dependent on assembly quality, particularly its contiguity and true resolution of high copy number repeats.

The first structure-based identification method, LTR_STRUC (LS) (McCarthy & McDonald 2003), set the trend for a range of related but increasingly sophisticated programs with equally unwieldy names. LS identifies putative LTRs by seeding random alignments between two points within a set distance constraint, and extending the alignment until scores drop beneath a predefined threshold. Alignment ends are fine-tuned by identification of the target site duplications (TSDs), which border the LTR, and the neighbouring Primer Binding Site (PBS) and Polypurine tract (PPT) at the inner edge of the 5' and 3' LTRs respectively. LS predominantly identifies LTR pairs with greater than 90 % homology, enabling high specificity, but with limited sensitivity. This can be improved with a poorly defined 'sensitivity scale', allowing identification of LTR pairs with 75 % identity, but with a large associated increase in the false positive rate.

Despite its release in 2003, LS continues to be used successfully for diverse LTR retrotransposon family annotation (McCarthy & McDonald 2004; Polavarapu et al. 2006; Huda et al. 2008; Garcia-Etxebarria & Jugo 2010; Garcia-Etxebarria & Jugo 2012), but is unable to complete exhaustive identification. Increased sensitivity and processing speed was introduced with the related LTR_par (Kalyanaraman & Aluru 2006) and the web-based LTR_FINDER (Xu & Wang 2007). These incorporated the use of suffix arrays for rapid seed extension, and the capability to manage multi-fasta files, span contig gaps and analyse the complementary strand; features lacking in LS which required additional accessory scripting.

LTR Harvest (LH) (Ellinghaus et al. 2008) advanced LTR retrotransposon annotation and largely replaced these latter two programs, using the search criteria of LTR_par but optimised for large genomes and memory efficiency. Additionally, users can specify a wide range of options to adapt the program's sensitivity, and completing the multiple analyses is fast, as the suffix array is created once and then stored. This enables the full analysis of gigabase size genomes with multiple parameter setups in minutes, rather than

hours or days, following one computationally expensive step. Superficially, the ease of performing multiple analyses may seem a bonus, but it is, in fact, a necessity. Slight parameter alterations have a dramatic effect on both the number of putative LTR retrotransposons identified and the observed false positive rate, the latter commonly reaching over 60 % (Lerat 2010), hence the 'suggestion' from the authors for 'thorough' putative element validation. Unlike all other available structure-based programs however, LH is actively curated and has associated validation software, LTR Digest (LD) (Steinbiss et al. 2009), which uses built-in and user-specified profile Hidden Markov Models (pHMMs) for the PPT and retroviral-like domains, and species-specific tRNA pHMMs to identify and categorise the PBS. Additionally, the identification (LH) and validation (LD) steps can be integrated into one analysis using the LTRsift pipeline (Steinbiss et al. 2012). However, validation is convoluted, requiring installation of multiple accessory programs, and remains highly subjective, based on arbitrary thresholds and the selection of 'appropriate' pHMMs. The latter is of great importance, as bias towards known sequence should be avoided. Ill-considered validation will likely increase the false negative rate without marked reduction of the false positive rate.

Besides this related family of programs, two others are commonly used. MGEScan_LTR (MGS; initially named LTR_Rho) (Rho et al. 2007) was developed to remove all sequence bias in LTR models and instead focus on motif structure, requiring 'maximal exact matches' (blocks of exact identity) of at least 40 bp between putative LTRs before extending the sequence alignment. This often limits LTR pair homology to 80 %, but all identified LTRs are clustered and pHMMs constructed for subsequent detection of related, but more divergent LTRs. The authors accept the potential for elevated false positive rates, but also identified that MGS detects a very different subset of LTR retrotransposons when compared with LS or LTR_par; a result also seen when later compared to LH (Lerat 2010).

Another, different identification approach is that of RetroTector (ReTe) (Sperber et al. 2007) which links retroviral-like motifs through 'fragment threading'. Whilst powerful, it is perhaps the most limited by existing sequence, as LTR searching depends on trained pHMMs and motif recognition is based on a series of sequence libraries. Initially designed for primate ERVs, the application of ReTe has diversified to other animal species (Garcia-Etxebarria & Jugo 2010; Garcia-Etxebarria & Jugo 2012), including

chicken (Bolisetty et al. 2012), but the reliance on motif libraries is evident: identification is biased towards young, structurally complete ERVs, rather than from any LTR retrotransposon clade. Perhaps because of this, ReTe has a relatively low false positive rate, at least in mammalian studies (Lerat 2010; Garcia-Etxebarria et al. 2014), aided by 'brooms' to mask ('sweep') LINEs and SINEs before the identification of putative LTRs. LTR searches favour speed over sensitivity, rapidly narrowing the search space for later computationally-intensive annotation of inner motifs. ReTe also identifies a markedly different element subset compared to MGS and the LS/LH family of programs.

Whilst these programs differ in the order and importance of the validating domains, and their computational complexity, they all share the initial requirement for candidate LTR pair identification. Truncated or solo LTR elements will not be detected by these methods and neither will LTR retrotransposons which lack the archetypal structure, such as DIRS elements for which a specific, homology-based approach has been developed (ReDoSt; Piednoël et al. 2011). Efforts to develop novel structure-based methods have not been hugely successful. Benachenhou and colleagues (2009a; 2009b) designed pHMMs to specifically identify retroviral LTRs alone. Whilst models fitted individual retroviral genera well, wider application showed that LTRs outside the training groups could not be identified. Furthermore, analyses were computationally complex and had high false positive rates as conserved LTR features, such as polyA regions and promoters, are common to other genomic features. In contrast, Ashlock and Datta (2010) focused on internal domain reading frame patterns and compared these to patterns observed in host coding and non-coding DNA. Sensitivity for retroviral domain detection was reported to be greater than 92 %, but interpretation was complex and limited by both frameshift mutations common in degraded repeats and the specific genera included in model training. Concordantly, the well-established structure-based methodologies remain the only viable option for genome-wide annotation of all LTR retrotransposon types, as long as they remain structurally intact.

Combining identification strategies

Homology-based identification is inherently biased towards known sequence, and whilst structure-based methods can identify diverse or lineage-specific LTR retrotransposons,

they also use motifs and models based on known sequence, and have a high false positive rate. It should also be noted that the implementation of all these tools is complicated, requiring installation of accessory software, management of SQL databases, and the writing of purpose-built scripts to control batch analysis and program-specific issues such as memory allocation. Analysis is further hindered by incomplete or confusing manuals and unavailability of the source code, as well as non-existent curation and user support, with LH a stark exception. As such, LTR retrotransposon annotation is daunting, particularly as the combination of multiple methodologies is necessary for complete annotation and the avoidance of program-specific biases (Lerat 2010; Garcia-Etxebarria et al. 2014).

Recent work combining two (Barrio et al. 2011; Bolisetty et al. 2012) or three (Garcia-Etxebarria & Jugo 2010; Garcia-Etxebarria & Jugo 2012) identification strategies in animal genomes has highlighted the limited redundancy between the elements identified by the different methodologies, commonly with more than 80 % of identified elements unique to one approach. Despite this, previous work in the chicken has had a strict *requirement* for redundancy between programs to validate identified LTR retrotransposons It is therefore highly likely that many genuine elements have been ignored rather than annotated, which may, in part, contribute to the previously reported 'deficit' of these elements in the chicken when compared with mammals and Neoaves.

3.2 Research Aims

This chapter covers three major research aims. Firstly, the development of a user-friendly bioinformatic pipeline to combine multiple LTR retrotransposon identification strategies into a single integrated approach. Secondly, the use of this pipeline to identify the LTR retrotransposons of the chicken genome to determine whether previous, incomplete annotation was responsible for the observed deficit of these repeats in chicken compared to other birds. Finally, to fully characterise the LTR retrotransposon distribution relative to genes, recombination rate and other LTR retrotransposons, and to assess the extent of LTR retrotransposon expression.

3.3 Statement of publication

The results presented in this chapter have been published: Mason AS, Fulton JE, Hocking PM & Burt DW (2016), A new look at the LTR retrotransposon content of the chicken genome, *BMC Genomics*, **17**:688. The paper has been included on the CD accompanying this thesis in Appendix 3: Paper2.

Figures, tables and text have been reworked to fit this thesis and further methodology optimisation, results, figures and tables have been included. The final section of the paper on the LTR retrotransposons of the avian phylogeny forms the preliminary lineage analysis in Chapter 4 of this thesis (section 4.5.1, page 5).

3.4 Development of the LocaTR identification pipeline

LTR retrotransposon identification was facilitated by the development of the LocaTR identification pipeline. This user-friendly, well documented pipeline consists of four broad parts: genome pre-processing, structure-based identification, homology-based identification, and the confirmation of putative LTR retrotransposons by removing false positives. LocaTR can be used to identify LTR retrotransposons in any assembled genome and can be easily adapted to include additional search programs, if required.

The combination of extensive documentation and intermediary scripts has made running the disparate identification programs easier, whilst retaining the ability for the user to alter individual parameters, although recommendations have been made in the documentation based on this work. Most of the programs and scripts have been written to work on a Linux server, but both LS and ReTe must be run on a desktop, specifically Windows for LS. If the homology and structure-based identification programs are run in parallel, a full LTR retrotransposon identification of a 1 Gb genome can be completed within 10 days.

LocaTR scripts and documentation

Seventeen intermediary scripts and six accessory scripts were written to manage the LocaTR identification pipeline. The names and functionalities of these scripts are

detailed in Table 3.1 and Table 3.2 respectively. Each script has its own extensive documentation, help messages and error catches, and additionally each of the four broad sections of LocaTR has its own documentation to aid the user. A detailed flow chart for LocaTR is shown in Figure 3.1. All scripts are on the CD accompanying this thesis and in a GitHub repository (https://github.com/andrewstephenmason/LocaTR).

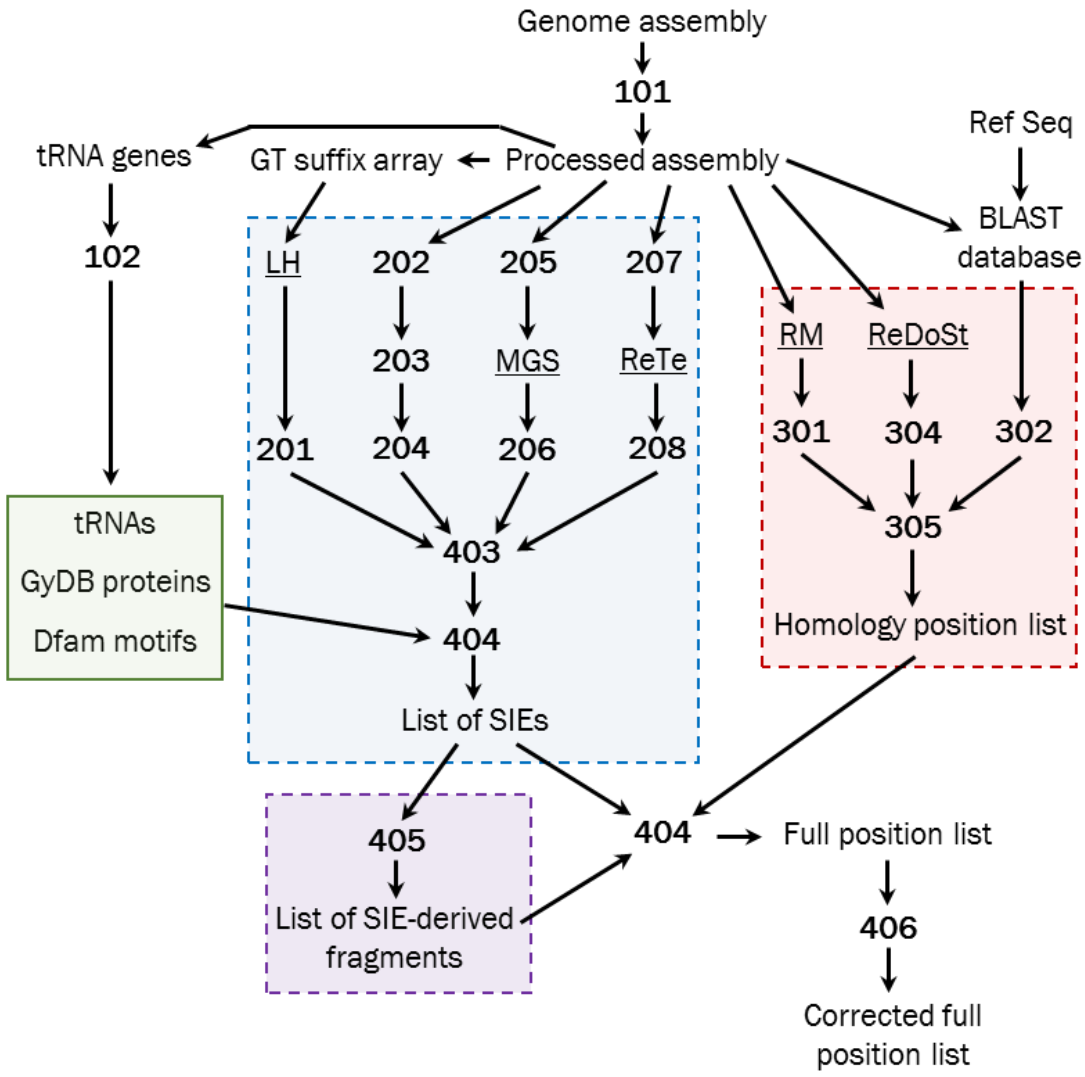


Figure 3.1 The LocaTR identification pipeline workflow. Each script run by the user is shown by its number (as described below in Table 3.1), and the independently run programs are underlined (LH = LTR Harvest; MGS = MGEScan_LTR; ReTe = RetroTector; RM = RepeatMasker). LTR_STRUC is not shown as its execution is completely controlled and hidden with script 203. Scripts for each LocaTR section are enclosed by boxes with dashed borders: homology (pink), structure-based (blue) and secondary BLAST (purple). The pHMMs used for the SIE feature tests are enclosed by the green box with the solid border.

Table 3.1 The LocaTR intermediary scripts.

Script name	Functionality
101_format_genome_file.py	Formats sequence headers ready for analysis
102_extract_tRNA_seq.py	Takes tRNAs identified by tRNAscan-SE, extracts sequences and create pHMMs
201_extract_LH_positions.py	Extracts positions and sequences following LTR Harvest analysis
202_LS_seq_formatter.py	Creates individual sequence files from the genome multi-fasta, and reverse complements
203_LS_ltrstruc_batch.py	Manages the LTR_STRUC process on Windows
204_LS_batch_pos_extract.py	Processes the sequences identified by LTR_STRUC and generates a positions file
205_MGS_seq_formatter.py	Creates individual sequence files from the genome multi-fasta
206_extract_MGS_positions.sh	Extracts positions and sequences following MGEScan_LTR analysis
207_rete_input_fasta_formatter.py	Creates individual sequence files from the genome multi-fasta and pads sequences shorter than 30kb
208_extract_ReTe_positions.py	Extracts positions and sequences following RetroTector analysis
301_extract_RM_positions.sh	Extracts positions and sequences following RepeatMasker analysis
302_refBLASTsearch.py	Performs the BLASTn/tBLASTx protocol and putative element cross-validation
304_extract_dirs_positions.py	Extracts positions and sequences following ReDoSt analysis
305_merge_homology_positions.sh	Merges all identified locations from the homology based searches
403_validate_SIEs.py	Performs feature analysis for putative SIE support and confirmation
404_merge_positions_list.sh	Merges all identified locations
405_secondary_BLAST_analysis.py	Performs the secondary BLASTn/tBLASTx protocol using confirmed SIE sequences
406_convert_to_original_contigs.py	Converts all sequence and position files back to original contig names

Table 3.2 The LocaTR accessory scripts.

Script name	Functionality
000_modify_paths.py	Matches LocaTR filepaths to host directory setup
001_seq_extract.py	Creates multi-fasta files from positions list
002_pos_merger.py	Merges positions lists
003_custom_rm_processor.py	Prepares custom RepeatMasker analysis output and for putative element cross-validation
004_rm_fragment_remover.py	Identifies non LTR retrotransposon repeats in putative sequence and removes them
static_functions.py	Series of functions for merging, sequence extraction, list formatting *etc.* called by scripts

3.4.1 Initial optimisation of individual identification programs

RepeatMasker

RepeatMasker (RM) analysis was performed on the Galgal4 assembly under default conditions, and by specifying the RepBase libraries for chickens and, more widely, for vertebrates. Under default conditions (human repeat libraries) 6.19 % of the genome was annotated as repeats, with only 0.13 % of the genome attributed to LTR retrotransposon derived elements. By specifying the chicken RepBase libraries, this value increased to 9.96 % of the genome for all repeats, and to 1.66 % of the genome for LTR retrotransposons, made up of some 32,674 independent elements, covering 17.38 Mbp of the genome. By widening the available RepBase libraries to all vertebrate entries, the annotated repeat content increased again to 10.41 %, and the LTR retrotransposon annotation increased to 1.78 % of the genome (Table 3.3). The addition of the vertebrate RepBase libraries increased the amount of LTR retrotransposon sequence identified, and annotated shorter, potentially more fragmented and divergent sequences, as the average identified repeat length fell by 22.5 % from 532 bp to 412 bp. Due to the identification of more LTR retrotransposon sequence, all further RM analyses used the '-species vertebrates' flag.

Table 3.3 Annotated Galgal4 repeat content using RepeatMasker with three different available RepBase libraries.

RepBase library	Total repeat content (%)	LTR Content (%)	LTR Content (Mbp)	# LTR elements
Default	6.19	0.13	1.34	2,929
Chicken	9.96	1.66	17.38	32,674
Vertebrates	10.41	1.78	18.64	45,223

Additionally, RM has three built-in sensitivity settings which were tested and compared to the default (Table 3.4). As reliable repeat discovery was the aim of this study, rather than genome masking for mapping or gene annotation, the reduction in computational time offered by the 'q' and 'qq' sensitivity settings did not justify the observed reductions in annotated content. Whilst the 's' (slow/sensitive) setting did identify more LTR retrotransposon-homologous sequence, it was accompanied by a large increase in computational time for relatively little gain. RM analyses are used several times in LocaTR so any increase to computational time would be multiplied through the process. Consequently, all RM analyses were implemented with default sensitivity.

Table 3.4 The effect of the different RepeatMasker specificity settings on detected LTR retrotransposon content and processing time. The predicted sensitivity and processing time effect from the manual are shown in columns 2 and 3. Increased sensitivity is shown with a '+', and decrease with a '-'. Processing time increase is denoted by 'S' (for 'slower') and decrease with 'F' (faster).

Setting	Sensitivity effect	CPU effect	Content (%)	Content (Mb)	# elements	CPU time
Slow (s)	+ 0-5%	2-3x S	1.80	18.82	46,663	>10 days
Default	0	0	1.78	18.64	45,223	2.5 days
Quick (q)	- 5-10%	2-5x F	1.74	18.17	41,902	1.5 days
Rush (qq)	- 10%	4-10x F	1.66	17.43	37,998	0.5 days

LTR_STRUC

The ten optional LS sensitivity settings had a marked impact on the number of identified elements, the false positive rate, and the processing time for each analysis during preliminary analysis of chromosomes 1 and Z (Table 3.5). Additional elements with feature support were identified at each sensitivity setting, but the processing time increased by a factor of 1.5-2.5 for each higher setting. However, as the highest setting still identified confirmed new LTR retrotransposons, setting 1 was used for all further LS analyses. On both chromosomes, confirmed elements were significantly younger, based on LTR identity, than the identified false positives ($t = 6.80$, $p < 0.0001$). The lengths of the confirmed elements were also significantly shorter than the identified false positives on chromosome 1 ($t = -3.43$, $p = 0.0006$), but there was no difference on the Z chromosome.

This preliminary work also identified that the 'default' sensitivity setting was the same as sensitivity setting 4, a piece of information which had been lost to the program curators.

Table 3.5 Sensitivity testing with LTR_STRUC using Galgal4 chromosome 1 and Z.

Sensitivity setting	Processing time (mins)	Identified elements	Confirmed elements	False positive rate (%)
1	2,523.45	101	46	54.4
2	1,286.13	99	45	54.5
3	636.17	87	43	50.6
4	268.52	77	41	46.8
default	261.12	77	41	46.8
5	163.37	64	39	39.1
6	85.00	52	34	34.6
7	45.05	43	31	29.5
8	24.78	33	22	33.3
9	14.87	21	13	38.1
10	9.75	10	5	50.0

Analysis of the full assembly with LS revealed that there were issues with memory allocation, as the program would finish after only analysing approximately 400 Mb of the genome. Additionally, the program did not consider the reverse strand, so elements were missed unless there was a clear signal in the reverse orientation. Concordantly, those elements which were detected on both strands were significantly younger insertions ($t = 8.86$, $p < 0.0001$), and were highly represented in the confirmed list after feature analysis. For example, a default analysis of chromosome 1 identified 38 LTR retrotransposons, but running on the reverse strand identified 45, where 24 were shared between the runs, giving a total of 59 putative elements, of which 25 (42.4 %) were confirmed by feature analysis. A script was written to separate the multi-fasta assembly file into individual sequence files, and to create reverse complemented sequences ready for analysis (202_LS_seq_formatter.py). A batch processing script was also created to manage passing multiple sequences to the LS executable (203_LS_ltrstruc_batch.py).

LTR Harvest

LTR Harvest (LH) identified the largest number of confirmed elements from the structure-based programs, but the false positive rates were strikingly high for all parameter settings tested, including the three sets presented in Table 3.6. Enabling LH to identify all potential LTR pairs for the identification of nested elements (set 2) did not identify any additional confirmed elements, but considerably increased the standard processing time for the feature annotation analysis. In fact, the seven cases of nested LTR retrotransposons identified in the Galgal4 assembly were all identified from the set 1 analysis. Additional parameter tests were completed (altering LTR identity and length parameters), but no other setup identified more confirmed LTR retrotransposons.

Table 3.6 Parameter optimisation with LTR Harvest

Parameter set	Identified elements	Confirmed elements	False positive rate (%)
Default	3,707	527	85.8
1	5,264	643	87.8
2	65,422	643	99.1

To determine whether the false positive rate was directly linked to putative LTR pair sequence identity, the set 1 parameters from above were used for ten additional runs where the LTR homology requirement was gradually increased (Table 3.7). False positive rates remained high throughout, despite the increasingly stringent requirements for near perfect LTR pair identity.

Table 3.7 The impact of increasing the LTR pair identity on the false positive rate of LTR Harvest

LTR identity (%)	Identified elements	Confirmed elements	False positive rate (%)
75 (set 1)	5,264	643	87.8
85	3,792	559	85.3
87	3,173	527	83.4
90	2,337	525	77.5
92	1,012	521	49.5
95	958	497	48.1
96	735	379	48.4
97	531	275	48.2
98	293	159	45.7
99	112	73	34.8
100	30	27	10.0

This again highlights the need for strict feature annotation for confirmation of the putative LTR retrotransposons. Near identical false positive rates were observed with the GenomeTools validation program LTR Digest (Steinbiss et al. 2009), suggesting that these are indeed issues with LH false positive rates, rather than over-conservatism in the feature annotation analysis. LTR Digest was not used as the validation methodology for all structure-based programs as it required putative LTR retrotransposons in strict GFF3 format, which proved difficult to reliably produce from the other programs. Additionally, construction of new scripts to confirm the putative LTR retrotransposons enabled a fresh review of the thresholds used for identifying these elements.

3.5 The LocaTR analysis of the Galgal4 chicken assembly

In total, 31.5 Mb of the chicken genome was identified as LTR retrotransposon-derived sequence, accounting for 3.01 % of the Galgal4 chicken genome assembly. This comprised 36,109 annotated regions of which 1,073 were structurally intact elements (SIEs): more than double the number previously reported (Bolisetty et al. 2012). Of this total, the expanded homology protocol identified over 20.3 Mb of sequence, almost 4 Mb more than was annotated by RM using the chicken RepBase libraries (Table 3.3). The structural-based methodologies identified 9.1 Mb, 45.8 % of which was 'novel', having not been identified in the homology-based searches. The secondary BLAST analysis for fragments related to annotated SIEs identified an additional 7.1 Mb of sequence (Figure 3.2).

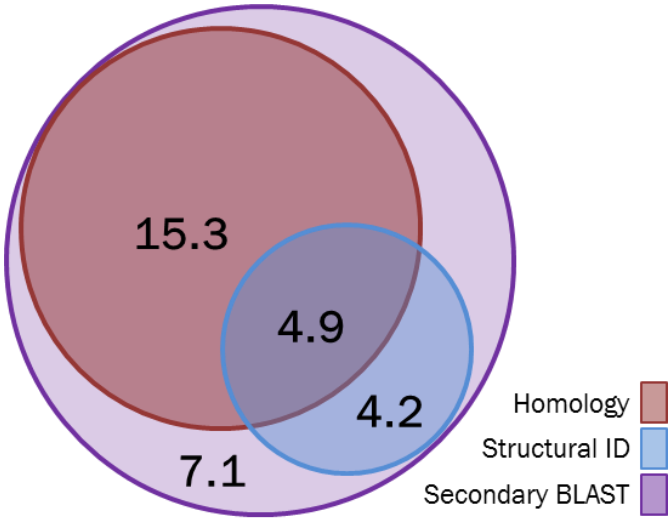


Figure 3.2 Performance of the homology and structure-based identification methodologies. Euler diagram representing the relative proportion of LTR retrotransposon content identified by the homology (red), structure (blue) and secondary BLAST (purple) modules of the LocaTR pipeline. Numbers represent the total length in megabases (Mb); 31.5 Mb in total. Homology methods identified 20.3 Mb of sequence, and structure-based ID methods 9.1 Mb including 4.9 Mb (54.2 %) overlap with the homology data. The secondary BLAST identified an additional 7.1 Mb based on elements from the structure-based methods.

Most of the 36,109 annotated LTR retrotransposon-derived regions are fragmented, with an average length of 0.9 kb and high standard deviation (2.1 kb) reflecting the large

sequence structural variability. SIEs also exhibit large element size variation (mean length 8.5 kb, standard deviation 6.5 kb), some of which can be accounted for by the seven examples of nested LTR retrotransposons, where one element has inserted within another resulting in the relative elongation of the outer element.

All Galgal4 structurally intact LTR retrotransposons were ERVs, predominantly from the betaretrovirus, gammaretrovirus and spumavirus retroviral genera. These are the shared ancestral ERV genera amongst vertebrates (Herniou et al. 1998; Borisenko 2003), and it is likely that the bias of mammalian sequences in the RepBase vertebrate libraries aided their identification and classification. A total of 65.7 % of SIEs could be classified by protein-coding domain homology. Of these, over a third were ERVs from the alpharetrovirus-betaretrovirus clade, generally the youngest and therefore most easily detectable group, including the assembled ALVE-RJF on chromosome 1. Consistent with previous publications, no elements were identified from either the Bel/Pao, DIRS, Gypsy or Copia groups of LTR retrotransposons (Huda et al. 2008; Piednoël et al. 2011; Bolisetty et al. 2012) and there was no evidence of deltaretrovirus or lentivirus ERVs.

3.5.1 Characterisation of the structurally intact LTR retrotransposons

Of the 1,073 SIEs, only 291 (27 %) were identified by two or more programs, with only seven SIEs identified by all four (Figure 3.3). With a strict requirement for identification by at least two structure-based programs, 2.8 Mb (67.14 %) of the novel annotated sequence identified by these four programs would have been missed.

Despite low cross-program corroboration, there appears to be no specific program biases for GC content, inner structural intactness, length, element age (shown in Table 3.8 as mean LTR identity), or classified genera, as had been proposed in the annotation of other species (Garcia-Etxebarria & Jugo 2010; Garcia-Etxebarria & Jugo 2012). All programs had SIE length distributions skewed by identification of very long LTR retrotransposons. The high percentages of SIEs unique to each program exemplifies the necessity of using multiple identification approaches. However, the high false positive rates mean stringent confirmation of putative SIEs is essential.

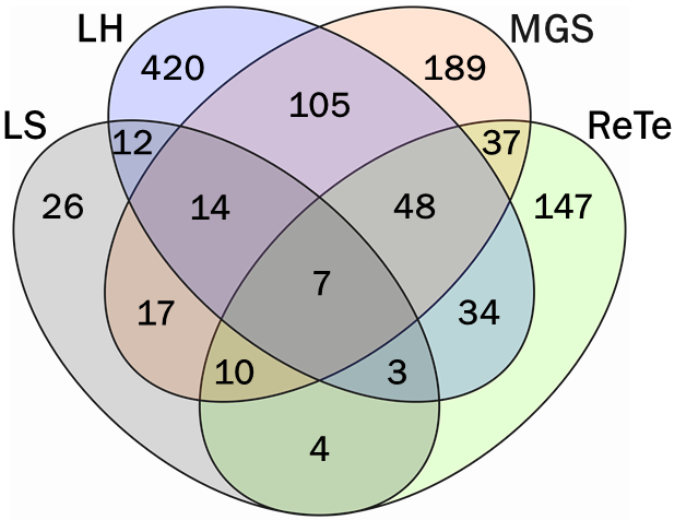


Figure 3.3 Four-set Venn diagram of the overlap between structurally intact LTR retrotransposons identified by the four structure-based identification methods. In general, there was little overlap between programs, with 782 elements (72.9 %) unique to one program. Each ellipse corresponds to one program (with LS in grey, LH in blue, MGS in orange and ReTe in green). The numbers represent the number of intact elements identified by that program or group of programs, with the ellipse overlaps representing shared identification.

Table 3.8 Comparison of intact LTR retrotransposons (SIEs) features identified by the four structure-based identification programs.

SIE program results	LTR_STRUC	LTR Harvest	MGEScan_LTR	RetroTector
Initial SIEs identified	299	5,264	523	567
SIEs with feature support	93	643	427	290
False positive rate (%)	68.9	87.8	18.4	48.9
Total SIE content (bp)	767,132	4,837,212	4,928,810	2,664,622
Mean SIE length (bp)	8,249	7,523	11,543	9,188
Median SIE length (bp)	6,144	6,047	7,889	7,477
Median SIE LTR identity (%)	95.8	95.2	91.5	94.0
Mean SIE GC content (%)	48.1	47.2	45.7	46.3
SIEs unique to program (%)	28.0	65.3	44.3	50.7

SIEs tend to represent recent insertions and accordingly exhibit LTRs with less than 10 % sequence divergence, supporting insertion less than 13.5 million years ago (MYA) (95 % confidence range: 12.5-14.7 MYA). Nearly 90 % of all identified SIEs have inserted since the separation of the chicken and turkey lineages (27.0 MYA, 95 % confidence range: 25.0-29.4 MYA) (Helm-Bychowski & Wilson 1986). In addition, SIE GC-content (46.9 %) was not significantly higher than the genome average of 41.8 %. However, variation in GC-content is explained by SIE insertion age, as absolute element GC-content deviance from the genomic mean decreases for older insertions ($r = 0.45$, $P < 0.001$, Figure 3.4).

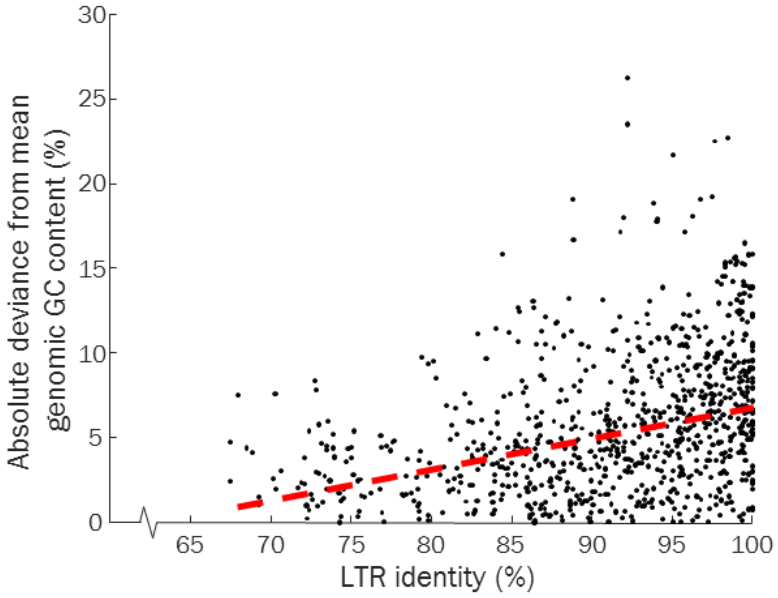


Figure 3.4 Absolute SIE GC content deviance from the genomic mean as a function of their age, measured by LTR pair identity. The red dashed line is the Pearson correlation ($r = 0.45$, $P < 0.001$). Younger insertions have a more deviant GC content, relative to the genomic average, than older elements.

3.5.2 Changes in LTR retrotransposon annotation since the previous genome assembly (Galgal3)

Some of the annotation improvement can be attributed to the improved quality of the Galgal4 assembly, as previous work used Galgal3 which had a four times higher proportion of ambiguous bases, a contig N50 of only 46.4 kb and significant assembly

errors on the Z chromosome. However, most improvements can be directly attributed to the use of the LocaTR pipeline, due to the enriched reference sequence database and reduced conservatism during SIE identification (Table 3.9). Improvements between the assemblies only accounted for an extra 2.5 Mb of annotated LTR retrotransposons. Use of LocaTR identified a further 14.1 Mb including an additional 587 SIEs not found by the previous RetroTector analysis (Bolisetty et al. 2012).

Table 3.9 Comparison of LTR retrotransposon annotations between chicken genome assemblies highlighting improvements made with the LocaTR pipeline.

Assembly feature	Galgal3	Galgal4 (chicken RM)	Galgal4 (LocaTR)
Assembly length (bp)	1,098,770,941	1,046,932,099	1,046,932,099
Scaffold N50 (bp)	11,063,745	12,877,381	12,877,381
Contig N50 (bp)	46,345	279,750	279,750
LTR content (bp)	14,870,595	17,369,358	31,490,117
LTR content (%)	1.35	1.66	3.01
Number of SIEs	492	-	1,073

3.6 Analysis of the LTR retrotransposons of Galgal4

3.6.1 LTR retrotransposon density

LTR retrotransposon density has a strong, positive correlation with chromosome size ($r = 0.91$, $P < 0.001$), and, consequently, a strong negative correlation with recombination rate ($r = -0.81$, $P < 0.001$) and gene density ($r = -0.72$, $P < 0.001$). Chromosome size was the only significant variable when fitted to the GLM ($P < 0.001$), but as recombination rate is scaled by sequence length (centimorgan per Mb; $cM.Mb^{-1}$) this remains an important contextual relationship. Chromosomes 16, 25 and W exhibited much higher than expected element density (Figure 3.5), but this was likely due to large amounts of missing sequence from the assemblies caused by unassembled, collapsed or unsequenced repetitive elements. However, chromosomes 27 and Z are much more complete and have high density relative to their length and recombination rate.

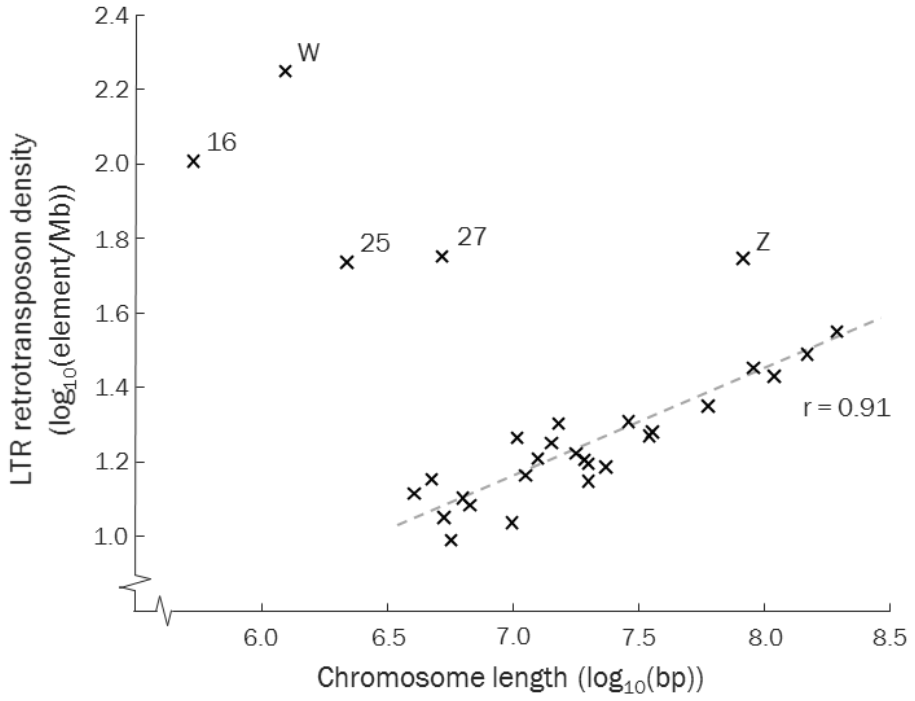


Figure 3.5 Correlation between the chromosome length and its LTR retrotransposon density, where both measures have been log_{10} transformed. The dotted grey line represents the strong, positive correlation when named outliers were removed ($r = 0.91$, $P < 0.001$). Chromosomes 16, 25 and W have shorter assembly lengths than the known physical distance (Masabanda et al. 2004). Chromosomes 27 and Z have densities greater than expected for their length.

Identification of intact LTR retrotransposon clusters

Whilst most LTR retrotransposons were found on the chicken macrochromosomes, intra-chromosomal element density was highly heterogeneous. There were large regions of the genome devoid of intact elements, but 40.3 % of all SIEs were found within clusters (432 elements in 28 clusters) unrelated by insertion age or genera (Appendix 2; AF03). This was significantly higher than expected under random integration, where only 6.49 % of SIEs fall within clusters ($P = 1.58 \times 10^{-30}$). Cluster size varied from the minimum defined of five, up to a cluster which contained all fifty-six identified SIEs on chromosome W (Table 3.10).

The suitability of using a cluster cut-off of 5 elements per Mb (chosen to match the work of Bolisetty et al. 2012) was tested by doubling the required frequency to 10 elements per Mb (matching the relative density cut-off of Bolisetty et al. 2012). This caused a

reduction of the number of SIEs within clusters (down to 343; 32.0 %) but all previously identified cluster regions were retained, suggesting that all were genuine clusters rather than random noise.

Table 3.10 Identified LTR retrotransposon clusters in the Galgal4 assembly. Cluster coordinates and contained SIE features are shown in Appendix 2; AF03.

Chromosome/Contig	# Clusters	# SIEs in clusters	Cluster sizes
1	7	107	32, 8, 7, 23, 9, 5, 23
2	2	52	46, 6
4	2	30	22, 8
5	2	21	15, 6
8	1	5	-
16	1	8	-
27	1	8	-
W	1	56	-
Z	5	103	15, 29, 10, 41, 8
AADN03011155.1	1	6	-
JH375232.1	1	6	-
JH375233.1	1	7	-
JH375236.1	1	11	-
JH376410.1	1	7	-
LGE64	1	5	-

Clusters were commonly associated with regions of elevated fragmented LTR retrotransposon density, suggesting the persistence of these seemingly favourable areas over time. There were also several examples of regions with a high density of fragmented LTR retrotransposons linking two separately identified SIE clusters, most notably the two clusters near the centromere of chromosome 4. Almost all clusters were in regions of low recombination relative to the chromosome average, which likely facilitated the structural longevity of elements in these regions (Table 3.11). However, this correlation

was limited by the lack of data for chromosome 16, the unassembled contigs and the non-recombining W. Clusters on chromosomes 1, 2, 4 and 8 encompassed the poorly recombining centromeric regions (asterisked clusters in Table 3.11). Contrastingly, the cluster on chromosome 5 which encompasses the centromere has a 500 kb bin

Table 3.11 Observed recombination rates in assembled chromosome clusters. The clusters indicated with an asterisk overlap the centromeres.

	Recombination rate (cM.Mb^{-1})	
Cluster location	500kb-bin average	Chromosome average
1: 72,775,206 - 76,073,001 *	0.16	2.10
1: 99,226,520 - 100,570,014	0.60	2.10
1: 140,707,522 - 141,575,378	0.16	2.10
1: 147,783,930 - 151,568,638	0.11	2.10
1: 152,941,395 - 154,373,881	0.00	2.10
1: 157,335,033 - 157,917,833	0.00	2.10
1: 158,996,141 - 161,747,651	0.12	2.10
2: 52,169,051 - 56,240,034 *	0.26	1.80
2: 132,517,171 - 133,448,919	0.10	1.80
4: 18,033,066 - 19,885,008 *	0.00	2.10
4: 21,732,969 - 22,756,182	0.00	2.10
5: 2,544,652 - 4,179,090	0.00	2.50
5: 5,395,023 - 5,835,541 *	3.77	2.50
8: 9,568,649 - 10,505,342 *	0.00	3.20
27: 42,531 - 857,048	0.70	10.80
Z: 26,957,545 - 27,971,718	1.70	3.00
Z: 41,956,846 - 44,554,708	0.13	3.00
Z: 48,504,870 - 49,571,015	0.40	3.00
Z: 72,996,715 - 78,405,920	0.60	3.00
Z: 79,462,456 - 80,646,626	0.00	3.00

recombination rate higher than the chromosomal average, but this could be due to averaging out across the region, or assembly issues with the centromere altering the observed distance.

Despite the apparent longevity of these cluster containing regions, there was no evidence of insertion age bias in the clusters compared to those without. However, there is an overall bias in the dataset due to the higher proportion of younger SIEs which might mask any age-related effects. Additionally, there was a significant under-representation of both sauropsid ($\chi^2 = 8.41$, $P = 0.004$) and amniote ($\chi^2 = 3.95$, $P = 0.047$) constrained elements within the cluster regions.

The eight clusters without recombination rate information were located at: 16: 230,428-498,782; W: 9,649-1,241,834; AADN03011155.1: 4,562-75,119; JH375232.1: 2,446-218,948; JH375233.1: 23,167-125,879; JH375236.1: 5,242-184,598; JH376410.1: 2,512-82,932; LGE64: 925-799,254.

Macrochromosome-like LTR retrotransposon density on chromosome 27

Chromosome 27 element density was four times higher than on the similarly sized chromosomes 26 and 28, and included eight SIEs. However, its overall length, GC-content and average recombination rate are consistent with these neighbouring chromosomes. Additionally, the chromosome shares a 1:1 synteny with the turkey (*Meleagris gallopavo*) chromosome 29 and zebra finch (*Taeniopygia guttata*) chromosome 27, so this relatively elevated LTR retrotransposon content was not a result of a recent macrochromosome fusion event.

On closer inspection, 89 % of all elements on chromosome 27 were within 1 Mb of the 5' telomere, including a cluster of all eight SIEs (Table 3.11). Whilst the average chromosome recombination rate is 10.8 $cM.Mb^{-1}$, the 5' 1 Mb has a recombination rate of 0.7 $cM.Mb^{-1}$, fifteen times lower than the chromosomal average. Interestingly, the gene density for this region is consistent with the rest of the chromosome, but almost all the genes are members of the *feather beta-keratin* family and share almost 100 % identity with each other. In birds, the main beta-keratin family is on chromosome 25, but recent work identified monophyletic expansions of the feather-specific family members on

chromosomes 2 and 27, shared among birds but absent in crocodilian outgroups (Greenwold & Sawyer 2010; Ng et al. 2014; Greenwold et al. 2014). 61 *beta-keratin* paralogues form a tandem array at the 5' end of chromosome 27, and these are surrounded by, but not overlapping with, the high density LTR retrotransposon-derived sequences. Greenwold & Sawyer (2010) originally predicted that the high microchromosome recombination rates could generate these large paralogue tandem arrays, but, as has been stated above, the recombination rate for this region is very low. It is therefore possible that retrotransposition events facilitated the expansion of these paralogous genes, and the low recombination rates enabled the LTR retrotransposons to retain their structural intactness.

Elevated LTR retrotransposon density on the Z chromosome

A total of 6.21 % of the Z chromosome was annotated by LocaTR as LTR retrotransposon-derived sequence. In contrast, chromosome 4 has a similar length to the Z, but only 1.78 % of its total length was identified as LTR retrotransposon-derived sequence. As a sex chromosome, the Z only recombines in males and the recombination rate is highly heterogeneous, with long regions of low recombination which are known to facilitate both repetitive element persistence and retention of structural identity (Bergero & Charlesworth 2009). This is supported by 103 of the Z chromosome SIEs (61.3 %) being in clusters, with recombination rates in these regions strikingly lower than the chromosomal average (Table 3.11). Based on the correlation between element density and chromosome length for the majority of the genome, the predicted density for the Z would be 27.4 elements per Mb, compared to the observed value of 55.3 elements per Mb. Therefore, the Z has 101.8 % more LTR retrotransposon-derived elements than would be expected on an autosome of equal length.

As a converse example of the relationship between density and recombination rate, the highly recombining pseudoautosomal region (PAR) of the Z chromosome has just 2.7 % of the expected LTR retrotransposon sequence given its length and the average chromosomal element density.

3.6.2 LTR retrotransposon distribution relative to transcriptional units

Assuming random integration, 51.3 % of all LTR retrotransposons would be expected within transcriptional units (TUs), of which 92.0 % would be within introns. However, there is a significant depletion of LTR retrotransposons in TUs in both the full data set (31.3 %; P = 1.94×10^{-3}) and SIE-only data set (35.8 %; P = 3.90×10^{-4}) (Figure 3.6). This skewed result is observed with the overall distribution relative to nearest gene annotation, with both the full (KS = 0.139, P = 1×10^{-100}) and SIE (KS = 0.146, P = 1.28×10^{-17}) data sets exhibiting significant shifts away from the TUs.

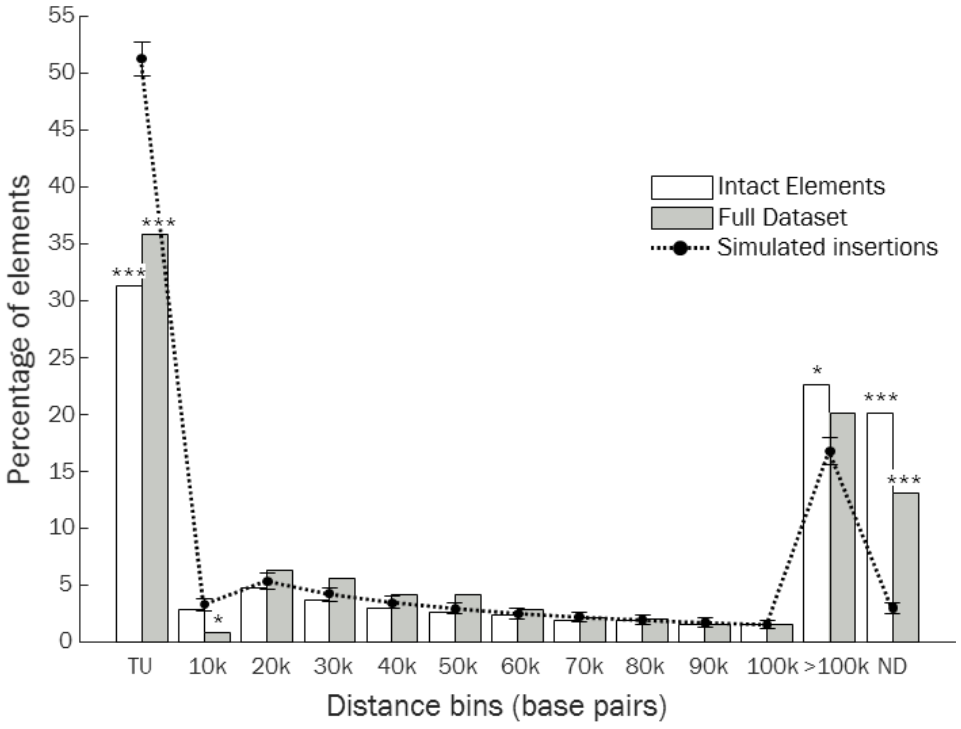


Figure 3.6 LTR retrotransposon distribution relative to the Ensembl genome annotations (v79). Shortest distance was measured from each element to the nearest annotated feature (irrespective of strand) and grouped into 10 kb bins, where the bin value represents the upper bin limit. For comparison, the dotted line represents 100,000 randomly generated distributions for each dataset, with standard deviation shown with the error bars. Only one line is shown because the randomly generated full and intact models gave the same results, and the standard deviation was equal between models when rounded to two decimal places. Significant differences between proportions in each bin are highlighted with asterisks, where * = P < 0.05 and *** = P < 0.0001. TU = Transcriptional Unit (incl. exons, introns, UTRs and 5 kb flanks up and downstream). ND = Non-Defined (elements on contigs without any Ensembl annotation).

Taken genome-wide, some of the distribution detail is overlooked. Microchromosomes generally follow the random integration distribution, but the longer macrochromosomes exhibit the depletion of elements within TUs (*e.g.* chromosome 1; 28.3 % in TUs compared to 47.2 % under random integration; $P = 3.88 \times 10^{-5}$). Additionally, chromosomes 1-5, 8 and Z have significant enrichment of elements greater than 100kb away from TUs (*e.g.* chromosome 1; 42.8 % compared to 23.0 % under random integration; $P = 1.17 \times 10^{-5}$). These chromosomes are also gene sparse and contain 73.6 % of the clustered SIEs (318 elements in 20 clusters). SIEs within clusters are also significantly depleted within TUs ($P = 1.60 \times 10^{-7}$) and enriched greater than 100kb away ($P = 0.001$) relative to the observed SIE distribution.

Together these data suggest that new insertions within TUs are selected against, and that accumulation is tolerated primarily in the poorly recombining, gene sparse regions of the genome where clusters can form and persist over long evolutionary timescales due to limited selective constraints. Consequently, SIE distribution should be age dependent, with new insertions following a random distribution and older elements skewed away from the TUs. Whilst there is some evidence that SIEs within clusters are generally older than those outside, there is no suggestion that SIE age distribution differs from randomly generated redistributions. This analysis is, however, confounded by the dominating proportion of 'young' SIEs (70 % LTR identity or greater; 96.52 % of all SIEs). Older elements alone exhibit depletion in TUs and enrichment greater than 100 kb away, but the sample size was too small for statistical robustness.

Overlaps with transcriptional units

Despite the evidence for enrichment away from transcribed regions, 31.3 % of all elements and 35.8 % of SIEs overlap TUs. Under random integration 4.9 % of insertions should be within exons, 3.1 % within UTRs, and the remaining 92.0 % within introns. However, the full dataset has significant enrichment of exon overlaps (10.1 %; $P = 0.015$), and the effect is greater in the SIE dataset with significant enrichment in both exons (38.3 %; $P = 4.1 \times 10^{-24}$) and 5' UTRs (5.5 %; $P = 0.005$), and consequent depletion within introns (51.6 %; $P = 1.3 \times 10^{-18}$). There was no evidence for significant sense/anti-sense differences, as has been previously reported (Bolisetty et al. 2012).

Many of the exonic overlaps were short (less than 10 bp), with most of the element within the neighbouring intron. It is possible that some of these definitions were therefore simply artefacts of poor element demarcation during identification. Despite this, the proximity to the exons is still of interest as it could affect, or even have created, splice sites in the containing gene. For those elements which completely contained exons or even complete, annotated genes, another observation was that most of these genes were categorised by Ensembl as "uncharacterised, known protein coding", which in most cases means that there was some RNAseq data which supported expression at this site, but little else is known. Additionally, most of these instances were single exon genes. It is therefore possible that these Ensembl 'genes' were only annotated due to expression (in one or more tissues for at least one submitted dataset) from the LTR retrotransposon itself, and are not 'true' genes.

It is, however, unlikely that all significant exon overlaps are a result of incorrect annotation. The two sets of constrained positions provide some evidence of biological significance. A total of 238 SIEs contain at least one constrained element and of these 82 have constrained elements from both lists, 82.9 % of which overlap exons. However, some overlaps are short and some of the constrained elements are themselves short, with a minimum accepted length of 10 bp. Exon overlaps may represent retrotransposon-derived exons or regulatory regions, or potential false positives which passed the feature tests. Most overlapped exons lacked any clear LTR retrotransposon homology through BLASTn, tBLASTx or domain pHMM analysis, even if the regions surrounding the overlapped feature have LTR retrotransposon homology. Whilst this may simply reflect selection and divergence of these retrotransposon derived sequences over time, it is worth noting that *Ovex1*, a co-opted gammaretrovirus in the chicken genome (discussed below in section 3.7.1), is found throughout the avian lineage and still retains clear LTR retrotransposon homology. The 5' UTR overlaps potentially represent more interest, even if the overlaps are short, as the proximity to the start of the gene might suggest impact on gene expression or regulation. However, only a fifth of these UTR overlaps are LTRs (which, when intact, have their own promoter activity).

Only 29 LTRs from the 1,073 SIEs contain constrained elements, with eight SIEs exhibiting constrained elements in both LTRs. These constrained LTRs largely overlap annotated exons rather than being potential standalone promoters under selection.

Overall, most constrained element overlaps with LTR retrotransposons appear to be with their internal regions. However, it is important to recognise that regions are only classed as being under constraint if they are conserved between multiple species groups, so any insertion since the divergence from the turkey would not appear constrained, but could still have function.

3.7 Analysis of structurally intact LTR retrotransposon expression

A total of 379 (35.3 %) SIEs have detectable RNA expression in the correct orientation, with robust transcript models, in at least one of the twenty-three tissues analysed. Expression was not biased towards younger elements or specific genera, but only 24.8 % of expressed SIEs are found within clusters. Expressed elements appear to follow a random distribution pattern relative to the Ensembl annotation, but those that overlap TUs are highly enriched in exons (47.1 %; $P = 4.1x10^{-24}$). Only 31 SIEs exhibit 'complete' expression, defined as a transcript extending at least the element length without the LTRs. Again, there was no apparent bias for genera or insertion age.

Two thirds of all complete transcripts can be found in at least one embryo stage, but there is no evidence for significantly elevated expression at the earliest stage. Incomplete transcription of LTR retrotransposons across all 379 SIEs may support a gradual decline of expression through the three analysed embryonic stages (186, 168 and 144 SIEs expressed respectively). Pancreas and ovary were the most represented adult tissues (each had 143 elements with at least fragmented expression, with 49.7 % overlap between the two tissues), but overall have the 6th and 15th highest transcript model coverage of the genome. This suggests element expression may be tissue specific, rather than simply related to the quantity of RNA expressed by a specific tissue across the genome.

Most identified LTR retrotransposon transcripts have high frequencies of closely interspersed stop codons in all three translated forward frames. However, there are examples of potentially full length protein products. One SIE on chromosome 8 (8: 10,499,515-10,505,342) has a long open reading frame (ORF) with high homology to the betaretrovirus *polymerase* polyprotein and was found to be expressed in embryo

stages HH14-15 and HH25-26. A second SIE (4: 85,449,603-85,458,772) has two long ORFs in the first reading frame with high homology to *gag* and *polymerase* respectively, and a third long ORF in the second reading frame with high homology for the *envelope* polyprotein, all from gammaretroviruses. Whilst the *gag* and *polymerase* putative proteins lack some domains, the *envelope* ORF encodes Ovex1 (GenBank: NP_001159385.1; Figure 3.7), a previously described 873 amino acid protein of known gammaretroviral origin (Carré-Eusèbe et al. 2009).

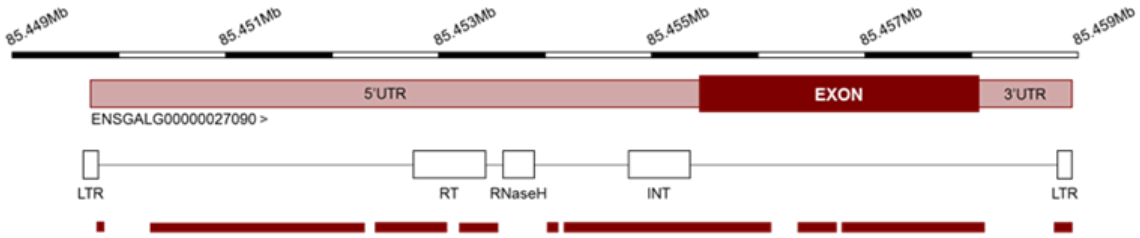


Figure 3.7 *Ovex1* schematic showing the long *gag-pol* 5'UTR and *envelope*-derived exon, promoted by the 5' LTR. The two LTR locations are shown, as are the recognisable reverse transcriptase (RT), RNaseH and integrase (INT) features identified by pHMM analysis. The lower line of red bars show the regions under constraint from the Ensembl alignment of seven sauropsid genomes. The location information shows the position of *Ovex1* on chromosome 4.

No putative transcripts were identified from alpharetroviral elements, although the many intact alpharetroviral LTRs could provide the basis for novel or alternative promoter activity, as could those from other retroviral genera. Both the intact *polymerase* putative proteins identified above have recognisable reverse transcriptase and RNaseH domains, which suggests they retain the ability to transpose other repeats, including non-autonomous elements.

3.7.1 Characterisation of Ovex1, the co-opted endogenous gammaretrovirus

In their original characterisation, Carré-Eusèbe and colleagues (2009) determined that chicken *Ovex1* RNA was limited to the gonads, but the RNAseq analysis described in the current study supports ubiquitous expression, with full-length transcript models

generated for ten adult tissues, including the ovary, and stage HH4-5 in the embryo data. Furthermore, the other RNAseq datasets had expression across the region, but at a level below the threshold required for transcript model construction in Cufflinks. Whilst expression in the ovary was the highest in the analysis (over 1000 times more read support than in spleen, the least supported intact transcript model), *Ovex1* expression was not solely limited to the gonad and may have a more general function.

InterPro analysis of the Ovex1 protein identified one transmembrane (TM) domain 22-47 residues away from the protein carboxyl-terminus, as well as several protein-protein interaction sites (Figure 3.8). In comparison, the envelope proteins of reference beta-, delta-, epsilon- and, most relevantly, gammaretroviruses exhibit two TM domains near the carboxyl-terminus: approximately 30-55 and 205-230 residues from the end. The Phobius predictions for Ovex1 and the reference envelope proteins identified the region from the N-terminus to the first TM domain as non-cytoplasmic. This is consistent with all envelope protein annotations, as the N-terminal two thirds of the protein forms the surface domain, and the carboxyl-end forms the TM domain. In exogenous retroviruses translated envelope protein is spliced into its two constituent domains which then form a heterodimer, and a subsequent homotrimer of these heterodimers forms the retroviral envelope subunits. Other examples of host co-opted gammaretroviruses, notably the mammalian *syncytin* placental genes and murine antiviral receptor genes *Fv4* and *Rcmf*, continue to form this homotrimer for their host function (Gong et al. 2005; Lavialle et al. 2013). The presence of putative protein-protein interaction sites suggests that *Ovex1* products could also form these functional homotrimer complexes.

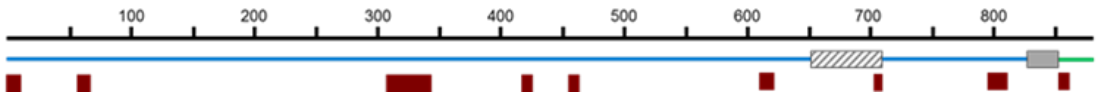


Figure 3.8 Domain analysis of the 873 amino acid Ovex1 protein. The blue line represents the Phobius-predicted non-cytoplasmic domain, and the short green line at the carboxyl-terminus is the cytoplasmic domain. The grey box near the carboxyl-terminus is the InterPro-predicted transmembrane (TM) domain, and the larger, upstream, diagonal-filled box represents the predicted location of the TM domain missing in Ovex1 that is present in gammaretroviral envelope proteins. The dark red boxes show the predicted protein-protein interaction sites.

Identification of Ovex1 homologues

BLASTp searches identified Ovex1 protein homologues across the sauropsid lineage, including in turkey (GenBank: XP_010708895.1; $1x10^{-200}$; 95 % identity), duck (*Anas platyrhynchos*; GenBank: XP_012958629.1; $1x10^{-200}$; 86 % identity), eighteen Neoaves, and four reptiles: *Anolis carolinensis*, *Pelodiscus sinensis*, *Python bivittatus* and *Thmanophis sirtalis*. The four reptile Ovex1 homologues were annotated as PPARD (peroxisome proliferator activated receptor delta) proteins in GenBank, but the alignment between these and the PPARD sequences from chicken (GenBank: NP_990591.1), mouse (*Mus musculus*; GenBank: NP_035275.1), and human (*Homo sapiens*; GenBank: AAH07578.1) was very poor (less than 8 % identity when the three PPARD proteins share 82.9 % identity), suggesting an incorrect annotation for these four reptile sequences. The avian Ovex1-homologue sequences were generally well conserved at the TM carboxyl-end, and three species (*Anser cygnoides*, *Serinus canaria* and *Zonotrichia albicollis*) had duplicated protein sequences which also retained high carboxyl-end identity. Analysis of the Ovex1-homologue protein alignment identified 207 sites (18.8 %) under purifying selection, including sites throughout the TM domain (supporting the carboxyl-end conservation) and regions correlating to the predicted protein-protein interaction sites.

