The evolutionary dynamics of endogenous retroviruses

a thesis submitted for the degree of D.Phil. at the University of Oxford

Trinity 2015

Patrick Gemmell
St. Hilda's College, University of Oxford

Abstract

The evolutionary dynamics of endogenous retroviruses—Patrick Gemmell, St. Hilda's College, University of Oxford—D.Phil. thesis, Trinity 2015.

This thesis studies the evolution, influence, and proliferation of endogenous retroviruses (ERVs) within animal genomes. First, a simple mathematical model is constructed to address the question of whether retroviral endogenizations occur most often in male or female hosts. The result of applying the model to a diversity of genomes suggests that there may be female risk factors to endogenization, or that selection may be acting on full-length ERVs. Second, a study of the divergence of orthologous full-length ERVs from human and chimpanzee is performed. It is found that highly transcribed members of the HERV-H family have been under directional selection in the last six million years. Third, the insertion and deletion activity of the largest ERV families in five primate species is studied. Using a phylogenetic model it is demonstrated that ERVs are likely to be deleted early if they are to be deleted at all. Notably, it is also shown that HERV-H is an outlier family that is unusually slowly deleted. Fourth, the HERV-H loci in the human genome are studied on an individual basis. It is found that the long terminal repeats of HERV-H affect the magnitude and specificity of its transcription. Surprisingly, a region of the retroviral *gag* gene is positively associated with transcription and it is argued that this association is a partial explanation for the preferential maintenance of HERV-H in a full-length form. In conclusion, it is argued that researchers should take seriously the notion that many ERVs have not drifted to fixation. It is also argued that taking account of solo-LTR formation is important to accurately assessing the historical activity of ERVs. Finally, it is hypothesised that the application of bioinformatics techniques like those developed in this thesis may be sufficient to identify exaptation events in species quite distant from the primates that are studied here.

Acknowledgements

I gratefully acknowledge the advice of my supervisors Aris Katzourakis and Jotun Hein. I also thank the DTC staff for their timely support and my fellow group members for some interesting discussions. This research was supported by an EPSRC studentship and travel grants from St. Hilda's College.

Supplementary files

This thesis refers to data that is already electronically archived or that is best kept in machine readable format.

The supplementary material is as follows:

- Supplementary File 2.1–2.5: original supplementary material for Chapter 2 as archived by BioMed Central as part of accession doi:10.1186/1471-2148-13-243.
- Supplementary File 3.1–3.2: original supplementary material for Chapter 3 as archived by BioMed Central as part of accession doi:10.1186/s12977-015-0172-6.
- Supplementary File 4.1: site patterns, genomic locations, and viral library referred to in Chapter 4.
- Supplementary File 5.1: genomic characteristics underlying the 1,225 loci examined in Chapter 5.

Contents

Chapter 1

Introduction, thesis structure, and background

1.1 Introduction

This thesis is made up of four data chapters, each in the format of a paper. The work is presented in chronological order, and the first two chapters have now been published (Gemmell et al., 2013, 2015). Despite the chronological presentation each chapter is self-contained and so the research can be read in any order. In each project an effort has been made to apply bioinformatics techniques to real genomes and therefore my conclusions always have a degree of empirical support.

The main theme of each chapter is the behaviour of endogenous retroviruses within animal genomes. How do they arrive and how are they destroyed? How do they evolve after fixation? When and how quickly are they deleted? What features define the most highly transcribed members of a co-opted family?

In the remainder of this chapter I will first provide a summary of the research questions underlying chapters 2–5. I will then provide some general background that further contextualizes my research.

As indicated above, the themes of invasion, control and co-option are important. These themes are part of a broader research program into selfish DNA, the population genetics of transposable elements and viral co-option by hosts. Each theme will be introduced before the start of Chapter 2.

1.2 Thesis structure

Chapter 2

In Chapter 2 there are two main questions of interest. First, do retroviral endogenizations tend to occur more frequently in male than female hosts? Second, can we see a strong link between meiotic recombination rates and endogenous retroviral deletion?

These questions are addressed by developing a simple mathematical model relating meiotic recombination rates, insertion rates, and deletion rates. There will be one parameter in the model, male bias. The remaining interactions are specified using assumptions based on the biological differences between the autosomes and the allosomes.

The male bias parameter will be estimated using 18 animal genomes. Somewhat surprisingly we will find that there does not seem to be a universal male bias in the origin of endogenous retrovirus insertions. I argue this may be due to the placental affinity of endogenous retroviruses. We will also see an excess of full-length endogenous retroviruses on the X chromosome which I suggest may be due to failing to account for selection or the timescale of deletion.

Since one of the conclusions of Chapter 2 will be that a blanket assumption of neutrality might be inappropriate when explaining the distribution of endogenous retroviruses it is natural to address selection in Chapter 3. The dynamics of deletion are comprehensively studied in Chapter 4.

Chapter 3

Chapter 3, the second data chapter in this thesis, asks one question only: is there evidence that full-length endogenous retroviruses have been evolving unusually since the divergence of the human and the chimpanzee?

This question will be addressed by considering the relative nucleotide diver-

gence of full-length endogenous retroviruses when compared to neighbouring selfish DNA. We will see that some endogenous retroviruses, those from the HERV-H family, have been diverging faster than expected. I will argue that this result suggests that HERV-H has been under under directional selection. This argument will be supported by the fact that divergence is significantly correlated with the transcription of the endogenous retroviruses in question. By assuming that selfish DNA evolves neutrally I will obtain coefficients that quantify the magnitude of this selection.

During the course of my D.Phil. exciting findings about the HERV-H family of retroviruses have helped make sense of my own results. Chapter 5 in particular will investigate some of the features of this unusual family of primate endogenous retroviruses.

Chapter 4

The third research project in this thesis is described in Chapter 4. This chapter adresses the question of whether endogenous retroviruses "die young." This rather odd sounding notion refers to the hypothesis that the deletion of viruses is an age dependent process rather than a constant one, as has often been assumed in the literature, my own Chapter 2 included.

To study deletion dynamics a methodological framework is required. Chapter 4 characterizes the insertion and deletion activity of many families of endogenous retroviruses using a maximum likelihood phylogenetic approach. By comparing the fit of hazard functions to data derived from multiple genome alignments of six primate species I will show that deletion is something that happens early in the lifetime of an endogenous retrovirus.

Within Chapter 4 we will also see that the HERV-H family is again found to be unusual among primate endogenous retroviruses. This is because members of the HERV-H family tend to be strikingly less likely to be deleted at any given age when compared with the members of other endogenous retroviral families.

Chapter 5

The final data chapter in this thesis is Chapter 5. Having shown that HERV-H has unusual evolutionary dynamics in the sense that it is deleted slowly (Chapter 4) and that highly transcribed loci evolve quickly (Chapter 3), this final chapter asks what we can learn about highly transcribed HERV-H in the human genome. Specifically, I ask what is the genomic structure of a highly transcribed HERV-H locus?

In Chapter 5 I will characterize the majority of HERV-H loci in the human genome with reference to a consensus sequence. By considering the regions of the consensus present at each locus, as well as some other genomic features, I will make observations on the specificity and magnitude of HERV-H transcription. We will see that the youngest HERV-H are the most highly transcribed and least intact loci examined. However, we will also see that the presence of a more complete LTR is associated with transcription, and that a subset of older loci with a particular repeat structure are especially active in embryonic cells.

A particularly surprising result of Chapter 5 will be a positive correlation between part of the HERV-H *gag* gene and transcription. I will argue that this region enhances the ability of a host to effectively silence a HERV-H, and that this ability is a definite advantage with respect to HERV-H co-option by a host. This finding goes some way to explaining the results of Chapter 4 because it suggests that the internal region of HERV-H is of some use to the host and therefore perhaps one reason why HERV-H is so slowly deleted.

1.3 Background

The subject of this thesis is endogenous retroviruses (ERVs). Much of what has been written about transposable elements (TEs) applies to ERVs and it is therefore appropriate to introduce TEs in general and then ERVs in particular. We will see that TEs can be thought of as selfish DNA that expand the genome. Much has been discovered

about the control and interaction of TEs and so a portion of this research will be introduced below. Towards the end of this introduction we will also see that ERVs have a positive side, and that some ERVs now act as genes or regulatory elements for their hosts. The true extent of the upside to ERVs is at present unknown.

In general, a large proportion of many eukaryotic genomes is made up of sequences of DNA known as TEs. TEs were discovered by Barbara McClintock who was later awarded the Nobel Prize for her work. As TEs are sequences that can move from one location to another within a genome they are sometimes referred to as "jumping genes" or mobile DNA. McClintock originally hypothesised that TEs were "controlling elements" that were crucial to cellular differentiation, though her particular hypothesis is now known to be incorrect.

TEs are currently categorized as belonging to one of two groups: class I or class II (Brown, 2006). Class II TEs use a conservative "cut and paste" mechanism and their intermediate form when moving between two genomic loci is DNA. Class I TEs use a replicative "copy and paste" mechanism and are actually transcribed into RNA via the usual host enzymes. This RNA can be reverse transcribed by the reverse transcriptase enzyme (RT) and the DNA that results can be reintegrated into the genome by enzymes referred to as integrases (Thain et al., 2004). Clearly, if a TE can copy and paste itself then it can proliferate within a genome; however, cut and paste TEs can also increase in number if they excise from DNA that has been replicated and integrate into a site ahead of a replication fork or into another chromosome that has yet to be duplicated (Charlesworth et al., 1994).

As class I TEs make use of RT and are retro-transcribed they are often referred to as retrotransposons. However, not all retrotransposons actually encode their own RT enzyme. For example, in mammals the so called short interspersed nuclear elements (SINEs) rely on the presence of RT produced from other sources such as the long interspersed nuclear elements (LINEs) (Rowold and Herrera, 2000). LINEs are autonomous retrotransposons in the sense that they encode their own RT and are therefore not de-

pendent on the presence of another kind of transposon for replication (Gonzalez and Petrov, 2012). As we will see below, analogous ideas also apply to ERVs. Both LINEs and SINEs are polyadenylated during replication and are known as poly-A or non-LTR retrotranpsons (Thain et al., 2004).

While retrotransposons such as SINEs and LINEs are polyadenylated some retrotransposons use a distinct replication mechanism that that involves the production of two long terminal repeats (LTRs). LTRs are direct repeats with a length of roughly several hundred to one thousand base pair (bp) depending on the TE family they originate from (Bannert and Kurth, 2006). In contrast to poly-A retrotransposons, retrotransposons that use LTRs are known as LTR retrotransposons (Thain et al., 2004) and are more similar to retroviruses (below).

The ERVs that are the focus of this thesis either once were, or are at least related to, exogenous retroviruses. Retroviruses are pathogens that integrate their single-stranded RNA genome into host cellular DNA via reverse transcription. This is an obligate part of their lifecycle. If a retrovirus happens to integrate into a germ line cell, and if the germ line cell leads to a successful zygote, than an endogenization has occurred: the retrovirus has been vertically transmitted in a Mendelian fashion and is now referred to as an endogenous retrovirus or ERV. The term HERV refers to endogenous retroviruses that have been characterized in humans.

ERVs were discovered in the late 1960s by observing that virological markers could be transmitted according to Mendelian laws (Weiss, 2006). Until RT was discovered these observations were difficult to understand. Today it is possible to locate thousands of ERVs in animal genomes and the Mendelian transfer of retroviruses is not controversial. ERVs can be studied using methods from molecular biology or with bioinformatics tools, and the discipline of paleovirology studies endogenous viral material of many kinds (Katzourakis, 2013). This thesis describes projects where tools such as LTRHarvest (Ellinghaus et al., 2008) or RepeatMasker (Smit et al., 2004) were used to identify ERVs. These tools allow one to delineate ERVs via homology or struc-

tural features. Once identified ERVs can then be studied like other genomic entities by making using of the software repertoire available to the modern biologist.

With regard to classification, the retroviruses are currently split into seven genera. As identified by van Regenmortel et al. (2000) the genera are: the alpharetroviruses from birds; the betaretroviruses from mammals including mice, primates and sheep; the gammaretroviruses, that were originally documented due to their association with murine cancers; the epsilonretroviruses that contain few endogenous members; the spumaviruses, or foamy viruses, whose relationship to disease is not well understood; and the deltaretroviruses and lentiviruses, that are associated with smouldering infections such as HIV/AIDS. The exact relationship between retroviruses and LTR retrotransposons is unclear. LTRs appear to have a common origin (Benachenhou et al., 2013), and although retroviruses may ultimately be derived from LTR retrotransposons, it also appears that viruses have donated genes (Malik et al., 2000) to LTR retrotransposons, enabling them to become horizontally transmissible (Jordan et al., 1999) like retroviruses. Figure 1.1 displays the exogenous retroviral genera and relates them to the ERV families that are discussed in this chapter.

Though retroviruses have been identified in a wide range of vertebrates, assessing their true host diversity is difficult and most knowledge comes from humans or domesticated animals (Gifford and Tristem, 2003). Here ERVs can help improve scientific understanding of retroviral diversity. Partly this is because ERVs are plentiful and are found in many species, including amphibians, birds, carnivores, fish, lagomorphs, primates, rodents and ungulates, and can therefore fill in the phylogenetic gaps between known exogenous retroviruses (e.g. Jern and Coffin, 2008; Bolisetty et al., 2012; Zhuo et al., 2013). However, ERVs also contribute information on ancient viruses so that they can be used to study, for example, the long-term co-evolution of hosts and viruses, whereby ERV abundance is related to changes in host genes e.g. Daugherty and Malik (2012).

The ERVs in the human genome are particularly well described. Using phylo-

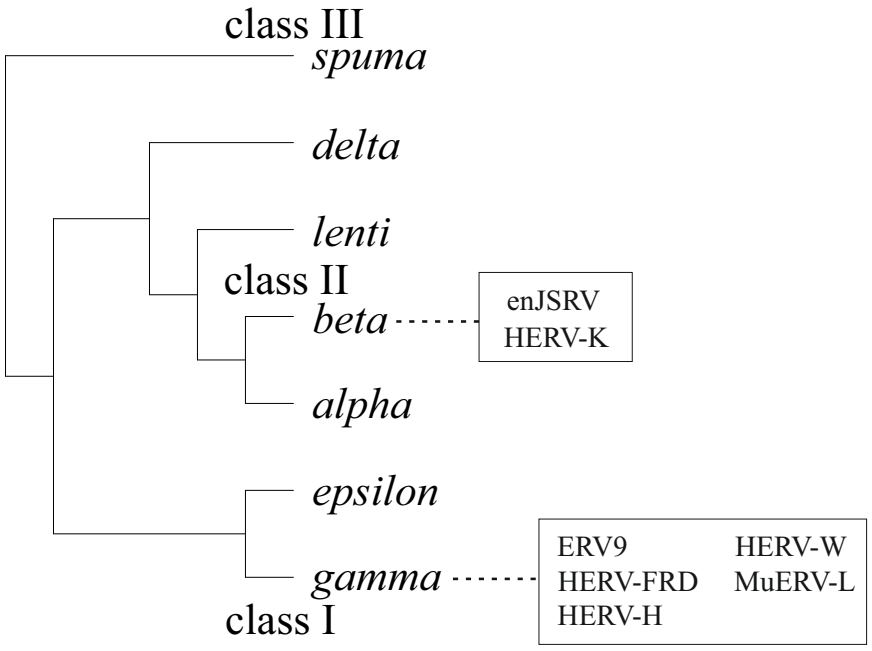

Figure 1.1: The relationship between retroviral genera as presented by Hayward et al. (2013). The ERV classes are shown and the ERV families mentioned in this chapter are related to their closest exogenous viral genus.

genetic methods, the HERVs have been divided into 31 families (Katzourakis and Tristem, 2005), some of which are very diverse (Tristem, 2000). These 31 families are traditionally grouped into three classes that relate them to exogenous retroviruses. Over two-thirds of the families are members of class I and are gammaretrovirus like, while the remaining families are equally split between the alpharetrovirus/betaretrovirus like class II, and the somewhat spumavirus like class III (Katzourakis and Tristem, 2005). Some HERVs are extremely old, perhaps over 60 Myr, though as the loci within each family appear to be derived from a relatively small number of bursts of activity it is still possible to relate them phylogenetically (Katzourakis and Tristem, 2005).

Although the HERVs have been grouped into families using phylogenetic techniques they are often named according to the tRNA thought to initiate their reverse transcription rather than by comparison to other ERV families. For example, HERV-K, a family recently active in humans, is primed by lysine tRNAs. The naming scheme

of HERVs is often argued to be problematic because it is redundant (unrelated HERVs share identical tRNA primers) and because it does not reflect the biological relationships between ERVs. However, so far the alternative efforts to classify HERVs (Blomberg et al., 2009) do not seem to have caught on.

Having defined ERVs I will now describe how they replicate. The origin and removal of replication competent ERVs is a research topic that I address in both Chapter 2 and Chapter 4. In Chapter 2 I am interested in opportunities for replication as well as the role of recombination in the destruction of ERV loci. In Chapter 4 I am interested in the timing of these phenomena. We will see that ERVs are actually replicated using a complex range of strategies, though I do not explicitly take these strategies into account in my research.

The ability of ERVs to self-replicate, by whatever means, is the crucial feature that allows them to be considered as selfish DNA. Selfish DNA is a powerful idea that we will come to in due course. However, while it is important to consider ERVs as selfish DNA in the first instance, in Chapter 3 we will see that not all the ERVs in the human genome are best viewed this way. This claim will be reinforced in Chapter 4 and Chapter 5, although its justification should also become obvious in light of the viral co-options that will be described below.

Exogenous retroviruses proliferate by gaining entry to the cells of host organisms and their viral lifecycle can be broken down into five steps (Sverdlov, 2000): first, a virus effects cell entry as receptors on its envelope interact with receptors on the host cell wall; second, a virus uncoats in order to be reverse transcribed from RNA into double stranded DNA; third, a virus enters the cell nucleus and integrates into the host genome as a provirus (Figure 1.2); fourth, the provirus is transcribed and processed, and viral products are transported to the cytosol; fifth, viral proteins are assembled to form many new viral particles which then bud from the host cell.

Cell entry is important to exogenous retroviruses but is not a prerequisite for the proliferation of ERVs (Belshaw et al., 2004). Although ERVs may replicate via cell exit

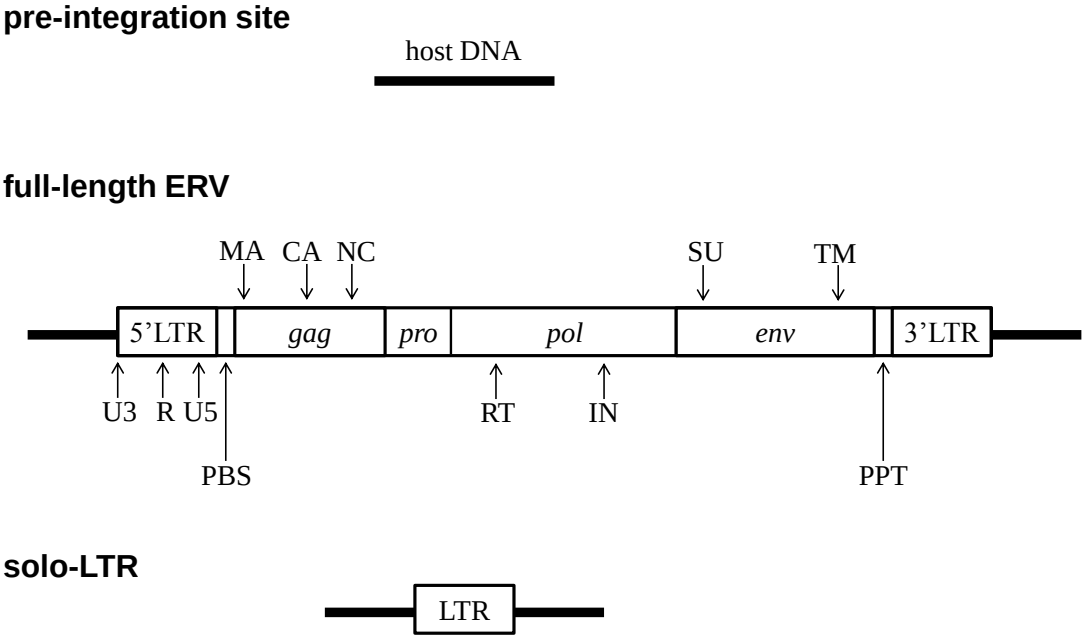

Figure 1.2: Pre-integration site: initially no virus is present at this locus in the host genome. Full-length ERV: the provirus is initially terminated by two identical LTRs that delineate: the PBS (primer binding site) where reverse transcription initiates; *gag* (group specific antigen) that contains structural proteins MA (matrix), CA (capsid) and NC (nucleocapsid); *pro* (protease) that cleaves polyproteins; *pol* (polymerase) that contains RT (reverse transcriptase) to convert RNA into DNA and IN (integrase) to incorporate viral DNA into the host genome; *env* (envelope) that contains SU (surface unit) that interacts with susceptible receptors and TM (transmembrane unit) that triggers the fusion of the virus with susceptible cells and that also contains an immunosuppressive domain; PPT (polypurine tract) that primes positive-strand DNA synthesis. Not shown in the diagram are the short 4–6 bp target site repeats that immediately flank the integrated provirus. Solo-LTR: most ERVs in primate genomes are found in a truncated form that is presumably the result of recombination between their 5′ and 3′ LTRs. Figure is after Bannert and Kurth (2006) and Ellinghaus et al. (2008).

and re-entry (reinfection) they can also replicate via retrotransposition, that involves processes that occur only within a single cell. The retrotransposition of ERVs implies that families of ERVs do not have to maintain a full complement of intact retroviral genes in order to proliferate. This is because the *env* gene is not required to duplicate within a single cell. By using the intracisternal A-type particle (IAP) as a model system

it has been shown that ERVs that lose their *env* gene and replicate via retrotransposition seem to become the most prolific families in the long term (Magiorkinis et al., 2012). In fact, the degradation of *env* is only one example of the way in which members of an ERV family can lose viral genes and yet continue to reproduce. Although it is possible for ERVs to replicate by complementation in *cis* (where the molecular products of the ERV's own sequence are used for its mobilization) it is also possible for an ERV to replicate by complementation in *trans* (whereby the ingredients required for mobilization are provided by other ERVs or exogenous viruses).

The differences between endogenous and exogenous viral replication are well emphasised by a study of the four largest HERV families in humans. There Belshaw et al. (2005b) found a tendency for ERVs to use mechanisms including complementation in *trans*, retrotransposition in *cis* and, in the case of HERV-W, the parasitization of LINEs. Belshaw et al. (2005b) further found that within the HERV-H family a large clade of elements had many inactivating deletions in its genes. While these ERVs could not have replicated themselves, they are nested within a clade of more intact elements exhibiting low dN/dS in their *env* genes. This evidence of purifying selection suggests that the clade of intact elements are re-infecting, as is typical of many of the smaller HERV families. These more intact ERVs are thought to have mobilized the less intact ERVs by complementation in *trans*. HERV-H is an important family with respect to this thesis and will be discussed further below. The mobilization patterns of HERV-H are in contrast to those of the relatively small HERV-K(HML2) family that is thought to be reproduced predominately via reinfection (Belshaw et al., 2004) as evidenced by the purifying selection that has been acting on its genes.

Once an endogenization event has occurred a family of ERVs may continue to proliferate until all members of its family are silenced by host defences (Johnson, 2007) or otherwise degraded via mutation or solo-LTR formation (Katzourakis et al., 2005). The term solo-LTR refers to a single LTR that is missing its associated paired LTR and proviral genes. Solo-LTRs are thought to be generated when LTRs undergo non-

allelic homologous recombination which results in a deletion and an acentric fragment (Stankiewicz and Lupski, 2002). Solo-LTRs are far more numerous than full-length ERVs in primate genomes (Bannert and Kurth, 2006) and most families of HERV exhibit a roughly 10:1 solo-LTR to full-length ERV ratio. It is not known whether the majority of solo-LTRs found in primate genomes are formed before or after endogenization although Belshaw et al. (2007) argued that ERVs from the HERV-K(HML2) family were less likely to form solo-LTRs with age. The issue of solo-LTR formation is directly addressed in Chapters 2 and 4 of this thesis.

Today the vast majority of ERVs in humans are thought to be inactive. Since the split between the chimpanzee and human only one ERV family, the aforementioned HERV-K(HML2), is known to have replicated. Some loci in this family are currently polymorphic, for example, loci HERV-K 113 and HERV-K 115. However, an analysis of American individuals suggests that the insertions occurred at least 0.8 Ma (Jha et al., 2009). If this integration date is correct then these loci were created some time before the origin of anatomically modern humans. Nevertheless, there has been and continues to be some speculation that the HERV-K family remains active today so that, for example, a recent study reported that a human individual contains an average of 6 HERV-K loci that are not present in the reference genome (Marchi et al., 2014). The possibility of active HERV loci in humans is exciting and further population data and improved genome assembly tools will ensure that the topic can be adequately addressed quite soon.

At present there is considerable interest in understanding the structure of eukaryotic genomes. An elementary aspect of genome structure is gene density, and it has been known for some time that the genomes of organisms having similar phenotypic complexity can vary massively in size. For example, the Norway spruce (*Picea abies*) has an enormous 20 Gbp genome (Nystedt et al., 2013) while the not so dissimilar bladderwort (*Utricularia gibba*) has a genome of only 82 Mb (Ibarra-Laclette et al., 2013). Both genomes contain the same number of genes, about 28,000.

The finding that the complexity of organisms does not scale with genome size is known as the C-value paradox (Eddy, 2012). I have described how TEs replicate but have not addressed the consequences of their proliferation. These consequences turn out to be profound because they explain the composition of the genome as well as highlight the importance of protecting it. As this thesis is motivated and contextualized by research on the mechanisms and effects of TEs it is necessary to describe this research in detail below.

One resolution to the C-value paradox is the observation that much of the difference in genome size between organisms is due to selfish DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Selfish DNA is sequence that is present largely due to its ability to self replicate within hosts rather than due to some ability to contribute to the replication of the host itself. The important point is that genomes can be swollen by sequence that generally provides no adaptive value. Certainly TEs in general, and ERVs in particular, are usually treated as selfish DNA, though there is no universal consensus quantifying the proportion of loci that provide a function (see Graur et al., 2015) to the host.

The original presentation of the selfish DNA concept included a number of hypotheses. These included the suggestion that selfish DNA would prefer not to transpose unrelated sequence and the suggestion that selfish DNA would evolve towards reliance on a limited number of genes. The original presentation also made an explicit call for a new kind of population genetics to describe the behaviour of selfish DNA. This call was partially answered by Hickey (1982) who showed how deleterious selfish DNA might spread within a population of organisms.

The model of Hickey (1982) was limited and considered only a single TE locus. However, it did emphasise the important point that even harmful TEs could spread within host populations unless they were somehow controlled. The idea that deleterious elements can spread is especially relevant to the study of ERVs as we do not have a complete picture of the costs or benefits associated with the acquisition of many loci

nor even a good idea about the frequency with which endogenizations might have occurred. It is possible that viral endogenizations are fairly frequent events so that almost every ERV we see has fixed via drift. But this is not something we know for sure.

More comprehensive population genetic models of TEs have also been developed. These include the models of Langley et al. (1983) and of Charlesworth and Charlesworth (1983). The latter work has been particularly influential, the main result being an analytical expression of the change in mean TE copy number per individual in terms of the total number of sites in a genome that can be occupied, the mean transposition and excision rates of TEs, and the mean fitness of the host as a function of element number. Importantly, this result was used to demonstrate that even in the absence of self restraint by TEs, a rapidly decreasing fitness function with respect to TE copy number could produce an equilibrium situation such that TEs would not fill up the genome. The distribution of occupation frequencies was also derived and has been empirically tested in, for example, fly (Charlesworth and Langley, 1989).

Over the past 25 years interest in the dynamics of TEs has been sustained. During this time a focus on the role of selective forces in controlling the proliferation of TEs has persisted. I will not discuss all the harmful effects of TEs in detail here, because I have done in Chapter 3. However, I will remark that ectopic recombination has frequently been a theme of theoretical work. As TEs are present in multiple copies it is quite possible for one TE to recombine with another similar TE from the same family during cell division. Charlesworth and Charlesworth (1983) highlighted the importance of the rate of reduction of host fitness as a function of TE number. Unlike many other potential fitness effects, the effect of ectopic recombination can a priori accumulate quickly enough to create an equilibrium situation. This is because ectopic recombination involves interaction between loci from the same TE family and is therefore expected to increase with the square of the number of elements in a genome.

Theoretical research into TEs has progressed since the 1980s. For example, the

effects of finite population sizes with respect to the control of transposable elements was studied by Brookfield and Badge (1997). The authors found that finite population sizes were associated with a decrease in the ability of a host to control transposable elements. This was due to a decreased variance in element copy number between individuals as well as particular loci becoming occupied at high frequencies. A highlight of this research was a model of the fitness effects of TEs in terms of the probability of aneuploid gamete formation due to recombination between heterozygous repeat loci. This idea was captured in the form of a recurrence.

More recently Le Rouzic and Capy (2005) have addressed the invasion of a new TE family into a diploid outbreeding species. Considering amplification, host-level selection, self-regulation and drift, the authors used simulations to argue that TEs will have an optimum transposition rate above which an invasion will do a great deal of damage to the host. The authors therefore proposed that successful genome invasions involve an initial burst of activity followed by a lower rate of activity or else a low rate of replication followed by fixation via drift.

Both finite population sizes and bursting seem particularly applicable when studying the fixation of ERVs. This is because exogenous retroviral infections would probably be both time limited and structured at the level of the host population. This suggests that effective transposition rates would be initially high, due to horizontal transmission, but would drop quite rapidly in the face of the response of the host population to an epidemic. Accordingly, bursts of retroviral replication in sub-populations could help drive a small number of ERVs to reasonable frequencies at which they would be much less likely to be lost from the overall population due to drift. These ideas are not a focus of this thesis but could be tested by studying the retroviral endogenizations that are currently occurring in koalas (Tarlinton et al., 2006).

A particular sub-thread of research into TEs has centred on the idea of the ecology of the genome (Brookfield, 2005; Venner et al., 2009). The main motivation for this analogy is that many of the features one wishes to understand about the TE com-

position of genomes have ecological parallels: What niches do TEs occupy? What determines TE diversity? What interactions occur between TE families? Though these questions may not be easy to answer a positive aspect of "genomic ecology" is that a history of interactions is preserved so that one does not need to focus only on the behaviour of currently active TEs.

There have been numerous responses to the ecological perspective. A straightforward application of the ecological metaphor was the construction of models to represent parasitism and obligate symbiosis between TE families (Le Rouzic et al., 2007). In the case of parasitism, and by analogy with LINEs and SINEs, one TE family produced resources that the other family needed to transpose. In the second case each TE family contributed a resource that the other family required to replicate. A related investigation modelled the dynamics of mutant TEs to show how replication incompetent TEs could parasitize replication competent ones in a way that could lead to cyclical population dynamics (Le Rouzic and Capy, 2006). Some ecological work has also focused on the relationship between host defences and TE abundance. Abrusán and Krambeck (2006) contributed a fairly technical model to predict TE diversity over time based on density dependent selection as effected by RNAi host defences. A metric characterizing the domination of TE communities by a small number of families (evenness) was related to changes in the number of families present.

It is not necessarily clear what distinguishes an ecological approach from an evolutionary one. Some clarification was provided by Linquist et al. (2013) who proposed an operational distinction: under an evolutionary approach one considers changes within classes of TEs over time whereas under an ecological approach one considers TEs as unchanging entities. This definition led the authors to suggest that factors such as GC content and genome size are ecological ones and most effective at describing the differences between observations taken from closely related host species; evolutionary approaches are better used when comparing observations from more distant hosts. To demonstrate their ideas the authors used the statistical method of redundancy anal-

ysis to show the relative roles of both kinds of factors on a dataset of 10 fly species (closer/ecological) and 13 mammalian species (more distant/evolutionary).