The phylogeny built from the alignment of retroviral envelope proteins and sauropsid Ovex1 protein homologues (Figure 3.9) was generally poorly supported, likely due to the extensive variability of the envelope sequences between retroviral genera. Whilst the Ovex1 homologues largely follow the known avian phylogeny, they do not nest with the gammaretroviral sequences as expected. However, the Ovex1 homologues do fall within the retroviral phylogeny (with the spumavirus group as a correct outgroup) and the combination of limited basal node support and excess of Ovex1 sequences (compared to other retroviral groups) may have affected the tree construction. The phylogeny alone would not be enough evidence to put the gammaretroviral classification in doubt, as all the *polymerase* domain classifications support gammaretroviral origin. Interestingly, the species-specific duplicated sequences group together as a sister clade to the reptile Ovex1 homologues, rather than falling within the main avian group, perhaps supporting an ancient duplication event within the avian lineage.

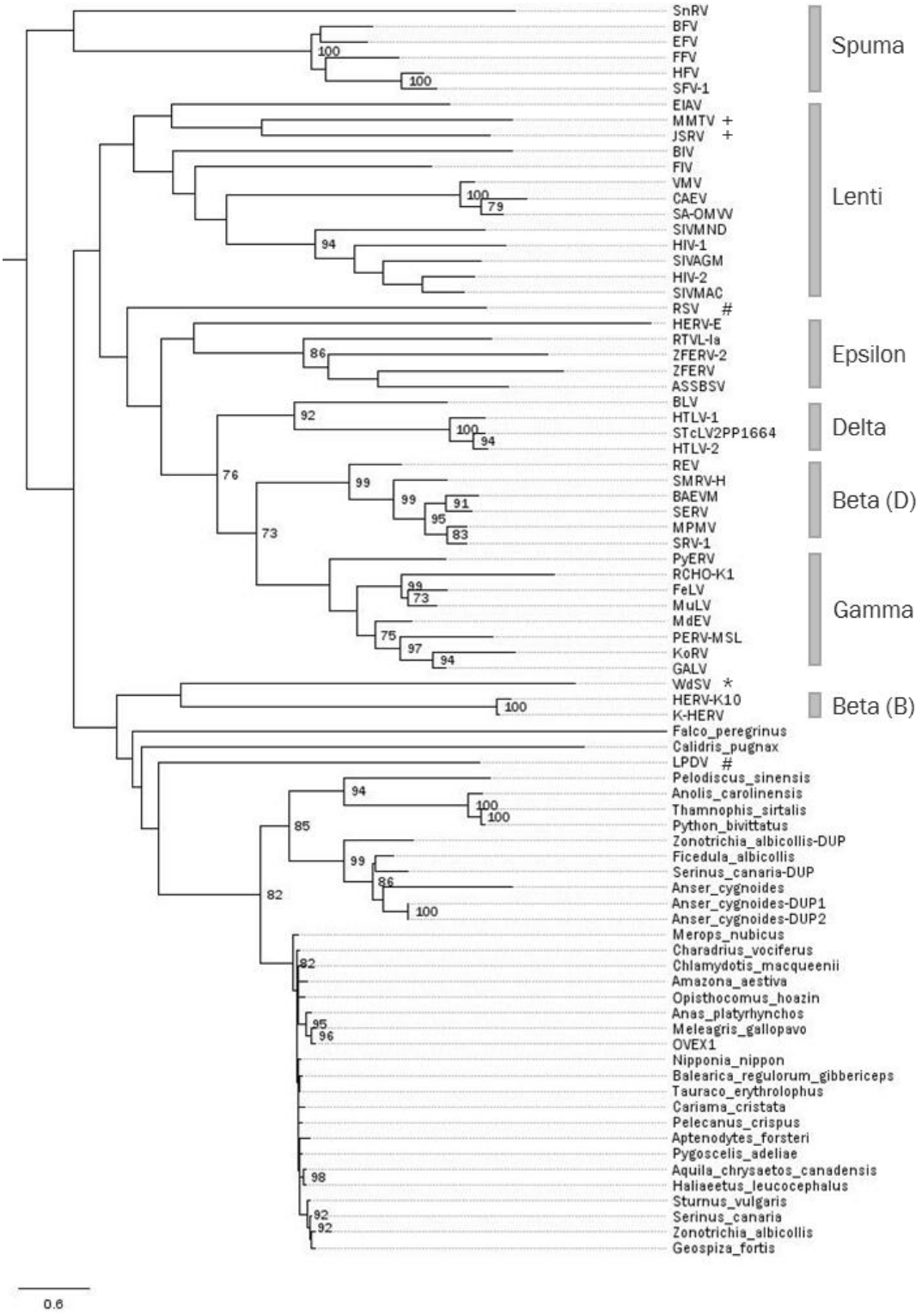


Figure 3.9 Phylogeny of retroviral envelope proteins, Ovex1 and the sauropsid Ovex1 homologues. Bootstrap values greater than 70 are shown on their respective nodes. The grey bars identify the retroviral clades, but exceptions are denoted by the symbols: '+' Beta (B), '#' Alpha, '*' Epsilon. The betaretrovirus clade is divided by its two *envelope* subclasses, B and D. The phylogeny generated

here largely matches the known envelope phylogeny published on GyDB, although it would be expected that deltaretroviruses would be basal to the epsilonretroviruses. In addition, the alpharetroviruses would be expected to cluster together as a sister group to the deltaretroviruses.

3.8 Discussion

This updated annotation of the LTR retrotransposon content of the chicken genome has identified 31.5 Mb of sequence, including 1,073 structurally intact elements. Improvements between the Galgal3 and Galgal4 genome assemblies accounted for almost 2.5 Mb of extra annotated sequence, but the expanded homology protocol and use of four structure-based identification programs implemented through LocaTR enabled this much more complete annotation, including an additional 587 SIEs not identified in the previous analysis. This work brings the total annotated LTR retrotransposon content to 3.01 % of the genome, matching the observed levels in Neoaves and even mammals, when scaled for genome size. However, this increase in annotated content is largely due to annotation effort, so this work alone is insufficient to refute the previously proposed deficit of LTR retrotransposons in the Galliformes. The LocaTR analysis of multiple genomes of the avian lineage presented in Chapter 4 more comprehensively addresses this evolutionary question.

An exhaustive list?

The question remains as to whether even this much-expanded annotation is a complete list of all LTR retrotransposons within the chicken genome. Put simply, it is highly unlikely that the work presented here is an exhaustive list. The chicken genome assembly is not perfect, with thousands of unplaced contigs, missing microchromosomes and poorly sequenced regions such as chromosomes 16 and W, both of which are known to have high repeat content. In addition, the new Galgal5 assembly, released in March 2016, is likely to have a similar effect on annotated content as was observed between the Galgal3 and Galgal4 assemblies (explored in Chapter 4).

Another consideration is the methodology itself. The LocaTR pipeline does not require redundancy between approaches, but a degree of conservatism is still required in element validation to reduce the risk of false positives. Furthermore, the identification programs themselves are inherently biased towards known, well described LTR retrotransposon reference sequences. This is most obvious with the homology-based approaches, but the structure-based programs are based on training sequences and 'typical' size constraints. However, it is also unlikely that most highly degraded or divergent sequences which are missed by these approaches are of great biological relevance to the host, unless such divergence was due to host co-option.

3.8.1 The development of the LocaTR identification pipeline

The initial installation, preliminary work and parameter optimisation completed for each of the disparate identification programs used in this study was time consuming and required the installation of multiple dependencies on the Roslin Linux servers. This was further complicated by generally limited documentation and a lack of ongoing author curation for several of the programs. As a result, every effort was made to ensure that the LocaTR identification pipeline was clear, well documented, and retained the ability for users to set program-specific parameters. Individual installations can still be difficult depending on the user's computer architecture and privileges, as well as the availability and continued compatibility of accessory packages. In addition, LocaTR requires access to multiple operating systems with LTR_STRUC limited to Windows, and RetroTector to Windows or macOS. However, LocaTR is modular, allowing users to 'pick and choose' individual identification programs and to add additional software if desired.

The programs used for the LocaTR analysis have been good choices, following their selection after an extensive review of the available software. There was generally low overlap between programs, and other tested programs had result subsets completely contained within other results, such as all elements identified by LTR_par and LTR_FINDER also identified by LTR Harvest. Other programs are available, but the current set of four structure-based programs offers good variety. The reference sequences used for the expanded homology-based protocol also give good phylogenetic

coverage, although analysis of plant genomes would benefit from the addition of plant-specific LTR retrotransposon groups to the sequence list.

The high false positive rate of the structure-based programs remains a concern, as does the potential for non-LTR retrotransposon sequences to be detected by the homology-based programs due to high identity between domains such as *reverse transcriptase*. However, the results from the feature tests give good support for putative LTR retrotransposons, and the high corroboration between the feature tests and LTR Digest was reassuring. There is a chance that some identified SIEs may be false positives that passed the feature tests, but every effort has been made to reduce this undesirable result.

Altogether, the pipeline development has been a success, not just for the results obtained with the analysis of the chicken, but also for its wider application.

3.8.2 LTR retrotransposon distribution in the chicken genome

Detailed analysis of the chicken LTR retrotransposons has shown that element distribution is non-random, with a significant depletion of elements within coding regions and an enrichment of elements in gene sparse areas, including significantly elevated LTR retrotransposon density on macrochromosomes. Genomic distribution is, therefore, dependent on insertion neutrality, as non-detrimental insertions are retained producing skewed distributions away from coding regions.

Over 40 % of structurally intact elements were within clusters which were unrelated by insertion age or retroviral genera, suggesting recurrent insertions into the same genomic locations, generally in gene-sparse regions. Bolisetty and colleagues (2012) found similar rates of clustering and proposed functional roles for these clusters as cytoskeletal binding regions during mitosis, or as hotspots for recombination. This analysis, however, has found no evidence of constraint within cluster locations, and that most clusters are within genomic regions of low or negligible recombination rates. It is therefore likely that clusters form as the regions where insertions elicit limited or negligible deleterious phenotypic effects increase in size, eventually self-perpetuating as new insertions grow the clusters further.

This concept may have a wider impact on genome size, as high repeat content promotes repeat content expansion. Previous work has already linked total repeat content to the greater genomic stability of avian genomes compared with mammalian genomes (Griffin et al. 2008; Ellegren 2010), and may also explain why avian genomes have a deficit of all repeat classes compared to mammals unless scaled by genome size (explored further in Chapter 4).

3.8.3 LTR retrotransposon activity within the chicken genome

The continued retrotransposition of chicken alpharetroviral ERVs (both ALVEs and EAVs) is well documented (Wragg et al. 2015; Rutherford et al. 2016), but the high LTR identity observed with some gammaretroviral and betaretroviral ERVs supports recent integration of these retroviral genera, with the potential for further retrotransposition. Only a minority of these younger insertions retain intact internal domains, but the expression of other retroviral-like proteins in cells has the potential to facilitate retrotransposition of these degraded sequences.

Transcribed LTR retrotransposons in the chicken were rare, and even these examples were largely fragmented or code for non-functional proteins. However, whilst a relatively diverse set of tissue types were used to assess expression they were from a limited age range and a single breed (crucially not the same bird used for the reference genome). Tissue-, temporal- and breed-specific expression is likely, and expression levels may be low enough to preclude transcript model construction. Despite this, transcripts were identified from *gag*, *polymerase* and *envelope* domains, with apparent tissue specificity. Of particular interest is the identified expression of *polymerase* in two cases, as both predicted products retain reverse transcriptase and RNaseH integrity, which suggests translated products could transpose other repeats, including non-autonomous elements. In addition, the integration of reverse transcribed host mRNA can form retrogenes; elements with huge evolutionary potential for the host through the introduction of intact domains to existing genes, or full gene duplication (Kaessmann et al. 2009).

The expression of intact, likely functional, domains also presents more opportunity for recombination with exogenous retroviruses. Identification here of both gamma- and

betaretroviral domains, rather than just the alpharetroviral sequences implicit in the formation of ALV-J (Fadly 2000; Borisenko 2003), also extends the range of potential recombinant retroviruses, especially as cross-genera recombination has been observed (Liu et al. 2011).

In addition, this work has enabled the further characterisation of *Ovex1*, including its much more ubiquitous pattern of expression compared to its initial characterisation (Carré-Eusèbe et al. 2009). The presence of putative protein-protein interaction domains suggests the Ovex1 protein may be able to form the functional homotrimer complexes observed in retroviral envelope proteins as well as other co-opted gammaretroviruses. Cell-cell cohesion, similar to that effected by the *syncytin* protein in the mammalian placenta, seems unlikely due to the ubiquity of expression. It is more likely that the protein may have a role in innate antiviral immunity through receptor interference, as has been widely documented in mouse and cat (*Felis catus*) with gammaretroviral envelope, in sheep (*Ovis aries*) with betaretroviral envelope, and in chicken with ALVE envelope (Lavialle et al. 2013; Smith et al. 1990a; Varela et al. 2009; Ito et al. 2013; Kozak 2014), by physically blocking retroviral entry receptors as a competitive inhibitor. Identification of duplicated avian homologues in three avian species also supports receptor interference, with duplicates potentially selected for defence to related, but distinct exogenous gammaretroviruses. This would be the first example of gammaretroviral *envelope*-mediated receptor interference in chicken.

3.9 Concluding remarks

This detailed annotation of LTR retrotransposons in the chicken reference genome provides a platform for further analysis. Within the scope of this project, it has enabled an evolutionary study of LTR retrotransposon content across the avian lineage (Chapter 4) and the evaluation of ALVE diversity across chicken subpopulations and commercial lines (Chapters 6 and 7). Beyond its application for chicken research, the development of the LocaTR identification pipeline provides a useful resource to other researchers for the identification of these retroelements in any assembled genome.

Chapter 4: Patterns in LTR retrotransposon content across the Avian lineage

4.1 Introduction

The work presented in the previous chapter showed that the chicken (super order Galloanserae) has a similar LTR retrotransposon content to that previously described in Neoaves. However, this could now be due to research effort rather than a true, biologically accurate representation of the difference in content between these birds. It should also be remembered that the initial proposal of a deficit of LTR retrotransposons within the Galliformes was based on comparisons with a limited selection of Neoaves genomes available at the time (Suh et al. 2011; Bolisetty et al. 2012). It is therefore possible that these conclusions were drawn based on lineage-specific LTR retrotransposon expansions.

During the second year of my PhD project a wide phylogenetic range of draft avian genomes were released, enabling large-scale comparative genomic studies for the first time (Jarvis 2014). This took the number of publicly available avian genomes to forty-eight, a number which has continued to increase over the last three years. Most of these were Neoaves genomes (reflecting avian extant diversity), with the chicken, turkey and duck of the Galloanserae, and the African ostrich (*Struthio camelus australis*) and white-throated tinamou (*Tinamus guttatus*) of the Paleognathae. Endogenous viral elements were studied almost immediately (Cui et al. 2014), and the greatest numbers of intact ERVs were identified in three oscine passerines: 725 in zebra finch, 785 in the medium ground finch (*Geospiza fortis*), and 1,032 in the American crow (*Corvus brachyrhynchos*). However, at least in terms of intact elements, there was no clear split in content between Galliformes and Neoaves. The average number of identified intact ERVs was 350: the chicken had 573 and turkey had 303. Furthermore, the work presented in chapter 3 identified 1,073 intact ERVs in chicken.

This variation in ERV content, especially as there is little evidence of shared avian ERVs with reptilian species, suggests lineage-specific expansion and contraction of many of these elements. Multiple studies published during and since my own analysis of LTR retrotransposons across the avian lineage (below) have supported the importance of lineage-specific repeat content, particularly in the Neoaves. This stems from the rapid

diversification of this group which has produced a hard polytomy at the base of the lineage, resulting in large-scale, incomplete lineage sorting of thousands of analysed genetic markers, including transposable elements (TEs) (Suh et al. 2015; Suh 2016).

The general evolutionary consensus for the consistently low TE content of avian genomes (5-10 %, except approximately 22 % in the downy woodpecker, *Dryobates pubescens*) has been that these elements are generally inactive (Shedlock 2006; Shedlock et al. 2007; Janes et al. 2010). However, recent work has shown that many TEs are highly active in birds, but that there is a counter balancing high rate of large genomic deletions which maintain the small avian genome size (Kapusta et al. 2017). It is also interesting to note that whilst the net size change in most lineages remains close to zero, neighbouring taxa can exhibit very different, yet balanced, rates of gain and loss. This may, again, facilitate highly lineage-specific repeat content, although some of this effect may be mitigated as new TE expansions or TEs with high copy number are more likely to be removed from the genome by processes such as non-allelic recombination (Kapusta & Suh 2017). All this potential for lineage-specificity likely means a solely homology-based approach to LTR retrotransposon identification is inadequate for complete annotation.

Another consideration for observed repeat lineage specificity is the highly variable quality of the avian genome assemblies. It was noted in chapter 3 that, even without the use of LocaTR, the analysis of the chicken Galgal4 assembly enabled identification of 2.5 Mb more LTR retrotransposon derived sequence than was identified in the same manner with the previous Galgal3 assembly. The original forty-eight avian genomes used short read sequencing data, but some of the more recent assemblies, including the Japanese quail (*Coturnix japonica*), hooded crow (*Corvus cornix*), great tit (*Parus major*) and the new Galgal5 chicken genome assembly, have made use of long read sequencing technology. This generates higher contiguity and is more able to sequence through repetitive regions. The Galgal5 assembly is 183 Mb longer than Galgal4, has ten times greater contig length, reduced contig number, a 95.7 % reduction in the number of scaffold spanned gaps, and three more assembled microchromsomes (30, 31 and 33) (Warren et al. 2017), summarised in Table 2.3. It is therefore likely that these long read based assemblies will produce more representative LTR retrotransposon content, and it is important that the impact of assembly quality is quantified.

4.2 Research Aims

This chapter covers two major research aims. Firstly, the identification of LTR retrotransposons in the new chicken genome assembly (Galgal5) using LocaTR, to determine whether recent improvements in assembly length and contiguity have effected total LTR retrotransposon content. Secondly, the identification of LTR retrotransposons across the avian lineage using LocaTR, to identify whether there is truly a deficit of these elements in Galliformes compared to the Neoaves, and to assess the extent of lineage-specific expansions and the impact of genome quality on repeat annotation.

4.3 Statement of publication

A summary of the LocaTR analysis of the Galgal5 chicken assembly was published within the paper describing the new genome build: Warren W, Hillier L, Tomlinson C *et al.* (2017), A New Chicken Genome Assembly Provides Insight into Avian Genome Structure, *G3*, **7**: g3.116.935923. This included a short section on the physical distribution of these elements, the comparison between Galgal4 and Galgal5, and relative improvements to total repeat content (Appendix 3: Paper3). At the point of publication, the RetroTector analysis had not been completed, so the total numbers presented here are slightly larger than in the Galgal5 paper.

As stated above (section 3.3), the initial, homology-based identification of LTR retrotransposons across a limited range of the sauropsid lineage was published as part of the LocaTR and Galgal4 annotation paper: Mason *et al.* 2016 (Appendix 3: Paper2).

4.4 Identification of the LTR retrotransposon content of the new Galgal5 chicken genome assembly

A total of 35.2 Mb of the Galgal5 assembly was identified as LTR retrotransposon-derived sequence. This accounts for 2.86 % of the genome. Whilst this is a smaller percentage than identified in Galgal4, this is due to the 17.51 % increase in total assembly

length. The actual annotated LTR retrotransposon sequence has increased by 3.7 Mb (11.91 % increase over the Galgal4 annotation; Table 4.1).

Table 4.1 Comparative summary of the LocaTR-annotated LTR retrotransposon content of the chicken genome from the Galgal4 and Galgal5 assemblies.

Assembly features	Galgal4	Galgal5
Total length (bp)	1,046,932,099	1,230,258,557
LTR content (bp)	31,454,008	35,200,607
LTR content (%)	3.01	2.86
Number of SIEs	1,073	1,295

This total content includes some 63,421 distinct sites of which 1,295 were structurally intact elements (SIEs). A total of 1,073 SIEs were identified in Galgal4 and the increase appears to be due to improvements on the W chromosome (SIEs increased from 56 in Galgal4 to 244 in Galgal5) and the sequencing of previously unrepresented regions. SIE content on the assembled macrochromosomes has generally reduced due to collapsing of repetitive regions and proper incorporation of unplaced contigs.

The effect of the improvements made in this assembly to increase both contiguity and total sequence were evident from a the RepeatMasker '-species vertebrates' analysis alone (Table 4.2). However, the use of LocaTR enabled identification of an additional 8.5 Mb of LTR retrotransposon-derived sequence (32.0 % extra). This observed increase using LocaTR was not solely due to identification of SIEs, as the LocaTR homology protocol identified 31.3 Mb of sequence, 4.6 Mb more than using RepeatMasker alone (Table 4.3). The secondary BLAST protocol (following the SIE identification) had a much more limited additive effect than in the Galgal4 analysis. This was due to most of these sites being identified by the homology protocol, following the inclusion of the Galgal4 results.

Table 4.2 Improvements in LTR retrotransposon content annotation across three chicken genome assemblies. The Galgal4 to Galgal5 sequence increase is almost four times more than the increase between Galgal3 and Galgal4.

Assembly features	Galgal3	Galgal4	Galgal5
Assembly length (bp)	1,098,770,941	1,046,932,099	1,230,258,557
Scaffold N50 (bp)	11,063,745	12,877,381	6,379,610
LTR content (bp)	14,870,595	17,369,358	26,660,513
LTR content (%)	1.35	1.67	2.17

Table 4.3 The contribution of the distinct LocaTR protocols to the overall total for both the Galgal4 and Galgal5 assemblies. Total identified content is shown for the homology and structure-based protocols, and just the additive effect is shown for the secondary BLAST. The use of the Galgal4 annotation as part of the Galgal5 analysis has reduced the additive effect of the secondary BLAST, limiting additional sequence to that homologous to the newly identified SIEs.

LocaTR protocol contributions	Galgal4	Galgal5
Homology protocol (bp)	20,322,178	31,290,565
Structure-based programs (bp)	9,114,835	9,683,365
Additive secondary BLAST (bp)	7,064,272	484,108
Total (bp)	31,454,008	35,200,607

Each of the four structure-based identification programs again exhibited high false positive rates (Table 4.4), but this time there was greater corroboration between the programs. In the Galgal4 analysis, 72.79 % (782/1073) of all SIEs were identified by a single program. For Galgal5, 532 of the 1,295 SIEs (41.08 %) were unique to a single program, and 609 (47.02 %) were identified by two programs. There was much greater overlap between the results of LTR Harvest, MGEScan_LTR and RetroTector for this analysis than was observed in Galgal4. Again, there were no detectable biases between the four programs, but only 0.31 % of SIEs were identified by all four.

Table 4.4 Structurally intact elements identified by the four structure-based identification programs. False positive rates are comparable to the Galgal4 analysis (Table 3.8), but the unique SIE percentages are lower for all programs.

Intact element features	LTR_STRUC	LTR Harvest	MGEScan_LTR	RetroTector
Initial SIEs identified	1,661	29,251	1,129	1,054
SIEs with feature support	379	800	687	539
False positive rate (%)	77.18	97.27	39.15	48.86
SIEs unique to program (%)	6.86	27.38	27.51	18.18

4.4.1 LTR retrotransposon density

In the Galgal5 assembly, like Galgal4, chromosome length had a strong negative correlation with both gene density ($r = -0.93$; $P < 0.001$) and recombination rate ($r = -0.91$; $P < 0.001$). In the Galgal4 assembly, LTR retrotransposon content was found to have the opposite correlation: a strong positive correlation with chromosome length and negative correlations with both gene density and recombination rate (Table 4.5). In Galgal5 these correlations are weaker, but remain significant.

Table 4.5 Correlations between LTR retrotransposon density and chromosome length, recombination rate or gene density for the chicken Galgal4 and Galgal5 assemblies. Each correlation is given as the Pearson 'r' and the associated P values are given below in brackets. Non-significant P values are underlined. Correlations were completed for all assembled Galgal4 chromosomes (Gg4), all assembled Galgal5 chromosomes (Gg5 All), the Galgal5 macrochromosomes (Gg5 Macro; 1-10) and the Galgal5 microchromosomes (Gg5 Micro; 11-15,17-26,28). Gg4 analysis excluded chromosomes 16, 25, 27, 32, W and Z (section 2.2.2), and Gg5 analysis excluded chromosomes 16, 27, 30-33, W and Z (section 2.3.2).

LTR density correlation	Gg4	Gg5 All	Gg5 Macro	Gg5 Micro
vs. chromosome length	0.91 (P < 0.001)	0.69 (P < 0.001)	0.97 (P < 0.001)	0.17 (P = <u>0.572</u>)
vs. recombination rate	- 0.81 (P < 0.001)	- 0.50 (P = 0.014)	- 0.88 (P = 0.001)	- 0.22 (P = <u>0.447</u>)
vs. gene density	- 0.72 (P < 0.001)	- 0.58 (P = 0.003)	- 0.88 (P = 0.001)	- 0.18 (P = <u>0.548</u>)

Reduction in the genome-wide correlation strength is due to the increase in annotated LTR retrotransposon content on the microchromosomes. The macrochromosomes and microchromosomes have very different LTR retrotransposon content compared to their chromosome length (Figure 4.1). Chromosomes 1 to 10 have correlations like those observed genome-wide in Galgal4, however, the microchromosomes (11-15, 17-26, 28) have no significant correlations between LTR retrotransposon content and chromosome length, recombination rate or gene density (Table 4.5).

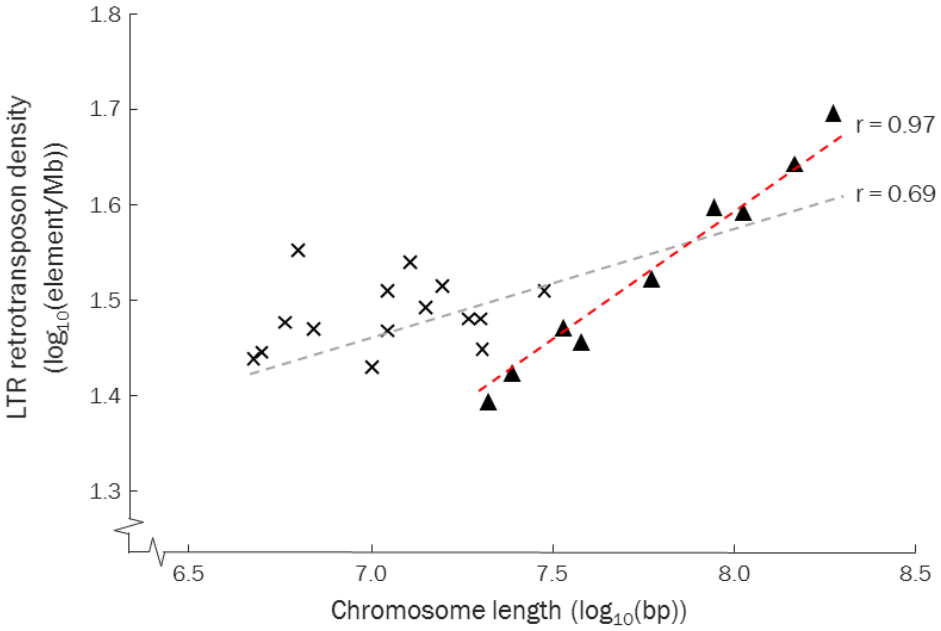


Figure 4.1 Correlation between chromosome length and LTR retrotransposon density, where both measures have been $\log_{10}$ transformed and all the outlier chromosomes (16,27,30-33,W,Z) have been removed. The grey dotted line (r = 0.69) shows the positive correlation for all data, and the red dotted line (r = 0.97) is the stronger positive correlation when just considering the macrochromosomes (chromosomes 1-10; triangles). The microchromosomes (crosses) alone have a non significant correlation.

The GLM fitted for the whole genome identified both chromosome length ($P < 0.001$) and recombination rate ($P = 0.027$) as significant variables. There was no significant interaction between the two variables even though recombination rate is scaled by chromosome length. Recombination rate was not significant in the Galgal4 GLM (section 3.6.1). The GLM fitted for just the macrochromosomes gave only chromosome

length as significant ($P < 0.001$), and the GLM for the microchromosomes had no significant variables.

Correlations between LTR retrotransposon content and chromosome length were all repeated with the addition of the unplaced, but localised, contigs for the assembled chromosomes. The lengths were included in the chromosome totals and elements identified on these contigs were added to the density calculation. The correlations were much the same, and the same split was observed between macrochromosomes and microchromosomes. These data are not shown here as the values for lengths and density were non-biologically representative approximations. However, it was important to check that the absent regions that were difficult to assemble, due to, for example, their repeat content, did not bias the correlations.

SIE clusters

A total of 521 SIEs (40.23 %) were identified within clusters which were unrelated by insertion age or genera (Appendix 2; AF05), proportionately almost identical to Galgal4 (40.26 %). All but two of the previously identified clusters were identified again. Both were clusters of five SIEs which have now reduced to four, removing them from the analysis. An additional cluster was identified on the Z chromosome, and neighbouring cluster pairs identified on both chromosome 1 and Z in Galgal4 were joined in this analysis. Despite these similarities it must be noted that several clusters on chromosomes 1, 2, 4 and Z reduced in size due to repeat collapsing and the incorporation of unplaced contigs which increased the region size. However, the numbers remain similar due to the expansion of the single W chromosome cluster from 56 SIEs in Galgal4 to 244 in this analysis.

4.4.2 The distribution of LTR retrotransposons relative to genomic features

In addition to the recent improvements of the chicken reference genome, the known annotation of coding features has also been updated using RNAseq data from multiple tissues and developmental stages, and the inclusion of PacBio IsoSeq data which

unambiguously identifies alternative transcripts (Gonzalez-Garay 2015). Consequently, much more of the genome has been annotated as coding, so now, under random integration, it would be expected that 58.32 % of insertions would fall within transcriptional units (TUs; the coding region and 5 kb up- and downstream). However, as with the Galgal4 analysis, there is a significant depletion of LTR retrotransposon derived sequence within coding regions (46.27 %; P = 0.017), and an enrichment of these in unplaced contigs (20.52 %; P = 0.020), when compared to a model of random integrations (Figure 4.2). These differences are more marked in the SIE dataset (P = 1.24×10^{-5} and P = 4.45×10^{-10} respectively).

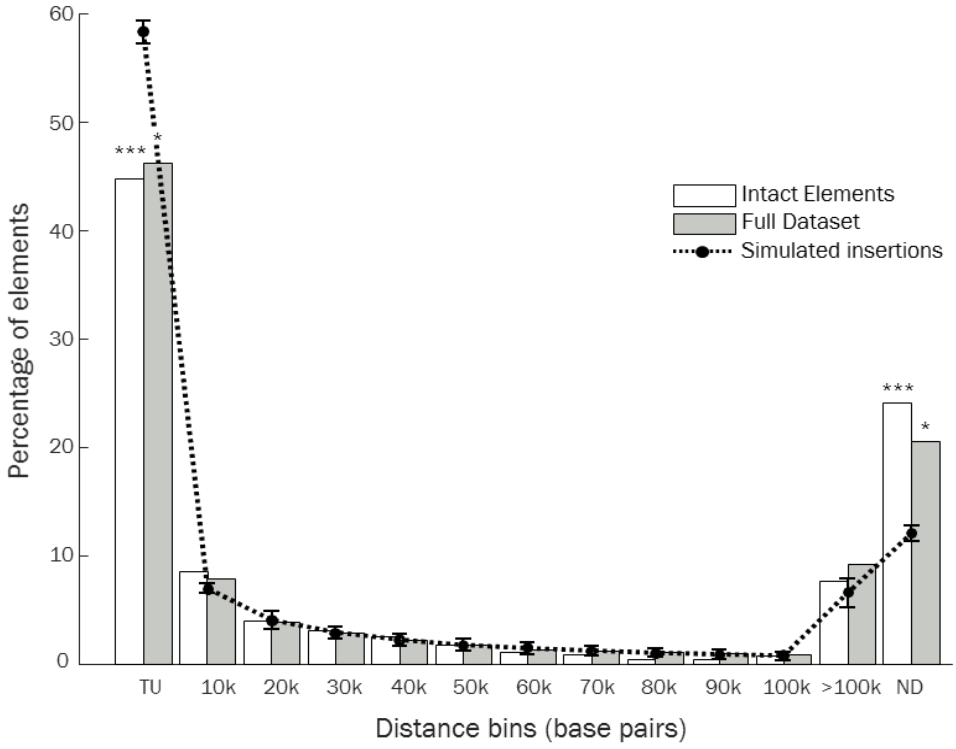


Figure 4.2 LTR retrotransposon distribution relative to the Ensembl genome annotations (v86). Shortest distance was measured from each element to the nearest annotated feature (irrespective of strand) and grouped into 10 kb bins, where the bin value represents the upper bin limit. The dotted line represents 100,000 randomly generated distributions for each dataset, with standard deviation error bars. One line is shown because the randomly generated full and intact models gave the same results, and the standard deviation was equal when rounded to two decimal places. Significant differences between proportions in each bin are highlighted with asterisks, where * = P < 0.05 and *** = P < 0.0001. TU = Transcriptional Unit (incl. exons, introns, UTRs and 5 kb flanks up and downstream). ND = Non-Defined (contigs without an Ensembl annotation).

In contrast to the Galgal4 analysis, there was no enrichment in either the full or SIE datasets in regions greater than 100 kb away from coding regions. However, when the macrochromosomes (1-10, Z) were considered alone, there was both significant depletion of elements within TUs (43.78 %; P = 0.027) and enrichment in regions greater than 100 kb away from coding regions (24.69 %; P = 0.032). On the microchromosomes, LTR retrotransposon distribution follows that of the random integration model.

Observed overlaps with long non-coding RNA (lncRNA) genes

Transposable elements have been shown to derive lncRNA genes in mammalian genomes (Kapusta et al. 2013), and work recently published from our group by Kuo and colleagues (2017) identified over 20,000 novel lncRNA genes in the chicken. Of the 782 structurally intact LTR retrotransposons on assembled chromosomes, 72 (9.21 %) were found to overlap with lncRNA genes by at least 50 bp. Of these, 32 (44.44 %) were where the lncRNA completely contained the SIE, and 21 (29.17 %) were SIEs which completely contained the lncRNA. A further 15 (20.83 %) had overlaps shorter than half the length of the smaller element.

Interestingly, 54.17 % of lncRNA (39/72) overlaps were with SIEs within clusters. One of these SIEs (4: 19,503,044-19,513,269) contained two lncRNAs in the same orientation: 4: 19,509,404-19,510,686 and 4: 19,511,604-19,513,631. Overall, 54.17 % of overlapping features were in the same orientation.

The output file of this intersectBed analysis is in Appendix 2: AF06.

4.5 LTR retrotransposon content across the avian lineage

4.5.1 Homology-based annotation using the LTR retrotransposon sequences identified in the analysis of the Galgal4 assembly

LTR retrotransposons were identified in each of the twenty-one analysed species with the generic RepeatMasker analysis for vertebrate repeats, but the addition of the Galgal4 custom library increased the annotated LTR retrotransposon content in all species. The additive influence of the chicken-derived sequences was substantial across the lineage

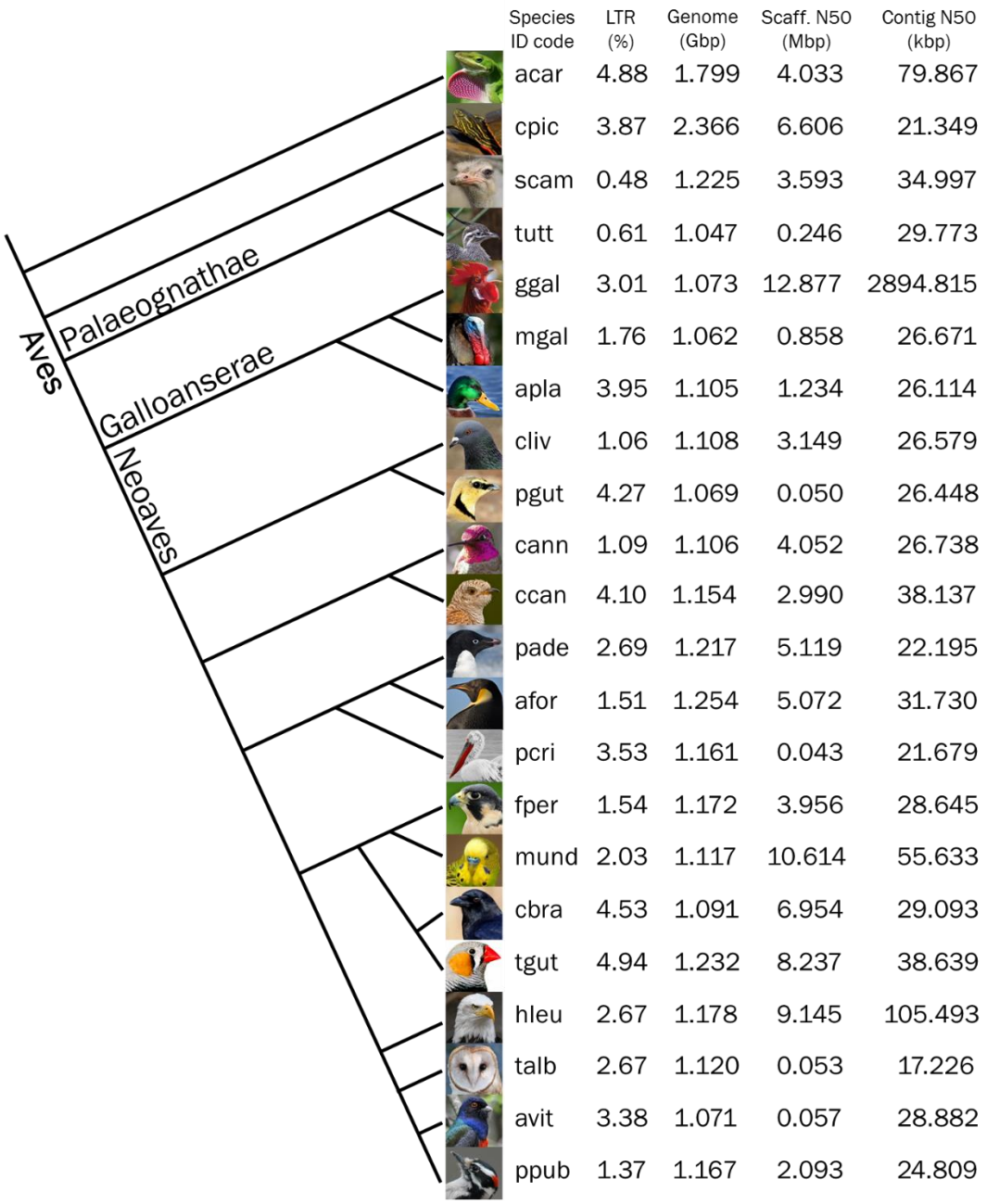
(Table 4.6), except in the phylogenetically distant Carolina anole (*Anolis carolinensis*). Eleven of the analysed species had an increase to their annotated repeat content greater than was seen in the chicken.

When these data were annotated against the known phylogeny (Figure 4.3) it was clear that LTR retrotransposon content is highly heterogeneous, and there was certainly no clear split between the Galloanserae and Neoaves. The highest LTR retrotransposon content was identified in the two oscine passerines analysed (American crow and zebra finch; cbra and tgua), matching the work of Cui and colleagues (2014). However, it is interesting to note that these two species differ in the content derived from each RepeatMasker analysis (Table 4.6), perhaps supporting lineage-specific repeat variation within this incredibly diverse phylogenetic group.

Sister taxa often have very different LTR retrotransposon content, such as the rock dove (*Columba livia*; cliv) and yellow throated sandgrouse (*Pterocles gutturalis*; pgut), and Anna's hummingbird (*Calypte anna*; cann) and the common cuckoo (*Cuculus canorus*; ccan). Interestingly, in both pairings it is the species with the 'poorer' genome quality (in terms of N50 length) which has the greater content. However, there were no significant correlations between LTR retrotransposon content and genome size, scaffold N50 or contig N50 (Figure 4.4). A striking example of this was that the same LTR retrotransposon content (2.67 %) was observed in the bald eagle (*Haliaeetus leucocephalus*; hleu) and barn owl (*Tyto alba*; talb) despite the bald eagle genome having a 180-fold higher scaffold N50 length. This likely represents true lineage-specific differences between the species.

Table 4.6 The identified LTR retrotransposon content in each analysed species showing the contribution of the two RepeatMasker analyses. Species are shown in alphabetical order with unique four letter codes. The LTR retrotransposon content is shown as percentages of the genome in all cases, with the first column the content identified from the '-species vertebrates' RepeatMasker analysis (RM-v), the second column from the custom Galgal4 library (RM-G4), and the third column shows the additive total. Annotated content overlapped between analyses in all cases, with the additive Galgal4 analysis effect in the final column.

Species name	Code	RM-v (%)	RM-G4 (%)	Total (%)	Effect (%)
Anas platyrhychos	apla	0.99	3.78	3.95	74.94
Anolis carolinensis	acar	4.52	0.43	4.88	7.38
Apaloderma vittatum	avit	0.86	3.10	3.38	74.56
Aptenodytes forsteri	afor	1.10	1.38	1.51	27.15
Calypte anna	cann	0.66	0.77	1.09	39.45
Chrysemys picta bellii	cpic	1.30	2.58	3.87	66.41
Columba livia	cliv	0.61	0.89	1.06	42.45
Corvus brachyrhynchos	cbra	1.97	3.48	4.53	56.51
Cuculus canorus	ccan	0.52	3.94	4.10	87.32
Falco peregrinus	fper	1.15	1.35	1.54	25.32
Gallus gallus	ggal	1.67	3.01	3.01	44.52
Haliaeetus leucocephalus	hleu	1.83	2.46	2.67	31.46
Meleagris gallopavo	mgal	1.06	1.62	1.76	39.77
Melopsittacus undulates	mund	1.52	1.59	2.03	25.12
Pelecanus crispus	pcri	1.73	3.38	3.53	50.99
Picoides pubescens	ppub	0.59	1.13	1.37	56.93
Pterocles gutturalis	pgut	1.11	3.97	4.27	74.00
Pygoscelis adeliae	pade	1.28	2.56	2.69	52.42
Struthio camelus australis	scam	0.17	0.36	0.48	64.58
Taniopygia guttata	tgua	3.54	2.07	4.94	28.34
Tinamus guttatus	tgus	0.21	0.49	0.61	65.57
Tyto alba	talb	1.51	2.09	2.67	43.45



Species ID code	LTR (%)	Genome (Gbp)	Scaff. N50 (Mbp)	Contig N50 (kbp)
acar	4.88	1.799	4.033	79.867
cpic	3.87	2.366	6.606	21.349
scam	0.48	1.225	3.593	34.997
tutt	0.61	1.047	0.246	29.773
ggal	3.01	1.073	12.877	2894.815
mgal	1.76	1.062	0.858	26.671
apla	3.95	1.105	1.234	26.114
cliv	1.06	1.108	3.149	26.579
pgut	4.27	1.069	0.050	26.448
cann	1.09	1.106	4.052	26.738
ccan	4.10	1.154	2.990	38.137
pade	2.69	1.217	5.119	22.195
afor	1.51	1.254	5.072	31.730
pcri	3.53	1.161	0.043	21.679
fper	1.54	1.172	3.956	28.645
mund	2.03	1.117	10.614	55.633
cbra	4.53	1.091	6.954	29.093
tgut	4.94	1.232	8.237	38.639
hleu	2.67	1.178	9.145	105.493
talb	2.67	1.120	0.053	17.226
avit	3.38	1.071	0.057	28.882
ppub	1.37	1.167	2.093	24.809

Figure 4.3 LTR retrotransposon genome content across the Avian lineage. This cladogram shows twenty species from the three major lineages of birds and two outgroup species: the Carolina anole and the Western painted turtle. The LTR (%) column shows the relative proportion of the genome annotated as LTR retrotransposon by the combined RepeatMasker protocol. The third column gives the genome size in gigabase pairs (Gbp). The fourth and fifth columns are indicative measures for assembly quality: the scaffold N50 in megabase pairs, and the contig N50 in kilobase pairs. The cladogram was constructed based on published avian phylogenies (Jarvis et al. 2015; Suh 2016). Species names are given as four letter codes which were defined above in Table 4.6.

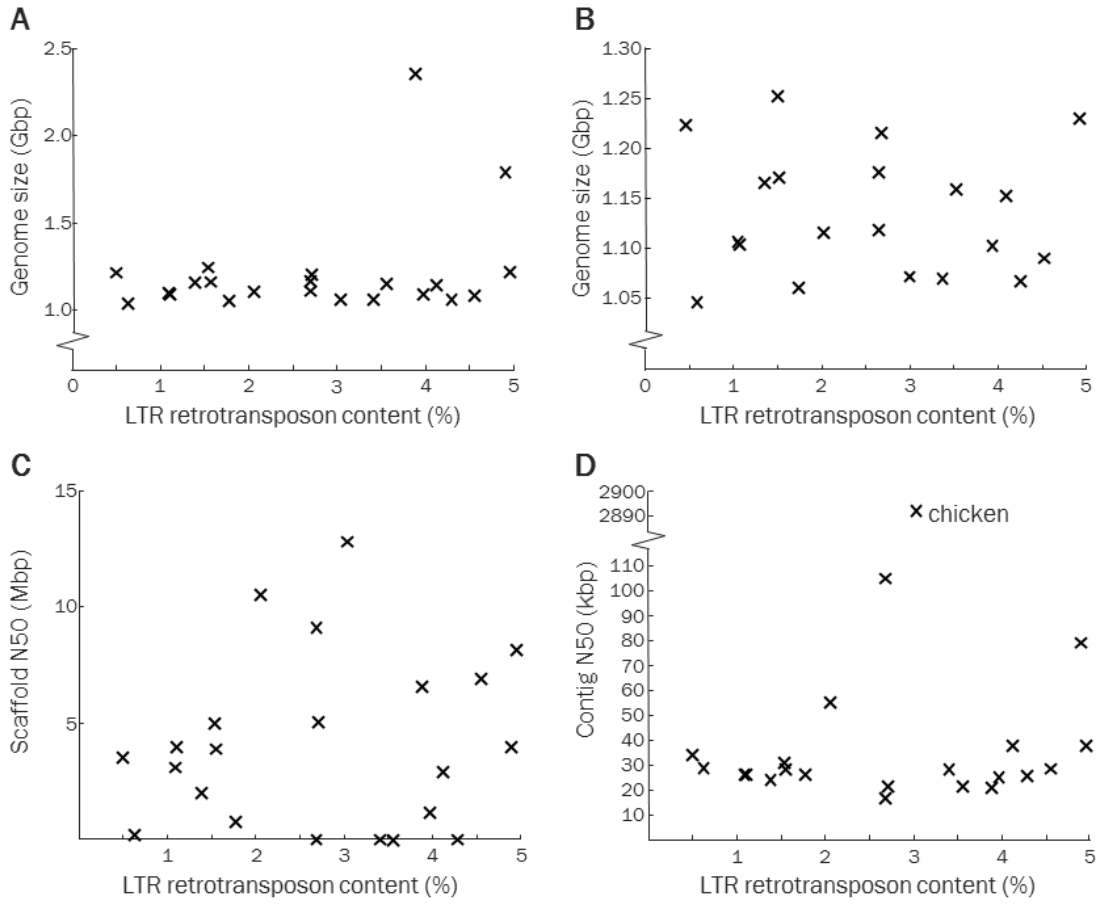


Figure 4.4 Scatter plots showing the absence of correlations between identified LTR retrotransposon content (%), genome size and genome quality. A) LTR content against genome size including the outlier larger reptile genomes (r = 0.31; P = 0.160). B) LTR content against avian genome size (r = -0.05; P = 0.821). C) LTR content against scaffold N50 length (r = 0.12; P = 0.584). D) LTR content against contig N50 (r = 0.16; P = 0.482). The chicken value was excluded from this last correlation as its contig N50 was eighty-fold greater than the average N50 value of the other genomes.

4.5.2 LocaTR analysis of the avian lineage

LTR retrotransposons were identified in sixty-seven avian genomes and six reptilian outgroups. Total LTR retrotransposon content ranged from 0.48 – 9.03 %, with a mean value of 2.68 % and a median of 2.09 %. Five of the reptiles exhibited the five largest contents, with *Python bivittatus* (pbiv) the thirteenth largest (4.29 %). Among the avian genomes, the maximum annotated content was 4.77 % in *Colius striatus* (cstr), the mean value was 2.31 %, and the median was 2.03 %. The number of identified SIEs varied

over three orders of magnitude, with 11 in *Ara macao* (amac) and 6,760 in *Anolis carolinensis* (acar) (mean = 506.14, median = 238). Predictably, the larger reptilian genomes had more SIEs than were observed in any bird (again, except pbiv). When the number of SIEs were scaled by genome size, five of the top ten densities were in bird genomes, the most being in chicken (ggal; fourth highest value overall).

The LocaTR pipeline identified more LTR retrotransposon-derived sequence in all species than was observed with a RepeatMasker analysis alone. However, the additive effect of LocaTR ranged from just 3.26 % in *Cathartes aura* (caur) to 1,000.00 % in pbiv. The mean additive effect was 144.97 %, but this was very skewed, as the median effect was 63.89 %. Those species with limited additive effect also had some of the lowest numbers of identified SIEs. In three species, the secondary BLAST protocol identified no additional sequence: *Eurypyga helias* (ehel; 81 SIEs), *Haliaeetus albicilla* (halb; 122 SIEs), and *Tyto alba* (talb; 115 SIEs).

There was no observable correlation between total LTR retrotransposon content and genome size ($r = -0.13$; $P = 0.297$), but content was positively correlated with the number of identified SIEs ($r = 0.38$; $P = 0.002$). There were positive correlations between the number of SIEs (when scaled for genome size) and contig N50 ($r = 0.48$; $P < 0.001$), and with scaffold N50 ($r = 0.75$; $P < 0.001$). Contig N50 was not correlated with total content ($r = 0.10$; $P = 0.388$), but there was a slight positive correlation between scaffold N50 and total content ($r = 0.29$; $P = 0.018$). Correlations with scaffold N50 must be treated carefully as the species used in this analysis have very different scaffold statistics. Over 50 % of species have a scaffold N50 over 2 Mbp, but in 40 % it is less than 65 kbp.

A comprehensive analysis of the contribution of different LTR retrotransposon groups to the overall annotated content was not completed in this study. However, DIRS elements were not identified in any bird genome, or in the genomes of *Alligator mississippiensis* (amis), *Crocodylus porosus* (cpor) or pbiv. These findings are corroborated by the literature (Piednoël et al. 2011), and support the loss of DIRS elements in the Archosauriformes, and an independent loss within the pbiv lineage.

The annotated LTR retrotransposon content is summarised below in Table 4.7, and mapped to the Sauropsida phylogeny in Figure 4.5.

Table 4.7 The identified LTR retrotransposon content in each analysed species showing the annotated content from a standard RepeatMasker analysis (RM-v), the annotated content from the LocaTR analysis, and the number of identified structurally intact elements (SIEs). Species are shown in alphabetical order with unique four letter codes used below in Figure 4.5, and the reptilian outgroups have been indicated by an asterisk. Any codes used above in Figure 4.3 and Table 4.6 were used again here. The LTR retrotransposon content is shown as percentages of the genome in all cases.

Species name	Code	RM-v (%)	LocaTR (%)	No. SIEs
Acanthisitta chloris	achl	1.26	4.57	88
Alligator mississippiensis *	amis	4.80	6.10	2,014
Amazona aestiva	aaes	1.66	4.28	714
Amazona vittata	avia	1.06	1.10	25
Anas platyrhynchos	apla	0.98	2.26	479
Anolis carolinensis *	acar	4.45	9.03	6,760
Anser cygnoides domesticus	acyg	0.97	2.02	182
Antrostomus carolinensis	acol	1.38	4.50	78
Apaloderma vittatum	avit	0.86	0.95	83
Aptenodytes forsteri	afor	1.08	1.77	155
Apteryx australis mantelli	aaus	0.15	0.48	284
Aquila chrysaetos canadensis	achr	1.85	2.03	442
Ara macao	amac	0.84	0.87	11
Balearica regulorum gibbericeps	breg	1.32	1.39	124
Buceros rhinoceros silvestris	brhi	0.74	1.52	44
Calidris pugnax	cpug	0.57	3.61	280
Calypte anna	cann	0.64	3.30	340
Cariama cristata	ccri	0.73	1.51	76
Cathartes aura	caur	0.92	0.95	19
Chaetura pelagica	cpel	0.67	3.72	312
Charadrius vociferus	cvoc	0.81	3.39	375
Chlamydotis macqueenii	cmac	1.20	2.09	68

Chrysemys picta belii *	cpic	1.30	8.96	3,209
Colinus virginianus	cvir	1.12	1.21	22
Colius striatus	cstr	1.76	4.77	109
Columba livia	cliv	0.59	2.54	264
Corvus brachyrhynchos	cbra	1.97	3.12	1,059
Corvus cornix cornix	ccor	1.61	2.29	701
Coturnix japonica	cjap	1.19	2.09	393
Crocodylus porosus *	cpor	5.14	7.68	1,818
Cuculus canorus	ccan	0.52	4.54	194
Egretta garzetta	egar	1.26	3.14	320
Eurypyga helias	ehel	1.41	1.47	81
Falco cherrug	fche	1.13	1.29	311
Falco peregrinus	fper	1.14	1.28	315
Ficedula albicollis	falb	1.53	1.82	238
Fulmarus glacialis	fgla	1.09	1.14	59
Gallus gallus	ggal	2.17	2.86	1,295
Gavia stellata	gste	0.63	1.53	42
Geospiza fortis	gfor	2.69	3.21	486
Haliaeetus albicilla	halb	1.58	1.65	122
Haliaeetus leucocephalus	hleu	1.83	1.95	276
Lepidothrix coronata	lcor	0.92	1.87	226
Leptosomus discolor	ldis	0.96	1.02	116
Manacus vitellinus	mvit	0.87	2.30	215
Meleagris gallopavo	mgal	1.06	2.80	411
Melopsittacus undulatus	mund	1.52	4.22	642
Merops nubicus	mnub	0.90	1.95	105
Mesitornis unicolor	muni	0.92	1.09	100
Nestor notabilis	nnot	1.22	3.63	85
Nipponia nippon	nnip	1.06	1.51	226

Opisthocomus hoazin	ohoa	1.02	2.09	492
Parus major	pmaj	1.49	2.10	613
Pelecanus crispus	pcri	1.73	3.00	225
Phaethon lepturus	plep	1.46	1.78	134
Phalacrocorax carbo	pcar	1.15	1.21	93
Phoenicopterus ruber	prub	0.96	1.86	56
Picoides pubescens	ppub	0.59	0.94	545
Podiceps cristatus	pcrs	1.28	2.84	111
Pseudopodoces humilis	phum	1.66	2.63	921
Pterocles gutturalis	pgut	1.11	1.26	50
Pygoscelis adeliae	pade	1.28	1.54	196
Python bivittatus *	pbiv	0.39	4.29	652
Serinus canaria	scan	3.48	4.31	934
Struthio camelus australis	scam	0.17	1.09	134
Sturnus vulgaris	svul	1.48	2.04	431
Taeniopygia guttata	tgua	3.54	4.50	1,109
Tauraco erythrolophus	tery	1.50	4.65	133
Thamnophis sirtalis *	tsir	0.83	6.14	2,614
Tinamus guttatus	tgus	0.21	1.46	243
Tyto alba	talb	1.51	1.56	115
Zonotrichia albicollis	zalb	2.29	2.82	338
Zosterops lateralis melanops	zlat	1.84	2.33	416

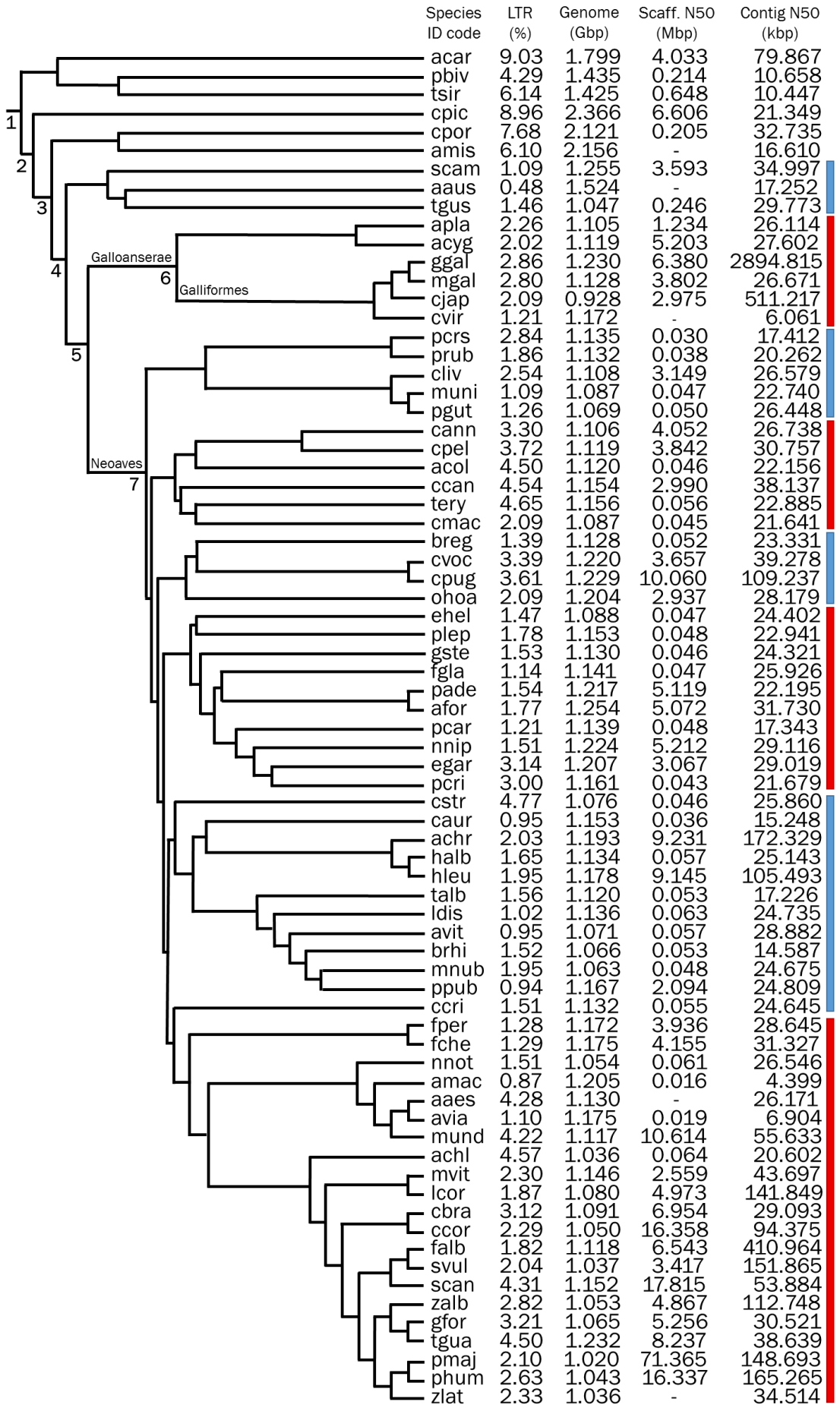

Species ID code
LTR (%)
Genome (Gbp)
Scaff. N50 (Mbp)
Contig N50 (kbp)
acar 9.03 1.799 4.033 79.867
pbiv 4.29 1.435 0.214 10.658
tsir 6.14 1.425 0.648 10.447
cpic 8.96 2.366 6.606 21.349
cpor 7.68 2.121 0.205 32.735
amis 6.10 2.156 - 16.610
scam 1.09 1.255 3.593 34.997
aaus 0.48 1.524 - 17.252
tgus 1.46 1.047 0.246 29.773
apla 2.26 1.105 1.234 26.114
acyg 2.02 1.119 5.203 27.602
ggal 2.86 1.230 6.380 2894.815
mgal 2.80 1.128 3.802 26.671
cjap 2.09 0.928 2.975 511.217
cvir 1.21 1.172 - 6.061
pcrs 2.84 1.135 0.030 17.412
prub 1.86 1.132 0.038 20.262
cliv 2.54 1.108 3.149 26.579
muni 1.09 1.087 0.047 22.740
pgut 1.26 1.069 0.050 26.448
cann 3.30 1.106 4.052 26.738
cpel 3.72 1.119 3.842 30.757
acol 4.50 1.120 0.046 22.156
ccan 4.54 1.154 2.990 38.137
tery 4.65 1.156 0.056 22.885
cmac 2.09 1.087 0.045 21.641
breg 1.39 1.128 0.052 23.331
cvoc 3.39 1.220 3.657 39.278
cpug 3.61 1.229 10.060 109.237
ohoa 2.09 1.204 2.937 28.179
ehel 1.47 1.088 0.047 24.402
plep 1.78 1.153 0.048 22.941
gste 1.53 1.130 0.046 24.321
fgla 1.14 1.141 0.047 25.926
pade 1.54 1.217 5.119 22.195
afor 1.77 1.254 5.072 31.730
pcar 1.21 1.139 0.048 17.343
nnip 1.51 1.224 5.212 29.116
egar 3.14 1.207 3.067 29.019
pcri 3.00 1.161 0.043 21.679
cstr 4.77 1.076 0.046 25.860
caur 0.95 1.153 0.036 15.248
achr 2.03 1.193 9.231 172.329
halb 1.65 1.134 0.057 25.143
hleu 1.95 1.178 9.145 105.493
talb 1.56 1.120 0.053 17.226
ldis 1.02 1.136 0.063 24.735
avit 0.95 1.071 0.057 28.882
brhi 1.52 1.066 0.053 14.587
mnub 1.95 1.063 0.048 24.675
ppub 0.94 1.167 2.094 24.809
ccri 1.51 1.132 0.055 24.645
fper 1.28 1.172 3.936 28.645
fche 1.29 1.175 4.155 31.327
nnot 1.51 1.054 0.061 26.546
amac 0.87 1.205 0.016 4.399
aaes 4.28 1.130 - 26.171
avia 1.10 1.175 0.019 6.904
mund 4.22 1.117 10.614 55.633
achl 4.57 1.036 0.064 20.602
mvit 2.30 1.146 2.559 43.697
lcor 1.87 1.080 4.973 141.849
cbra 3.12 1.091 6.954 29.093
ccor 2.29 1.050 16.358 94.375
falb 1.82 1.118 6.543 410.964
svul 2.04 1.037 3.417 151.865
scan 4.31 1.152 17.815 53.884
zalb 2.82 1.053 4.867 112.748
gfor 3.21 1.065 5.256 30.521
tgua 4.50 1.232 8.237 38.639
pmaj 2.10 1.020 71.365 148.693
phum 2.63 1.043 16.337 165.265
zlat 2.33 1.036 - 34.514
1
2
3
4
5
6
7
Galloanserae
Galliformes
Neoaves

Figure 4.5 Cladogram of the avian lineage, including reptilian outgroups, with the annotated LTR retrotransposon content of each genome. Species names are given with a four letter code as defined in Table 4.7. The total LTR retrotransposon content is given as a percentage of the genome size following annotation by LocaTR. The final three columns are for the genome assembly size, the scaffold N50 length, and the contig N50 length. Five of the assemblies do not contain scaffolds, so the scaffold N50 value is shown as '-'. The cladogram was constructed based on the known phylogeny (Kan et al. 2010; Meiklejohn et al. 2014; Jarvis et al. 2015; Suh 2016), and seven of the nodes are numbered to give their approximate dates of divergence. The cladogram itself is not to scale, but the long branches after the Neoaves radiation have been chosen to show the length of separation of these major groups. Numbered nodes: 1) Separation of the Lepidosauromorpha (lizards and snakes) and Archosauromorpha ~265 MYA, 2) separation of the Pantestudines (Turtles) and Archosauriformes ~260 MYA, 3) separation of the crocodilian and avian lineage ~250 MYA, 4) separation of the Palaeognathae and Neognathae ~105 MYA, 5) separation between the Galloanserae and Neoaves ~100 MYA, 6) split between the Galliformes and Anseriformes ~60 MYA, and 7) the Neoaves radiation ~65 MYA following the K-T mass extinction event. The short branch lengths after node 7 have been drawn to represent the incomplete lineage sorting following rapid diversification. The eight taxonomic groupings used for the GLM were: Paleognathae (scam-tgus), Galliformes (apla-cvir), Columbea (pcrs-pgut), Caprimulgiformes and Otidimorphae (cann-cmac), Cursorimorphae and Opisthocomiformes (breg-ohoa), Aequornithia and Phaethantimorphae (ehel-pcri), Afroaves (cstr-ppub), and Australaves (ccri-zlat). The eight groupings are indicated in order by alternating blue and red blocks.

Total LTR retrotransposon content is highly variable across the avian genomes, and even sister species are strikingly different. *Falco peregrinus* (fper) and *F. cherrug* (fche) match closely (1.28 and 1.29 % respectively), but the two other sister species groupings of *Amazona aestiva* (aaes; 4.28 %) and *A. vittata* (avia; 1.10 %), and *Corvus brachyrhynchos* (cbra; 3.12 %) and *C. cornix* (ccor; 2.29 %) show marked differences. Some of this variation is due to differences in the number of identified SIEs. The two *Falco* species have 315 and 311 SIEs respectively, but aaes had 714 whilst avit had 25, and cbra had 1,059 whilst ccor had 701. Such variation between closely related species makes the identification of any phylogenetic effect difficult, as the number of SIEs (shown above to be correlated with genome quality) confounds observable patterns.

The fitted GLM showed that both contig N50 and genome size were non-significant variables in explaining the variation in total LTR retrotransposon content. However,

both the broad taxonomic groupings (Figure 4.5) and scaled SIE numbers were significant variables in the model (both P = 0.001), although there was no significant interaction term. Taxonomic groupings explained 28.49 % of the observable variation, and the scaled SIE number explained a further 12.95 %. This supports lineage specificity in LTR retrotransposon content in broad taxonomic groupings separate since the K/T extinction event. Interestingly, when all Neoaves were grouped as a single taxon in the model, the variable was non-significant (P = 0.068).

Therefore, this expanded analysis of the avian lineage corroborates the results of the homology-based analysis, and does not support a significant difference in LTR retrotransposon content between the Galliformes and Neoaves. However, it is possible that improvements to assembly quality could significantly increase the total identified LTR retrotransposon content in individual species and alter these conclusions.

Comparison with the results of the purely homology-based annotation

The LTR retrotransposon content of twenty-one species (excluding the chicken due to the change in assembly) were analysed with all three methodologies: the standard RepeatMasker analysis, the homology-based protocol (section 4.5.1), and the full LocaTR analysis (above). The homology-based protocol increased the annotated LTR retrotransposon content in all species compared to RepeatMasker alone, and it was predicted that the LocaTR analysis would either match this content or increase it further, as lineage-specific or divergent SIEs could be detected. However, as Table 4.8 shows, the LocaTR analysis gave lower values for over half the analysed species (red shading).

For each species where the value 'reduced' in the LocaTR analysis, the element position files were overlapped, and the 'missing' sequences were analysed with RepeatMasker. In each case the sequences responsible for the higher content in the homology-based analysis had at least partial CR1 homology, as elements were commonly fragments matching the *pol* gene. Strict filtering was used in LocaTR to reduce the potential for CR1 contamination, but some false positives were called in the homology-based analysis.

Table 4.8 A comparison of the annotated LTR retrotransposon content based on three search methodologies. These twenty-one species were analysed using a standard RepeatMasker analysis (RM-v), the purely homology-based identification from section 4.5.1 (Homology), and the full LocaTR analysis. Species are in alphabetical order with their four letter codes, and the two reptilian outgroups have been indicated by an asterisk. The annotated values are as a percentage of total genome length in all cases, and the values in the LocaTR column have been shaded red if they are less than the annotated value from the homology-based identification.

Species name	Code	RM-v (%)	Homology (%)	LocaTR (%)
Anas platyrhynchos	apla	0.98	3.95	2.26
Anolis carolinensis *	acar	4.45	4.88	9.03
Apaloderma vittatum	avit	0.86	3.38	0.95
Aptenodytes forsteri	afor	1.08	1.51	1.77
Calypte anna	cann	0.64	1.09	3.30
Chrysemys picta belii *	cpic	1.30	3.87	8.96
Columba livia	cliv	0.59	1.06	2.54
Corvus brachyrhynchos	cbra	1.97	4.53	3.12
Cuculus canorus	ccan	0.52	4.10	4.54
Falco peregrinus	fper	1.14	1.54	1.28
Haliaeetus leucocephalus	hleu	1.83	2.67	1.95
Meleagris gallopavo	mgal	1.06	1.76	2.80
Melopsittacus undulatus	mund	1.52	2.03	4.22
Pelecanus crispus	pcri	1.73	3.53	3.00
Picoides pubescens	ppub	0.59	1.37	0.94
Pterocles gutturalis	pgut	1.11	4.27	1.26
Pygoscelis adeliae	pade	1.28	2.69	1.54
Struthio camelus australis	scam	0.17	0.48	1.09
Taeniopygia guttata	tgua	3.54	4.94	4.50
Tinamus guttatus	tgus	0.21	0.61	1.46
Tyto alba	talb	1.51	2.67	1.56

These findings do not undermine the general results presented as part of my paper published last year (Mason et al. 2016), but they do highlight the need for strict filtering of LINE elements during annotation of LTR retrotransposons. Even under the stricter filtering used in LocaTR, identification of CR1 elements remained an issue, and further work is needed to reduce the likelihood of false positives without limiting the detection ability of the pipeline.

4.6 Discussion

4.6.1 The LTR retrotransposons of the chicken genome

The LocaTR analysis of the new chicken Galgal5 assembly identified an additional 3.7 Mb of LTR retrotransposon-derived sequence in the genome compared to Galgal4, taking the total content to 35.2 Mb (2.86 % of the genome). Overall, this means that the annotated LTR retrotransposon content of the chicken has been doubled during this PhD project. This includes the identification of 1,295 structurally intact LTR retrotransposons (222 more than in Galgal4), 40.23 % of which are in clusters unrelated by insertion age or genera.

Compared to Galgal4, the Galgal5 assembly is 183.3 Mb longer, with a ten-fold greater contig N50 length, and now includes chromosomes 30, 31 and 33, as well as improved assemblies of chromosomes 16, 25 and W (Warren et al. 2017). Given these improvements, and the previously observed effects on repeat annotation between the Galgal3 and Galgal4 assemblies, it was unsurprising that the LTR retrotransposon content was increased. However, the totals identified on the macrochromosomes (particularly of SIEs) were lower in Galgal5 than in Galgal4, likely due to repeats collapsing, and unplaced contigs being correctly included. The improved W chromosome assembly counteracted this effect with its high repeat content, as did the elevated LTR retrotransposon content observed on the microchromosomes. This created a bimodal relationship between chromosome length and total LTR retrotransposon content, which, in Galgal4, was a simple linear relationship, where longer chromosomes had higher content. In Galgal5, the macrochromosomes exhibited

this effect, but this was lost on the microchromosomes, with content independent of chromosome length.

This bimodal result is much more biologically representative than the 'one model fits all' Galgal4 result, especially when you consider each integration as an independent event, rather than only looking at the current, total distribution and content. As the macrochromosomes account for a large proportion of the total genome size individually, recurrent independent integrations will accumulate based on the probability of inserting within a sequence of that size. However, for microchromosomes which account for approximately 1 % of the genome each, the number of integrations is largely due to chance. The observation that microchromosome integrations follow a random distribution relative to genes is related to this insertion probability, as well as the higher microchromosome gene density.

Overall, the distribution of LTR retrotransposons relative to gene features was very similar between the two assemblies, even with the recent augmentation of the annotation file. Interestingly, seventy-two SIEs significantly overlapped with recently identified lncRNA genes (Kuo et al. 2017), and over half of these were SIEs within clusters, significantly more than expected by chance. In their Galgal3 analysis of LTR retrotransposons, Bolisetty and colleagues (2012) hypothesised that the clusters had cellular functions, either as conserved cytoskeletal binding regions or sites which promoted recombination. In chapter 3, I instead suggested that these sites were simply regions where insertions had limited negative effects on the hosts, so could be maintained and expanded over time. Furthermore, absence of constrained sites and low recombination rates in the clusters provided no support for Bolisetty's hypotheses. However, it is possible that the long-term maintenance of structural integrity in the clusters facilitates the co-option of these integrations as cellular lncRNAs. As lncRNAs are generally species-specific, this may explain the functional role of clusters which lack evidence of phylogenetic constraint (Kapusta et al. 2013; Kuo et al. 2017). Further work is needed to fully characterise the cellular effects of these LTR retrotransposon-derived lncRNAs, and it is likely that some non-SIEs also overlap with lncRNA genes.

4.6.2 Heterogeneous LTR retrotransposon content across the avian lineage

In birds, the LTR retrotransposon content is highly heterogeneous, ranging from 0.48 % to 4.77 %. In general, the reptilian outgroups had more LTR retrotransposon-derived sequence, although this distinction was lost when content was scaled for genome size. The number of structurally intact elements also varied considerably across the analysed genomes, ranging from 11 to 6,760 (equating to 9.1 SIEs per Gb to 3,757.3 SIEs per Gb when scaled to genome size), and much of this variation was likely due to the highly variable assembly quality (see below).

The distribution of genomic content is highly lineage specific, with much of the variation explained by broad taxonomic groupings present at the K/T extinction event approximately 65 million years ago. Representation of species within these groups is not equal, but neither is the extant diversity of these groups. The observed broad phylogenetic effect is reduced in some lineages by very different LTR retrotransposon content values, even between very closely related species. This may be an accurate biological representation of very narrow lineage-specific expansions and contractions, but could also be due to issues with the assembly. Further lineage-specific resolution could be gained through construction of a phylogenetic tree based on multiple sequences and features, although such trees have recently been shown to exhibit a hard polytomy at the base of the Neoaves, corresponding to the rapid superorder diversification after the K/T extinction event (Suh et al. 2015; Suh 2016). Further analysis is needed to quantify the specific ERVs responsible for contraction/expansion in each lineage, and to date these periods of potentially rapid genome evolution.

The observed heterogeneity of LTR retrotransposon content across the avian lineage does not support the previously proposed hypothesis that there is a deficit of these elements in galliform birds (Bolisetty et al. 2012). Whilst the average content in the galliforms (and the Galloanserae more widely) is lower than the average values in three other broad taxonomic groupings (the Caprimulgiformes and Otidimorphae, Cursorimorphae and Opisthocomiformes, and Australaves), there is no consistent pattern to support higher general LTR retrotransposon content in the Neoaves.

Consistent with the literature on wider repeat content (Kapusta & Suh 2017), the Paleognathae exhibit some of the lowest annotated LTR retrotransposon values, despite

having some of the largest avian genomes. Recently published work by Kapusta and colleagues (2017) related the larger genome sizes in flightless birds (including penguins) to reduced rates of medium and large scale deletions, and the presence of generally older transposable elements, rather than increased genome size reflecting new transposable element activity. The results obtained here support these findings, although the LTR retrotransposon content for penguins (afor and pade; Figure 4.5) appears representative within its taxonomic grouping. However, loss of flight in penguins occurred following the Neoaves diversification (the potential time of repeat group expansion), and the penguin genomes exhibit moderate LTR retrotransposon levels within their group, but have relatively contiguous genomes. Again, further characterisation of the representative ERVs and their ages would clarify periods of retrotransposon activity and itheir potential contribution to genome evolution.

The effect of assembly quality on annotated LTR retrotransposon content

Across the avian genomes, LTR retrotransposon content was not correlated with genome size, but there were positive correlations between the number of identified SIEs and genome assembly contiguity. Identification of few SIEs limits the total annotated content directly, but also reduces the additive effect of the secondary BLAST protocol. This effect was evident between chicken genome assemblies, and sister species with markedly different SIE number and total LTR retrotransposon content. Due to lineage-specific effects, there was no clear difference in LTR retrotransposon content based purely on genome contiguity, but it was a significant variable in the fitted GLM, second to the taxonomic grouping.

Many of the analysed genomes were first draft publications based on short read sequencing technology, and consequently exhibited low N50 values (Jarvis 2014). Some of the more recent avian genome sequencing projects (including Japanese quail, great tit, and the Galgal5 assembly) made use of PacBio long read sequencing which generates much more contiguous assemblies. Long read sequencing is now one of a series of tools, including Hi-C (sequencing of sites with chromatin interactions) and BioNano high resolution optic mapping (creation of long physical maps for scaffold joining), which are being implemented to create high quality, contiguous genome assemblies at a fraction of

the cost required for the 'finishing' of genomes such as human and mouse (Nagano et al. 2013; Cao et al. 2014; Howe & Wood 2015; Mak et al. 2016; Worley 2017). These techniques have recently been used on a diverse range of species to create 'gold standard' genomes (Shi et al. 2016; Bickhart et al. 2017; Jiao et al. 2017; Mohr et al. 2017; Zimin et al. 2017).

Similar improvements to many of the avian genomes would greatly facilitate the accurate annotation of repetitive DNA, as these sections often 'collapse' during *de novo* short read assembly, but are resolved in long read sequencing or optical mapping (Bergman & Quesneville 2007; Alkan et al. 2011; Zhang et al. 2011; Schatz et al. 2012; Howe & Wood 2015; Michael & VanBuren 2015; Weissensteiner et al. 2017). It is therefore possible that future analyses of LTR retrotransposon content in birds could reveal a deficit in the Galliformes compared to other lineages. However, the evidence presented here does not support that hypothesised deficit, and it is likely that future work would further characterise the lineage-specificity of LTR retrotransposon repeat content across different phyla. Further characterisation of the proportion of different ERV genera, and their subsequent expansion and contraction across the lineage, is needed to better understand retrotransposon dynamics, periods of co-infection with exogenous retrovirus relatives, and the role these elements play in genome evolution.

4.6.3 Analysis of multiple genomes with LocaTR

No issues were experienced moving from analysis of just the chicken genome to analysis of multiple species. All custom-built scripts in the pipeline were written to ensure that variable sequence naming systems were handled appropriately, although the large numbers of small contigs in some assemblies did increase the processing time of the structure-based identification methods, particularly LTR_STRUC. Validating the identified SIEs was the most computationally intensive part of each analysis, usually due to the large numbers of putative sites identified by LTR Harvest, where the false positive rate (as with chicken) was commonly over 90 %.

Most sections of the pipeline were highly parallel, but RetroTector and LTR_STRUC required desktop architecture. The programs could be run simultaneously on the same

machine, but only one genome at a time. Consequently, the homology-based protocols, and the MGEScan_LTR and LTR Harvest identifications were completed within a fortnight, but the LTR_STRUC and RetroTector identifications took over six months using multiple machines. For such large, multi-species analyses it would be beneficial to re-write these programs as Linux-based tools, but this is unlikely to happen due to issues with source code availability, and the requirement for access and management of SQL databases for RetroTector. If a quick analysis of multiple genomes is required, it may be necessary to host multiple virtual machines on a cloud-based service to complete the analyses in parallel. Alternatively, these program analyses could be ignored, but, as seen with the chicken analysis in chapter 3, both LTR_STRUC and RetroTector identified LTR retrotransposons not found by other programs.

The unwanted identification of LINEs (such as CR1) remains an issue with LocaTR. This was particularly evident during the tBLASTx searches of the homology-based protocols, as these detect *reverse transcriptase* from all retrotransposon classes. Great effort has been taken to remove LINE-homologous sequences, but some are still annotated as LTR retrotransposons. E-value thresholds could be adapted to see if this reduces LINE detection. It would also be beneficial to add LINE-specific pHMMs to the validation process, as these could highlight sites which have a greater homology to these elements than LTR retrotransposons. As with any annotation process, manual checks and tests by the user are recommended. This should include identification of the percentage of ambiguous bases (Ns) in any annotated element, and plotting the distribution of putative element lengths.

More generally, the LocaTR pipeline would benefit from some code revisions to improve efficiency. Much of the coding was completed towards the start of this PhD project, and could certainly be improved. This would include better memory, loop and function management, and the handling of intermediary files. Any improvements would avoid the use of accessory software, as one of the major benefits of LocaTR is that it is largely self-contained. One major revision would be to use standardised BED formats for positions files, rather than the custom file organisation used throughout. This would make the output files more standardised and immediately transferrable to other software or to genome browsers such as Ensembl.

4.7 Concluding remarks

This broad analysis of LTR retrotransposons across the avian lineage does not support the previously proposed deficit of these elements in galliform birds compared to the Neoaves. Overall, LTR retrotransposon content is highly variable, but shows the effects of lineage specificity as far back as the rapid diversification after the K/T mass extinction. However, these findings may be confounded by the variability in genome assembly quality, so annotated content will likely increase as assembly quality improves. Concordantly, the updated Galgal5 chicken genome assembly had a higher total LTR retrotransposon content and more identifiable intact elements than in Galgal4.

The LocaTR pipeline was used successfully across multiple genomes, although such a large-scale analysis highlighted several areas for pipeline improvement. Specifically, the implementation of LTR_STRUC and RetroTector, and comprehensive handling of false positives exhibiting LINE homology.

Chapter 5: Materials and Methods (ii)

This chapter outlines the methodology used for chapters 6 and 7. These chapters move away from the annotation of all LTR retrotransposon elements in assembled genomes, to instead focus on a single class of endogenous alpharetroviruses - Avian Leukosis Virus subgroup E (ALVEs) – and their identification in unassembled, whole genome (re)sequencing (WGS) data.

5.1 Development of an ALVE identification pipeline using Hy-Line and Roslin J-Line DNA re-sequencing data

5.1.1 Genomic Resources

Whole genome resequencing data from eight Hy-Line (HL) elite commercial lines was used with permission from Hy-Line International following its use in the development of the 600K chicken SNP array (Kranis et al. 2013). Sequencing data for each line came from pooled libraries of ten individuals. All sequenced lines were from one of three well described commercial breeds: White Leghorn (WL; named WL1-5), White Plymouth Rock (WPR; sister lines named WPR1 and WPR2), and Rhode Island Red (RIR). Three of the sequenced lines (WL2, WL3 and WPR1) used pools of solely females, and the other five lines used only males.

In addition, individual whole genome resequencing data was used from nine Roslin J-Line (JL) females (EBI ENA: PRJEB15189) and a ten male JL pool (sequenced for the 600K chicken SNP array). This line is a Brown Leghorn (BL) created from a mixture of six experimental lines (originally derived from the same ancestral population) which underwent differential selection for over fifty years, now retained at the Roslin Institute, UK (Blyth 1954; Blyth & Sang 1960).

All sequencing data were Illumina paired-end 101 bp reads with a 500 bp insert size. Data were quality checked with FastQC v0.11.2 (Andrews 2012) and no trimming was required. Each dataset was individually mapped to the Galgal5 reference genome (GenBank: GCF_000002315.4) using BWA-mem v0.7.10 (Li 2013) and average genome coverage was calculated using the samtools v0.1.19 mpileup tool (Li et al. 2009).

5.1.2 ALVE identification with a custom pipeline

A custom pipeline was created to identify ALVE insertions in paired-end sequencing data. This pipeline makes use of existing, commonly used and freely available data manipulation software, coded through seven independent scripts. The code files are available on the CD associated with this thesis (Appendix 1) and online in a GitHub repository: https://github.com/andrewstephenmason/ALVE_ID_pipeline.

The pipeline identifies and extracts paired reads from resequencing data where at least one of the reads has ALVE homology. Extracted read pairs are mapped to the reference genome, enabling identification of the genomic insertion coordinates. Known genomic locations for assembled alpharetroviral elements are filtered out and the remaining putative insertions sites are manually checked. The pipeline, and its use with the HL and JL data, is described more completely in the following subsections.

Pseudochromosome mapping, read subtraction and reference genome mapping

Reads from each dataset were individually mapped to an ALV-derived 'pseudochromosome' using BWA-mem. The pseudochromosome was constructed from eleven ALV-derived reference sequences (Appendix 2: AF08) joined by 1,000 bp of Ns, with the pseudochromosome terminating in 1,000 bp of Ns (ambiguous bases).

All pseudochromosome-mapped reads were subtracted from the original FASTQ files, along with their read pairs (but filtering out secondary alignments), to produce new 'reduced' FASTQ files (Figure 5.1). Read subtraction was completed using tools from samtools and BASH commands, filtering based on the SAM flag. The reduced FASTQ files were mapped to the chicken Galgal5 reference genome assembly, again with BWA-mem, and reads with a mapping quality lower than 20 were removed. The resulting BAM file was converted to BED-6 format using the BEDTools v2.23.0 (Quinlan & Hall 2010) bamToBed tool, splitting clipped reads into distinct BED entries. Entries 12 bp or fewer apart were merged with the BEDTools mergeBed tool, and these putative ALVE insertion site regions were removed if they were shorter than 200 bp.

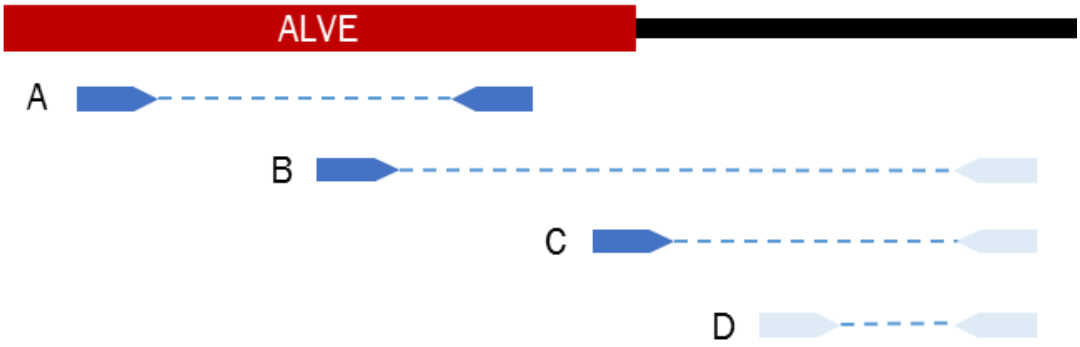


Figure 5.1 Reads mapped to the retroviral pseudochromosome. For read pairs, there are four possible mapping scenarios. Both reads in the pair can map to the pseudochromosome (A), but this gives no positional information relative to the host genome. One read of the pair can map fully (B) with the mate presumably mapping to the host genome. This locates the read, providing general positional information. In a similar manner, if one read of the pair maps at the very end of the retroviral sequence some of the read is clipped (C). The clipped sequence is host genome, so provides exact coordinates for the insertion point. Finally, most read pairs will be unmapped as they have no retroviral sequence (D). Mapped reads are in dark blue and unmapped reads in light blue. The dotted lines represent the distance between read pairs. The black bar represents the 1 kb of Ns which separate the pseudochromosome retroviral sequences.

Filtering and manual confirmation of putative insertions

Putative ALVE insertion site regions were compared to known alpharetroviral locations in the Galgal5 reference genome using the BEDTools intersectBed tool. Putative insertion site regions were removed if they overlapped with known assembled locations. Assembled alpharetroviral-derived sequence locations were identified using a BLASTn (Altschul et al. 1990) protocol with an E-value threshold of 10^{-10}, and identified genomic coordinates were converted into BED-6 format. Thirty reference sequences were used in total, including all those used in the pseudochromosome (Appendix 2: AF07). The extra sequences were added to identify EAV and ART-CH (avian retrotransposon of chickens) alpharetroviral sequences, as these share significant homology with ALV-derived sequence; enough for multi-mapping.

The remaining putative ALVE insertion site regions were filtered based on the presence of clipped reads. These reads map correctly to the reference genome for only part of their length, so may indicate the exact insertion site if the remainder of the read is

homologous to ALVE sequence. Higher confidence scores were given to those regions exhibiting both 5' and 3' clipped reads, as these may support the putative insertion site from both ends of the insertion (Figure 5.2). Regions were checked manually using IGV Desktop for Windows v2.3.60 (Thorvaldsdóttir et al. 2013). Clipped read sequences were used as queries for the NCBI BLASTn megablast and blastn algorithms to check for homology to ALV-derived sequences, and the insertion hexamer and was noted.

Figure 5.2 Clipped read support for ALVE insertion sites. Scenarios A, B and C represent reads (black arrows) across a putative insertion hexamer sequence, with clipped ALVE insertion sequences angled off and in red. A) a site with no insertion showing the reads mapping correctly across the region. B) a site with an insertion where there are clipped reads at both the 5' and 3' ends. The reads correctly map to the reference genome at the start but then become clipped due to the presence of the insertion. As the hexamer sequence is duplicated it is represented at both ends of the insert, so reads cover it from both directions. With an intact insertion, the red arrow sequence would be the LTRs at opposite ends of the ALVE. C) a site where the sequencing data only provides evidence for clipped reads at one end of the insertion. These have lower confidence than those sites exhibiting scenario B.

Putative insertion site existing nomenclature and nearby gene features

Most previously characterised ALVEs lack published genomic coordinates, but published insertion hexamer sequence, immediate flanking sequence and diagnostic PCR primer sequences were used as BLASTn queries against the Galgal5 assembly to identify known ALVE insertion sites (Benkel 1998; Chang et al. 2006; Smith & Benkel 2009a; Smith & Benkel 2008; Chen et al. 2014). In addition, Professor Bernhard Benkel (Dalhousie University, Nova Scotia, Canada) kindly shared validated insertion sites from his recent 'Chickens of the World' study of ALVE diversity, built on forty years of ALVE research (Rutherford et al. 2016, and additional manuscripts in preparation).

In addition, the 100 bp of 5' genomic sequence flanking identified insertion sites was used as a BLASTn query against the Galgal4 assembly (GenBank: GCF_000002315.3)

to identify the predicted insertion site locations in the previous assembly. These positions were overlapped with the Ensembl Galgal4 version 79 annotation file using the BEDTools intersectBed and closestBed tools to identify overlapped genes or the shortest distance from an ALVE to the nearest coding feature. Well described ALVEs are often within or near genes, so this annotation also added confidence to assigning an existing name to an identified insertion, predicting its size and structural integrity, and suggesting the potential impact of any insertion on the phenotype of the bird.

Developing nomenclature for novel insertions

Names were given to novel insertions following the format "ALVE_ros001", where 'ros' is an abbreviation for The Roslin Institute. Attempts were made to develop names specific for each insertion using chromosomal locations, insertion hexamer sequences, location relative to genes, and ALVE intactness, but each of these was not deemed future-proof (Table 5.1). In addition, names including these details became very cumbersome, whereas the priority was to classify these insertions as ALVEs and then refer the researcher to this work describing their locations. As such, new names have been applied to previously identified insertions that lacked any clear sign that they were ALVEs, such as the Benkel 'N4' and 'New11' elements identified in the Hy-Line birds.

Table 5.1 Identified future-proofing concerns with ALVE-specific nomenclature.

Naming feature	Issues for naming
Location	Chromosome unlikely to change, but coordinates would shift with each revised build. Names would also be long: ALVE3 would include Gal5:20:10309347 as part of its name.
Insertion hexamer	Some hexamers contain line-specific SNPs and there is inconsistency in the literature for whether hexamers are reported as observed in genome assemblies, or in the same orientation and from the same strand as the ALVE.
Gene feature	Gene names and exon numbers change over time, particularly with identification of additional isoforms. There may also be overlapping features, such as a gene and non-coding RNA.
ALVE intactness	This is unknown for most ALVEs, and may differ between lines.

5.2 Characterisation of the ALVEs identified in the Hy-Line and Roslin J-Line DNA resequencing data

5.2.1 Genotyping of identified ALVEs in the Hy-Line commercial flocks

Bird DNA samples

Genomic DNA samples were collected by Hy-Line for male birds in each of the eight elite layer lines. Data was collected for each line for 15 generations between 1996 and 2011. The birds used for the sequencing detailed above were from the 2008 generation. Sample sizes for each line varied across the generations, but over nine thousand DNA samples were available for ALVE genotyping. In each case, genomic DNA samples were extracted using a standardised DNA spooling protocol following salt and ethanol extraction from wing vein blood samples.

Diagnostic KASP assay design

Flanking chicken genome sequences were given to Hy-Line for the design of the KASP assay primers. For each insertion this required 100 bp of up- and down-stream flanking sequence as well as all the available sequence from the insert, obtained from the 5' and 3' clipped reads. Primers were designed using the Kraken Primer Picker software with an optimum length of 20 - 25 bases, aiming for equal GC content between the primers, ideally between 40 % and 50 % GC. Assays used a four-primer approach with two fluorophore-labelled, allele-specific primers for the presence or absence of the ALVE insert, each with their own reverse primer pair (Figure 5.3).

Assays were conducted on 1,536-well plates using 1μl total reaction volumes in each well. Reactions used dehydrated DNA samples, primers and the LGC KASP 2x Mastermix V4.0 1,536 formulation, following the original KBiosciences KASP protocol. PCR used a 61˚C to 55˚C annealing temperature touchdown protocol for ten cycles and then 55˚C for twenty further cycles. Plates were then read using the PHERAstar Plus SNP plate reader software which called the individual genotypes. To gain increased resolution, or to ensure samples were correctly grouped by genotype, additional PCR cycles were completed in groups of three and the plates read again.

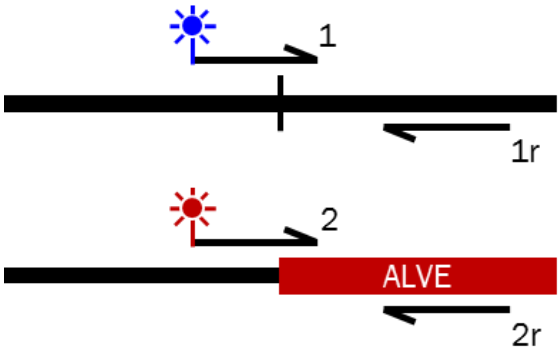


Figure 5.3 KASP assay primer rationale for wildtype and insert-containing sites. Primer 1 (wildtype) and primer 2 (insert) are fluorophore-labelled primers and their amplification enables genotyping direct from solution. The starting sequence for the genotype-specific primers is often the same, but they differ when they cross the insertion hexamer, with primer 1 continuing through host genome sequence and primer 2 entering the ALVE insertion.

The primers were redesigned for assays which failed or produced unexpected results in initial tests. Redesigns included identification of, and controlling for, SNPs within the primer binding regions, and moving primers which bound highly repetitive regions of the genome, identified by lower case letters in the masked genome build.

ALVE frequencies were recorded for each of the identified ALVEs across all the lines. Due to commercial sensitivity, exact insert frequencies will not be reported here, instead the following categories will be used: absent ($f = 0$), rare ($0 < f \leq 0.1$), low ($0.1 < f \leq 0.25$), medium ($0.25 < f \leq 0.75$), high ($0.75 < f < 1$) and fixed ($f = 1$).

Diagnostic PCR assay design

In addition to the KASP assays, standard PCR assays were also developed for each of the identified ALVE insertions. Existing assays were used for ALVE3, ALVE9, ALVE15, ALVE21, ALVEB5 (Benkel 1998), ALVE-TYR (Chang et al. 2006), ALVE-NSAC1 (Smith & Benkel 2009b), and ALVE-NSAC3 (Smith & Benkel 2008). The upstream primer of the published ALVE1 assay (Benkel 1998) was redesigned for a higher melting temperature (T_m), but the other two primers were retained.

New assays were designed for the remaining insertions using the 500 bp 5' and 3' insertion flanking sequence, and the sequence from the soft-clipping reads supporting the insertion site. Primers were designed using the Primer3 software (Rozen & Skaletsky 2000) hosted on the Biology Workbench 3.2 (Subramaniam 1998), requiring product sizes of 100–500 bp, a GC clamp size of 1 base, optimal primer length of 22 bases, and minimum primer length of 20 bases. Most assays used three primers, but ALVE_ros005 and ALVE_ros007 only used a two primer assay.

For all assays, PCR was conducted in 10µl reaction volumes, with equal concentrations of primers, using the Roche FastStart Taq kit (04738357001). Each PCR reaction began with a 4 minute activation at 95°C, then had 35 cycles of 30 seconds denaturing at 95°C, 30 seconds annealing at 60°C and 45 seconds elongation at 72°C, and then finished with a 7 minute final extension at 72°C. The ALVE15 and ALVEB5 PCR reactions had an annealing temperature of 50°C due to the low primer T_m. In addition, the 45 second elongation within each cycle was extended to 3 minutes for ALVE_ros007. Samples were run on a 1% agarose gel with Invitrogen SYBR Safe DNA gel stain (S33102), using the Bioline Hyperladder I (BIO-37045) as the marker ladder.

5.2.2 Genotyping of identified ALVEs in the Roslin J-Line

Bird sampling and sample preparation

Blood samples were taken from the wing veins of all thirty-two JL individuals in the current flock. Blood sampling for DNA extraction was conducted after ethical approval under project licence PPL60/4056. Genomic DNA was extracted from whole blood using the ThermoFisher Scientific DNAzol protocol, following the manufacturer's instructions. DNA concentration was quantified and samples diluted to a concentration of 10ng/µl.

Diagnostic PCR assay design

Existing PCR primers were used for the ALVE3 and ALVE15 assays (Benkel 1998). A three-primer PCR assay was developed for ALVE_ros011 using the 500 bp 5' and 3'

insertion flanking sequence, and the sequence from the 5' soft-clipping reads supporting the insertion site. As above, primers were designed using Primer3, requiring product sizes of 100-500 bp, a GC clamp size of 1 base, optimal primer length of 22 bases, and minimum primer length of 20 bases.

For all three assays, PCR was conducted in 10µl reaction volumes, with equal concentrations of primers, using the Roche FastStart Taq kit. ALVE3 and ALVE_ros011 were cycled 35 times with an annealing temperature of 60°C. As stated above, the ALVE15 PCR reaction had 35 cycles, but an annealing temperature of 50°C. Samples were run on a 1% agarose gel with Invitrogen SYBR Safe DNA gel stain, using the Bioline Hyperladder I as the marker ladder.

5.2.3 Probability of missing an ALVE insertion within the WGS datasets

Probability of missing an ALVE insertion in the JL individual sequencing data

A model was constructed in Python to determine the probability of missing an insertion of a given frequency when choosing the nine individuals for the sequencing project (Figure 5.4). For a given ALVE frequency, individuals from a modelled population of 32 birds were randomly assigned an insertion genotype based on Hardy-Weinberg equilibrium. Nine birds were chosen at random without reselection and the observed frequency noted. As coverage for the JL was >18X for all individuals it was assumed that heterozygote insertions would always be identified within the sequencing data if they were present, and that variability in read coverage or allele specific amplification in library preparation would have little impact on insertion discovery.

The model was run for all possible insertion frequencies from a 32 individual flock (from an allele frequency of 1/64, incrementing in equal steps to 1), repeated one million times, and the probability of missing the insert was calculated in each case. This detection calculation follows a binomial distribution. Additionally, the model was run varying the modelled sample size from one individual to the whole flock to determine the number of individuals required to give 90 %, 95 % or 100 % detection probability for all possible insertion frequencies.

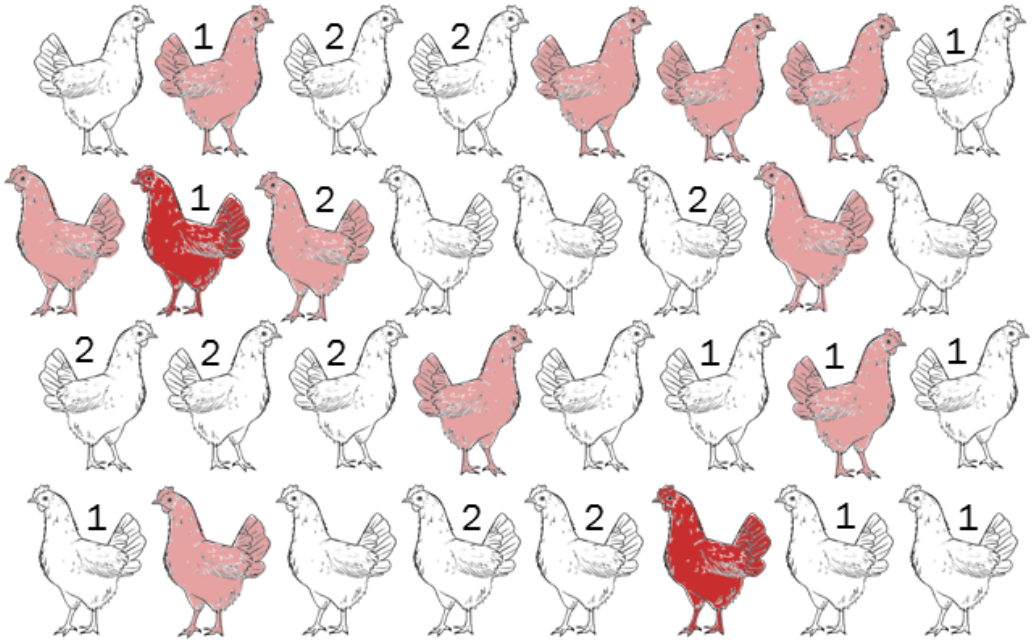


Figure 5.4 Schematic highlighting the issues with sampling bias for insertion discovery from individual sequencing datasets. Under Hardy-Weinberg equilibrium, a flock of thirty-two birds with an insert frequency of 0.2 would have ten heterozygote birds (pink) and two birds homozygous for the insert (red). Depending on the nine birds chosen for sequencing, inserts could be missed (from the 28,048,800 possible ways to choose nine birds from a thirty-two bird flock, an individual carrying the insert could be missed 167,960 times (0.60 %)) or the observed frequency could differ from the actual frequency. For example, if the nine birds with a number 1 were chosen, the observed frequency would be 0.22, but if the birds with a number 2 were chosen, the observed frequency would be 0.06.

Probability of missing an ALVE insertion in the HL and JL pooled sequencing data

The model used for the pooled datasets was largely the same as above, but the number of sampled individuals was always ten and the population sizes were changed for each dataset to reflect each line. The JL population was again 32, but the HL populations cannot be stated here due to commercial sensitivities. However, it must be noted that as the pools for WL2, WL3 and WPR1 were solely from females, these sampled populations were much larger than those from males alone.

Additionally, as the pooling methodology was to pool the DNA samples to make a single sequencing library (rather than libraries constructed for each sample, then pooled), there was a high probability that allele specific amplification in the library preparation would

remove alleles from the sequencing project. This was a particular concern as coverage in the pooled data was 11X-18X, so on average even if PCR amplification was perfectly distributed across the ten samples, two to nine of the available alleles would have been missed. As coverage is so important for allele representation in the pooled data, variability in the coverage across the genome was also modelled using a Poisson distribution. Furthermore, an underlying error rate for read mapping was defined to potentially disrupt mapping rate. This value was based on the proportion of unmapped reads for each dataset when mapped to the reference genome. Another known issue for allele detection in pooled data is variability in the 'sequencability' of the genome. However, as there is limited literature on this (Li 2011b; Li 2011a; Li 2015) and the value is largely arbitrary, for this model it was assumed that all regions of the genome could be sequenced equally well, especially as the majority of the chicken genome falls within the optimum GC % observed with Illumina sequencing (Warr et al. 2015).

Given these additional parameters, the model was run as follows: for a given frequency the flock of a given size was randomly assigned genotypes based on Hardy-Weinberg equilibrium, individuals and alleles were sampled according to a binomial distribution, and samples scaled for genome-average coverage and Poisson-varied coverage, both of which were further scaled by an underlying error rate. Models were run one million times and the probability of missing the insertion was calculated. Probabilities were calculated for each line, for each observable insertion frequency within the sample of ten birds used for sequencing, using the Poisson-coverage-scaled probabilities. Probabilities for insertions with a frequency ≤ 0.75 were negative $\log_{10}$ transformed, plotted against insertion frequency, and linear gradients were calculated and correlated with genome-average coverage.

The model was also run specifically for three cases when the ALVE insertions could be detected in the sampled birds, but were missed in the sequencing, and for the case with the lowest observed frequency identified in the identification pipeline (section 6.4.3). As the specific coverage for these sites was known, the probabilities without Poisson-scaled coverage variation were recorded.

5.2.4 Sequencing and characterisation of ALVEs identified in the Hy-Line lines

Selection of samples for sequencing

Hy-Line provided stock DNA samples to cover at least three birds per ALVE insertion per breed. Samples were chosen based on their KASP assay results and were homozygous for the insertion where possible. To achieve this, samples from an unsequenced HL line (sister line to WL4) were used to give homozygous ALVE_ros008 individuals. DNA samples were diluted to a concentration of 10ng/µl.

Amplification of ALVE inserts by PCR and purification of PCR product

The external, 'no insert' diagnostic PCR primers designed for this study were used to amplify the inserts for sequencing. PCRs were performed with the Takara PrimeSTAR® GXL DNA Polymerase kit (R050A) using the standard 30 cycle 3 step PCR protocol with 50µl reaction volumes. For primer T_m values less than 55°C, the annealing PCR step was at 55°C but otherwise at 60°C. The 68°C extension step was done for 8 minutes, following the manual's advice of extension for 1 minute per kb of expected product. Following PCR, 10µl of PCR product was run on a 1 % agar gel with Invitrogen SYBR Safe DNA gel stain, using the Bioline Hyperladder I as the marker ladder, to check for approximate band sizes.

For inserts with a PCR product greater than 1kb, the remaining 40µl of PCR product was run again on a 1 % agarose gel with wide wells. Gels were run until the end of the marker was off the gel to give good band separation, and the longest band matching the expected insert size was cut from the gel and weighed. DNA was purified from these excised bands using the Invitrogen PureLink™ Quick Gel Extraction kit (K210012), following the protocol for purification using a centrifuge. DNA was further purified by ethanol precipitation, and resuspended in 12µl tris elution buffer.

For inserts shorter than 1 kb, 10µl of the remaining PCR product was cleaned to remove primers and excess dNTPs using the Exo/SAP protocol: Exonuclease I (NEB M0293L) and shrimp alkaline phosphatase (GE Healthcare E70092Y) incubated at 37°C with the PCR product for 15 minutes, then at 80°C for a further 15 minutes.

Cloning of inserts greater than 1kb in length and extraction of plasmids

Purified insert DNA was cloned into the Invitrogen ZeroBlunt TOPO pCR®4 Blunt-TOPO® vector (45-0031) with a 24 hour incubation. 3µl of cloned product was used to transform One Shot® Mach1™-T1^{R} Competent *E. coli* cells, which were grown overnight on kanamycin selective plates. Positive clone colonies were selected and added to 100µl of lysogeny broth (LB) supplemented with ampicillin (LB-amp), and checked for successful transformation by PCR using the protocol described above. 40µl of successfully transformed LB-amp colony solution was transferred to 3ml LB-amp and incubated in the horizontal shaker at 37˚C for 48 hours.

Colony incubations were homogenised and 2ml of solution spun down to leave a bacterial pellet. Plasmids were extracted from the pellet using the Invitrogen PureLink™ Quick Plasmid Miniprep kit (K210011), following the protocol for purification of DNA with a centrifuge, including the additional ethanol wash (W10). Purified plasmid DNA was eluted into ambient temperature tris elution buffer and the concentration measured.

ALVE sequencing and characterisation

Purified ALVE DNA was amplified for sequencing using the Applied Biosystems BigDye Terminator v3.1 Cycle Sequencing kit (4337454) following the manufacturer's instructions. The initial primers used for each amplification were the external, 'no insert' primers used to produce the PCR product. These reactions were then submitted to Edinburgh Genomics (University of Edinburgh, UK) for Sanger sequencing.

Sequences produced for each insert in this first sequencing run were mapped to the flanking DNA for each insertion as well as to the ALVE1 reference sequence (GenBank: AY013303.1) using the 'Map to Reference' tool in Geneious v7.0.4 (Kearse et al. 2012). For inserts shorter than 1kb, one sequencing run from each of the original primers was sufficient to cover the whole insertion, so a consensus sequence was generated and observed SNPs or indels were annotated. For the longer inserts, the initial sequencing runs were used to observe any insert-terminal deletions (such as loss of the *envelope*-LTR section of the insert) and to check that the start of each sequencing run matched the known flanking DNA.

Sixteen further primers were designed based on the ALVE1 reference sequence to amplify the interior of the ALVE inserts with the BigDye Terminator protocol. Fourteen of these were evenly spaced every 500 bp along the ALVE1 sequence, and the other two were designed to cover each LTR in the direction of the genomic flanking DNA (Figure 5.5). Each primer sequence was designed to be 22 bases in length and have terminal GC clamps (Table 5.2). The Benkel LTR primers were not used as these are present twice in intact elements. In each case, following amplification with the BigDye protocol, samples were submitted to Edinburgh Genomics for Sanger sequencing.

All additional sequencing runs for each insert were mapped to the respective flanking DNA sequences and ALVE1 to form contiguous consensus sequence using the Geneious 'Map to Reference' tool. Consensus sequences were then used as BLASTn queries against the NCBI non-redundant database to check for identity with other ALVE sequences, and to identify the extent of any retroviral domain deletions.

LTR pairs from intact elements were aligned using MUSCLE, with default settings, to identify LTR divergence and suggest the ALVE age. All 3' LTRs, the two solo LTRs and the 3' LTRs from reference ALV-A (GenBank: M37980.1) and ALV-J (GenBank: JF951728.1) exogenous retroviruses were aligned using MUSCLE with default settings. The alignment was trimmed to remove the insertion hexamers and a tree constructed using RAxML with a generalised time reversible (GTR) gamma nucleotide model with 100 bootstraps. In addition, transcription factor binding sites were identified using the EMBOSS v6.6.0 tfscan tool (Rice et al. 2000), and the sequences scanned for the miR-155 target sequence (AGCATTA) (Hu et al. 2016) using the EMBOSS fuzznuc tool.

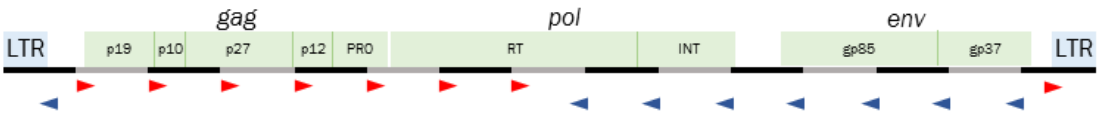


Figure 5.5 ALVE1 reference sequence domains and sequencing primer locations. The alternating black and grey backbone represents 500 bp stretches of the 7.5 kb element. The triangles show the sequencing primer locations and directions. The 5'-3' order of the primers is the same as is in Table 5.2 below. Domains: matrix (p19 and p10), capsid (p27), nucleocapsid (p12), protease (PRO), reverse transcriptase (RT), integrase (INT), surface (gp85), and transmembrane (gp37).

Table 5.2 Generic ALVE sequencing primers. Positional information based on the alignment against the ALVE1 reference sequence.

Name	Sequence	Primer location
ALV_5LTRrc	5'-ACCACTATTCCCTAACGATCAC	290 – 311
ALV_500	5'-CGACGACTGAGCAGTCCACCCC	500 – 521
ALV_1000	5'-CGTTGGCACATCCTGCTATCAG	1,000 – 1,021
ALV_1500	5'-TACAGACGGTTATAGCGGCAGC	1,500 – 1,521
ALV_2000	5'-ATCCAGCCCTTAGTTATGGCAG	2,000 – 2,021
ALV_2500	5'-CATGCGAAAATCCCGGGATATG	2,500 – 2,521
ALV_3000	5'-CAAGGATTGCTTCTTTTCTATTC	3,000 – 3,021
ALV_3500	5'-CCTTTTATGAGCAGTTACGAGG	3,500 – 3,521
ALV_4000rc	5'-CAGGGTGGTCGGTAACCCTCAC	3,979 – 4,000
ALV_4500rc	5'-GGTCTGAACAACCTCCCTAGCC	4,479 – 4,500
ALV_5000rc	5'-TCACCACGCTCAAAGTGATTGAG	4,978 – 5,000
ALV_5500rc	5'-GAGAGGCAGAAATCCGTTTGGC	5,479 – 5,500
ALV_6000rc	5'-ATGCACCGCAGTACTCACTCCC	5,978 – 6,000
ALV_6500rc	5'-ATCTGAGCATGTATCATCCAGG	6,479 – 6,500
ALV_7000rc	5'-CATCTTTCGGATGCTACTGGAC	6,983 – 7,004
ALV_3LTR	5'-GATATAGTAGTTGCGCTTTTGC	7,215 – 7,236

Geneious-hosted GLIMMER3 (Delcher et al. 2007) was used to predict retroviral ORFs with default settings. ORFs were translated into the appropriate frame and aligned to well annotated ALV protein sequences (GenBank:Q04095.2, AAK13201.1).

5.2.5 Characterisation of the *K* locus in the Hy-Line elite layer lines

Identification of the Galgal5 K locus and design of the diagnostic K locus assay

Previously published PCR primer sequences (Tixier-Boichard et al. 1994; Tixier-Boichard & Boulliou-Robic 1997; Elferink et al. 2008) and the ***K*** locus bridging

sequence (Bu et al. 2013) were mapped to the Galgal5 assembly using the 'Map to Reference' tool in Geneious, to characterise the size of the *K* locus tandem duplication. The predicted *K* locus sequence was used to design a KASP assay for the duplicated region bridging sequence, to be used in tandem with the ALVE21 KASP assay.

Optic mapping of the K locus

Blood samples were taken from one male individual from each of five Hy-Line lines, using the same blood extraction protocol as described in section 5.2.1. Samples were taken from the slow feathering WL4 and WPR1, as well as the sister line of the RIR used for the sequencing of this project (RIR-sf). Samples were also taken from the fast feathering WL3 and WPR2. Blood samples were mixed with ThermoFisher Scientific UltraPure Low Melting Point Agarose, formed into stable plugs and then analysed with the BioNano Irys platform at The Earlham Institute, UK. Samples were extracted from the agarose plugs and quality controlled before analysis. The restriction enzyme Nt.BspQ1 (recognition site: GCTCTTC) was used for mapping due to its suitable frequency of occurrence in the *K* locus target region.

Analysis of BioNano results

Molecule object files generated by the Earlham Institute were quality checked and submitted to Kees-Jan Francoijs (KJF), the European support scientist at BioNano Genomics. KJF performed an *in silico* Nt.BspQ1 digest of the Galgal5 Z chromosome using the BioNano Knickers v1.5.5.0 software, to generate a consensus map (CMAP) file with the predicted fluorescence patterns. Molecule object files were then viewed in IrysView v2.5.1 and assembled to the CMAP using the Refaligner and Assembler v5122 packages (available: https://bionanogenomics.com/support/software-downloads/). All viewing and interpretation was undertaken locally in an IrysView v2.5.1 instance.

Analyses were undertaken by KJF due to large memory requirements for performing *de novo* assemblies using IrysView (recommended >48 GB RAM limited to Windows architecture). The new BioNano Access (which can connect to a dedicated Linux server

for memory-intensive analyses) was still in beta testing when analysis was undertaken, and there were difficulties using the Access pipeline due to university regulations on SQL databases and remote Linux server access. Consequently, the analysis undertaken in this thesis has been to view the generated alignments across the ***K*** locus region, rather than running the BioNano structural variant detection pipelines. The molecule objects for each analysed line were viewed separately with the mapping confidence set at 20.

5.3 Identification of ALVEs from the DNA resequencing data of various commercial, experimental and 'wild' chicken populations

5.3.1 Pipeline implementation with single-end reads

Required adjustments to pipeline for single-end reads

For user ease, separate scripts were written for use with single-end data. The changes made within the script files were relatively minor, adapting input/output to handle a single FASTQ file rather than two, and adjusting the required region size to 80 bp compared to the 200 bp used with paired end data.

Effects on pipeline sensitivity using single-end rather than paired-end datasets

The HL and JL paired-end data were used to create pseudo-single-end FASTQ files by concatenating paired files and removing pair information from the sequence headers. This was completed with a custom Python script (pseudo_fastq.py; Appendix 1). Pseudo-FASTQ files were mapped to the Galgal5 reference genome using BWA and average genome coverage calculated with samtools mpileup.

Pseudo-single-end datasets were then analysed for ALVE insertions using the scripts adapted for single-end data analysis, and the identified locations were manually checked as described above (section 5.1.2). Identified ALVE insertion sites were compared to the data from the paired-end analysis and the relative sensitivity of the single-end pipeline was assessed.

5.3.2 Genomic resources

Chicken resequencing data was analysed from nineteen projects, enabling ALVE identification of eighty-nine datasets. Most datasets were publicly available through the ENA or DDBJ Short Read Archive, but some were kindly provided by collaborators.

Most datasets were sequenced with Illumina technology. However, the Andersson and Kauai Feral Chicken datasets were sequenced using the Applied BioSystems SOLiD platform, which produces output not directly compatible with BWA. For both datasets, the raw FASTA and quality score (QUAL) files had already been merged to create colorspace FASTQ files with Sanger quality scores. A custom Python script (color-fastq2sanger-fastq.py; Appendix 1) was used to convert these to standard 'basespace' FASTQ files ready for use in the pipeline.

Datasets were quality checked with FASTQC and reads were trimmed if base quality score dropped beneath 20, and read pairs removed if over half a read was trimmed, or if the read was mostly Ns. Trimming was completed with TrimGalore v0.4.0 (Krueger 2013) using Cutadapt v1.4 (Martin 2011). The analysed projects are detailed below.

Paired end sequencing datasets

Most datasets were paired end, Illumina sequencing data. The lines are presented in Table 5.3 and the accession numbers for the datasets are shown in Table 5.4.

Table 5.3 Paired end WGS datasets analysed for this study. The datasets and the lines they contain are listed. Each line was given a coded name which is either a shortening of the full line name or includes the breed, where BL = Brown Leghorn, Br = broiler, RIR = Rhode Island Red, RIW = Rhode Island White, RJF = red jungle fowl, WL = White Leghorn, and WPR = White Plymouth Rock. The RJF abbreviations include their countries (C = China, J = Java, S = Sumatra) and the Ethiopian village birds are abbreviated with the international code for Ethiopia (ETH). The Tibetan highland breeds are abbreviated as TIB-HL. The library type is shown for each line, showing either the number of individuals used in each pool, or the number of individual (indiv) sequencing libraries available. The final column assigns each line a group number (1-5) for the GLM analysis in section 7.4.1. Group 1 is commercial white egg layers, group 2 is commercial brown egg layers, group 3 is broilers, group 4 is generalist and 'native' breeds, and group 5 is RJFs and 'village' chickens.

Dataset	Line	Code	Library type	Group
Arkansas vitiligo model	Brown	BL-Br	Pool (10)	1
	Smyth	BL-Sm	Pool (10)	1
Cobb heritage broiler	-	Br-Cobb	Indiv (20)	3
Commercial 1995 broiler	-	Br-REL	8 x Pool (10)	3
Egg shell strength	High	RIW-ESH	Pool (8)	2
	Low	RIW-ESL	Pool (8)	2
Ethiopian village birds	Horro	ETH-Horro	Indiv (6)	5
	Jarso	ETH-Jarso	Indiv (5)	5
Fat-Lean broilers	Fat	Br-VLDL-F	Indiv (4)	3
	Lean	Br-VLDL-L	Indiv (4)	3
High and Low antibody	High	WL-HA	Pool (16)	1
	Low	WL-La	Pool (16)	1
INRA high/low fat cross	-	Br-INRA	Indiv (16)	3
Indonesian natives	Black java	Bl-java	Pool (10)	4
	Black sumatra	Bl-sum	Pool (10)	4
	Kedu Hitam	Kedu hitam	Pool (10)	4
	Sumatera	Sumatera	Pool (5)	4
	Java RJF	RJF-J	Pool (2)	5
	Sumatra RJF	RJF-S	Pool (3)	5
	WL	WL-NU	Indiv (1)	1
Iowa State	Fayoumi	Fayoumi	Pool (16)	4
	Leghorn	WL-IS	Pool (16)	1
Korean domestic	Araucana	Araucana	Indiv (3)	4
	Korean	Korean	Indiv (3)	4
	WL	WL-K	Indiv (3)	1
Lohmann layers	RIR	RIR-L	Indiv (25)	2
	WL	WL-L	Indiv (25)	1
	WL	WL-Lp	Pool (10)	1

	WPR	WPR-L	Pool (10)	2
Pirbright inbred lines	15	WL-PB-15	Pool (10)	1
	6	WL-PB-6	Pool (10)	1
	7	WL-PB-7	Pool (10)	1
	C	WL-PB-C	Pool (10)	1
	N	WL-PB-N	Pool (10)	1
	P	WL-PB-P	Pool (10)	1
	Wellcome	WL-PB-W	Pool (10)	1
	Zero	WL-PB-Z	Pool (10)	1
Roslin experimental blind	BEG blind	BL-BEGb	Pool (10)	1
	BEG sighted	BL-BEGs	Pool (10)	1
	RGE blind	BL-RGEbm	Indiv (1)	1
	RGE blind	BL-RGEbp	Pool (10)	1
	RGE sighted	BL-RGEsf	Indiv (1)	1
	RGE sighted	BL-RGEsp	Pool (10)	1
SPF commercials	A	WL-SPFa	Pool (14)	1
	B	WL-SPFb	Pool (11)	1
Taiwanese domestic	Silkie	Silkie	Indiv (1)	4
	Taiwan Country	Taiwan	Indiv (1)	4
Tibetan highland/lowland	Chahua	Chahua	Indiv (1)	4
	Highland1	TIB-HL1	Indiv (1)	4
	Highland2	TIB-HL2	Indiv (1)	4
	Highland3	TIB-HL3	Indiv (1)	4
	Lhasa White	Lhasa white	Indiv (1)	1
	Lindian	Lindian	Indiv (1)	3
	WL	WL-B-D	Indiv (1)	1
	WL	WL-B-E	Indiv (1)	1
Tibetan fighting	RJF	RJF-C	Indiv (6)	5
	Xishuangbanna	Xishuang	Indiv (8)	4

Table 5.4 Accession numbers and references for the paired end WGS data. All accession numbers are for the European Nucleotide Archive (ENA) except the Indonesian natives stored in the DNA Databank of Japan (DDBJ). Some datasets were not publicly available but were kindly shared by collaborators. Collaborators have been named with their affiliation.