The ecology of the genome has played a conceptual role in recent work on selfish DNA. Though the main chapters of this thesis do not specifically reference ecological ideas it is definitely the case that in chapters 2–5 the families of ERV are treated as fixed categories. Therefore in some sense Chapter 2 discusses whether the X chromosome is a niche for ERVs, while Chapter 4, where I investigate the mortality of the larger HERV families, in some sense discusses the life history of ERVs. However, it is also true that in these chapters the hosts themselves are hypothesized to be evolving under the influence of a subset of the TEs. Moreover, at no point do I take account of any potential interactions between members of different ERV families. For this reason I would argue that there is no strong link between my work and the ecological metaphor.

Up until now I have emphasised the similarities between all TEs, especially with respect to their nature as selfish DNA that contribute little to the host and that need to be controlled. In contrast to this emphasis there have been relatively recent discoveries that demonstrate some surprising ways in which ERVs can confer benefits upon a host. The benefits of these viral co-option events include: providing immunity to mammals via receptor interference and the sabotage of exogenous viral particles (Varela et al., 2009); the contribution of placentally expressed genes to many mammalian species (Lavialle et al., 2013); and in all likelihood, a yet to be fully elucidated contribution to stem cell functionality (Robbez-Masson and Rowe, 2015). I will consider the latter two examples further here. The finding that ERVs can have substantial effects upon host phenotypes was part of the personal motivation behind the research undertaken in Chapter 3, where I look at the divergence of orthologous ERVs in humans and chimpanzees. It is also relevant to the results reported in chapters 3–5, in which knowledge of the co-option of HERV-H is essential background information.

Of the contributions made by ERVs to host biology their role in placental bi-

ology is probably the most surprising: exogenous retroviruses use the fusion peptide and immuno suppressive domain of the *env* gene to effect virus-cell fusion but these features have also been domesticated by their hosts (Lavialle et al., 2013). This relationship was first postulated after *env* genes with open reading frames (ORFs) in the human genome were found to exhibit low dN/dS and to be maintained across the ape lineage. These observation led to the identification of a 30 Myr old HERV-W *env*, now named *syncytin-1* (Voisset et al., 1999; Blond et al., 1999; Mi et al., 2000; Blond et al., 2000; Mallet et al., 2004) and a 45 Myr old HERV-FRD *env*, now named *syncytin-2* (Bénit et al., 2001; Blaise et al., 2003; Esnault et al., 2008). As the two syncytins have been shown to be specifically expressed in the placenta and to have fusogenic properties they are currently believed to play a role in fusing the outer layer of the trophoblast (placenta) to the wall of the uterus.

That *env* has been co-opted on two occasions is surprising. However, the evidence in favour of *env* co-option is broader than *syncytin-1* and *syncytin-2* and rather sublime. Since the work in human was undertaken two further *env* genes, *syncytin-A* and *syncytin-B*, have been identified in the mouse (Dupressoir et al., 2005). Gene knockout tests subsequently demonstrated that *syncytin-A* is necessary on the maternal side (Dupressoir et al., 2009) and that *syncytin-B* is important on the fetal side if runty pups are to be avoided (Dupressoir et al., 2011).

Further syncytins have been discovered in Lagomorpha (Heidmann et al., 2009), Caviomorpha (Vernochet et al., 2011), Carnivora (Cornelis et al., 2012), Ruminantia (Cornelis et al., 2013), Marmota (Redelsperger et al., 2014) and marsupials (Cornelis et al., 2015). The approximate age of these syncytins is 12–30 Myr (Lagomorpha), >30 Myr (Caviomorpha), 65–80 Myr (Carnivora), >30 Myr (Ruminantia), >45 Myr (Marmota), and >80 Myr (marsupials). Though knockout experiments have not been performed beyond the mouse, the weight of evidence (including fusion assays, the duration over which ORFs are maintained, and expression analyses) is fairly conclusive. Indeed, considering the many occasions on which syncytins have been independently

captured one might suggest that, given enough time, the placental co-option of ERVs is an almost inevitable consequence of the interaction between retroviruses and mammals.

The discovery of the syncytins, combined with the observation that ERVs and retroviral promotors are highly transcribed or exclusively expressed in the placenta, has led Chuong (2013) to speculate that ERVs and the placenta are in symbiosis. The argument is that ERVs are tolerated in the placenta because they were involved in the creation of the trophoblast cell lineage and that, since then, ERVs have often acted as a source of promotors that can rewire gene expression. This is an interesting but controversial hypotheses because from the point of view of a virus, placental expression could allow ERVs to be transmitted more effectively from mother to offspring (Haig, 2012, 2013).

If Chuong (2013) is correct one might expect sexually antagonistic selection to act on ERVs that are expressed in the placenta. This is because assuming that first, a large number of proviruses and retroviral LTRs have played an important role in the evolution of the placenta, and that second, some of the evolution of placental phenotypes is driven by positive selection due to parent offspring conflict (Trivers, 1974), then not only is it in the interest of fetal genes to transfer nutrients to a level in excess of the mother's preferences, but it can also be in the father's interest to transfer maternal resources to his offspring at a cost to other (potentially unrelated) ones (Haig, 1993). This implies that insofar as the average effect of a provirus is to increase the transfer of nutrients beyond the optimum from the mother's perspective, and insofar as the average substitution into a provirus mitigates this effect, substitutions into proviruses will benefit the female but harm the male.

Indeed, antagonism might be seen even if proviruses did not increase the transfer of nutrients beyond a level compatible with the preferences of both the mother and the offspring under ideal circumstances. This is because while both the male and the female fetus would (presumably) reap the benefits of parental investment via the

placenta equally, it is only the female that will come into contact with the potentially harmful effects of the placental activity of ERVs as an adult. Therefore, if the cost of adult contact with the placenta increases to a level that balances the benefit provided to an offspring then substitutions that oppose any further transfer will be of benefit to the female and not to the male. A desire to test these hypothesis was the motivation behind comparing the substitution rates of autosomal and X-linked ERVs in Chapter 3, though ultimately no substantial supporting evidence was found.

If the role of ERVs in placentation was a highlight of ERV research in the last decade then the role of ERVs in pluripotency may turn out to be the highlight of ERV research in this one. This is because in the last two years the ERV family HERV-H has been found to be essential to the maintenance of stem cell identity in humans (Lu et al., 2014; Wang et al., 2014).

HERV-H is a large family that we shall study in detail in Chapter 5. There are over 1,200 full-length loci integrated into the human genome (Bannert and Kurth, 2006). A succession of studies have identified various characteristics of the family including several LTR subtypes, the phylogeny of many loci, and a consensus sequence for the largest clade (Mager, 1989; Goodchild et al., 1993; Jern et al., 2004, 2005). The majority of integrations of HERV-H occurred perhaps 25–30 Ma, though the first integration occurred prior to the split of the New World Monkey and Old World Monkey lineages (Mager and Freeman, 1995). Unusually for an ERV family, HERV-H appears to have been maintained at a roughly 1:1 full-length to solo-LTR ratio in the human genome (Bannert and Kurth, 2006) though, as previously mentioned, the HERV-H family is dominated by elements containing large deletions in their genes, particularly in *env*.

In the last two years interest in the HERV-H family has redoubled as its relationship with stem cells has become known. In 2014 several important results were published. First, Wang et al. (2014) showed that 550 of the 1,225 full-length copies of HERV-H in the human genome are actively transcribed in human pluripotent stem

cells. These transcripts are mostly chimeric or long non-coding RNAs. Interfering with the LBP9 driver or the associated HERV-H transcripts destroyed stem cell identity. Second, Lu et al. (2014) showed that HERV-H transcription is necessary for both the creation and maintenance of stem cell identity and that HERV-H knockdown down-regulates pluripotency markers. Finally, Fort et al. (2014) suggested (via a large scale analysis of mouse and human transcriptomes) that stem cell specific transcription factors directly control transcripts originating from LTRs. As the HERV-H family has turned out to be relevant to my research, these results will be revisited in Chapter 3, Chapter 4, and particularly Chapter 5, where they are central.

The aforementioned host roles for HERV-H and the syncytins are not in conflict with the concept of selfish DNA, yet they are different in emphasis. They have led to retroviruses being described as regulatory elements that evolve especially quickly due to the adaptive arms race between host and virus (Schlesinger and Goff, 2015) and as lineage specific mobile promotors that rapidly rewire entire gene regulatory networks (Chuong, 2013). This demonstrates that the cutting-edge perception of ERVs by some experimentalists is distinctly at odds with a view that retroviruses behave solely as pathogens or selfish elements. In some sense this contemporary conception of the role of ERVs remains closely aligned to the spirit, if not the mechanism, of McClintock's original theory of controlling elements. Clearly many details of viral co-option are unknown and elucidating the dynamics of co-option remains an open topic for biologists. Bearing these thoughts in mind, it is appropriate to continue on to the research itself.

Chapter 2

Sex-specific aspects of endogenous retroviral insertion and deletion

2.1 Abstract

Background

We wish to understand how sex and recombination affect endogenous retroviral insertion and deletion. While theory suggests that the risk of ectopic recombination will limit the accumulation of repetitive DNA in areas of high meiotic recombination the experimental evidence so far has been inconsistent. Under the assumption of neutrality, we examine the genomes of eighteen species of animal in order to compute the ratio of solo-LTRs that derive from insertions occurring down the male germ line as opposed to the female one (male bias). We also extend the simple idea of comparing autosome to allosome in order to predict the ratio of full-length proviruses we would expect to see under conditions of recombination linked deletion or otherwise.

Results

Using our model, we predict the ratio of allosomal to autosomal full-length proviruses to lie between $\frac{3}{2}$ and $\frac{2}{3}$ under increasing male bias in mammals and between 1 and 2 under increasing male bias in birds. In contrast to our expectations, we find that a pattern of male bias is not universal across species and that there is a frequent over-abundance of full-length proviruses on the allosome beyond the ratios predicted by our model.

Conclusions

We use our data as a whole to argue that full-length proviruses should be treated as deleterious mutations or as effectively neutral mutations whose persistence in a full-length state is linked to the rate of meiotic recombination and whose origin is not universally male biased. These conclusions suggest that retroviral insertions on the allosome may be more prolific and that it might be possible to identify mechanisms of replication that are enhanced in the female sex.

2.2 Background

As an obligate part of their life-cycle, retroviruses integrate genetic information into their host's cellular DNA. If such an integration occurs in a germ line cell and is not sufficiently harmful to its host then it is possible for viral DNA to pass vertically from parent to progeny. Over time, endogenized viral DNA may become fixed within populations and it is therefore possible to detect the traces of ancient viral infections, often as fragments, by trawling modern genomes e.g. (Katzourakis et al., 2007b; Zhuo et al., 2013).

Most endogenous retroviruses (ERVs) are not observed in their original full-length proviral form. Immediately after successful integration, a provirus will consist

of a pair of long terminal repeats (LTRs) that flank the open reading frames for several retroviral genes, typically *gag*, *pol* and *env*. As flanking LTRs are identical at the time of insertion (Bannert and Kurth, 2006), a persistent similarity between viral extremities over generations means that there is a strong possibility of illegitimate recombination between LTRs from the same or similar ERVs. Recombinational deletion is said to occur when the internal region of a provirus is eliminated by recombination between LTRs and only a solitary or solo-LTR is left behind (Belshaw et al., 2007). As ERVs may replicate within the genome via reinfection or retrotransposition (Belshaw et al., 2004, 2005b; Katzourakis et al., 2005), recombinational deletion is one of the forces shaping both the retention and proliferation of selfish genetic elements (Orgel and Crick, 1980) and is therefore of interest to those concerned with the accumulation of repetitive DNA over time.

It has previously been shown that recombinational deletion in human is correlated with local meiotic recombination rate but that the fixation of ERVs is not (Katzourakis et al., 2007a). These findings are consistent with work examining transposons in worms (Duret et al., 2000) and retrotransposon specific evidence from flies (Rizzon et al., 2002), but are also in contrast to theory and experimental evidence that suggests that transposable elements in general are more frequent in chromosomal regions with lower rates of recombination (Charlesworth and Langley, 1989; Charlesworth et al., 1994).

In the aforementioned work, Katzourakis et al. (2007a) proposed that the majority of retroviral insertions are acquired down the male germ line due to the relatively high number of cell divisions involved in the production of sperm in the male as opposed to eggs in the female. As many exogenous viruses require cell division in order to cross the nuclear membrane (Roe et al., 1993), or are at least more efficient at infecting dividing cells (Suzuki and Craigie, 2007), it is reasonable to hypothesize that a deeper germ line will offer more opportunities for retroviral infection than a shallow one. This male bias hypothesis was supported by data showing an excess of ERVs

on the Y chromosome, even after the chromosome's low gene density was taken into account. The reasoning behind such a hypothesis is similar to original arguments for male mutational bias (Miyata et al., 1987) in which cell division is associated with error prone DNA replication: in both cases cell division is thought to be correlated with changes in germ line DNA. Although estimates of male mutation bias vary considerably (Li et al., 2002), it is generally thought that male bias correlates to life history traits, with longer lived animals tending to exhibit a higher male bias than shorter lived ones (Sayres et al., 2011).

The work of Katzourakis et al. (2007a) is robust but limited in two ways. First, recent evidence suggests that the rate of recombination along the length of chromosomes can vary rapidly (Myers et al., 2005) and therefore we are not sure how closely recent recombination rates correlate with those in the distant past. Second, we are interested in species beyond humans, including those for which we do not have a recombination map or even an assembly of the Y chromosome. To address these challenges we develop a straightforward model relating recombinational deletion, sex specific ERV integration rates and meiotic recombination at a chromosome level and then use it to examine whether genomic data from several species supports the conclusions of previous work.

2.2.1 Model

We want to consider how a sex specific ERV integration rate interacts with a recombinational deletion process that is either independent of or strongly linked to the background rate of meiotic recombination. To do so we will consider retroviral insertions under the XY and ZW sex-determination systems. We do this with the intention of comparing the density of ERVs on the allosome (X and Z chromosomes only) to those on the autosome using publicly available mammalian and bird genomes.

Assume that retroviral integrations into host genes are highly deleterious or lethal and that the insertions we see today are effectively neutral and fixed by drift.

We will consider p_i, the proportion of full-length proviruses per unit length of chromosome i. We write

$$p_i = n_i/(l_i - g_i)$$

where n_i is the number of full-length proviruses on chromosome i, l_i is the length in base pairs (bp) of chromosome i, and g_i is the number of bases on chromosome i that are part of a gene. We will use subscript a to refer to autosomal DNA and subscript x or z to refer to allosomal DNA.

Let f be the rate of proviral integration for females and let $m = \beta f$ be the rate of proviral integration for males, where β is a positive real number used to model male bias i.e. the ratio of viral integrations occurring down the male germ line to those viral integrations occurring down the female germ line. As the X chromosome spends twice the amount of time in the female as the male the average rate of proviral integrations on the X chromosome will be

$$\frac{1}{3}\beta f + \frac{2}{3}f$$

while the average rate of proviral integrations on an autosomes will be

$$\frac{1}{2}\beta f + \frac{1}{2}f.$$

We are interested in knowing whether recombinational deletion is linked to meiotic recombination and whether male bias has been an important factor in integrations. Let r_x and r_a represent the intensity of the process that deletes full-length proviruses from the X chromosome and the autosomes respectively. The rate of accumulation of full-length proviruses on the X chromosome is given by

$$\dot{n}_x = \frac{1}{3}\beta f + \frac{2}{3}f - r_x n_x \tag{2.1}$$

while the rate of accumulation of full-length proviruses on an autosome will be

$$\dot{n}_a = \frac{1}{2}\beta f + \frac{1}{2}f - r_a n_a. \tag{2.2}$$

We will use the ratio $p = \frac{p_x}{p_a}$ to make predictions under various scenarios. Equations 2.1 and 2.2 have a straightforward analytical solution (Supplementary File 2.1) so that in general we have

$$\lim_{t\to\infty} \frac{n_x}{n_a} = \frac{r_a}{r_x}\frac{2}{3}\frac{(\beta+2)}{(\beta+1)}.$$

As the X chromosome (excluding pseudoautosomal regions) recombines at only $\frac{2}{3}$ the rate of the autosomes we have $r_x = \frac{2}{3}r_a$ in the case that the deletion process is linked to meiotic recombination and $r_x = r_a$ otherwise. Therefore we arrive at Equation 2.3 and Equation 2.4 which give asymptotic values of p as a function of β in the presence or absence of recombination linked deletion respectively.

$$p = \frac{(\beta+2)}{(\beta+1)} \tag{2.3}$$

$$p = \frac{2}{3}\frac{(\beta+2)}{(\beta+1)} \tag{2.4}$$

These functions are similar to those originated by (Miyata et al., 1987) in the context of molecular evolution and are plotted in Figure 2.1. In the case that recombination is linked to deletion we expect to see more ERVs on the X chromosome due to its reduced rate of meiotic recombination. As the X chromosome also spends less time in males, the expected excess will be reduced in line with male bias so that as β increases the value of p will tend to unity. In the case that recombination is not linked to deletion then ERVs on the X chromosome are no more or less effectively deleted by the host. Now we do not expect any excess of ERVs unless male bias is strong, in which case the autosomes will receive more insertions than the X chromosome and p will tend to $\frac{2}{3}$ as β increases.

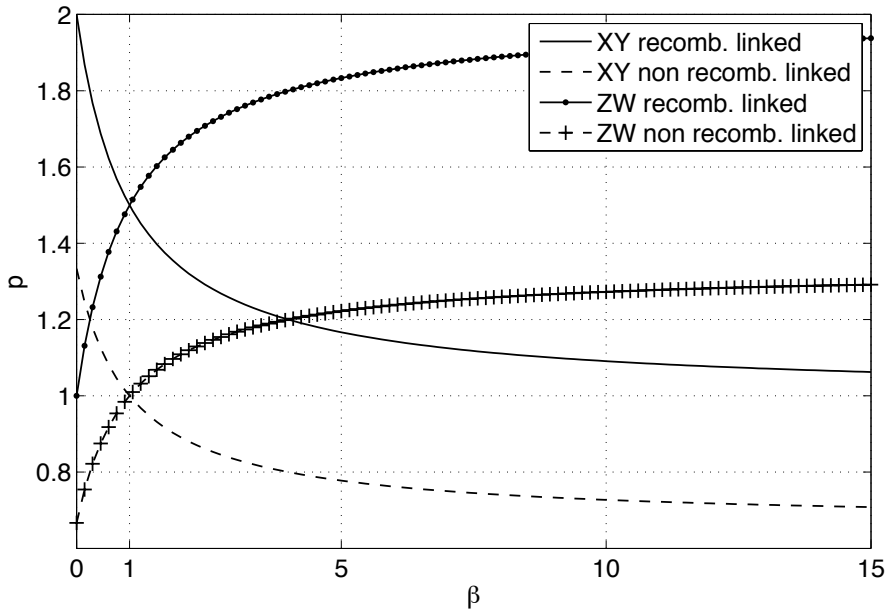

Figure 2.1: Predicted ratios. Predicted ratios of full-length proviruses on the allosome (X or Z) to the autosome under recombination linked and non recombination linked deletion scenarios. Predictions are shown for both the XY sex-determination system and the ZW sex-determination system. For any given bias (β) and sex-determination system we make two predictions as to the allosome-to-autosome ratio of full-length proviruses (p). A value of β greater than one is a male bias and a value of β less than one is a female bias. Under the ZW system (e.g. in birds) both male bias and a lack of recombination may contribute to an excess of ERVs on the Z chromosome when compared to the autosome.

In birds, males are the homogametic sex, each having two Z chromosomes, and therefore our model makes different predictions. Under the ZW sex-determination system

$$\dot{n}_z = \frac{2}{3}\beta f + \frac{1}{3}f - r_z n_z$$

so that by using a similar argument as before p_z/p_a is given by

$$p = \frac{(2\beta + 1)}{(\beta + 1)} \tag{2.5}$$

in the case of recombination linked deletion and

$$p = \frac{2}{3}\frac{(2\beta + 1)}{(\beta + 1)} \tag{2.6}$$

otherwise. When calculating $p = p_z/p_a$ male bias and reduced meiotic recombination both act in the same direction to increase the expected excess of ERVs on chromosome Z. As shown in Figure 2.1, in this case we expect p to be $\frac{3}{2}$ tending to 2 as β increases when recombination is linked to deletion and p to be 1 tending to $\frac{4}{3}$ otherwise.

2.3 Methods

In order to compare our model with reality we obtained an estimate of p_i for the chromosomes of eighteen species. This was done by counting full-length proviruses on each chromosome of the genome of each species and then using gene annotation information to calculate l_i and g_i.

Eighteen soft-masked animal genomes were obtained from the Ensembl project (Flicek et al., 2012): cat, chicken, chimp, cow, dog, gorilla, horse, human, macaque, marmoset, mouse, opossum, orangutan, pig, rabbit, rat, turkey and zebra finch. A collection of 771 viral *pol* sequences was used to locate as many potential endogenized

pol sequences as possible from across all eighteen genomes. The 771 probe sequences were selected to represent the full diversity of exogenous and endogenous retroviral genes and are the same as those used in previous studies (Katzourakis et al., 2007b; Magiorkinis et al., 2012). Application of tBLASTn (Altschul et al., 1990) identified putative *pol* hits which were used to extract 49,928 non-overlapping 15kb regions each centred on a match. These 49,928 regions were processed using LTRharvest (Ellinghaus et al., 2008), a tool designed to detect full-length LTR retrotransposons based on structural features alone. Thus, a large set of BLAST results were reduced to 18,203 putative full-length sequences of which we filtered the 16,661 that occur in sequence that is assembled into chromosomes of interest.

For some genomes LTRharvest was inclined to report sequences made up of a large amount of unknown nucleotide sequence, that is sequence recorded with `N`s in the genome, as retrotransposon like. These `N`s between LTRs are doubly problematic as they lead us to question whether LTRs are genuinely physically associated and also make it harder to confirm that the internal regions contain viral genes. To be more certain that we were dealing with genuine full-length proviruses we discarded any sequence containing more than five consecutive unknown nucleotides or comprising more than five percent unknown nucleotides overall. These particular cutoff values are conservative and were chosen with caution in mind. We then used the LTRdigest annotation tool (Steinbiss et al., 2009) to further discard any full-length proviral sequences that did not contain at least one retro-viral gene beyond the *pol* previously identified by homology. This filtering process left 7,299 full-length sequences for analysis as is recorded explicitly in Supplementary File 2.2.

From Ensembl genes69 we estimated g_i for each chromosome of interest using the BioMart section of the website. As gene annotations can overlap we post-processed the downloaded results to ensure that each base pair of annotation contributed at most once by using an algorithm that incrementally merged overlapping annotations. The total length of each chromosome l_i was available both from Ensembl and also directly

from the genomes themselves. Each putative provirus occurring on known chromosomal DNA contributed to the total count n_i for the chromosome.

As we are interested in any overall bias in retroviral insertions we also performed a survey of solo-LTRs across all eighteen genomes. In this case we compiled a query library containing the 5′ LTR region of each of the 7,299 full-length proviruses and performed a BLASTn search against every genome. Alignments of at least 95% identity and covering at least 95% of the query sequence were retained and multiple overlapping alignments were merged to give 926,894 putative LTRs. As the purpose of this search was to identify solo-LTRs, any putative LTRs separated by less than 15kb of intermediate sequence were discarded leaving 508,811 merged alignments that we consider represent the solitary remnants of proviruses.

We use solo-LTRs as a proxy for total insertions which is justifiable given the fact that they are so much more numerous than full-length proviruses, as is recorded in Supplementary File 2.3. However, as we detect solo-LTRs based solely on their similarity to LTRs of full-length proviruses we will not identify solo-LTRs that have no full-length proviral representatives.

For both full-length proviruses and solo-LTRs, we also checked that the ratio of allosomal (X or Z) to autosomal ERVs is not correlated to GC content (Supplementary File 2.4).

2.4 Results and discussion

We wish to see whether the predictions of our model are borne out by genomic data and so for eighteen genomes we aggregate retroviral insertions into two groups for easy display and analysis: those on the allosome (X or Z) and those on the autosome. That is to say, the ratio p is estimated by calculating $n_a = \sum_i n_i$ such that $i \notin \{x, y, w, z\}$ and using n_x or n_z as appropriate. We plot the ratio p in Figure 2.2 and present the raw data in Supplementary File 2.3.

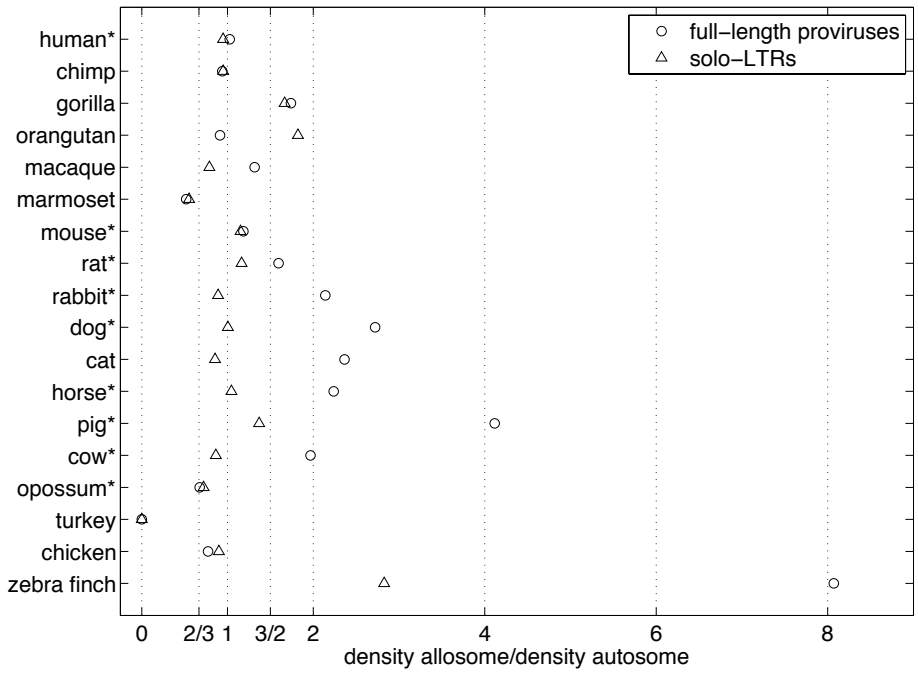

Figure 2.2: Observed ratios. Observed ratios of full-length proviruses and solo-LTRs on the allosome (X or Z) to the autosome for the genomes of 15 mammals and 3 birds. Vertical lines mark the key ratios $\frac{2}{3}$, 1, $\frac{3}{2}$ and 2. Asterisks mark the genomes we consider as trustworthy and discuss in the Results section. Mammalian full-length provirus ratios typically lie beyond 3/2, the maximum predicted value under an assumption of male-bias. Mammalian solo-LTRs are generally more evenly distributed between autosome and allosome. Mammalian solo-LTRs are also generally relatively less abundant on the allosome than full-length proviruses.

Under our model, every ratio p implies two bias values β, one for each deletion scenario. For each genome studied we make three point estimates of β, one based on solo-LTR ratios and two based on full-length proviral ratios under the scenario of recombination linked deletion and otherwise. We record the results in Table 2.1.

	Mammals								
	solo-LTRs			**Full-length recomb.**			**Full-length no recomb.**		
	point	**lower**	**upper**	**point**	**lower**	**upper**	**point**	**lower**	**upper**
human*	1.36	1.03	1.79	38.24	2.56	∞	0.86	0.09	3.32
chimp	1.36	0.89	2.08	NA	1.48	∞	1.44	0	∞
gorilla	NA	0	0.05	0.35	0	16.13	NA	0	0.7
orangutan	NA	NA	NA	NA	0.63	∞	1.73	0	∞
macaque	4.47	3.35	6.28	2.18	0.13	∞	0.03	0	1.65
marmoset	NA	NA	NA	NA	3.39	∞	NA	0.19	∞
mouse*	0.37	0.28	0.48	4.39	1.73	33.87	0.29	0	0.84
rat*	0.34	0.23	0.46	0.68	0	3.74	NA	0	0.23
rabbit*	1.99	1.53	2.62	NA	0	2.85	NA	0	0.12
dog*	0.97	0.78	1.2	NA	0	1.79	NA	NA	NA
cat	2.56	1.17	7.12	NA	0	2.02	NA	0	0
horse*	0.76	0.01	3.18	NA	0	4.98	NA	0	0.33
pig*	NA	0	0.07	NA	NA	NA	NA	NA	NA
cow*	2.38	2.02	2.82	0.03	0	1.06	NA	NA	NA
opossum*	11.22	5.12	167.79	NA	4.39	∞	77.19	0.28	∞
	Birds								
	point	**lower**	**upper**	**point**	**lower**	**upper**	**point**	**lower**	**upper**
turkey	NA	0	∞	NA	0	∞	NA	0	∞
chicken	0.53	0	∞	NA	0	∞	0.19	0	∞
zebra finch	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 2.1: Point estimates and intervals on bias β implied by measurement of: solo-LTRs distribution (left); full-length proviral distribution under the recombination linked deletion model (middle); full-length proviral distribution under the non recombination linked deletion model (right). Although each model implies a single bias β, we also ask what bias values delineate the range (lower and upper) under which we could expect to measure our observed ratios with a probability of less than 0.05. We use 'NA' to mark those situations in which no point estimate or boundary value of β can be computed. Asterisks mark the genomes we consider as trustworthy and discuss in the Results section.

As our point estimates would vary if we had counted differing numbers of viruses on the autosome or allosome, we also use equations from the Model section to identify the range of bias outside of which our ratios would be observed with probability less than 0.05. We do this by solving our equations for n_x or n_z, the number

of viruses expected on the allosome, and then finding the range of β for which the Chi-square statistic is less than 3.841, the 0.05 p-value for a 1 d.o.f. goodness of fit (Supplementary File 2.5). Where β would be less than 0 we consider a prediction non-applicable (NA).

Three aspects of our results are immediately striking. First, the ratio of full-length proviruses and solo-LTRs shows a great deal of variation between species, with both over and under abundance on the allosome represented in our results. Second, solo-LTR ratios tend to fall within the range $\frac{2}{3}$ to $\frac{4}{3}$ that make sense in the context of our model. Third, our results show that full-length proviruses are more abundant on the allosome than the autosome with the exception of orangutan, opossum, marmoset, chimp and chicken. We elaborate on these three observations in turn below.

2.4.1 LTR detection, genome variation and phylogenetic independence

Our ERV counting methodology relies on de-novo LTR retrotransposon detection. We favour the approach in this study as we expect de-novo prediction to work well on both familiar and lesser studied ERVs, an important consideration when we examine genomes that are relatively distant from human. It is important to note that we do not expect to detect all full-length proviruses but merely to detect a consistent proportion across either allosomal or autosomal DNA. A similar expectation holds for false positives: our requirement of our tool is consistency.

The actual number of proviruses we use in our study is less than can be detected in principle and our results can be improved by using more effective counting methods. In the present study we are keen to retain a structural model for ERVs and therefore reject the use of repeat masking tools that are designed to detect repetitive nucleotide sequences, that may well be fragmented, rather than accurately count individual proviruses. We proceed with the knowledge that studies combining the frame-

work from the Model subsection with better tools may well provide better estimates of bias. We certainly do not consider the data we use exhaustive but do think it a reasonable sample.

The genomes we examine vary in the number of ERVs they contain but also in how often we are willing to trust the results we obtain from them. For example, LTRharvest identifies 1,228 ERVs in the orangutan genome but we must throw roughly two-thirds away because they contain many consecutive unknown nucleotides (Ns); for the dog we identify just 177 ERVs but need discard only six percent. For this reason we consider the genomes of the cow, dog, horse, human, mouse, opossum, pig, rabbit and rat as trustworthy for our purposes as we discard less than one third of potential proviral sequences due to unknown nucleotides. We consider the remaining nine genomes less informative as the opposite is the case and we are particularly disappointed that so few full-length proviruses could be recovered from bird genomes. We largely restrict the remainder of our discussion to results from the more trustworthy genomes and mark those genomes with asterisks when appearing in tables or figures.