Dataset	Accession No.	Reference
Arkansas vitiligo model	PRJNA256208	Jang et al. 2014
Cobb heritage broiler	PRJEB15276 (non-public)	Khoo et al. in prep.
Commercial 1995 broiler	Douglas Rhoads, University of Arkansas	Pavlidis et al. 2007
Egg shell strength	PRJNA231017	Zhang et al. 2015
Ethiopian village birds	Olivier Hanotte, University of Nottingham	Wragg et al. 2015
Fat-Lean broilers	PRJEB15288 (non-public)	Griffin et al. 1991; Khoo et al. in prep.
High and Low antibody	Chris Ashwell, North Carolina State	Kuehn et al. 2006
INRA high/low fat cross	PRJNA247952	-
Indonesian natives	DDBJ DRA003951	Ulfah et al. 2016
Iowa State	Susan Lamont, Iowa State	-
Korean domestic	PRJNA291174	Oh et al. 2016
Lohmann layers	Rudolf Preisinger, Lohmann	Pooled data used in Kranis et al. 2013
Pirbright inbred lines	Roslin Institute	Kranis et al. 2013
Roslin experimental blind	Roslin Institute	Hocking & Guggenheim 2014
SPF commercials	Marc Eloit, Pasteur Institute	Gagnieur et al. 2014
Taiwanese domestic	PRJNA202483	Fan et al. 2013
Tibetan highland/lowland	PRJNA309581	Zhang et al. 2016
Tibetan fighting	PRJNA241474	Guo et al. 2016

Single end sequencing datasets

Two datasets were used which had single end datasets, both of which were sequenced with the Applied Biosystems SOLiD platform. The first dataset was a collection of lines sequenced to identify signals of domestication. This comprised pooled sequencing data for eight red jungle fowl, ten each from two commercial broilers, eight RIR, eleven and eight from two commercial WL lines, eleven each from two WPR sister lines bred for differential growth rates (high and low), and ten from an obese WL experimental line (ENA SRP001870) (Cole 1966; Dunnington & Siegel 1996; Rubin et al. 2010). The second dataset was for feral chickens on the Hawaiian island of Kauai (Gering et al. 2015). This included individual sequence data from twenty-three feral chickens, as well as one laboratory RJF (ENA PRJNA272379).

5.3.3 ALVE identification

Each dataset was analysed in turn using either the paired-end or single-end ALVE identification pipeline, as appropriate. Putative ALVE insertion regions were checked manually as described above (section 5.1.2) and the insertion hexamer noted. With the release of the Galgal5 feature annotation from Ensembl (v87), insertion locations were not also identified in Galgal4 as with the HL and JL data. Instead, insert coordinates were compared directly to the Galgal5 annotation using BEDTools closestBED and intersectBED. Insertion sites were compared with the HL and JL data, as well as other known ALVE locations, to assign existing nomenclature where appropriate. Novel insertions, and those lacking clear ALVE-based names, were assigned new names based on the ALVE_ros001 nomenclature developed above.

Due to the short, 35bp reads of the Andersson dataset, mapping was completed with BWA-aln followed by BWA-samse in addition to the standard BWA-mem protocol, as BWA-aln is optimised for short read mapping (Li & Durbin 2009; Li 2013). The results were compared between these two mapping approaches.

Characterisation of insertion locations

The insertion hexamers of the identified ALVEs were checked for sequence over-representation and GC content distribution. These observed values were compared to a model which simulated equal numbers of hexamers randomly distributed across the genome, repeated one million times. The GC content distributions were compared using a two-sample t-test and plotted using MATLAB.

Insertion bias due to GC content was also assessed by calculating the GC content in the insertion regions in windows of 100 bp, 1 kb, 10 kb and 100 kb. Values were compared to the genome average as well as to the chromosomal GC content, and patterns in GC deviance were assessed. In addition, the number of insertions identified on each chromosome was $\log_{10}$ transformed for normality and correlated with the $\log_{10}$ transformation of chromosome length.

5.3.4 Cluster analysis based on ALVE content

Presence/absence data for each line for all identified ALVEs was used to create a 1/0 matrix for clustering the lines based on their ALVE content. Euclidean distances were calculated for the matrix using the MATLAB pdist function. These distances were used to form a hierarchical binary cluster tree with the MATLAB linkage function using average distances, and the complete tree plotted using the dendrogram function.

Analysed datasets differed on library type (individuals or pools) and genome coverage. A general linear model (GLM) was used to identify whether these factors significantly influenced ALVE identification compared to the observed chicken breed categories. Lines were grouped into five categories: white egg layers, brown egg layers, broilers, native breeds, and 'wild' (including RJF samples). The GLM was fitted with identified ALVE number as the response variable, line category and library type as categorical variables, and average genome coverage as a covariate. For individual libraries, genome coverage was calculated per dataset and then averaged between datasets.

To ensure the generated dendrogram was due to shared ALVEs, rather than similar numbers of identified ALVEs in disparate lines forcing relatedness, a model was generated to randomly redistribute the presence/absence data. The model data set was

limited to lines with twelve ALVEs or fewer to ensure the large numbers of lineage-specific ALVEs observed in some datasets did not bias the dendrogram construction. For each line, the number of identified ALVEs was kept the same, but the modelled ALVEs were randomly assigned. One hundred randomly generated dendrograms were created, manually inspected, and compared to a dendrogram constructed using the observed presence/absence data for the same lines.

Principal Coordinate Analysis

In addition to the hierarchical dendrogram construction, the lines were also clustered using classical multidimensional scaling, also known as Principal Coordinate Analysis (PCoA). Traditional Principal Component Analysis (PCA) was not appropriate as the presence/absence data was mainly zeros. The PCoA was completed on the Euclidean distance matrix calculated above using the MATLAB cmdscale function, which also calculated the eigenvalues. Eigenvalues were plotted using a scree plot and the contribution of each was compared. The MATLAB plot function was used with various pairs and trios of the eigenvalues to identify clusters between the analysed lines.

Chapter 6: The discovery and characterisation of Avian Leukosis Virus subgroup E (ALVE) insertions using whole genome (re)sequencing (WGS) data

6.1 Introduction

Avian Leukosis Virus subgroup E (ALVE) insertions are the youngest chicken endogenous retroviruses (ERVs), and are diverse across various chicken lines (Benkel 1998; Weiss 2006). Whilst their genomic copy number is often low, as contemporary insertions many ALVEs retain high structural integrity.

Complete, transcriptionally active ALVEs produce replication-competent viral particles which can be shed, facilitating horizontal transmission of the virus within and between flocks. For the hosts, viral particles induce viremia which causes persistent immunological and physiological stress. This has been shown to significantly delay and reduce antibody production to exogenous Avian Leukosis Virus (ALV) infection (Gavora et al. 1995), as well as detrimentally affect commercial traits in both broilers (*e.g.* early growth rate, total body weight) and layers (*e.g.* egg weight and density, total egg count) (Fox & Smyth 1985; Kuhnlein et al. 1989; Gavora et al. 1991; Ka et al. 2009). Despite this, multiple intact and transcriptionally active ALVEs remain in commercial lines due to their association with desirable phenotypic traits. These include ALVE-TYR, responsible for white feather colour through the recessive white mutation (Fox & Smyth 1985; Chang et al. 2006; Chang et al. 2007), and ALVE21, which is closely associated with the slow feathering ***K*** locus used to differentiate bird gender at hatch (Bacon et al. 1988; Iraqi & Smith 1995; Elferink et al. 2008). In addition, many lines harbour intact ALVEs, such as ALVE1, which have been transcriptionally silenced by DNA methylation, but can reactivate under certain infection conditions or during embryonic development (Conklin 1991; Hu et al. 2016).

Not all ALVEs are structurally intact, but the expression of individual retroviral domains can still elicit a significant effect on the host. Production of gag glycoproteins (*e.g.* ALVE3) has been shown to induce tolerance to novel ALV infections, resulting in a delayed immune response and higher incidence of lymphoid tumours (Astrin & Robinson 1979; Crittenden et al. 1984). Contrastingly, expressed envelope glycoproteins (*e.g.* ALVE3, ALVE6, ALVE9) confer resistance to novel ALV infection through

competitive interference, as the envelope proteins physically block the TVB (tumour virus binding) receptors used by ALVEs to enter host cells (Robinson et al. 1981; Yu et al. 2008). These opposed effects confer complex infection dynamics in hosts with multiple ALVE insertions. For example, ALVE21 positive birds which also contain ALVE6 or ALVE9 have reduced viremia, higher ALV-antibody titres, and reduced viral shedding when compared to birds with ALVE21 alone (Smith et al. 1990a; Gavora et al. 1995). This cross-reactivity can lead to unexpected results when lines with different ALVE complements are crossed.

These detrimental, and often complex, ALVE-related effects have led many commercial companies to select against these loci, aided by the development of a range of diagnostic PCR assays and an ELISA for p27, a *gag* component found in all ALV subgroups (Smith et al. 1979; Benkel 1998; Chang et al. 2006; Smith & Benkel 2009a; Rutherford et al. 2013). However, changing focus in selection programmes, the expense of testing all known ALVEs for entire commercial flocks using gel based PCRs, the close association between many ALVEs and desirable traits (particularly ALVE21), the incomplete identification of all ALVEs within individual lines, and the high frequency of many ALVEs, means that ALVEs have not been eradicated in commercial lines. This enables further ALVE retrotransposition events, which could produce novel mutagenic effects, and recombination between ALVEs and exogenous ALV.

The current detrimental impact of ALVEs in commercial chicken production goes beyond the described effects on productivity. For example, all commercial flocks are tested for exogenous ALV using the p27 ELISA, but ALVEs expressing *gag* generate false positive results and these birds must be culled. Recent research into the increasingly virulent Marek's Disease Virus (MDV; alpha herpesvirus) has shown that whilst the ALVE background does not influence MDV incidence or survival, MDV itself induces ALVE expression (even of normally silenced elements such as ALVE1), provoking viral shedding (Chang et al. 2015; Hu et al. 2016; Hu et al. 2016a). There was also found to be a significantly higher incidence of spontaneous lymphoid tumours in lines containing ALVE21 following MDV vaccination (Fadly et al. 2014; Cao et al. 2015). This extra context of biologically representative co-infection is important as individual infections or traits are commonly only considered in isolation.

The range of current and future concerns makes ALVE identification and characterisation within commercial lines essential for improvements in both productivity and animal welfare. This information can then be used to design new, high throughput diagnostic assays for flock genotyping to advise better selective breeding programmes, or even identify targets for CRISPR/Cas9 deletions.

Existing methods for identification of ALVE insertions

Traditionally, ALVEs were identified almost exclusively in White Leghorn (WL) chickens by performing full genomic DNA digests with several restriction enzymes, followed by fragment hybridisation to ^{32}P-labelled RNA from Rous Sarcoma virus (RSV) (Astrin 1978). This generated band size patterns specific to known ALVEs. As WLs generally have few ALVEs (Benkel 1998), *in situ* hybridisation with RSV RNA probes often yielded general genomic locations. In addition, each analysed line had known phenotypes for ALVE *gag* and *envelope* expression which could then be assigned directly to individual ALVE insertions (Astrin & Robinson 1979). These methods were laborious, but did enable the identification of twenty-three ALVEs in WLs and a further twenty to twenty-five once the methodology was expanded to other chicken breeds including commercial broilers (Sabour et al. 1992; Benkel 1998; Hunt et al. 2008).

The specificity of this detection method improved with the availability of a chicken reference genome, as primers could be designed to amplify the restriction enzyme digest fragments for sequencing. This enabled identification of exact genomic coordinates for the insertion (Smith & Benkel 2009). Further improvements were predicted with the cost-effective availability of short read whole genome (re)sequencing (WGS) data. As hypothesised, this data facilitated identification of huge numbers of single nucleotide polymorphisms (SNPs) across various chicken lines (Kranis et al. 2013). Larger structural variants (SVs), particularly transposable element copy number variants (CNVs), could not be detected so easily, usually requiring the identification of incongruent read mapping events relative to the reference genome (Alkan et al. 2011). However, incongruent read mapping is relatively common, especially when assemblies are incomplete and repetitive sections of the genome remain poorly resolved.

Unfortunately, *de novo* assemblies from short read technology are not the solution, as they are limited by their inability to assemble through repeated regions greater than 1 kb. Repetitive element insertions generally 'collapse', with all element-homologous reads mapping to a single assembled region. Long read sequencing technology, such as PacBio, now generates high throughput data with reads long enough to sequence through most repeat classes. Whilst these technologies are now becoming more commonplace, most sequencing projects have been completed with short read technology, and, consequently, have not been mined fully for their structural variants.

Discovery of novel chicken ALVEs from WGS data faces these same issues. The Galgal5 reference genome contains two ALVEs: ALVE-RJF on chromosome 1 and the partially assembled ALVE6 (Benkel & Rutherford 2014). When WGS data is mapped to the reference genome, all ALVE-homologous sequences map to the assembled ALVE-RJF. Only reads that map to the very edge of ALVE-RJF, or who's read pair maps to 'true' chicken genomic DNA, provide novel insertion positions. However, detection of these insertion junctions is heavily dependent on genome coverage.

In an effort to improve coverage of insert junction sites, protocols have been developed which echo the hybridisation strategies of traditional repeat identification. These 'target capture' sequencing projects use repeat sequences to form a 'bait panel' for sheared genomic DNA. Bound DNA fragments are then sequenced to high coverage and the resultant data mapped to the reference genome. This has been used successfully with other retrotransposon classes such as human L1 elements (Baillie et al. 2011), but during the course of this project a methodology was also published specifically for ALVE target capture sequencing (Rutherford et al. 2016).

As an alternative to WGS, target capture sequencing would reduce overall costs. However, these data are very specific and unnecessary if identified ALVEs were previously known or at a high frequency in a population, particularly as sequencing cost is dropping all the time. One of the major advantages of WGS data is the ability to apply a single dataset to multiple research questions. In addition, whilst identifying structural variants from existing WGS datasets is more difficult than identifying SNPs, it is possible, as demonstrated below. This means existing WGS data can be used for ALVE identification, rather than requiring additional sequencing.

6.2 Research Aims

This chapter covers four major research aims. Firstly, the development of a new bioinformatics pipeline for the identification of ALVEs within existing WGS datasets, without requiring further, targeted sequencing. Secondly, the validation of the findings of this pipeline by the development of diagnostic assays specific to each ALVE insertion which were subsequently tested on the original lines. Thirdly, the annotation of each ALVE insertion by assigning existing or new nomenclature, describing the insert location relative to annotated genes, and characterising the intactness of the insert sequence. Finally, the further characterisation of the commercially important slow feathering ***K*** locus using high resolution optic mapping.

6.3 Development of the ALVE identification pipeline

6.3.1 Initial approaches

With the emergence of more cost effective WGS datasets, multiple bioinformatics tools have been developed in recent years to identify viral insertions, particularly in the field of human cancer genomics. These tools, such as Virana (Schelhorn et al. 2013), VirusSeq (Chen et al. 2013), VirusFinder (Q. Wang et al. 2013; Wang et al. 2015) and ViralFusionSeq (Li et al. 2013), create a viral pseudochromosome which is appended to the reference genome enabling bridging of read pairs between genomic and viral DNA.

Initially, identification of ALVEs was completed with custom scripting following this approach. Both VirusFinder and VirusSeq were trialled, but the ability to directly tailor a pipeline to discovery of retroviral insertions (rather than any viral genera) and reduce the computational effort meant that the time creating a new identification pipeline was well spent. However, it was quickly observed that there were limited mapping events between the genomic DNA and viral pseudochromosome. Additionally, any mapping was well within the viral sequence, rather than at either terminus. Whilst this enabled detection of an insertion within an approximately 500 bp region, it did not enable exact insertion site identification.

Another approach was to first map all reads to the reference genome, then extract any incongruently or unmapped reads and map these to the pseudochromosome. Any read

pairs with one read mapping to the pseudochromosome would then be mapped back against the reference genome, enabling identification of the insertion site. Whilst this methodology worked, it was computationally expensive as many low quality or non-ALVE reads were carried through to the pseudochromosome mapping. The approach was changed to begin with the pseudochromosome mapping, removing the most computationally expensive step without any impact on detection sensitivity. This approach is described below.

6.3.2 Pipeline implementation

This ALVE identification approach was wrapped into seven scripts, written to be executed in Linux using either standard BASH tools or Python 2.7 (Table 6.1). The overall pipeline is shown in Figure 6.1. Following identification of putative ALVE sites, the BAM files can be viewed at the command line or in a much more visual manner using a genome viewer such as IGV or Golden Helix. All scripts are on the CD accompanying this thesis (Appendix 1) and in the GitHub ALVE identification repository (https://github.com/andrewstephenmason/ALVE_ID_pipeline).

Table 6.1 ALVE identification pipeline scripts and functionality.

Script name	Functionality
S1_run_blast_ref_seq.sh	Identifies known alpharetroviral sites
S2_make_pseudochromosome.py	Creates the pseudochromosome
S3_run_bwa_alignment.sh	Maps sequencing reads to the pseudochromosome
S4_extract_ref_seq_mapped_reads.sh	Extracts viral mapped reads and pairs
S5_run_bwa_alignment.sh	Maps sequencing reads to the reference genome
S6_extract_putative_sites.py	Identifies the putative insertion sites
S7_merge_lists_and_reduce_ref_genome.sh	Merges identified sites between sequencing files and prepares a reduced reference genome

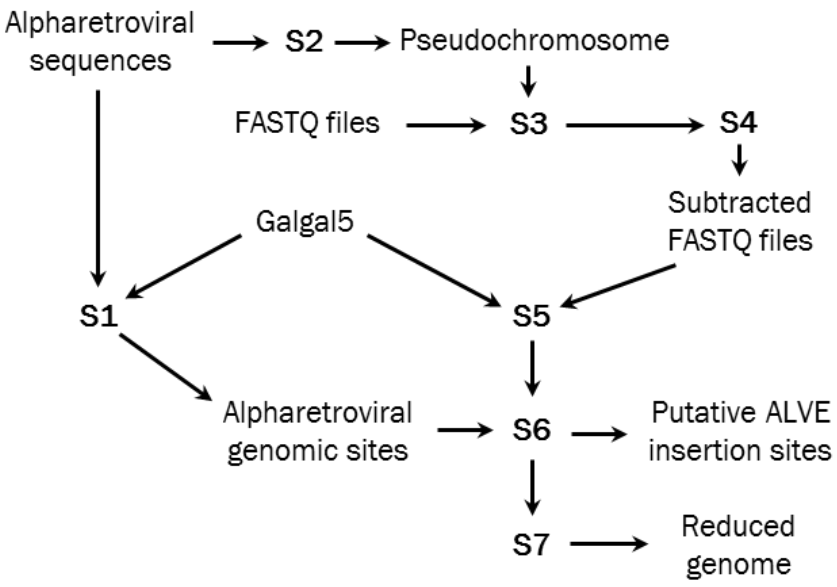


Figure 6.1 ALVE identification pipeline workflow. Script names have been abbreviated to their order as in Table 6.1 and input/output files are shown.

Each script file manages the naming prefixes of all linked files, enabling the processes to be completed in a largely parallel manner. However, data analysis is limited to the number of processors and amount of writable memory available. As BAM files are very large, the most memory intensive steps were always when these files had to be sorted, such as after a mapping, hence the setting of sorting threads and memory in the two mapping scripts (S3 and S5). Additionally, the samtools sort program creates huge temporary files during sorting and thus parallel runs were usually limited to eight datasets at a time, where the total compressed FASTQ size for the dataset was 20-30 GB, corresponding to 200-300 million paired end reads.

Once the reference alpharetroviral sites have been identified (S1_run_blast_ref_seq.sh) and the pseudochromosome created and indexed (S2_make_pseudochromosome.py), these two processes do not have to be completed again. The slowest step is always the mapping of the full FASTQ files to the pseudochromosome (S3_run_bwa_alignment.sh; 12-24 hours). Read subtraction usually completes in less than two hours (S4_extract_ref_seq_mapped_reads.sh) and the mapping of the reduced dataset is often minutes (S5_run_bwa_alignment.sh). Similarly, the putative site identification (S6_extract_putative_sites.py) and reference genome processing for

viewing (S7_merge_lists_and_reduce_ref_genome.sh) take only minutes. Depending on the number of identified sites, and whether sites match known ALVEs, the manual validation of putative insertion sites can take minutes or hours. Altogether, this means that once the FASTQ files are processed and ready for analysis, a novel dataset can have its ALVE content characterised within 24 hours.

6.3.3 Unwanted issues with sequence homology

Undesirable identification of other alpharetroviral classes

Preliminary mapping work showed that ALVE sequences retain enough homology to map to other alpharetroviral classes such as EAVs and ART-CH sequences. As a result, these sequences were also identified in the Gal5 reference genome and used to filter out putative sites that were due to spurious mapping.

However, evJ and ALV-J reference sequences were initially retained in the alpharetroviral pseudochromosome used for the first mapping process, and the *envelope* gene of these sequences is derived from EAV-HP (Benson et al. 1998; Smith et al. 1999). This enabled the mapping of EAV homologous reads to the pseudochromosome which were then mapping to EAV sites in the analysed lines. EAVs are more numerous than ALVE insertions, so initially made up over 80% of putative insertion regions. These sites were identified as EAVs by BLASTn against the NCBI non-redundant database, and the causative evJ and ALV-J sequences were removed from the pseudochromosome, preventing unwanted mapping events.

Dubious mapping on chromosomes 20 and 23

Initial mapping of the Hy-Line data routinely identified several putative insertion sites within an 8 kb region on chromosome 20 (20: 5,221,295-5,229,299) and two regions on chromosome 23 of 7.5 kb (23: 2,053,213-2,060,796) and 4.2 kb respectively (23: 5,152,304-5,156,520). Manual inspection of these sites suggested no ALVE homology, at least to ALVE LTRs, but the regions were of real interest as all the putative sites were intronic, but very close to the gene exons.

In each of the three cases the intronic sites were within genes from the proto-oncogene tyrosine protein kinase *SRC* family, including the *sarcoma* (*c-SRC*) gene itself at the chromosome 20 site. This gene was transduced by, and mutated in, the proto-ALV, Rous Sarcoma Virus (RSV), to become *v-SRC* (Swanstrom et al. 1983). In these first mappings, the pseudochromosome included the RSV reference sequence which enabled read mapping to *c-SRC* causing the false matches. In a similar manner, the closely related genes of *YRK* and *LCK*, sited at the respective chromosome 23 regions, also facilitated read mapping.

RSV was removed from the pseudochromosome and the regions added to the filtering list during putative region identification.

Hybrid ALVE-genome reference sequences

A final identification issue was that the reference sequences for ALVE1 and ALVE21 also included the insertion hexamer sequence and flanking genomic DNA. As a result, when these sequences were used to identify existing alpharetroviral sites in the reference genome they also masked flanking genomic DNA. Consequently, ALVE1, known to be in at least some of the WL HL lines, and ALVE21, known to be in the two slow feathered HL lines, were not identified at any point in initial mapping.

The initial response was to remove these sequences from the reference sequence identification step. Whilst this enabled ALVE1 identification in the 'correct' lines, it also incorrectly caused ALVE21 identification in every line. This was simply because these two reference sequences had not been removed from the pseudochromosome, so non-viral genomic reads could map, be retained, and support a putative insertion. These two sequences were edited to remove the flanking DNA using the GenBank annotations, and all other reference sequences were checked in a similar manner.

6.3.4 Thresholds for suitable insertion site support

Preliminary identification analyses required putative ALVE insertion sites to have at least 360 bp of supporting flanking sequence, as well as evidence of both 5' and 3' soft-clipped

reads supporting the insert. However, use of these stringent thresholds meant that known ALVEs in some of the HL lines were missed during these analyses.

360 bp was chosen to ensure good read coverage supporting the site, but did not consider the reduction of coverage due to removal of genomic DNA reads or the heterogeneity of coverage across such sites. Putative site regions were defined by reads overlapping when mapped to the reference genome, but if two nearby reads were separated by more than 12 bp they would not represent one contiguous region, potentially reducing the extent of the defined region furthest from the insertion site. This threshold was reduced to 200 bp to reflect a scenario where two reads were end-to-end at the insertion site. In partnership with the requirement that putative regions were supported by at least three reads, this kept mapping noise to a minimum without reducing sensitivity. Additionally, many initially identified sites with poor support were supported by multi-mapped reads with mapping qualities of less than 20. Filtering was used to remove these low quality mappings and false positive noise was further reduced.

The requirement for representation of both 5' and 3' soft-clipped reads at an insertion site does give high confidence to a putative ALVE insertion. However, this requires high enough coverage of an insertion so that reads can be identified that map to the genome-ALVE junction from both ends, which in turn requires significant enough mapping to both the pseudochromosome and reference genome with the read pair mapping in the most optimal locations. When read coverage is insufficiently high, or insertion alleles are poorly represented in the pool, it is unlikely that 5' and 3' support will be achieved. Additionally, more complicated insertions which are associated with genome deletions or translocations would not necessarily have both 5' and 3' soft-clipped reads for the one insertion in the same genomic location. The filtering was adjusted to require identification of either 5' or 3' soft-clipped support of a junction, but sites with both directions of support were highlighted during detection.

Automated ALVE insertion detection remains imperfect. Manual viewing and confirmation of sites is still required to confirm an ALVE insertion and to remove those sites with limited or no obvious support. Regions of apparent noise are still identified in some analyses, potentially due to reads mapping to highly conserved alpharetroviral regions such as the *pol* gene for novel EAV or ART-CH insertions. However, manual

confirmation of sites is generally very quick, especially when sites are shared between lines, and viewing is still required for identification of the insertion hexamer as well as the extraction of soft-clipped sequence used in diagnostic assay design.

6.4 The ALVEs of the Hy-Line elite layer lines

Twenty ALVEs were identified across the eight Hy-Line lines (Table 6.2). The WLs consistently have the fewest ALVEs (2-4), followed by the two WPR sister lines (5-7) and the RIR with 10 (Table 6.3).

The WLs have the literature predicted ALVEs of ALVE1 (all 5 lines), ALVE3 (3 of the 5), ALVE9 (1 of the 5) and ALVE15 (3 of the 5). In addition, WL4 has ALVE21 as predicted, and ALVE_ros008, an ALVE previously identified in other lines by the Benkel group (BK-59). Furthermore, the identification pipeline identified another putative ALVE insertion site region in WL4 (1: 66,120,264 - 66,121,075), but there were only five soft-clipped reads in the region and there was no significant ALVE homology, so this site was discarded.

The WPR sister lines share five of their ALVEs, including ALVE-TYR, the causative factor in recessive white (Chang et al. 2006), hence the white plumage of the WPR breed. Three other shared ALVEs were all originally identified within heritage breeds at the Nova Scotia Agricultural College (Smith & Benkel 2009a; Rutherford et al. 2013): ALVE-NSAC1, ALVE-NSAC3 and ALVE-NSAC7. Unexpectedly, WPR2 also has AVE21, even though it has the fast feathered phenotype. As well as the five ALVEs shared between the WPRs, WPR2 also had ALVEB5 and the novel ALVE_ros009.

The RIR had the most ALVEs and shared only ALVEB5 and ALVE-NSAC1 with the WPR lines. Four other sites had previously been described by the Benkel group: ALVE_ros001 (COTW55), ALVE_ros002 (COTW69), ALVE_ros005 (New11), and ALVE_ros006 (N4). The remaining four identified ALVEs were novel intergenic sites: ALVE_ros003, ALVE_ros004, ALVE_ros007 and ALVE_ros010.

Table 6.2 The twenty ALVEs identified across the eight Hy-Line elite layer lines, with their Galgal5 location, insertion hexamer and overlapped feature. The six Benkel-defined ALVEs which lacked clear ALVE nomenclature are also shown.

Name	Old name	Location	Hexamer	Feature
ALVEB5		1: 10,637,460	GGTGGT	
ALVE1		1: 65,993,542	ACGGTT	SOX5 intron 1
ALVE-ros001	COTW55	1: 101,668,931	GTTGTG	
ALVE-ros002	COTW69	1: 158,775,708	ATAAGT	
ALVE-ros003	SGT-24	1: 163,248,553	CCTACT	
ALVE-TYR		1: 187,921,213	ACACTG	TYR intron 4
ALVE-NSAC1		2: 120,868,843	CCTGTT	
ALVE-ros004		2: 124,432,997	CTTGAC	
ALVE-ros005	New11	2: 142,480,536	TTGATA	
ALVE-NSAC3		3: 53,639,776	ATAAAA	
ALVE-ros006	N4	3: 57,337,987	GGACTC	
ALVE15		3: 70,384,294	GTTTAT	GRIK2 intron 16
ALVE-ros007		4: 59,843,015	AATAGA	
ALVE-ros008	BK-59	4: 62,680,158	CTGTAG	
ALVE-ros009		4: 71,095,932	GTCCAG	
ALVE9		6: 33,153,441	CTCAAA	DOCK1 intron 35
ALVE-NSAC7		9: 11,714,130	CTTCTC	
ALVE-ros010		9: 11,871,576	TCGGAT	
ALVE3		20: 10,309,347	AACCAC	HCK intron 6
ALVE21		Z: 10,681,671	GGGTAG	

The initial characterisation of ALVE_ros007 was confusing as the pipeline identified it as two independent sites 1,939 bp apart. The 5' site (4: 59,843,021) had only 3' soft-clipped reads, and the 3' site (4: 59,844,960) had only 5' soft-clipped reads. Additionally, in the filtered BAM file there were no mapped reads in the space between these two sites, but the full RIR BAM file had mapping throughout the region (Figure 6.2). This

suggested that there had been both an ALVE insertion and a genomic deletion at the site, with the two occurrences linked in one genotype. Further evidence for this came from BLASTn searches against the NCBI non-redundant database with the soft-clipped read sequence. The 3' soft-clipped reads at the 5' site were homologous to the 3' end of the ALVE LTR, suggesting an insertion in the negative orientation. The 5' soft-clipped

Table 6.3 The ALVEs identified in each of the Hy-Line elite layer lines using the ALVE identification pipeline.

Name	WL1	WL2	WL3	WL4	WL5	WPR1	WPR2	RIR
ALVEB5							✓	✓
ALVE1	✓	✓	✓	✓	✓			
ALVE-ros001								✓
ALVE-ros002								✓
ALVE-ros003								✓
ALVE-TYR						✓	✓	
ALVE-NSAC1						✓	✓	✓
ALVE-ros004								✓
ALVE-ros005								✓
ALVE-NSAC3						✓	✓	
ALVE-ros006								✓
ALVE15	✓	✓			✓			
ALVE-ros007								✓
ALVE-ros008				✓				
ALVE-ros009							✓	
ALVE9			✓					
ALVE-NSAC7						✓	✓	
ALVE-ros010								✓
ALVE3		✓	✓	✓				
ALVE21				✓		✓	✓	

reads at the 3' site, however, were homologous to the ALVE *envelope* domain, suggesting a deletion of over 6 kb of the ALVE insertion as well as almost 2 kb of the genomic DNA (section 6.6.1).

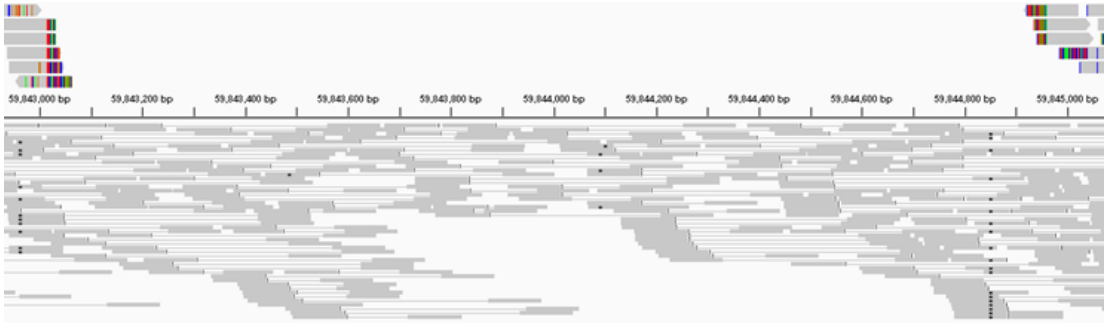


Figure 6.2 The ALVE_ros007 genomic region showing the ALVE homologous reads and the full RIR BAM file. Above the chromosome 4 scale bar are the ALVE filtered reads. No reads mapped between the two clipped sites when filtered, but the full region (below the scale, collapsed showing lines between pairs) has mapping throughout.

Genomic distribution of the identified ALVEs

Five of the twenty identified ALVEs are intronic (Table 6.2), and these were all previously well described and characterised. ALVE1 is within the first intron of SOX5 (SRY-box 5 transcription factor; 15 exon gene), ALVE3 is within the sixth intron of HCK (tyrosine-protein kinase; 11 exon gene), ALVE9 is within the thirty-fifth intron of DOCK1 (dedicator of cytokinesis 1; 52 exon gene), ALVE15 is within the final intron of GRIK2 (glutamate ionotropic receptor kainite type subunit 2; 16 exon gene), and ALVE-TYR is within the final intron of TYR (tyrosinase; 5 exon gene).

There is no correlation between the ALVE insertion site and the observed GC content of that region. Additionally, there are no apparent correlations or patterns in the insertion hexamer sequences. However, with the relatively few ALVEs identified, it is unlikely that any true patterns could be observed.

6.4.1 KASP development

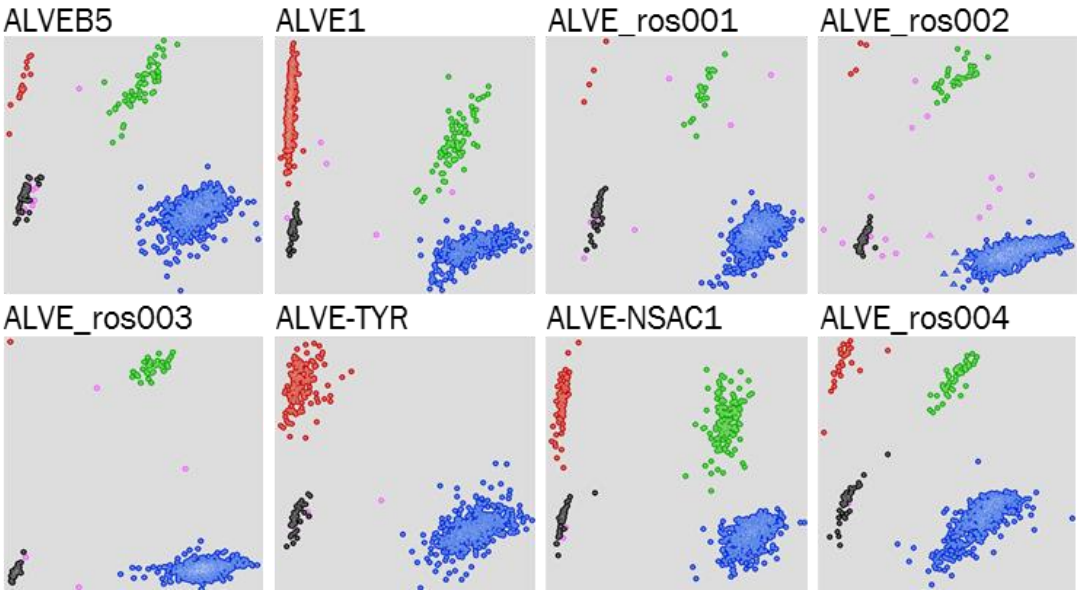
KASP assays were successfully developed for all twenty identified ALVE insertions from the HL lines. The four primers for each assay are presented in Table 6.4, and assay results for the 2010 generation males are shown in Figure 6.3.

Table 6.4 Primers used for the diagnostic Hy-Line-based KASP assays. There are four primers for each assay. The first pair is for the 'no insert' genotype and the second pair is for the insert. Primers 1 and 3 for each assay have the fluorescent tags (sequences not shown). Base ambiguities are shown using the IUPAC codes.

ALVE	KASP primers
ALVEB5	5'-AATAAACAATTCTCAGCTTAACCACCC
	5'-GAATGTTTAAGTCYGTGGATCTAATGA
	5'-AAAATAAACAATTCTCAGCTTAACCACCT
	5'-TTGCGAACACCTRAATGAAGCAGAA
ALVE1	5'-AAATTGACTTTAATATCTGTCAACGGTTC
	5'-CCAAGCAAGTAGCCCAAAACACAGTA
	5'-AAATTGACTTTAATATCTGTCAACGGTTG
	5'-GAACACCTGAATGAAGCAGAAGGCTT
ALVE-ros001	5'-TCATACACATGTTGTGYTCCCTG
	5'-ATGTCCTGACTATGATCTGGCAGCT
	5'-TTCATACACATGTTGTGTGAAGCCTT
	5'-TAACGATTGCGAACACCTGAATG
ALVE-ros002	5'-TTCTGTGATTTAGTGATTCTATGATAAGTC
	5'-CTACAATTCTGTGATTCTGTGATTCTG
	5'-ATTCTGTGATTTAGTGATTCTATGATAAGTG
	5'-CAAGTTGCCTCTGGCTCTATTTGACTA
ALVE-ros003	5'-TAGATCCTGATATCTTCATCCCTATCA
	5'-ATTTATACAAATGTATGTGGTGAGAATGAT
	5'-TAGATCCTGATATCTTCATCCCTATCT
	5'-GAGTTGCCTCTGGCTCTATTTGACTA
ALVE-TYR	5'-TTTCACTCTGAGCCTTCCAGTGTTA
	5'-TGCAGTACCAGTGTATACAGATTGTGTAA
	5'-TTCACTCTGAGCCTTCCAGTGTTG
	5'-GAACACCTGAATGAAGCTGAAGGCTT

ALVE-NSAC1	5'-GTAAGCCTGGAGATGTCCTGTTC
	5'-CAATACACAAGACTGAAAGCAGTCCAT
	5'-GTAAGCCTGGAGATGTCCTGTTT
	5'-GAACACCTGAATGAAGCTGAAGGCTT
ALVE-ros004	5'-CTCTTGCTTAATGTTTTGTTATCTTGACC
	5'-TGGCAAATCGTTTCTGAGTCCAATTAGAT
	5'-TCTCTTGCTTAATGTTTTGTTATCTTGACT
	5'-CAAGTTGCCTCTGGCTCTATTTGACTA
ALVE-ros005	5’-CTAAATATGTCTTTTTTGTCTCCTTGATAC
	5’-GGGTAAACAATAGCACTGCTCCTTAT
	5’-ATTCTAAATATGTCTTTTTTGTCTCCTTGATAT
	5’-TAGCGATTGCGAACACCTGAATGAA
ALVE-NSAC3	5'-AAACAGCTGATGGTATATCTTTTCATAAAATA
	5'-CATTTCCATACACATCACAGAGATGAAATT
	5'-ACAGCTGATGGTATATCTTTTCATAAAATG
	5'-GAACACCTGAATGAAGCAGAAGGCTT
ALVE-ros006	5'-ATATTCAGACACAGGACTCC
	5'-AGCCAGTCTGTATCTTCTGT
	5'-CATATTCAGACACAGGACTCT
	5'-TCTGGCTCGATTTGACTAC
ALVE15	5'-TAGAATATATTTACAAAAATCTCCATRTTTATG
	5'-CTTTGAGGATCTACTTGATGAAAACATGTT
	5'-TAGAATATATTTACAAAAATCTCCATRTTTATT
	5'-TAACGATTGCGAACACCTGAACGAA
ALVE-ros007	5'-TCATAACAATGGAGATGTGGGAATAGATA
	5'-TTGATCCAAGGCAGGTAGTATTATCTGTT
	5'-TTCATAACAATGGAGATGTGGGAATAGATT
	5'-CGTTGAGTCCCTAACGATTGCGAA
ALVE-ros008	5'-TATGATAAGAATTTTCTGTAGG
	5'-AATATCTGAGACAGAGAATAAA
	5'-CTATGATAAGAATTTTCTGTAGT
	5'-AGACTATTCAAGTTGCCTCT
ALVE-ros009	5’-GGGTTCATGCTGTGTCCAGTG
	5’-GAACACCTGAATGAAGCAGAAGGCTT
	5’-GGGTTCATGCTGTGTCCAGTC

	5’-AGATTTGCAGAGGGTGCACTCCAT
ALVE9	5'-TCCCTCTTGAGTCTCAAAAGAGTTC
	5'-GAAAGCCTGTGTATTATTAAAGGCCC
	5'-ATTCCCTCTTGAGTCTCAAATGAAGCC
	5'-TCCCTAACGATTGCGAACACCTGAA
ALVE-NSAC7	5'-TGGATGAACAAGTTCACTTCTCTAAGA
	5'-ACCAGACTGCATGTGTGTTAGCTTAACA
	5'-TTGGATGAACAAGTTCACTTCTCTGAAGC
	5'-ACGATTGCGAACACCTGAATGAAGCA
ALVE-ros010	5'-CTTATTTAGGAGAAATGCAAATGTAGGCTA
	5'-TTTGTGTAATACCCCATTAGGGGCATAT
	5'-ATTTAGGAGAAATGCAAATGTAGGCTG
	5'-AAGTTGCCTCTGGCTCTATTTGACTA
ALVE3	5'-GCTCCACGGTCCGTGGTTG
	5'-TGCAGTAATGGTGTTTGACACCAATTGAT
	5'-GGCTCCACGGTCCGTGGTTT
	5'-GAACACCTGAATGAAGCAGAAGGCTT
ALVE21	5'-AAAACCAAACACTTTTGTATATGGGTAGTT
	5'-CTGACTTCACTACTCAGCATCACCAA
	5'-ACCAAACACTTTTGTATATGGGTAGTG
	5'-GAACACCTGAATGAAGCTGAAGGCTT



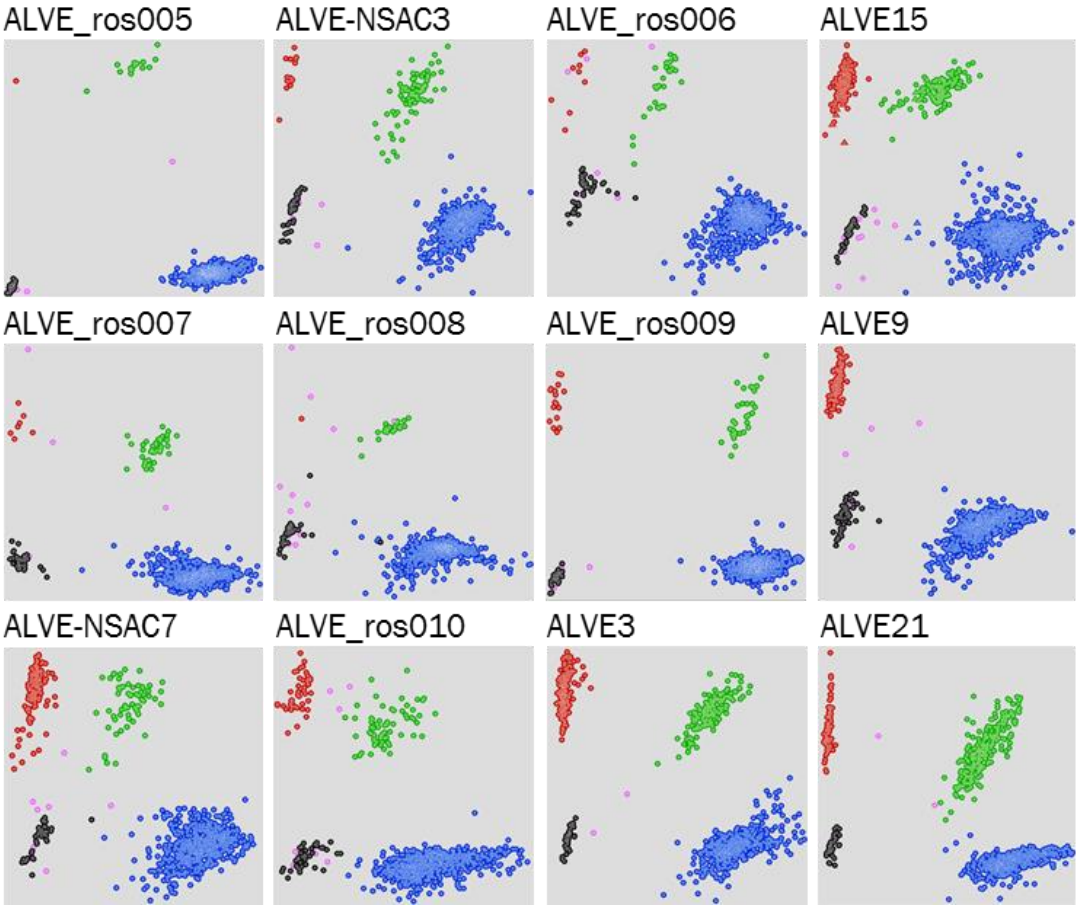


Figure 6.3 KASP diagrams for each of the identified ALVEs using the 2010 males of all eight Hy-Line lines. The red data points are individuals homozygous for the ALVE, greens are heterozygotes, blues are wildtype homozygotes, and blacks are the controls. Pink data points were ambiguous either due to their location outside genotype clusters or the path taken during cycling (such as the pink data points within the genotype clusters of ALVE_ros006). For assays such as ALVE_ros003 (where the insert allele was rare) groups have been called using data from multiple generations (not shown). ALVE-TYR and ALVE9 lack any heterozygote calls due to these ALVEs being fixed in some lines and absent in all others.

Most assays needed multiple designs to fix issues caused by SNPs in the primer binding regions (Figure 6.4) or non-amplification of one of the alleles, usually the ALVE insert (Figure 6.5). This usually involved identification of SNPs so that they could be included as ambiguous wobbles in the primers or avoided completely. Primers ordered with ambiguous bases included both primer variants in the indicated locations, ultimately limiting the number of manageable SNPs to two within each primer. SNP-related issues were often with the wildtype group, as assays were designed based on ALVE+ lines, so

SNPs in the ALVE- lines were overlooked. One example of this was ALVE_ros005 which had a group consisting of individuals from all three brown egg layer lines which did not leave the origin in the original assay design (Figure 6.6A). The assay was successfully redesigned (Figure 6.6B) after sequence from origin group individuals from all three lines showed a 6 bp deletion in the 'no insert' reverse primer binding location (Figure 6.6C). Whilst individuals from all three lines exhibited the deletion, ALVE_ros005 was only identified in WPR1 and the RIR, both at rare frequencies.

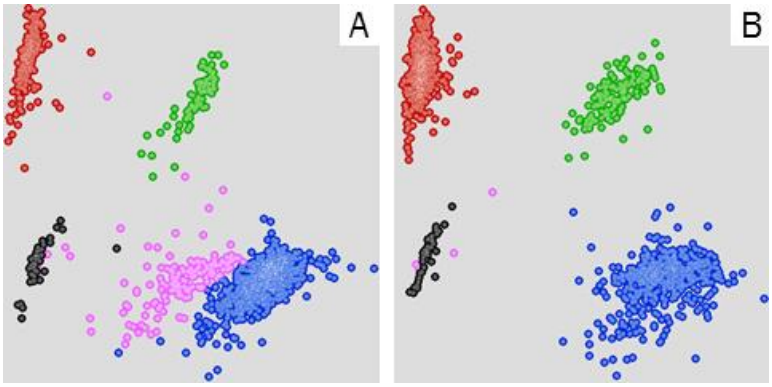


Figure 6.4 ALVE15 KASP assay redesign based on a SNP the base before the insert hexamer. A) The pink group to the left of the main blue group were from brown egg layer lines which contained a G/A SNP in the fluorescent-tagged primer binding region. This reduced wildtype fluorescence and created this 'slow' group. B) The redesign included the SNP ambiguity (G/A = R) and enabled blue group resolution.

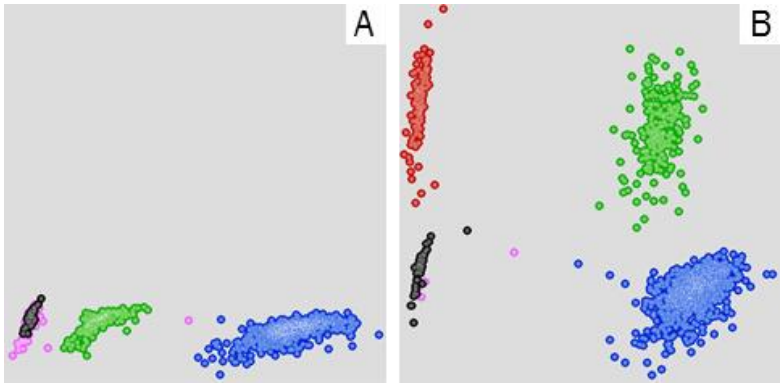


Figure 6.5 ALVE-NSAC1 KASP assay redesign due to non-amplification of the insert primers. A) Pink group at the origin did not move due to non-amplification of the insert-specific primers. This also caused the green heterozygote group to remain on the *x*-axis. B) Redesign of the insert-specific primers enabled amplification creating the red group and better resolving the green and blue groups.

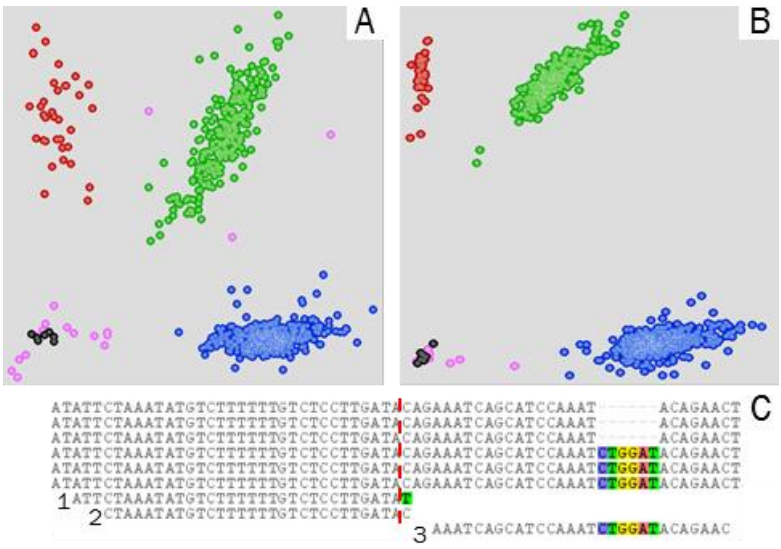


Figure 6.6 ALVE_ros005 assay redesign due to presence of an origin group consisting of individuals from both WPR lines and the RIR. In the original design (A) a group of pink individuals did not leave the origin, suggesting that there was a problematic SNP or other variant in one of the primer binding regions. The assay was redesigned (B) following identification of a 6 bp deletion in the original 'no insert' reverse primer location in all three lines by Sanger sequencing (C). The top three sequences in C were from the pink 'fail' group in A, and the next three were from individuals which were correctly called in A. The insertion site is indicated by the red dashed line, and the KASP primer locations are shown by the numbered primer sequences, where 1 = insert fluorescent primer (hence the highlighted T after the insertion site), 2 = 'no insert' fluorescent primer, and 3 = the 'no insert' reverse primer, showing the impact of the 6 bp deletion. The redesign moved primer 3 to avoid the deletion, producing tight genotype groups and the removal of most pinks. Remaining pinks were likely due to low quality DNA as those samples also failed in several other KASP assays. These DNAs should be replaced, as two of the eight RIR pinks in the redesign were ALVE_ros005 heterozygotes when checked by gel-based PCR assay.

An advantage of the KASP system was the ability to PCR the plates for a given number of cycles, read the plates, and then continue PCR if necessary. This aided assay improvement for the optimum number of cycles (Figure 6.7), and enabled tracking of individual samples during PCR to ensure the path taken was consistent with other individuals in the genotype cluster (*e.g.* pink calls in ALVE_ros006; Figure 6.3).

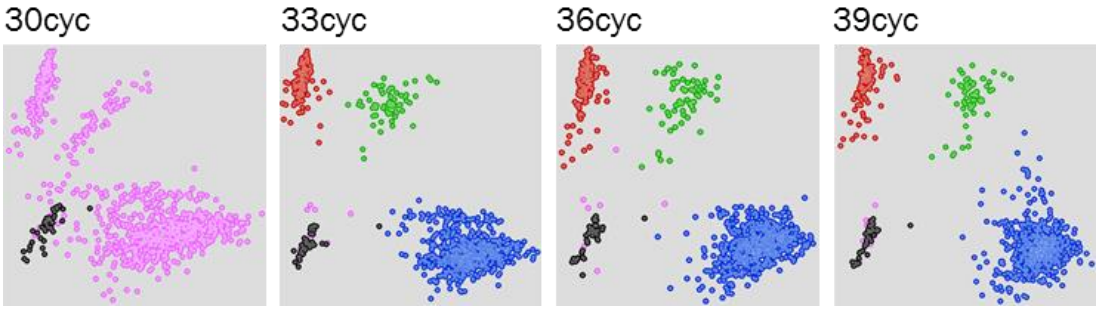


Figure 6.7 ALVE-NSAC7 genotype cluster resolution during PCR cycling, with cycles increasing from left to right. Initially the groups are not distinct enough to be called (30cyc), but resolution gradually improves through the 33cyc and 36cyc. 36cyc was chosen as optimum due to the better resolution between the red and green groups (several intermediates changed genotype call between the two). As the plates become 'over-cycled' (39cyc) the blue cluster loses cohesion, with a group moving up the plot towards the heterozygotes.

6.4.2 Diagnostic PCR development

In addition to the KASP assays used for large-scale genotyping, diagnostic PCR assays were also developed as these are more suitable for the non-industrial genotyping of smaller flocks. Of the twenty identified sites, ten already had published PCR assays: ALVE1, ALVE3, ALVE9, ALVE15, ALVE21, ALVEB5, ALVE-TYR, ALVE-NSAC1 and ALVE-NSAC3. However, the original ALVE1 upstream primer had a T_m of just 49.6°C which caused amplification problems for the longer PCR runs used in the ALVE sequencing (sections 5.2.4 and 6.6), so it was redesigned to give a T_m of 67.7°C.

Twenty-eight new primers were designed for the remaining ten identified ALVEs, with ALVE_ros005 and ALVE_ros007 having a two primer assay protocol rather than the standard three. When designing primers for ALVE_ros005 it was observed that read pairs were spanning the predicted insertion site, suggesting that it was a solo LTR like ALVE15. Concordantly, the same primers could amplify all three possible genotypes. An alternative, 'insert' primer was initially designed, but it became redundant following identification of the site as a solo LTR. The ALVE_ros007 design was inhibited by not knowing whether the insertion was always truncated and associated with the genomic deletion. Although the KASP assay worked on the assumption that individuals were either wildtype or had both the truncated insertion and genomic deletion at this locus, the PCR assay would potentially be used on lines where this was not the case, so band

Table 6.5 Primers used for the diagnostic Hy-Line-based PCR assays. For each element, the first sequence is the forward, upstream primer and the second is the reverse, downstream primer. In three primer assays the third primer is the alternative primer. Any original Benkel LTR primers are named (LTRA *etc.*). The papers for previously published primers are numbered where: 1 = Benkel, 1998; 2 = Chang *et al.*, 2006; 3 = Smith & Benkel, 2009; 4 = Smith & Benkel, 2008.

Name	Primers	ALVE+ (bp)	ALVE– (bp)
ALVEB5	5'-CAGTCATATATCCGAATGTTTAAGTCT (1) 5'-GGAGCCATAATTTCATAATGAA (1) 5'-CGCCCATATGTCCTTGCGTC (LTRD; 1)	123	241
ALVE1	5'-CGGTTATAATGAGGGTTGTGCTTTTC 5'-GCACCAAACAATCTAGTCTGTGC (1) 5'-CCTGAATGAAGCAGAAGGCTTC (LTRA; 1)	260	368
ALVE-ros001	5'-TTCCTCTCAGGTCTTCTTGGCG 5'-TGCCCTACGTATGACAATGCTG 5'-AACGATTGCGAACACCTGAATG	275	460
ALVE-ros002	5'-TCAGCAGCAACAGAAATCCAGC 5'-CAAGACCCACCTGGATGCCTAC 5'-GGCTATTCAAGTTGCCTCTGGC	270	361
ALVE-ros003	5'-TCTCACAAACCCCACAGGTGTC 5'-TTTGTCATCTCTGTGCCCTTGG 5'-GTTGCCTCTGGCTCTATTTGAC	186	498
ALVE-TYR	5'-TTGAGATACTGGAGGTCTTTAGAAATG (2) 5'-CAAAACCATAAATAGCACTGGAAATAG (2) 5'-CCTCTGGCTCTATTTGACTACACAGT	345	481
ALVE-NSAC1	5'-GGTTTGGAGAGCGTTAGCAG (4) 5'-TGACGTCTGTTTTCCCATGA (4) 5'-TGTAGTCAAATAGAGCCAGAGG (LTRC; 1)	340	519
ALVE-ros004	5'-CTAAGTAGCTGTCATCCCCACC 5'-AAAGTGTTTCCAGCAGTTTTCC 5'-AAGTTGCCTCTGGCTCTATTTGAC	273	336
ALVE-ros005	5'-TGGGGAAGTTGTGCTTTTCCAC 5'-TCTTTGCAACACAGCTTGGGAG	668	388
ALVE-NSAC3	5'-TGCTATCTCCCTGCTCATTTG (3) 5'-CACCCAGATCCTTTTCCTCA (3)	163	500

	5'-GAGTCCCTAACGATTGCGAACAC (LTRF; 1)		
ALVE-ros006	5'-TAACTTCTCTTCCAGCCTCAGC 5'-TTTTTCAAGGAGCAGAAAATCC 5'-GAACCCCTAAATGAAGCTGAAG	333	417
ALVE15	5'-CAAATGAGGGTAATAAGGGAG (1) 5'-CACTACCAAATATAATTCTGTAG (1)	460	180
ALVE-ros007	5'-TCAGCATAAAACCACAGCAAAG 5'-GTGGATTTTGGGCTACTTTCAG	1,970	2,511
ALVE-ros008	5'-GCACAGAGAAGGATATGTGCTG 5'-CTGTAAAAGAATCCCATGCCTC 5'-GACTATTCAAGTTGCCTCTGGCTC	336	428
ALVE-ros009	5'-ACAGCCTCTCTGGACAACCTGG 5'-GCCCATGTCAAACATCATCAGG 5'-TGTTTTTCCCTTATTTGGTCTTCAG	248	366
ALVE9	5'-CATTCTCCATGCACCTGAAGTG (1) 5'-TAGTGCACATATAATTTCAGATGAGTT (1) 5'-ACCTGAATGAAGCTGAAGGCTTC (LTRB; 1)	115	450
ALVE-NSAC7	5'-ACACCATCTCCATACACTTCCC 5'-GAAATGCACGTAAGCACAAAAG 5'-TGAATGAAGCAGAAGGCTTCAGAG	172	288
ALVE-ros010	5'-CCAAGCTCTGAACATACACTGC 5'-CTGGGTAACAGAAGAGTGGTCC 5'-CTAACGATTGCGAACACCTGAATG	228	433
ALVE3	5'-GAAATGCCTGCCCCATGCCAGTG (1) 5'-CTTCTCCAGCTTCAGTGACGC (1) 5'-CCTGAATGAAGCAGAAGGCTTC (LTRA; 1)	190	270
ALVE21	5'-CATTTCAAGCAAGGGACTGGC (1) 5'-GTGGGAATGGTACTACAGAGAAGG (1) 5'-ACCTGAATGAAGCTGAAGGCTTC (LTRB; 1)	390	510

size variations between individuals had to be observable. Consequently, only the two external primers were used for this PCR assay, with a band length of 2,511 bp in the wildtype, and 1,970 bp in an individual with both the truncated insertion and genomic deletion. Any major deviations of these band sizes, or their absence, would therefore

require further investigation. To account for these longer PCR products, the elongation step in each PCR cycle was extended to 3 minutes.

All primers designed for this project are summarised in Table 6.5 along with the previously published primers and predicted band sizes for products with and without the ALVE insertion. Each primer set was tested on the HL DNA samples and the observed genotypes corroborated the KASP assay results.

6.4.3 Genotyping of the 2008 birds used for resequencing

DNA samples from the eighty individuals used for the 2008 HL line resequencing project were genotyped using all the developed KASP assays. Interestingly, there were marked differences between frequencies of ALVEs shared between lines of the same breed, and many of the 'common' layer ALVEs, such as ALVE1, ALVE3 and ALVE15, were not fixed in WLs as has been previously suggested.

WL ALVE frequencies were generally high, with ALVE1 fixed in three lines, ALVE9 fixed in WL3, and ALVE3 and ALVE15 found at medium or high frequencies in all cases (Table 6.6). Only the WL4 insertion ALVE_ros008 was at a rare frequency, with two heterozygotes in the original pool. ALVE_ros008 had the lowest observed frequency of any insert identified by the identification pipeline. This region had 10.8X coverage, consistent with the genome average of 11.1X, but the individual site modelling suggests that there was a 74.5 % chance of missing this insertion in the data.

As expected, both WPRs were fixed for ALVE-TYR. The other insertions were of medium frequency, except ALVE-NSAC7 which was fixed in WPR1 and at a high frequency in WPR2. Interestingly, as the KASP assays were applied across all the lines, ALVEB5 was also identified in WPR1. This insertion was only seen in one homozygous WPR1 individual and there was no evidence for it in the full, reference genome mapping. Coverage over this site was 11.7X, 15 % lower than the genome average, and modelling suggested only a 28.6 % chance of this insertion being detected. Following detection of ALVEB5 in WPR1, the only observed difference between the WPR ALVE contents was the novel chromosome 4 insertion ALVE_ros009 in WPR2 (Table 6.7).

Table 6.6 Observed frequency categories for each identified White Leghorn ALVE in the five Hy-Line white egg elite layer lines, using the 2008 resequenced birds.

Insert	WL1	WL2	WL3	WL4	WL5
ALVE1	Fixed	Fixed	Medium	High	Fixed
ALVE15	High	Medium	-	-	Medium
ALVE_ros008	-	-	-	Rare	-
ALVE9	-	-	Fixed	-	-
ALVE3	-	Medium	Medium	High	-
ALVE21	-	-	-	Fixed	-

Table 6.7 Observed frequency categories for each identified ALVE in the three Hy-Line brown egg elite layer lines, using the 2008 resequenced birds. Entries indicated by asterisk were identified by KASP assay, not the identification pipeline.

Insert	WPR1	WPR2	RIR
ALVEB5	Rare *	Medium	Medium
ALVE_ros001	-	-	Medium
ALVE_ros002	-	-	Medium
ALVE_ros003	-	-	Low
ALVE-TYR	Fixed	Fixed	-
ALVE-NSAC1	Medium	Medium	Medium
ALVE_ros004	-	-	High
ALVE_ros005	-	-	Low
ALVE-NSAC3	Medium	Low	-
ALVE_ros006	-	-	Medium
ALVE_ros007	-	-	Medium
ALVE_ros009	-	Medium	-
ALVE-NSAC7	Fixed	High	-
ALVE_ros010	-	-	Medium
ALVE3	-	-	Rare *
ALVE21	Fixed	Fixed	-

The RIR exhibited no fixed inserts, with most inserts having medium frequency and two with low frequency (Table 6.7). As with WPR1, the application of all developed KASPs to all lines enabled the detection of ALVE3 in the RIR. This is an insert shared with three WLs and not with the WPRs, the other brown egg layer lines. However, the frequency of this insert is very low, as it was only found in one heterozygous individual. Again, full genome mapping did not support this insertion, despite the region having 36.5X coverage, more than double the genome average. This exemplifies the potential for allele specific amplification during sequencing library preparation, as homogeneous amplification should result in one or two ALVE-homologous reads. Detection modelling for this site suggested a 44.0 % chance of identifying ALVE3 in the data.

Probability of detecting insertions with the HL pooled sequencing data

The probability of insert detection in the pooled datasets is almost perfectly, positively correlated with the average genome coverage ($r = 0.999$, $P < 0.00001$). However, the combination of sampling bias, variability in allele specific amplification, and relatively low coverage generates quite low detection probabilities from the model.

Across the eight lines, 90 % detection confidence is only achieved with insertion frequencies between 0.3 (WL1) and 0.5 (WL4). At ALVE frequencies as low as 0.1, there is only a 28 - 45 % chance of detecting that insert across the lines. This means that there is a high probability that rare ALVEs will be missed from the pooled sequencing data. This is particularly important when ALVE insertions are line specific, and therefore cannot be identified using assays developed from detection in other lines. As coverage and sequencing variability largely determine detection probability, increasing the number of sampled individuals in the pool does not improve the rate of detection.

6.4.4 Multi-generation genotyping of line males

In addition to the eighty individuals used for resequencing, DNA was also available for all the males of each of the eight lines for fifteen generations, totalling over nine thousand samples. This enabled the detection of insert frequency changes over time, as well as the

potential discovery of inserts present in lines when the birds used for resequencing did not contain those inserts. However, this would only detect inserts found in at least one of the lines from the resequencing data, not unidentified line-specific inserts, as a KASP assay would not have been developed for them.

The multi-generation data for the WLs matched the sequenced individuals very well, with each frequency category matching those from the 2008 resequenced individuals (Table 6.6). One difference was that the observed ALVE1 frequency for WL3 was half that observed in the sequenced individuals. This frequency remains in the 'medium' category, but is rarer than the original sampling would suggest. Interestingly, ALVE3 frequency has been gradually increasing in WL2, WL3 and WL4, but not in the RIR. ALVE3 is an intronic insertion, within the sixth intron the HCK tyrosine protein kinase gene, and is centred in a 200 kb region containing twenty genes (Figure 6.8). This region includes the immune gene IRF9 (interferon regulatory factor 9), and the apoptosis regulator BCL2L1 (B-cell lymphoma 2 like protein 1), which has been identified as a candidate in differential response to MDV (Ohashi et al. 1998).

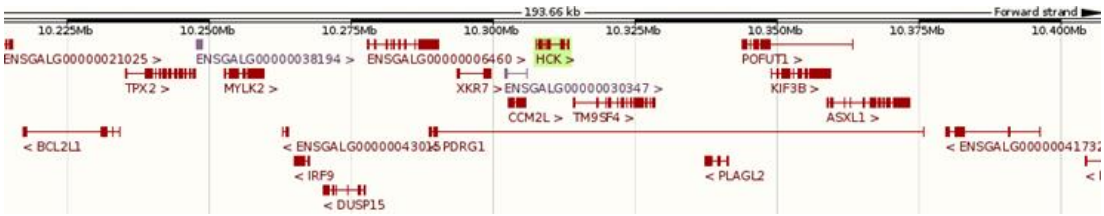


Figure 6.8 Ensembl view of the 200 kb genomic region centred on the ALVE3-containing HCK gene. Tracks have been collapsed down to genes (red) and predicted lncRNA (grey). The image is a screenshot from the Ensembl website.

The WPR sequenced individuals were far less representative of the multi-generation data. ALVEB5 was rarer overall in both lines than in the sequenced individuals, and ALVE-NSAC7 was not fixed in WPR1, as was suggested above, but was 25% rarer, with a frequency bordering medium to high, lower than observed in WPR2 (Table 6.8). Additionally, three insertions identified in the RIR were identified in the WPRs. Both lines had ALVE_ros004 at a rare frequency and WPR1 also exhibited ALVE_ros005 at a rare frequency and ALVE_ros010 at a low frequency. In WPR1 the frequencies of

all three had remained consistent over the sampled generations, but the WPR2 ALVE_ros004 frequency dropped rapidly over the last eleven generations from medium to rare frequency. ALVE_ros004 is intergenic: over 40kb upstream from the reverse strand gene MMP16 (matrix metallopeptidase 16), and 450kb upstream of the forward strand gene RIPK2 (receptor interacting serine/threonine kinase 2). Additionally, ALVE_ros009 was also found in WPR1 in the 1996-1998 generations at rare frequencies, but was then lost from the population, so was not counted in the 2010 frequency data. The frequency of ALVE_ros009 in WPR2 did not have a similar decline, so the loss from WPR1 was likely due to drift.

Table 6.8 Observed frequency categories for each identified ALVE in the three Hy-Line brown egg elite layer lines, using the most recently available, 2010 full generation of male birds. Entries indicated by asterisk were ALVEs present in the lines but absent from the 2008 resequenced birds, and the entries indicated by the circumflex were ALVEs with frequency categories different from Table 6.7.

Name	WPR1	WPR2	RIR
ALVEB5	Rare	Medium	Low ^
ALVE_ros001	-	-	Low ^
ALVE_ros002	-	-	Low ^
ALVE_ros003	-	-	Low
ALVE-TYR	Fixed	Fixed	-
ALVE-NSAC1	Medium	Medium	Medium
ALVE_ros004	Rare *	Rare *	Medium
ALVE_ros005	Rare *	-	Rare ^
ALVE-NSAC3	Medium	Low	-
ALVE_ros006	-	-	Medium
ALVE_ros007	-	-	Medium
ALVE_ros009	-	Medium	-
ALVE-NSAC7	High ^	High	-
ALVE_ros010	Low *	-	High
ALVE3	-	-	Low ^
ALVE21	Fixed	Fixed	-

The RIR frequencies matched the resequenced individuals well, despite category changes for four ALVEs (Table 6.8). ALVE3, which was not originally identified in the bioinformatics pipeline, had a higher frequency than in the resequenced individuals, but the observed frequency was consistently low across the generations.

6.4.5 Final list of identified ALVEs in the Hy-Line lines

Overall, two to four ALVEs were identified in the five WL lines, nine were identified in WPR1 and eight in WPR2, seven of which were shared, and eleven ALVEs were identified in the RIR. This totals twenty different ALVEs with forty-two occurrences. The predicted gain and loss of ALVEs within the lines is represented in Figure 6.9.

The two brown egg breeds (WPR and RIR) shared five ALVEs: ALVEB5, ALVE-NSAC1 and the novel ALVE_ros004, ALVE_ros005 and ALVE_ros010. The RIR and three WL lines shared ALVE3, and there was at least one documented occasion in the RIR line history where there has been a WL cross, likely introducing this typically white egg layer ALVE into the RIR. ALVE21 is the only other cross-egg-colour ALVE, but as this is associated with the commercially important slow feathering locus it was likely introduced when breeding for this phenotype.

Unidentified ALVEs within the Hy-Line lines

BAM files for the Hy-Line lines were also manually inspected for a whole range of additional ALVEs with known insertion sites (Table 6.9). These were identified using existing diagnostic assay primers, flanking information and insertion hexamers, as well as information kindly shared by Professor Bernhard Benkel. However, no additional ALVEs were identified beyond those discovered by the pipeline.

There was no evidence in any of the eight lines for the two ALVEs in the chicken reference genome assembly: ALVE6 and ALVE-RJF (Benkel & Rutherford 2014). Whilst ALVE-RJF (Figure 6.10) has not been identified in any 'modern' chicken breeds (Benkel & Rutherford 2014), the literature for ALVE6 states that it is common in layers (Benkel 1998), so it was surprising that it was not identified by the pipeline, particularly

after successful detection via preliminary gel-based PCR assays completed by Hy-Line in WL4 and WPR1. However, the ALVE6 insertion site is at the very start of chromosome 1 (Figure 6.11), and, concordantly, read mapping is poor even for normal

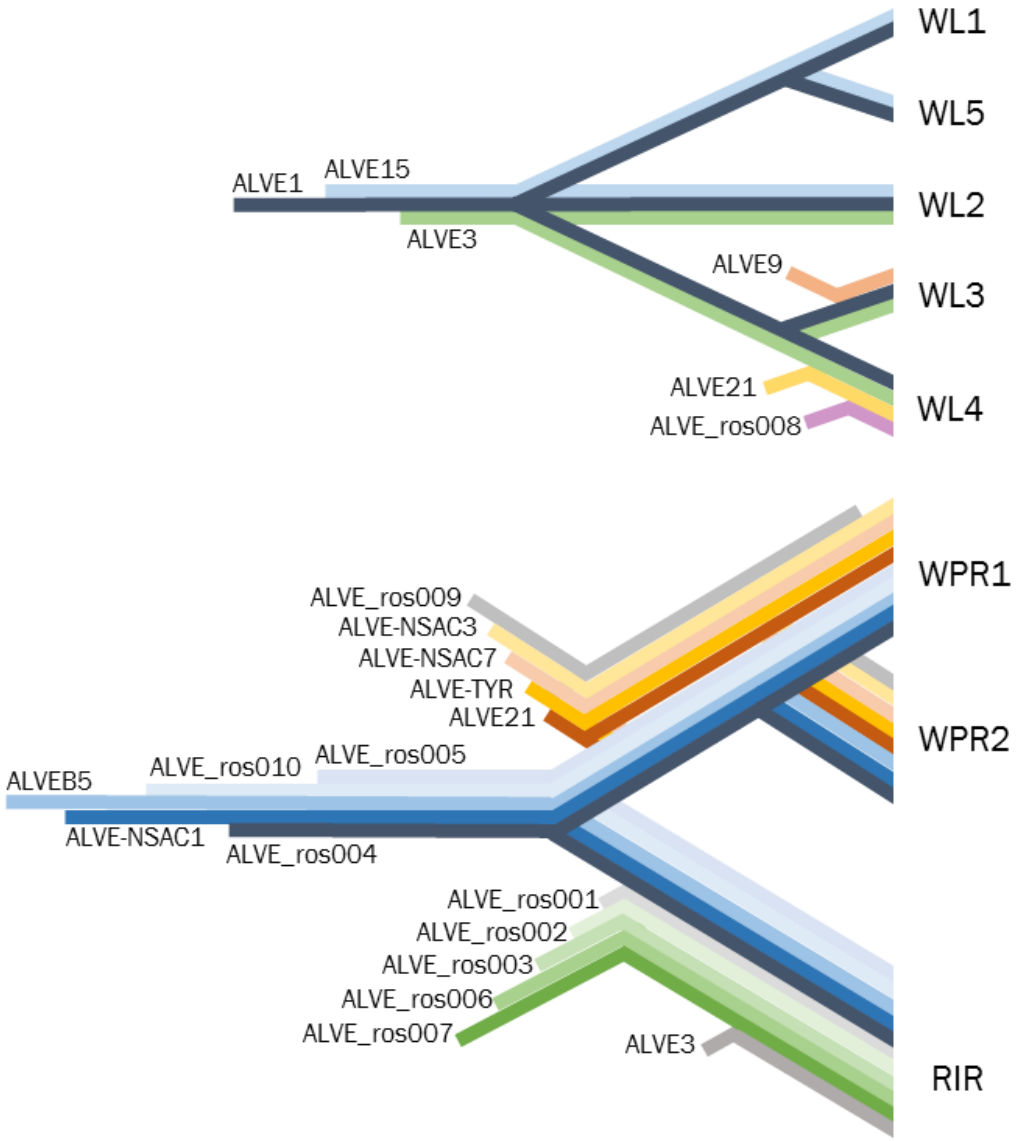


Figure 6.9 Network map of shared ALVEs in the Hy-Line elite layer lines. Gains and losses have been plotted in the most parsimonious manner, so the white and brown egg layer groups are independent. Each ALVE is shown by a different colour in each lineage, with the number of different coloured lines reaching each HL line representing the total number of ALVEs identified. No line inter-relatedness has been included for the WLs and the presence/absence of ALVE3 and ALVE15 was assumed to be due to incomplete lineage sorting. The ALVE_ros009 WPR1 line does not reach the end as it was present in the 1996-1998 generations, then lost.

Table 6.9 ALVEs with previously published insertion sites, insertion hexamers or diagnostic assays which were manually checked across the Hy-Line lines.

Name	Location	Hexamer
ALVE6	1: 576	GGCGCT
ALVE-RJF	1: 32,603,304	GGCTTG
ALVE2	1: 36,872,528	GAGGGG
ALVE16	1: 67,449,967	CATGGC
ALVE12	1: 122,259,940	GTGTTG
ALVE-NSAC2	1: 146,108,486	GGGTCC
ALVEB11	2: 62,587,235	AGAGGA
ALVEB2	2: 95,058,383	GACCAT
ALVE-NSAC5	3: 73,338,411	GGCTGA
ALVEB10	4: 27,829,406	GCATTC
ALVEB4	4: 88,793,982	ATGTTT
ALVEB9	5: 12,126,222	GGGGAC
ALVEB1	5: 23,238,400	GTTATT
ALVE-NSAC6	5: 57,261,506	AAAACT
ALVE4	6: 33,827,722	GCTGCC
ALVEB3	7: 20,479,059	GTAGTC
ALVE-NSAC4	12: 17,625,255	CCTGGG
ALVEB6	14: 9,367,708	GTGTCT
ALVEB8	20: 1,468,850	GACTAC
ALVE7	Z: 14,471,852	ACCCTC

genomic reads. It is therefore possible that ALVE6 was missed during the identification process, as were any other insertions located at the telomeres, centromeres or other difficult to sequence regions of the genome. When combined with the high probabilities of missing low frequency or line specific insertions when using pooled datasets, it is highly likely that the twenty ALVEs across the eight HL lines are an under-representation.

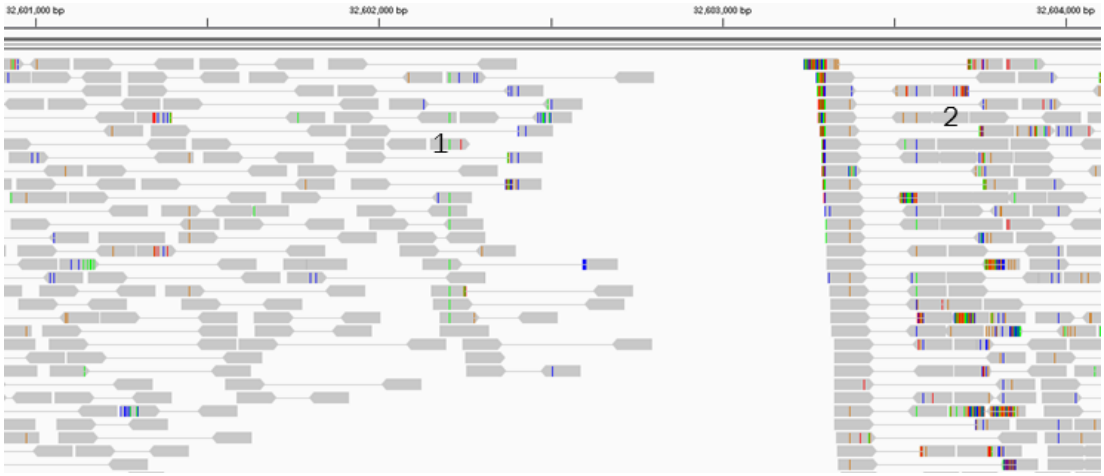


Figure 6.10 Hy-Line WPR1 read mapping (with linked read pairs) around the Galgal5-assembled ALVE-RJF. There was a clear demarcation between reads mapping to the reference genome 5' of the assembled ALVE-RJF, and ALVE-homologous reads mapping to ALVE-RJF itself. As no reads bridge this gap, WPR1 does not contain ALVE-RJF. In addition, the mate of read 1 maps 13,954 bp downstream, again supporting the absence of ALVE-RJF in WPR1. The mate of read 2 maps to the 5' genomic region which flanks the ALVE-TYR insertion.

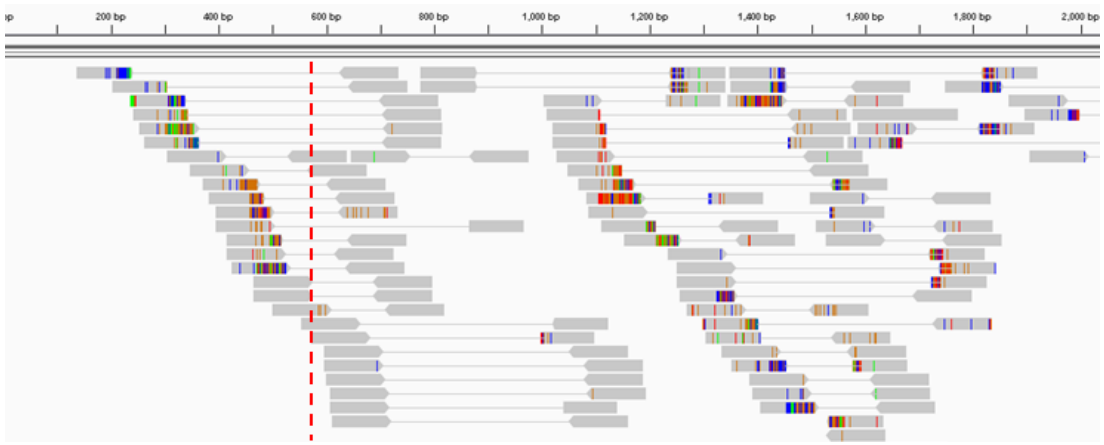


Figure 6.11 Hy-Line WPR1 read mapping (with linked read pairs) around the putative ALVE6 insertion site (red dashed line) in the first 2,000 bp of the Galgal5 chromosome 1. Read mapping is generally poor and there is no evidence of an ALVE insertion at the indicated site.

6.5 Further characterisation of ALVE21 and the *K* locus

ALVE21 was identified in WPR1 and WL4 as expected, but also in the fast feathered WPR2 (all three lines are fixed for their feathering phenotypes). The BAM files for the

region clearly showed support for the ALVE21 insertion in these three lines, especially when compared to the ALVE21-, fast feathered WL5 (Figure 6.12).

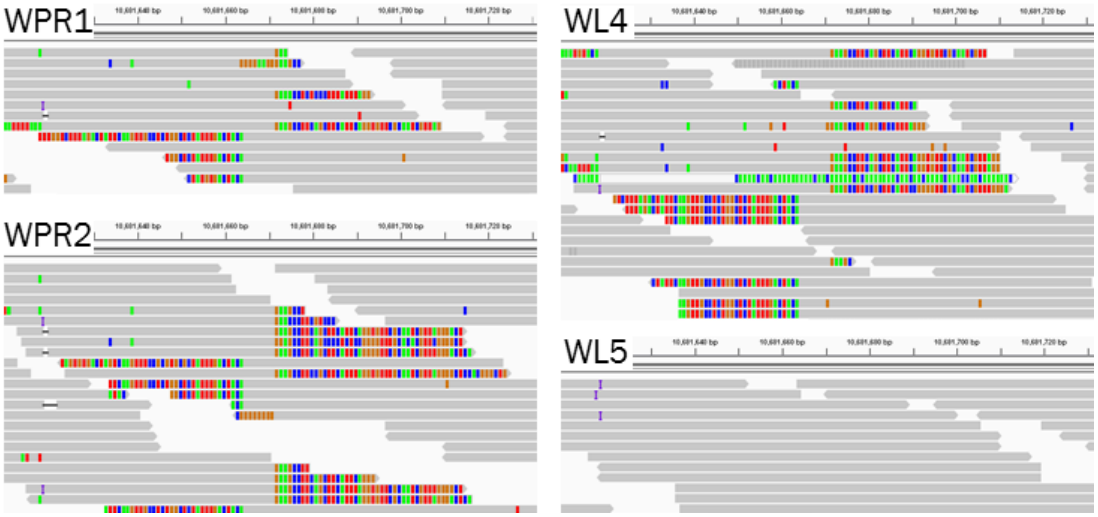


Figure 6.12 BAM file support for ALVE21. IGV screenshots show the clipped reads at the ALVE21 insertion site in WPR1, WPR2 and WL4, whilst reads map correctly across the region in WL5 (ALVE21-, fast feathered). Strikingly, no reads map through the insertion site in WPR2 without clipping. However, in both WPR1 and WL4 (the expected ALVE21+ lines) approximately half the reads covering the insertion site map through the region without clipping. This supports the presence of the locus duplication in WPR1 and WL4 (where one of the duplicates contains an unoccupied ALVE21 site) but not in WPR2, matching the feathering phenotype.

Whilst ALVE21 is closely associated with the slow feathering phenotype, it is the ***K*** locus duplication (Figure 6.13) which is the causative factor. However, there have been reported cases of phenotype reversion to fast feathering caused by recombination between the tandem repeated sections of the ***K*** locus (Figure 6.14) (Levin & Smith 1990). WPR1 and WPR2 are sister lines which were separated based on the fast/slow feathering phenotype. The presence of ALVE21 in the fast feathering WPR2 suggests that the fast/slow feathering phenotype was not segregating in the WPR population at the point of separation. Rather, that fast feathering revertants were selected as the founders of the WPR2 line.

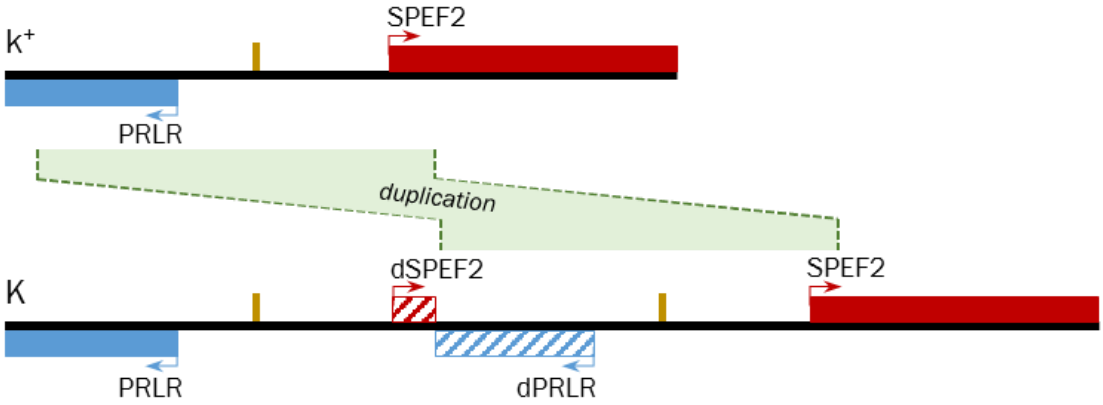


Figure 6.13 Schematic for the feathering locus. The wildtype k^+ has the PRLR gene on the reverse strand, SPEF2 on the forward strand and the ALVE21 insertion site (gold bar). The duplicated region is shown with the duplicated gene sections (dSPEF2 and dPRLR) and duplicated insertion site. In the *K* allele one of the ALVE21 sites is occupied and the other is empty (as in k^+), but it is unknown which site is occupied. The only unique sequence is the link between dSPEF2 and dPRLR.

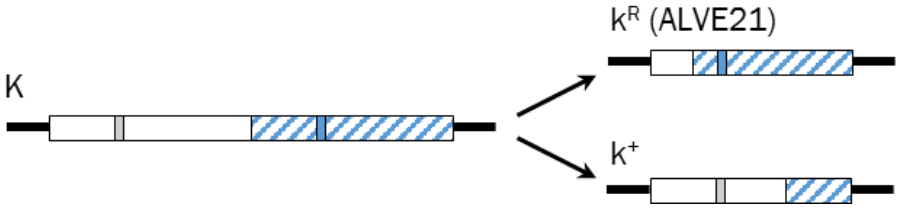


Figure 6.14 Potential *K* allele revertants. The tandem repeats of the *K* allele retain greater than 99 % homology so will readily recombine producing phenotypic revertants. With respect to ALVE21, there are two possible revertant genotypes, with (k^R) or without (k^+) the insertion. These genotypes will depend on where recombination crossing over occurs.

Duplication of the ALVE21 insertion site in the slow feathered ***K*** allele meant that the KASP assay identified all WPR1 and WL4 individuals as heterozygotes, as the 'empty' insertion sites were picked up by the 'no insert' primer pair (Figure 6.15A). In addition, the assay provided more support for the hypothesised feathering phenotype reversion (k^R) in WPR2, as all these individuals were homozygous for ALVE21. Furthermore, the ALVE21 KASP was used on male progeny of a WL4 (♂) x WL3 (♀) cross to show that these were 'genuine' heterozygotes for ALVE21 (Figure 6.15B, grey, K/k^+ genotypes).

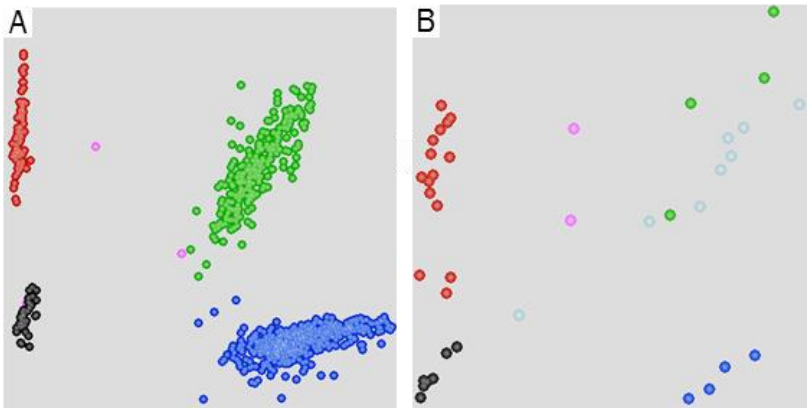


Figure 6.15 KASP assay results for ALVE21. A) Results for the 2010 males from all eight Hy-Line lines showing WPR1 and WL4 as heterozygotes, WPR2 as homozygous for the insert, and the remaining five, fast feathered lines as homozygous wildtype. B) Results for WL4 males (green, *K/K*), WL3 males (blue, k^+/k^+), the male progeny of a WL4 (♂) x WL3 (♀) cross (grey, K/k^+) and the female progeny of the same cross (red, *K*/-). Theoretically more samples would better resolve the grey group towards the *x*-axis, as the ALVE21 insertion site ratio is 2 empty to 1 occupied (*K* has 1 of each, k^+ has 1 empty). There were no samples available for 'true' ALVE21 heterozygotes (k^R/k^+, a WPR2 crossed with a non-WPR2 fast feathered bird).

6.5.1 K locus bridging sequence KASP development

Whilst the ALVE21 KASP assay was viable for use with lines where the feathering phenotype was known to be fixed, it cannot be used reliably on segregating lines. Even with perfect data, genotype clusters would likely resolve poorly and it would be impossible to distinguish between ***K/K*** and k^R/k^+ individuals (Figure 6.16). However, the genotypes can be reliably called when the ALVE21 KASP is used in tandem with an assay specific to the duplication bridging region within the ***K*** allele.

As Figure 6.13 shows, the only unique sequence within the ***K*** locus is the bridge between duplicated and wildtype sequence. Using primer sequences designed by Elferink and colleagues (2008) and the published bridging sequence (Bu et al. 2013), the exact bridging point was identified at Z: 10,800,130, meaning the duplicated region is 188,265 bp (Z: 10,611,865-10,800,129), almost 12 kb longer than previously reported. The duplicated region within the ***K*** locus therefore contains most of PRLR (up to 575 bases of exon 12) and the first four exons and 1,540 bases of the fourth intron of SPEF2 (378 bp short of exon 5). The potential ALVE21 insertion sites could therefore be at Z:

10,681,670 or at position 258,691 of the full ***K*** locus (measured as 441,255 bp from the outer points of PRLR and SPEF2).

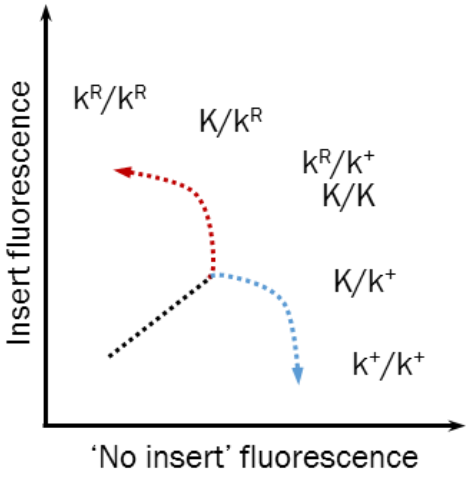


Figure 6.16 The continuum of possible ALVE21 KASP results from the six *K* locus genotypes. KASP resolution to five groups based on relative intensity has been shown (Fulton et al. 2016), but is unlikely with limited numbers of each genotype. In addition, no resolution would be possible between the k^R/k^+ and *K*/*K* genotypes.

A composite sequence for the ***K*** allele was created and used to design the duplication KASP assay (Table 6.10). The results from the use of this assay on all HL lines from the 2010 generation were as expected, with the slow feathering WPR1 and WL4 positive for the duplicated region and all others negative, including WPR2 (Figure 6.17).

Table 6.10 KASP primers for the *K*-duplication assay. As the result of this assay is the presence/absence of the entire unique sequence, the first primer pair is an internal control (commonly used by Hy-Line). The second primer pair is for the duplication sequence. The first primer in each pair carries the fluorescent tag.

Assay	KASP primers
K-duplication	5'-CCACGGTCCGTGGTTG
	5'-ATTGACAGATTGAGAGCTCTTTCTCGATT
	5'-ACTAGGGCTAGCATTTAATATAACCCCT
	5'-TGAAACCATCCCTGGAGAGATGGAA

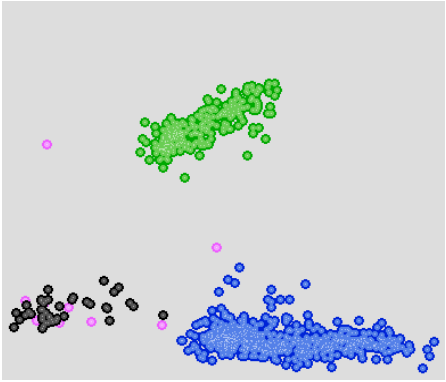


Figure 6.17 K-duplication KASP assay results. Due to the use of an internal control individuals homozygous for the duplication appear as heterozygotes (green). Some preliminary work (not shown) suggests 'true' duplication heterozygotes form an intermediary group. Here, the green group contains all WPR1 and WL4 individuals.

6.5.2 Characterisation of the *K* locus with BioNano optic mapping

The KASP assay results and original BAM file for WPR2 support the fast feathering phenotypic reversion in this line. However, additional data was needed to confirm this result, and to identify which of the ALVE21 insertion sites was occupied in the ***K*** locus. Coverage across this region was low for all lines and varied to the extent that it was impossible to infer higher coverage of the duplicated sections of the ***K*** allele from existing sequence data. Sequencing alone cannot resolve this locus due to the high homology (99 %) between duplicate regions and the gaps of at least 70 kb between unique sequences. Even long read sequencing technology would be unable to reliably resolve this region. However, the high resolution optic maps produced by the BioNano Irys technology have average N50s greater than 200 kb so can concatenate regions of unique sequence (Figure 6.18).

Optic maps for the Hy-Line samples

A total optic map length of 140 – 200 Gb was predicted for all samples, but far less data was generated for the WPR1, WPR2 and RIR-sf samples, and molecule N50 values were shorter than the 200 kb average in four cases (Table 6.11). Furthermore, the success of the *de novo* assemblies was generally poor, with no consensus maps generated for WPR2, very limited genome coverage for the other samples, and no consensus maps

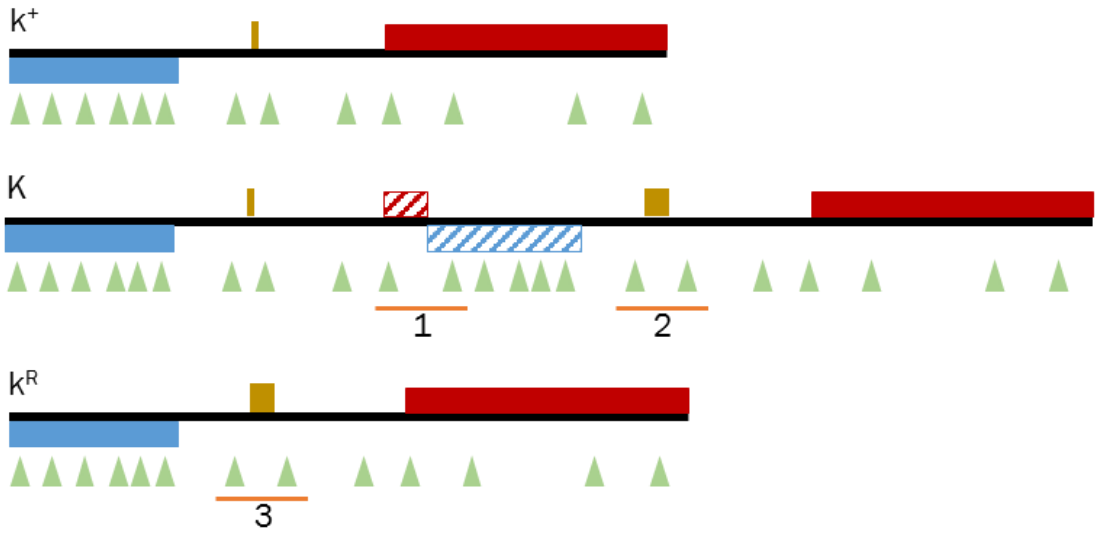


Figure 6.18 BioNano strategy for optic mapping of the *K* locus. The BioNano Irys technology creates optic maps by incorporating fluorescent tags using modified restriction enzymes (green triangles). Observed tag patterns between different genotypes are then compared to identify the structural variants (locations in figure not biologically representative). The *K* allele-containing Z will not just be longer, but will also have differential tag patterns: 1) altered spacing due to unique duplicated region bridge; 2) increased spacing due to insertion of ALVE21 at one of the two unoccupied sites. The revertant k^R should have a pattern very like the wildtype k^+ allele, but will have the increased spacing due to ALVE21 (3).

Table 6.11 Optic map statistics for the five Hy-Line samples. WL3 was wildtype fast feathered (k^+), WPR2 was the predicted fast feathered revertant line (k^R), and WL4, WPR1 and RIR-sf were slow feathered (*K*). Statistics for total map length, number of optic map molecules, and map molecule N50 is shown for each analysed line. The final column is the total consensus map length after *de novo* assembly with the average map coverage in brackets. No consensus could be formed for WPR2 due to limited data, and the other sample consensus maps did not cover the entire Galgal5 genome (1.2 Gbp). The absence of long contiguous consensus maps limited SV detection, and contiguous maps were absent across the *K* locus.

Line	Map length (Mb)	Total molecules	Map N50 (kb)	Consensus map (Mb)
WL3	222,923.5	1,240,879	180.7	355.9 (23.7X)
WL4	288,517.2	1,637,620	180.1	551.0 (39.7X)
WPR1	126,395.3	568,595	230.5	159.9 (14.8X)
WPR2	41,146.6	253,579	159.0	No consensus
RIR-sf	49,009.3	331,003	141.9	5.2 (62.6X)

generated for any sample across the ***K*** locus region of the Z chromosome. Consequently, this data was insufficient to fully characterise the duplication within the ***K*** allele of the slow feathered samples or validate the ALVE21 insertion site in WPR2.

Visualisation of the WL3 optic maps mapped to the Z chromosome (Figure 6.19) revealed limited coverage across even the wildtype k^+ allele, although the corroboration between predicted and observed Nt.BspQ1 sites was generally good. However, there were multiple maps which exhibited incongruence against the *in silico* reference digest.

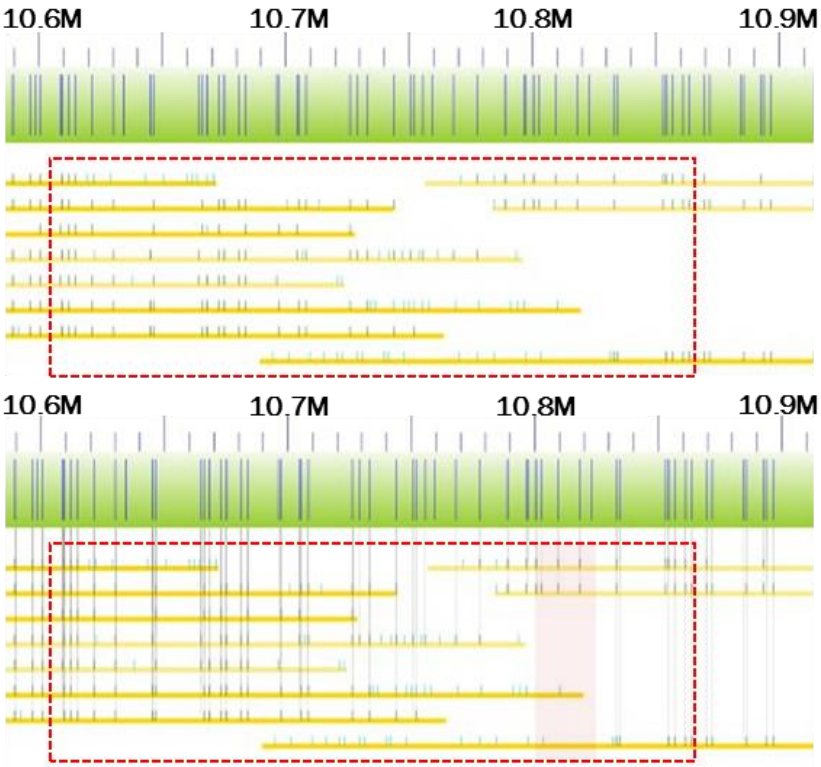


Figure 6.19 Optic maps for the WL3 (k^+) samples. In each panel the red dotted box outlines the k^+ allele, the top scale represents the location on the Z chromosome, and the blue vertical lines in the green bar represent the predicted Nt.BspQ1 sites. Both panels are the same, but the lower panel shows the matched sites between maps and the *in silico* digest. The red shaded section in the lower panel represents the bridging region in the duplicated *K* allele. Dark orange maps map to the positive strand, and the paler orange to the negative strand. The fluorescent tags (green marks on the maps) are darker green when the tags are more confident. Most tags match the *in silico* predictions well, and there are reads mapping through the bridging region, supporting the wildtype prediction. However, there are maps with poor matches (such as the very bottom dark orange map).

Importantly for this sample, there were optic maps which mapped through the bridging region which would generate breaks in the ***K*** allele (shaded red in Figure 6.19), supporting the wildtype k^+ genotype. None of the other samples had sufficient coverage across the region to draw any confident biological conclusions.

The primary aim of the BioNano optic mapping was to validate the presence of the ALVE21 insertion in the WPR2 sample, providing additional support for the phenotypic reversion of this line. As the original data was insufficient to address this question, the Earlham Institute generated an additional 220 Gb of optic map molecules which greatly enriched the original dataset. This data was sufficient to generate a consensus map across the ***K*** locus, and identified the ALVE21 insertion due to the augmented spacing of Nt.BspQ1 markers around the known insertion site (Figure 6.20). This corroborates the mapping scenario predicted in Figure 6.18, and provides additional evidence for WPR2 line origin as a fast feathering revertant.

It is likely that the generation of additional data for the slow feathering samples would enable the further characterisation of the ***K*** allele.

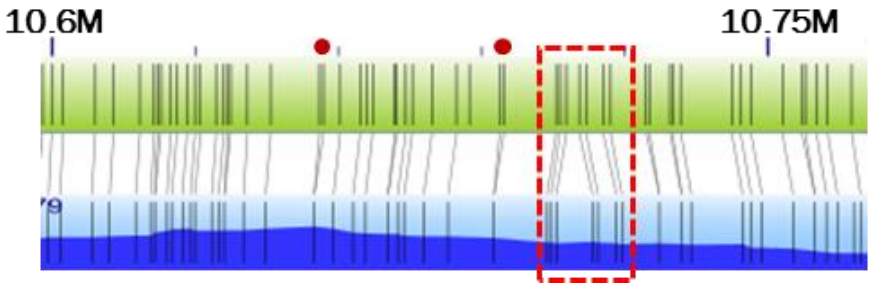


Figure 6.20 Consensus map for the WPR2 (k^R) allele. The scale bar is for the Z chromosome. The vertical lines in the green bar represent the *in silico* digest Nt.BspQ1 predicted sites, the two-tone blue bar represents the consensus assembly for WPR2 with the vertical lines showing the supported Nt.BspQ1 sites, and the connecting lines show matched sites. The dark blue within the two-tone blue bar shows the optic map coverage across the region. The red circles mark examples of tandem sites which have collapsed to a single site in the consensus due to cross-reacting fluorescence. The red dotted box marks the ALVE21 insertion site (Z: 10,681,671), and the expanded region between tags in the consensus shows the ALVE21 insertion in WPR2. This matches the Figure 6.18 prediction.

6.6 Characterisation of the identified ALVE sequences and assessment of their recombination, expression and retrotransposition potential

Fifteen of the twenty ALVEs identified across the HL lines were fully sequenced in this study (Table 6.12). Of these, ALVE15 and ALVE_ros005 were solo LTRs, and ALVE9, ALVE-NSAC1 and ALVE_ros007 had varying degrees of 5' truncation. The remaining ten were 'intact' elements with both 5' and 3' LTRs, although, as expected, the ALVE3 sequence had no *reverse transcriptase* domain.

Table 6.12 Key features of the fifteen sequenced ALVEs

ALVE name	Length (bp)	Orientation	Structure	LTR Identity (%)
ALVE1	7,530	-	Intact	100
ALVE3	5,848	+	Intact, no RT	100
ALVE9	5,077	-	*pol-env*-3'LTR	-
ALVE15	280	-	Solo LTR	-
ALVE21	7,529	-	Intact	100
ALVEB5	7,530	+	Intact	99.6
ALVE-NSAC1	4,838	-	*pol-env*-3'LTR	-
ALVE-NSAC7	7,531	-	Intact	100
ALVE-TYR	7,534	+	Intact	100
ALVE_ros001	7,531	+	Intact	100
ALVE_ros003	7,528	+	Intact	100
ALVE_ros004	7,530	+	Intact	100
ALVE_ros005	280	-	Solo LTR	-
ALVE_ros007	1,400	-	*env*-3'LTR	-
ALVE_ros008	7,529	+	Intact	100

No sequence information was obtained for ALVE-NSAC3 or the novel ALVE_ros002, ALVE_ros006, ALVE_ros009 or ALVE_ros010. Due to the length of ALVE insertions, both PCR amplification and sequence cloning were difficult. PCR amplification often

resulted in band 'smears' on the gel, or the presence of multiple, varying length bands for the same assay. Due to this, confirmed and purified insert DNA was never obtained for ALVE_ros006, ALVE_ros009 or ALVE_ros010. The diagnostic PCR assays described in section 6.4.2 worked as expected for these ALVEs, so further optimisation of the long-range PCR protocol may enable amplification of these sites. Troubleshooting was attempted, but it is possible that further primer redesigns are required, as was successfully completed with ALVE1.

Cloning success with the intact ALVEs was reduced by 96-99 % compared to test runs with the short, solo LTR ALVE15 sequence, even when the cloning vector reaction was extended to 24 hours from the '5 minute' protocol. For both ALVE-NSAC3 and ALVE_ros002, several repeats of the cloning and transformation protocol yielded only two to five colonies, and it is possible that these were contaminants rather than genuinely transformed colonies. PCR checks of the colony preparations were ambiguous, but sequencing reactions submitted with multiple internal primers as well as the external primer pair produced no sequence. Whilst no sequence information was obtained, preliminary PCRs suggested that all the non-sequenced ALVEs were full length. Some support for this prediction comes from the original characterisation of ALVE-NSAC3 as full length (Smith & Benkel 2008). However, for the novel insertions these are very preliminary suggestions as the PCRs themselves may be part of the problem.

6.6.1 Structure of the incomplete ALVE insertions

Whilst some of the nine completely intact ALVEs contain mutations which disrupt their protein coding potential (section 6.6.3), all nucleotide domains are covered. The six incomplete ALVEs are described below. All ALVE sequences are in Appendix 2: AF09.

Solo LTRs: ALVE15 and ALVE_ros005

Both ALVE15 and ALVE_ros005 are full length, solo LTRs in the negative orientation. ALVE15 is widespread in layers (Benkel 1998), and is within the final intron of the negative strand GRIK2 gene, 800 bp upstream from the final exon. Whilst there is no

literature supporting any impact of ALVE15 on this gene, Ensembl predicts two transcripts for GRIK2, the only difference being the presence or absence of the final 26 amino acid exon. However, there is no available data with which to correlate ALVE15 presence/absence with prevalence of these transcript variants. From InterPro analysis, loss of the final exon does not remove any functional domains or motifs from GRIK2, but there could be impacts on protein structure. Conversely, ALVE_ros005 is intergenic, more than 150 kb from the nearest gene.

Internal deletion: ALVE3

ALVE3 has been well studied due to its expression of both *gag* (Crittenden et al. 1984) and *envelope* (Robinson et al. 1981) glycoproteins. However, ALVE3 is non-autonomous as it lacks approximately 1,370 bp of sequence, encompassing nearly all the *protease* and *reverse transcriptase* domains. Either side of this internal deletion, ALVE3 is largely complete and the *gag-pol* single open reading frame has been maintained, producing a single transcript spanning the entire *gag* domain, *RNaseH* and *integrase*. The expression of the ALVE3 retroviral constituents is likely facilitated by its location in the sixth intron of the HCK gene, however there is no available literature on the impact ALVE3 presence has on HCK expression or splicing.

3' truncation: ALVE9, ALVE-NSAC1 and ALVE_ros007

ALVE9, ALVE-NSAC1 and the novel ALVE_ros007 are all negative orientation insertions which are truncated from their 5' ends, but to varying degrees. ALVE9 is intact from 185 bp into the *protease* domain, ALVE-NSAC1 is intact from 74 bp into the *reverse transcriptase* domain, and ALVE_ros007 is far shorter, intact from 280 bp into the *envelope surface* domain. Whilst all three lack the promoter and enhancers within the 5' LTR, retroviral expression is possible, as ALVE9 expresses high levels of envelope protein, known to inhibit ALVE infection (Robinson et al. 1981). However, ALVE9 is intronic which could facilitate its expression, but ALVE-NSAC1 and ALVE_ros007 are both intergenic.

The truncations for ALVE9 and ALVE-NSAC1 are both clean deletions with no removal of surrounding genomic sequence. ALVE_ros007, however, is associated with a genomic deletion of 1,941 bp (4: 59,843,021-59,844,960). Relative to ALVE1, the first 6,123 bp of ALVE_ros007 has been truncated, so the total deletion event removed 8,064 bp of sequence (Figure 6.21). Whilst there is some evidence of avian constrained sequence in this deleted genomic region, the constrained region is short and has no protein coding potential or promoter/enhancer activity, although the latter is from ChIP-seq data limited to the adult liver (Eory et al., manuscript in preparation).

Figure 6.21 Schematic for the predicted ALVE_ros007 insertion and subsequent genomic deletion. Starting from the wildtype (A) there was an ALVE insertion in a negative orientation (B). There was then a deletion (C; dotted box) of 6,123 bp from the ALVE and 1,941 bp from the host genome, leaving the fragmented insertion of the ALVE 3' LTR and partial *envelope* domain (D).

6.6.2 ALVE LTR alignment and phylogeny

The alignment of all twenty-five sequenced LTRs had 98.6 % identity, and LTRs were all 274 bp long, except the 3'LTR of ALVE_ros007 which was degraded. Of the ten intact ALVEs, nine had LTRs with 100 % pair identity, with the LTRs of ALVEB5 sharing 99.6 % identity due to one SNP (G262T) in the 5' LTR (3'LTR was identical to other ALVE LTRs at this site). All LTRs contained an intact TATA box motif (146-152) and the nearby transcription start site (168-174; U3-R boundary) which was also identified as a binding site for the SRF (serum response factor) transcription factor (Figure 6.22). In addition, a second, upstream SRF site (43-52) was identified in all LTRs.

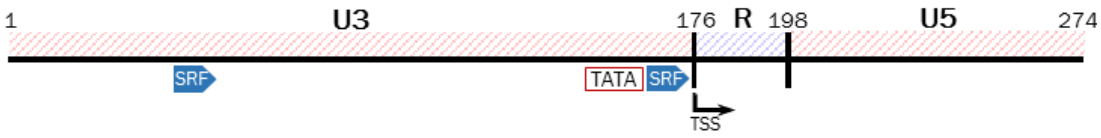


Figure 6.22 The features of the ALVE LTR. LTRs are split into three main sections: U3, R and U5, where the U3-R boundary marks the transcription start site (TSS). The U3 contains enhancer/promoter sequences, including the TATA box motif (146-152) and SRF transcription factor binding sites (43-52, 168-174).

The alignment of all ALVE 3'LTRs, solo LTRs and the ALV-A and ALV-J 3'LTRs had 88.6 % sequence identity. Alignment was best from the TATA box through to the 3' end of the LTR (R-U5 domains). This was expected as the U3 LTR domain is known to be more divergent as well as contain variable promoters, enhancers and transcription factor binding sites that effect tissue-specific expression (Benachenhou et al. 2013). The ALVE U3 domain is also known to be much shorter than the U3 of exogenous ALV LTRs, as ALVEs lack sequences essential for expression enhancement which causes a two to three order of magnitude reduction in overall expression (Norton & Coffin 1987; Conklin 1991; Ruddell 1995). Concordantly, the ALV-A and ALV-J LTRs are 324 bp in length, with all 50 'additional' bases within the U3 region.

However, the exogenous ALV LTRs provide a good outgroup for the ALVE phylogeny (Figure 6.23). The internal ALVE phylogeny itself is poorly resolved, with limited levels of bootstrapping support. This is not helped by the identical ALVE21, ALVE-TYR, ALVE-NSAC1, ALVE_ros004 and ALVE_ros008 LTR sequences, which force the polytomy at the base of the lineage. However, the general view is that the ALVEs originating from the brown egg layers are basal to the white egg layer-specific ALVEs.

LTR alignments were chosen as LTRs were common to all obtained sequences, but the short length does not facilitate well resolved phylogenies. However, even alignments with internal domains, such as a 1.3 kb stretch of *RNaseH* and *integrase* shared between all sequences except ALVE15, ALVE_ros005 and ALVE_ros007, exhibit this same problem. Given their total length, each of the ALVE sequences are very similar, likely due to their recent integration into the genome, and are evolving independently in the different lines. This combination makes a truly representative phylogeny hard to resolve.

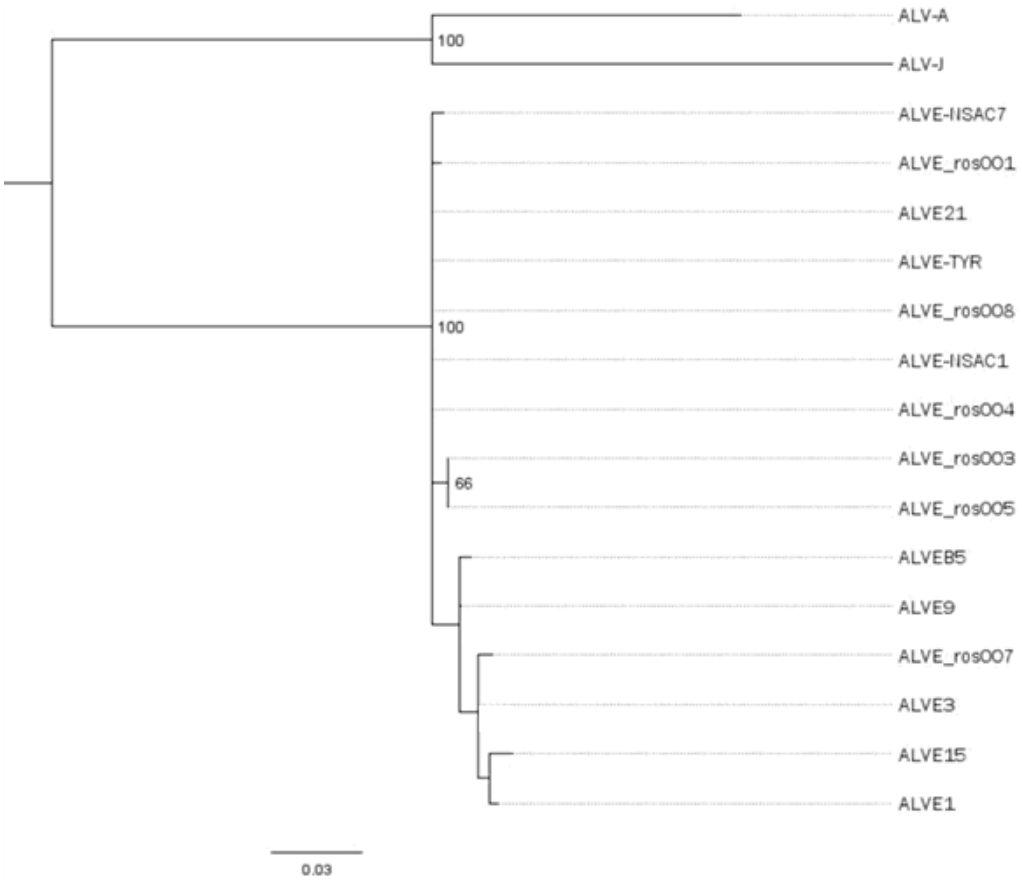


Figure 6.23 ALVE LTR phylogeny with the exogenous ALV-A and ALV-J 3' LTRs forming the outgroup. Bootstrap support greater than 60 % is shown at the relevant nodes, but support is generally limited due to the high identity between the sequences. There are several sites which distinguish the generally brown egg layer ALVEs from the white egg layer ALVEs, but the LTRs have high identity.

6.6.3 ALVE open reading frame integrity and potential expression

Exogenous ALV is transcribed from the 5' LTR transcription start site into two ORFs which are translated, then later cleaved into separate peptides by the retroviral protease. The first ORF is *gag-pol* and the second ORF is *env*, and is usually phased into the next reading frame. However, this compact genomic organisation is vulnerable to frame shift mutations once under the genetic regulation of the host when the retroviral selective constraints are lost. Each of the sequenced ALVEs was annotated for ORFs and their integrity assessed below. The *gag-pol* analysis is summarised in Figure 6.24, and the *env* analysis in Figure 6.25.

ORF1: gag-pol

Six of the ten ALVEs with *gag* nucleotide coverage have one or more mutations in the p10 or p27 domains which truncate any transcripts. In most cases, a second viable ORF is then possible from amino acid 113 within the p27. ALVEB5, ALVE-NSAC7, ALVE_ros001, ALVE_ros003, ALVE_ros004 and ALVE_ros008 all do this with various causative mutations. All these ALVEs, except ALVE_ros003, exhibit a viable ORF up to the various frameshift or nonsense mutations in p10/p27. In addition, ALVE_ros001 has a frameshift mutation after 39 amino acids of integrase, truncating the transcript, and ALVE_ros004 has a frameshift after 72 amino acids of RNaseH with a secondary, 5' truncated *RNaseH-integrase* ORF.

ALVE3 has a full ORF across *gag-pol*, but both the *protease* and *reverse transcriptase* are absent from the element. Likewise, ALVE9 and ALVE-NSAC1 have intact ORFs, but due to their 5' truncation these start at the end of *protease* and 39 amino acids into *reverse transcriptase* respectively. The ALVE1 sequence matches the GenBank reference exactly with a single ORF into the *protease* domain ending with a frameshift mutation. A second ORF overlaps the first but truncates after 214 amino acids into integrase, 100 amino acids before the end of the gag-pol product. ALVE-TYR has a single ORF from *gag* until the very start of *RNaseH*, where there is a frameshift-causing deletion, and a second ORF which begins 53 amino acids into integrase. Only ALVE21 has an intact ORF spanning all the domains.

Due to the number of mutations observed in the p10 and p27 *gag* regions, intact p27 was only detected from ALVE1, ALVE3, ALVE21 and ALVE-TYR. Whilst ALVE1 is not normally expressed (Conklin et al. 1982), ALVE3, ALVE21 and ALVE-TYR are (Gavora et al. 1991; Chang et al. 2006). These four ALVEs are common and, with respect to the HL lines, are either fixed or at high frequencies. Furthermore, as partial p27 coverage was observed at both the 5' and 3' ends of the other six sequenced ALVEs, it is possible that, if translated, partial protein folding could be enough to create the epitope detected by the industry-standard p27 ELISA (Smith et al. 1979).