We do not perform a phylogenetic analysis on our results as we know that most full-length proviruses are not shared among species as closely related as human and macaque (for example, 70% of full-length sites in macaque are not present in solo or full-length form in any of human, chimp, gorilla or orangutan based on our own unpublished analysis) and because we do not draw conclusions that involve making detailed comparison between species. Rather, we examine a diverse set of animal genomes and recognise that some applications of our method, such as those on the primates, are pseudoreplications that produce non-independent results.

The heterogeneity of ERV replication affinity in a genome may be a confounding factor in our study. If some types of virus are better at infecting male germ line cells and others are better at infecting female germ line cells then the former variety will show more male bias than the latter variety. As various kinds of ERVs may have different biases it is important to note that our model treats bias (and rates of provi-

ral insertion and deletion) as averages. Similarly, as the ERVs in a genome are only derived from a relatively small number of ERV lineages, any replication affinity of particular lineages could, in principle, bias the result. For example, in humans one-third of all ERVs are descended from thirty-one to forty distinct colonizations (Bannert and Kurth, 2006; Katzourakis and Tristem, 2005).

2.4.2 Ratios of solo-LTRs

Using the results from our more informative genomes we want to address the role of male bias and recombination linked deletion in ERV biology. Our intervals for solo-LTR biases are tight (Table 2.1) and our results suggest that cow, human, opossum and rabbit all have a male biased insertion history. On the other hand, mouse and rat exhibit a female biased insertion history while dog and horse give ambiguous results. These results are surprising because, as discussed in the introduction, we expect ERV integration bias to be male oriented and positively correlated with generation time in the same way that mutational bias is.

Here our results appear to unlink deep germ lines and ERV proliferation in general, perhaps suggesting that ERV integration tends to take place during a short window of time that is unrelated to the protracted process of germ line cell division. We also think it possible that integrations might be driven by ERV expression in placental tissue (Malik, 2012; Haig, 2012). While transmission from or via placenta to progeny will affect both sexes of embryo equally, placental tissue could also re-infect maternal germ line cells. Therefore placental expression of ERVs could well have the effect of reducing male bias overall. The effect would be stronger for species in which the females spend a greater proportion of their lives bearing offspring.

Nevertheless, it is generally thought that conventional mutational male bias should increase with generation time, metabolic rate and sperm competition. A comprehensive study (Sayres et al., 2011) used age at sexual maturity, maximum life span, and interlitter interval as proxies for generation time; basal metabolic rate, body mass

and body temperature as proxies for metabolic rate; and testes-to-body mass ratio and mating patterns (polyandrous/polygynandrous versus monogamous/polygynous) as proxies for sperm competition. The conclusion was that that generation time was a powerful predictor of mutational bias but that metabolic rate was of less use. Sperm-competition appeared to be unexplanatory. While we would not necessarily expect the same results it could be argued that all of the above factors should also be positively correlated with ERV bias. The availability of closely related animal genomes means there is potential for an analogous study of the effects of life history traits on ERVs.

2.4.3 Ratios of full-length proviruses

Our results for full-length proviruses are interesting in the extent to which ERVs are over represented on the allosomes. We expect to see ratios in the range $\frac{3}{2}$ to 1 or 1 to $\frac{2}{3}$, which correspond to scenarios of male bias under recombinational deletion or otherwise. Instead, what we observe is that, among our more informative genomes, all ratios besides those for the human and opossum lie beyond the range of values predicted by our model. (We note that the opossum X chromosome is unusual in that it receives *more* recombination than the autosomes (Mikkelsen et al., 2007).) This does not mean our model is useless but instead that we must examine it more closely in order to interpret our results. Therefore we consider three general reasons that we might see an overabundance of proviruses on the allosome: dynamics, a lack of neutrality or stochasticity.

First, ratios close to the asymptotes of our model may not have been reached. Under a recombination-linked deletion scenario it is mathematically possible for LTRs to accumulate on the allosome while recombination 'catches up' and restores our predicted ratios. Although we would eventually expect to see the ratios described in the Model section enough time may not yet have passed that we actually do so. This explanation highlights a limitation of our model that can not be addressed solely by examining older proviral insertions.

Second, we may be mistaken in assuming that full-length proviruses are effectively neutral and drift to fixation. In this case factors including linkage disequilibrium, differing mutation rates between sexes, the reduced relative population size of the allosome, the heterozygosity of proviral mutations or sexual antagonism mean that we cannot say whether we would expect to see higher or lower ratios than our neutral model predicts.

For example, considering mammals, we expect the female nucleotide substitution rate to be lower than the male substitution rate, in which case proviruses on the X chromosome will receive less nucleotide substitutions than those on the autosome. Therefore, we expect that proviruses on the autosome are more likely to be made benign by random mutation than those on the X chromosome and thus are more likely to reach high frequency or fixation via drift or draft. Furthermore, as proviruses will initially be found at low frequencies, any harmful recessive effects will be felt most strongly in the hemizygous sex (Vicoso and Charlesworth, 2006) and so selection against proviruses may be more effective on the X chromosome, also enhancing the relative number of proviruses we might expect to see on the autosome. Both these effects would act in the same direction and increase apparent male bias.

However, the extent to which the above effects hold is not known. For example, while ectopic recombination might be a major harmful consequence of carrying a proviral insertion, it is an open question as to whether it is generally healthier for a host to be homozygous for a proviral insertion or whether other factors dominate; is an ERV best modelled as a recessive harmful mutation? Furthermore, any sexual antagonism in the effects of proviral insertions can shift our expectations of relative abundance of proviruses in either direction. For example, we would expect to see more fixation of proviruses on X for recessive mutations that are of benefit to males but harmful to females or for dominant mutations that are of benefit to females but harmful to males (Rice, 1984). Dominance and antagonism effects are examples of unknown factors that can decrease any apparent male bias or lead to an apparent female

bias instead.

Overall, these complexities are such that we cannot incorporate selection into our framework without knowing more about the harm full-length proviruses cause. The explanation that ERVs are non-neutral implies a misapplication of our model and might possibly be supported by the recent observation that in mouse (Nellåker et al., 2012) there are about 75% of the expected number of polymorphic (unfixed) ERVs on the X chromosome yet close to the expected amount of fixed ones. As an apparent underabundance of TEs on the X chromosome is reduced over time this evidence suggests that polymorphic ERVs are more likely to fix on the X chromosome than the autosome. In this case we note that if solo-LTRs are also deleterious, or if the process of proviral deletion often occurs when proviruses are fixed or at a high frequency, then our estimates of bias obtained from solo-LTRs will also be an underestimate.

Third, our results may genuinely reflect the processes described by our model in many cases. On a species-by-species basis the overrepresentation of full-length proviruses that we see on the X chromosome is often statistically compatible with a range of positive bias under both recombination linked and non recombination linked deletion. As Table 2.1 shows, under the recombination linked scenario fourteen of the fifteen observed ratios are statistically acceptable and ten of the fifteen have a bias that is compatible with that obtained from the corresponding solo-LTR ratios. Under the non recombination linked scenario twelve of the fifteen ratios are statistically allowable and eight of these are compatible with the corresponding solo-LTR ratios. We acknowledge that we find these wide intervals an uncomfortable shortcoming of an approach relying on comparing a small allosome to a large autosome.

Of these three explanations, the first two are applicable in the case that a lack of recombination is in one way or another responsible for the overabundance of full-length proviruses that we highlighted above. With respect to the third explanation, we find our observations are more often suggestive of recombination linked deletion than otherwise. Given we know that opossum X chromosome biology is unusual,

and also that when our ratios are statistically problematic they tend to be too large, it is reasonable for us to favour a scenario of recombination linked deletion and to question the assumption that ERVs are neutral alleles.

Of course, we may also see ratios that fall outside of our range of predictions as a reason to reject our model entirely, in which case we are obliged to look for some other explanation of what exactly it is about the X chromosome that results in full-length proviruses being more abundant there. Nevertheless, in either case, if full-length proviruses can persist for longer on the X chromosome then it is likely that if we look more closely we will find that virus that integrate there are more successful replicators than those who arrive elsewhere.

2.5 Conclusions

We predicted the allosomal to autosomal ratio of full-length proviruses we would expect to see under a neutral model given recombination linked deletion or otherwise. Using bioinformatics tools we detected an excess of full-length endogenous retroviruses on the sex chromosomes of eleven mammals and one bird. We also observed overall patterns of endogenous retroviral abundance that, under a neutral model, are not universally male biased. We suggest that a recombination linked deletion process or non-neutral alleles best explain our observations and that, in future, population data and a better quantification of the harm caused by full-length proviruses can help us more accurately explain their relative frequencies on sex chromosomes.

Chapter 3

Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split

3.1 Abstract

Background

Endogenous retroviruses (ERVs) are often viewed as selfish DNA that do not contribute to host phenotype. Yet ERVs have also been co-opted to play important roles in the maintenance of stem cell identity and placentation, amongst other things. This has led to debate over whether the typical ERV confers a cost or benefit upon the host. We studied the divergence of orthologous ERVs since the chimp-human split with the aim of assessing whether ERVs exert detectable fitness effects.

Results

ERVs have evolved faster than other selfish DNA in humans and chimp. The divergence of ERVs relative to neighbouring selfish DNA is positively correlated with the

length of the long terminal repeat of an ERV and with the percentage of neighbouring DNA that is not selfish. ERVs from the HERV-H family have diverged particularly quickly and in a manner that correlates with their level of transcription in human stem cells. A substitution into a highly transcribed HERV-H has a selective coefficient of the order of 10^{-4}. This is large enough to suggest these substitutions are not dominated by drift.

Conclusions

ERVs differ from other selfish DNA in the extent to which they diverge and appear to have measurable effects on hosts, even after fixation. The effects are strongest for HERV-H and suggest that the HERV-H transcriptome has recently evolved under the influence of directional selection. As there are many HERV-H loci distributed across the ape lineage, our results suggest that in future this family can be used to model the evolutionary consequences of ERV exaptation in primates and other mammals.

3.2 Background

As an obligate part of their lifecycle, retroviruses integrate their genomes into their host's nuclear DNA. This integrated retroviral genome is referred to as a provirus. Sometimes integration occurs in a germ line cell, and if the integration is not too damaging to the host, then it becomes possible for proviral DNA to be passed in a vertical (Mendelian) way from parent to offspring. An initial vertical transmission is known as an endogenization and the inherited proviral DNA is known as an endogenous retrovirus (ERV). Over time, some ERVs reach high frequencies or fixation in a host population and it is therefore possible to detect the traces of ancient viral infections, often in fragmented form, by examining modern genomes.

As transposable elements (TEs) with an RNA intermediate form, ERVs can be thought of as selfish DNA (Orgel and Crick, 1980; Doolittle and Sapienza, 1980). The

term 'selfish DNA' refers to sequences that are present in genomes in multiple copies largely due to their ability to replicate themselves rather than because they provide any benefit to the host. Although selfish entities can replicate, they do not seem to expand genomes indefinitely, probably because they impose a selective cost (Boissinot et al., 2006; Petrov et al., 2011). Selection will act against individual TEs, especially if they are very harmful. This selective cost has lead to the evolution of host defences (Johnson, 2007). Host defences are not perfect however, and TEs can still saturate a genome unless selection against them increases sufficiently quickly with respect to mean element copy number per individual (Johnson, 2007; Charlesworth et al., 1994). In other words, as TEs do not fill up our genomes, population genetics suggests they must be harmful, and as TEs can increase their copy number, some fraction of TEs may fix, even when they have a cost to their host.

The cost of harbouring TEs is often categorized as arising in three ways (Gonzalez and Petrov, 2012), all of which apply to ERVs. The first cost of TEs is due to the fact that they can be present in many copies in the genome. As repetitive sequence they may increase the occurrence of ectopic recombination whereby meiotic crossover occurs between TEs from the same family that are located in non-homologous parts of the genome (Hughes and Coffin, 2001). The probability of ectopic recombination between two sequences is thought to be related to length of uninterrupted similarity between them (Opperman et al., 2004), and as ERVs are longer than typical TEs, two particular ERVs may be more likely to ectopically recombine than, say, two particular SINEs. The second cost of TEs is due to the possibility that an element may insert itself into a functional region of the genome in a way that disrupts the ability of the host to survive. Insofar as ERVs retain their ability to retrotranspose (i.e. insert a copy of themselves into a new chromosomal location within a cell) or to reinfect (i.e. insert a copy of themselves in a potentially different cell after performing a cell exit and subsequent cell entry), it is clear that ERVs present the same risks as other TEs in this respect. The third cost of TEs is the cost to the host due to the mechanism of repli-

cation itself. For ERVs, particularly recently integrated ones, this cost may be severe, as ERVs contain viral genes that were selected to allow exogenous viruses to circulate between hosts. This means that in addition to the side-effects that are common to all retrotransposons, such as those due to the production of an intermediate RNA form, ERVs can have additional effects. An example of an additional effect is virion formation, the costs of which can include immune responses or the infection and mutagenization of cells throughout the body (Young et al., 2013). Indeed, it is ERVs that mitigate the consequences of their history as horizontally infectious agents by losing their envelope gene that are exactly those that proliferate most effectively in the long term (Magiorkinis et al., 2012).

Despite the ways in which ERVs can be harmful, there are an increasing number of described cases where ERVs may be conferring some benefits to their host. For example, recent debate has occurred over the significance of the fact that ERVs exhibit relatively high levels of placental transcription (Sanford et al., 1985; Red-Horse et al., 2004; Golding et al., 2010; Rowe and Trono, 2011; Smith and Meissner, 2013), the fact that some retroviral promoters are exclusively expressed in the placenta (Schulte et al., 1996; Bi et al., 1997), and the fact that genes derived from ERVs have frequently been co-opted for placental function (Lavialle et al., 2013). One suggestion, as proposed by Chuong (2013), is that ERVs and the placenta are in symbiosis: placental expression of ERVs is tolerated because ERVs were involved in the origin of the placenta via the creation of the trophoblast cell lineage and because, since then, ERVs have continued to play important roles in placental function. It is argued that long terminal repeats (LTRs) of ERVs act as mobile promoters that can rapidly rewire gene regulation networks in a way that may be crucial to the origin and evolution of a new cell type. This hypothesis is interesting but controversial (Haig, 2012, 2013) as from a viral perspective placental expression may allow ERVs to segregate with greater than even odds from heterozygous mothers and also provide a mechanism by which a father can infect a mother and all of her future offspring.

A more concrete example of exaptation also hinges on the ability of ERVs to facilitate widespread transcriptional rewiring and comes from studies that highlight the participation of ERVs in the initiation and maintenance of stem cell identity. It has been shown that of 1,225 full-length copies of HERV-H in the human genome, 550 are actively transcribed in human pluripotent stem cells at levels that are positively correlated with the integrity of their 5′ LTRs (Wang et al., 2014). In human embryonic stem cells, the transcription factor LBP9 has been shown to drive production of stem cell specific HERV-H associated chimeric transcripts and long non-coding RNAs (lncRNAs), the latter having been shown to be essential for the maintenance of a stem cell like state (Wang et al., 2014). Elsewhere it has been independently shown that HERV-H knockdown downregulates pluripotency markers, and that HERV-H transcription is necessary for both the creation and maintenance of stem cell identity (Lu et al., 2014). Furthermore, a large scale analysis of both the mouse and the human stem cell transcriptome suggests that LTR derived transcripts are under the direct control of the main stem cell specific transcription factors (Fort et al., 2014). Research on mouse has produced related results, and the MuERV-L family of ERVs has been shown to produce chimeric transcripts originating from over 300 LTR loci, the activity of which appear to grant some totipotent like properties to induced and embryonic stem cells (Macfarlan et al., 2012). The weight of evidence from these studies does suggest that, at least for some part of their history, a proportion of ERVs have contributed in important ways to host function.

In this paper, we consider the degree to which ERVs in general are active parts of the genome rather than inert sequences that lost their effects on hosts prior to fixation. Given viruses and TEs can be so disruptive to the host, ERVs that are observed in contemporary genomes have often been assumed to be effectively harmless and to evolve neutrally. However, we do not have a clear picture of the costs, benefits and frequency of ERVs in ancient populations that are necessary to support such assumptions. At one extreme, some ERVs we observe today may be members of families that

were both prolific and harmful in ancestral populations, so that the fixation of some deleterious ERVs was an inevitable consequence of their ability to replicate quickly. On the other hand, ERVs may have been frequently co-opted due to the pre-packaged functions they provided, with the benefits of these functions balancing out any deleterious side effects. In this study we examine orthologous ERVs in human and chimp genomes and compare their divergence since the split between the two species. If ERVs are indeed inert they should have evolved neutrally after they reached fixation. On the other hand, if ERVs had an effect on the host they should have evolved at rates that differ from the neutral rate. In particular, ERVs that are conserved will have evolved more slowly than the neutral rate while ERVs should only have evolved more quickly than expected if they were useful to the host and underwent adaptation, or if they were still harmful to the host and were degraded.

3.3 Results

We wanted to see if recently integrated proviruses accumulated mutations more quickly than neighbouring DNA. Our approach was to examine substitutions into ERVs and their neighbouring genomic sequence that lead to differences between human and chimp. To achieve this goal we identified ERVs and their flanking DNA from both species. Using bioinformatics tools, we searched the human and chimp genomes for full-length ERVs using a broad spectrum of retroviral probes. We then attempted to associate the results of our search process in terms of orthology: by using a two stage pairwise alignment process we deemed sufficiently similar sequences originating from syntenic chromosomes in different species as paired orthologues. In the rare case that there was evidence of paralogy we excluded all the paralogous regions from the study. Overall, we identified 336 chimp-human pairs of sequence from a variety of genomic locations (Table 3.1). The ERVs in the sequence were from a variety of families (Table 3.2). We carefully pairwise aligned these ERV containing sequences, masking regions

that were badly aligned and could not be safely included in the study.

linkage	count
1	29
2	29
3	38
4	22
5	11
6	25
7	36
8	22
9	9
10	11
11	16
12	10
13	8
14	10
15	7
16	5
17	2
18	2
19	19
20	2
21	8
22	0
X	15

Table 3.1: Chimp-human orthologue linkage. We detected 336 pairs of ERV containing sequence from chimp and human genomes. Note: ch2a/2b (chimp) were paired with ch2 (human).

Each of the 336 pairs of ERVs in our study are contained in a 40 kb region of DNA. Inspection of these regions reveals they are mostly comprised of repetitive elements. Some of these repetitive elements are typically selfish (e.g. DNA transposons) whereas a minority (e.g. tRNA) are essential to the host. Substitution into regions that are useful to the host will generally be constrained as mutations in these regions are likely to be deleterious. We are interested in whether substitutions into ERVs are more common than substitutions into other selfish elements. To determine this we classified all columns of our alignments as one of: provirus (PV); repetitive and selfish DNA (RM^{+}); and non-repetitive or repetitive but non-selfish (RM^{-}). The sequence

family	autosomal	X-linked
ERV-9	57	1
HERV-ADP	1	0
HERV-E	10	1
HERV-F type_b	2	0
HERV-H	58	6
HERV-I	23	0
HERV-K(HML2)	101	4
HERV-K(HML5)	11	1
HERV-K(HML6)	12	0
HERV-K(HML9)	1	0
HERV-L	1	0
HERV-P	2	0
HERV-R	5	0
HERV-T	4	0
HERV-U3	1	0
HERV-W	19	0
HERV-XA	1	0
RRHERV-I	6	1
Unclassified	6	1

Table 3.2: Chimp-human orthologue family, by linkage. ERV family was assigned using the best matching viral pol probe (see Detecting ERVs in Methods).

classified as PV was the result of our original search for ERVs and the categories RM^+ and RM^- were assigned to the flanking regions of ERVs by using RepeatMasker annotations. Because CpG sites are known to mutate quickly, we censored these sites in our analyses; all results pertain to censored analyses unless we explicitly state otherwise. Overall, the following site patterns were observed for each of the three categories of sequence (Table 3.3, Table S1 in Supplementary File 3.1).

chimp:human	**autosomal**			**X-linked**		
	PV	**RM^+**	**RM^-**	**PV**	**RM^+**	**RM^-**
A:A	591392	1603551	1145366	26509	83350	43123
A:T	888	2496	1390	38	105	45
A:G	3650	8749	5067	112	335	155
A:C	1104	2678	1593	41	111	46
A:?	0	0	0	0	0	0
T:A	984	2572	1456	28	90	41
T:T	597941	1609580	1151417	26035	87639	43695
T:G	1197	2816	1527	57	109	62
T:C	3636	8534	4949	127	338	172
T:?	0	0	0	0	0	0
G:A	3294	8415	4950	131	389	144
G:T	1143	3034	1688	35	108	51
G:G	450769	1135140	703983	22571	56854	24790
G:C	974	2586	1439	29	105	51
G:?	0	0	0	0	0	0
C:A	1206	2935	1652	26	102	46
C:T	3394	8432	4957	120	328	183
C:G	986	2623	1400	31	91	55
C:C	460550	1128491	698783	18236	56537	25202
C:?	0	0	0	0	0	0
?:A	5	0	2	0	0	0
?:T	0	1	2	0	1	0
?:G	1	0	1	0	1	0
?:C	1	4	0	0	0	0
?:?	0	0	0	0	0	0
total	**2123115**	**5532637**	**3731622**	**94126**	**286593**	**137861**

Table 3.3: Site patterns observed across CpG censored alignments. Patterns were observed at sites classified as one of: ERV (PV); selfish DNA (RM^+); or non-repetitive or repetitive but non-selfish (RM^-).

Hoping to take account of any differences in local mutation rates in the genome, we first considered each of the 336 pairs of virus containing sequences individually i.e.

due to their physical co-location, we considered PV and RM^+ as paired measurements. We found that PV divergence is significantly greater than RM^+ divergence for autosomal ERVs (Wilcoxon signed-rank test, $W = 32602.5$, $p < 0.0001$) with a small median difference of 0.001 substitutions per site. We also found that median PV divergence was greater for autosomal ERVs than for X-linked ERVs (Wilcoxon signed-rank test, $W = 3178$, $p = 0.018$) by a distance of 0.002 substitutions per site.

As we found that proviruses diverged faster than other selfish DNA we wanted to see if this effect was related to the age of viral integration. To do this we searched for each ERV's full-length representative in the gorilla, orangutan, and macaque genomes, using the same method as that for the human and chimp. For each chimp-human orthologue we aimed to identify the lineage that split earliest from the lineage leading to chimp/human that also contained the particular ERV in question. In other words, we identified a minimum age bound for each ERV by examining progressively more distant relatives. As this approach relied on the ability of LTR detection software to detect a full-length ERV in more than one species, the age classification was approximate. There were 187 ERVs for which no additional orthologue was found (CH) and 149 ERVs that were confirmed to be at least as old as the gorilla split (CH^+): there were 77 ERVs for which gorilla was the earliest split (CHG), 70 ERVs for which orangutan was the earliest split and 2 ERVs for which macaque was the earliest split ($CHGO^+$). Considering the difference between PV divergence and RM^+ divergence we found that the potentially youngest ERVs (CH) had diverged significantly more since the chimp-human split than those that were confirmed to be at least as old as the gorilla split (CHG) but that there was no significant difference between PV divergence and RM^+ divergence for the CHG and $CHGO^+$ categories (Figure 3.1). We therefore report that the potentially youngest ERVs (CH) had diverged significantly more (Wilcoxon signed-rank test, $W = 16110$, $p < 0.01$) since the chimp-human split than those that were confirmed to be at least as old as the gorilla split (CH^+). The median difference in divergence between PV and RM^+ was 0.0012 substitutions per site for alignments

in CH and 0.0003 substitutions per site for alignments in CH^+.

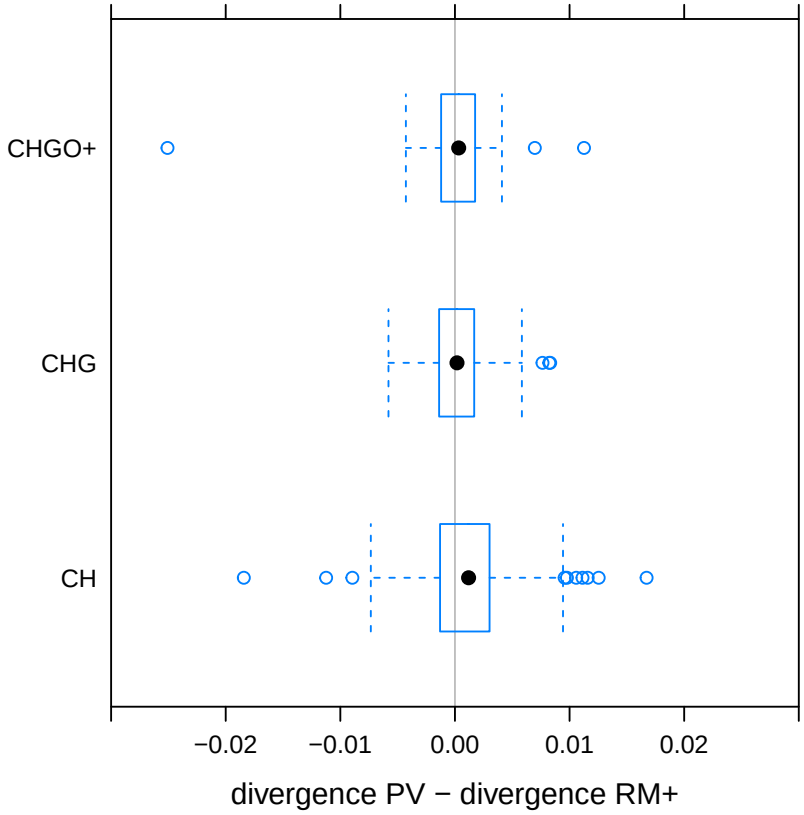

Figure 3.1: Difference in divergence between PV (ERVs) and RM^+ (selfish DNA) for chimp-human orthologues, aggregated by age. Age categories were assigned to chimp-human orthologues by identifying the most distant primate relative in which the orthologous sequence could also be found. Category CH contains ERVs detected in chimp and human only (187 ERVs); category CHG contains ERVs for which gorilla was the most distant relative in which an ERV was detected (77 ERVs); category $CHGO^+$ contains ERVs for which orangutan (70 ERVs) or macaque (2 ERVs) was the most distant relative in which an ERV was detected. (Note: whiskers estimate 95% confidence intervals, filled dots represent median values, unfilled dots represent outliers.)

As can be seen in Figure 3.2, it appears as if HERV-H is responsible for much of the divergence in the CH category. This was confirmed by re-running our analyses with the 64 HERV-H removed from our dataset. In this case, a significant age effect was no longer observed. Further investigation showed that the difference in divergence between PV and RM^+ is significantly greater for HERV-H than for ERVs that

are not classified as HERV-H (Wilcoxon signed-rank test, $W = 12675$, $p < 0.0001$) with a median difference between PV and RM^+ of 0.0026 substitutions per site for HERV-H and 0.0003 substitutions per site for ERVs that are members of any other family. Assuming that substitutions into RM^+ are the result of neutral semi-dominant mutations, the ratio of these divergence values suggests a median selection coefficient of 2.3×10^{-5} for younger CH ERVs. Moreover, the upper quartile (16 out of 64) of all HERV-H selection coefficients are not small ($2Ns > 1$), ranging from 5.0×10^{-5} to 2.0×10^{-4}.

The differences we observed between PV and RM^+ were quite large. For this reason we examined how divergence related to transcription, for HERV-H orthologues only, and to virus length, LTR length and the percentage of an ERV's environment that was selfish (RM^+) for all orthologues. Pairing our orthologues with transcription activity data (Wang et al., 2014) we found that the log ratio of PV divergence to RM^+ divergence was significantly correlated with the log of the average transcription level of HERV-H in human embryonic stem cells (hESC) and induced pluripotent stem cells (hiPSC) using both linear models ($R^2 = 0.23$, $p < 0.0001$) (Figure 3.3) and nonparametric tests (Kendall's rank correlation, $\tau = 0.315$, $p < 0.001$). Using the transcription activity categories of (Wang et al., 2014), we further found that this divergence ratio was higher for 12 "highly-active" ERVs than for 22 "moderately active" ERVs (Wilcoxon signed-rank test, $W = 197$, $p < 0.01$), the 22 "moderately active" ERVs in turn had a higher divergence ratio than the 29 "inactive" ERVs (Wilcoxon signed-rank test, $W = 424$, $p = 0.023$) (Figure 3.4). The median selection coefficients for transcriptionally "highly active", "moderately active" and "inactive" HERV-H ERVs are 5.7×10^{-5}, 2.6×10^{-5} and 1.3×10^{-5} respectively. We further found that, across all ERVs, the log ratio of PV divergence to RM^+ divergence was significantly positively correlated with LTR length (Kendall's rank correlation, $\tau = 0.121$, $p < 0.001$) and significantly positively correlated with the percentage of the flanking DNA of an ERV that is non-selfish (RM^-) (Kendall's rank correlation, $\tau = 0.140$, $p < 0.0001$). These

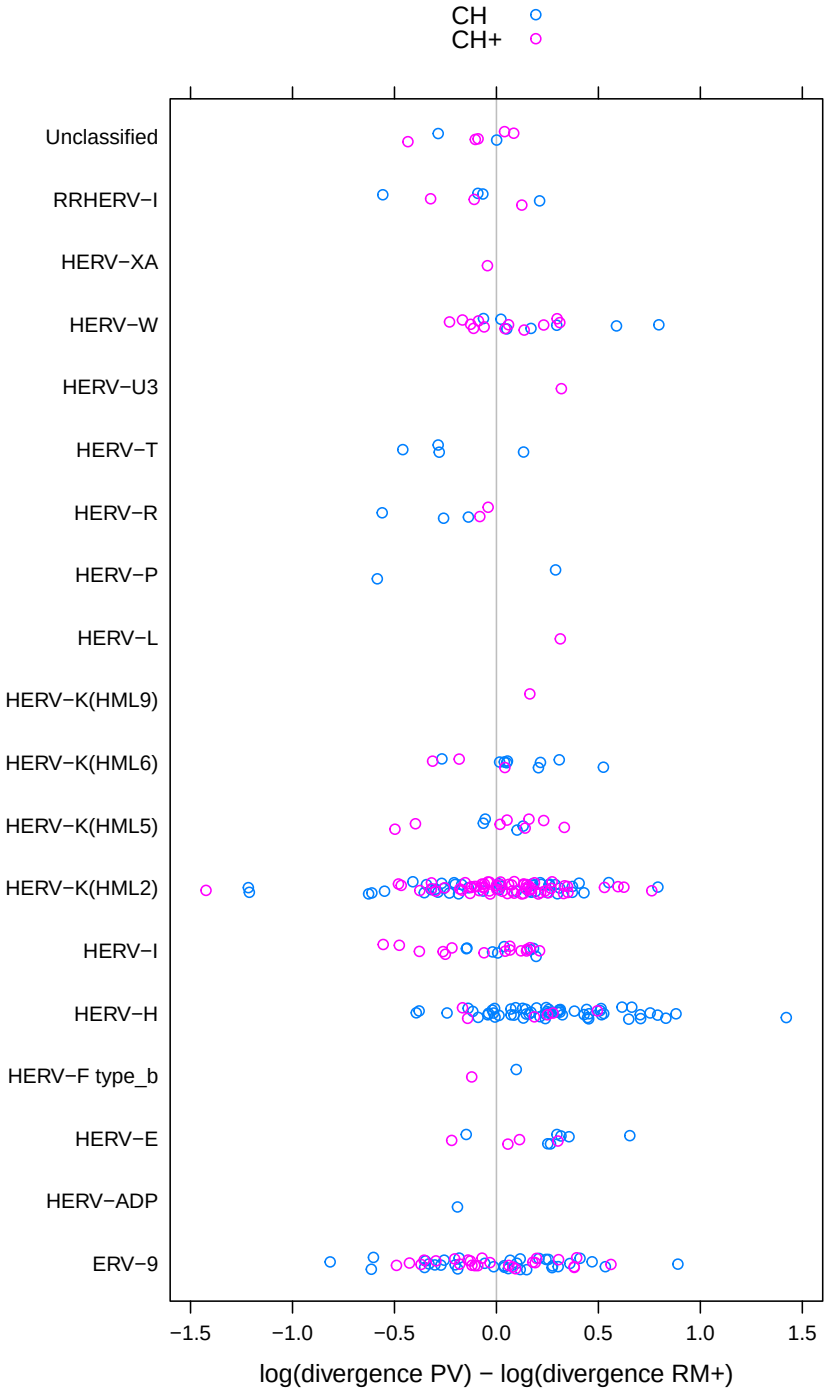

Figure 3.2: Relative divergence of PV (ERVs) and RM$^+$ (selfish DNA) aggregated by age and ERV family. ERV family was assigned using the best matching viral pol probe (see Detecting ERVs in Methods). Age categories were assigned to chimp-human orthologues by identifying the most distant primate relative in which the orthologous sequence could also be found. Category CH contains ERVs detected in chimp and human only (187 ERVs); category CH$^+$ contains ERVs detected in a primate beyond human and chimp (149 ERVs).

correlations remained significant ($p < 0.01$) even if HERV-H were excluded from our dataset. We did not find a positive correlation between virus length and divergence (Kendall's rank correlation, $\tau = 0.10$, $p = 1.00$).