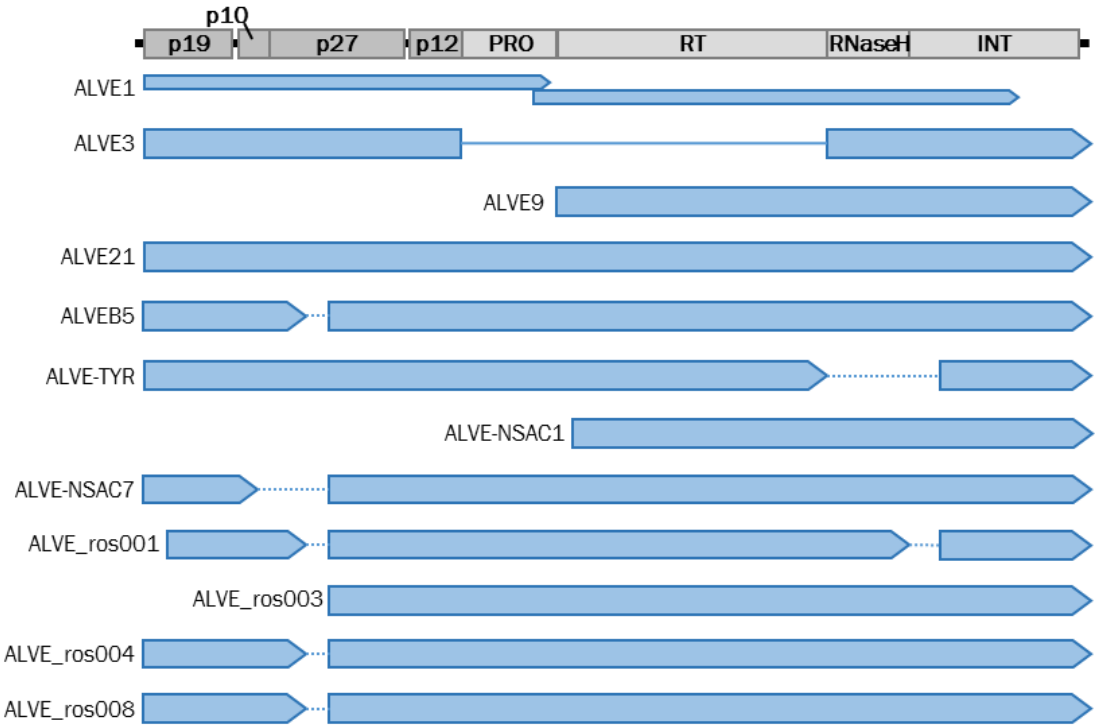


Figure 6.24 Schematic showing ALVE *gag-pol* domains and the open reading frames identified in each of the twelve ALVEs containing *gag* or *pol* sequence. Non-continuous ORFs are shown with dotted lines. The solid line connecting the ORF blocks of ALVE3 shows contiguity but the absence of PRO and RT. Abbreviated *pol* domains are: PRO = *protease*, RT = *reverse transcriptase*, INT = *integrase*.

ORF2: env

Ten of the sequenced ALVEs exhibited intact *envelope* ORFs spanning both domains. These were ALVE1, ALVE3, ALVE9, ALVE21, ALVE-TYR, ALVE-NSAC1, ALVE-NSAC7, ALVE_ros003, ALVE_ros004 and ALVE_ros008. On average, each of these exhibited four to six non-synonymous changes across this region.

ALVE_ros001 had a single ORF which covered most of *envelope*, but truncated 47 amino acids short of the 3' end. Despite this, the transmembrane motif itself was not affected, so the protein may retain function if the cytoplasmic tail is not required for folding. ALVEB5 had two ORFs over the region. The first is truncated so misses 147 amino acids of the surface (SU) domain, but the second ORF covers the entire transmembrane (TM) domain. The largely truncated ALVE_ros007 also has two ORFs. The first starts within 75 bp of the truncated element and covers 67 amino acids of SU

and 33 amino acids of TM before an indel-caused frame shift mutation. The second ORF covers the final 153 TM amino acids.

The majority of sequenced *envelope* appears intact, however expression may be inhibited as all sequences contain the full miR-155 target site (Hu, Zhu, Chen, Liu, Sun, Geng, Wang, et al. 2016).

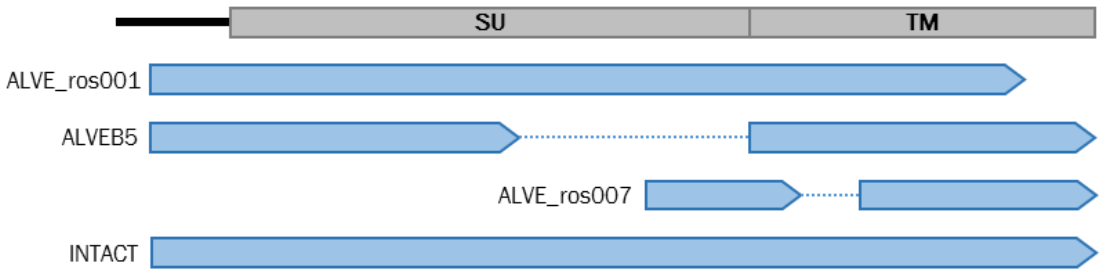


Figure 6.25 ALVE *envelope* schematic showing surface (SU) and transmembrane (TM) domains and the open reading frames identified in the three ALVEs with non-intact *envelope* ORFs compared to the intact ORF.

6.7 The ALVEs of the Roslin J-Line

Three ALVEs were identified in the Roslin J-Line: ALVE15 (3:70,384,294; GTTTAT), ALVE3 (20:10,309,352; AACCAC) and a novel insertion on chromosome 15 named ALVE_ros011 (15:7,599,053; CTCACT). As has been described above, ALVE15 and ALVE3 are within gene introns (GRIK2 intron 16 and HCK intron 6 respectively), but the novel ALVE_ros011 was intergenic. As the J-Line sequencing data was for individual birds, the ALVE frequencies could be estimated directly from the BAM files. ALVE3 was fixed across the individuals, ALVE15 had a frequency of 0.66 and ALVE_ros011of 0.44. There was no observed sequencing bias between ALVE and wildtype alleles.

The insertion detection modelling suggests that, given the sampling of nine individuals from the population of 32, insertions with frequency greater than or equal to 0.36 would be detected 100% of the time. Insert frequencies of 0.125 can be detected 95% of the time, and frequencies of 0.1 would be detected 90% of the time. This means that any rare ALVEs in the JL could have been missed by sequencing only nine individuals. However, sampling more individuals has a limited effect on increasing the confidence

of detecting rarer insertions. Specifically, 95% confidence of detecting an insert with frequency of 0.05 is only achieved by sampling 20 individuals, and 31 individuals would be needed to achieve 95% confidence of identifying any insertion.

6.7.1 PCR assay development and genotyping

New gel-based PCR primers were developed for ALVE_ros011 genotyping. All three designed primers were twenty-two nucleotides long, had GC content of 45 - 55 %, and T_m of approximately 60°C. The common forward primer sequence was 5'-GTTCAGGCTAACCAACAAAACC, the 'no insert' reverse primer was 5'-AGACACTTCACACACCTTGTGC, and the 'insert' reverse primer was 5'-GACAGACCGTTGAGTCCCTAAC and derived from clipped LTR sequence. This design produced bands of 455 bp for no insert and 289 bp when the ALVE was present. Primers and product sizes for ALVE3 and ALVE15 are presented above in Table 6.5.

All flock individuals were tested for the identified ALVEs. ALVE3 was fixed within the line. ALVE15 had a frequency of 0.63, with twelve of the thirty-two individuals homozygous for the insert, and sixteen heterozygote individuals. ALVE_ros011 had a lower frequency of 0.44, also with twelve heterozygotes but only six homozygous individuals. These observed frequencies for the entire flock from 2016 closely matched the frequencies from the 2013 sequencing dataset. Only one individual was homozygous for all three inserts, and only one lacked both the ALVE15 and ALVE_ros011 inserts.

6.7.2 Comparison of results to the J-Line pool sequenced for the 600K paper

In 2008, ten J-Line individuals were used for an Illumina sequencing pool used in the development of the 600K SNP array (Kranis et al. 2013). Analysis of this dataset with the ALVE identification pipeline identified ALVE3, with the BAM file suggesting the ALVE was homozygous in all individuals, and ALVE15, where there was clear evidence of 'no insert' alleles in the population. However, ALVE_ros011 was not detected.

Ten sequencing reads covered the insertion site of ALVE_ros011 (average genome-wide coverage was 11.27X) with none supporting an ALVE insertion. As the observed

ALVE_ros011 frequency was 0.44 in both the 2013 individual sequencing data and full flock 2016 samples, it is unlikely that the insertion could have arisen and reached this frequency within five years, particularly with no observed increase between 2013 and 2016. However, in 2016 there were ten individuals which were homozygous for no insert and sixteen heterozygotes, so, by a chance, a higher proportion of individuals lacking the insertion could have been chosen for the pool, but only a 4.27×10^{-13}% chance it could have been missed altogether.

Genotyping of the original DNA samples used for sequencing revealed that, again, ALVE3 was fixed in the line, but both ALVE15 and ALVE_ros011 had a much lower frequency in these individuals than was observed in the more recent sampling. ALVE15 frequency was 0.40, with one homozygous individual and six heterozygotes, and ALVE_ros011 frequency was just 0.15, with one homozygous individual and one heterozygote. ALVE_ros011 frequency in the pool was therefore approximately only a third of the predicted line frequency, making it very likely that those alleles were not sampled for sequencing. Analysis of this site with the pooled data model for detection probability suggested that there was a 64.7 % chance of missing this insertion.

6.8 Discussion

The new ALVE identification pipeline developed here has been successfully used to identify twenty-one different ALVEs across nine chicken layer lines, without the need for additional targeted sequencing. Of these, six were novel to this study. The white egg layer Leghorn lines (including the Roslin J-Line) had two to four ALVEs. As expected (Sabour et al. 1992; Benkel 1998), the three brown egg layer lines had more identified ALVEs, some of which have been previously identified in broiler lines. The WPR sister lines had eight and nine ALVEs (seven of which were shared), and the RIR had eleven.

Diagnostic assays were developed for each of the ALVEs and used to obtain insert frequencies for multiple generations of each line. ALVE1 was found to be fixed in three WLs (1, 2 and 5), ALVE3 was fixed in the JL, ALVE9 was fixed in WL3, ALVE21 was fixed in WL4 and both WPRs, and, as predicted, ALVE-TYR was fixed in both WPRs. No ALVEs were fixed in the RIR. Most ALVEs had variable frequencies which

fluctuated randomly over time. It is possible that ALVEs which were rare at population level could have been missed from the identification (discussed below).

Five previously characterised ALVEs (ALVE1, ALVE3, ALVE9, ALVE15 and ALVE-TYR) were intronic insertions, but no novel ALVEs were found within or near genes. There was no observable insertion site bias related to GC content, and no patterns were observed in the insertion site hexamer sequences.

Full ALVE sequence was obtained for fifteen of the twenty Hy-Line insertions. Two of these (ALVE15 and ALVE_ros005) were solo LTRs, and three (ALVE9, ALVE-NSAC1 and ALVE_ros007) had varying 5' truncations. As previously reported in the literature, the sequence for ALVE3 was intact except for *protease* and *reverse transcriptase* domains. The other nine sequenced ALVEs were full length, but varied in terms of their domain intactness when ORFs were predicted. Specifically, most ALVEs exhibited mutations within the p10 or p27 *gag* domains which disrupted putative ORFs, however most appeared to contain an intact *envelope* domain. The expression status of the ALVEs remains unknown, but all identified *envelope* domains contained the target site for the miR-155 microRNA which marks these transcripts for degradation before translation. The five non-sequenced ALVEs were also likely to be full length, although this is based purely on band length predictions from problematic long range PCRs.

The completeness of ALVE identification, wider use of the diagnostic assays, and the relevance of this data for commercial improvement are discussed below.

6.8.1 Critical assessment of the ALVE identification pipeline

The preliminary analyses attempted with published viral insertion detection software involved the installation of many dependencies (BioPerl, BioPython *etc.*). Furthermore, the programs were difficult to customise beyond the hard-coded settings for analysis of the human genome for certain viral genera. In contrast, the new ALVE identification pipeline developed here makes use of common NGS data manipulation software such as samtools, and scripts were written in BASH and Python 2.7, without reliance on modules outside the standard Python library. This has been done to make analysis as easy, user-friendly and adaptable as possible. Developed scripts have logical names and

include clear documentation, built in help messages, and output file name management. Additionally, multiple analyses can run in parallel, and a complete analysis of a single WGS chicken dataset typically takes less than two days.

The success of the pipeline is limited by the completeness of the reference genome and the average coverage of the sequencing data. For example, ALVE6 was not detected in any of the nine datasets, despite being a common ALVE in layers and having previously been found to have a high frequency in WPR1 from gel-based PCRs performed in the Hy-Line Molecular Genetics laboratory. However, ALVE6 is located at the very 5' end of the assembled chromosome 1 sequence and is poorly covered by short read sequencing projects, prohibiting identification of incongruently mapped reads. This would likely be an issue for any ALVE found within telomeres, centromeres or other incompletely assembled genomic regions. Even when assembled in the reference genome, regions which are difficult to sequence often have a much lower coverage than the genome-wide average. Reduced coverage can lead to putative insertion sites not reaching set support thresholds, or insert allele sequence being missed altogether. Initially ALVE insertions were not called by the pipeline unless there was both 5' and 3' clipped support, but this reduced the number of identified sites as low frequency inserts in regions of low coverage may only have support for one end of the insert. Altering this parameter did not increase false positive 'noise' in the results, but did enable the identification of thirty-six ALVE instances rather than the original twenty-seven.

Missing ALVEs due to sampling

Testing of all the Hy-Line lines with the KASP assays developed in this project identified five additional ALVE instances which were missed by the pipeline. Three of these could not have been detected as the ALVE was not present in any individual chosen for the sequencing pool. In the case of ALVE_ros004 in the WPRs, this was because the insert was rare in the population. However, ALVE_ros010 was missed in WPR1 despite being more common, with there being a 75 % chance that at least one heterozygous individual was included in the pool. This highlights the issue with sampling a population for the detection of rarer inserts, and this would not be improved by using individual sequencing libraries rather than pools, or by increasing ALVE coverage through additional targeted

sequencing. Experimental design for sequencing is therefore vital, as researchers and companies need to consider the lower limit of frequencies they need to detect reliably and then sample accordingly, particularly if flock sizes are large.

The other two missed ALVEs (ALVE3 in the RIR, and ALVEB5 in WPR1) were present in the sequencing pool individuals, but no ALVE reads were detected at the sites, even in the full BAM file. This is an issue created by using sequencing pools for variant discovery, as not all individuals will be represented at each site (depending on coverage) and some alleles will be missed entirely due to random amplification in the PCR stage of sequencing library preparation. These issues can be mediated by increasing coverage in the pooled data or, more comprehensively, by using individual sequencing libraries rather than pools. This is again a case of experimental design. However, just as sequencing has almost completely moved from single end to paired end data, new WGS projects are now commonly completed using individual sequencing libraries, likely due to reductions in sequencing cost. These data generally exhibit higher coverage than older pool datasets, and in areas of lower coverage, there are only two alleles to represent.

These two issues combined means that it is possible that ALVEs which were relatively rare in a line, or were line-specific, could have been missed during this project. This was an issue with the analysed WGS data, rather than with the identification pipeline itself.

The role of targeted ALVE sequencing

Additional targeted sequencing is no better for identifying ALVE insertions than using data obtained from individual sequencing libraries with standard coverage. Both methods are affected by sampling bias from the flock, but if the sequencing data was to be used solely for the identification of ALVE insertions, targeted sequencing projects would be more cost effective. However, datasets from targeted sequencing are very specific and have limited extra uses, whereas full genome, individual WGS data can be used for a variety of research questions. Targeted ALVE sequencing may be useful to groups with existing pooled sequencing project data who want to more confidently identify the ALVEs in their lines.

Wider application of the ALVE identification pipeline

Beyond the Hy-Line elite layer lines and Roslin J-Line, the pipeline has been used in this project to analyse other chicken datasets to more completely assess ALVE diversity across various lines (Chapter 7). The pipeline was also adapted to perform a preliminary identification of the more numerous EAV insertions in the Hy-Line lines by altering the sequences in the pseudochromosome used for initial mapping.

The pipeline is very versatile and could be used for the identification of any viral insertion in any species relative to its reference genome. This would simply require the user to create a FASTA file of one or more viral reference sequences. This file would be used to identify assembled sites in the reference genome (S1_run_blast_ref_seq.sh) and to construct the pseudochromosome (S2_make_pseudochromosome.py). The pipeline could therefore be used for the identification of any endogenous viral elements (EVEs), including those responsible for causing cancers in other species, such as human. As mentioned above, any application would be limited by reference genome intactness and coverage. The pipeline may therefore be unable to identify telomeric insertions by herpesviruses such as Marek's Disease Virus in chickens.

6.8.2 Development of diagnostic ALVE assays

Over the last twenty-five years diagnostic gel-based PCR assays have been developed for most well described ALVE insertion sites, generally enabling reliable genotyping. Exceptions to this included ALVE2, ALVE6, ALVEB4 and ALVEB8, where published assays could only determine presence or absence of the ALVE without distinguishing between heterozygous or homozygous individuals, and ALVE21 where results were ambiguous due to the presence of the *K* locus duplication (Benkel 1998). In this project, twelve new, gel-based PCR assays were developed, including an adapted version of the previously published ALVE1 assay to account for the large difference in T_m values between primers. Assays developed for long ALVE insertions used three primers, where one was within the insertion, enabling unambiguous genotyping. Assays for the truncated ALVE_ros007 and solo LTR ALVE_ros005 used only two primers, as standard PCR conditions can amplify through these insert lengths.

Gel-based PCR assays continue to be a laboratory standard, as they are generally easy to run and interpret, and are applicable across diverse lines, breeds and populations. However, when used on large commercial flocks, gel-based assays become laborious and expensive, making them an impractical large-scale genotyping tool. The KASP diagnostic assay system was originally developed for high-throughput, inexpensive SNP genotyping, and exhibits limited interference from negative controls and low error rates compared to chip-based genotyping (Semagn et al. 2014). Traditionally the KASP system has not been used for genotyping large structural variants, but in this project we have successfully developed diagnostic KASP assays for the twenty ALVEs identified in the Hy-Line elite layer lines, and an assay to detect the unique sequence in the *K* locus duplication. The development of these high throughput assays enabled the largely automated genotyping of almost ten thousand individuals for all twenty-one variants. This would have been incredibly expensive and time consuming with gel-based assays.

Whilst the KASP system works best with large numbers of samples (for genotype group confidence), it can also be performed with small sample numbers on a qPCR machine. This means that the assays developed here could be used on other chicken lines, but lineage-specific SNPs in the primer binding regions would likely disrupt the assay. This is an unavoidable issue resulting from the original purpose of this technology as a SNP genotyper, with chemistry sensitive to ambiguities in primer binding regions. This has the potential to limit assay application to lines with available sequence data, where the primers could be checked and modified where applicable. This was observed when these assays were applied to non-sequenced Hy-Line lines (data not shown), and even within those lines with available sequence data (ALVE_ros005; section 6.4.1). However, as KASP assays are generally used for high throughput genotyping, it is likely that sequence data would be available for the lines of interest, or obtaining whole genome resequencing data would be a viable option. In addition, as the ALVE_ros005 work showed, targeted Sanger sequencing around a known insertion site in 'problematic' samples can be used to advise modifications in existing assay primers.

6.8.3 ALVE21 and the slow feathering *K* locus

The slow feathering locus has provided commercial breeding companies with a cost-effective method for sexing day old chicks. However, the close association of this locus with ALVE21 renders breeding for slow feathered birds an imperfect solution, as ALVE21 is replication competent and known to increase mortality rates, facilitate viral shedding, and detrimentally effect productivity traits such as muscle growth rate and total egg count (Smith et al. 1990a; Smith et al. 1990b; Fadly & Smith 1991; Gavora et al. 1995; Hamoen et al. 2001; Khosravinia 2009). Full characterisation of this locus, and the development of diagnostic assays, has been hindered due to the large tandem duplication in the causative slow feathered ***K*** allele (Levin & Smith 1990).

In this project, two high-throughput diagnostic KASP assays have been developed to facilitate unambiguous ALVE21 and ***K*** locus genotyping when used in tandem. As discussed above, it is possible that these assays may not be applicable to all commercial stock due to SNPs or indels. This is particularly relevant for the duplication bridge assay, as multiple slow feathering phenotype variants have been observed, suggesting that the extent of the underlying duplication may differ between lines (Iraqi & Smith 1995; Wimmers et al. 1996; Tixier-Boichard & Boulliou-Robic 1997; Kansaku et al. 2011). If application of the assay developed here gave unexpected results across different lines, it would be prudent to fully characterise the bridging location following the methodology previously used to identify the site in layers (Elferink et al. 2008; Bu et al. 2013).

It was disappointing that the BioNano high resolution optic mapping generated during this project was unable to fully characterise the duplicated region in the ***K*** locus due to insufficient data. However, further analysis is underway and the provision of more data by the Earlham Institute should facilitate the full elucidation of the locus and identification of the ALVE21 site which is occupied. In addition, as the new BioNano Access software becomes more widely used it will be much easier to analyse the data.

Optic maps currently represent the only technology capable of fully characterising a variant as long as the 180 kb duplication within the ***K*** allele. Complete characterisation would advise breeding companies on how to best mitigate the ALVE21 insertion, perhaps through CRISPR/Cas9-mediated deletions (discussed below in section 6.8.4). However, another consideration is that the partially duplicated genes within the ***K*** allele

may also contribute to the detrimental effects on productivity traits, particularly as PRLR SNPs have previously been linked, in wildtype fast feathering birds, to reductions in layer success (Zhang et al. 2012). Recent studies have identified that both PRLR and SPEF2 mRNA levels were elevated in slow feathered individuals (with no evidence of antisense RNA interference), and that the duplicate transcripts exhibited a similar, and very broad, spatiotemporal distribution (Luo et al. 2012; Bu et al. 2013; Zhao et al. 2016). This may reflect a diverse range of phenotypic effects due the slow feathering locus, and it may be better for commercial breeders to pursue a different sex marker.

WPR2 is a fast feathering revertant line

Development of the ALVE identification pipeline was facilitated by the positive controls of known associations between ALVEs and specific phenotypes. Both WPR lines were known to contain the recessive white mutation caused by ALVE-TYR, and the slow feathered WPR1 and WL4 were expected to contain ALVE21. It was therefore surprising that WPR2, the fast feathered sister line to WPR1, was also fixed for ALVE21. This finding was validated by manual inspection of the genome alignment maps (BAM files), positive ALVE21 KASP assay results, and the visualisation of the insertion using BioNano optic mapping. However, the fast feathered phenotype was also validated by the homozygous ALVE21 KASP result, the negative ***K*** locus bridging KASP result, and the absence of reads mapping through the insertion site.

Together, these results support a fast feathered revertant origin for the WPR2 line. The WPR sister lines were separated based on their feathering rates, and it is therefore likely that the ancestral WPR line was slow feathering but the occasional fast feathered revertant made it seem like the line was still segregating. WPR2 therefore contains an ALVE insertion which has a wide range of detrimental phenotypic effects, including on productivity traits, but also lacks the commercially useful slow feathering allele. This result should be considered by Hy-Line when planning future breeding programmes.

6.8.4 The commercial response to ALVE loci

This work has characterised the ALVE content of eight Hy-Line elite layer lines. Whilst it is possible that line-specific ALVEs may have been missed due to population sampling or incomplete allele amplification, this data enables Hy-Line to develop a management programme for these variants. Some possibilities and considerations are explored below.

Evidence for current ALVE-related selection

Multi-generational KASP assay genotyping has shown that the frequencies of most ALVEs are not changing in a directional manner. This is unsurprising, as there has been no targeted breeding programme designed for the elimination of ALVEs, and selection against p27 limits selective pressure to ALVEs with intact *gag* expressed at the point of testing (prior to their selection as breeder birds). Most ALVE frequencies are fluctuating randomly due to drift, which means ALVEs at rare frequencies could be lost (*e.g.* ALVE_ros004 in either WPR line), or those at very high frequencies could become fixed (*e.g.* ALVE15 in WL1).

Only ALVE3 had any directional selective effect, with frequency increasing in all three WLs (WL2, WL3, WL4) inheriting this insertion, but only random fluctuations were observed in the RIR. ALVE3 is a well described element, and it has been long known to express both gag and envelope proteins, with the latter at a particularly high titre (Astrin & Robinson 1979; Robinson et al. 1981). Concordantly, ALVE3 does elicit a regulatory effect on infection by both exogenous and endogenous ALV through receptor interference (Robinson et al. 1981; Smith et al. 1990a; Smith et al. 1990b). However, the HL lines are not under recurrent ALV selective pressure and ALVE3 produces p27, so it is unlikely that ALVE3 is being directly selected for, particularly as this is not seen in the RIR. It is possible that the proximity of ALVE3 to two genes with immune roles relevant to MDV infection success (which is under direct selection) has resulted in increasing frequency due to linkage disequilibrium. This effect may have been broken in the RIR, rather than selection specifically differing between the two breeds.

The impact of selecting against p27 expression may also be apparent from the general *gag* sequence degradation observed in six of the structurally intact sequenced ALVEs.

Assessing ALVE phenotypic effects

The high structural identity of many of the sequenced ALVEs suggests that they may retain the ability to retrotranspose, even if individual ALVE insertions do not contain all retroviral domains. For example, any line containing expressed ALVE21 would be able to retrotranspose any other ALVE insertion. ALVE polymerase proteins could also facilitate the movement of other autonomous and non-autonomous retrotransposons, increasing the potential for insertional mutagenesis in the individual bird. However, as the steady-state methylation of all the insertions is unknown, it remains unclear how likely retrotransposition is under normal conditions. During early embryogenesis (when methylation patterns are removed) ALVEs may be able to retrotranspose, although recent work has detected piRNA-mediated defence against ALVE retrotransposition in domesticated chickens (Sun et al. 2017). Any full length ALVE or intact ALVE domain poses a future genomic threat for recombination, retrotransposition or reactivation of expression after line crosses or epigenetic modifications.

The full expression status for many of the identified ALVEs is unknown, despite the sequence data obtained in this study. Structurally intact sequences may not be expressed due to methylation or genomic location, but this cannot be fully or uniquely elucidated without expression data in a range of tissues from a range of developmental stages. Due to high sequence identity between ALVE insertions, any transcriptomics data would need to be long read sequencing, such as PacBio, to uniquely describe expression levels. It would also be pertinent to characterise microRNA expression, as many of the sequenced ALVE *envelope* domains are intact at the nucleotide level, but contain the miR-155 recognition sequence which would target these transcripts for degradation prior to translation. Additional, novel microRNAs may be present which regulate expression from other ALVE domains.

Even if an ALVE is not expressed, presence of an insertion alone may elicit a phenotypic effect. For example, an ALVE LTR insertion modulates the expression profile of the *aromatase* gene, causing the *henny-feathering* mutation (Matsumine et al. 1991), and the insertion of ALVE-TYR into the final intron of the *tyrosinase* gene causes truncation of the final exon, producing the recessive white mutation (Chang et al. 2006; Chang et al. 2007). The phenotypic effect of the five intronic insertions identified in the Hy-Line

elite layer lines have been well described, so further analysis may not be necessary, but this should be a consideration for any novel ALVEs identified in future.

Even with this extra data on ALVE activity and local impact, a full association analysis with phenotypic data would be needed to assess the individual and cumulative effects of the ALVEs on the host. Hy-Line collects phenotypic data for many commercially relevant productivity traits, and all individuals can now be genotyped for their ALVE content using the KASP assays from this work. This kind of study may identify specific ALVE insertions as priorities for eradication from the lines. However, the effect of any ALVE which is fixed within a line cannot be detected as there is no variation at that site.

Methods for ALVE eradication from the Hy-Line elite layer lines

ALVE insertions pose current and future threats, but as non-essential genomic components (Zhang et al. 2008), they can be eliminated from the genome without negative developmental or physiological effects. In commercial lines which are otherwise ALV free, this would likely halt any further ALVE integrations.

Traditional breeding methods could be used to gradually reduce ALVE frequency in each line, focusing on those with greatest predicted effects in an association analysis, or those inserts which are structurally intact or in regions of the genome where they elicit a phenotypic effect. However, such methods are slow and would have to be integrated within existing selective breeding programmes. Any fixed ALVE could also not be removed in this manner, requiring crosses with other lines which may create varied, undesirable phenotypic effects. The CRISPR/Cas9 system makes it possible to perform targeted genetic modification without the need for generations of selective breeding. Large, targeted deletions are possible (Zhou et al. 2014), and recent work has made CRISPR/Cas9 possible in chickens following initial problems with accessing zygotes without disturbing development (Dimitrov et al. 2016; Oishi et al. 2016).

CRISPR/Cas9-mediated deletions could facilitate the removal of all ALVEs in a single generation, but, for commercial stock, this would need to be managed correctly to maintain genetic diversity within the flock. Furthermore, recent studies into the public perception of genetically modified (GM) organisms show that people remain concerned

about the development and consumption of GM animals ((Bawa & Anilakumar 2013; Frewer et al. 2013; Frewer et al. 2014; Lucht 2015)), and only the AquAdvantage salmon has been approved for consumption to date (United States Food and Drug Administration, 2015). However, the CRISPR/Cas9 system would provide a particularly neat solution to the negative effects of ALVE21, as the retrovirus could be removed without disrupting the slow feathering phenotype. A similar approach could also be used with ALVE-TYR, and the final exon could also be removed if the white feathering phenotype was an essential requirement.

6.9 Concluding remarks

The bioinformatic pipeline developed here has enabled the identification of ALVE insertions using existing WGS datasets. Compared with WGS datasets from pools, it is likely that ALVE-specific target capture sequencing would facilitate the detection of rare insertions. However, WGS data generated from individuals is just as sensitive for ALVE detection, and has a much wider application. Identification and characterisation of ALVEs in the Hy-Line elite layer lines will also enable the development of a breeding programme to remove ALVE insertions, using traditional or modern techniques. This would likely improve the productivity of the birds, resulting in greater commercial gains, and improve the flock-level animal welfare.

Within this project, this pipeline has enabled the detection of ALVEs in a wide range of chicken subpopulations, not just layers, and has facilitated the widest characterisation of ALVE diversity to date (Chapter 7). Beyond its application for chicken research, this approach could be used to identify any viral integration into genomic DNA, in any species. This has great potential for use in diagnosing oncornavirus-induced cancers, such as human adult T cell leukaemia (caused by Human T-Lymphotropic Virus; HTLV), and aiding the development of personalised therapeutics.

Chapter 7: Discovery of the wider diversity ALVE insertions across commercial and non-commercial chickens using whole genome (re)sequencing data

7.1 Introduction

At the end of the 1980s White Leghorn (WL) chickens had been extensively characterised for their ALVE content with at least twenty-three independent insertions identified. Research initially focused on these prominent commercial layer chickens due to the well described detrimental effects on egg laying success (Gavora et al. 1991). Detrimental effects of ALVEs had also been identified in brown egg and broiler chickens, albeit indirectly, when it was found that the recessive white mutation (later found to be caused by ALVE-TYR; Chang et al. 2006) was linked to viremia and reductions in muscle growth rate and total muscle mass (Fox & Smyth 1985). However, targeted study of broiler ALVEs was hindered by their greater number when compared to WLs. A series of studies in the early 1990s found that whilst some ALVEs were shared between white and brown egg layers and broilers (notably ALVE3 and ALVE6) many ALVEs were novel and there was no suggestion of a clear ancestral ALVE complement (Gudkov et al. 1986; Iraqi et al. 1991; Sabour et al. 1992; Grunder et al. 1995). Furthermore, the standard method for ALVE classification at the time, restriction fragment length polymorphisms (RFLPs), became harder to interpret with greater numbers of ALVEs, especially after RFLPs were found to vary between breeds for the same insertions (Aarts et al. 1991; Boulliou et al. 1991).

Despite these issues with identification, it was generally found that white egg layers contained one to three ALVEs, and brown egg layers and broilers contained six to ten (Benkel 1998). However, ALVE research has largely utilised experimental lines held at research institutes and universities. This has biased work towards leghorns (as the eggs could, originally, be sold on to recoup research costs) and heritage breeds, rather than commercially relevant lines. These birds are also highly inbred, so ALVEs considered 'common' (such as ALVE1 or ALVE3) are likely to become fixed, promoting their common status in the literature, and rarer ALVEs are more likely to be lost, reducing apparent diversity. It is therefore unclear whether these generally accepted numbers of ALVEs for layers and broilers are accurate. The work presented in chapter 6 highlighted

this, as the WLs had two to four ALVEs and the brown egg layer WPRs both had eight and the RIR had eleven. Whilst these elite Hy-Line layer lines have undergone selection to mitigate the impacts of ALVEs, they have a much higher effective population size than any university population, and are likely more representative of ALVE diversity in these breeds. The larger effective population sizes of commercial broilers, non-commercial 'wild' chickens, and RJF populations may elicit larger numbers of ALVE insertions, particularly in the absence of intensive artificial selection. However, broiler genetics are rarely made public due to fears about commercial sensitivities, and non-commercial birds and RJFs have not been analysed to date. It is highly likely that the presence of just two ALVEs in the chicken reference genome is significantly under-representative of the wider ALVE diversity (Benkel & Rutherford 2014).

In 2016 Rutherford and colleagues used their ALVE target capture NGS protocol to analyse over thirty European and North American heritage breeds. A total of 137 novel ALVE insertions were identified, tripling the existing numbers in the literature. However, the development of the ALVE identification pipeline (Chapter 6) enables analysis of any existing WGS dataset. Whilst commercial broiler data remain unavailable, WGS data for experimental lines, heritage broilers, broiler crosses, 'wild' chickens and RJF individuals is publicly available in short read archives, or has been kindly shared by collaborators. The analysis of these datasets, presented below, has enabled a much more thorough characterisation of ALVE diversity across chickens, without any further targeted sequencing.

7.2 Research Aims

This chapter covers three major research aims. Firstly, the expansion of the ALVE identification pipeline developed in Chapter 6 to identify ALVE insertions in publicly available single end WGS data. Secondly, the identification of ALVEs in WGS data from a wide range of chicken lines, including layers, broilers, 'wild' non-commercial birds, and RJF datasets, to identify trends in ALVE distribution and determine how artificial selection has affected ALVE diversity and abundance in commercial lines. Finally, an assessment of a phylogeny generated by using ALVEs as genetic markers.

7.3 ALVE identification pipeline adaptation for single end WGS data

7.3.1 New pipeline scripts for use with single end data

The majority of changes required to make the ALVE identification pipeline applicable for use with single end sequencing data were simply alterations to script input/output: to expect a single FASTQ file rather than two. The only change made to the parameters was to reduce the minimum region length required for an insert to be identified from 200 bp in the paired end data to 80 bp with single end data. This reduction accounts for the fact that single end data will have no read mates to increase the area of interest around the ALVE hexamer. As significant pseudochromosome mapping will not occur with less than 20 bp of ALVE homologous sequence, some sites could be represented by just 80 bp of reference genome mapped sequence.

For user ease, separate scripts were created for use with single end data (Table 7.1; Appendix 1). However, as already stated, and explained below, the changes were minor and Figure 7.1 shows how parallel the pipelines remain. se_S1_run_bwa_alignment.sh replaced both S3_run_bwa_alignment.sh and S5_run_bwa_alignment.sh and was altered so that the BWA mapping only accepted one FASTQ file. This changed the user specification (lines 2 and 6), how the output file prefixes are generated (line 11) and the BWA mem command (line 13). se_S2_extract_ref_seq_mapped_reads.sh replaced S4_extract_ref_seq_mapped_reads.sh with the only alteration made to output a single FASTQ file following read subtraction (line 30). se_S3_extract_putative_sites.py replaced S6_extract_putative_sites.py and included the reduction of minimum required region length to 80bp (line 28). All additional scripts are on the CD accompanying this thesis (Appendix 1) and in the GitHub ALVE identification pipeline repository: https://github.com/andrewstephenmason/ALVE_ID_pipeline.

Table 7.1 ALVE identification pipeline scripts for use with single end WGS data.

Script name	Functionality
se_S1_run_bwa_alignment.sh	Maps sequencing reads to i) the pseudochromosome and ii) the genome
se_S2_extract_ref_seq_mapped_reads.sh	Extracts viral mapped reads
se_S3_extract_putative_sites.py	Identifies putative insertion sites

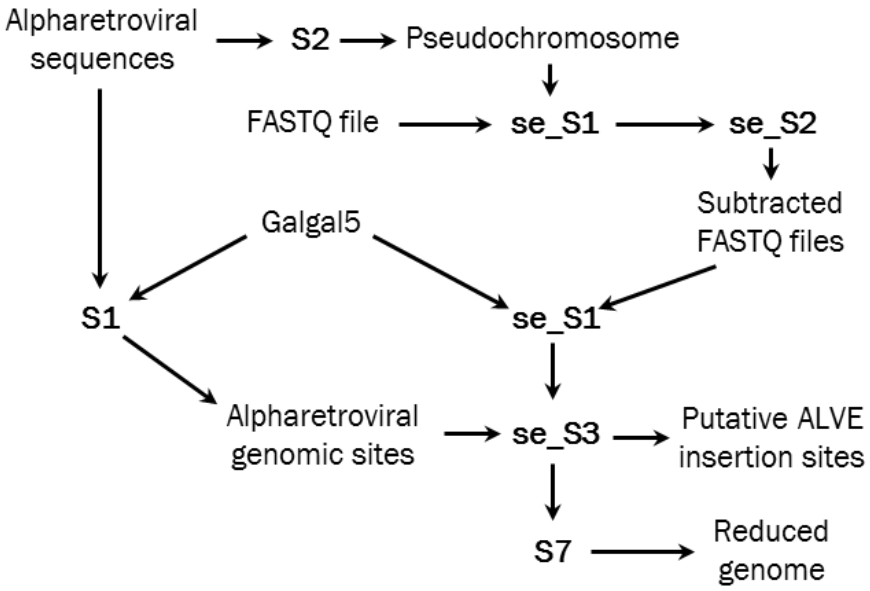


Figure 7.1 ALVE identification pipeline workflow showing how the single end (se) WGS data scripts fit. The schematic is based on Figure 6.1 using script abbreviations from Tables 6.1 and 7.1. Scripts S1, S2 and S7 remain the same.

7.3.2 Assessing pipeline sensitivity using pseudo single end FASTQ files derived from the Hy-Line and J-Line paired end sequencing data

Impact of single end reads on read mapping and overall coverage

BAM files produced by mapping the pseudo single end FASTQ files to the reference genome had a significantly greater genome-wide average coverage than those generated from the original paired end FASTQ files for both the Hy-Line (repeated measures 't' test; $t = 5.88$; $P = 6.14 \times 10^{-4}$) and J-Line (repeated measures 't' test; $t = 10.97$; $P < 1 \times 10^{-5}$) datasets. Consequently, read coverage was deemed unlikely to cause any problems in the detection of ALVEs within the datasets.

Use of pseudo single end FASTQ files with the Hy-Line and J-Line datasets

Thirty-five instances of ALVEs were identified across the eight Hy-Line lines when the original paired end sequencing data was used in the identification pipeline. However, use of the pseudo single end FASTQ files missed fourteen of those instances (40 %), with only lines WL1 and WL3 having all ALVEs identified (Table 7.2).

Table 7.2 Identified and 'missed' ALVEs using the pseudo single end WGS data for each of the eight Hy-Line lines. The majority (21/35) of ALVE instances were identified, but eight of the individual ALVEs were missed completely.

Line	Identified ALVEs	Missed ALVEs
WL1	ALVE1, ALVE15	-
WL2	ALVE1, ALVE3	ALVE15
WL3	ALVE1, ALVE3, ALVE9	-
WL4	ALVE1, ALVE3, ALVE21	ALVE_ros008
WL5	ALVE1	ALVE15
WPR1	ALVE-TYR, ALVE-NSAC1, ALVE-NSAC7	ALVE-NSAC3, ALVE21
WPR2	ALVEB5, ALVE-TYR, ALVE21, ALVE-NSAC3, ALVE-NSAC7	ALVE_ros009
RIR	ALVE_ros004, ALVE_ros010	ALVEB5, ALVE-NSAC1, ALVE_ros001, ALVE_ros002, ALVE_ros003, ALVE_ros005, ALVE_ros006, ALVE_ros007

Whereas most ALVEs were identified, use of single end data clearly reduced the sensitivity of the identification pipeline. For example, WPR1 had five identified ALVEs, three of which were identified (ALVE-TYR, ALVE-NSAC1, ALVE-NSAC7), and two were missed (ALVE-NSAC3, ALVE21). Even with the three ALVEs that were identified, comparison of the insertion region BAM files outputted from the identification pipeline shows a reduction in both the coverage and clipped support for the insertion site (Figure 7.2). There was a complete absence of reads which have only short (< 25 bp) clipped sections, as these were too short for significant mapping to occur on the pseudochromosome. These reads were observed in the paired end data as their read mates mapped within the insert, so both reads were retained during read subtraction. In comparison, the two ALVEs that were missed had limited support in the paired end data (Figure 7.3), and further reduction in read support caused these sites to be missed by the pipeline. Both ALVE-NSAC3 and ALVE21 were identifiable in the pipeline output, but did not reach the required threshold level.

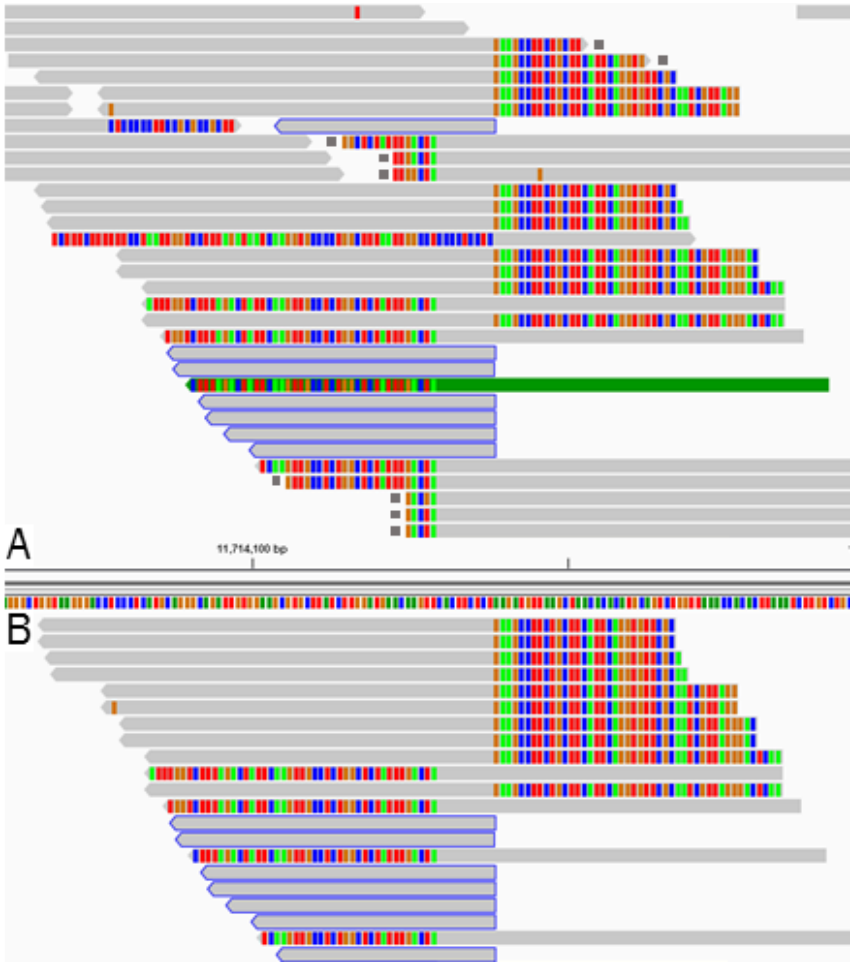


Figure 7.2 Read coverage of the ALVE-NSAC7 insertion site in Hy-Line WPR1 using the paired end (A) and pseudo single end (B) WGS data. ALVE-NSAC7 was detected from both analyses, but the read coverage was reduced in B. All reads with soft-clipped regions less than 25 bp (marked with grey squares in A) were not in the final filtered BAM files. These reads were found in A as their read pair mapped completely to the pseudochromosome.

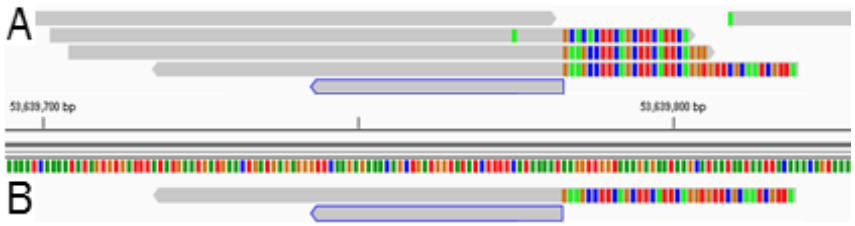


Figure 7.3 Read coverage of the ALVE-NSAC3 insertion site in Hy-Line WPR1 using the paired end (A) and pseudo single end (B) WGS data. ALVE-NSAC3 was missed due to the limited coverage in the pseudo single end data. The two soft-clipped reads absent in B have only 20 and 23 soft-clipped bases respectively.

When this approach was applied to the other five Hy-Line lines with 'missing' ALVEs (WL2, WL4, WL5, WPR2, RIR), all instances were identified but simply did not meet the threshold level. This was particularly problematic in cases where the paired end data only supported one end of the insert site with soft-clipped reads, such as ALVE_ros002, ALVE_ros008 and ALVE_ros009, all of which were detrimentally effected by the loss of reads with only short soft clipped sections. Furthermore, the complex RIR insertion ALVE_ros007 only had one read supporting the 3' end of the insert in the single end data, so from this information alone, it would have been impossible to identify that this ALVE is associated with the genomic deletion described above (section 6.6).

In contrast, all ALVEs identified in the JL individuals using paired end data were identified successfully using the single end sequencing data and the adapted pipeline scripts. Like the HL data, there was loss of reads with only short sections homologous to ALVE sequences, but the overall higher coverage of the JL datasets reduced the chance effect that the only supporting reads had short clipped sections. Additionally, as the JL samples were of individuals rather than pools, the potential for sampling bias reducing the representation of insert alleles in the dataset was less of a concern.

7.4 ALVE content of diverse chicken WGS datasets

A total of 322 different ALVEs were identified across the analysed datasets, including those identified in the Hy-Line elite layer lines and the Roslin J-Line, and 81.1 % were previously uncharacterised. Only sixty-two (19.3 %) of ALVEs identified in this study were found in multiple lines. When combined with the ALVEs identified by Rutherford and colleagues following their target-capture sequencing of thirty heritage breeds (2016), this brings the total known ALVE sites to at least 430.

The following section describes the ALVEs identified in each line. Lines have been grouped into five broad classes based on their general 'use' or breed management to enable easier comparison of ALVE diversity and identification of shared insertions. All ALVEs are listed in Appendix 2: AF10 with new nomenclature, insertion location, insertion hexamer and gene overlaps. The full presence/absence data for each ALVE for each analysed line is a 322 x 65 matrix, and has been included on the CD

accompanying this thesis (Appendix 2: AF11). No ALVEs were identified in either of the single end sequencing datasets, and the reasons for this are explored in section 7.4.2.

7.4.1 Identified ALVE insertions

Commercial white egg layers

Including the eight Hy-Line lines and Roslin J-Line, thirty-three white egg layer datasets were analysed. Typical layer ALVEs were well represented across the lines. ALVE1 was in twenty-six lines, ALVE3 in nineteen, ALVE4 in seven, ALVE9 in thirteen, ALVE15 in fifteen, and ALVE21 in seven.

The WLs had one to six ALVEs, apart from the WL-PB-Z 'Zero' line which was bred to have no ALVEs (Bacon et al. 2000), and none were detected (Table 7.3). The BLs had three to eight ALVEs, most of which were shared with the WLs, although there is evidence that the two blind bird experimental datasets were derived from a cross between a BL and a brown-egg layer (discussed below). Few novel ALVEs were identified in these lines. Interestingly the Chinese white egg layer Lhasa white line had ALVE1, ALVE3, ALVE9 and ALVE15: textbook ALVE content for a white leghorn.

The Lohmann commercial leghorn (WL-L) has four ALVEs: ALVE1, ALVE3, ALVE15 and ALVE_ros282, a novel insertion found on an unplaced contig (NT_464277.1), which is shared with two of the inbred Pirbright lines (WL-PB-N, WL-PB-P). The Pirbright WL lines were originally housed at ADOL (Avian Disease and Oncology Laboratory, USA) where their ALVE contents were well described (Bacon et al. 2000). However, only WL-PB-7 matches up completely with the literature (ALVE1 and ALVE2), even when accounting for the absence of ALVE6 in any line due to its insertion site. For example, ALVE10 is supposed to be in at least WL-PB-15 and WL-PB-C, but the insertion site is unknown, and the two lines do not share any 'novel' sites. WL-PB-N is even more deviant from the literature reported ALVE1, ALVE3 and ALVE6. This pipeline identified ALVE1, ALVE4, ALVE15, the novel ALVE_ros282 mentioned above and ALVE_ros034 which is shared with WL-D and blue egg layer Aracauna. These significant ambiguities could be due to changed breeding management when the lines moved to Pirbright, likely with reduced effective population sizes.

Table 7.3 Number of ALVEs identified in the commercial white egg layer lines. The shared ALVEs column represents how many of the ALVEs in that line were shared with any other analysed dataset (not just this grouping). The novel ALVEs column represents how many of the identified ALVEs were previously unknown.

Line name	No. ALVEs	Shared ALVEs	Novel ALVEs
BL-BEGb	8	8	2
BL-BEGs	8	8	2
BL-RGEbm	5	4	0
BL-RGEbp	5	5	0
BL-RGEsf	4	3	0
BL-RGEsp	5	5	0
BL-Br	4	3	1
BL-Sm	3	3	1
JL	3	2	0
Lhasa white	4	4	0
WL1	2	2	0
WL2	3	3	0
WL3	3	3	0
WL4	4	4	0
WL5	2	2	0
WL-L	4	4	1
WL-B-D	5	5	1
WL-B-E	5	4	0
WL-NU	3	2	1
WL-K	5	5	0
WL-IS	2	2	0
WL-HA	6	4	1
WL-LA	4	4	0
WL-PB-15	4	3	1
WL-PB-6	1	1	0

WL-PB-7	2	1	0
WL-PB-C	2	1	0
WL-PB-N	5	5	2
WL-PB-P	5	5	1
WL-PB-W	3	2	1
WL-PB-Z	0	0	0
WL-SPFa	2	2	0
WL-SPFb	3	3	0

WL-SPFa and WL-SPFb were lines from separate companies bred to provide pathogen free eggs for research. Both lines have ALVE1, which is not expressed under normal conditions, and ALVE9, which expresses envelope glycoproteins. Both these lines would be classed as ALV free despite this envelope expression due to testing solely with the p27 ELISA. In addition, SPFb also contains ALVE_ros010 (which is also present in brown egg layers, broilers and generalist breeds such as the Silkie) for which the sequence and expression profile are unknown.

The WL-HA and WL-LA lines were selectively bred for high and low antibody response to infection respectively. The two lines share ALVE1, ALVE3 and ALVE15, and the low line also has ALVE4. The high line has ALVE9 and two sites shared with generalist breeds and broilers. The pool data came from sixteen individuals for both lines, but average coverage was 14.7 X for the high and 16.4 X for the low, meaning that on average at least half of the alleles would not be represented at any one site. As a result, it is difficult to assign any selective effect on the observed differences in ALVE content, especially without expression data for the additional sites in the high line.

The other analysed WLs are all held at research institutes and exhibit typical layer ALVEs. Novel ALVEs were also identified in WL-D (ALVE_ros034) and WL-NU (ALVE_ros046).

The BL-Br line is the parental line for the Smyth line (BL-Sm) which was bred as a model for autoimmune vitiligo. ALVEs in both lines were analysed in 2000 using RFLPs, which suggested both lines had the same three ALVEs (Sreekumar et al. 2000).

Localisation using FISH suggested one was on an unknown microchromosome, one on chromosome 1 (cytogenetic band p2.5) and the other on chromosome 2 (q2.6). Identification from the WGS data identified ALVE3 (chromosome 20), ALVEB5 (chromosome 1) and the novel ALVE_ros173 (chromosome 4) in both lines and ALVEB6 (chromosome 14) in BL-Br. ALVEB5 matches the chromosome 1 banding location well, but clearly there is no chromosome 2 insertion, and no traditional layer ALVEs are in the 2q2.6 region. It is therefore likely that the FISH identification was incorrect or incomplete, as it would be unlikely that an ALVE on chromosome 2 would be lost in both lines independently (unless there has been active selection against it) accompanied by the introduction of a novel ALVE in both lines on chromosome 4.

Both blind chicken datasets have more ALVEs than would be expected in BLs, and many are shared with broilers and brown egg layers such as ALVE-NSAC1, ALVE-NSAC2, ALVE-TYR, ALVE_ros004, ALVE_ros005 and ALVE_ros010, reflecting RIR and WPR crosses early in the respective line development (Hocking 2017, personal communications). BL-BEG lines have eight ALVEs each, sharing seven including a novel lineage-specific insertion on chromosome 1 (ALVE_ros040) and a novel insertion shared with both RIW lines (ALVE_ros066). The only differences are that the blind birds have ALVE15 and the sighted birds have ALVE9, ALVEs which are common in layers without impacting sight. It is likely that the observed difference is due to sampling effects from the sequencing pool. Similarly, each of the Br-RGE lines has four or five ALVEs, but there were nine total ALVEs in the dataset. It is likely that ALVEs were simply missed from some lines, rather than representing true biological differences.

Commercial brown egg layers

The commercial brown egg layer breeds had between six and eleven ALVEs (Table 7.4). All seven lines shared ALVEB5, and all lines except the RIRs shared ALVE-TYR, ALVE-NSAC7 and ALVE21. Three novel ALVEs were discovered across the datasets, of which the Hy-Line RIR and WPR2 novel ALVEs (ALVE_ros007 and ALVE_ros009 respectively) were described above (section 6.4). The RIR-L novel insertion (ALVE_ros211; *5*: 56,988,767) is within the eighth intron of MDGA2 (MAM domain containing glycosylphosphatidylinositol anchor 2).

Table 7.4 Number of ALVEs identified in the commercial brown egg layer lines. The shared ALVEs column represents how many of the ALVEs in that line were shared with any other analysed dataset (not just this grouping). The novel ALVEs column represents how many of the identified ALVEs were previously unknown.

Line name	No. ALVEs	Shared ALVEs	Novel ALVEs
RIR	11	10	1
RIR-L	7	6	1
RIW-ESH	11	10	0
RIW-LSH	11	11	0
WPR1	8	8	0
WPR2	8	7	1
WPR-L	6	6	0

Compared with the sister HL WPRs, the Lohmann WPR has no additional ALVEs. The Lohmann WPR sequence was from a pooled sequencing library so it is possible that rare ALVEs could have been missed, such as ALVE_ros004 which was present in all other brown egg layer lines, but was only found at rare frequencies in the HL WPRs. Despite the similar genetics, the Lohmann RIR has fewer ALVEs than the HL RIR and the two lines only share five ALVEs. Sequencing does not appear to be the issue here as data was available for a pool of ten, and for twenty-five individuals, so it is highly unlikely that any ALVEs with a frequency greater than approximately 0.1 were missed.

The sister RIW lines were under differential selection for egg shell strength. The two lines do differ in their ALVE content, but this is likely due to incomplete lineage sorting or dropped alleles from the pooled sequencing libraries, rather than due to selection. No ALVEs that differ between the two lines are within or near genes.

Broilers

The Western heritage broiler lines and the Lindian Chinese broiler had between thirteen and thirty ALVEs (Table 7.5). Seventy-two different ALVEs were found between the six lines, forty-one of which were novel to this study. The Western lines

share most of their ALVEs with other analysed birds. ALVEB6, ALVE_ros010 and ALVE_ros020 are shared between all five Western lines. The common 'meat type' ALVEs of ALVEB5, ALVEB10, ALVE-NSAC1, ALVE-NSAC5 and ALVE-TYR are also common across these five lines. Despite similar total numbers, the Lindian is very different from the other broilers. Only two of its fourteen ALVEs are shared with any other dataset, and only ALVE_ros220 is shared with another broiler (Br-INRA). It is also unlikely that the full ALVE diversity was represented here as the Lindian sequencing data came from one individual. Two Lindian insertions were within introns.

Table 7.5 Number of ALVEs identified in the broiler lines. The shared ALVEs column represents how many of the ALVEs in that line were shared with any other analysed dataset (not just this grouping). The novel ALVEs column represents how many of the identified ALVEs were previously unknown.

Line name	No. ALVEs	Shared ALVEs	Novel ALVEs
Br-Cobb	15	12	7
Br-INRA	20	12	10
Br-REL	30	24	13
Br-VLDL-F	15	12	6
Br-VLDL-L	13	11	7
Lindian	14	2	11

Five of the fifteen ALVEs identified in Br-Cobb overlapped with genes. Four were intronic, but the lineage-specific ALVE_ros231 is within the 3'UTR of LPHN2 (Latrophilin 2; also known as ADGRL2). Seven of the twenty Br-INRA ALVEs were intronic. This included ALVE_ros072 in the third intron of FRY (Furry homologue; shared with Br-Cobb, Br-VLDL-L, Silkie and RJF-C), ALVE_ros234 in the fourteenth intron of ADGRL4 (Adhesion G Protein-Coupled Receptor L4; shared with Bl-java), and three lineage-specific insertions.

The high ALVE count seen in the Br-REL line was unexpected, but it is possible that matings outside the breed have occurred, as the line includes the typically layer ALVE1 and ALVE15. The REL dataset is also more likely to include the full diversity of ALVEs

within that line as eighty individuals (in 8 pools of 10 birds) were used for sequencing. Seven of the thirty ALVEs are intronic, including ALVE1, ALVE15, ALVEB6 and ALVE-TYR. The other three are all novel to this study, and only ALVE_ros142 is shared (with ETH-Horro).

The Br-VLDL-F and Br-VLDL-L lines are sister lines differentially selected since the 1980s for the concentration of very low density lipoprotein (VLDL) in blood plasma, an indicative marker for fatness or leanness in chickens (Griffin et al. 1989; Griffin et al. 1991). Despite this recent shared ancestry, only five of the twenty-three total ALVE instances identified in these two lines are shared. This could be due to the sequencing of only four individuals for each line, so some of the rarer insertions could have been missed from one line but identified in the sister line. Nine intronic ALVE insertions were identified, four of which were found only in the fat (Br-VLDL-F) line (ALVE15, ALVEB6 and the novel, line-specific ALVE_ros083 and ALVE_ros273), but none of the effected genes were associated with gene ontology terms for lipid/amino acid biosynthesis, and none of the regions were associated with recent selection signature analysis (Khoo et al. manuscript in prep). However, the novel ALVE_ros083 is in the second intron of the CaSR (Calcium-sensing receptor) gene, which regulates calcium, sodium, potassium and water reabsorption in kidney by regulating the release of parathyroid hormone (D'Souza-Li 2006; Vezzoli et al. 2009). Interestingly, this insertion is associated with a genomic deletion of 334 bp (in a similar manner to ALVE_ros007 in the HL RIR). A combination of the presence of an ALVE insertion within the intron, and disruption of the intron genomic sequence, might impact the expression of CaSR.

Generalists and 'native' breeds

This diverse group of fourteen datasets exhibited very varied ALVE content, both in terms of total number and the specific sites identified. Lines ranged from having a single, novel, unshared ALVE (Chahua; ALVE_ros271) to twenty-two ALVEs in the Black Java (Bl-java) breed (Table 7.6). Eighty-three different ALVEs were identified across the lines, fifty-six of which were new to this study (67.5 %).

Table 7.6 Number of ALVEs identified in the generalist or native breeds. The shared ALVEs column represents how many of the ALVEs in that line were shared with any other analysed dataset (not just this grouping). The novel ALVEs column represents how many of the identified ALVEs were previously unknown.

Line name	No. ALVEs	Shared ALVEs	Novel ALVEs
Araucana	12	8	6
Bl-java	22	12	9
Bl-sum	2	1	1
Chahua	1	0	1
Fayoumi	6	2	6
Kedu hitam	12	5	8
Korean	10	9	5
Silkie	9	3	5
Sumatera	2	2	0
Taiwan	7	4	3
TIB-HL1	4	1	3
TIB-HL2	2	2	0
TIB-HL3	2	0	2
Xishuang	16	5	12

The ALVE content of some lines may be incompletely represented due to the sequencing methodology used. For example, sequencing data for the Chahua, three Tibetan highland breeds, Silkie and Taiwan country chicken were each derived from only a single individual, so are unlikely to be truly representative of the breed or region.

RJFs and 'village' chickens

This group represents the most diverse lines when compared to all other analysed datasets. These five lines have 178 ALVE instances (over 55 % of the total ALVEs identified) and 165 ALVEs are novel characterisations in this study (92.7 %; Table 7.7). Within the group, RJF-J and RJF-S have markedly fewer ALVEs, but the identified

ALVEs are nearly all lineage-specific. These two lines had pooled sequencing datasets for two and three individuals respectively, so this may have limited the number of identified ALVEs compared to RJF-C (six individual sequencing datasets).

Table 7.7 Number of ALVEs identified in the RJFs and 'village' chickens. The shared ALVEs column represents how many of the ALVEs in that line were shared with any other analysed dataset (not just this grouping). The novel ALVEs column represents how many of the identified ALVEs were previously unknown.

Line name	No. ALVEs	Shared ALVEs	Novel ALVEs
ETH-Horro	50	11	44
ETH-Jarso	59	8	52
RJF-C	52	4	49
RJF-J	11	1	11
RJF-S	12	0	12

Insertions within genes are common within these lines, except for RJF-S where only one ALVE is intronic (8.3 %; ALVE_ros107). ETH-Horro has twelve intronic overlaps (24.0 %) and ETH-Jarso has fourteen (23.7 %). RJF-C has sixteen gene overlaps (30.8 %), with ALVE_ros074 overlapping with the 3'UTR of ELMOD1 (ELMO domain containing 1). RJF-J has six gene overlaps (54.5 %), including ALVE_ros012, which is the only exonic insertion in the entire study.

ALVE_ros012 is within the eighth exon of the eleven exon CPA5 (carboxypeptidase A5 precursor). The insertion is at 1: 963,340 and the exon eight coordinates are 1: 963,280-963,370. If this insertion causes a transcript truncation, the final 170 amino acids (40.6 % of the total protein length) will be lost. On length alone this would likely destroy the protein functionality, but InterPro analysis also shows that the insertion region is non-cytoplasmic, so protein folding is likely important, and that the insertion disrupts a core PRINTS peptidase domain. For the host RJF this may have limited phenotypic effect as there are multiple paralogues which potentially facilitate redundancy, including the immediately neighbouring CPA2 (upstream) and CPA1 (downstream). However, tissue specific expression or paralogue protein specialisation is unknown.

7.4.2 Failings with the Kauai and Andersson datasets

No ALVEs were identified in either of the single end datasets analysed in this study: the Kauai feral chickens or the multiple Andersson-sequenced lines. Whereas no previous analyses for ALVE content have been conducted with the Kauai chickens, the Andersson-sequenced lines included WLs, WPRs and RIRs, shown in this study and previous work to contain multiple ALVE insertions. Problems with the identification scripts were ruled out due to the successful identification of ALVE insertions in the J-Line and Hy-Line pseudo single end FASTQ files using the single end script pipeline (section 7.3.2). However, the analysis of pipeline sensitivity with those pseudo single end FASTQ files did highlight the need for at least 25 bp of viral clipped sequence to successfully map to the pseudochromosome, and how coverage can affect insertion calling, particularly from pooled libraries.

The Kauai chickens were all sequenced from individual libraries but only with an average coverage of 2.7–4.8X. In addition, the reads were 75 bp compared to the standard 101 bp generated in Illumina sequencing, so the probability of a single read mapping significantly to both host genome and pseudochromosome correctly was reduced. For the Andersson lines the average coverage was even lower at 1.7–3.1X, the data was all from pooled libraries, and the reads were only 35 bp in length, making the calling of ALVE insertions impossible from this identification pipeline.

To further check pipeline sensitivity, each of the Andersson-sequenced BAM files was checked manually for ALVEs found to be common for those breeds, such as ALVE1, ALVE3 and ALVE15 for WLs. No ALVEs were detected this way either, suggesting that structural variant calling of this kind may not be possible with single end 35 bp reads. Testing was done using both the standard BWA mem protocol and the short read optimised BWA aln followed by BWA samse protocol, but this made no significant difference to mapping rates. The Kauai BAM files were also checked manually, again without success. However, this may be because these lines do not have ALVEs common with previously analysed birds, rather than no inserts being detectable with 75 bp reads.

7.5 ALVE insertion patterns across all analysed datasets

7.5.1 ALVE insertion sites

There was no observable bias across the 322 ALVE insertion sites for any kind of GC skew when compared to chromosome or genome-wide GC levels, at any calculated window size. No hexamer insertion sequences were over-represented, and the hexamer GC content was normally distributed around a mean value of 49.9 % GC with five occurrences of entirely A/T hexamers, but no entirely G/C hexamers. The modelled hexamer sequence GC content was also normally distributed, but shifted towards the genome-wide GC content with a mean value of 42.4 %, significantly less GC rich than the observed sequences ($t = 4.66$; $P = 3.84 \times 10^{-6}$).

A total of eighty-six identified ALVE insertion sites were within genes (26.7 %): one within an exon, two within 3' UTRs, and the remainder in introns. If all insertions were random events in the same line you would expect 51.8 % to be within introns. However, each line has had independent insertion events, so it is unwise to over interpret these observed distributions, particularly as it hides the patterns in individual lines. For example, six of the WL lines (WL1, WL5, WL-IS, WL-L, WL-PB-C, WL-SPFa) only have intronic ALVE insertions. Despite this deficit of insertions overlapping genes, 32.9 % of insertions fall within 10 kb of a gene, compared to 4.1 % under random integration.

ALVE insertions, like LTR retrotransposons (Chapter 3), are most common on the largest chromosomes, with identified ALVE number positively correlated with chromosome length ($r = 0.959$, $P < 0.001$). Insertion sites therefore appear random, but the chance an individual insertion will be retained will depend on genetic drift and the selective pressure, if any, caused by that insertion.

7.5.2 ALVEs as genetic markers

The dendrogram constructed from the ALVE presence/absence data (Figure 7.4) generally follows the accepted pattern of ALVE numbers in commercial stock, where white egg layers have fewer than brown egg layers, which have fewer than broilers.

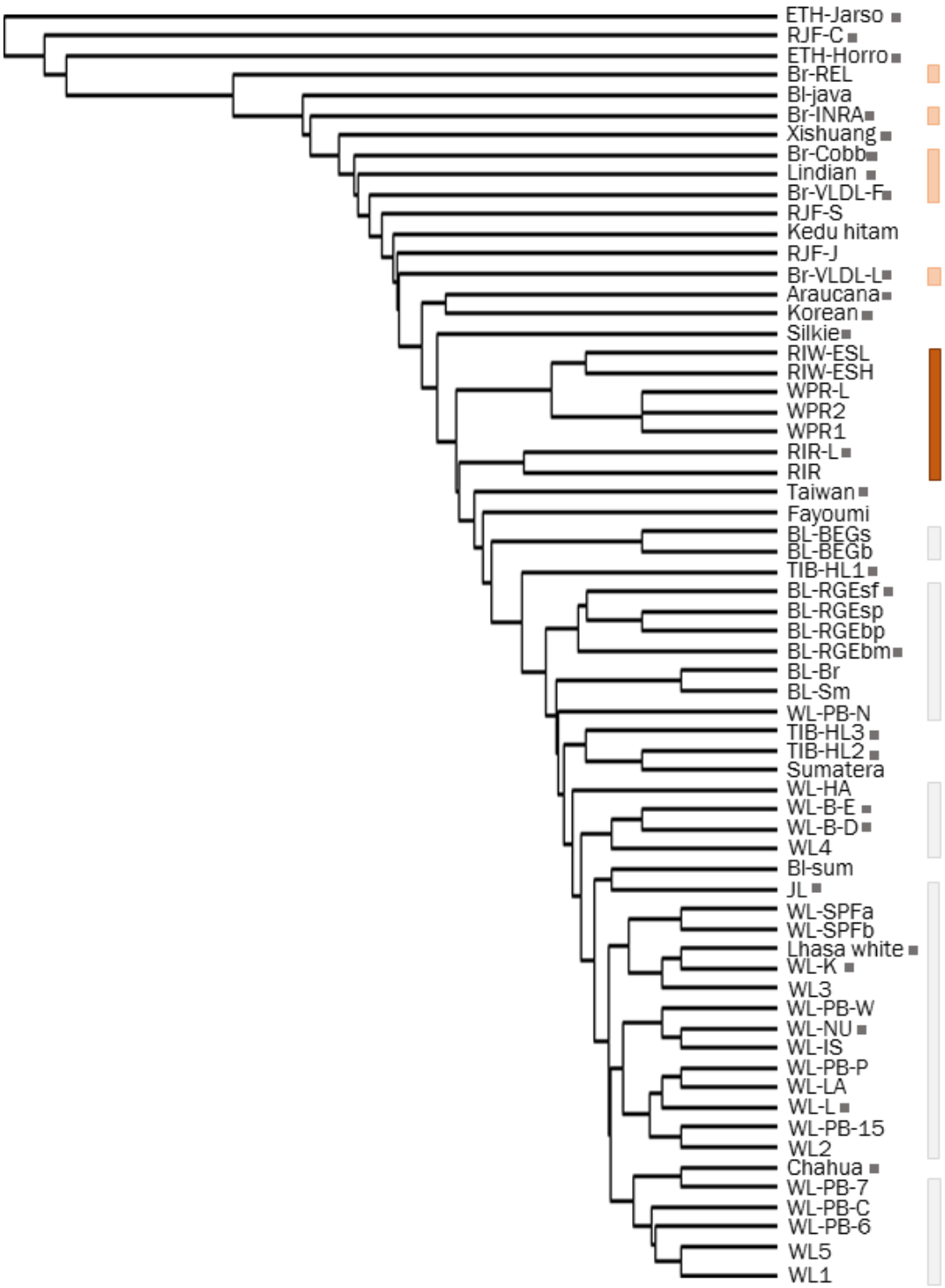


Figure 7.4 Dendrogram of relatedness between successfully analysed chicken lines using ALVE presence/absence data. Broilers are marked with pink boxes, brown egg layers with brown boxes and white egg layers with white boxes. Lines marked with grey squares had individual WGS data. All unmarked lines used pooled data. All datasets were paired end Illumina data. Line name abbreviations were defined in Table 5.3.

The RJF and 'village' birds appear most basal due to their high overall ALVE count and large number of lineage-specific inserts. Whilst RJF-J and RJF-S are not at the base of the dendrogram with RJF-C or the Ethiopian village chickens, this appears to be due to the number of identified ALVEs, rather than the observed diversity. RJF-S has no shared ALVEs and RJF-J shares only one ALVE with the Black Java breed, which makes geographical sense. As considered above, it is unclear whether the broiler lines with large ALVE numbers (such as Br-REL and Br-INRA) are truly representative of commercial broilers as they are pedigree/heritage lines now with relaxed selection, random mating and effective population sizes smaller than historical levels. However, the five Western broiler lines share most of their ALVEs, suggesting limited lineage-specific effects. A dendrogram based on shared ALVE sites alone places these five broiler lines as the most basal branches.

The 'generalist' group are found throughout the dendrogram, many dominated by novel and/or lineage-specific ALVEs, perhaps representing the bias in the analysed datasets, and ALVE literature, towards Western breeds. It is likely that the position of these lines within the dendrogram is more numerical than due to shared sites (discussed below).

The dendrogram internal structure is very sequential, rather than a limited number of clear clades. Where clades do exist, such as the final WL group (WL-SPFa down to WL1) or the WPR/RIW group, this is due to the high numbers of shared ALVEs between these lines and the absence of ALVEs shared outside these clades. ALVEs such as ALVE1, ALVE3, ALVE21, ALVEB5, ALVE-NSAC1, ALVE-TYR and ALVE_ros010 were found commonly and identified in different breeds and groupings, narrowing calculated distances between lines. However, relatedness is modulated further as the large number of lineage-specific ALVEs increased relative distance between lines.

Numerical clustering bias

Whilst Figure 7.4 does appear to reconstitute the predicted general relationships between the analysed lines, the number of identified ALVEs in each line does appear to be a confounding factor. For those lines with few ALVEs this is an inherent bias created by calculating relatedness from a distance matrix using presence/absence data,

as lines with few ALVEs group together based on the high prevalence of the 'absence' genotype at most sites. Modelled dendrograms invariably cluster solely based on the number of ALVEs in each line, where the lowest numbers have the shortest distance (Figure 7.5). The real data does not follow this pattern completely, but instead has structure at least partially based on known line and breed relatedness.

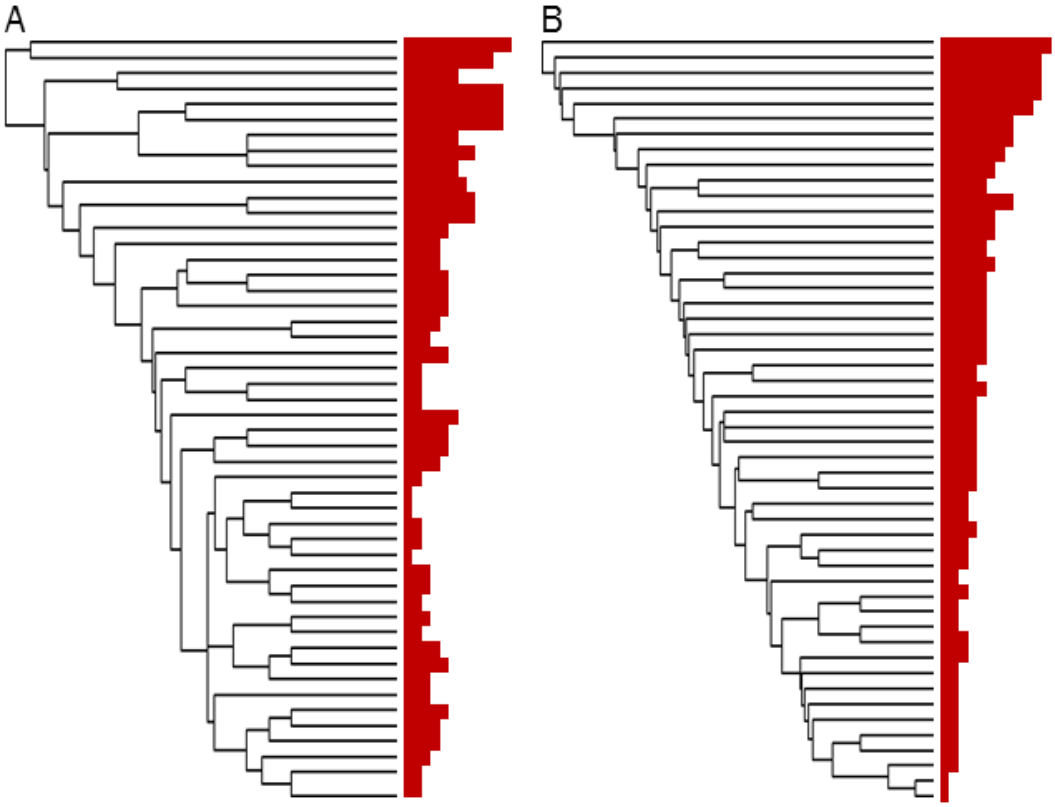


Figure 7.5 Dendrograms constructed for lines with twelve ALVEs or fewer. Line interrelatedness is shown by the dendrogram and line ALVE number is shown by the red bars, where longer bars are more ALVEs. Dendrogram A was constructed using the real data and shows structure based on relatedness rather than number of discovered ALVEs. Dendrogram B was an example of one of the hundred modelled dendrograms with randomly assigned ALVE insertions. Its structure is based solely on the number of ALVEs in each line, with almost uniform number decrease down the dendrogram.

Despite the role of genetic structure, the number of identified ALVEs clearly still has an impact on some of the dendrogram nodes. This point has already been hinted at with the positioning of RJF-J and RJF-S. These lines have one and zero shared ALVEs respectively, yet due to their ALVE number are less basal than heritage broilers which

contain the traditional layer insertions ALVE1 (Br-REL, Br-VLDL-L), ALVE3 (Br-INRA) and ALVE15 (Br-REL, Br-VLDL-F).

Multiple numerical-biased examples can be found within the 'generalist' birds. The Chahua has only one ALVE (a novel insertion on chromosome 18; ALVE_ros271) which is shared with no other lines, but it groups most closely with WLs which also only contain one or two ALVEs. The Fayoumi line has six ALVEs, all of which were novel, but two were shared with the ETH-Horro 'village' chickens. Despite this, the Fayoumi sits in the middle of the dendrogram due to the number of ALVEs, rather than more basally. Similarly, the TIB-HL1 has four ALVEs, three of which are novel and lineage-specific. The remaining ALVE_ros276 is shared with multiple lines including the basal ETH-Horro, ETH-Jarso and RJF-C, but TIB-HL1 still groups within the BLs due to the number of insertions. Finally, the Sumatera groups with TIB-HL2 and TIB-HL3 within the WLs, and all three lines have two identified ALVEs. The Sumatera has ALVE-NSAC1 and ALVE-TYR, both of which are commonly found in all groups excluding the commercial white egg layers. TIB-HL2 shares ALVE-TYR and the previously mentioned ALVE_ros276, but TIB-HL3 has only lineage-specific insertions.

Shared ALVEs appear responsible for most dendrogram groupings, but the number of ALVEs, particularly lineage-specific ALVEs, has a notable effect on the node order. In the absence of shared ALVEs due to common ancestry, lines are simply arranged by the number of ALVEs, as was seen in the modelled dataset (Figure 7.5B).

The impact of sequencing library and average coverage on ALVE identification

Sequencing library type adds an additional level of complexity, as the modelling completed in chapter 6 showed that ALVEs which are rare at the population level are much more likely to be detected from individual sequencing libraries rather than pools. Concordantly, eight of the ten lines with the most identified ALVEs were sequenced from individual sequencing libraries (Table 5.3; Figure 7.4). These lines were, however, predicted to have a greater number of ALVEs than any of the layer-type chickens which exhibited consistently lower ALVE numbers. Use of individual sequencing library datasets with layer individuals, such as the J-Line (JL; 9 individuals) or Lohmann WL

(WL-L; 25 individuals), did not produce an increase in ALVEs relative to other leghorn lines.

When these data were analysed by the GLM, the results showed that line category (based on breed or 'use'; Table 5.3) was the only significant explanatory variable ($P < 0.0001$), accounting for 90.3 % of the observed variation in identified ALVE number once the other variables were accounted for. Neither library category or average genome coverage were significant variables, although these were highly correlated with each other. Within the GLM analysis they cumulatively accounted for 6.1 % of the observed variation, but 97.5 % of this was common between the two variables. This suggests that the use of different sequencing library types has not biased the generated dendrogram.

Clustering by Principal Coordinate Analysis (PCoA)

PCoA was conducted to provide support for the conclusions from the hierarchical clustering presented above. This methodology takes multivariate datasets and attempts to reduce the number of dimensions used to explain dataset variance and relatedness. However, the highly lineage specific ALVE content meant that the first thirteen eigenvalues significantly contributed to explaining the observed variance, meaning that a plot would need to have thirteen dimensions to best visualise the relatedness between the data. This is impossible on paper, so multiple plot versions were trialled using two or three of the largest eigenvalues, but none created good low-dimensional visualisation of the data.

Reduced datasets which included either shared ALVEs across all datasets, only ALVEs found in layers, or only shared ALVEs found in layers were also analysed by PCoA, but these had large contributions from at least eight eigenvalues. This made clear plotting impossible and likely reflects the large variation in lineage-specific ALVE content. Further dataset reductions were not trialled as this would defeat the aim of clustering.

7.6 Discussion

7.6.1 Applying the ALVE identification pipeline to analyse multiple datasets

The ALVE identification pipeline was used successfully on a wide range of paired end WGS datasets. The only notable issue with pipeline performance was with the RJF-C dataset, where the pipeline detected a very large number of putative insertion sites which were then dismissed after visual inspection in IGV. This seemed to be due to the presence of other structural variation in the genome which resulted in a large proportion of incongruently mapped reads. The manual insertion site check was the most laborious step of the pipeline, particularly with the large number of detected sites in some of the analysed datasets when compared with the Hy-Line and J-Line analysis in Chapter 6. A possible improvement to the S6_extract_putative_sites.py script would be to check clipped sequence homology to ALVE sequence, rather than automatically including any sites which have clipped reads. Short sequence BLAST parameters and appropriate thresholds would be required, and sites would still need to be checked manually, but this may reduce some of the noise.

Analysis of the single end WGS datasets was far less successful. No sites were detected in either the Kauai or Andersson datasets, and sensitivity was much reduced in the Hy-Line pseudo single end data. Reduction in sensitivity was due to lower insertion site coverage when reads with only a short clipped section were lost (compared to the Hy-Line and J-Line paired end analyses) as these had no viral-mapped mate to retain the read during read subtraction. This effect was compounded in the Kauai and Andersson datasets by their shorter read lengths, as at least 25 bp was needed for BWA mapping. The Andersson reads (35 bp) could therefore not map to both the reference genome and retroviral pseudochromosome, and the chance of mapping was reduced for the Kauai 75 bp reads compared to the 101 bp Illumina reads of the pseudo single end data. Detection was further limited by the overall low average coverage in both these data sets (< 5X). This is a common problem for the generally older single end datasets, as these were completed when sequencing chemistry was more expensive and less productive.

Paired end WGS datasets are far more commonplace now than single end data, and it would be unwise for any researcher to commission new single end sequencing. The

single end ALVE identification pipeline is therefore useful for Illumina single end data, or perhaps very high coverage 75 bp SOLiD reads.

Future applications: adapting to long read sequencing data

Long read technologies offer an exciting opportunity for the study of ALVE insertions, and structural variants more widely. Long reads facilitate unique repetitive read mapping (producing better assemblies) and may include the full ALVE sequence within a read. This would enable initial ALVE characterisation without the need for the additional sequencing, as performed for the Hy-Line sites (section 6.6).

Recent work has successfully used long read technology to identify retrotransposon insertions and deletions, with the authors developing the PBHoney program to automate detection of all observed variants (English et al. 2014). I would recommend the construction of an identification approach specific to ALVEs to reduce search effort and post-PBHoney filtering. This would follow a very similar approach to the pipeline for short read data developed in this project: mapping reads to a pseudochromosome using a long read aligner such as BLASR (Chaisson & Tesler 2012), alignment of mapped reads to the reference genome, and then filtering by the alignment CIGAR strings to identify clipped reads or reads with large internal insertions (likely the whole ALVE sequence). Secondary alignments may need to be considered if reads containing long intact ALVEs preferentially mapped to the assembled ALVE-RJF sequence leaving the true, flanking genomic DNA as a clipped alignment.

7.6.2 ALVE diversity across chicken populations

A total of 322 different ALVEs were identified in this study. When combined with the recent targeted sequencing approach by Rutherford and colleagues (2016), this brings the total number of identified ALVEs to over 430, almost nine times as many as were known at the start of this project. This highlights the diversity of ALVEs across chickens and the large potential for phenotypic host effects, given the high structural integrity of these young retroviral insertions. This number is likely also an underestimate, even of

the lines which were analysed, due to population sampling and the use of pooled sequencing libraries. In my work, only 19.3 % of identified ALVEs were shared between multiple lines (some of which were experimental sister lines) and 81.1 % were previously uncharacterised.

Overall, ALVEs showed no bias for the GC content of their insertion site region, but insertion hexamers were significantly more GC rich than would be expected with random insertions. Interestingly, alpha- and betaretroviruses have been found to exhibit the lowest integration site preference amongst retroviruses, appearing almost random (Serrao et al. 2015). However, like other retroviruses, ALV site choice is dependent on the ease of access for the integrase, with insertion sequences exhibiting A-philicity, where the DNA has A-DNA-like structure rather than typical B-DNA, producing a larger minor groove in the DNA helix (Wu et al. 2005; Serrao et al. 2015; Grawenhoff & Engelman 2017).

One observed bias for ALV insertion is an apparent preference for open chromatin, particularly near regions transcribed by RNA polymerase II, such as protein-coding genes (Narezkina et al. 2004; Serrao et al. 2015). As gene density is positively correlated with GC content, this may explain the elevated GC content of the insertion site sequences. The data appears to support this preference, as 59.6 % of identified ALVEs (192 of 322) either overlapped with or were within 10 kb of a protein-coding gene. Only eighty-six of these overlapped genes, but this is not unexpected, as we are observing only the insertions that were retained by the host. It is possible that many genic insertions were lethal or highly detrimental to the reproductive or commercial success of the host, resulting in this apparent deficit within genes, despite the retrovirus preference.

ALVE trends between different breeds

This study has identified an increased number of ALVEs across the Western commercial breeds, but the overall pattern of white egg layers having fewer than brown egg layers, which have fewer than broilers remains true. These breeds also reconstitute good phylogenetic relationships based on their ALVE content due to the number of

shared sites, with the brown egg layers appearing like a hybrid between broilers and white egg layers, reflecting their shared ancestry (Muir et al. 2008).

The difference between these three groups is partially due to greater selection in layers, particularly white egg layers, against the negative effects of ALVE expression on commercial traits such as egg laying success. However, commercial white egg layers were also derived from a much narrower genetic background than brown egg layers and broilers, and therefore exhibit much lower levels of genetic diversity (Muir et al. 2008). In all commercial stocks, a relatively small effective population (N_e) size and an elevated probability of inbreeding could lead to the reduction of the number of ALVEs in a population, but it could also lead to more rapid fixation of detrimental variants in the population, retaining ALVEs as fixed loci (Charlesworth 2009). Most highly inbred lines, such as the Pirbright WLs or other institute-housed WLs with N_e as low as two, were also derived from just a few individuals of more outbred commercial stock following their initial intensive selection. WLs therefore share very similar, low ALVE content, but their population allele frequencies and likelihood of fixation or loss of an insert are dependent on N_e and the strength of selection. It is also possible that reduced N_e has effected the ALVE content of the broilers analysed for this work. These lines are no longer commercially relevant, but are retained as heritage breeds or experimental, institute lines, both of which will have significantly reduced N_e. Analysis of commercial broilers would be of great interest, particularly to see if dam lines (bred solely for reproductive success) also exhibit reduced ALVE content relative to the sire lines, due to the negative effects of ALVE expression on layer success (Muir et al. 2008).

The large numbers of ALVEs within wild or RJF populations superficially suggests that there was an originally high diversity of ALVEs within the RJF ancestor birds before domestication. However, many of these ALVEs are lineage specific, likely reflecting the complex domestication of the chicken, with multiple origins and backcrosses, coming from an ancestrally very large N_e (Ellegren 2005; Rubin et al. 2010). This likely suggests that all domesticated breeds were derived from high, possibly very different, ALVE backgrounds. Birds were then naturally, indirectly selected for reduced ALVE content, due to the impact on productivity and the potential for infection of other birds in the flock. Parallel selection from different origins most parsimoniously explains the similar number, but highly divergent insertion sites, of non-Western breeds, such as the

Fayoumi and Silkie. It is possible that with additional sequencing of geographically disparate RJFs and more non-Western, non-commercial breeds, that ALVEs could be used to track breed origins and interrelatedness.

This work also supports the recent research by Ulfah and colleagues (2016) which showed that the RJF reference genome bird was a result of extensive introgression with WLs, rather than truly representative of the domestic chicken ancestor. This sequence should therefore not be used as a tool for identifying changes since domestication. One such case was the recent publication of piRNA-mediated ALVE control in the germline which used the reference genome as the ancestral baseline (Sun et al. 2017).

7.7 Concluding remarks

ALVE identification has generally been limited to commercial breeds or experimental lines, with a large bias towards Leghorns. In this work, 322 ALVEs were identified from almost one hundred chicken WGS datasets, including village chickens, non-Western breeds and wild red jungle fowl, and over 80 % of these were previously unknown insertions. This has enabled a better characterisation of the wide diversity of ALVEs across chicken populations, but it is likely that this only scratches the surface.

Chapter 8: Discussion

This PhD project consisted of two broad aims. Firstly, the production of an updated annotation of the LTR retrotransposon content of the chicken genome, including a wider evolutionary study of LTR retrotransposon content across the avian lineage. Secondly, the development of a bioinformatics approach for the identification of novel ALVE insertions from chicken whole genome resequencing data, focusing on characterisation of ALVEs in the Hy-Line elite layer lines.

Chapters 3 and 4 addressed the first aim. The available methodologies for identifying LTR retrotransposons were critically assessed, and the LocaTR identification pipeline was subsequently developed. LocaTR combines multiple identification strategies from existing programs to enable the most comprehensive annotation of LTR retrotransposons currently available. LTR retrotransposons were identified firstly in the chicken Galgal4 assembly, almost doubling the previously annotated content. Identified elements were exclusively ERVs, and analysis of element distribution revealed the effect of selection on these elements to reduce their phenotypic impact on the host genome. ERVs were also assessed for their age, completeness and expression, which enabled the further characterisation of *Ovex1* (a co-opted chicken gene of gammaretroviral origin), perhaps supporting a different, wider function for this gene than was previously hypothesised. LocaTR was then used to annotate the LTR retrotransposon content of seventy-three sauropsid genomes (including the new chicken Galgal5 assembly), enabling characterisation of lineage-specific effects and the role of genome quality in repeat annotation. This work did not support a previous hypothesis that galliform birds had a deficit of LTR retrotransposons compared to other avian groups.

Chapters 6 and 7 addressed the second aim. A scripting pipeline was developed to identify ALVEs from WGS data using standard software and tools. This pipeline was initially used on resequencing data from eight Hy-Line elite layer lines, and twenty different ALVEs were identified including five novel sites. As expected, white egg layers were found to have fewer ALVEs than brown egg layers, but many white egg layer ALVEs were fixed in individual lines. Diagnostic KASP and traditional gel-based PCR assays were developed to identify all insertions, and fifteen of the identified ALVEs were sequenced and characterised. Additionally, the fast feathered WPR line was identified

as a *K* locus revertant, which still contained the replication competent ALVE21. The pipeline was then applied to almost one hundred diverse chicken datasets and over three hundred different ALVEs were identified, 81.1 % of which were novel to this study. This wider study revealed some of the ALVE diversity in non-commercial lines, and how unrepresentative the reference genome was for RJF ALVE diversity. The pipeline was critically assessed for detection sensitivity, and particularly how this changed depending on read length, read library construction and observed coverage.

In this final chapter, I will evaluate the relevance of my findings in the wider context, considering limitations with the methodology, the evolutionary roles for these repetitive elements in avian genomes, and the practical consequences of this study, including the proposals for further research.

8.1 The development of novel identification pipelines for the identification of LTR retrotransposon-derived repetitive elements

8.1.1 LocaTR

The LocaTR identification pipeline consists of three homology-based identification programs, four structure-based identification programs and twenty-three custom scripts written in BASH and Python 2.7. Multiple programs and strategies were required as individual programs have been shown to identify different subsets of LTR retrotransposons (Lerat 2010; Garcia-Etxebarria & Jugo 2010; Garcia-Etxebarria & Jugo 2012). The use of so many identification programs covering both homology and structure-based strategies, had not been completed before this project. In addition, some of the programs were difficult to run, either due to limited documentation and support, or more systematic issues with memory allocation. LocaTR mitigates these issues through the accessory scripts. All scripts are on the CD accompanying this thesis (Appendix 1) and on GitHub (https://github.com/andrewstephenmason/LocaTR).

LocaTR accessory scripts use common Linux-hosted software (particularly for a bioinformatics group) such as BLAST, EMBOSS and HMMER, but are largely self-contained to avoid complications with accessory programs and dependencies. LocaTR can control for variation in input files (such as long or variable sequence header names)

and is highly adaptable, as additional identification software, reference sequences or validation profiles can be added at the user's discretion, and the individual parameters for identification software can still be set by the user. This makes the pipeline applicable for LTR retrotransposon identification of any assembled genome, with full characterisation of a one to two gigabase genome in approximately ten days.

Analysis with LocaTR has increased the annotated LTR retrotransposon content of all genomes studied within this project. However, the incorrect identification of non-LTR retrotransposons, such as CR1, has remained a concern due to two main issues. Firstly, LTR retrotransposons are highly diverse elements and although the main groups share an archetypal structure, the lengths of individual features (such as the LTRs) can vary across three orders of magnitude (Benachenhou, Jern, et al. 2009; Llorens et al. 2011). Consequently, parameters for structure-based programs must be kept unconstrained, leading to the inclusion of non-LTR retrotransposons within blocks of repetitive sequence which superficially represent LTRs. Secondly, high homology between the *polymerase* domains of retrotransposons leads to detection of these regions during the BLAST protocols of LocaTR. Future work needs to address the careful validation of putative elements to guarantee their correct annotation, particularly when the pipeline is applied to more diverse species with different non-LTR retrotransposon complements. This should include better handling of fragmented sequences identified in the BLAST protocols by expanding alignments to the neighbouring regions, and including non-LTR retrotransposon domain pHMMs in the validation scripts, to test for more significant domain matches than are observed with LTR retrotransposon pHMMs.

The major limitation to the success of LocaTR is genome contiguity. This does not significantly influence the initial homology-based identification, but fewer intact elements are identified in less contiguous genomes. This has a multiplying effect, as fewer intact elements result in a smaller contribution from the secondary BLAST protocol, where lineage-specific divergent and degraded sequences are detected. This is a consequence of repetitive genomic regions being most difficult to assemble, and should improve as genome assemblies are updated with long read sequencing and high resolution optic mapping, as was discussed on page 5. Researchers should therefore be careful when comparing repeat content between species unless both genomes are highly contiguous. Concordantly, future work could identify LTR retrotransposon patterns among the bird

genomes, including the previously hypothesised deficit in the Galliformes which was not supported by this work.

Additional improvements to LocaTR are required to improve its efficiency and scalability. This should include investigating cloud computing methods to parallelise identification with RetroTector and LTR_STRUC (as these require desktop architecture), and the standardisation of intermediary files to match existing file formats such as BED6.

8.1.2 ALVE identification pipeline

The ALVE identification pipeline was originally designed for the analysis of paired end sequencing data and consists of seven scripts written in BASH and Python 2.7. The pipeline uses standard bioinformatic tools for the processing, alignment and manipulation of NGS data, avoiding reliance on complex accessory software packages. The pipeline's application was extended with alternative scripts for the analysis of single end sequencing data, and an accessory script to convert SOLiD 'colorspace' FASTQ files into standard 'basespace' FASTQ format. The ALVE identification pipeline is therefore applicable to most publicly available WGS datasets, although additional scripts would be required for the analysis of long read sequencing data (discussed on page 5). The pipeline is also highly versatile as users select the type of inserts to be detected by providing the appropriate reference sequences. This makes the pipeline applicable to any species, and any viral insertion. Depending on the size of the FASTQ files, identification of ALVE insertions was completed in one to three days. All script files are on the CD accompanying this thesis (Appendix 1) and in a GitHub repository (https://github.com/andrewstephenmason/ALVE_ID_pipeline).

The major limiting factor for the success of the ALVE identification pipeline is total coverage. Soft-clipped reads at the insertion site are required to give the exact insertion co-ordinates, as well as validate the insertion by providing ALVE-homologous sequence. In datasets with limited coverage, there are few soft-clipped reads to support the insertion site, and in single end datasets reads must contain significant matches to both the insert and reference genome. This issue is compounded by short sequencing reads (no ALVEs

were detected with read lengths shorter than 100 bp during this project) and the use of pooled sequencing libraries, where inserts may be at a low frequency within the sample population or the insert alleles may become underrepresented after PCR amplification. More recent sequencing projects which use paired end reads of at least 100 bp, generate at least 20x coverage, and use individual sequencing libraries, avoid issues with insufficient coverage. Consequently, targeted ALVE sequencing to generate higher coverage of genome-insert boundaries should only be used to complement existing datasets or to conduct ALVE diversity studies in a more cost effective manner if whole genome sequencing is not required. However, as the cost of individual chicken genome sequencing to 30x coverage has fallen to approximately £300 (and will continue to fall), it is worth completing whole genome sequencing rather than generating datasets with limited wider research applications. Careful consideration is required to select an appropriate number of individuals for sequencing based on the desired confidence for detecting insertions of a given frequency, and the size of the flock. Individual sequencing libraries are best, as variation in PCR amplification in pooled sequencing libraries reduces detection confidence and can reduce sensitivity (sections 6.4.3 and 7.3.2).

Another limitation is the difficulty in identifying insertions in poorly assembled regions. This certainly hindered the identification of ALVE6 in any sequencing data analysed in this PhD project. Future improvements to the reference genome should eventually mitigate these issues, but this is worth remembering if the pipeline is applied to viruses which preferentially integrate at the telomeres or centromeres. Interestingly, this thesis has shown that the chicken reference genome is particularly under-representative of ALVE diversity among chickens and wild RJF, corroborating recent analysis by Ulfah and colleagues (2016). However, this has actually facilitated the identification of novel ALVE insertions during this project, as the reference genome provided an almost ALVE-free background, particularly as the only fully assembled reference genome ALVE (ALVE-RJF) has not been identified in any other chicken (Benkel & Rutherford 2014). Known alpharetroviral sites were filtered out during the identification pipeline, but each line was manually checked for the presence of ALVE-RJF to ensure mapped reads were not filtered out incorrectly. This is an important consideration if the pipeline is applied to more ancient and numerous insertions, such as EAVs, as there will be assembled EAVs which are shared with some lines, but absent in others. It may aid novel

insertion detection if assembled insertions are masked after their identification (*e.g.* converted to ambiguous bases, Ns), and each detected site treated individually.

The most laborious part of the pipeline is the manual validation of putative integration sites using a genome viewer such as IGV. Whilst this was straightforward with the low numbers of ALVEs detected in the commercial layers, it was more time consuming with the non-commercial and RJF datasets. Future work could develop the filtering of putative sites, perhaps even identifying soft-clipped reads, checking these against the given reference sequences, and extracting the insertion site and hexamer for comparison against a database of previously identified ALVEs. Such changes would likely limit manual checking rather than replacing it entirely, to ensure putative sites are not filtered out incorrectly.

8.2 Evolutionary roles for LTR retrotransposon-derived sequences in avian genomes

8.2.1 Repetitive elements in avian genome evolution

This project has increased the annotated LTR retrotransposon content of all analysed genomes. Despite this, avian genomes remain repeat sparse. Whilst the avian karyotype likely closely resembles that of the amniote ancestor (Ellegren 2005; Ellegren 2010), it is interesting that birds exhibit only a subset of the transposable elements present in that ancestor. Within the LTR retrotransposons, this has included the loss of both DIRS and Gypsy elements, with only ERVs represented in the lineage. However, this is a wider phenomenon, as birds contain only one of the autonomous non-LTR retrotransposon groups (CR1) which have been observed across other amniotes (Kapusta & Suh 2017).

Bayesian inference of extinct dinosaur genome sizes has suggested that genome contraction in the avian lineage began approximately 230 million years ago, following the divergence from the crocodilian lineage (Organ et al. 2007). As flight evolved much later, it is likely that initial genome contraction was due to the physiological pressures of endothermy, with similar reductions avoided in mammalian genomes due to the evolution of enucleated erythrocytes (Hughes & Hughes 1995; Waltari & Edwards 2002; Cavalier-Smith 2005; Organ et al. 2007). Even in avian groups where flight has been lost,

relaxation of the constraint on genome size has not necessarily had a consequential increase in repetitive content (Briggs 2003; Phillips et al. 2010; Kapusta et al. 2017).

The paucity of all repeat classes in Paleognathae birds (Kapusta & Suh 2017), including LTR retrotransposons (Chapter 4), may support the significant depletion of repetitive elements during the genome contraction of the avian ancestor, corroborating the modelling of Organ and colleagues (2007). This likely suggests a limited role for repetitive elements in the diversification of birds within the Archosauriformes, even though these elements remained active in the genome. However, Neognathae birds exhibit higher relative repeat content despite continued constraint on genome size, supporting a secondary, more recent, evolutionary role for transposable elements. Large differences in repetitive content have been shown to increase speciation rates due to hybrid sterility, whilst additionally facilitating divergence by providing novel promoters, splice variants and entire coding regions (Ginzburg et al. 1984; de Boer et al. 2007; Jern & Coffin 2008; Stoye 2012). This project has generated much more complete LTR retrotransposon annotations for many avian species, so could be analysed further to examine whether periods of elevated LTR retrotransposon activity match known speciation events, such as the rapid diversification of Neoaves at the K/T boundary. This would involve the characterisation of the age and genera of the identified LTR retrotransposons, likely following the methodology of the existing avian repeat study of Kapusta and colleagues (2016). This work would also aid the understanding of lineage-specific LTR retrotransposon content and group dynamics, and the role of exogenous retroviruses in stimulating lineage-specific expansions.

In addition to characterising the genera and ages of LTR retrotransposons across the avian lineage, further work should describe the distribution of these elements compared to the work presented in this thesis with the chicken. In chicken, LTR retrotransposon distribution is non-random, and represents the effects of selection to mitigate the impact of integrations. Consequently, LTR retrotransposon density is greater on the gene sparse macrochromosomes, and distribution on these chromosomes is skewed away from coding regions (sections 3.6 and 4.4). Whilst many of the analysed avian genomes lack comprehensive gene annotations, it would be of great interest to characterise their LTR retrotransposon distribution, particularly the presence of intact element clusters and whether any of these share synteny between closely related species. The presence of

large genomic regions which differ between otherwise syntenic chromosomes may have contributed to hybrid sterility and speciation (Ginzburg et al. 1984; Nuzhdin 1999).

8.2.2 Co-opted LTR retrotransposon-derived sequences in chicken

Protein coding genes of LTR retrotransposon origin

Protein coding genes co-opted from LTR retrotransposon-derived sequences are rare occurrences, although more transient roles in antiviral immunity have been identified in multiple species, including chicken (see below) (Crittenden et al. 1984; Smith et al. 1991; Sacco et al. 2004; Aswad & Katzourakis 2012; Stoye 2012; Hurst & Magiorkinis 2014). In this study, the chicken co-opted gammaretroviral *envelope* gene, *Ovex1*, was shown to have a wide distribution in the Sauropsida, and to exhibit expression beyond the ovaries, contrary to its initial characterisation (Carré-Eusèbe et al. 2009). This may support a much more general function for *Ovex1*, perhaps as a competitive inhibitor against exogenous gammaretrovirus infection, however this requires further functional characterisation.

Ovex1 expression was ubiquitous, but was far greater in the ovaries than in other identified tissues. It is therefore possible that minimal levels of Ovex1 protein was produced outside the ovaries, therefore limiting its range of effects. Furthermore, whilst *in silico* domain identification was carried out in this project, isolation and identification of the Ovex1 protein would provide evidence as to its functionality, as fully functional retroviral envelope proteins form homotrimers. *Ovex1* knockouts would also identify whether Ovex1 is an essential protein, particularly whether it is required in ovary development as was originally hypothesised (Carré-Eusèbe et al. 2009), or whether *Ovex1* mutants are more susceptible to current exogenous gammaretrovirus infection, such as CSV. As *Ovex1* homologues are present throughout the sauropsids, it would be of great interest to fully understand the evolutionary impact of this co-opted ERV.

It is possible that other chicken genes have incorporated LTR retrotransposon-derived sequences as exons or promoters, particularly as overlaps with protein-coding genes were observed in the Galgal4 analysis. Further work should confirm and characterise overlaps with annotated exons, and identify their contribution, if any, to gene function.

Long non-coding RNA genes of LTR retrotransposon origin

Recent improvements to the chicken genome annotation has enabled the identification of over 20,000 long non-coding RNA (lncRNA) genes (Kuo et al. 2017). As a group, lncRNAs are not necessarily functional, but merely evidence that a genomic locus is expressed, even if in a tissue-specific or temporally-specific manner. However, those that have been well characterised have a wide range of functions, from subcellular organisation and epigenetic remodelling, to regulating transcription and facilitating alternative splicing, and are often species-specific (Mercer et al. 2009; Baker 2011; Pontier & Gribnau 2011; Wang & Chang 2011).

As many mammalian lncRNAs are known to be of transposable element origin (Kapusta et al. 2013), it was interesting that seventy-two intact chicken LTR retrotransposons (5.6 %) significantly overlapped (> 50 bp) or fully contained a lncRNA gene. Future work needs to assess LTR retrotransposon regions which are overlapped, whether any degraded LTR retrotransposon sequences overlap with lncRNA genes, and derive the function, if any, of the overlapped loci. There is no general, comprehensive methodology for the functional characterisation of lncRNA genes, and loci generally need to be evaluated independently with extensive experimental evidence (Mercer et al. 2009). Recent *in silico* efforts have been made to assign direct lncRNA functional relatedness by assessing co-expression with well described protein coding genes using RNAseq datasets over multiple tissues (Xiao et al. 2015). However, this requires a lncRNA to have a direct effect, and expression that can be accurately quantified over multiple tissues. Initially, research should focus on identifying the functional 'class' of any overlapped lncRNA, with individual loci targeted for further research. Functional characterisation of lncRNA genes is a new and rapidly expanding field, and new predictor tools are currently under development (Kuo 2017, personal communication).

8.2.3 Transient ALVE-derived immunity in chickens

Many recent ALVE insertions retain a high degree of structural integrity, and the role of well described elements such as ALVE6 and ALVE9 in mitigating exogenous ALV infection through receptor interference has been well described (Robinson et al. 1981;

Smith et al. 1990a; Smith et al. 1991; Gavora et al. 1995). In commercial flocks, where exogenous ALV infections are controlled and largely absent (at least in Western settings), these viral-protein producing loci are undesirable and have the potential to cause complications when lines are interbred. However, in wild RJF and non-commercial birds it is likely that these insertions would be transiently beneficial to the host, as the physiological cost of ALV infection outweighs the negative effects of endogenous expression. This potentially explains the large diversity of ALVEs in the village birds and RJF datasets, with cumulative flock immunity likely caused by many individual ALVEs at low population frequencies, as the rate of adoption will be far greater than the rate of fixation (Taylor et al. 2011; Aswad & Katzourakis 2012). Individual insertions which remain useful will be retained, and selection will only act to keep insertions whilst they provide a selective advantage.

ALVE insertions which confer a selective advantage to the host are expected to be under constraint, exhibiting fewer stop codons and low dN/dS ratios (non-synonymous/synonymous changes) in the advantageous domains, such as the *envelope* gene. These values could be compared with other ALVE insertions to identify domains under selection, and to exogenous relatives to observe how selection pressures change when under host control. Observed selection effects can be quantified using the McDonald-Kreitman test, or the more robust Distribution of Fitness Effects (DFE) test. DFE tests account for changes in effective population size over time, and the impact this has on the effectiveness of selection (Aswad & Katzourakis 2012). Such analyses would require the sequencing of multiple inserts from individual birds, however with long read sequencing it may be possible to generate individual reads which cover the entire insertion. Whilst it is unlikely that the selective role of recent insertions could be studied in this manner, these analyses would enable characterisation of ALVE degradation over time, without the influence of artificial selection against P27 expression.

8.2.4 Further areas for research

In chickens, further work is needed to fully characterise the diversity of ALVE insertions across RJF and non-commercial chickens, as well as identify whether ALVE numbers in commercial broiler dam lines are as similarly reduced as in commercial layers.

Furthermore, ALVEs are not the only alpharetroviral sequences which remain mobile in the chicken genome. EAVs are older and generally less complete, but are present in greater copy numbers and still modulate phenotypes. During this PhD project I collaborated on an EAV-HP diversity study in a range of chicken populations (Wragg et al. 2015), but further analysis of all EAV classes is required. Preliminary testing has been completed using the ALVE identification pipeline adapted for EAVs, and this would be a straightforward application. Beyond endogenous alpharetroviruses, further work is needed to characterise the effects of the betaretrovirus and gammaretrovirus *polymerase* transcripts identified from the J-Line RNAseq data (page 5). If translated, these proteins could facilitate the retrotransposition of non-autonomous repetitive elements, and the formation of retrogenes.

This PhD project has also generated updated LTR retrotransposons for seventy-two additional sauropsid species, including commercially relevant and key indicator species across the avian lineage. Most identified LTR retrotransposon sequences will exhibit limited phenotypic influence on the host due to their location or degree of degradation, but expression and epigenetic data from a range of germline and somatic tissues is required to fully characterise these sequences. These annotations provide the basis for future evolutionary studies into transposable element activity and phenotypic influence across the avian lineage, as well as a baseline for characterising the diversity of recently inserted elements between populations, as was completed here with chicken ALVEs. Whilst ALVEs are limited to *Gallus gallus*, other endogenous ALVs are found across galliforms, and may present future targets for recombination. It is also likely that the study of other avian species will reveal recurrent retroviral infections and integrations.

8.3 Practical applications from this CASE PhD project

8.3.1 ALVEs in the Hy-Line elite layer lines

The ALVE identification pipeline identified twenty different ALVEs across eight elite layer lines. As expected, the white egg layer WLs had consistently lower numbers of insertions than were detected in the brown egg layer WPRs and RIR. However, it is highly likely that not all ALVEs within these commercial flocks were identified. ALVEs

at low frequency could have been missed due to individual sampling or under-representation in the pooled sequencing libraries, particularly if insertions were line specific and could not be detected with subsequent KASP assays. To completely characterise the ALVE content of the lines, additional individual sequencing is required, and must be designed to maximise the probability of detecting all ALVE insertions. However, it is possible that very rare insertions would remain undetected, particularly if they are on the W chromosome. Furthermore, any insertion in a poorly assembled region of the genome, such as ALVE6 at the 5' end of chromosome 1, could not be reliably detected by this methodology. Work is ongoing to identify the ALVE6 insertion by extending the chromosome 1 assembly using long read sequencing data.

Traditional, gel-based PCR and high-throughput KASP assays were developed for each of the identified ALVE insertions. It is likely that all gel-based PCR assays could be used successfully with other chicken lines, enabling the genotyping of ten ALVEs which lacked assays prior to this study. The SNP-level specificity of the KASP system might mean that assays need to be adjusted before their successful application in other populations. In the absence of sequencing data to direct these adjustments, a similar approach could be taken as was completed with ALVE_ros005, where local Sanger sequencing was used to identify problematic sequence. The KASP assay system is a powerful tool for high-throughput genotyping, and this work represents one of the first uses of the system for detecting large structural variants rather than SNPs.

Further characterisation of the ALVE insertions

The work presented in chapter 6 of this thesis focused on the detection of ALVE insertions from WGS data, the development of diagnostic assays, and the elucidation of the insert sequence. However, further work is needed to fully quantify the phenotypic effects, if any, of each ALVE insertion. Firstly, examples of each ALVE from each line need to be sequenced, rather than a single representative sequence for each ALVE across the Hy-Line elite layer lines. This was the original intention of the work presented in section 6.6, but difficulties in long range PCR and PCR product cloning limited the scope of sequence characterisation, including the absence of sequences for five identified ALVEs. It is possible that as these insertions are relatively recent there will be

few significant changes between different lines, however any large-scale changes could drastically affect the phenotypic impact of an ALVE, especially if a full length insert in some lines is a solo LTR in others.

Prior to any additional sequencing (see below), the phenotypic effects of each ALVE could be assessed through an association analysis with productivity trait data. These analyses compare the genotypes for genetic variants with observed differences in productivity traits, to associate either individual or cumulative genotypes with observed phenotypic differences (Haley & Knott 1992; Haley et al. 1994; Korte & Farlow 2013). Such analyses cannot observe the effect of fixed insertions (such as ALVE-TYR in the WPRs), but could identify other associations. It is possible that such an analysis would reveal no negative associations between ALVEs and productivity traits, but any highlighted ALVEs would be prime targets for eradication from the lines. In addition, gene expression of the five genes where ALVEs are within introns (ALVE1, ALVE3, ALVE9, ALVE15 and ALVE-TYR) should be quantified using RT-qPCR in wildtype, heterozygote and homozygote individuals, to identify whether the presence of the ALVE influences gene expression. This should be completed for the exons immediately before and after the insertion, as previous work with ALVE-TYR found that the insertion did not affect total expression but caused transcript truncation (Chang et al. 2007).

To completely characterise the phenotypic effects of each ALVE, more sequencing data is needed across a whole range of tissues at different ages, including developmental stages. This should include whole genome bisulfite sequencing for characterisation of the methylation status of each ALVE insertion, and ChIPseq to identify protein interactions with the ALVEs such as enhancer or TF binding, or the presence of histones. Additionally, the expression of each ALVE should be assessed using long read RNA sequencing, such as PacBio IsoSeq, to enable unique identification of transcripts from each site. This data would also confirm whether the *envelope*-regulating microRNA miR-155 is active within these lines. Together, these data would identify whether ALVEs are transcriptionally active under normal conditions, and whether they could be reactivated under stress, as has been shown during coinfection with MDV (Fadly et al. 2014; Cao et al. 2015).

ALVEs in other commercial lines

The approaches for ALVE identification and characterisation outlined in this thesis could easily be applied to the identification of ALVE insertions in other commercial datasets. Related analyses are currently underway in other companies of the EW group (the Hy-Line parent company; J Fulton & S Tyack 2017, personal communications), but it is unlikely that the broiler results will be made public. The analysis of current commercial broiler data, particularly the differences between sire and dam lines, would be of great interest, as controlling for the likely more numerous ALVE insertions in those lines could greatly influence productivity and animal welfare.

In chapter 7, the Lohmann (also in the EW group) WL, WPR and RIR layer lines were analysed for their ALVE content, revealing very similar complements to the Hy-Line lines. It is possible that KASP assays developed here could be used successfully with the Lohmann stock, but adjustments may have to be made for SNPs in the primer binding regions. It is also possible that shared ALVEs contain company-specific SNPs or even larger structural variation. Further sequencing is needed to identify the line and breed-specific differences within companies, as well as any company-specific variants. These may reflect differences in selective breeding programmes, including against P27 expression, and could result in varied phenotypic effects on the host.

8.3.2 Wider application of the identification pipelines

The two identification pipelines developed in this PhD project are not limited to the work presented here. As has been shown, LocaTR can be used to annotate LTR retrotransposon content in any assembled genome, providing the basis for evolutionary studies in any eukaryotic organism. In practical terms, researchers would need to adjust the list of reference sequences and select identification program parameters to best suit the chosen study species.

The ALVE identification pipeline also has a much wider potential application, again depending on the reference sequences chosen by the researcher. A modified ALVE identification pipeline could enable the identification of any viral insertion relative to a reference genome, in any species. Beyond the study of ERV integrations in agricultural

animals, the pipeline could therefore be applied to viral-induced cancer genomics, including in humans.

8.4 Concluding remarks

Compared with other eukaryotes, avian genomes contain a limited diversity of LTR retrotransposons, restricted solely to endogenous retroviruses. However, the work presented in this PhD thesis has increased the known abundance of these retroviral sequences in all sauropsid genomes analysed, and enabled a thorough characterisation of ERV distribution, expression and putative functions in the chicken genome. My research then focused on ALVE insertions, the youngest chicken ERVs, enabling a wider characterisation of their diversity across chicken populations, including an in-depth analysis of their location, intactness and frequency in the Hy-Line elite layer lines.

This work has both academic and practical applications for avian biology. These range from evolutionary studies into the role of transposable elements in the rapid diversification of birds, to directly advising commercial breeding companies about the phenotypic impacts of ALVE insertions and identifying which sites should be prioritised for eradication. Future work proposed in this thesis will further characterise the diverse roles of these repetitive elements, and expand the applications of the tools developed during this PhD project. As ALVEs remain a recurrent problem in commercial flocks, further understanding of these loci is critical for improving animal welfare and guaranteeing long term global food security.

References

Aarts, H. J. *et al.* (1991) Variations in endogenous viral gene patterns in White Leghorn, medium heavy, White Plymouth Rock, and Cornish Chickens, *Poultry Science*, 70(6), p. 1281–1286.

Aguilera, A. and Gómez-González, B. (2008) Genome instability: a mechanistic view of its causes and consequences, *Nature Reviews Genetics*, 9(3), p. 204–217.

Alkan, C., Coe, B. P. and Eichler, E. E. (2011) Genome structural variation discovery and genotyping., *Nature Reviews Genetics*, 12(5), p. 363–76.

Altschul, S. F. *et al.* (1990) Basic local alignment search tool., *Journal of molecular biology*, 215(3), p. 403–10.

Altschul, S. and Madden, T. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25(17), p. 3389–3402.

Andrews, S. (2012) FastQC. "A quality control tool for high throughput sequence data".

Arkhipova, I. R. *et al.* (1986) The steps of reverse transcription of Drosophila mobile dispersed genetic elements and U3-R-U5 structure of their LTRs., *Cell*, 44(4), p. 555–63.

Ashlock, W. and Datta, S. (2010) Using Fourier phase analysis on genomic sequences to identify retroviruses, *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology - BCB 10*. New York, New York, USA: ACM Press.

Astrin, S. M. (1978) Endogenous viral genes of the White Leghorn chicken: common site of residence and sites associated with specific phenotypes of viral gene expression., *PNAS*, 75(12), p. 5941–5945.

Astrin, S. M. and Robinson, H. (1979) Gs, an Allele of Chickens for Endogenous Avian Leukosis Viral Antigens, Segregates with ev 3, a Genetic Locus That Contains Structural Genes for Virus, *Journal of Virology*, 31(2), p. 420–425.

Aswad, A. and Katzourakis, A. (2012) Paleovirology and virally derived immunity., *Trends in ecology & evolution*, 27(11), p. 627–36.

Axelsson, E. *et al.* (2004) Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey., *Molecular Biology and Evolution*, 21(8), p. 1538–47.

Backström, N. *et al.* (2010) The recombination landscape of the zebra finch *Taeniopygia guttata* genome, *Genome Research*, 20(4), p. 485–95.

Bacon, L. D. *et al.* (1988) Association of the Slow Feathering (K) and an Endogenous Viral (ev21) Gene on the Z Chromosome of Chickens, *Poultry Science*, 67(2), p. 191–197.

Bacon, L. D., Hunt, H. D. and Cheng, H. H. (2000) A review of the development of chicken lines to resolve genes determining resistance to diseases., *Poultry Science*, 79(8), p. 1082–1093.

Bai, J., Payne, L. N. and Skinner, M. a (1995) HPRS-103 (exogenous avian leukosis virus, subgroup J) has an env gene related to those of endogenous elements EAV-0 and E51 and an E element found previously only in sarcoma viruses., *Journal of Virology*, 69(2), p. 779–84.

Baillie, J. K. *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain., *Nature*, 479(7374), p. 534–7.

Baker, M. (2011) Long noncoding RNAs: the search for function, *Nature Methods*, 8(5), p. 379–383.

Barrio, Á. M. *et al.* (2011) The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships., *PloS one*, 6(5), p. e19832.

Bawa, A. S. and Anilakumar, K. R. (2013) Genetically modified foods: safety, risks and public concerns — a review, *Journal of Food Scienc and Technology*, 50(6), p. 1035–1046.

Benachenhou, F. *et al.* (2009) Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data., *PloS one*, 4(4), p. e5179.

Benachenhou, F. *et al.* (2013) Conserved structure and inferred evolutionary history of long terminal repeats (LTRs), *Mobile DNA*, 4(5).

Benachenhou, F., Blikstad, V. and Blomberg, J. (2009) The phylogeny of orthoretroviral long terminal repeats (LTRs)., *Gene*, 448(2), p. 134–8.

Benkel, B. F. (1998) Locus-specific diagnostic tests for endogenous avian leukosis-type viral loci in chickens., *Poultry Science*, 77(7), p. 1027–35.

Benkel, B. and Rutherford, K. (2014) Endogenous avian leukosis viral loci in the Red Jungle Fowl genome assembly, *Poultry Science*, 93, p. 2988–2990.

Benson, S. *et al.* (1998) The unique envelope gene of the subgroup J avian leukosis virus derives from ev/J proviruses, a novel family of avian endogenous viruses, *Journal of Virology*, 72(45), p. 10157–64.

Bergero, R. and Charlesworth, D. (2009) The evolution of restricted recombination in sex chromosomes, *Trends in Ecology and Evolution*, 24(2), p. 94–102.

Bergman, C. M. and Quesneville, H. (2007) Discovering and detecting transposable elements in genome sequences., *Briefings in bioinformatics*, 8(6), p. 382–92.

Bickhart, D. M. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome, *Nature Genetics*, 49(4), p. 643–650.

Blyth, J. (1954) Notes on the poultry research centre flock of brown leghorns, *Worlds Poultry Science journal*, 10(2), p. 140–143.

Blyth, J. and Sang, J. (1960) Survey of line crosses in a Brown Leghorn flock, p. 408–421.

Bock, M. and Stoye, J. P. (2000) Endogenous retroviruses and the human germline., *Current opinion in genetics & development*, 10(6), p. 651–5.

Boeke, J. D. (2003) The unusual phylogenetic distribution of retrotransposons: a hypothesis., *Genome research*, 13(9), p. 1975–83.

de Boer, J. G. *et al.* (2007) Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids, *BMC Genomics*, 8(1), p. 422.

Bolisetty, M., Blomberg, J. and Benachenhou, F. (2012) Unexpected Diversity and Expression of Avian Endogenous Retroviruses, *MBio*, 3(5).

Borisenko, L. (2003) Avian endogenous retroviruses., *Folia biologica*, 49(5), p. 177–82.

Borysenko, L., Stepanets, V. and Rynditch, A. V (2008) Molecular characterization of full-length MLV-related endogenous retrovirus ChiRV1 from the chicken, Gallus gallus., *Virology*, 376(1), p. 199–204.

Boulliou, A. *et al.* (1991) Restriction fragment length polymorphism analysis of endogenous avian leukosis viral loci: determination of frequencies in commercial broiler lines., *Poultry Science*, 70(6), p. 1287-96.

Boyce-Jacino, M. T., ODonoghue, K. and Faras, A. J. (1992) Multiple complex families of endogenous retroviruses are highly conserved in the genus Gallus., *Journal of Virology*, 66(8), p. 4919–29.

Briggs, J. C. (2003) Fishes and Birds: Gondwana Life Rafts Reconsidered, *Systematic Biology*, 52(4), p. 548–53.

Bromham, L. (2002) The human zoo: endogenous retroviruses in the human genome, *Trends in Ecology & Evolution*, 17(2), p. 91–7.

Bu, G. *et al.* (2013) Characterization of the novel duplicated PRLR gene at the late-feathering K locus in Lohmann chickens., *Journal of molecular endocrinology*, 51(2), p. 261–76.

Bushman, F. (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons, *Cell*, 115, p. 135–8.