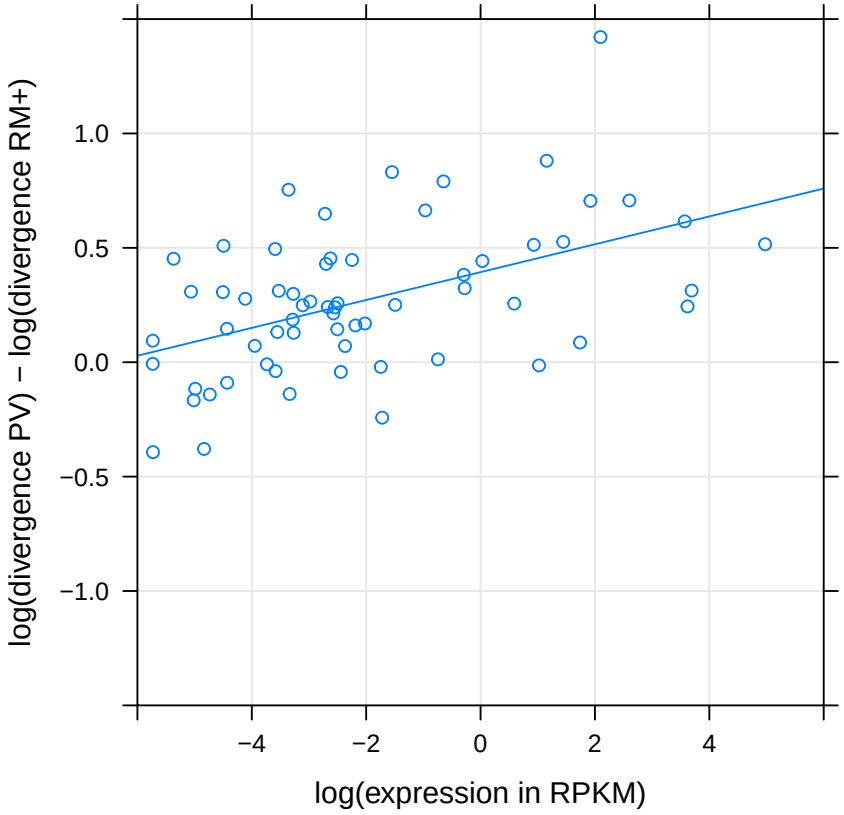

Figure 3.3: Relative divergence of HERV-H loci and RM^+ (selfish DNA) correlates with stem cell transcription. The log of the average transcription level (in reads per kilobase of transcript per million reads mapped) (Wang et al., 2014) of 63 HERV-H loci across hESC and hiPSC is correlated ($R^2 = 0.23$, $p < 0.0001$) with their divergence since the chimp-human split.

Our results show that ERVs (PV) experience faster evolution than nearby selfish DNA (RM^+), particularly if the ERVs are potentially younger (CH), and particularly if they are HERV-H. Our results also show that ERVs evolve faster if they have longer LTRs and are located regions of the genome with less selfish DNA, and that autosomal ERVs evolve faster than X-linked ERVs. The faster evolution of ERVs than nearby selfish DNA might be due to selective forces or to mechanistic factors.

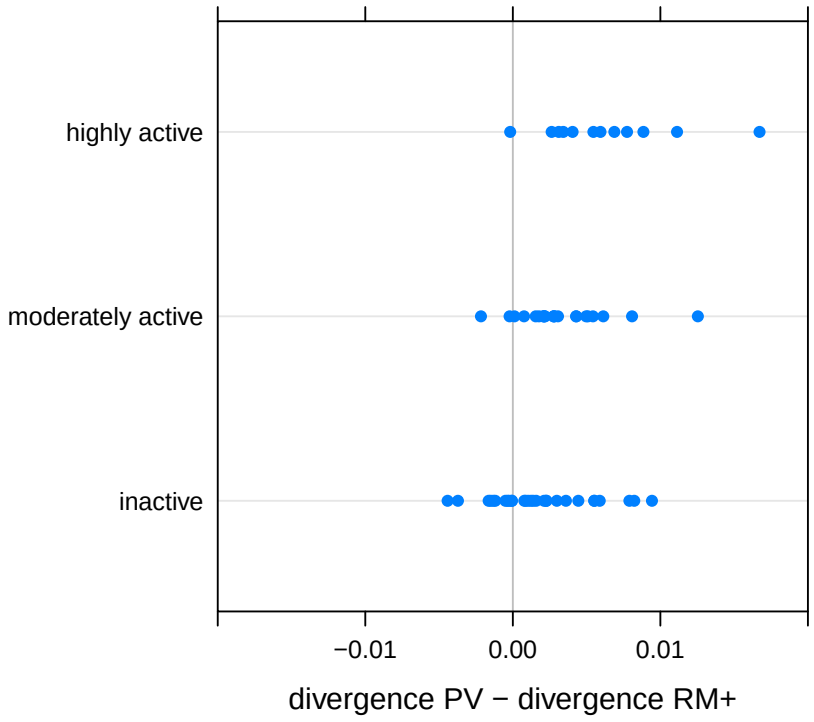

Figure 3.4: Fig. 4 Excess divergence of HERV-H grouped by categorical transcription levels in human stem cells. The difference in divergence between PV (ERVs) and RM^{+} (selfish DNA) of 63 HERV-H loci (12 "highly active", 22 "moderately active", 29 "inactive") increases with their categorical transcription levels (Wang et al., 2014) across hESC and hiPSC.

To investigate sex-effects and dominance, as well as the aforementioned mechanistic factors, we aggregated the sequence from our 336 orthologous stretches of ERV containing DNA, combining sequence based on its linkage (autosomal or X-linked) and its classification (PV, RM^+ or RM^-). We found that ERVs (PV) diverged more quickly than repetitive and selfish flank (RM^+), that in turn diverged more quickly than non-repetitive or repetitive but not selfish flanking DNA (RM^-) (Table 3.4, Figure 3.5). This was true for the autosome and the X-chromosome, whether or not we censored CpG sites. The divergence values in (Table 3.4) imply selection coefficients of 1.3×10^{-5} and 2.4×10^{-5} for autosomal and X-linked ERVs before the censoring of CpG sites and 4.7×10^{-6} and 6.7×10^{-6} after censoring. We observe that in all cases these are small forces ($2Ns < 1$) and that for both censored and uncensored sites the ratio of autosomal to X-linked relative divergence suggests that mutations into ERVs are recessive.

		uncensored (CpG^+)			censored (CpG^-)		
linkage	class	$EQ^{+/-}$	EQ^+	EQ^-	$EQ^{+/-}$	EQ^+	EQ^-
A	PV	0.01649	0.0143	0.00222	0.01066	0.00885	0.00182
A	RM^+	0.01446	0.0123	0.00217	0.01017	0.00831	0.00187
A	RM^-	0.01144	0.00969	0.00177	0.00865	0.00712	0.00154
X	PV	0.01365	0.01182	0.00184	0.00829	0.00694	0.00135
X	RM^+	0.01087	0.00929	0.00159	0.00776	0.00639	0.00137
X	RM^-	0.01005	0.00841	0.00166	0.00767	0.00627	0.0014

Table 3.4: Divergence aggregated by class, linkage, CpG censoring, and differences used. Differences used were classified as: all differences ($EQ^{+/-}$); CG equilibrating differences only (EQ^+); and non CG equilibrating differences only (EQ^-). Sites were classified as one of: ERV (PV); selfish DNA (RM^+); or non-repetitive or repetitive but non-selfish (RM^-).

In our study we make comparisons between ERVs (PV) and repetitive and selfish DNA (RM^+) that are paired as we expect pairs of sequence to share a similar genomic environment e.g. similar mutation rates. We also compare the aggregate of all ERVs in our study to the aggregate of all repetitive and selfish DNA in our study. This aggregation disassociates paired ERV and flanking sequence, yet an elevated divergence effect is still visible for ERVs. We found the difference between PV and RM^+

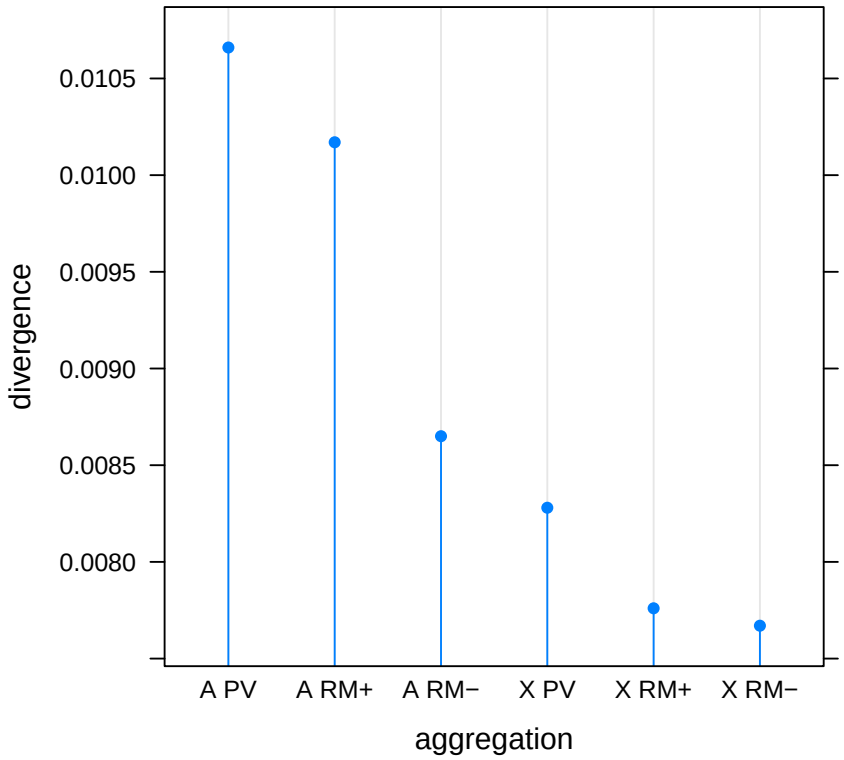

Figure 3.5: Divergence aggregated by linkage and sequence classification. ERVs (PV) diverge faster than selfish DNA (RM^+) which diverges faster than non-repetitive or repetitive but non-selfish DNA (RM^-). Autosomal loci (A) diverge faster than X-linked (X) loci.

under aggregation to be 0.0005 substitutions per site i.e. effectively the same as the small median difference between paired autosomal PV and RM^+ sequence of 0.001 substitutions per site that we mention above. Nevertheless, all repetitive and selfish DNA discussed so far originated from a location within 40 kb of a full-length ERV by experimental design.

Given the high divergence of HERV-H orthologues, we conducted an additional analysis targeting the six highly active HERV-H orthologues that could be located in long primate alignments. Our motivation was to explore whether ERVs drawn from the fastest diverging group in our study could still be considered to be diverging quickly if we compared them to RM^+ regions located at greater distances. This analysis revealed that HERV-H orthologues were local divergence maxima (Figure 3.6) and also that an equivalent or greater divergence occurs only when analyzing regions centered on 1–13% of the loci in these alignments. Furthermore, examining the neighbourhood of the ERVs it is clear that they are not located exclusively in regions that are otherwise slowly evolving (plots for ch5 and ch7 reveal nearby sequence that diverges at greater than the alignment mean) but neither are they located exclusively in regions that evolve quickly as a whole (plots for ch14 and chX reveal nearby sequence that diverges at less than the alignment mean). These analyses suggest that our results are not a consequence of ERVs (PV) depressing the divergence of nearby repetitive and selfish flank (RM^+). Additionally, as these results indicate that we could find regions that diverged either faster or slower than any particular ERV if we looked far enough away, they support our decision to consider regions that are close to and of a comparable length to ERVs in our other analyses.

Other factors beside selection can influence substitution rates. These include a mutation bias that means that GC nucleotides preferentially decay into AT nucleotides and biased gene conversion. The effect of biased gene conversion may be quite small, but it can be expected to favour the segregation of GC over AT nucleotides. We investigate these two effects below. Interestingly, for RM^+ sequences, we found that

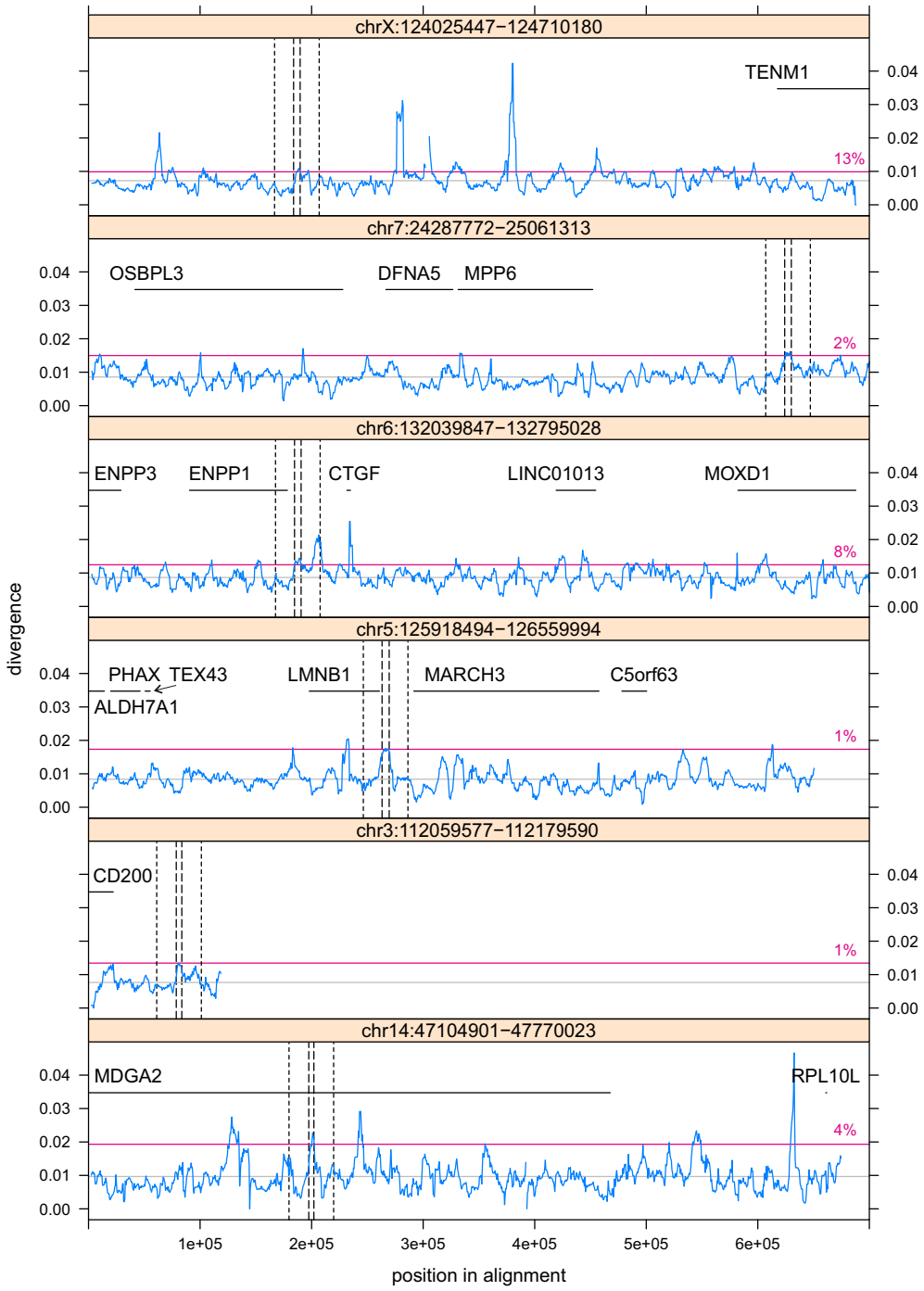

Figure 3.6: Divergence of six long alignments containing "highly active" HERV-H loci. The divergence of RM^+ (selfish DNA) including HERV-H sequence (PV) is plotted (blue line) against alignment coordinates using a sliding window of the same length as the HERV-H in each alignment. The grey horizontal line represents the mean divergence of RM^+ across the alignment. The magenta horizontal line is a reference line indicating the divergence of the window centred on the HERV-H (i.e. the divergence of PV); the associated percentage gives the percentage of windows for which divergence is at least as great as the divergence of the HERV-H. Inner vertical dashed lines mark a window centred on the HERV-H. Outer vertical dotted lines mark a region of length 40 kb that is centred on the HERV-H. RefSeq gene annotations appear in black.

divergence was not significantly correlated with GC content for both the CH category (Pearson's product-moment correlation, $r = 0.13$, $p = 0.07$) and for the CH^+ category (Pearson's product-moment correlation, $r = 0.01$, $p = 0.86$). We further found that, for PV sequences, divergence was not significantly correlated with GC content for either the CH category (Pearson's product-moment correlation, $r = 0.05$, $p = 0.48$) or for the CH^+ category (Pearson's product-moment correlation, $r = 0.07$, $p = 0.37$). This demonstrates that GC content has not driven the divergence of the ERVs or nearby selfish DNA in our dataset. (Indeed, it is visually clear that different ERV families maintain distinct GC compositions on the timescale of our study as is shown in Figure S1 in Supplementary File 3.1) Nevertheless, as we observed that a large fraction of young CH ERVs with larger differences between PV and RM^+ divergence were classified as HERV-H (Figure 3.2), a family with relatively high GC content, we also performed AIC forward-backward stepwise model selection with the log of the ratio of PV divergence to RM^+ divergence as a response variable and age (CH/CH^+), ERV family (HERV-H or not HERV-H), and the log ratio of PV to RM^+ GC content as explanatory variables. We found that ERV family was the only significant predictor retained by this process, further evidence that the faster evolution of ERVs (PV) compared to their neighbouring selfish DNA (RM^+) was not due solely to differences in GC content.

As both mutation bias and gene conversion would act to introduce differences that changed GC content, we also divided all substitutions ($EQ^{+/-}$) into equilibrating mutations (EQ^+) between G or C and A or T and non-equilibrating mutations between G and C or A and T (EQ^-) (Table 3.4). Consistent with the above results, for mutations that were EQ^+, we found that PV sequence evolved faster than RM^+ sequence that in turn evolved faster than RM^- sequence. In contrast, we found that in the EQ^- category, RM^+ sequence actually diverged slightly more than PV sequence on both the autosome and the X-chromosome. We note that transitions are excluded from the non-equilibrating EQ^- category, and that sequences diverge roughly ten-times less

when only these substitutions are considered.

As we had observed that censoring CpGs reduced divergence by up to 0.006 substitutions per site we examined the dinucleotide composition of our data (Table S2 and Figure S2 in Supplementary File 3.1). We found that the dinucleotide composition of PV sequence differed significantly from RM^+ and RM^- sequence and so we cannot formally rule out the possibility that the greater divergence of PV versus RM^+ is due to unidentified context dependent effects.

3.4 Discussion

We have shown that endogenous retroviruses (ERVs) have diverged more at the nucleotide level than other selfish DNA since the chimp-human split. We have further shown that this effect is positively correlated with both the length of an ERV's LTR and with the percentage of an ERV's neighbouring DNA that is non-repetitive or non-selfish. The faster evolution of ERVs is especially noticeable for younger members of the HERV-H family, in which case the relative divergence of an ERV when compared to neighbouring selfish DNA correlates well with the level of transcription of the ERV in human stem cells. Our results show a hierarchy of divergence, with ERVs having diverged more than selfish DNA, which in turn has diverged more than non-repetitive or repetitive but non-selfish sequence. We have attempted to rule out mechanistic explanations for our observations and suggest that directional selection is responsible for our results. If the higher divergence of ERVs when compared to other selfish DNA is due to selection then the relative rate of evolution on the autosome compared to the X-chromosome suggests that the mutations that are acted upon are, on average, recessive in nature.

One explanation for selection leading to a faster substitution rate into ERVs than other selfish DNA relates to the cost of an ERV's mechanism of replication. More than a dozen ERVs in the human genome contain open reading frames (Young et al.,

2013; Lavialle et al., 2013) but none of the consensus sequences from the ERVs we examined (where present in more than two species) did. This accords with the notion that ERVs are generally fragmented. However, ERVs can have many effects that do not depend on complete coding genes. In general, ERVs can act as promoters or enhancers in opposition to the interests of the host by recruiting transcription factors and interfering with the regulation of nearby host genes (Gonzalez and Petrov, 2012; Isbel and Whitelaw, 2012). The effect of this family of disruption can be severe, as is the case for Hodgkin's lymphoma, which appears to be conditional upon the derepression of MaLR LTRs (Lamprecht et al., 2010). The transcription of ERVs also diverts RNA polymerase from host genes and produces mRNA that may interfere with the preferred regulatory dynamics of the host cell (Young et al., 2013). In some cases such transcripts are known to trigger harmful autoimmune responses such as those that occur in TREX1 deficient mice (Gall et al., 2012) while in other cases transcripts have been shown to hybridize to produce replication competent (pathogenic) viruses (Bartosch et al., 2004; Young et al., 2012). For fixed ERVs, these kinds of disruption are likely relatively rare or of mild effect, and this is consistent with the observation that in general, the relative divergence of ERVs (as compared with selfish DNA) implies only small selective coefficients. Our observation that ERVs that are surrounded by more selfish DNA diverge more slowly than those surrounded by more non-selfish or non-repetitive DNA is consistent with the idea that the extra mutations we observe in ERVs may be mitigating the transcriptional disruption ERVs cause to nearby host sequence. So, some of the excess divergence we see in ERVs may be due to their remaining capability to recruit transcription machinery and produce transcripts.

There is another reason we might expect selection for substitutions into ERVs, and this relates to an ERV's repetitive nature, a property shared by all selfish DNA. As repetitive sequences, ERVs can increase the probability of harmful ectopic recombination (Hughes and Coffin, 2001; Campbell et al., 2014). The effects of such recombination can be catastrophic to the host, for example, infertility (Kamp et al., 2000; Sun

et al., 2000). Using population data, it has been concluded that negative selection acting against full-length polymorphic members of the human specific L1 Ta1 subfamily of LINEs is roughly 2×10^{-4} (Boissinot et al., 2006). This is one order of magnitude larger than the largest median selective coefficient we derive using the same effective population size. We do not expect fixed ERVs to cause as much harm via ectopic recombination as LINEs that are removed from the population before fixation, however, we do find that the relative divergence of (whole) ERVs increases with LTR length. Our finding might be due to longer LTRs acting as better promoters, however, it is also consistent with the hypothesis that longer LTRs are more likely to ectopically recombine. This is an idea supported by evidence that purifying selection against TEs in *Drosophila melanogaster* increases with element length (Petrov et al., 2011). The fact that we found no similar correlation between ERV length and divergence may reflect the fact that the probability of ectopic recombination increases with the number of possible pairings of near-identical elements present in an individual, and therefore roughly with the square of element number. As most ERVs are present only as solo-LTRs, and as each full-length ERV includes two LTRs, the probability of recombination between LTRs is expected to be very much greater than the probability of recombination between viral regions. Therefore, in short, there is both evidence and reason to believe that ectopic recombination may make some contribution to increasing the rate of divergence between orthologous ERVs.

In this study, we have made comparisons between ERVs (PV) and selfish DNA (RM^+). This seemed like a pragmatic way to obtain selection coefficients that characterized the differences between ERVs and sequence that is usually assumed to evolve neutrally. However, it should be noted that our assignment of sequence to one of three categories is crude and suggests that the differences we have reported between ERVs and their surrounding DNA may underestimate the selective forces acting upon ERVs. We have argued that ERVs diverge at faster than neutral rates because they sometimes have an effect on the host, even after fixation. Some of these effects are due to prop-

erties shared by most TEs, particularly the potential for ectopic recombination or the disruption of transcription. If ERVs diverge faster than other selfish DNA in part because of properties they share with other TEs, then some portion of TEs should also be expected to diverge at faster than neutral rates. These TEs are assigned to the RM^+ category and therefore we compare ERVs to sequence that is, on average, potentially also evolving at faster than neutral rates. For this reason we consider our selection coefficients conservative lower bounds.

The primary goal of this study was to determine whether, on aggregate, ERVs (PV) have had a measurable effect on their hosts. Under our assumptions this could have been seen in one of two ways. First, ERVs could have been conserved relative to neutral (RM^+) rates. Second, ERVs could have diverged more quickly than neutral rates. In fact, we observed the second possibility. It is interesting that this is the case but this is not the whole story. We can compare the divergence of ERVs (PV) and selfish DNA (RM^+) to non-repetitive or repetitive but non-selfish flank (RM^-). Doing so reveals that the distribution of RM^+:RM^- is shifted to the left of and more peaked than that of PV:RM^- (Figure S3 in Supplementary File 3.1). In other words, relative to non-repetitive or non-selfish DNA, some ERVs diverge more slowly than most other selfish DNA, even though the average ERV is a faster evolver. (The syncytins (Blond et al., 2000; Blaise et al., 2003) are not part of our dataset but are ERVs that would presumably exhibit such behaviour.) These issues have not been a focus of our study but warrant further investigation because if fixed ERVs have a different distribution of effects to other TEs then they probably have different kinds of effects too. In particular, they may be more often co-opted than other TEs.

Not all of the effects we observed were small. In particular, we observed that the median relative divergence of highly transcribed HERV-H implies a selection coefficient of 5.7×10^{-5}. This is closer to the selective force acting on a polymorphic LINE and is large enough to be of interest. This is particularly true as we know that highly transcribed HERV-H ERVs are functional components with respect to the reg-

ulation of stem cell identity (Wang et al., 2014). As we have shown that the relative divergence of HERV-H increases with their transcriptional activity we suggest that the excess substitutions we observe are tuning the transcription levels of these ERVs in stem cells. What is less clear is whether such tuning is associated with adapting pre-existing, necessary and stable host functions (Wittkopp and Kalay, 2012), or whether it is instead alleviating the cost of transcription as a side effect of the co-option of a subset of HERV-H (Young et al., 2013). For example, it may be that the HERV-Hs that we observe evolving quickly are doing so because they promote functional lncRNAs or chimeric transcripts at a level that needs to be adjusted. Such adjustment might have been necessary due to differences between the biological challenges faced by human, chimp and their common ancestor. On the other hand, it may be that the co-option of some functional HERV-H loci brought with it the unfortunate side effect of the transcription of some different and purely selfish HERV-H loci. These loci would not be at all useful to the host yet could, at an early stage of a host's lifecycle, introduce any of the previously discussed costs of ERVs. Selection on the host population would be expected to attenuate these costs over time. These two possibilities will in future need to be disentangled, but whatever the reality, we can see that actively transcribed HERV-H has been diverging particularly quickly at the sequence level since the chimp-human split and conclude that our selective coefficient provides a lower bound on the magnitude of the forces acting upon it.

3.5 Conclusions

Endogenous retroviruses (ERVs) have evolved faster than other selfish DNA in humans and chimp. The divergence of ERVs relative to neighbouring selfish DNA is positively correlated with the length of the long terminal repeat of an ERV and with the percentage of neighbouring DNA that is non-repetitive or non-selfish. Members of the HERV-H family evolve particularly fast and in a manner that correlates with their

level of transcription in human stem cells. Assuming faster evolution is due to directional selection, the typical substitution into an ERV is recessive and a substitution into a highly transcribed HERV-H has a selective coefficient of the order of 1×10^{-4}, which is not small. This suggests that the HERV-H transcriptome has recently evolved under the influence of directional selection. Further work is needed to discover whether HERV-H is the subject of adaptive regulatory change or whether co-opting some proportion of ERVs has opened up the genome to the harmful effects of other unwelcome retrovirally derived guests.

3.6 Methods

3.6.1 Detecting ERVs

A library of 771 viral pol genes were used as probes in a tBLASTn (Altschul et al., 1990) search against five soft-masked primate genomes: human (*Homo sapiens*), chimp (*Pan troglodytes*), gorilla (*Gorilla gorilla gorilla*), orangutan (*Pongo abelii*) and macaque (*Macaca mulatta*). The genomes were obtained from the Ensembl project (Flicek et al., 2012). The viral probes were selected to represent endogenous and exogenous retroviruses from a broad range of sources and are the same as those used in previous studies (Magiorkinis et al., 2012; Katzourakis et al., 2007b; Gemmell et al., 2013). The aim was to identify as many ERVs as possible and a summary of the diversity of probes is available in Tables S3 and S4 in Supplementary File 3.1. The 15kbp of sequence centred on each of the resulting collection of 19,945 putative pol hits was processed using the LTR detection and annotation software LTRharvest and LTRdigest (Steinbiss et al., 2009). The original genomic location of the 5′ start and 3′ finish of each LTR was recorded for those regions containing paired LTRs. Locations containing at least one retroviral gene (as detected by LTRdigest) beyond the pol identified by tBLASTn were assumed to contain full-length proviruses and were retained for further processing. Our goal was not to identify novel ERVs and confirmation that the location of our

ERVs overlap with another study, as well as details of the locations identified by our study, are contained in machine readable form in Additional file 2. Detecting orthology between proviruses

Orthologue detection proceeded in two stages. First, the 20kbp surrounding each putative full-length provirus (hereafter 20kbp excerpt) was used as a BLASTn query in a search against every other syntenic 20kbp excerpt from every primate species. Synteny mapping was based on chromosome name and therefore pairings could be made between ERVs on human chromosome 2 and ERVs on chimp chromosomes 2a or 2b. A local BLASTn alignment of length at least 7500 nucleotides and of at least 95% identity between two 20kbp excerpts was considered suitable to qualify pairs of 20kbp sequence as potentially orthologous. Second, the aforementioned candidate orthologies were investigated in detail by performing Needleman-Wunch pairwise global alignment using the stretcher program (gap-open penalty 16, gap-extend penalty 4 and matrix EDNAFULL) from the EMBOSS software suite (Rice et al., 2000). A sample of over fifty candidate orthologies, picked uniformly at random, were examined by hand. Upon inspection of these pairwise alignments it was determined that choosing a minimum global identity of 85% and minimum global similarity of 85% would sufficiently capture our intuition of orthology. That is to say, a lower threshold would run the risk of pairing non orthologous sequence but a higher threshold would unnecessary exclude genuinely orthologous provirus and flank from our study. Alignments of this kind (i.e. alignments indicating orthology) were noted. In the rare event that two or more 20kbp excerpts were orthologous within the same species (a potential paralogy) all homologous 20 kb excerpts across all species were excluded from further analyses. This resulted in the removal of 32 paralogous pairs.

3.6.2 Annotating aligned provirus and flanking DNA

Once orthology had been determined we switched to using 40kbp excerpts (this did not involve discarding any data). Orthologous 40kbp excerpts were pairwise aligned

with the stretcher program using the same settings as mentioned above. Each 40 kb alignment was annotated as follows. We classified each column of our alignment as one of PV, RM^+ or RM^-. Membership of PV was determined by taking the union of the two contiguous regions identified as an ERV due to running LTRharvest on each of the chimp and human sequences in an alignment. The outermost 25 bp of this union region was excluded from all analyses to take account of uncertainty over the ability of LTRharvest to sharply identify the precise endpoints of 5′ and 3′ LTRs. The remaining flanking columns of each alignment were then classified based on their RepeatMasker annotation. We obtained RepeatMasker annotations for all of our 40kbp excerpts by submitting them to repeatmasker.org using settings `cross_match` and `speed/sensitivity slow`. The category RM^+ contained sequence classified as `DNA`, `LINE`, `Low_complexity`, `LTR`, `RC`, `Retroposon`, `Satellite`, `Simple_repeat`, `SINE` or `Unknown`; the category RM^- contained unmasked sequence or sequence classified as `RNA`, `rRNA`, `scRNA`, `snRNA`, `srpRNA` or `tRNA`. All dinucleotide pairs in an alignment were annotated as CpG sites if they were zero or one mutation away from CG:CG or GC:GC, i.e. exhibited a potentially mutated cytosine or guanine, or if they were of the form TG:CA or CA:TG, i.e. exhibited a potential common double transition at both cytosine and guanine.

3.6.3 Alignment quality

When performing distance calculations we were concerned with ensuring that, as far as possible, differences between sequences did not result from regions of bad alignment. To mitigate this possibility we excluded gapped and low complexity regions from our final analysis using a program (available on request) that implemented the following heuristic method. Alignments were broken into blocks separated by gaps or low complexity regions of eight or more consecutive columns in length. Low complexity sequence was defined as that masked by the dustmasker program of the BLAST suite (Altschul et al., 1990). The edges of blocks of ungapped and unmasked sequence

were examined six nucleotides at a time. If these six nucleotide regions contained any mismatched bases the appropriate block had the six nucleotide region removed. This process was repeated until blocks started and finished with regions containing six identical nucleotides or were removed entirely. Only blocks of at least 20 nucleotides in length were used in our analyses.

3.6.4 Calculating distances

All distances were calculated using PAML 4.8 (Yang, 2007). For per-alignment comparisons the K80 method was used. For aggregate comparisons both the K80 and the GTR model were applied, though we found the two methods produced identical distances beyond the precision reported in our study. The overall number of patterns used to calculate distances appear in Table 3.3 and Table S1 in Supplementary File 3.1.

3.6.5 Calculating selection coefficients

Assuming substitutions into RM^+ are neutral then a measure of the rate of substitution in the RM^+ flank is also a measure of the neutral mutation rate μ. We write the elevated substitution rate into ERV DNA that we obtain from measures of divergence of PV as γ. It is well known that the ratio $\lambda = \frac{\gamma}{\mu}$ is directly related to the selection coefficient s acting on substitutions. Therefore, under the assumption of weak selection, a Wright-Fisher model of drift and semi-dominant mutations ($h = \frac{1}{2}$) we have: $\lambda = 2N(1 - \exp(-s))/(1 - \exp(-2Ns))$. As the diffusion equation from which the previous equation is derived assumes a small s, it is common and numerically convenient to use the approximation $\lambda = 2Ns/(1 - \exp(-2Ns))$ (Charlesworth and Charlesworth, 2010). We take effective population size $N = N_e$ to be 10,000 in our calculations (Boissinot et al., 2006).

3.6.6 Calculating dominance

By calculating the relative divergence of autosomal and X-linked ERVs it is possible to make statements about dominance (Vicoso and Charlesworth, 2006). Denote the rate of substitution of mutations on the autosome as $K_A = 2N\nu_A\mu_A$, where $2N$ is the number of copies of the autosome in a population, ν_A is the probability of fixation of a beneficial mutation, and μ_A is the mutation rate. For the X chromosome the analogous expression is $K_X = \frac{3}{2}N\nu_X\mu_X$, where we allow substitutions on the X chromosome to derive from a process with its own mutation rate and probability of fixation.

Alignments of orthologous sequence provide chimp-human divergence values K_At (autosomal PV), μ_At (autosomal RM^+), K_Xt (X-linked PV) and μ_Xt (X-linked RM^+), where t is the evolutionary time for which chimp and human have been separated. Let ratios of divergence be denoted by A and X so that $A = K_A/\mu_A$ and $X = K_X/\mu_X$. Using aggregated data we find that $X > A$ (see Results section).

Assuming weak directional selection, and the population genetic framework in Table 3.5, which allows separate selective coefficients s_m in males and s_f in females, the probabilities of fixation ν_A and ν_X are well approximated by $\frac{1}{2}h(s_f + s_m)$ and $\frac{1}{3}(2hs_f + s_m)$ respectively (Charlesworth et al., 1987). These weak selection approximations allow one to make statements about dominance and sexually antagonistic selection. Based on our divergence data we are interested in cases when $2h(s_f + s_m) < 2hs_f + s_m$. For positive s_m, this occurs when (dominance) $h < \frac{1}{2}$.

			male			**female**		
linkage	**A**	**genotype**	A_1A_1	A_1A_2	A_2A_2	A_1A_1	A_1A_2	A_2A_2
		fitness	1	$1 + hs_m$	$1 + s_m$	1	$1 + hs_f$	$1 + s_f$
	X	**genotype**	A_1		A_2	A_1A_1	A_1A_2	A_2A_2
		fitness	1		$1 + s_m$	1	$1 + hs_f$	$1 + s_f$

Table 3.5: The model of fitness effects of mutations into ERVs (PV) used in this study.

3.6.7 Transcription data

We paired genomic coordinates located in the supplementary material of (Wang et al., 2014) with the genomic coordinates of our 40kbp excerpts from human. Each HERV-H locus in (Wang et al., 2014) was paired with its nearest syntenic 40kbp excerpt from human if the distance between the centroids of the two sets of coordinates (theirs and ours) was less than 2500 bp. This resulted in the association of 63 of the 64 of the previously identified HERV-H ERVs in our dataset with 63 sets of transcription data. No association between transcription data and ERVs from any other family was made. The nominal transcription levels "highly active", "moderately active" and "inactive" are the same as those referred to in the main text and figures of (Wang et al., 2014) and were read directly from the supplementary data. The continuous levels we discuss were obtained by taking the mean of the expression levels across all stem cell measurements in the supplementary data (Wang et al., 2014).

3.6.8 Long distance analysis

To examine the divergence of regions greater than 40 kb in length we searched the six-way EPO multiple alignments available from the Ensembl project for regions that contained the coordinates of the 12 "highly active" HERV-H orthologues in our study. Alignments for six of the 12 orthologues could be identified. We removed sequence that was gapped in both chimp and human. We then annotated the chimp and human sequence in the same way as our 40 kb alignments (described above). For each of the six alignments we computed the divergence of sites classified as RM^{+} or PV using a sliding window. For any particular alignment we used a natural window size of the same length as the HERV-H region the alignment contained.

Chapter 4

A phylogenetic maximum-likelihood analysis of endogenous retroviral insertion and deletion in primates

4.1 Abstract

Background

As much as 8% of the human genome is thought to be made up of endogenous retroviruses (ERVs). ERVs are commonly found in two forms, the proviral form and the more numerous solo-LTR form that is thought to be the result of homologous recombination. We introduce a phylogenetic framework to study ERV insertion and solo-LTR formation. We then apply the framework to ERV site patterns generated by applying a heuristic to a set of long alignments covering six primate genomes.

Results

We study six categories of ERVs and quantitatively recapitulate patterns of ERV insertional activity that are often described in qualitative terms in the literature. We observe

a slowdown in most ERV groups but suggest that HERV-K activity may have increased in the last 8 Myr. We find that the rate of solo-LTR formation decreases rapidly as a function of ERV age. We show that an age-dependent model of solo-LTR formation describes the history of ERVs more accurately than the commonly used exponential decay model. We find HERV-H loci are markedly less likely to form solo-LTRs than ERVs from other families.

Conclusions

A previous finding that solo-LTR formation occurs rapidly can be phylogenetically formalized and generalizes to the majority of ERVs in primate genomes. The use of an exponential decay process to describe solo-LTR formation should be abandoned in most cases. ERVs that are not solo-LTRs by an age of 5 Ma are unlikely to change state as they age. Our findings are compatible with the hypothesis that solo-LTR formation is prevented by mutational divergence. We suggest the lower probability of solo-LTR formation for HERV-H loci supports a long-term host role for many elements from this family.

4.2 Introduction

By definition, endogenous retroviruses (ERVs) are the result of a Mendelian (vertical germ line) transmission of retroviruses from parent to progeny. Over many generations it is possible for an ERV to fix in a host population so that in humans, for example, as much as 8% of the genome is thought to be retrovirally derived (Paces et al., 2002). Studies have identified ERV activity dating back over millions of years and in many species (Katzourakis, 2013). This activity is often classified by dividing ERVs into particular groups or families. Some of these families, such as HERV-K, have been shown to have been replicationally active until very recently (Subramanian et al., 2011). Because demographic stochasticity and natural selection may act to keep novel

ERV insertions at low frequency, it is possible that HERV-K may still be replicating in humans today (Belshaw et al., 2005a; Marchi et al., 2014).

Successful retroviral insertions or proviruses are known to possess a common structure consisting of viral genes flanked by direct repeats known as long terminal repeats (LTRs). ERVs that retain this characteristic viral structure are commonly described as full-length ERVs. In addition to full-length ERVs, endogenized viruses are also found in a second, dramatically different form, referred to as a solo-LTR. A solo-LTR is a solitary LTR that is missing its associated partner LTR and adjacent proviral genes. Solo-LTRs are thought to be generated when paired LTRs undergo non-allelic homologous recombination which results in a deletion and an associated acentric fragment (Stankiewicz and Lupski, 2002), a piece of chromosomal material lacking a centromere that is unlikely to persist across many cell divisions. Clearly, like other genomic DNA, both forms of ERVs are also subject to ordinary mutational processes so that over time they may become degraded or fragmented due to point mutations or indel events.

Although most ERVs are thought to be functionally inert, several high-profile studies have identified important biological roles for loci originating from several families. As an example, ERVs from the apparently replicationally inactive family HERV-H have been shown to be essential to the definition of a naive-like stem cell state in humans (Lu et al., 2014; Wang et al., 2014). Other very recent work on human specific HERV-K insertions shows that some loci from this family have phenotypic effects including the modulation of cellar mRNAs and the production of viral like particles (Grow et al., 2015). Across the mammalian clade the co-option of ERVs appears to be an important and recurring phenomenon in placental evolution (Lavialle et al., 2013). These headline discoveries rightly attract attention, but the repetitive and selfish nature of the typical ERV loci also suggests that many will have had (potentially deleterious) consequences for their hosts at some point in the past. These consequences include participation in ectopic recombination events, the generation of insertions into

functional regions of the host genome, and some sort of transcriptional cost (Gonzalez and Petrov, 2012).

When attempting to describe the replicational activity of viruses on evolutionary timescales many studies start by searching for full-length ERVs in a host genome. The resulting ERVs are then dated on an individual basis by comparing the divergence of paired LTRs that are assumed to have been identical at integration time. When it is important to demonstrate a minimum age for a particular locus then orthologous sites in additional genomes are examined until the identification of a pre-integration site is made. The pre-integration site can then be used to establish an upper bound on an insertion time with respect to host speciation events. Examples of recent studies that use LTR dating include work by Chong et al. (2014) and by Mata et al. (2015). The aforementioned strategy is certainly reasonable but does have at least three limitations.

First, it is known that ectopic recombination or gene conversion between repetitive sequences is relatively common and can confound inferences of evolutionary distance. For example, Hughes and Coffin (2005) suggest that one-third of a set of 15 HERV-K elements were involved in gene conversion or ectopic recombination events. If we cannot trust inferences of evolutionary distance then we have a degree of uncertainty over the activity of families of ERVs over time.

Second, many ERVs may be too fragmented to contribute to studies that use LTR dating to establish the replicational activity of a family. For example, an analysis of the supplementary data from (Wang et al., 2014) suggests that 16% (158 of 1057) of HERV-H proviruses in the human genome have zero or only one associated LTR. The degree of fragmentation of an ERV may be age-dependent or have had a bearing on the fitness of a host. If this is true then ignoring fragmented ERV insertions may systematically bias our reconstructions of their evolutionary activity.

Third, most ERVs are present in solo-LTR form (Bannert and Kurth, 2006) and therefore do not contribute to descriptions of replicational activity at all. An investigation into polymorphic HERV-K (HML-2) loci in human finds that the majority of

loci are represented by pre-integration sites or solo-LTRs (Belshaw et al., 2005a). This in turn suggests that the solo-LTR formation process is very rapid. If this is true for most ERV families for most of the time then it is quite reasonable to treat a full-length ERV as a proxy for a constant proportion (perhaps one-sixth to one-tenth) of all ERVs. However, if the result does not generalize then such an assumption is not reasonable as the proportion of ERVs that are present in full-length versus solo-LTR form will vary given the age of an infection.

In this study we investigate the effect of considering solo-LTRs when describing ERV activity. Like other authors, we bring together insertion rates from several viral groups that are present in primates. Unlike other authors we do this by systematically sampling both full-length ERVs and solo-LTRs—in the same way for different ERV families and different host species—and relating these ERVs explicitly via a host genome alignment. By combining our sampling process with a host phylogeny we can then place insertion rates in quantitative comparison. In this study we are particularly interested in the ratio of full-length ERVs to solo-LTRs. This question has been tackled before (Belshaw et al., 2005a) but we extend prior work in several ways. This is achieved by using a likelihood framework that allows model comparison, by examining a wider variety of ERV families, and by considering ERVs over a larger host phylogeny. Using this approach we find that HERV-H has unusual dynamics and that previously used models of deletion should probably be abandoned where possible. Based on our findings we argue that a phylogenetic approach to characterizing ERV activity is a sensible approach for future studies, particularly as researchers are able to leverage better annotation tools, additional genomes, and more complete genome alignments.

4.3 Materials and methods

In overview, our method was to collect a sample of ERV site patterns from a variety of primates, and to use those patterns, in combination with a host phylogeny, to find the maximum likelihood parameter values for two variations of insertion and deletion processes. This allowed us to decide which model process fit our data the best. Below we break our methods section into four parts. First we describe the process of sampling site patterns. Second we describe our phylogenetic insertion and deletion models. Third we describe maximum likelihood estimation. Fourth we describe a simulation we refer to in our discussion section.

4.3.1 Alignment and repeat annotation

We obtained the six-way Enredo-Pecan-Ortheus (EPO) whole genome multiple alignment of primate species that forms part of Ensembl Release 71 (Flicek et al., 2012). The six species included in the alignment are: human (*Homo sapiens*), chimp (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo abelii*), macaque (*Macaca mulatta*) and marmoset (*Callithrix jacchus*).

The EPO alignment contains 7,224 individual alignments containing exactly one sequence for each of the six species. (These alignments comprise approximately 32% of the alignments present in the EPO dataset, which also covers duplicate regions and regions that are present or alignable in only a subset of the six primates.) The 7,224 individual alignments were usually of the order of 10^5 to 10^6 columns in length and had a median LTR content of 8%. To identify LTR retroelements we ran RepeatMasker 3.3 (Smit et al., 2004) on each ungapped sequence from each individual alignment using the `-species mammal -no_is -pa 4 -q -nolow -norna` options.

The annotations made by RepeatMasker are often fragmented such that an LTR element originating from a single insertion event is referenced using several distinct annotations. For this reason we applied REannotate 26.11.2007 (Pereira, 2008) using

options `-c -n -f` to our RepeatMasker results. This resulted in the identification of complete ERVs, truncated ERVs, and solo-LTRs, entities corresponding to distinct insertional events as defined by Pereira (2008). The application of REannotate also conveniently mapped synonymous RepeatMasker identifiers to appropriate canonical identifier e.g. identifiers `HERVH`, `LTR7`, `LTR7Y`, `RTVL-H`, `RTVL-H2` and `RGH` were all mapped to identifier `HERV-H`.

The result of the above repeat masking and annotation processes was data giving the location, repeat type and structural status of LTR elements in ungapped coordinates. These ungapped coordinates could be mapped back to the appropriate locations in the original EPO multiple alignment files.

To check that REannotate had assigned the correct identifiers to ERVs we performed a BLAST (Altschul et al., 1990) alignment of a representative sequence underlying each repeat locus against a library of 51 viral sequences drawn from the Retrovirus reference sequence library version 2013-12-26, located at http://tinyurl.com/odk2ukh. For any given alignment locus, a full-length representative sequence was preferred to a solo-LTR where available. Representative sequences were drawn from human or the closest primate to human in preference to those that were more distant. For each repeat assigned to families ERV9, HERVK11, HERV-H and HERV-K, we also submitted the sequences to Dfam (Wheeler et al., 2013), and as BLAST queries to NCBI located at http://blast.ncbi.nlm.nih.gov, to examine their structure in detail. This resulted in the removal of 8 SVAs that would have otherwise been erroneously included in our study.

4.3.2 Construction of site patterns

To obtain site patterns from alignments we needed to first designate regions of the alignments as either full-length ERVs, solo-LTRs, or pre-integration sites on a per-species basis. To achieve this we started by adopting a heuristic that broke each alignment into regions of contiguous columns that contained identical features in each row. We retained regions of at least 50 bp in length in our analysis. (For purposes of effi-

ciency and simplicity we ignored overlapping insertions and restricted our analysis to non-nested ERVs.) The intuition behind our approach is that, moving along an alignment, we are interested in gross changes of state in the relationship between each of the sequences comprising an alignment—so that we can detect insertions and deletions—but that we are not interested in the fine details. For example, when processing an alignment we are interested in the appearance of a 5,000 bp gap in a marmoset sequence aligned to regions annotated as full-length HERV-H in the other primates but we are not at all interested in the fact that a 12 bp gap might exist this ERV at position 434 in the chimp. This idea is described more formally below.

Consider a six-way alignment of length m. We form a corresponding 6 by m classification matrix $A = \{a_{i,j}\}$ in order to combine the information output from the REannotate program with information contained in the alignment. Each element $a_{i,j}$ is conceptually of one of the following kinds: an unannotated nucleotide coded as `d`; an unannotated gap coded as `g`; the ith solo-LTR having identifier id coded as `s`$-i-id$; or the ith partial or full-length ERV having identifier id and coded as `c`$-i-id$. For example, positions annotated as belonging to the fourth full-length HERV-H in an alignment would be given the classification `c-HERV-H-4`.

We partition a classification matrix A into the minimal number of ℓ adjacent sub-matrices $A^{(1)}, \ldots, A^{(\ell)}$ having 6 rows and $m_1, \ldots, m_\ell$ columns such that $A^{(i)}_{j,k_1} = A^{(i)}_{j,k_2}$ for j in $1 \ldots 6$ and k_1, k_2 in $1, \ldots, m_\ell$. This is to say, our partition of A ensures all columns in submatrix $A^{(i)}$ are identical and that two consecutive submatricies differ. Clearly, the first column of any submatrix also characterises the whole submatrix. We ignore sub-matrices of less than 50 columns in length and refer to the sequence of the first columns of the remaining submatrices as the sequence of n pre-patterns $P = (A^{(i)}_{(1\cdots6,1)} : \text{ncols}(A^{(i)}) \geq 50) = P^{(1)}, P^{(2)}, \ldots, P^{(n)}$. We form these pre-patterns for every six-way alignment in the EPO dataset.

Our ultimate goal was to place patterns describing the status of ERVs located at orthologous positions in their host genomes at the tips of a phylogenetic tree. To

achieve this objective the sequences of pre-patterns described above were parsed into bona-fide site patterns characterising the state of orthologous ERV loci in each species. The first stage of this parsing process utilized a modified version of Thompson's Construction Algorithm (Thompson, 1968) and had the effect of identifying all contiguous subsequences of pre-patterns of length 3 or more having the following properties: (i) anchored at both ends by a pre-pattern of all-`d`s; (ii) containing a common solo-LTR or full-length ERV in the same row of every pre-pattern between the aforementioned anchors; and (iii) containing an entry coded as a gap in the same row of every pre-pattern between the aforementioned anchors. More precisely, using $P_q^{(i)}$ to denote the qth element of the ith vector in P, we identify the set S containing all subsequences of pre-patterns $P^{(i)}, P^{(i+1)}, \ldots, P^{(j)}$ of length $j - i + 1 \geq 3$ such that: (i) $P^{(i)} = P^{(j)} = (\mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d})^\top$; (ii) there exists q such that $P_q^{(k)} = P_q^{(k+1)} = \mathtt{s}\text{–}x\text{–}y$ or $P_q^{(k)} = P_q^{(k+1)} = \mathtt{c}\text{–}x\text{–}y$ is satisfied for all $k : i < k < j - 1$; and (iii) there exists q such that $P_q^{(k)}$ is gapped for all $k : i < k < j$. These conditions are sufficient to identify contiguous subsequences that can be interpreted as site patterns.

The set S of subsequences of pre-patterns that we have identified so far are unambiguously gapped in one species and unambiguously contain an ERV in one species. One example of such a subsequence is the following:

$$S^{(i)} = (\mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d})^\top, (\mathtt{s}, \mathtt{s}, \mathtt{s}, \mathtt{c}, \mathtt{g}, \mathtt{g})^\top, (\mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d}, \mathtt{d})^\top.$$

As we always write our pre-patterns in phylogenetic order, one can see that this example pattern describes an ERV that is missing in macaque and marmoset (the 5th and 6th elements), is present as a solo-LTR in human, chimp, and gorilla (the 1st – 3rd positions), and is present as a partial or full-length ERV in orangutan (the 4th position). In fact, we are interested coding patterns that are gapped in marmoset (our outgroup) and present in at least one other species. We code our site patterns using the following notation: an absent sequence is coded as `0`; a solo-LTR is coded as `x` (the letter `x` being

reminiscent of both deletion and recombination); and an ERV present in either partial or full-length form is coded as a `1`. It is clear that in our example we would want to assign the site pattern $(\mathtt{x}, \mathtt{x}, \mathtt{x}, \mathtt{1}, \mathtt{0}, \mathtt{0})^\top$ to our subsequence. Indeed, it is clear that for any subsequence in S, we can reasonably assign a `1` or an `x` as appropriate to the row fulfilling condition (ii) and also assign an `g` to the row fulfilling condition (iii). Nevertheless, in general, we have constructed S requiring unambiguity in only two of six positions. To assign the correct coding to the remaining four positions we adopted the following heuristic to subsequences in S.

Consider subsequence of pre patterns $S^{(i)} = P^{(j)}, \ldots, P^{(k)}$. We wish to form site pattern $U^{(i)} = (u_1, \ldots, u_6)$. We know by construction that $u_6 = \mathtt{g}$ (absence in marmoset) and that $u_q = \mathtt{1}$ or $u_q = \mathtt{x}$ for some q (evidence of ERV insertion in some species). We wish to assign $u_{q'}$ for $q' \neq q$ in $1 \ldots 6$. To do this we apply the following procedure to each subsequence in S:

1. if $P^{(k)}_{q'} = \mathtt{s}$ for some $i < k < j$ then we set $u_{q'} \leftarrow \mathtt{x}$;
2. if $P^{(k)}_{q'} = \mathtt{c}$ for some $i < k < j$ then we set $u_{q'} \leftarrow \mathtt{1}$;
3. if $P^{(k)}_{q'} = \mathtt{d}$ for every $i < k < j$ then we set $u_{q'} \leftarrow \mathtt{1}$ if we have already assigned a `1` in the previous steps, otherwise we set $u_{q'} \leftarrow \mathtt{x}$.
4. if $P^{(k)}_{q'} = \mathtt{g}$ for every $i < k < j$ then we set $u_{q'} \leftarrow \mathtt{0}$;

The above methods convert pre-patterns to site patterns by combining homology and processed RepeatMasker annotations but do not take any account of the length of sequences. Each pattern assigns an integration state of `0` (absence), `1` (presence) or `x` (presence in solo-LTR form) to orthologous ERV loci in six species. In a final post-processing phase we make use of the underlying alignment to remove any patterns in which (i) the 5′ flank $P^{(i)}$ of a pattern is backed by less than 100 nucleotides, (ii) the 3′ flank $P^{(j)}$ of a pattern is backed by less than 100 nucleotides, or (iii) the inter-flank region $P^{(i+1)}, \ldots, P^{(j-1)}$ of a pattern is backed by less than 250 nucleotides. In

addition, where patterns are mixed, that is patterns contain both 1s and xs, we apply a procedure that converts any particular x to a 1 if 85% of the called sequence backing the interflanking region of position annotated as x is identical to called sequence in any of the corresponding aligned sequence backing a position annotated as 1. The rationale for this action is that our alignment allows us to apply reasoning by homology that is not available to RepeatMasker or REannotate which only operate on independent sequences. Every such mixed pattern was examined by hand to ensure that such reasoning was appropriate. We then use these patterns as an input to our phylogenetic model described in the following section.

4.3.3 Phylogenetic insertion and deletion model

Consider the primate phylogeny T reproduced in Figure 4.1 where the branch length immediately below node i is denoted T_i and takes the value given by (Steiper and Young, 2006). We wish to relate subsets of our sampled site patterns U (described above), for example, the subset of patterns relating to HERV-H, to an insertion process on the tree T as well as to one of two potential deletion processes, also on T, between which we wish to discriminate. The insertion process is assumed to be Poisson and the deletion process is assumed to be either Weibull or exponential. The exponential deletion model is a nested submodel of the Weibull deletion model. We first discuss insertion and then discuss deletion.

A site in a genome can be assigned one of three states at any particular time: absent or 0 (a pre-integration site); insertion or 1 (contains an ERV); and deleted or x (contains a solo-LTR). Over time, a site may transition from the absent state to the insertion state to the deleted state. That is, the following state transitions can occur: $0 \rightarrow 1 \rightarrow x$. Five of the branches of T are external and their tips relate to observations U. Because we use marmoset as an outgroup, in our patterns u_6 is always equal to 0 and we will ignore it from now on. Given the permitted state transitions, it is clear the first five positions of a pattern unambiguously identify the branch of T on which an

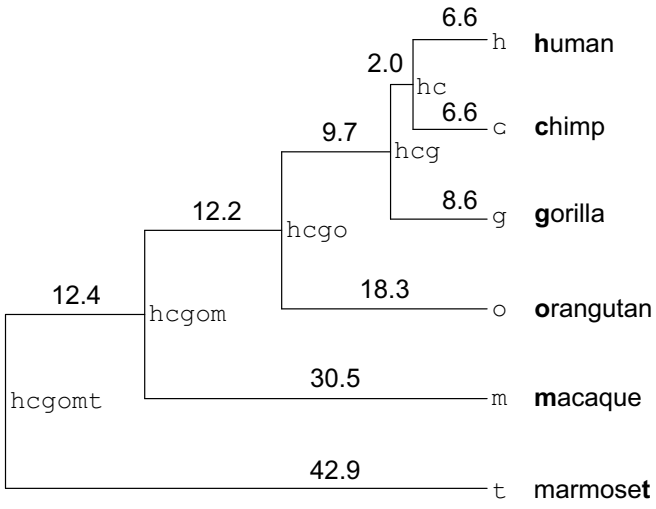

Figure 4.1: The primate host phylogeny used in this study. Nodes are labelled (fixed-width typeface) and branch lengths are given in millions of years (Myr) in accordance with Steiper and Young (2006).

ERV first appeared. Beyond this the patterns provide no further temporal information. We consider the insertion process on any given branch i of T to be a Poisson process that is fully described by rate parameter Φ_i. If, for a dataset under analysis, there are N_i insertions on branch i, then writing $\phi_i = \Phi_i T_i$, the likelihood of all insertions on T is given by

$$\prod_{i=1}^{9} \frac{e^{-\phi_i} \phi_i^{N_i}}{N_i!}. \tag{4.1}$$

We now discuss deletion. After entering the insertion state `1`, a site may transition to the deleted state `x`. A site pattern provides an observation of the final state of a site in any particular linage but it will not necessarily describe the state of a site at internal nodes. For example, the site pattern $(\texttt{x}, \texttt{0}, \texttt{0}, \texttt{0}, \texttt{0})^\top$ implies that the state at the corresponding site was `0` at all internal nodes. On the other hand, the pattern $(\texttt{x}, \texttt{x}, \texttt{0}, \texttt{0}, \texttt{0})^\top$ implies that the state at the corresponding site was `0` at all internal nodes except for `hc`, at which a state of `1` or `x` are both consistent with the evidence: in the former case a deletion (transition from `1` to `x`) occurred independently on the branches `h` and `c`, while in the latter case a single deletion occurred on the branch `hc` prior to the human-chimp split. We use a deletion process to consider the probability

of transition from 1 to x under all possible assignments of states to internal nodes.

To describe our deletion model we need to consider two kinds of branch, the insertion branch and the post-insertion branch. Figure 4.2 provides an abstract visualization of this concept. For any particular pattern, the insertion branch is uniquely identified as all nodes beneath the insertion branch have state 0 and all nodes above the insertion branch have state 1 or state x. The corresponding post-insertion branches are all those branches above the insertion branch in the tree.

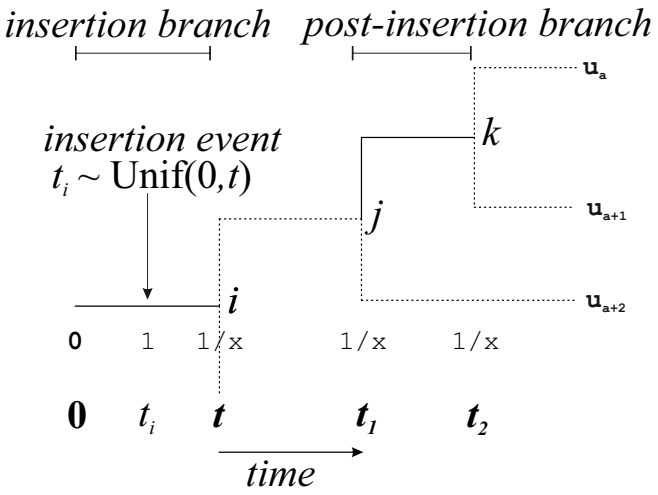

Figure 4.2: The phylogenetic abstraction on which we describe insertion and deletion processes. For a given site pattern an insertion branch is uniquely determined. Where applicable, we consider all possible state transitions on all post-insertion branches up to and including the external nodes of the tree.

We will first consider the deletion process under an exponential model whereby deletion occurs with a constant probability over time. The exponential decay process is common to all branches and parameterized by rate parameter ψ. Consider an insertion at time t_i on branch leading to node j. Let $t_0 = 0$ be the time at the origin of the brach leading to node j and time $t = T_j$ be the termination of the branch. Under an exponential deletion model the probability of deletion given an insertion at t_i is simply $\Pr(0 \to x|t_i) = 1 - e^{-\psi(t-t_i)}$. As insertion follows a Poisson process the time of insertion t_i of any particular ERV on the branch will be uniformly distributed between t_0 and t. Therefore, taking advantage of the fact that $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_x(x)\,\mathrm{d}x$

for random variable X with density function f_x and $g : \mathbb{R} \to \mathbb{R}$, the probability of changing from state `0` to state `x` during interval 0 to t is

$$\Pr(\texttt{0} \to \texttt{x}) = \int_0^t \frac{1}{t}\left[1 - e^{-\psi(t-t_i)}\right] \mathrm{d}t_i = 1 + \frac{e^{-\psi t} - 1}{\psi t}, \tag{4.2}$$

while the probability that an ERV is not deleted on the insertion branch is $\Pr(\texttt{0} \to \texttt{1}) = 1 - \Pr(\texttt{0} \to \texttt{x})$.