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(8), p. 521–25.

Cam, H. P. *et al.* (2008) Host genome surveillance for retrotransposons by transposon-derived proteins., *Nature*, 451(7177), p. 431–6.

Cao, H. *et al.* (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology, *GigaScience*, 3(1), p. 34.

Cao, W. *et al.* (2015) Further observations on serotype 2 Mareks disease virus-induced enhancement of spontaneous avian leukosis virus-like bursal lymphomas in *ALVA6* transgenic chickens, *Avian Pathology*. Taylor & Francis, 44(1), p. 23–7.

Capy, P. (2005) Classification and nomenclature of retrotransposable elements., *Cytogenetic and genome research*, 110, p. 457–61.

Carbone, L. *et al.* (2009) Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution, *PLoS Genetics*, 5(6).

Carré-Eusèbe, D., Coudouel, N. and Magre, S. (2009) OVEX1, a novel chicken endogenous retrovirus with sex-specific and left-right asymmetrical expression in gonads., *Retrovirology*, 6, p. 59.

Cavalier-Smith, T. (2005) Economy, speed and size matter: Evolutionary forces driving nuclear genome miniaturization and expansion, *Annals of Botany*, 95(1), p. 147–175.

Chaisson, M. J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory., *BMC Bioinformatics*, 13, p. 238.

Chang, C.-M. *et al.* (2006) Complete association between a retroviral insertion in the tyrosinase gene and the recessive white mutation in chickens., *BMC Genomics*, 7, p. 19.

Chang, C. M. *et al.* (2007) Quantitative effects of an intronic retroviral insertion on the transcription of the tyrosinase gene in recessive white chickens, *Animal Genetics*, 38(2), p. 162–7.

Chang, S. *et al.* (2015) Genetic susceptibility to and presence of endogenous avian leukosis viruses impose no significant impact on survival days of chickens challenged with very virulent plus Marek's disease virus, *Annals of Virology and Research*, 1(2), p. 1007.

Chaparro, C. *et al.* (2007) RetrOryza: A database of the rice LTR-retrotransposons, *Nucleic Acids Research*, 35, p. 66–70.

Charlesworth, B. (2009) Effective population size and patterns of molecular evolution and variation, *Nature Reviews Genetics*, 10(3), p. 195–205.

Chen, W. *et al.* (2014) Polymorphism of Avian Leukosis Virus Subgroup E Loci Showing Selective Footprints in Chicken, *Biochemical Genetics*, 52(11–12), p. 524–37.

Chen, Y. *et al.* (2013) VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue., *Bioinformatics*, 29(2), p. 266–7.

Chesters, P. M. *et al.* (2001) Acutely Transforming Avian Leukosis Virus Subgroup J Strain 966: Defective Genome Encodes a 72-Kilodalton Gag-Myc Fusion Protein, *Journal of Virology*, 75(9), p. 4219–25.

Chiu, Y.-L. and Greene, W. C. (2008) The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements., *Annual review of immunology*, 26, p. 317–53.

Chung, H.-C. *et al.* (2014) Inhibition of porcine endogenous retrovirus in PK15 cell line by efficient multitargeting RNA interference., *Transplant international*, 27(1), p. 96–105.

Chuong, E. B. *et al.* (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta., *Nature Genetics*, 45(3), p. 325–9.

Coffin, J. M. *et al.* (1983) Genomes of endogenous and exogenous avian retroviruses., *Virology*, 126(1), p. 51–72.

Cohen, C. J., Lock, W. M. and Mager, D. L. (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment., *Gene*, 448(2), p. 105–14.

Cole, R. K. (1966) Hereditary Hypothyroidism in the Domestic Fowl, *Genetics*, 53, p. 1021–1033.

Conklin, K. (1991) Activation of an endogenous retrovirus enhancer by insertion into a heterologous context., *Journal of Virology*, 65(5), p. 2525–32.

Conklin, K. F. *et al.* (1982) Role of methylation in the induced and spontaneous expression of the avian endogenous virus ev-1: DNA structure and gene products, *Molecular Cell Biology*, 2, p. 638–52

Cordaux, R. and Batzer, M. a (2009) The impact of retrotransposons on human genome evolution., *Nature Reviews Genetics*, 10(10), p. 691–703.

Crittenden, L. B., Smith, E. J. and Fadly, A. M. (1984) Influence of endogenous viral (ev) gene expression and strain of exogenous avian leukosis virus (ALV) on mortality and ALV infection and shedding in chickens., *Avian diseases*, 28(4), p. 1037–56.

Cui, J. *et al.* (2014) Low frequency of paleoviral infiltration across the avian phylogeny, *Genome Biology*, 15(12), p. 539.

Cusack, B. P. and Wolfe, K. H. (2007) Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates, *Molecular Biology and Evolution*, 24(3), p. 679–86

D'Souza-Li, L. (2006) The calcium-sensing receptor and related diseases, *Journal for the Brazilian Society of Endocrinology and Metabolism*, 50(4), p. 628–39.

Dalloul, R. A *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis., *PLoS biology*, 8(9).

Dawson, D. A. *et al.* (2007) Gene order and recombination rate in homologous chromosome regions of the chicken and a passerine bird., *Molecular Biology and Evolution*, 24(7), p. 1537–52.

Delany, M. *et al.* (2003) Telomeres in the chicken: genome stability and chromosome ends, *Poultry Science*, 82, p. 917–26.

Delcher, A. L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer, *Bioinformatics*, 23(6), p. 673–9.

Desfarges, S. and Ciuffi, A. (2010) Retroviral integration site selection., *Viruses*, 2(1), p. 111–30.

Dimcheff, D. and Drovetski, S. (2000) Cospeciation and horizontal transmission of avian sarcoma and leukosis virus gag genes in galliform birds, *Journal of Virology*, 74(9), p. 3984–95.

Dimcheff, D., Krishnan, M. and Mindell, D. (2001) Evolution and characterization of tetraonine endogenous retrovirus: a new virus related to avian sarcoma and leukosis viruses, *Journal of Virology*, 75(4), p. 2002–9.

Dimitrov, L. *et al.* (2016) Germline gene editing in chickens by efficient crispr-mediated homologous recombination in primordial germ cells, *PLoS ONE*, 11(4), p. 1–10.

Doolittle, R. and Feng, D. (1989) Origins and evolutionary relationships of retroviruses, *Quarterly Review of Biology*, 64(1), p. 1–30.

Dorus, S. *et al.* (2008) Recent origins of sperm genes in Drosophila, *Molecular Biology and Evolution*, 25(10), p. 2157–66.

Dunn, C. *et al.* (2005) Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5., *Gene*, 364, p. 2–12.

Dunn, C. *et al.* (2006) Transcription of two human genes from a bidirectional endogenous retrovirus promoter., *Gene*, 366(2), p. 335–42.

Dunnington, E. a and Siegel, P. B. (1996) Long-term divergent selection for eight-week body weight in white Plymouth rock chickens., *Poultry Science*, 75(10), p. 1168–79.

Dupressoir, A. *et al.* (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene, *PNAS*, 106(29), p. 12127–32.

Eddy, S. R. (2009) A New Generation of Homology Search Tools Based on Probabilistic Inference, *Genome Informatics*, 23(1), p. 205–11.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput., *Nucleic Acids Research*, 32(5), p. 1792–7.

Ekarius, C. (2007) *Storeys Illustrated Guide to Poultry Breeds*. Storey Publishing.

Elferink, M. *et al.* (2010) Regional differences in recombination hotspots between two chicken populations, *BMC Genetics*, 11(11).

Elferink, M. G. *et al.* (2008) Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken., *BMC Genomics*, 9, p. 391.

Ellegren, H. (2005) The avian genome uncovered., *Trends in ecology & evolution*, 20(4), p. 180–6.

Ellegren, H. (2010) Evolutionary stasis: the stable chromosomes of birds., *Trends in ecology & evolution*, 25(5), p. 283–91.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons., *BMC Bioinformatics*, 9, p. 18.

English, A. C., Salerno, W. J. and Reid, J. G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping., *BMC Bioinformatics*, 15, p. 180.

Fadly, A. (2000) Isolation and identification of avian leukosis viruses: a review, *Avian Pathology*, 29(6), p. 529–35.

Fadly, A. *et al.* (2014) Role of endogenous avian leukosis virus and serotype 2 Mareks disease virus in enhancement of spontaneous lymphoid-leukosis-like tumors in chickens, in *American Veterinary Medical Association Annual Convention, July 25-29, 2014, Denver, Colorado.*, p. Abstract No. 16756.

Fadly, A. M. and Smith, E. J. (1991) Influence of Maternal Antibody on Avian Leukosis Virus Infection in White Leghorn Chickens Harboring Endogenous Virus-21 (EV21), *Avian diseases*, 35(3), p. 443–51.

Fan, W. L. *et al.* (2013) Genome-wide patterns of genetic variation in two domestic chickens, *Genome Biology and Evolution*, 5(7), p. 1376–92.

Farkašová, H. *et al.* (2017) Discovery of an endogenous Deltaretrovirus in the genome of long-fingered bats (Chiroptera: Miniopteridae), *PNAS*, 114(12), p. 3145–50.

Faulkner, G. J. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells., *Nature Genetics*, 41(5), p. 563–71.

Feschotte, C. and Gilbert, C. (2012) Endogenous viruses: insights into viral evolution and impact on host biology., *Nature Reviews Genetics*, 13(4), p. 283–96.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) Structural variation in the human genome, *Nature Reviews Genetics*, 7(2), p. 85–97.

Flavell, A. J. *et al.* (1997) The evolution of Ty1-copia group retrotransposons in eukaryote genomes., *Genetica*, 100, p. 185–95.

Fox, W. and Smyth, J. R. J. (1985) The effects of recessive white and dominant white genotypes on early growth rate., *Poultry Science*, 64(3), p. 429–33.

Frewer, L. *et al.* (2014) Attitudes towards genetically modified animals in food production, *British Food Journal*, 116(8), p. 1291–1313.

Frewer, L. J. *et al.* (2013) Public perceptions of agri-food applications of genetic modification - A systematic review and meta-analysis, *Trends in Food Science & Technology*, 30(2), p. 142–52.

Fridolfsson, A.-K. K. *et al.* (1998) Evolution of the avian sex chromosomes from an ancestral pair of autosomes., *PNAS*, 95(14), p. 8147–52.

Frisby, D. P. *et al.* (1979) The distribution of endogenous chicken retrovirus sequences in the DNA of galliform birds does not coincide with avian phylogenetic relationships., *Cell*, 17(3), p. 623–34.

Fulton, J. E. *et al.* (2016) A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex B region between BG2 and CD1A1, *Genetics Selection Evolution*, 48(1), p. 1–15.

Gagnieur, L. *et al.* (2014) Analysis by high throughput sequencing of Specific Pathogen Free eggs, *Biologicals*, 42(4), p. 218–9.

Ganapathy, G. *et al.* (2014) High-coverage sequencing and annotated assemblies of the budgerigar genome, *GigaScience*, 3(1), p. 11.

Garcia-Etxebarria, K. and Jugo, B. M. (2010) Genome-wide detection and characterization of endogenous retroviruses in Bos taurus., *Journal of Virology*, 84(20), p. 10852–62.

Garcia-Etxebarria, K. and Jugo, B. M. (2012) Detection and characterization of endogenous retroviruses in the horse genome by in silico analysis., *Virology*, 434, p. 59–67.

Garcia-Etxebarria, K., Sistiaga-Poveda, M. and Jugo, B. (2014) Endogenous Retroviruses in Domestic Animals, *Current Genomics*, 15, p. 256–65.

Gavora, J. S. *et al.* (1991) Endogenous Viral Genes: Association with Reduced Egg Production Rate and Egg Size in White Leghorns, *Poultry Science*, 70(3), p. 618–23.

Gavora, J. S. *et al.* (1995) Endogenous viral genes influence infection with avian leukosis virus., *Avian Pathology*, 24(4), p. 653–64.

Gering, E. *et al.* (2015) Mixed ancestry and admixture in Kauais feral chickens: Invasion of domestic genes into ancient Red Junglefowl reservoirs, *Molecular Ecology*, 24(9), p. 2112–24.

Gilbert, C. *et al.* (2009) Parallel germline infiltration of a lentivirus in two Malagasy lemurs., *PLoS genetics*, 5(3), p. e1000425.

Ginzburg, L. R., Bingham, P. M. and Yoo, S. (1984) On the theory of speciation induced by transposable elements, *Genetics*, 107(2), p. 331–41.

Gogvadze, E. and Buzdin, A. (2009) Retroelements and their impact on genome evolution and functioning., *Cellular and molecular life sciences*, 66(23), p. 3727–42.

Gong, R. *et al.* (2005) Structural characterization of the fusion core in syncytin, envelope protein of human endogenous retrovirus family W., *Biochemical and biophysical research communications*, 331(4), p. 1193–1200.

Gonzalez-Garay, M. (2015) Introduction to Isoform Sequencing Using Pacific Biosciences Technology (IsoSeq), in *Translational Bioinformatics: Transcriptomics and Gene Regulation*, p. 141–160.

Goodier, J. L. (2016) Restricting retrotransposons: a review., *Mobile DNA*, 7:16.

Goodwin, T. J. and Poulter, R. T. (2001) The DIRS1 group of retrotransposons., *Molecular Biology and Evolution*, 18(11), p. 2067–82.

Grawenhoff, J. and Engelman, A. N. (2017) Retroviral integrase protein and intasome nucleoprotein complex structures, *World Journal of Biological Chemistry*, 26(81), p. 32–44.

Greenwold, M. J. *et al.* (2014) Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles, *BMC Evolutionary Biology*, 14(1), p. 249.

Greenwold, M. J. and Sawyer, R. H. (2010) Genomic organization and molecular phylogenies of the beta (beta) keratin multigene family in the chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*): implications for feather evolution., *BMC Evolutionary Biology*, 10(1), p. 148.

Gregory, T. R. (2005) Synergy between sequence and size in Large-scale genomics, *Nature Reviews Genetics*, 6(9), p. 699–708.

Gregory, T. R. *et al.* (2009) The smallest avian genomes are found in hummingbirds, *Proceedings of the Royal Society B: Biological Sciences*, 276(1674), p. 3753–7.

Gregory, T. R. (2017) Animal Genome Size Database. http://www.genomesize.com.

Gremme, G., Steinbiss, S. and Kurtz, S. (2013) Genome Tools: a comprehensive software library for efficient processing of structured genome annotations, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3), p. 645–56.

Griffin, D. K. *et al.* (2008) Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution., *BMC Genomics*, 9, p. 168.

Griffin, H. *et al.* (1989) Plasma lipoprotein metabolism in lean and in fat chickens produced by divergent selection for plasma very low density lipoprotein concentration., *Journal of lipid research*, 30(8), p. 1243–50.

Griffin, H. D., Windsor, D. and Whitehead, C. C. (1991) Changes in lipoprotein metabolism and body composition in chickens in response to divergent selection for plasma very low density lipoprotein concentration., *British Poultry Science*, 32(1), p. 195–201.

Grunder, A. A. *et al.* (1995) Characterization of eight endogenous viral (ev) genes of meat chickens in semi-congenic lines., *Poultry Science*, 74(9), p. 1506–14.

Gudkov, A. *et al.* (1986) Genetic structure of the endogenous proviruses and expression of the gag gene in Brown Leghorn chickens, *Folia Biologica*, 32(1), p. 65–72.

Guo, X. *et al.* (2016) Whole-genome resequencing of Xishuangbanna fighting chicken to identify signatures of selection, *Genetics Selection Evolution*, 48(1), p. 62.

Haley, C. S. and Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers., *Heredity*, 69(4), p. 315–24.

Haley, C. S., Knott, S. A. and Elsen, J. M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares, *Genetics*, 136(3), p. 1195–1207.

Hamoen, F. F. *et al.* (2001) Detection of genes on the Z-chromosome affecting growth and feathering in broilers., *Poultry Science*, 80(5), p. 527–34.

Han, M. V *et al.* (2009) Adaptive evolution of young duplicated genes in mammals, *Genome Research*, 19, p. 859–67.

Hanken, J. and Wake, D. B. (1993) Miniaturization of Body Size: Organismal Consequences and Evolutionary Significance, *Annual Review of Ecology and Systematics*, 24(1), p. 501–19.

Havecker, E., Gao, X. and Voytas, D. (2004) The diversity of LTR retrotransposons, *Genome Biology*, 5(6), p. 225.

Havenstein, G. B., Ferket, P. R. and Qureshi, M. a (2003) Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets., *Poultry Science*, 82(10), p. 1500–8.

Hedges, S. *et al.* (1996) Continental breakup and the ordinal diversification of birds and mammals, *Nature*, 381, p. 226-9.

Heidmann, O. *et al.* (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new “syncytin” in a third order of mammals, *Retrovirology*, 6(1), p. 107.

Helm-Bychowski, K. M. and Wilson, A. C. (1986) Rates of nuclear DNA evolution in pheasant-like birds: Evidence from restriction maps, *PNAS*, 83(3), p. 688–92.

Henzy, J. E. *et al.* (2014) A novel recombinant retrovirus in the genomes of modern birds combines features of avian and mammalian retroviruses., *Journal of Virology*, 88(5), p. 2398–405.

Herniou, E. *et al.* (1998) Retroviral diversity and distribution in the vertebrates., *Journal of Virology*, 72, p. 5955-66.

Hillier, L. *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution., *Nature*, 432(7018), p. 695–716.

Hocking, P. M. and Guggenheim, J. A. (2014) The chick as an animal model of eye disease, *Drug Discovery Today: Disease Models*, 10(4), p. e225–30.

Howe, K. and Wood, J. M. (2015) Using optical mapping data for the improvement of vertebrate genome assemblies, *GigaScience*, 4(1), p. 10.

Hu, X., Zhu, W., Chen, S., Liu, Y., Sun, Z., Geng, T., Wang, X., *et al.* (2016) Expression of the env gene from the avian endogenous retrovirus ALVE and regulation by miR-155, *Archives of Virology*, 161(6), p. 1623–32.

Hu, X., Zhu, W., Chen, S., Liu, Y., Sun, Z., Geng, T., Song, C., *et al.* (2016) Expression patterns of endogenous avian retrovirus ALVE1 and its response to infection with exogenous avian tumour viruses, *Archives of Virology*. 162(1), p. 89-101.

Hu, Y. *et al.* (2013) Comparison of the genome-wide DNA methylation profiles between fast-growing and slow-growing broilers., *PloS one*, 8(2), p. e56411.

Huang, Y. *et al.* (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species, *Nature Genetics*, 45(7), p. 776–83.

Huda, A. *et al.* (2008) Endogenous retroviruses of the chicken genome., *Biology Direct*, 3(9).

Hughes, A. L. and Hughes, M. K. (1995) Small genomes for better flyers, *Nature*, 5, p. 391.

Hughes, J. F. and Coffin, J. M. (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution., *Nature Genetics*, 29(4), p. 487–489.

Hunt, H. *et al.* (2008) Survey of endogenous virus and TVB* receptor status of commercial chicken stocks supplying specific-pathogen-free eggs., *Avian diseases*, 52(3), p. 433–40.

Hurst, T. and Magiorkinis, G. (2014) Activation of the innate immune response by endogenous retroviruses, *Journal of General Virology*, 96(6), p. 1207-18.

Iraqi, F. and Smith, E. J. (1995) Organization of the sex-linked late-feathering haplotype in chickens., *Animal Genetics*, 26(3), p. 141–6.

Iraqi, F., Soller, M. and Beckmann, J. S. (1991) Distribution of endogenous viruses in some commercial chicken layer populations., *Poultry Science*, 70(4), p. 665–79.

Isbel, L. and Whitelaw, E. (2012) Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes., *BioEssays*, 34(9), p. 734–8.

Ito, J. *et al.* (2013) Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses., *Journal of Virology*, 87(22), p. 12029–40.

Jacques, P.-É., Jeyakani, J. and Bourque, G. (2013) The majority of primate-specific regulatory sequences are derived from transposable elements., *PLoS genetics*, 9(5), p. e1003504.

Jadin, L. *et al.* (2008) Skeletal and hematological anomalies in HYAL2-deficient mice: a second type of mucopolysaccharidosis IX?, *The FASEB Journal*, 22(12), p. 4316–26.

Janes, D. E. *et al.* (2010) Genome Evolution in Reptilia, the Sister Group of Mammals, *Annual Review of Genomics: Human Genetics*, 11, p. 239–64.

Jang, H.-M. *et al.* (2014) Genome resequencing and bioinformatic analysis of SNP containing candidate genes in the autoimmune vitiligo Smyth line chicken model., *BMC Genomics*, 15(1), p. 707.

Jarvis, E. D. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds, *Science*, 346(6215), p. 1126–38.

Jarvis, E. D. *et al.* (2015) Phylogenomic analyses data of the avian phylogenomics project., *GigaScience*, 4, p. 4.

Jern, P. and Coffin, J. M. (2008) Effects of retroviruses on host genome function, *Annual review of genetics*, 42, p. 709–32.

Jern, P., Sperber, G. O. and Blomberg, J. (2005) Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy., *Retrovirology*, 2(50).

Jiao, W.-B. *et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data, *Genome Research*, 27(5), p. 778–86.

Jurka, J. *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements., *Cytogenetic and genome research*, 110, p. 462–7.

Justice, J. and Beemon, K. L. (2013) Avian retroviral replication., *Current opinion in virology*, 3, p. 1–6.

Ka, S. *et al.* (2009) Proviral integrations and expression of endogenous avian leucosis virus during long term selection for high and low body weight in two chicken lines., *Retrovirology*, 6, p. 68.

Kaessmann, H., Vinckenbosch, N. and Long, M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights., *Nature Reviews Genetics*, 10(1), p. 19–31.

Käll, L., Krogh, A. and Sonnhammer, E. L. (2004) A Combined Transmembrane Topology and Signal Peptide Prediction Method, *Journal of Molecular Biology*, 338(5), p. 1027–36.

Kalyanaraman, A. and Aluru, S. (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons, *Journal of Bioinformatics and Computational Biology*, 4(2), p. 197–216.

Kan, X.-Z. *et al.* (2010) Phylogeny of major lineages of galliform birds (Aves: Galliformes) based on complete mitochondrial genomes., *Genetics and molecular research*, 9(3), p. 1625–33.

Kanda, R., Tristem, M. and Coulson, T. (2013) Exploring the effects of immunity and life history on the dynamics of an endogenous retrovirus, *Philosophical Transactions of The Royal Society B: Biological Sciences*, 368(1626).

Kansaku, N. *et al.* (2011) Sequence Characterization of K-gene Linked Region in Various Chicken Breeds, *Journal of Poultry Science*, 48, p. 181–6.

Kapusta, A. *et al.* (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs., *PLoS genetics*, 9(4), p. e1003470.

Kapusta, A. and Suh, A. (2017) Evolution of bird genomes-a transposons-eye view, *Annals of the New York Academy of Sciences*, 1389, p. 164–85.

Kapusta, A., Suh, A. and Feschotte, C. (2017) Dynamics of genome size evolution in birds and mammals, *PNAS*, p. E1460-9.

Katz, R. A. and Skalka, A. M. (1990) Generation of diversity in retroviruses., *Annual review of genetics*, 24, p. 409–45.

Katzourakis, A. *et al.* (2007) Discovery and analysis of the first endogenous lentivirus., *PNAS*, 104(15), p. 6261–5.

Katzourakis, A. and Gifford, R. J. (2010) Endogenous viral elements in animal genomes., *PLoS genetics*, 6(11), p. e1001191.

Katzourakis, A., Rambaut, A. and Pybus, O. G. (2005) The evolutionary dynamics of endogenous retroviruses., *Trends in microbiology*, 13(10), p. 463–8.

Kawakami, T. *et al.* (2014) A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution, *Molecular Ecology*, 23(16), p. 4035–58.

Kearse, M. *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, *Bioinformatics*, 28(12), p. 1647–9.

Khosravinia, H. (2009) Effect of the slow (K) or rapid (k +) feathering gene on carcass-related traits of broiler chickens selected for breast and thighs weight, *Russian Journal of Genetics*, 45(1), p. 98–104.

Kijima, T. E. and Innan, H. (2010) On the estimation of the insertion time of LTR retrotransposable elements., *Molecular Biology and Evolution*, 27(4), p. 896–904.

Kim, D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions., *Genome Biology*, 14(4), p. R36.

Kordis, D. (2005) A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses., *Gene*, 347(2), p. 161–73.

Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review., *Plant methods*. Plant Methods, 9(1), p. 29.

Kozak, C. (2014) Origins of the Endogenous and Infectious Laboratory Mouse Gammaretroviruses, *Viruses*, 7(1), p. 1–26.

Kranis, A. *et al.* (2013) Development of a high density 600K SNP genotyping array for chicken., *BMC Genomics*, 14(1), p. 59.

Krueger, F. (2013) Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Buisulfite-Seq) libraries.

Kuehn, L. *et al.* (2006) Antibody response of chickens to sheep red blood cells: Crosses among divergently selected lines and relaxed sublines, *Poultry Science*, 85(8), p. 1338.

Kuhnlein, U. *et al.* (1989) Influence of Selection for Egg Production and Mareks Disease Resistance on the Incidence of Endogenous Viral Genes in White Leghorns, *Poultry Science*, 68(9), p. 1161–7.

Kuo, R. *et al.* (2017) Normalized long-read RNA sequencing in chicken reveals transcriptome complexity similar to human, *BMC Genomics*. BMC Genomics, 18(323), p. 1–19.

Langille, M. G. I. and Clark, D. V. (2007) Parent genes of retrotransposition-generated gene duplicates in Drosophila melanogaster have distinct expression profiles, *Genomics*, 90(3), p. 334–43.

Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2., *Nature Methods*, 9(4), p. 357–9.

Lavialle, C. *et al.* (2013) Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation, *Philosophical Transactions of the Royal Society B: Biological sciences*, 368(20120507).

Lepperdinger, G., Müllegger, J. and Kreil, G. (2001) Hyal2 - Less active, but more versatile?, *Matrix Biology*, 20(8), p. 509–14.

Lerat, E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs., *Heredity*, 104(6), p. 520–33.

Leung, D. C. and Lorincz, M. C. (2012) Silencing of endogenous retroviruses: when and why do histone marks predominate?, *Trends in biochemical sciences*, 37(4), p. 127–33.

Levin, E. and Smith, E. J. (1990) Molecular Analysis of Endogenous Virus ev21-Slow Feathering Complex of Chickens. 1. Cloning of Proviral-Cell Junction Fragment and Unoccupied Integration Site, *Poultry Science*, 69, p. 2017–2026.

Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 25(16), p. 2078-9.

Li, H. (2011a) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*, 27(21), p. 2987-93.

Li, H. (2011b) Improving SNP discovery by base alignment quality, *Bioinformatics*, 27(8), p. 1157–8.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv*, 1303.3997.

Li, H. (2015) BFC: Correcting Illumina sequencing errors, *Bioinformatics*, 31(17), p. 2885–2887. doi: 10.1093/bioinformatics/btv290.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform., *Bioinformatics*, 25(14), p. 1754–60.

Li, J. *et al.* (2012) Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance., *Genome research*, 22(5), p. 870–84.

Li, J.-W. *et al.* (2013) ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution., *Bioinformatics*, 29(5), p. 649–51.

Liu, C. *et al.* (2011) Detection and molecular characterization of recombinant avian leukosis viruses in commercial egg-type chickens in China., *Avian Pathology,* 40(3), p. 269–75.

Llorens, C. *et al.* (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees., *Biology Direct*, 4, p. 41.

Llorens, C. *et al.* (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0., *Nucleic Acids Research*, 39, p. D70-4.

Llorens, C., Fares, M. a and Moya, A. (2008) Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three king's hypothesis., *BMC Evolutionary Biology*, 8, p. 276.

Lowe, T. M. and Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence., *Nucleic Acids Research*, 25(5), p. 955–64.

Lucht, J. M. (2015) Public Acceptance of Plant Biotechnology and GM Crops, *Viruses*, 7, p. 4254–81.

Luo, C. *et al.* (2012) Differences of Z chromosome and genomic expression between early- and late-feathering chickens, *Molecular Biology Reports*, 39(5), p. 6283–8.

Lynch, C. and Tristem, M. (2003) A Co-opted gypsy-type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats, *Current Biology*, 13, p. 1518–23.

Magiorkinis, G. *et al.* (2012) Env-less endogenous retroviruses are genomic superspreaders., *PNAS*, 109(19), p. 7385–90.

Magiorkinis, G., Belshaw, R. and Katzourakis, A. (2013) "There and back again": revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era,Philosophical Transcations of The Royal Society B: Biological Sciences, 368, p. 20120504.

Mak, A. C. Y. *et al.* (2016) Genome-wide structural variation detection by genome mapping on nanochannel arrays, *Genetics*, 202(1), p. 351–62.

Malik, H. and Eickbush, T. (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons, *Journal of Virology*, 73(6), p. 5186–90.

Malik, H. S. and Eickbush, T. H. (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses., *Genome Research*, 11(7), p. 1187–97.

Malik, H. S., Henikoff, S. and Eickbush, T. H. (2000) Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses, *Genome Research*, 10(9), p. 1307–18.

Marco, A. and Marín, I. (2009) CGIN1: A retroviral contribution to mammalian genomes, *Molecular Biology and Evolution*, 26(10), p. 2167–70.

Martin, J. *et al.* (1997) Human Endogenous Retrovirus Type I-Related Viruses Have an Apparently Widespread Distribution within Vertebrates, *Journal of Virology*, 71(1), p. 437–43.

Martin, J. *et al.* (1999) Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses, *Journal of Virology*, 73(3), p. 2442–9.

Martin, J. A. and Wang, Z. (2011) Next-generation transcriptome assembly, *Nature Reviews Genetics*, 12(10), p. 671–82.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads.

Masabanda, J. S. *et al.* (2004) Molecular Cytogenetic Definition of the Chicken Genome: The First Complete Avian Karyotype, *Genetics*, 166, p. 1367-73.

Mason, A. S. *et al.* (2016) A new look at the LTR retrotransposon content of the chicken genome, *BMC Genomics*. BMC Genomics, 17(1), p. 688.

Matsubara, K. *et al.* (2006) Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes, *PNAS*, 103(48), p. 18190-5.

Matsumine, H., Herbst, M. and Ou, S. (1991) Aromatase mRNA in the extragonadal tissues of chickens with the henny-feathering trait is derived from a distinctive promoter structure that contains a segment of a retroviral long terminal repeat, *Journal of Biological Chemistry*, 266(30), p. 19900–7.

Mattick, J. S., Taft, R. J. and Faulkner, G. J. (2010) A global view of genomic information--moving beyond the gene and the master regulator., *Trends in genetics*, 26(1), p. 21–8.

McCarthy, E. M. *et al.* (2002) Long terminal repeat retrotransposons of *Oryza sativa*., *Genome biology*, 3(10), p. R53.

McCarthy, E. M. and McDonald, J. F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons, *Bioinformatics*, 19(3), p. 362–7.

McCarthy, E. M. and McDonald, J. F. (2004) Long terminal repeat retrotransposons of *Mus musculus*., *Genome biology*, 5(3), p. R14.

McKay, J. (2009) The Genetics of Modern Commercial Poultry, in Hocking, P. (ed.) *Biology of Breeding Poultry - Poultry Science Symposium Series Volume Twenty-nine*. CAB International, p. 3–9.

Meiklejohn, K. A. *et al.* (2014) Incongruence among different mitochondrial regions: A case study using complete mitogenomes, *Molecular Phylogenetics and Evolution*, 78(1), p. 314–23.

Meisel, R. and Connallon, T. (2013) The faster-X effect: integrating theory and data, *Trends in genetics*, 29(9), p. 537–44.

Meisler, M. H. and Ting, C.-N. (1993) The Remarkable Evolutionary History of the Human Amylase Genes, *Critical Reviews in Oral Biology & Medicine*, 4(3), p. 503–9.

Mercer, T. R., Dinger, M. E. and Mattick, J. S. (2009) Long non-coding RNAs: insights into functions, *Nature Reviews Genetics*, 10(3), p. 155–9.

Mi, S. *et al.* (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis, *Nature*, 403(6771), p. 785–9.

Michael, T. P. and VanBuren, R. (2015) Progress, challenges and the future of crop genomes, *Current Opinion in Plant Biology*, 24, p. 71–81.

Mitchell, A. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years, *Nucleic Acids Research*, 43(D1), p. D213–21.

Mohr, D. *et al.* (2017) Improved de novo Genome Assembly: Synthetic long read sequencing combined with optical mapping produce a high quality mammalian genome at relatively low cost, *bioRxiv*, p. 128348.

Muir, W. M. *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds, *PNAS*, 105(45), p. 17312–7.

Nagano, T. *et al.* (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure, *Nature*, 502(7469), p. 59–64.

Nam, K. and Ellegren, H. (2008) The chicken (Gallus gallus) Z chromosome contains at least three nonlinear evolutionary strata, *Genetics*, 180(2), p. 1131–6.

Narezkina, A. *et al.* (2004) Genome-Wide Analyses of Avian Sarcoma Virus Integration Sites, *Journal of Virology*, 78(21), p. 11656–63.

Nätt, D. *et al.* (2012) Heritable genome-wide variation of gene expression and promoter methylation between wild and domesticated chickens., *BMC Genomics*, 13(1), p. 59.

Ng, C. S. *et al.* (2014) Genomic organization, transcriptomic analysis, and functional characterization of avian α- and β-keratins in diverse feather forms, *Genome Biology and Evolution*, 6(9), p. 2258–73.

Norton, P. A. and Coffin, J. M. (1987) Sarcoma Virus Sequences Essential for Viral Gene Expression, *Journal of Virology*, 61(4), p. 1171–9.

Nuzhdin, S. V. (1999) Sure facts, speculations, and open questions about the evolution of transposable element copy number, *Genetica*, 107(1–3), p. 129–137.

Nystedt, B. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution, *Nature*, 496(7451), p. 579-84.

Oh, D. *et al.* (2016) Whole Genome Re-Sequencing of Three Domesticated Chicken Breeds., *Zoological science*, 33(1), p. 73–7.

Ohashi, K. *et al.* (1998) Expression of bcl-2 and bcl-x genes in lymphocytes and tumor cell lines derived from MDV-infected chickens., *Acta virologica*, 43(2–3), p. 128–132.

Oishi, I. *et al.* (2016) Targeted mutagenesis in chicken using CRISPR/Cas9 system., Nature *Scientific Reports*, 6, p. 23980.

Ono, R. *et al.* (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality, *Nature Genetics*, 38(1), p. 101–6.

Organ, C. L. *et al.* (2007) Origin of avian genome size and structure in non-avian dinosaurs, *Nature*, 446(7132), p. 180–4.

Ozsolak, F. and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities, *Nature Reviews Genetics*, 12(2), p. 87–98.

Paces, J. *et al.* (2004) HERVd: the Human Endogenous RetroViruses Database: update., *Nucleic Acids Research*, 32, p. D50.

Patel, M. R., Emerman, M. and Malik, H. S. (2012) Paleovirology - Ghosts and gifts of the past, *Current Opinion in Virology*, 1(4), p. 304–9.

Pavlidis, H. O. *et al.* (2007) Divergent selection for ascites incidence in chickens, *Poultry Science*, 86(12), p. 2517–29.

Payne, L. N. *et al.* (1991) A novel subgroup of exogenous avian leukosis virus in chickens, *Journal of General Virology*, 72(4), p. 801–7.

Payne, L. N. *et al.* (1992) Host range of Rous sarcoma virus pseudotype RSV(HPRS-103) in 12 avian species: Support for a new avian retrovirus envelope subgroup, designated J, *Journal of General Virology*, 73(11), p. 2995–7.

Payne, L. N. (1998) Retrovirus-induced disease in poultry., *Poultry Science*, 77(8), p. 1204–12.

Payne, L. N. and Nair, V. (2012) The long view: 40 years of avian leukosis research., *Avian Pathology*, 41(1), p. 11–9.

Peterson-Burch, B. D. and Voytas, D. F. (2002) Genes of the Pseudoviridae (Ty1/copia retrotransposons)., *Molecular Biology and Evolution*, 19(11), p. 1832–45.

Phillips, M. J. *et al.* (2010) Tinamous and moa flock together: Mitochondrial genome sequence analysis reveals independent losses of flight among ratites, *Systematic Biology*, 59(1), p. 90–107.

Piednoël, M. *et al.* (2011) Eukaryote DIRS1-like retrotransposons: an overview., *BMC genomics*, 12(1), p. 621.

Polavarapu, N., Bowen, N. J. and McDonald, J. F. (2006) Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses., *Genome Biology*, 7(6), p. R51.

Pond, S. L. K. *et al.* (2008) A Maximum Likelihood Method for Detecting Directional Evolution in Protein Sequences and Its Application to Influenza A Virus, *Molecular biology and evolution*, 25(9), p. 1809-24.

Pond, S. L. K. and Frost, S. D. W. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments, *Bioinformatics*, 21(10), p. 2531–3.

Pontier, D. B. and Gribnau, J. (2011) Xist regulation and function eXplored, *Human Genetics*, 130(2), p. 223–36.

Pontius, J. U. *et al.* (2007) Initial sequence and comparative analysis of the cat genome, *Genome Research*, 17(11), p. 1675–89.

Poulter, R. T. M. and Goodwin, T. J. D. (2005) DIRS-1 and the other tyrosine recombinase retrotransposons., *Cytogenetic and genome research*, 110(1–4), p. 575–88.

Promislow, D. E., Jordan, I. K. and McDonald, J. F. (1999) Genomic demography: a life-history analysis of transposable element evolution., Proceedings of The Royal Society B: Biological Sciences, 266(1428), p. 1555–60.

Pruitt, K. *et al.* (2002) The Reference Sequence (RefSeq) Database, in *The NCBI Handbook*. Bethseda (MD): National Library of Medicine (US), National Center for Biotechnology Information.

Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features., *Bioinformatics*, 26(6), p. 841–2.

Reiss, D. *et al.* (2010) Variable DNA methylation of transposable elements: the case study of mouse Early Transposons., *Epigenetics*, 5(1), p. 68–79. A

Reiss, D. and Mager, D. L. (2007) Stochastic epigenetic silencing of retrotransposons: does stability come with age?, *Gene*, 390(1–2), p. 130–5.

Rho, M. *et al.* (2007) De novo identification of LTR retrotransposons in eukaryotic genomes., *BMC Genomics*, 8, p. 90.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite, *Trends in genetics*, 16(6), p. 276–7.

Rigal, M. and Mathieu, O. (2011) A "mille-feuille" of silencing: epigenetic control of transposable elements., *Biochimica et biophysica acta*, 1809(8), p. 452–8.

Roberts, V. (2009) *British Poultry Standards*.

Robinson, H. L. *et al.* (1981) Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections., *Journal of Virology*, 40(3), p. 745–51.

Romanish, M. T. *et al.* (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution., *PLoS genetics*, 3(1), p. e10.

Romanish, M. T., Cohen, C. J. and Mager, D. L. (2010) Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer, *Seminars in Cancer Biology*, 20(4), p. 246–53.

Roth, G., Nishikawa, K. C. and Wake, D. B. (1997) Genome size, secondary simplification, and the evolution of the brain in salamanders., *Brain, behavior and evolution*, 50(1), p. 50–9.

Rowe, H. M. *et al.* (2013) De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET., *Development*, 140(3), p. 519–29.

Rowe, H. M. and Trono, D. (2011) Dynamic control of endogenous retroviruses during development., *Virology*, 411(2), p. 273–87.

Rozen, S. and Skaletsky, H. (2000) *Primer3 on the WWW for general users and for biologist programmers.*, *Methods in molecular biology (Clifton, N.J.)*.

Rubin, C.-J. *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication., *Nature*, 464(7288), p. 587–91.

Ruddell, A. (1995) Transcription regulatory elements of the avian retroviral long terminal repeat., *Virology*, 206(1), p. 1–7.

Rutherford, K. *et al.* (2016) Discovery of an expended set of avian leukosis subroup E proviruses in chickens using Vermillion, a novel sequence capture and analysis pipeline, *Poultry Science*, 95, p. 2250–8.

Rutherford, K., McLean, N. and Benkel, B. (2013) A Rapid Profiling Assay for Avian Leukosis Virus Sub-group E Proviruses in Chickens., *Avian Diseases*, 58(1), p. 34-8..

Sabour, M. P. *et al.* (1992) Endogenous Viral Gene Distribution in Populations of Meat-Type Chickens, *Poultry Science*, 71(8), p. 1259–70.

Sacco, M. *et al.* (2004) Assessing the roles of endogenous retrovirus EAV-HP in avian leukosis virus subgroup J emergence and tolerance, *Journal of Virology*, 78(19), p. 10525–35.

Sacco, M. A. and Nair, V. K. (2014) Prototype EAV Endogenous Retroviruses of the Gallus Genus., *The Journal of general virology*, 5, p. 1–27.

Sacco, M. and Flannery, D. (2000) Avian endogenous retrovirus EAV-HP shares regions of identity with avian leukosis virus subgroup J and the avian retrotransposon ART-CH, *Journal of Virology*, 74(3), p. 1296–1306.

Sacco, M., Howes, K. and Venugopal, K. (2001) Intact EAV-HP endogenous retrovirus in Sonnerat's jungle fowl, *Journal of Virology*, 75(4), p. 2029–32.

Sacco, M. and Venugopal, K. (2001) Segregation of EAV-HP ancient endogenous retroviruses within the chicken population, *Journal of Virology*, 75(23), p. 11935-8.

Schatz, M. C., Witkowski, J. and McCombie, W. R. (2012) Current challenges in de novo plant genome sequencing and assembly, *Genome Biology*, 13(4), p. 243.

Schelhorn, S.-E. *et al.* (2013) Sensitive detection of viral transcripts in human tumor transcriptomes., *PLoS computational biology*, 9(10), p. e1003228.

Sekita, Y. *et al.* (2008) Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta, *Nature Genetics*, 40(2), p. 243–8.

Semagn, K. *et al.* (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement, *Molecular Breeding*, 33(1), p. 1–14.

Serrao, E. *et al.* (2015) Key determinants of target DNA recognition by retroviral intasomes., *Retrovirology*, 12, p. 39.

Shedlock, A. M. (2006) Phylogenomic Investigation of CR1 LINE Diversity in Reptiles, *Systematic Biology*, 55(6), p. 902–11.

Shedlock, A. M. *et al.* (2007) Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome, *PNAS*, 104(8), p. 2767–72.

Shi, L. *et al.* (2016) Long-read sequencing and de novo assembly of a Chinese genome, *Nature Communications*, 7, p. 12065.

Smeds, L. *et al.* (2014) Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes., *Nature communications*, 5, p. 5448.

Smit, A., Hubley, R. and Green, P. (2013) RepeatMasker Open-4.0.3.

Smith, A. and Benkel, B. F. (2008) A diagnostic assay for the endogenous ALV-type provirus *ALVE-NSAC*-3 of chickens, *Animal Genetics*, 39(5), p. 574–5.

Smith, A and Benkel, B. F. (2009) Novel avian leukosis virus-related endogenous proviruses from layer chickens: characterization and development of locus-specific assays., *Poultry Science*, 88(8), p. 1580–5.

Smith, E. J. *et al.* (1991) The Influence of ev6 on the Immune Response to Avian Leukosis Virus Infection in Rapid-Feathering Progeny of Slow- and Rapid-Feathering Dams,*Poultry Science*,70(8),p.1673-8.

Smith, E. J., Fadly, A. M. and Crittenden, L. B. (1990a) Interactions Between Endogenous Virus Loci ev6 and ev21.: 1. Immune Response to Exogenous Avian Leukosis Virus Infection, *Poultry Science*, 69(8), p. 1244–1250.

Smith, E. J., Fadly, A. M. and Crittenden, L. B. (1990b) Interactions Between Endogenous Virus Loci ev6 and ev21.: 2. Congenital Transmission of EV21 Viral Product to Female Progeny from Slow-Feathering Dams, *Poultry Science*, 69(8), p. 1251–6.

Smith, E. J., Fadly, A. and Okazaki, W. (1979) An enzyme-linked immunosorbent assay for detecting avian leukosis-sarcoma viruses., *Avian diseases*, 23(3), p. 698–707.

Smith, L. M. *et al.* (1999) Novel endogenous retroviral sequences in the chicken genome closely related to HPRS-103 (subgroup J) avian leukosis virus, *Journal of General Virology*, 80(1), p. 261–8.

Sperber, G. O. *et al.* (2007) Automated recognition of retroviral sequences in genomic data--RetroTector., *Nucleic Acids Research*, 35(15), p. 4964–76.

Sreekumar, G. P. *et al.* (2000) Analysis of the effect of endogenous viral genes in the Smyth line chicken model for autoimmune vitiligo., *The American journal of pathology*, 156(3), p. 1099–1107.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, 30(9), p. 1312–3.

Steinbiss, S. *et al.* (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons., *Nucleic Acids Research*, 37(21), p. 7002–13.

Steinbiss, S., Kastens, S. and Kurtz, S. (2012) LTRsift: a graphical user interface for semi-automatic classification and postprocessing of de novo detected LTR retrotransposons, *Mobile DNA*, 3, p. 18.

Stoye, J. P. (2001) Endogenous retroviruses: still active after all these years?, *Current biology*, 11(22), p. R914-6.

Stoye, J. P. (2012) Studies of endogenous retroviruses reveal a continuing evolutionary saga., *Nature reviews. Microbiology*, 10(6), p. 395–406.

Subramaniam, S. (1998) The biology workbench - A seamless database and analysis environment for the biologist, *Proteins: Structure, Function and Genetics*, 32(1), p. 1–2.

Suh, A. *et al.* (2011) Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds, *Nature Communications*, 2(1), p. 443.

Suh, A. (2016) The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves, *Zoologica Scripta*, 45, p. 50–62.

Suh, A., Smeds, L. and Ellegren, H. (2015) The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds., *PLoS biology*, 13(8), p. e1002224.

Sun, C. *et al.* (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders, *Genome Biology and Evolution*, 4(2), p. 168–183.

Sun, Y. H. *et al.* (2017) Domestic chickens activate a piRNA defense against avian leukosis virus, *eLife*, 6, p. 1–24.

Swanstrom, R. *et al.* (1983) Transduction of a cellular oncogene: the genesis of Rous sarcoma virus., *PNAS*, 80(9), p. 2519–23.

Tarlinton, R. E., Meers, J. and Young, P. R. (2006) Retroviral invasion of the koala genome., *Nature*, 442(7098), p. 79–81.

Taylor, D. J. *et al.* (2011) Evolutionary maintenance of filovirus-like genes in bat genomes, *BMC Evolutionary Biology*, 11(1), p. 336.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration, *Briefings in Bioinformatics*, 14(2), p. 178–192.

Tixier-Boichard, M. and Boulliou-Robic, A. (1997) A deleted retroviral insertion at the ev21-K complex locus in Indonesian chickens, *Poultry Science*, 76, p. 733–42.

Tixier-Boichard, M. H. *et al.* (1994) Screening Chickens for Endogenous Virus ev21 Viral Element by the Polymerase Chain Reaction, *Poultry Science*, 73(10), p. 1612–6.

Trapnell, C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation., *Nature Biotechnology*, 28, p. 511–5.

Ulfah, M. *et al.* (2016) Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatera and Kedu Hitam, *BMC Genomics*, 17(1), p. 320.

Varela, M. *et al.* (2009) Friendly Viruses, *Annals of the New York Academy of Sciences*,1178,p.157–72.

Varriale, A. (2014) DNA Methylation, Epigenetics, and Evolution in Vertebrates: Facts and Challenges, *International Journal of Evolutionary Biology*, 2014:475981.

Venugopal, K. (1999) Avian leukosis virus subgroup J: a rapidly evolving group of oncogenic retroviruses., *Research in veterinary science*, 67(2), p. 113–9.

Vezzoli, G., Soldati, L. and Gambaro, G. (2009) Roles of calcium-sensing receptor (CaSR) in renal mineral ion transport, *Current pharmaceutical biotechnology*, 10(3), p. 302–310.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model, *Cytogenetic and Genome Research*, 110(1–4), p. 91–107.

Volff, J.-N. (2009) Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes., *Annals of the New York Academy of Sciences*, 1178, p. 233–43.

Volff, J. N. (2006) Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes, *BioEssays*, 28(9), p. 913–22.

Waltari, E. and Edwards, S. V (2002) Evolutionary dynamics of intron size, genome size, and physiological correlates in archosaurs., *The American Naturalist*, 160(5), p. 539–52.

Wang, K. C. and Chang, H. Y. (2011) Molecular Mechanisms of Long Noncoding RNAs, *Molecular Cell*, 43(6), p. 904–14.

Wang, Q., Jia, P. and Zhao, Z. (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data., *PloS one*, 8(5), p. e64465..

Wang, Q., Jia, P. and Zhao, Z. (2015) VERSE: a novel approach to detect virus integration in host genomes through reference genome customization, *Genome Medicine*, 7(1), p. 2.

Wang, Z. *et al.* (2013) An EAV-HP insertion in 5 Flanking region of SLC01B3 causes blue eggshell in the chicken., *PLoS genetics*, 9(1), p. e1003183.

Warr, A. *et al.* (2015) Identification of low-confidence regions in the pig reference genome (Sscrofa10.2), *Frontiers in Genetics*, 6, p. 1–8.

Warren, W. C. *et al.* (2010) The genome of a songbird., *Nature*, 464(7289), p. 757–62.

Warren, W. C. *et al.* (2017) A New Chicken Genome Assembly Provides Insight into Avian Genome Structure., *G3*, 7(1), p. 109-17.

Weber, M. and Schübeler, D. (2007) Genomic patterns of DNA methylation: targets and function of an epigenetic mark., *Current opinion in cell biology*, 19(3), p. 273–80.

Weiss, R. (2006) The discovery of endogenous retroviruses., *Retrovirology*, 3(67).

Weissensteiner, M. H. *et al.* (2017) Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications, *Genome Research*, p. 697–708.

Wheeler, T. *et al.* (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models, *Nucleic Acids Research*, 41(D1), p. D70-82.

Wimmers, K. *et al.* (1996) Molecular analysis of a new variant of the ev21 insertion/K-gene complex in the super slow feathering Nunukan chicken, *Journal of Animal Breeding and Genetics*, 113(4–5), p. 323-9.

Worley, K. C. (2017) A golden goat genome, *Nature Genetics*, 49(4), p. 485–6.

Wragg, D. *et al.* (2013) Endogenous Retrovirus EAV-HP Linked to Blue Egg Phenotype in Mapuche Fowl., *PloS one*, 8(8), p. e71393.

Wragg, D. *et al.* (2015) Genome-wide analysis reveals the extent of EAV-HP integration in domestic chicken, *BMC Genomics*, 16(1), p. 784.

Wright, A. E. *et al.* (2015) Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution, *Molecular Ecology*, 24(6), p. 1218–35.

Wright, N. A., Gregory, T. R. and Witt, C. C. (2014) Metabolic "engines" of flight drive genome size reduction in birds, *Proceedings of the Royal Society B: Biological Sciences*, 281(1779), p. 20132780.

Wu, X. *et al.* (2005) Weak Palindromic Consensus Sequences Are a Common Feature Found at the Integration Target Sites of Many Retroviruses, *Journal of Virology*, 79(8), p. 5211–4.

Xiao, Y. *et al.* (2015) Predicting the Functions of Long Noncoding RNAs Using RNA-Seq Based on Bayesian Network, *BioMed Research International*, 2015, p. 839590.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons., *Nucleic Acids Research*, 35, p. W265-8.

Youngson, N. A. *et al.* (2005) A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting, *Journal of Molecular Evolution*, 61(4), p. 481–90.

Yu, Y. *et al.* (2008) Quantitative evaluation of DNA methylation patterns for ALVE and TVB genes in a neoplastic disease susceptible and resistant chicken model., *PloS one*, 3(3), p. e1731.

Zhang, H., Bacon, L. D. and Fadly, A. M. (2008) Development of an endogenous virus-free line of chickens susceptible to all subgroups of avian leukosis virus., *Avian diseases*, 52(3), p. 412–8.

Zhang, L. *et al.* (2012) Genetic effect of the prolactin receptor gene on egg production traits in chickens, *Genetics and Molecular Research*, 11(4), p. 4307–15.

Zhang, L. *et al.* (2014) The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*, *Virulence*, 5(6), p. 655–64.

Zhang, Q. *et al.* (2015) Integrating transcriptome and genome re-sequencing data to identify key genes and mutations affecting chicken eggshell qualities, *PLoS ONE*, 10(5), p. 1–16.

Zhang, Q. *et al.* (2016) Genome Resequencing Identifies Unique Adaptations of Tibetan Chickens to Hypoxia and High-dose Ultraviolet Radiation in High-altitude Environments, *Genome Biology and Evolution*, 8(3), p. 765-76.

Zhang, W. *et al.* (2011) A practical comparison of De Novo genome assembly software tools for next-generation sequencing technologies, *PLoS ONE*, 6(3), p. e17915.

Zhao, J. *et al.* (2016) Identification of candidate genes for chicken early- and late-feathering Gene in Chicken Strand-specific Transcription of the Fusion Gene in Chicken, *Poultry Science*, 95(7), p. 1498-1503.

Zhou, H. *et al.* (2014) Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice, *Nucleic Acids Research*, 42(17), p. 10903–14.

Zimin, A. V *et al.* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm., *Genome research*, 27(5), p. 787–92.

Appendices

Appendix 1: Code repositories

All scripts for the two developed pipelines have been included on the CD accompanying this thesis, and are also hosted in individual GitHub repositories:

LocaTR	https://github.com/andrewstephenmason/LocaTR
ALVE_ID_pipeline	https://github.com/andrewstephenmason/ALVE_ID_pipeline

In addition, a third GitHub repository has been created which contains additional code which has been useful during this thesis. All scripts contain appropriate comments, and a general README file has been created to explain individual script functionality:

ASM_PhD_extras	https://github.com/andrewstephenmason/ASM_PhD_extras

Appendix 2: Additional files

This appendix consists of eleven additional files of results and reference sequences. Each file has been included on the CD accompanying this thesis, and static links have also been provided. This section briefly describes the files including their format.

1) Reference sequences used in LocaTR

A total of 717 reference sequences were used as part of the LocaTR expanded homology protocol. Sequences were chosen to give good phylogenetic coverage of known LTR retrotransposons, but there is a bias towards ERVs and Avian repeats. Sequences were downloaded from RepBase, Gypsy Database and NCBI. The file is a standard FASTA.

Filename:	AF01_LocaTR_reference_sequences.fa
Static link:	https://tinyurl.com/ydeux9ef

2) LTR retrotransposons identified in the Galgal4 assembly

Identified LTR retrotransposon positions have been given in BED6 format. This is a tab spaced file with columns: chromosome, start position (0 indexed), end position, name, score and strand. Line names were 'FL' meaning full list, or 'SIE' meaning

structurally intact element. All SIEs were contained within FL lines, but some SIEs are within larger FL regions. The score column was a placeholder, with values were set to 0.

Filename: AF02_Galgal4_LTR_retrotransposons.bed
Static link: https://tinyurl.com/ydgh7ep3

3) Structurally intact LTR retrotransposon clusters in the Galgal4 assembly

This file is the output of cluster_counter.py (ASM_PhD_extras repository), showing the analysis of contigs which contained five or more SIEs and whether these were in clusters. For contigs with clusters, the output shows number and proportion of SIEs within clusters, cluster sizes and locations, and the LTR homology of each SIE LTR pair. Homology values of 0.0 represent sequences where LTRs had significantly diverged in length (through insertions or deletions) so the homology scores were uninformative.

Filename: AF03_Galgal4_SIE_cluster_locations.txt
Static link: https://tinyurl.com/ya2k8kmd

4) LTR retrotransposons identified in the Galgal5 assembly

As AF02, but with LTR retrotransposons identified in the Galgal5 assembly.

Filename: AF04_Galgal5_LTR_retrotransposons.bed
Static link: https://tinyurl.com/yan9s5qq

5) Structurally intact LTR retrotransposon clusters in the Galgal5 assembly

As AF03, but with clusters identified from the LTR retrotransposons in the Galgal5 assembly. LTR pair homology was not included in this output.

Filename: AF05_Galgal5_SIE_cluster_locations.txt
Static link: https://tinyurl.com/yb7of5gh

6) Structurally intact LTR retrotransposon overlaps with Galgal5 lncRNA genes

This file is in modified BED format. Columns 2-7 are the SIEs which overlap with lncRNA genes in BED6 format, where the name is 'SIE' and the score is the element

length, and columns 8-13 are the same but for the lncRNA genes, with the name as 'RK_lnc' for Richard Kuo lncRNA gene annotation. The first column is the length of overlap between each feature on the line.

Filename: AF06_Galgal5_lncRNA_overlaps.txt
Static link: https://tinyurl.com/yb894o34

7) Alpharetroviral reference sequences used to mask the Galgal5 assembly

A total of 31 alpharetroviral sequences were used to identify alpharetroviral-homologous regions in the Galgal5 reference genome. The file is a standard FASTA with the header name containing the sequence name, GenBank accession and sequence length.

Filename: AF07_alpharetroviral_reference_sequences.fa
Static link: https://tinyurl.com/yafun5tf

8) ALVE reference sequences used to construct the viral pseudochromosome

A total of 11 ALVE sequences were used to construct the viral pseudochromosome. The file is a standard FASTA with the header name containing the sequence name, GenBank accession and sequence length. This file is a subset of AF07.

Filename: AF08_ALVE_reference_sequences.fa
Static link: https://tinyurl.com/y7cq5wev

9) The sequenced Hy-Line ALVEs

A standard FASTA file with the fifteen sequenced ALVEs from the Hy-Line elite layers including terminal hexamers. Sequence headers include the name and orientation.

Filename: AF09_HL_ALVE_sequences.fa
Static link: https://tinyurl.com/ybydbhes

10) The locations of all identified ALVEs in the Galgal5 assembly

Each identified ALVE is listed with its chromosome, insertion hexamer start position, hexamer sequence and gene feature overlap. Previous and new nomenclature is shown.

Filename: AF10_ALVE_locations.xlsx

Static link: https://tinyurl.com/yaunavym

11) Presence/Absence matrix for all identified ALVEs

All datasets are listed as columns matching the order outlined in section 7.4.1. The ALVE complement of each line is shown as 1 (presence) or 0 (absence) for all identified ALVEs. ALVEs are shown with their insertion site and new nomenclature.

Filename: AF11_ALVE_presence_absence_matrix.xlsx

Static link: https://tinyurl.com/ybsavjza

Appendix 3: Published papers

During this PhD project, I have been involved in the publication of three papers. PDF versions of each have been included on the CD accompanying this thesis, and static links have also been provided. Each paper was published in an open access journal.

1). Wragg D, Mason AS, Yu L, Kuo R, Lawal RA, Desta TT, Mwacharo JM, Cho CY, Kemp S, Burt DW & Hanotte O (2015), Genome-wide analysis reveals the extent of EAV-HP integration in domestic chicken, *BMC Genomics*, **16:** 784.

I analysed whole genome resequencing data from the eight Pirbright inbred lines for novel EAV-HP integrations. This analysis was carried out using a pipeline developed by David Wragg with additional scripting completed by me and Richard Kuo. I reviewed the manuscript before final submission, but was not involved in its writing.

Filename: Paper1_Wragg2015.pdf

Static link: https://tinyurl.com/y79ecswp

2). Mason AS, Fulton JE, Hocking PM & Burt DW (2016), A new look at the LTR retrotransposon content of the chicken genome, *BMC Genomics*, **17:** 688.

This project was initially conceived by my primary supervisor, Dave Burt, with participation from Paul Hocking, Janet Fulton and myself. I performed all the analysis

presented in the paper, wrote the manuscript and constructed all figures and tables. The manuscript was reviewed by all authors before submission. All scripts written for the LocaTR identification pipeline were my own work.

Filename: Paper2_Mason2016.pdf

Static link: https://tinyurl.com/ybm4z86q

3). Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, Schneider V, Mansour TA, Brown CT, Zimin A, Hawken R, Abrahamsen M, Pyrkosz AB, Morisson M, Fillon V, Vignal A, Chow W, Howe K, Fulton JE, Miller MM, Lovell P, Mello CV, Wirthlin M, Mason AS, Kuo R, Burt DW, Dodgson JB & Cheng HH (2017), A New Chicken Genome Assembly Provides Insight into Avian Genome Structure, *G3: Genes, Genomes, Genetics*, **7(1)**: 109-117.

I analysed the new chicken genome assembly with the LocaTR identification pipeline, and performed a preliminary analysis of the distribution of LTR retrotransposons in the assembly. Additionally, I performed a RepeatMasker annotation of the total assembly repeat content. I wrote my sections of the manuscript, including the construction of Table 2, and reviewed the manuscript before final submission.

Filename: Paper3_Warren2017.pdf

Static link: https://tinyurl.com/y7xyjjxh