On a post-insertion branch the situation is much simpler as the memoryless property of an exponential decay process ensures that we do not need to consider the time during which a site has been in a particular state when calculating whether it will change state over any subsequent time period. Consider the post-insertion branch from node j to node k of length $t_2 - t_1$. If the site is in state `1` at node j then the probability of no state change is $\Pr(\texttt{1} \to \texttt{1}) = e^{-\psi(t_2-t_1)}$ and the probability of a deletion is $\Pr(\texttt{1} \to \texttt{x}) = 1 - e^{-\psi(t_2-t_1)}$. As the state `x` is absorbing $\Pr(\texttt{x} \to \texttt{x}) = 1$ while all other transitions on post-insertion branches have zero probability.

We now consider a Weibull process, under which the probability of a deletion occurring during any small time interval may increase or decrease given the age of an insertion. The Weibull process is again common to all branches but requires two parameters, a rate parameter ψ and a shape parameter ω. For an insertion branch, the appropriate probability of seeing two state transitions is

$$\Pr(\texttt{0} \to \texttt{x}) = \int_0^t \frac{1}{t}\left[1 - e^{-\left[\frac{t-t_i}{\psi}\right]^{\omega}}\right] \mathrm{d}t_i = 1 + \frac{\psi\left[\Gamma\left(\frac{1}{\omega}, \left(\frac{t}{\psi}\right)^{\omega}\right) - \Gamma\left(\frac{1}{\omega}\right)\right]}{\omega t}, \tag{4.3}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t$ and $\Gamma(\alpha, z) = \int_z^\infty t^{\alpha-1}e^{-t}\,\mathrm{d}t$. The probability than an ERV is not deleted on the insertion branch remains the only other possibility so that $\Pr(\texttt{0} \to \texttt{1}) = 1 - \Pr(\texttt{0} \to \texttt{x})$.

The essential feature of a Weibull process is that it can describe deletion rates that vary given the age of an ERV. This means that under a Wiebull model the proba-

bility of a state change from 1 to x on the post-insertion branch from node j to node k is no longer independent of insertion time t_i. Consider again the possibility of a state change on a branch of length $t_2 - t_1$. Under a Weibull process the probability of deletion takes into account uncertainty over insertion time t_i giving

$$\Pr(1 \rightarrow 1|t_i) = \int_0^t \frac{1}{t} \left\{ \frac{e^{-\left[\frac{t_2 - t_i}{\psi}\right]^\omega}}{e^{-\left[\frac{t_1 - t_i}{\psi}\right]^\omega}} \right\} \mathrm{d}t_i \tag{4.4}$$

and $\Pr(1 \rightarrow x|t_i) = 1 - \Pr(1 \rightarrow 1|t_i)$. These values can be numerically evaluated for all branches and distances on our tree T. As before, $\Pr(x \rightarrow x) = 1$ while all other transitions on post-insertion branches are impossible.

We have now completely specified the state transition probabilities for individual branches under a strict exponential model and under a Weibull model. To compute the likelihood of a site pattern given a tree and a deletion model we use a dynamic program algorithm that is essentially Felsenstein's pruning algorithm (Felsenstein, 1973). In the case that we use a Weibull deletion model, the algorithm must be modified to keep track of the insertion branch when considering transitions on post-insertion branches.

4.3.4 Combining site patterns and the phylogenetic model for maximum likelihood estimation

The above description has described how to calculate the probability of insertions as well as the probability of all post-insertional state transitions that might occur. We have also described how to construct site patterns from primate genomes. Therefore we are ready to describe how to calculate the probability of a set of site patterns U given a tree T, an insertion model M_i and a deletion model M_d. For each of n site patterns $U^{(j)}$ we can identify the insertion branch for that pattern, and hence the subtree $T^{(j)}$ of T that includes only the insertion branch and its descendants i.e. any post-

insertion branches. To calculate the likelihood of the site patterns we compute:

$$\Pr(U|M_i, M_d) = \prod_{i=1}^{9} \frac{e^{-\phi_i}\phi_i^{N_i}}{N_i!} \prod_{j=1}^{n} \Pr(U^{(j)}|M_d, T^{(j)}), \tag{4.5}$$

where the first product gives the likelihood of the insertions and the second product uses the aforementioned dynamic programming method to sum over all possible post-insertional transitions.

The insertion model M_i has nine parameters, the 9 insertion rates in Φ. The deletion model has one parameter if it is strictly exponential (the rate parameter ψ_e) or two parameters (ψ_w and the additional shape parameter ω) in the case that it is Weibull. By repeatedly computing the likelihood of our site patterns we can numerically maximize the logarithm of $\Pr(U|M_i, M_d)$ using code written in the MATLAB language. In practice we performed simulated annealing using `simulannealbnd` (limited to 5 minutes per replicate during bootstrapping) followed by gradient descent using `fmincon`. As the insertion process is independent of the deletion process we were able to verify our maximum likelihood results using grid search and gradient descent from random starting points.

4.3.5 Simulating mutations into LTRs

We wished to address two questions via simulation. First, given the arrival of an ERV into a population, how long would it take for the first mutation to occur in its LTRs? Second, if we considered the arrival of an ERV into a population, would we expect to observe the fixation of an allele with zero, one, two, or three or more mutations in its LTRs? To answer these questions we performed forward simulation under a Wright-Fisher model with an effective population size of 10,000 individuals, assuming 2,000 bp of LTR per ERV. We assumed a Poisson mutation process with a rate of 10^{-8} events per site per generation. To address the first question we tracked the number of ERVs in a population and terminated the simulation when the first mutation event occurred.

Mutations could occur in any LTR in the population. To address the second question we tracked the number of alleles in the aforementioned four mutational classes, allowing alleles to progress from one mutation class to the next: zero to one, one to two, and so on. This second simulation was terminated only when alleles from one mutational class fixed in the population. As with the first simulation, mutations could occur in any LTR in the population.

4.4 Results

We obtained site patterns describing ERV integrations and deletions that had occurred in the primate lineage (Figure 4.1) since the split between macaque and marmoset roughly 40 million years ago (Ma). The site patterns were obtained by relating post-processed RepeatMasker annotations to a six-way genome scale alignment of human, chimp, gorilla, orangutan, macaque and marmoset sequence. These annotations were then converted to site patterns using a polynomial time heuristic method. Our intention was to quantitatively describe the insertion rate of ERVs across branches of the primate phylogeny and also to investigate what is thought to be a fairly predictable process of ERV deletion that converts full-length ERVs into solo-LTRs.

Applying the process sketched above, we identified 1,197 distinct insertion events that had occurred on the branch `hcgom` or later. These distinct insertions could be split naturally into groups based on the type of ERV they involved. We investigated the properties of insertions from the four largest families: ERV9 (245 insertions); HERVK11 (197 insertions); HERV-H (116 insertions); and HERV-K (59 insertions). Based on a BLAST search against a library of 51 viral sequences, many ERVs from smaller families were assigned to group-I (131 insertions) or group-II (112 insertions). These patterns of insertion and deletion may be useful to other researchers and are and are presented in machine readable format in Supplementary File 4.1. Below we present the following results: (i) the insertion rate parameters obtained; (ii) the

deletion rate parameters obtained under two different models of deletion; (iii) a comparison of the two competing deletion models; and (iv) the results of a simulation that tests the adequacy of the most appropriate deletion model.

We applied our phylogenetic method (see Materials and methods) to obtain the maximum likelihood relative insertion rates per million years (Myr) for the nine branches in our tree (Figure 4.3, Table 4.1). These estimates of insertion activity are independent of the deletion model used. Our results are broadly compatible with descriptions contained in commonly cited studies on ERV dynamics such as those by Sverdlov (2000) or by Bannert and Kurth (2006). For example, the HERV-K insertion rates for branches `h` (1.36 relative insertions per Myr) and `c` (0.61 relative insertions per Myr) capture the commonly reported fact that HERV-K has been recently active in both human and chimp and also that the activity in human specific ancestors appears to have been at least 50% greater than the activity in chimp specific ancestors. In general, more detailed comparison is difficult as individual studies vary considerably in methodology and reporting style, a state of affairs that partially motivated us to perform the analysis reported in this paper.

	h	c	g	o	m	hc	hcg	hcgo	hcgom
ERV9	0.00	0.00	0.12	2.02	2.56	0.00	1.75	6.64	2.50
HERVK11	0.00	0.00	0.00	0.38	2.95	0.00	0.41	5.66	2.18
HERV-H	0.00	0.00	0.00	0.87	1.77	0.50	0.93	1.89	1.05
HERV-K	1.36	0.61	0.23	0.33	0.56	0.00	0.52	0.66	0.65
group-I	0.00	0.15	1.05	0.05	2.16	0.00	0.21	1.15	3.06
group-II	0.00	0.00	0.00	0.16	1.34	0.00	0.41	2.13	3.06

Table 4.1: Branch and group specific maximum likelihood relative insertion rates in relative insertions per Myr. Node names are as per Figure 4.1.

As well as insertion, we are also interested in the process by which ERVs are deleted. The simplest conceivable model of a deletion process is one which has a constant hazard over time i.e. a process under which the probability of deletion in an infinitesimally small period of time is constant. The unique process with this property is an exponential decay process. Under the exponential model, the probability of

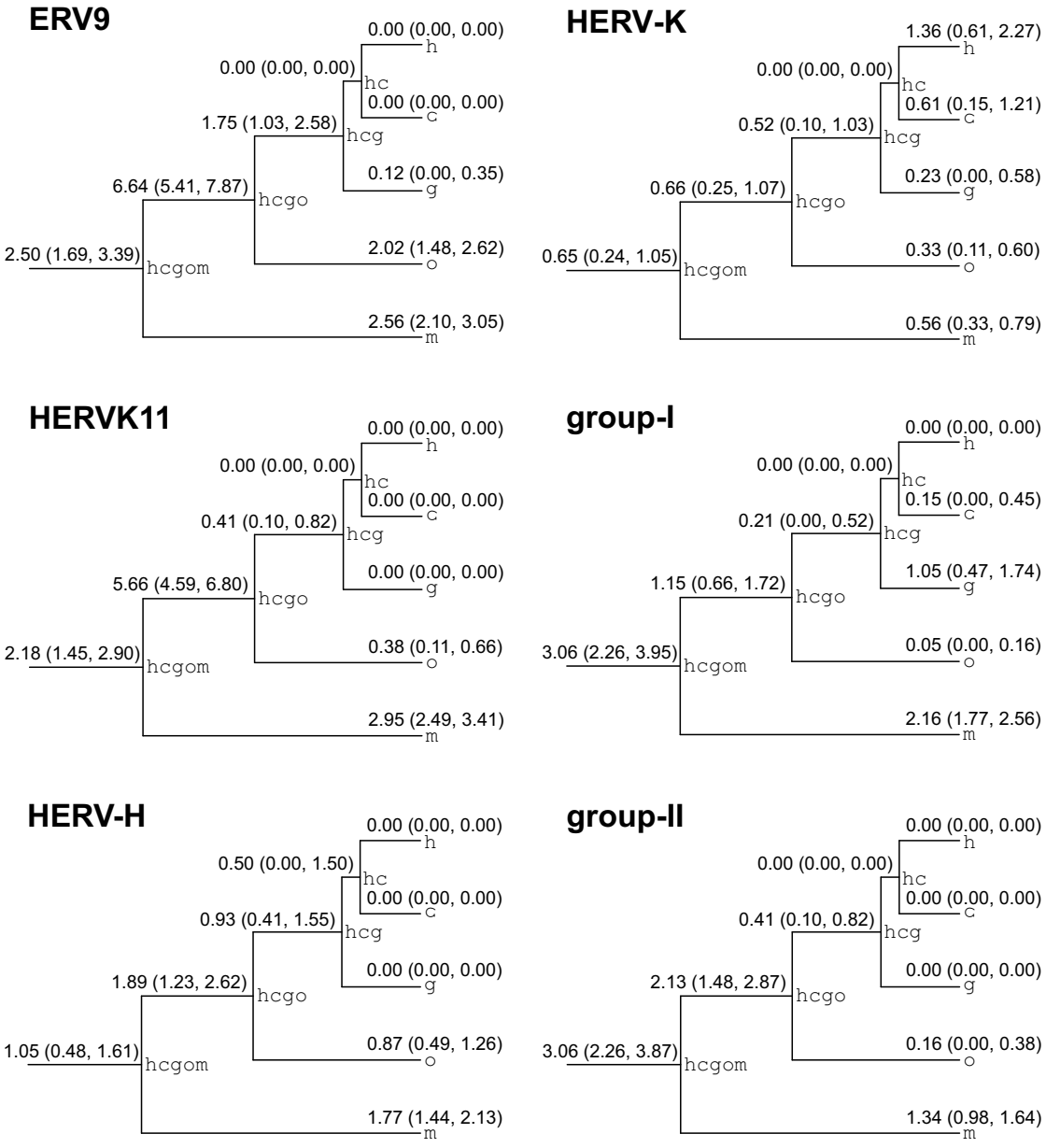

Figure 4.3: Branch and group specific maximum likelihood relative insertion rates in relative insertions per Myr. Bootstrap derived 95% confidence intervals are displayed in parentheses.

deletion of an ERV is independent of its age. Such a model is appropriate if the probability of deletion of an ERV is small and fairly constant across generations, and if the probability of deletion of an ERV has nothing to do with the process by which an ERV ages. For each of the six groups of ERV we obtained a maximum likelihood estimate of exponential rate parameter ψ_e (Table 4.2).

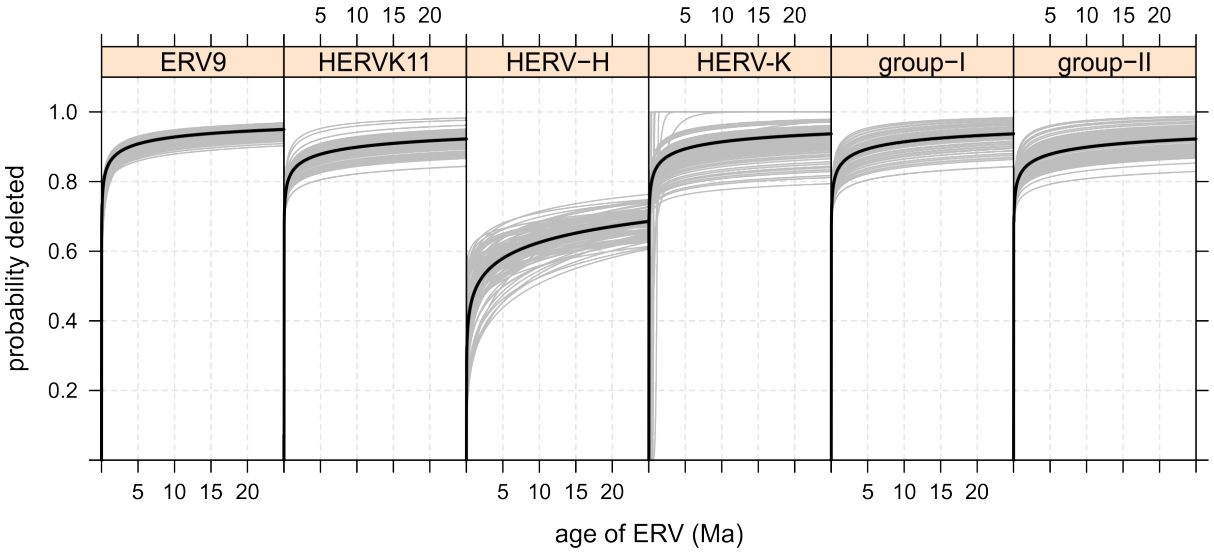

Figure 4.4: Group specific CDFs describing the probability of deletion of an ERV given its age. The CDFs are derived using maximum likelihood parameter estimates under a Weibull deletion process (thick black line) and from bootstrap resampled site patterns (thin grey lines).

Under the exponential model, full-length elements from the HERV-K family would be deleted most quickly, with full-length loci having an average pre-deletion lifetime of approximately 6.25 Myr. Under the same model, the most long-lived group would be HERV-H, for which the average full-length lifetime would be roughly 20 Myr. At an age of 400,000 yr (the expected fixation time of a neutral ERV given an effective population size of 10,000 and a generation time of 10 years), the exponential model predicts that 94–98% of ERVs would retain their full-length form. By an age of 25 Myr, a time period comparable to the scope of our phylogeny, the exponential model predicts that only 2% of HERV-K insertions would remain in full-length form while a much larger 29% of HERV-H insertions would.

Beyond the simplest possible deletion scenario, we are also interested in the

hypothesis that the formation of solo-LTRs is governed by a process that is not independent of the age of an insertion i.e. a process with a variable hazard. For example, it may be that as ERVs age, substitutions and gene conversion introduce differences between paired LTRs that substantially reduce their chance of pairing and producing solo-LTRs. A process with a hazard rate that changes with time is often modeled using a Weibull distribution. Under this process the rate of deletion is proportional to a power of time so that the probability of the removal of a full-length ERV can decrease with age, given a shape parameter $\omega < 1$, or increase with age, given a shape parameter $\omega > 1$. A shape parameter of $\omega = 1$ implies an exponential decay process so that the exponential model is a nested submodel of the Weibull model. For each of the six groups of ERV we obtained a maximum likelihood estimate of scale parameter ψ_w and shape parameter ω (Figure 4.4, Table 4.2).

Under a Weibull model, we find that at an early age it is again ERVs from the HERV-H family that are most likely to remain in full-length form. We find that an age of 400,000 yr 58% of HERV-H would remain in full-length form whereas only 19–21% of the other five groups would. These predictions differ by 40–76 percentage points from those of an exponential model. At the longer time scale, maximum likelihood parameter estimation suggests that at an age of 25 Myr we expect 31% of HERV-H to remain in full-length form while we expect less of ERV-9 (5%), HERVK11 (8%), HERV-K (6%), group-I (6%) or group-II (8%) to do so. Summarizing a Weibull deletion process requires considering the role of the shape parameter. The maximum likelihood estimate of shape parameter ω is less than 1 for all six groups of ERV, and therefore suggests the rate of ERV deletion does decrease monotonically with time. Bootstrap replicates suggest this is unambiguously true for all families apart from for HERV-K, for which 12% of bootstrap replicates identify a shape parameter >1. This implies that there is a degree of uncertainty over whether HERV-K has qualitatively different dynamics than the other ERV groups, with these exceptional bootstrap estimates suggesting peak deletion rates occur at ages of up to just over 4 Myr.

	ℓ_e	ℓ_w	ψ_e	${\psi_w}^{-1}$	ω	$2\Delta\ell$
ERV9	-158.54	-107.55	0.14	> 100.00	< 0.14	101.98
HERVK11	-144.83	-74.68	0.11	> 100.00	< 0.12	140.31
HERV-H	-140.88	-121.79	0.05	$= 0.09$	=0.18	38.18
HERV-K	-58.46	-32.33	0.16	> 100.00	< 0.13	52.27
group-I	-109.17	-47.54	0.10	> 100.00	< 0.13	123.26
group-II	-89.25	-50.62	0.10	> 100.00	< 0.12	77.27

Table 4.2: Group specific log likelihoods and corresponding maximum likelihood estimates of parameter values under exponential and Weibull deletion processes.

We have described the results of fitting two competing models, formalized as the null hypothesis H_e that the deletion process is an exponential decay, and as the alternative hypothesis H_w that the deletion process is age dependent. To decide which of the models is more appropriate we performed a likelihood ratio test. As H_e is nested within H_w, which has one additional parameter, we compare $2\Delta\ell = 2(\ell_w - \ell_e)$ with a χ^2 cutoff of 10.83. We find that $2\Delta\ell$ implies that the Weibull age specific model is clearly more appropriate for all six groups of ERV (p <0.001).

Our likelihood values show that a Weibull model is a much better description of the ERV deletion process than an exponential model. However, likelihood ratio tests do not provide an assessment of the adequacy of the Weibull model itself. For this reason we conducted a simulation to see whether the Weibull model could explain the empirical site patterns we observed for each of the six ERV groups. Our simulation proceeded as follows. For each of the six ERV groups, we generated 10,000 insertions on branch `hcgom`, the deepest branch in our phylogeny. We then simulated the history of these insertions according to the Weibull model operating under group specific maximum likelihood estimates of ψ_w and ω. This generated the group specific frequency distribution of site patterns at the tips of the tree. We performed a goodness of fit test comparing the distribution of empirical site patterns with the frequencies obtained via simulation. We found no statistical evidence that the distribution of observed site patterns differed from those expected under the Weibull model for any of the six groups (Table 4.3). This suggests a Weibull model is an adequate one.

	n	H_e	H_w
ERV9	31	$p < 0.01$	$p = 0.48$
HERVK11	27	$p < 0.01$	$p = 0.37$
HERV-H	13	$p = 0.60$	$p = 0.28$
HERV-K	8	$p < 0.01$	$p = 0.53$
group-I	38	$p < 0.01$	$p = 0.25$
group-II	38	$p < 0.01$	$p = 0.56$

Table 4.3: Group specific two-sided goodness of fit tests (Fisher exact test) show that the distribution of observed site patterns is not significantly different from those expected under Weibull deletion model H_w. This is not true in general for the exponential decay model H_e.

4.5 Discussion

In this paper we obtained site patterns sampled from the primate phylogeny. These patterns characterize the insertion and deletion process of ERVs and may be a helpful resource for other researchers. In addition, we also present a phylogenetic model which we show captures the deletion process of ERVs in a way that is superior to existing descriptions. Applying this model to site patterns from six large groups of ERVs we find that HERV-H is the most slowly deleted group of ERVs across the primate lineage and that, with the potential exception of HERV-K, ERVs appear to "die young." Below we discuss the biological implications of our findings, the limits of our approach, and why we think our approach is preferable to that used in other studies in the field.

Previous studies of ERV activity (i.e. insertions) have often proceeded by enumerating the full-length ERVs, perhaps of a specific type, from one host species. Meta-studies will then collate the results of primary studies and attempt a synthesis of their contents e.g Sverdlov (2000); Bannert and Kurth (2006). Meta studies face the difficult problem of relating various sampling (search) methodologies. They also face the impossible problem of relating counts of full-length ERVs between species when the overlap between counts is unknown. Here we think our approach is helpful. Consider Figure 4.3, where we provide estimates that allow one to answer quantitative

questions about the insertional activity of ERVs from different families and species. Assuming our sampling of viruses has been effective, we can be confident that both group-I and group-II ERVs became less active after the split of the macaque lineage from the ape lineage, but also that ERV9 and HERVK11 were more active in the ape lineage after this split than before. As we measure insertions using solo-LTRs as well as full-length ERVs, we also suggest that, contrary to Magiorkinis et al. (2015), an apparent speedup in HERV-K(HML2) is not an artefact of considering only full-length ERVs. We think our results complement existing research. We also think that similar approaches will become more useful as newly sequenced genomes give better resolution within phylogenies and allow for improved genome scale alignments.

The topic of deletion has been less widely studied in the past, perhaps because it is assumed to be unimportant to host phenotype, or perhaps because an adequate treatment requires phylogenetic data. What is true is that when a deletion process is explicitly mentioned it is often assumed to be exponential or constant over time e.g. Pereira (2004) or Lynch (2007, pg. 161). Our results show that estimates produced assuming a constant deletion model differ dramatically from those produced using an age dependent model. We also show that an assumption of a constant deletion process is clearly inappropriate for five of the six groups of ERVs we examined in primates. Therefore, pending further research, we conclude an exponential model is inappropriate more generally.

The question of whether ERV deletion rates may vary with age was addressed by Belshaw et al. (2007) in which it was reported that the deletion rate for recent integrations was 200 fold higher than for integrations that occurred over 6 Ma. This result is implicitly conditional on loci being retained in full-length form in at least one of human or chimp. To our knowledge this research, while commonly cited, has not been followed up elsewhere. Our results generalize the qualitative conclusions of Belshaw et al. (2007) to the primate phylogeny and to a variety of ERV families. This is important because the results of Belshaw et al. (2007) rely on elements from the HERV-K

category, which are unusual both because this family of ERVs have been recently insertionally active and also because some recent HERV-K insertions have been shown to have biochemical effects, including, virion formation, during early stages of human development (Grow et al., 2015). In addition, our results surpass previous ones as we provide a description (the Weibull model) that gives an indication of the expected survival function of an ERV at various points in its lifetime.

We have shown that an age dependent deletion effect occurs rapidly and is strong enough to be obvious at short time scales e.g. well before 1 Myr. A relevant question to ask is why this should be the case? Idiosyncratic factors affecting the cost, and therefore the fixation probability, of ancient full-length and solo-LTR alleles will likely always remain mysterious. However, whatever role these uncertain factors may play, the fact remains that for a solo-LTR to fix it must have been generated in the first place. Research has focused on two main causative factors in the creation of solo-LTRs: the background recombination rate and the mutational divergence between paired LTRs.

Initially it does seem reasonable to assume that the background mutation rate could account for ERVs appearing to die young. This could happen if ERVs in regions of high recombination were more likely to undergo deletion than ERVs elsewhere in the genome. The result of such a process would be a bimodal distribution of lifetimes, with ERVs in regions of high recombination being deleted very quickly and ERVs elsewhere persisting in full-length form almost indefinitely. Indeed, using the human genome, it has been demonstrated that the ratio of full-length ERVs to solo-LTRs correlates well with local genomic recombination rates (Katzourakis et al., 2007a). However, since this observation was made, it has also become widely known that local genomic recombination rates can evolve quickly (Myers et al., 2005). If this is the case, genomic recombination rates alone cannot be responsible for an age dependent deletion effect as, when the local recombination rate of the genome increased, ERVs that were present in full-length form would become more likely to be deleted,

no matter what their age.

Given the inability of local recombination rates to completely explain an age dependent deletion effect it seems that mutational divergence remains a leading hypothesis. At first this might seem surprising because for mutational divergence to be a persuasive explanation (i) mutations must introduce an allelic variant of a full-length ERV quickly and (ii) the newly introduced allele must have a substantially lower chance of undergoing recombinational deletion than the original ERV.

In answer to point (i), forward simulation under a Wright-Fisher model shows that in a population of 10,000 individuals, and assuming 2,000 bp of LTR per ERV, a (conservative) mutation rate of 10^{-8} per site per generation will introduce a mutation into a neutrally segregating pair of LTRs by 215 generations on average. Indeed, 25% of the time, a pair of LTRs will differ by the 75th generation after insertion and a full 75% of the time a difference will arise by the 300th generation. The expected frequency of the original full-length ERV at the time a difference is introduced is less than 1% and a full-length ERV can be expected to have a complementary mutant by the time it has reached a frequency of 4% at the most.

The answer to point (ii), as to whether a single difference can have a substantial effect, comes from experimental evidence. For example, Datta et al. (1997) find that a single nucleotide difference between two 350 bp substrates can lower recombination 3-fold in yeast, while Opperman et al. (2004) reach similar conclusions, finding a 4-fold reduction when using 618 bp substrates in arabidopsis. Both of these studies find that additional differences have relatively little effect. The above reasoning suggests, in agreement with the conclusions of Belshaw et al. (2007), that mutational divergence between LTRs can play an important role in preventing recombination deletion soon after an ERV has integrated.

We have suggested that mutations can arise quickly and prevent recombination. This, however, is not the whole story and we must also address stochastic effects that mean a neutrally segregating pair of LTRs that do have a mutation are very likely

to be lost from a population. This can also be done via simulation. If one tracks a such alleles until fixation occurs (we group alleles based on whether they possess 0, 1, 2, or 3 or more mutations) the allele that ultimately fixes is mutation free on only 33% of occasions. Otherwise, its LTRs contain 1 (24% of occasions), 2 (15% of occasions) or 3 or more mutations (28% of occasions). Therefore, if mutational divergence plays a significant role in reducing recombinational deletion then, in agreement with our main results, recombinational deletion must usually occur quickly. If this were not the case then most ERVs that fixed in a population would already contain mutations that prevented solo-LTR formation. Such an outcome would be inconsistent with the empirical fact that most ERVs are found in solo-LTR form. We therefore conclude that the assumption that mutational divergence between LTRs can prevent solo-LTR formation theoretically supports our statistical findings that ERVs are usually deleted very soon after integration.

Not all ERVs are deleted equally quickly. Amongst the groups of ERVs we examined we found that HERV-H were unusually slowly deleted. This is interesting for several reasons. HERV-H is notable as a family because there is very strong evidence that HERV-H associated stem cell transcripts are essential for the maintenance of stem cell identity in humans (Lu et al., 2014; Wang et al., 2014). It has also been demonstrated that highly transcribed HERV-H loci have recently evolved faster than other ERVs and other repetitive DNA (Gemmell et al., 2015). Here our results show that HERV-H loci are more likely to be preserved in a full-length state than any of the five other groups of ERVs we examined. This may suggest that HERV-H are useful to the host and preserved in a full-length state. This may also suggest that a side effect of their rapid evolution is a degree of divergence between paired LTRs that means they are unlikely to be subject to recombinational deletion. As discussed by Gemmell et al. (2015), a faster than expected evolution of LTRs is evidence of a phenotypic effect on the host, but not necessarily evidence of an ERV providing a benefit to the host.

It is important to discuss the limits of our model. While phylogenetic assign-

ment of insertions to branches is probably more precise than assignments based on age estimates, such assignment is limited by the resolution of the phylogeny used. This is clear from Figure 4.3. The branches leading to human and chimp generally reflect a decrease in ERV activity while the branch leading to macaque can only reflect the average activity over a 30 Myr period. The average is reassuringly similar to the average over internal branches but it is clear our approach is of less use for distantly related species. Our approach assumes that ERVs arrive in full-length form. This assumption was necessary, but it is reasonable to point out that some site patterns (those with xs at every tip) might never have been passed vertically from generation to generation in full-length form. From a scientific perspective studies of active ERV families, such as KoRV (Ishida et al., 2015), can potentially tell us how many loci endogenize in solo-LTR form. Finally, our model is limited by the availability of full-length insertion data. Branches for which all insertions have resulted in solo-LTRs provide no upper bounds on the rate of deletion. Larger datasets or improved sampling of site patterns can resolve this problem.

Limitations notwithstanding, we hope to show that ERV activity is well described using phylogenetic techniques that avoid problematic LTR dating and that previous arguments that younger ERVs are more quickly deleted than older ones can be correctly formalized. We also suggest that our finding that HERV-H is removed slowly across the primate phylogeny supports a long term biological role for the family. New questions can be formulated using our approach, for example it is simple to use our model to perform a maximum likelihood reconstruction of the ancestral state of an ERV. We hope that similar studies outside the mammalian phylogeny might be undertaken using some of our ideas.

4.6 Conclusion

We provide 1,197 phylogenetic patterns describing the state of endogenous retroviruses (ERVs) in six primate species. We introduce a phylogenetic framework to quantitatively study the process of insertion and deletion, without relying on LTR dating. We place the activity of six categories of ERV in direct quantitative comparison for the first time. We find that ERVs "die young" but that HERV-H loci are markedly more long-lived than ERV9, HERVK11, HERV-K, or other class-I and class-II loci. We show that a probabilistic age dependent (Weibull) process is sufficient to describe the ERV deletion process and suggest our approach may be useful to other researchers in the paleovirological field.

Chapter 5

The genomic structure of highly transcribed HERV-H loci

5.1 Abstract

Background

HERV-H is a prolific and unusual family of primate endogenous retroviruses that has recently been found to play an essential role in the maintenance of stem cell identity in humans. Though HERV-H has been studied for several decades, there has been no work that thoroughly relates the transcription of the majority of HERV-H loci to their detailed structure. Here we identify the relationship between the transcription of human loci and their age, their LTR repeat type, the characteristic deletions they possess, their presence in other primate genomes, and their distance to the nearest host gene.

Results

We find that the most transcribed HERV-H loci are younger, more fragmented, and less likely to be present in other primate genomes, so that most of the highly tran-

scribed sequence found in human is missing from chimpanzee, gorilla, orangutang or macaque. Though the most highly transcribed HERV-H contain many deletions in their genes, the presence of the final 3' region of *gag* is positively correlated to transcription in several cell types. The repeat types within an LTR are important to the cell type in which a HERV-H is transcribed: type-I HERV-H are highly transcribed in stem cells while HERV-H with type-II repeats are more highly transcribed in embryonic cells.

Conclusion

We suggest that the finding that type-II repeats are correlated to the earlier transcription of HERV-H may be important when attempting to reliably identify populations of naive like stem cells in culture. We further argue that the surprising positive correlation between the zinc finger protein binding region of *gag* and HERV-H transcription indicates that an ability to be effectively repressed was a facilitator of HERV-H cooption. If this hypothesis is correct, it goes some way towards explaining why an unusually large proportion of HERV-H loci have been preserved in a full-length form.

5.2 Introduction

Endogenous retroviruses (ERVs) are the result of germ line retroviral integrations that are passed from generation to generation in a Mendelian fashion (Bannert and Kurth, 2006). At integration time, a typical ERV locus contains the viral genes *gag*, *pol* and *env*, as well as two flanking long terminal repeats (LTRs) that are identical. Once present in a population, ERVs may increase in number via reinfection or retrotransposition until all active members of the family are silenced by host defences, degraded by mutations, or truncated by solo-LTR formation (Katzourakis et al., 2005; Johnson, 2007). Though many human endogenous retroviruses (HERVs) are thought to be essentially irrelevant to our biology, notable exceptions include the *syncytins* (Lavialle et al., 2013)

and members of the HERV-H family, that have recently been found to be crucial to the maintenance of stem cell identity (Lu et al., 2014; Wang et al., 2014).

The HERV-H family, formerly known as RTVL-H, has been studied down to the level of individual insertions on many occasions in the past 30 years. Early research quickly identified variation in the LTR regions of HERV-H and divided the family into type-I and type-II subtypes, both of which polyadenylated transcripts in human cells (Mager, 1989). Later, a third type of LTR, type-Ia was identified. In structure, all HERV-H LTRs appeared to share common sequence in the U3, R and U5 regions. However, whereas type-I LTRs typically had two copies of a 49 bp type-I repeat, and type-II LTRs had one copy of a type-I repeat followed by a variable number of 27–32 bp type-II repeats, the type-Ia LTRs had both a type-I repeat and a type-II repeat, as well as some further type-I typical sequence (Goodchild et al., 1993). For this reason, type-Ia LTRs are thought to have arisen via a recombination between type-I and type-II LTRs (Goodchild et al., 1993). Type-Ia HERV-H was originally thought to have expanded recently and to have stronger transcription than HERV-H with pure type-I or type-II LTRs (Goodchild et al., 1993). However, the relative youth of type-Ia repeats was later brought into question by Anderssen et al. (1997) who discovered type-Ia repeats in the marmoset, a New World monkey species.

The sequencing of the human genome has been important for more complete analysis of HERV-H. A study of most full-length HERV-H integrations was performed by Jern et al. (2004) who clustered HERV-H into two groups, the larger HERV-H group of 926 full-length elements and the smaller HERV-H like group of 92 full-length elements. Within the larger HERV-H group, Jern et al. (2004) further defined two subgroups: an older group of 77 RGH2-like elements having a fairly intact *pol* and more frequently containing *env*; and a younger group of 705 RTVLH2-like elements having more *pol* deletions and less frequently containing *env*. The 926 bona-fide HERV-H elements in the human genome were subsequently studied in some detail resulting in an annotated consensus sequence (Jern et al., 2005). In an unrelated study on ERV

replication mechanisms, a small number of relatively intact HERV-H (presumably the RGH2-like elements) were found to have low *env* dN/dS and proposed to be responsible for copying the less intact members of the HERV-H family (Belshaw et al., 2005b). Belshaw et al. (2005b) also observed a common deletion in the *gag* gene to add to the four common *pol* deletions and one common *env* deletion previously described by Mager and Freeman (1995).

In the last two years HERV-H has again come under study as a relationship between retrotransposon transcription and stem cells has become a focus of research. In the case of HERV-H, there does not seem to be documented evidence suggesting a role for viral proteins, though there is evidence that HERV-H transcripts are upregulated in, and necessary for, the maintenance of stem cell identity. This evidence derives from studies including those of Wang et al. (2014) who show that nearly half of the full-length HERV-H in the human genome are bound by pluripotency associated transcription factors NANOG, OCT5 and LBP9, and that, as such, produce a number of chimeric hESC (human embryonic stem cell) and hiPSC (human induced pluripotent stem cell) specific transcripts and long non-coding RNAs. By disrupting HERV-H or LBP9, Wang et al. (2014) could demonstrate that some HERV-H play an essential biological role: differential markers were upregulated while pluripotency-associated transcription factors were downregulated, so that an ability for self-renewal was shown to be impaired. The same year, Lu et al. (2014) reached a similar conclusion on the necessity of HERV-H to stem cell identity after discovering that interfering with HERV-H transcripts in hESC led to modified cells becoming more fibroblast like, with concomitant changes in the appropriate transcriptional markers. These two studies are supported by the transcriptomic work of Kunarso et al. (2010) and of Fort et al. (2014).

It is likely that the authors who conducted early analyses on HERV-H would, if in fact conducting their analyses in 2015, have some interest in relating the features they described at the genomic level to the phenotypic effects of HERV-H in stem cells.

It is perhaps surprising that recent studies that investigate these phenotypic effects have not yet made reference to such comprehensive genomic research. This is particularly true given the fact that although the LTR of HERV-H has been the focus of attention, an unusual proportion of HERV-H loci are actually present in full-length form rather than as solo-LTRs (Bannert and Kurth, 2006). The extraordinary state and dynamics of HERV-H (see Chapter 4) raise the question of whether the internal region of HERV-H has also played a role in primate physiology too.

In this study we explore the intrinsic genomic factors associated with measurements of HERV-H transcription in humans as made available by Wang et al. (2014). The intrinsic features we refer to are the systematic deletions and idiosyncratic decay of particular HERV-H loci with respect to the known consensus (Jern et al., 2005), and also LTR subtypes, as used to classify HERV-H in the pre-genomic era (Anderssen et al., 1997). We also consider the phylogenetic placement of loci, their age, and their state in other primate species.

It is reasonable to assume that at some point most of what we identify as HERV-H today would have performed some biological function as part of an exogenous virus. What is good for a virus would not usually be expected to be good for a host, and it is as yet unclear precisely which characteristics of HERV-H were important for its co-option in stem cells. Below we identify the features of HERV-H that are significantly correlated with transcription and show how different subtypes of viruses are transcribed in different types of cell. We also find an interesting relationship between *gag* and transcription, and conjecture that it is this region of *gag* that is partially responsible for the maintenance of so many HERV-H loci in a full-length state.

5.3 Method

5.3.1 Multiple sequence alignments of HERV-H loci

We obtained the genomic sequence underlying the the 1,225 full-length HERV-H loci described by Wang et al. (2014) from the UCSC Genome Browser database (hg19, GRCh37) available at http://genome.ucsc.edu (Rosenbloom et al., 2015). We also obtained the corresponding RepeatMasker (Smit et al., 2004) gene annotations from the same database.

Before aligning the HERV-H loci we first used RepeatMasker annotations to identify any non HERV-H repeats within the underlying sequences. Such nuisance repeats, for example SINEs or LINEs, would, if ignored, introduce spurious indels into our analysis. For this reason all but the outermost 20 bp of the nuisance repeats were removed from the 1,225 HERV-H sequences before constructing alignments. The remaining 40 bp or less of nuisance repeats were flagged so that we could remove them by hand, thereby ensuring that we did not remove mistakenly RepeatMasked genuine HERV-H sequence. Nuisance repeats were found to be rare, so that 80% of the sequence underlying the 1,225 loci remained completely unmodified. The remaining 20% of loci had a median of 12% of their underlying sequence removed.

To identify the *gag*, *pol*, and *env* genes within the 1,225 HERV-H sequences we used tBLASTn (Altschul et al., 1990). We searched each of the 1,225 loci using a previously published (Jern et al., 2005) HERV-H consensus sequence as a query. Hits of at least 25 bp in length and with an expect value no more than 10^{-6} were merged in a way that maintained fragment order and sense. The result of this reconstruction was 1,080 *gag*, 1,126 *pol* and 1,081 *env* genes.

Only 20 of the 1,225 HERV-H loci were not matched by one of the three tBLASTn searches. An inspection using Dfam (Wheeler et al., 2013) revealed that these loci were sequences with short internal regions, that contained other retroviral insertions, or that were short overall (one full-length region was only 44 bp).

A similar search and merge process was performed for the 5′ (LTR5) and 3′ LTRs (LTR3) of each of the 1,225 HERV-H loci. In this case the search was performed using BLASTn (Altschul et al., 1990) and resulted in the reconstruction of 908 LTR5 and 932 LTR3.

To construct multiple sequence alignments of the genes and LTRs of the HERV-H loci we first pairwise aligned the reconstructed sequences to the appropriate part of the consensus. We then progressively combined these alignments to create five multiple sequence alignments, one for each gene and one for each LTR. Pairwise alignment was conducted with Stretcher (Rice et al., 2000) and progressive multiple alignment was conducted using MUSCLE (Edgar, 2004).

5.3.2 Distance to nearest gene

The RefSeq gene annotation track was downloaded from the appropriate UCSC Genome Browser database (as detailed above). The distance between the centroid of each HERV-H locus and its nearest neighbouring gene was calculated and recorded.

5.3.3 Characterization of LTR subtypes

Several examples of HERV-H LTR subtypes (Mager, 1989; Goodchild et al., 1993) are provided by Anderssen et al. (1997). We constructed consensus type-I and type-II repeats as well as consensus unique-I and unique-II sequences based on Figure 2 of the study by Anderssen et al. (1997). These consensus sequences were used as queries in a BLASTn search against the 1,225 HERV-H loci. Hits with expect values of no more than 10^{-6} were treated as indicating the presence of the appropriate sequence. Data on the presence of LTR subtype sequences at HERV-H loci are recorded in Supplementary File 5.1.

5.3.4 Pairing loci to EPO multiple alignments

To examine the status of the 1,225 HERV-H full-length loci in other primates we obtained the Enredo-Pecan-Ortheus (EPO) genome scale multiple sequence alignment from Ensembl Release 71 (Flicek et al., 2012). We then used BLASTn to locate the HERV-H loci in the human row of the EPO alignments. Of the 1,225 HERV-H loci, 847 were unambiguously located (unique and exact matches) within a six-way EPO alignment. The remaining loci were not present in the alignments in an unambiguous form e.g. the region of the human genome they were located in might not have been included in the EPO alignments, or the region they were located in may have been duplicated one or more times in a primate other than human.

5.3.5 Tree building and phylogenetic GLS

A supermatrix concatenation of *gag*, *pol* and *env* alignments was produced. The alignment was edited by hand to remove short or badly aligned regions and sequences that could not reasonably be assumed to be homologous. The tree building software RAxML 8.2.3 (Stamatakis, 2014) was used to produce a maximum likelihood (ML) tree relating the 834 sequences in the supermatrix alignment. Tree inference was performed under the GTR + gamma substitution model.

The supermatrix based tree was rooted by constructing an auxiliary phylogeny of the reverse-transcriptase (RT) region of *pol* (nucleotides 82-576 of the *pol* consensus). The 569 HERV-H sequences with relatively complete RT, having over 140 of 165 possible codons, were first translated and then combined with a panel of 15 HERV-W RTs. An ML tree was constructed from the resulting alignment using the RAxML `PROTGAMMAAUTO` option.

All regression analysis was conducted with the R system (R Development Core Team, 2008). Regressions taking into account phylogeny were performed using a phylogenetic generalized least squares (PGLS) approach as introduced by Grafen (1989).

The λ method (Pagel, 1997) was used to assess phylogenetic signal. Analyses were conducted using the APE (Paradis et al., 2004) implementation of these methods.

5.4 Results

In this section we first describe the retrieval and representation of the genomic features of the HERV-H loci in our study. We then describe the statistical models that we use to relate HERV-H transcription to these features.

5.4.1 Genomic features of HERV-H loci

We retrieved the 1,225 full-length HERV-H loci described by (Wang et al., 2014) from the human genome. We then produced multiple sequence alignments of the genes *gag* (1,080 sequences), *pol* (1,126 sequences) and *env* (1,081 sequences) using the consensus HERV-H described by Jern et al. (2005) as a guide. We used the same consensus to form alignments of the 5′ (908) and 3′ (932) LTRs (henceforth LTR5 and LTR3).

Examining the sequence in each alignment it was possible to find regions that were commonly deleted in subsets of viral sequences. The presence of sequence in these regions was bimodally distributed and is represented in Figure 5.1, where the regions are named via enumeration from 5′ to 3′. We identified four regions in the alignment of HERV-H LTRs (L51–L54 for the four regions in LTR5 and L31–L34 for the four regions in LTR3), four regions in *gag* (G1–G4), eight regions in *pol* (P1–P8) and four regions in *env* (E1–E4). As per Belshaw et al. (2005b), if less than 5% of a region was aligned for a particular HERV-H the region was marked as absent at that locus while if more than 55% of a region was aligned it was marked as present. Ambiguous regions were coded as missing values, and the frequency of the presence and absence of regions is given in Table 5.1. The 25 regions we defined were included as indicator variables (i.e. 1 or 0) when constructing statistical models (below).

Some features of HERV-H LTRs have previously been described in detail. In

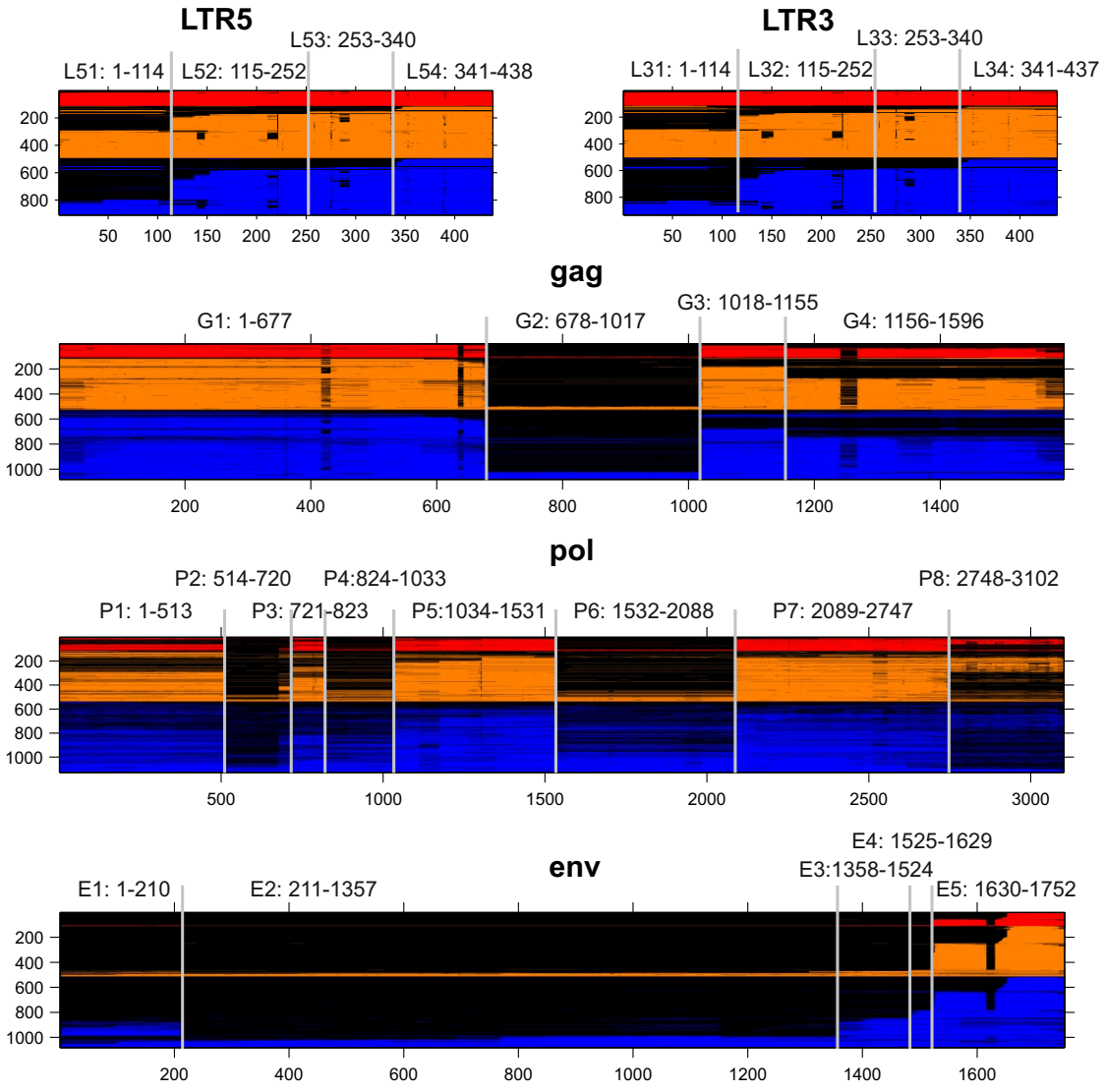

Figure 5.1: A representation of HERV-H sequences recovered from the human genome and aligned to a known consensus. Loci are coloured by the categorical transcription levels of Wang et al. (2014): red represents highly active loci; orange represents moderately active loci; and blue represents inactive loci.

feature	present	absent	ambiguous
L51	432	454	22
L52	758	133	17
L53	786	114	8
L54	875	31	2
G1	974	48	58
G2	113	950	17
G3	823	227	30
G4	684	326	70
P1	671	375	80
P2	159	731	236
P3	661	428	37
P4	363	724	39
P5	854	112	160
P6	367	706	53
P7	870	119	137
P8	391	649	86
E1	144	890	47
E2	94	928	59
E3	272	729	80
E4	704	334	43
E5	1027	39	15
L31	452	462	18
L32	778	131	23
L33	812	118	2
L34	891	39	2
TI	1047	178	0
TII	184	1041	0
UI	998	227	0
UII	124	1101	0

Table 5.1: Frequencies of HERV-H features across the 1,225 full-length HERV-H loci.

particular, members of the HERV-H family have previously been categorized into type-I, type-Ia, and type-II elements (Mager, 1989; Goodchild et al., 1993; Anderssen et al., 1997). We used BLASTn to search for the appropriate characteristic sequences at the 1,225 HERV-H loci. The overall number of TI and TII repeats, as well as the overall number of UI and UII subtype specific unique repeats detected is shown in Table 5.1. These LTR subtype indicator variables are considered in our analyses below.

Most of the features we analyze were defined with respect to the sequence of the HERV-H loci themselves. However, to place the loci in context we also obtained their distance from the nearest gene, as presented by the RefSeq gene annotation track of the UCSC Genome Browser database. It was also possible to locate 847 of the 1,225 HERV-H loci in a six-way alignment of primate genomes, thereby giving some evolutionary perspective on the present day structural state of the loci in other primates. Finally, we were able to determine the genetic distance (K80) between the paired LTRs of 627 of the 1,225 HERV-H loci. Though the LTRs of ERVs may not evolve at strictly neutral rates (Hughes and Coffin, 2005; Gemmell et al., 2015), the divergence between paired LTRs does provide approximate information on the age of an insertion. Therefore both genetic distance and orthology information contribute information on the age of the loci under investigation.

5.4.2 Characteristics associated with present-day HERV-H transcription

HERV-H transcription data for 1,225 HERV-H loci is provided by (Wang et al., 2014). The most highly transcribed of these loci have been diverging quickly, and are presumably under directional selection (Gemmell et al., 2015). An unanswered question, and the crux of this study, is as follows: what is the relationship between the genomic characteristics of a HERV-H locus and the level at which the locus is transcribed?

To answer this question, we used multiple regression to quantify the associa-

tion between HERV-H transcription and the per locus genomic features of HERV-H described above. The response variable in our regression is the logarithm of per locus mean transcription in reads per kilobase per million reads (RPKM) across the 5 cell culture types defined by (Wang et al., 2014): somatic (32 lines); cancer (8 lines); embryonic (114 lines); hESC (55 lines) and hiPSC (25 lines). Of the 6,125 measurements available (5 cell types $\times$ 1,225 HERV-H loci), 6,025 had non-zero transcription data and were included in our regression. The resulting model M1 is shown in Table 5.2.

coefficient	**estimate**	**2.50%**	**97.50%**	**std. err.**	t **value**	p **value**
somatic (intercept)	-7.21	-7.87	-6.56	0.33	-21.69	$p < 0.01$
cancer	0.94	0.64	1.23	0.15	6.30	$p < 0.01$
embryo	2.57	1.20	3.94	0.70	3.68	$p < 0.01$
hESC	3.19	2.43	3.95	0.39	8.24	$p < 0.01$
hiPSC	3.48	2.72	4.24	0.39	8.99	$p < 0.01$
L51	1.37	1.18	1.57	0.10	13.72	$p < 0.01$
L52–L54	-0.20	-0.38	-0.01	0.09	-2.09	$p = 0.04$
L52–L54:embryo	0.46	-0.01	0.92	0.24	1.93	$p = 0.05$
G4	0.49	0.24	0.74	0.13	3.82	$p < 0.01$
P4	-0.57	-0.81	-0.32	0.12	-4.56	$p < 0.01$
P5	-0.78	-1.18	-0.38	0.20	-3.83	$p < 0.01$
P6	-0.96	-1.24	-0.67	0.14	-6.64	$p < 0.01$
E3	-0.49	-0.84	-0.13	0.18	-2.70	$p = 0.01$
LTR divergence	-9.05	-14.27	-3.83	2.66	-3.40	$p < 0.01$
stem:type-I	1.16	0.78	1.54	0.19	6.03	$p < 0.01$
embryo:TII	3.47	2.47	4.48	0.51	6.78	$p < 0.01$

Table 5.2: Model M1, a non phylogenetic representation of HERV-H transcription in the form of a multiple regression; $R^2 = 0.57$, Adjusted $R^2 = 0.57$, F-statistic = 198.7 on 15 and 2243 DF, $p < 0.01$.

To construct model M1 we used AIC based model selection and plots to explore model space and then formulated a minmum adequate model to explain features of our dataset i.e. we retained only significant and marginally significant explanatory variables at the expense of model fit.

Transcription by cell type: as model M1 demonstrates, HERV-H transcription is distinct among cell types, with transcription being lowest in somatic cells (the intercept term). Transcription increases significantly in turn for each cell type: cancerous

cells, embryonic cells, hESC cells and hiPSC cells.

Genic regions: the presence of genic regions G1–G4, P1–P8 and E1–E5 was generally found to be negatively associated with transcription of HERV-H loci in all cell types. The regression shows this negative correlation was significant for regions P4, P5, P6 and E3 in particular. Notably, genic region G4 was positively correlated with transcription.

Transcription and LTRs: although non-genic LTR region L51 was found to be significantly positively correlated with HERV-H transcription, regions L52–L54 (indicators L52 + L53 + L54) appeared to hinder transcription. However, as model M1 shows, the negative correlation between L52–54 and transcription did not apply in embryonic cells, where the presence of these three LTR regions appears to correlate in the opposite direction. Note that variables L31–L34 were essentially collinear with variables L51–L54 and were therefore dropped from the regression model—results applying to L51–L54 were confirmed as applying equally to L31–L34.

We observed the presence of type-I and type-II repeats to be correlated with transcription in particular cell types. Type-I HERV-H were particularly associated with stem cell transcription. This is confirmed by the significant positive correlation between transcription and the interaction term between composite variable type-I (indicators TI + UI) and indicator variable stem (1 for measurements from hESC or hiPSC cells and 0 otherwise). LTRs containing type-II repeats (i.e. type-Ia and type-II LTRs) are particularly associated with embryonic transcription. This association is demonstrated by the positive correlation between measurements from embryonic cells and the variable TII.

Effect of other factors: HERV-H transcription was found to decrease with the age of a locus as approximated by the nucleotide divergence between its paired LTRs. The distance between a HERV-H locus and the closest gene to that locus was negatively correlated with transcription in somatic cells and is investigated below.

Summary of regression: An R^2 value of 0.57 indicates that the contempo-

rary state of HERV-H transcription is explained well by the intrinsic characteristics of HERV-H loci; the residual error in model M1 is largely due to the under fitting of a number of loci that are exceptionally highly transcribed as can be seen in Figure 5.2.

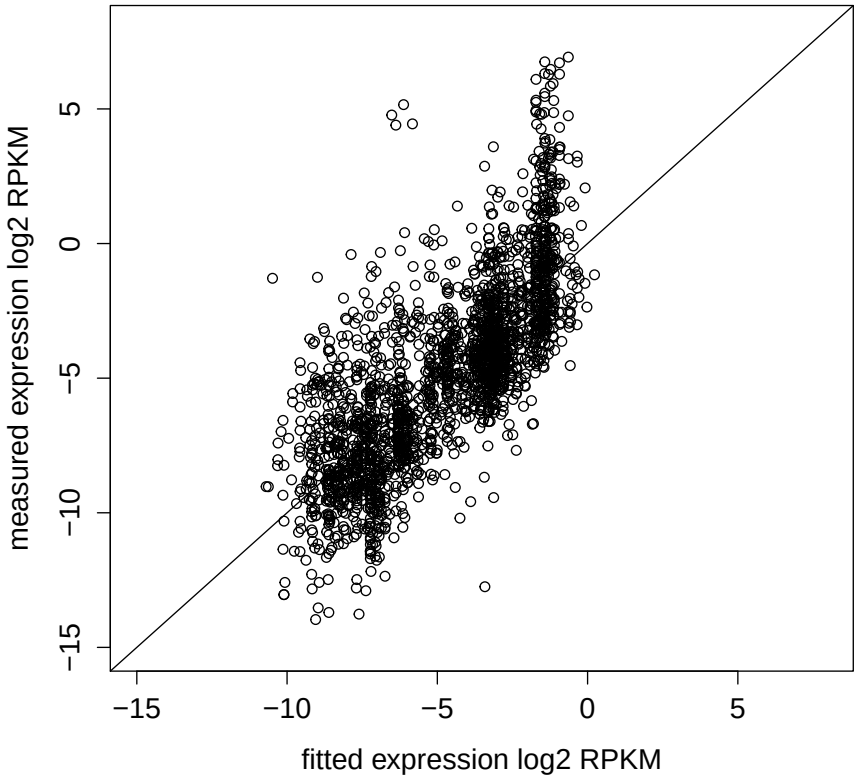

Figure 5.2: Fitted versus actual transcription for non phylogenetic model M1.

5.4.3 Phylogenetic analysis of HERV-H transcription

The coefficients in Table 5.2 describe, at face value, the association between the genomic features of HERV-H loci and their transcription: each transcription measurement in model M1 is considered independent of every other and, when describing the current state of HERV-H transcription in humans, this is certainly a useful approach. However, when robustly testing a hypothesis it is preferable to control for non-independence between measurements due the evolutionary relationship between

the entities they are derived from. To do so one performs a phylogenetic regression (see Methods).

A phylogenetic regression requires a tree that relates HERV-H loci. We were able to create a supermatrix nucleotide concatenation of the *gag*, *pol* and *env* genes of 847 of the 1,225 HERV-H loci. A maximum likelihood (ML) phylogeny was then produced (see Methods). Of the loci in the phylogeny, 409 had a complete collection of genomic features, i.e. no missing values, and could therefore be used in our phylogenetic analysis. We first describe the placement of the genomic features of HERV-H with respect to our ML tree. We then detail three phylogenetic multiple regressions: models M2 (somatic), M3 (embryonic), and M4 (stem).

Figure 5.3: shows a ML phylogeny of the aforementioned 409 HERV-H loci. Tips of the tree are coloured red, orange, or blue according to the categorical transcription level assigned to them by Wang et al. (2014). (Categorical transcription is an assignment of transcription based upon a hierarchical clustering of loci by transcription across all cell lines.) It can be seen that loci annotated as highly transcribed by Wang et al. (2014) are located towards the bottom of this ladderized tree. We confirmed that this phenomenon was also true of phylogenies built using LTRs and from amino-acid or nucleotide trees of individual HERV-H genes, though we do not show these trees here.

The columns somatic, embryo, and stem of Figure 5.3 show the average transcription of loci across cells categorized as somatic, embryonic or stem respectively. Considering these columns in conjunction with the column marking LTR type one can see that type-Ia/II loci are distributed through the tree and are associated with higher levels of embryonic transcription, even if a locus is categorically characterized as inactive or moderately active by Wang et al. (2014).

Columns L5, *gag*, *pol*, *env*, and L3 indicate the presence or absence of the individual viral regions L51–L54, G1–G4, P1–P8, E1–E5 and L31–34. Highly active loci towards the bottom of the tree can be seen to have highly intact LTRs while some less

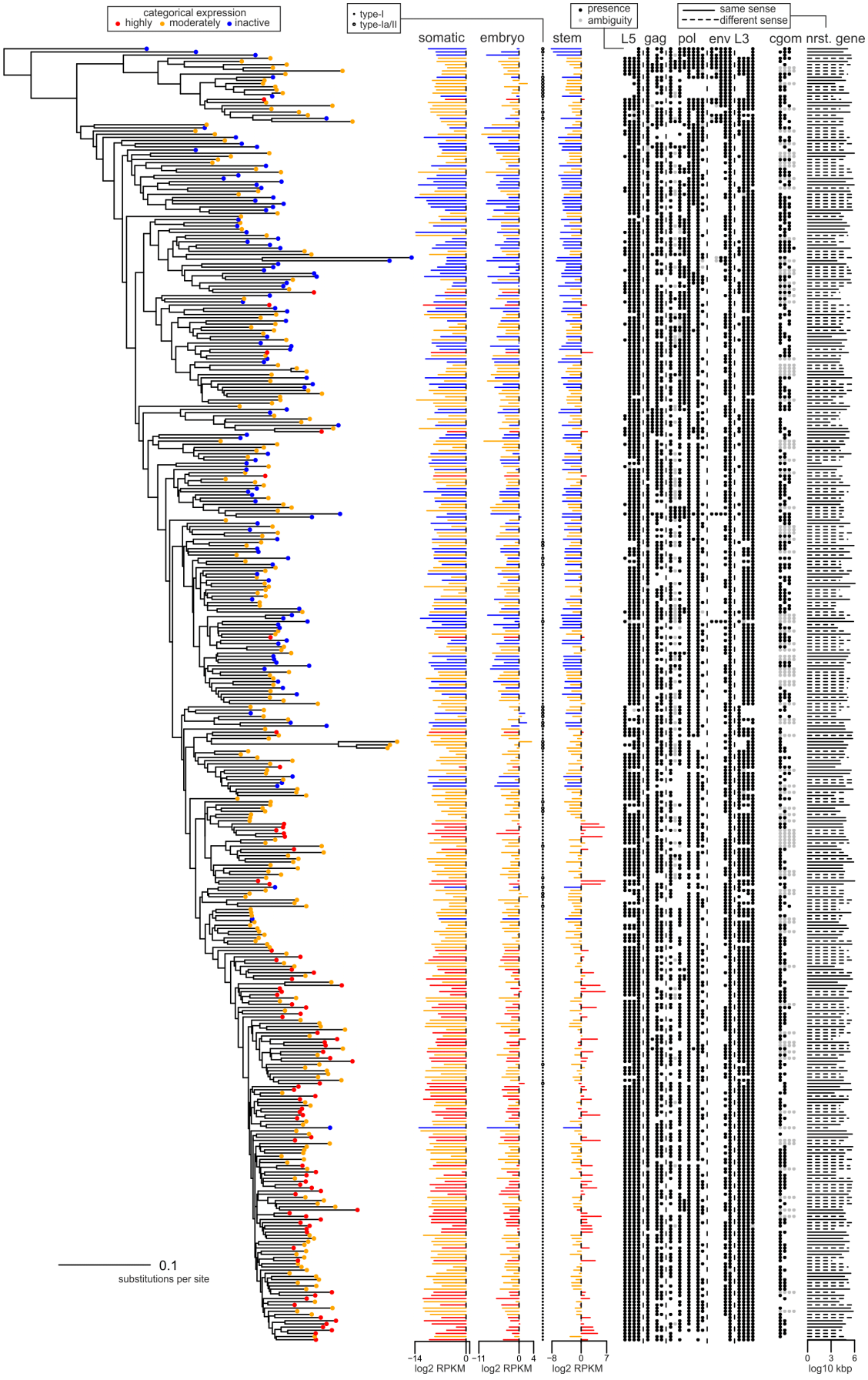

Figure 5.3: Phylogeny, cell type specific transcription levels, and per locus genomic features of 409 full-length HERV-H loci from the human genome.

transcribed loci towards the top of the tree have less intact LTRs. Less active older loci towards the top of the tree tend to have relatively complete internal regions when compared to the active loci at the bottom of the figure.

The sequences used to build the phylogeny in Figure 5.3 come from the human genome. An indication of the age and status of HERV-H loci in other primates is given by the column *cgom*. This indication was obtained by searching for each HERV-H loci within the six-way EPO alignment of primate genomes (see Methods). Loci present in chimpanzee, gorilla, orangutan or macaque at a minimum of 25% of the level that they appear in human are then marked with a dot as appropriate. Surprisingly, it is clear from this column that that less derived and less active HERV-H loci located towards the top of the tree appear to more likely to be present in a substantive way in other primates. The active and derived loci towards the bottom of the tree tend to be absent or degraded in other primates.

A simplified version of Figure 5.3 appears as Figure 5.4, and highlights the main findings of this study.

Phylogenetic regression: the PGLS method uses branch lengths of a tree to specify the variance-covariance structure of regression residuals. A λ parameter may then be introduced to specify the strength of the interaction between phylogeny and residuals. The ML estimate of λ can therefore be used to describe the degree of phylogenetic signal in a regression. The PGLS technique is thought to be relatively robust to phylogenetic error (Garamszegi, 2014, p. 122). Three minimally adequate PGLS regressions are reported in Table 5.3: model M2 has mean somatic transcription as a response variable; model M3 has mean embryonic transcription as a response variable; and model M4 has mean transcription across hiPSC and hESC as a response variable.

Model M2: suggests that the presence of genic region P7 is positively correlated with HERV-H transcription in somatic cells. The model suggests a negative correlation of similar magnitude for genic region P5. Finally, the model suggests that the distance between a HERV-H and its nearest gene is negatively correlated with its tran-

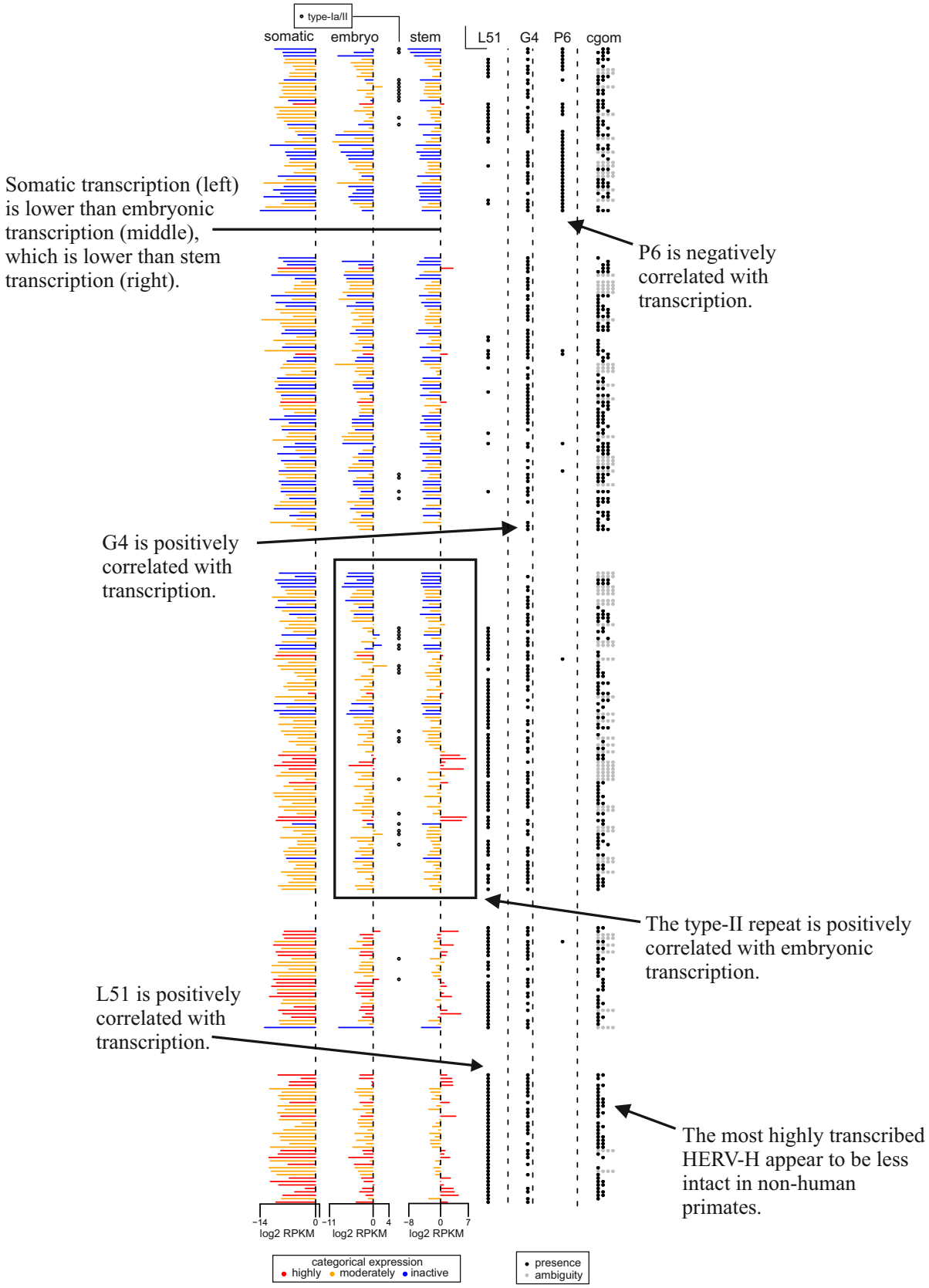

Figure 5.4: A simplified version of Figure 5.3, highlighting the main findings of this study. Note that many loci are omitted for the sake of clarity.

	model M2 — somatic cells					
coefficient	**estimate**	**2.50%**	**97.50%**	**std. err.**	t **value**	p **value**
(Intercept)	-8.09	-10.27	-5.91	1.11	-7.27	$p < 0.01$
P5	-1.60	-3.13	-0.08	0.78	-2.06	$p = 0.04$
P7	1.94	0.29	3.59	0.84	2.31	$p = 0.02$
nrst. gene	-3.00×10^{-6}	-3.58×10^{-6}	-1.48×10^{-6}	0.00	-4.74	$p < 0.01$
	model M3 — embryonic cells					
coefficient	**estimate**	**2.50%**	**97.50%**	**std. err.**	t **value**	p **value**
(Intercept)	-5.26	-6.92	-3.60	0.85	-6.22	$p < 0.01$
L51	0.76	0.23	1.29	0.27	2.82	$p < 0.01$
G4	0.46	0.10	0.82	0.18	2.50	$p = 0.01$
P6	-0.98	-1.72	-0.23	0.38	-2.57	$p = 0.01$
type-II	2.24	1.32	3.16	0.47	4.76	$p < 0.01$
type-Ia	0.71	-0.55	1.97	0.64	1.11	$p = 0.27$
	model M4 — stem cells					
coefficient	**estimate**	**2.50%**	**97.50%**	**std. err.**	t **value**	p **value**
(Intercept)	-6.23	-8.17	-4.29	0.99	-6.30	$p < 0.01$
L51	1.28	0.72	1.85	0.29	4.46	$p < 0.01$
G4	0.75	0.21	1.29	0.27	2.72	$p = 0.01$
P6	-0.91	-1.75	-0.06	0.43	-2.11	$p = 0.04$
LTR divergence	-21.21	-32.54	-9.89	5.78	-3.67	$p < 0.01$
type-I	0.88	0.44	1.33	0.23	3.87	$p < 0.01$

Table 5.3: Three phylogenetic regressions relating transcription to genomic characteristics of HERV-H loci. Summary, 409 degrees of freedom: M2 (somatic) $\lambda = 0.49$ (0.12, 0.87), $R^2 = 0.01$; M3 (embryonic): $\lambda = 0.48$ (0.17, 0.78), $R^2 = 0.24$; M3 (stem) $\lambda = 0.50$ (0.27,0.73), $R^2 = 0.13$.

scription in such that that a 1 Mbp increase in distance results in a 3–11 fold drop in transcription.

Model M3: confirms several of the features seen to be important across the larger dataset or in Figure 5.3 are robust to phylogenetic correction. The presence of genic region P6 is negatively correlated with embryonic transcription whereas the presence of genic region G4 is positively correlated with embryonic transcription. The positive correlation between the presence of region L51 and embryonic transcription is confirmed, as is the positive correlation between type II repeats and embryonic transcription.

Model M4: agrees with model M3 on the role of region L51, and genic regions G4 and P6. The model also confirms that type-I LTRs are correlated with stem cell transcription, as suggested by the non-phylogenetic regression. Additionally, model M4 suggests that older ERVs that have a higher divergence between LTRs are less active in stem cells. This negative correlation between stem cell transcription and LTR divergence would be expected based on the clustering of highly transcribed loci towards the bottom of the tree in Figure 5.3.

Summary of regression: In contrast with model M1, models M2–M4 contain fewer significant effects. Using a coefficient of determination corrected to take account of correlation structure (Paradis, 2011, p. 224) we find that for model M2, $R^2 = 0.01$ and therefore that model M2 explains almost nothing. As somatic transcription of HERV-H is low in absolute terms and decreases with the distance of a loci from genes we suspect that it is ectopic noise i.e. some transcription of HERV-H is registered in somatic cells but this is due to the transcription of neighbouring genes and imperfect silencing on the part of the host. For both these reasons we shall not consider model M2 further. Model M3 has $R^2 = 0.24$ and therefore explains nearly a quarter of the variance in embryonic transcription of HERV-H using only a few features of the HERV-H loci themselves. Model M4 has $R^2 = 0.13$ and as with model M1, seems to under fit the most highly transcribed HERV-H loci (Figure 5.5). All PGLS models pos-

sessed an intermediate level of phlyogenetic residuals: M2 $\lambda = 0.49$ (95% confidence interval 0.12–0.87); M3 $\lambda = 0.48$ (95% confidence interval 0.17–0.78); M4 $\lambda = 0.50$ (95% confidence interval 0.27–0.73).

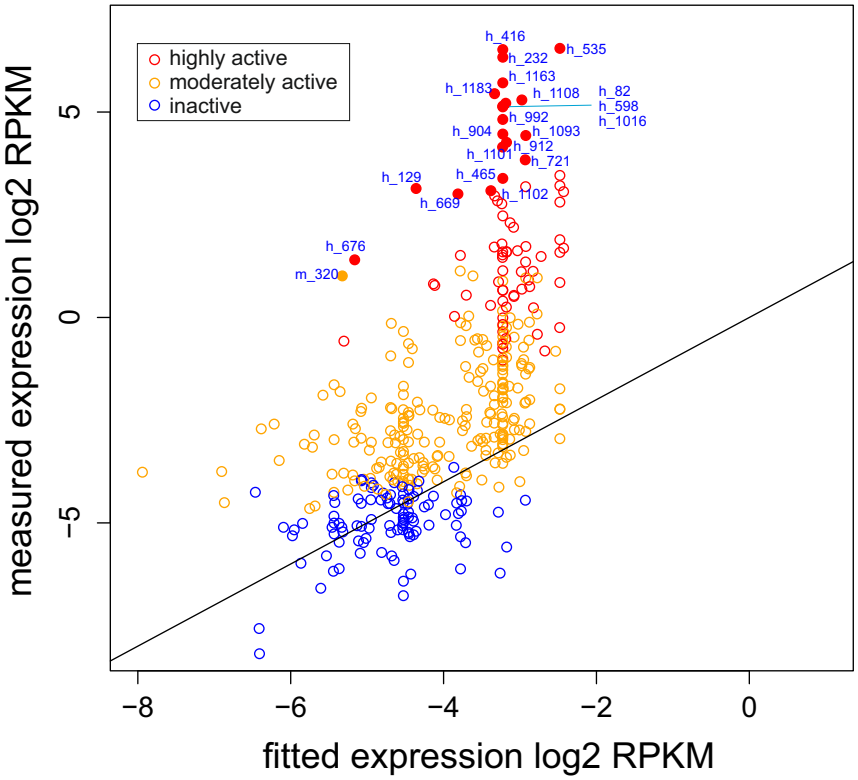

Figure 5.5: Fitted versus actual transcription for model M4 (stem). Points are colour coded according to the categorical transcription levels of Wang et al. (2014). Outliers are marked using solid circles and identified by a serial number (see Supplementary File 5.1).

The characteristics of the most under fitted 5% of HERV-H loci from model M4 were examined. These 21 outliers (Figure 5.5) were found to have similar genomic features to the sample as a whole except that they were more likely to possess region L51 and were located slightly further away from genes than the sample average. Only the presence of region L51 was statistically significant (Wilcoxon rank sum test, W = 5361, $p < 0.01$). The 21 outliers were also examined in the UCSC Genome Browser. A minority of 9 loci were found to be adjacent to regions marked as highly conserved

across vertebrates (PhyloP track). Some 8 loci were marked as part of spliced ESTs as would be expected given their transcription in cultured stem cells. As might be expected from the results of Gemmell et al. (2015), in general there also appeared to be a higher density of SNPs in the HERV-H regions than in the corresponding flanks.

5.5 Discussion

We investigated 1,225 full-length human HERV-H loci for which transcription data is available (Wang et al., 2014) and identified a hierarchy, whereby HERV-H transcription is lowest for somatic cells, and higher for cancerous, embryonic, hESCs and hiPSCs cell lineages in turn. Comparing these full-length HERV-H loci to a published consensus (Jern et al., 2005) we also determined the properties of HERV-H that are significantly correlated with transcription.

We found the integrity of HERV-H LTRs to be important to the magnitude of HERV-H transcription. In particular, transcription of HERV-H is consistently correlated with the presence of the first 114 bp (L51) of the consensus LTR. This short sequence is part of the larger U3 region and is known to contain a MYB binding site (76–80) as well as to finish with an Sp1 binding site (105–114). The positive correlation between the presence of L51 and transcription can be seen in models M1, M3 and M4. The correlation makes sense given Sp1 binding sites have been shown to be important in previous assays (Sjøttem et al., 1996; Anderssen et al., 1997).

We also found LTRs to be correlated with where HERV-H is transcribed. HERV-H loci classified as type-I were roughly twice as highly transcribed in stem cells (models M1 and M4) when compared to other loci. On the other hand we found the presence of a type-II repeat increased transcription of HERV-H loci by at least a factor of 2 (model M1 and M3) in embryonic cells. The correlation between embryonic transcription and type-II repeats is not exclusively due to type-Ia loci as the indicator variable type-Ia is not a significant factor in model M3. These results are contrary to previous

suggestions that type-II loci are inactive (Anderssen et al., 1997). It is possible that type-II loci have been thought to be inactive due to their relatively low transcription when compared to type-I HERV-H in some cell types.

The finding that type-II repeats are transcribed in embryonic cells is potentially important in light of the widespread interest in cultivating naive human cells. Mouse ESC cultures have been shown to contain sub-populations of naive-like cells that have totipotent properties (Macfarlan et al., 2012), and these totipotent cells have transcriptional characteristics that are usually associated with 2-cell embryos. Our analyses show that the type-II repeat is correlated with HERV-H transcription in early human embryonic cells. At the same time, these loci with type-II repeats do not appear to be a focus of (Wang et al., 2014) as they do not usually cluster into the highly transcribed category (Figure 5.3). As the work of Macfarlan et al. (2012) shows that the natural timing of the transcription of repeat loci is relevant for identifying naive-like cells in mouse, it is worth considering whether a similar situation is true for humans also. If so, it would be the type-II HERV-H loci that would be most relevant as totipotency markers, making it important to distinguish between type-I and type-II loci when conducting experimental work. To our knowledge such a distinction has not been made recently. In future, focussing on type-II repeat containing HERV-H might be more useful than treating all HERV-H as identical entities.

A different feature of our results is the fragmented nature of the *pol* and *env* genes of many HERV-H loci (Figures 5.1 and 5.3). It has previously been shown that the decay of *env* is a common feature of the most prolific families of retrotransposons (Magiorkinis et al., 2012) and HERV-H seems to be usual in this respect. In addition it is reasonable to assume that a more complete (longer) *pol* or *env* will contain more sequence that triggers host defences or is more likely to be involved in ectopic recombination events (Gonzalez and Petrov, 2012). In particular we find here that the P6 region of *pol* is negatively correlated with transcription to the extent that the presence of P6 is associated with a roughly 2 fold decrease in HERV-H transcription (models

M1, M3, M4). This result shows that even today more complete HERV-H loci are subject to more restricted transcription in their host.

In light of the preceding discussion it is striking that the *gag* gene is relatively complete in the majority of HERV-H and that the presence of the G4 region is positively correlated with a 48-66% increase in HERV-H transcription (models M1, M3, M4). The consensus HERV-H *gag* gene has previously been annotated and was found to contain only two transcription factor binding sites (Jern et al., 2005). Both of these sites bind zinc finger proteins (ZFPs) and are located in region G4 (1390–1431 and 1459–1497). Their presence may provide an explanation of the relationship between G4 and transcription.

There are hundreds of KRAB zinc finger binding proteins (KZNFs) coded for in the human genome (Birtle and Ponting, 2006) and their evolutionary role as mammalian suppressors of ERVs is just beginning to be described in detail (Thomas and Schneider, 2011; Lukic et al., 2014; Jacobs et al., 2014). KZNFs bind retroelement DNA with their zinc finger domain and a TRIM28 complex with their KRAB domain, and it has recently been shown that they operate in human stem cells, where they are a way for a host to control ERVs via epigenetic silencing (Turelli et al., 2014). As loci containing region G4 can be bound by host ZFPs/KZNFs we would expect any loci that retain G4 to be more co-optable. This is because loci that retain G4 possess two additional features that allow them to be more easily controlled by pre-existing mechanisms where transcription is appropriate, and more effectively silenced elsewhere. In contrast, the transcription of loci without G4 would need to be controlled via another mechanism, probably involving mutation or fragmentation followed by selection. Such a mechanism might be less plastic and would also take some time to evolve, as each individual locus would have to be modified through a series of intermediate states. As a consequence, new HERV-H loci without G4 would be more likely to be degraded or removed from the population via purifying selection some time before they could accumulate enough changes to act in a way that conferred a net benefit to

the host, and in this sense they would be less co-optable.

A synthesis of the above discussions begins to suggest a natural perspective on HERV-H that assesses its structure in three ways: first as an exogenous virus approximated by the consensus; second as a prolific endogenous retrovirus as characterized by the majority of loci in the genome; and third as strategically located and co-opted host sequence. The point of such elaboration is that it clarifies the structural features of HERV-H that we see today.

In the distant past exogenous HERV-H was a complete retrovirus adapted to horizontal transmission. As such it was characterized by a full complement of retroviral genes and a range of pathogenic effects. The traces of this history are visible in the aggregated structure of the entirety of the HERV-H loci in the human genome.

At some point roughly 35 Ma (Mager and Freeman, 1995) an endogenization of HERV-H occurred. It is possible that HERV-H then drifted to fixation. It is also plausible that HERV-H immediately provided some benefit to its primate hosts. Indeed, by assuming that ancestral HERV-H provided antiviral protection of the kind provided by enJSRV in sheep we might even explain the relatively intact nature of the *gag* gene at the majority of HERV-H loci. This is because, in the case of sheep, endogenous *gag* is thought to provide protection by competing with exogenous *gag* for receptor HYAL2 and also by interfering with exogenous *gag* during virion formation (Varela et al., 2009).

The intact nature of *gag* is in stark contrast to the fragmented *pol* and *env* genes that are the hallmarks of prolific retrotransposons. Fragmented HERV-H clearly multiplied successfully and some loci fortuitously retrotransposed into regions where, as long as they were properly controlled, they could benefit the host. The precise role of HERV-H remains elusive but involves an ability to be transcribed at the appropriate time in the appropriate cell type. If our hypothesis about control is correct then the importance of G4 would go some way towards resolving the puzzle of why HERV-H is present at a roughly 1:1 full-length to solo-LTR ratio in the human genome. This

ratio is extraordinary (Bannert and Kurth, 2006) and is not an artefact of the loci being young (see Chapter 4). In this case the subset of HERV-H that have been co-opted can be viewed in a third way: an aggregation of LTRs and the G4 region of *gag*, both embedded in an ancestral ERV packaging, the remainder of which may be of little functional consequence today. This final viewpoint happens to respect the fact that HERV-H loci have persisted in an unusual state for so many years.

5.6 Conclusion

We examine the full-length HERV-H loci in the human genome and characterize them using genomic features including LTR subtype, approximate age, and large deletions. HERV-H transcription is highest in stem cells, and lower in embryonic and somatic cells in turn.

We find that younger type-I HERV-H are the most transcribed members of the HERV-H family as they are particularly active in hESC and hiPSC cell lines. These active loci tend to be highly internally fragmented though integrity of the LTR, in particular the first 114 bp, is significantly correlated to transcription.

Although HERV-H transcription is highest in stem cells, it is still roughly six times higher in embryonic cells than somatic ones. In these embryonic cells we have shown that loci containing type-II repeats are most active. By analogy to mouse, we suggest that type-II loci may be more relevant to totipotency than the younger type-I repeats, given their earlier natural transcription.

Finally, we find that the presence of the last 441 bp of *gag* is significantly positively correlated to HERV-H transcription in both embryonic and stem cells. This is surprising and we suggest that while HERV-H LTRs drive transcription, binding sites in the last third of *gag* made some HERV-H loci more amenable to control via host silencing and consequently more likely to be first tolerated and later co-opted for a role in stem cell identity.

Chapter 6

Summary of main results and closing thoughts

Below I will summarize the results and conclusions of chapters 2–5 as well as introduce some additional points that only make sense in the context of all four of these studies. I will finish with some closing thoughts and some ideas on possible future directions for ERV research.

6.1 Female risk-factors and non-neutral ERVs

In Chapter 2, I investigated the relationship between sex, meiotic recombination, and the conversion of full-length ERVs into solo-LTRs. Creating a mathematical model allowed me to discuss this relationship as well to ask about the the origin of ERVs: did more ERV loci originate in males than females?

The motivation for these investigations was due to two questions. First, is the deeper male germ line responsible for more endogenizations than the shallower female one ? Second, does meiotic recombination play a role in promoting solo-LTR formation, and therefore limit the spread of ERVs as selfish DNA? Additionally, I wanted to examine patterns of solo-LTR formation in a wide variety of species and in a way

that circumvented shifting local recombination rates.

Having examined several animal genomes, the five results of Chapter 2 were as follows: (i) that solo-LTRs were often found at an allosomal to autosomal ratio of $\frac{2}{3}$ to $\frac{4}{3}$ in accordance with the expectations of the model; (ii) that the distribution of solo-LTR ratios suggested that ERVs do not exhibit a universally male biased integration pattern (half did, one quarter did not, one quarter were ambiguous); (iii) that full-length ERVs were often found at a ratio greater than 2, exceeding the maximum ratio expected under a deletion process strongly linked to recombination or otherwise; and (iv) that only in the case of the opossum, that has unusual X chromosome biology, was a full-length ERV ratio found to be less than 1, i.e. indicative of male bias. Taking the results relating to full-length ERVs to be more suggestive of a female bias than of a male one, a pair of conclusions were also drawn.

First, the ERV integration process is not strongly linked to the depth of the male germ line. This conclusion is based on the full-length and the solo-LTR data, and makes sense in light of wider biology for two reasons: (i) ERVs are often expressed (and therefore are likely to integrate) at a very early point in the lifecycle of a host, and the difference between the depth of male and female germ lines is simply irrelevant at this time; (ii) ERVs are known to be highly expressed in the placenta and therefore females are likely to come into contact with active ERVs due to their role in pregnancy, whereas males are not.

Second, although the full-length ratio of allosomal to autosomal ERVs is most consistent with a female integration bias, the higher than expected ratios that were observed could be due to at least three factors: (i) that full-length ERVs are not deleted quickly enough that the predicted ratio has been reached; (ii) that ERVs are non-neutral mutations and that therefore they do not drift to fixation; or (iii), that the evidence is entirely consistent with a deletion process that is strongly linked to meiotic recombination rates and that the larger than expected point estimates are statistical artefacts.

With respect to subsequent investigations from later chapters some additional conclusions can also be formulated. The results of Chapter 4 suggest that we can rule out the idea that in general ERVs are not deleted quickly enough for an equilibrium to be set up. This is because it seems that ERVs are in fact generally deleted extremely quickly. However, if it is the case that there are often large families with unusual dynamics, like HERV-H, then in future the members of these exceptional families should be censored from studies like that performed in Chapter 2. If they are not, they may dominate results in an unpredictable way.

The results of chapters 3 and 5 also have something to add, for they suggest that we cannot rule out the idea that many ERVs do not drift to fixation. Indeed, we should take the idea that a reasonable proportion of ERVs are not neutral mutations quite seriously. This is because, if we look at HERV-H, we see that full-length loci evolve differently at the nucleotide level when compared to ERV loci from other families and to other selfish DNA. We also see that HERV-H has its own, slower, deletion dynamics, and that part of the internal region of HERV-H appears to be positively associated with its transcription, and therefore presumably its function in humans. For the reasons mentioned in Chapter 2, unless some very specific facts about the fitness effects of ERVs are known, it is not possible to predict how selection would impact on the relative fixation probability of ERVs. However, this is no argument against further investigation.

Beyond the censoring of particular families, an additional methodological improvement could be made to address the sex-specific origin of ERVs in the future. This would involve taking account of the life histories of the hosts in which ERVs are studied as there may be interesting confounding factors. For example, since the research in Chapter 2 was conducted, Hayward et al. (2015) have studied ERV transmission with respect to a number of ecological variables, finding that internal fertilization is correlated with ERV abundance. A study by Katzourakis et al. (2014) has suggested that as body size increases retroviral activity decreases. Though it does not appear that

body size is a strong predictor for conventional mutational bias (Sayres et al., 2011), ecological confounds should be accounted for nevertheless.

6.2 Directional selection on highly transcribed ERVs

The speculation that a proportion of ERVs might be in symbiosis with their hosts (Chuong, 2013), as well as the knowledge that even harmful selfish DNA can fix under some circumstances (Charlesworth and Charlesworth, 1983), motivated me to investigate the evolution of orthologous ERVs from human and chimpanzee in Chapter 3. This involved categorizing paired sites from DNA alignments as one of full-length ERV, selfish DNA, or non-selfish DNA. The hypothesis of Chapter 3 was that some ERVs might evolve significantly faster (directional selection) or slower (purifying selection) than the neutral rate of evolution, as estimated by sites from the category selfish DNA.

The eleven findings of the study are as follows: (i) that there is a hierarchy of divergence, such that ERVs diverge faster than other selfish DNA, that in turn diverges faster than non-selfish DNA; (ii) that there is a hierarchy of divergence, such that the autosome diverges faster than the X chromosome; (iii) that the difference between the divergence of a typical ERV and typical selfish DNA is a statistically significant, but very small, 10^{-4} substitutions per site; (iv) that the relative divergence of an ERV is significantly positively correlated with the length of its LTRs; (v) that the relative divergence of an ERV is significantly positively correlated with the amount of neighbouring sequence that is non-selfish DNA; (vi) that potentially younger HERV-H diverge significantly faster than other ERVs; (vii) that the divergence of HERV-H is significantly positively correlated with categorical or continuous measures of its transcription in human stem cells; (viii) that the ratio of relative ERV divergence is greater for X-linked ERVs than for autosomal ones, suggesting substitutions into ERVs have a recessive effect; (ix) that the divergence of highly transcribed HERV-H implies a selection coeffi-

cient of the order of 10^{-4} that is not small; (x) that highly transcribed HERV-H are not diverging more quickly than other ERVs only because they are located in especially conserved or rapidly evolving regions of the genome; (xi) that the faster evolution of ERVs than other selfish DNA is not due to a CpG effect, but that a role for arbitrary higher order effects cannot be ruled out.

The first two findings are reassuring from a methodological standpoint, as is the finding that increased divergence is not due to a CpG effect. The following four conclusions were drawn in Chapter 3.

First, the correlation between the length of the LTRs of an ERV and the relative divergence of the ERV support the idea that substitutions into ERVs act to reduce the harmful effects of transcription (Young et al., 2013) or ectopic recombination (Campbell et al., 2014). In the former case, LTRs are known to have promotional activity (e.g. Mager, 1989) and it seems reasonable to assume that longer LTRs may be better promotors. In the latter case, longer LTRs might be more likely to ectopically recombine than shorter ones (Petrov et al., 2011). In both cases, it does not seem that an ORF is necessary in order for an ERV to have some effect on the host.

Second, the relative divergence of HERV-H loci increases with respect to the intensity of their transcription in human stem cells. This supports the notion that the higher relative divergence of the most transcribed HERV-H loci is adaptive. This is particularly true given the recent discovery of the importance of HERV-H to stem cell biology. As the selective coefficients for the most highly transcribed loci are of the order of 10^{-4} then it is reasonable to argue that the HERV-H transcriptome has recently evolved under the influence of directional selection.

Third, if directional selection is acting on highly transcribed HERV-H loci, it is not clear whether it is acting to alleviate unwanted transcription or whether it is acting to adaptively tune pre-existing host functions.

Fourth, logic suggests that the selection coefficients obtained in Chapter 3 are lower bounds on the effect of an ERV. This is for two reasons: (i) the coefficients ap-

ply to single substitutions whereas an ERV is thousands of nucleotides long; and (ii), many of the potentially harmful properties of ERVs are common to other TEs, and although other TEs were assumed to evolve neutrally, they too may exhibit accelerated divergence in some cases.

The results of chapters 4 and 5 suggest that an additional conclusion should be drawn. Chapter 4 shows that full-length HERV-H loci are deleted more slowly in primates than full-length ERVs from other families. Chapter 5 shows that internal regions of HERV-H are associated with HERV-H transcription in humans. Both these results further support the idea that full-length HERV-H is genuinely under selection. In the former case this is because it appears that full-length HERV-H may be preferentially retained in primate hosts. In the latter case this is because an internal region of HERV-H is correlated with transcription levels, and the effect of this transcription, whether good or bad, can presumably be modified via nucleotide substitution into full-length loci.

6.3 The dynamics of ERV deletion and the importance of solo-LTRs

The investigations in Chapter 4 were motivated by two issues. First, I wanted to quantitatively relate the activity of various HERV families over time, which I do not think has been done satisfactorily before. Second, I wanted to investigate the persistence of HERV loci in a full-length form, to test the hypothesis that ERVs are less likely to be deleted as they age (Belshaw et al., 2007). These questions were addressed by introducing a method to process multiple sequence alignments, and a maximum likelihood phylogenetic framework with which to interpret the resulting site patterns.

The main findings of Chapter 4 were: (i) that insertion rates obtained via sampling recover common qualitative facts from the literature e.g. the relatively greater activity of HERV-K in chimpanzee versus human or the relative timing of the major

bursts of HERV activity; (ii) that the constant hazard (exponential) model suggests that HERV-K are the most quickly deleted family and that HERV-H are the most slowly deleted family; (iii) that a variable hazard (Weibull) model suggests that HERV-H is slowly deleted, with an expected 58% chance of remaining full-length after 400,000 yr, as opposed to a 19–21% chance for loci from other families—after 25 Myr the appropriate probabilities are 31% and 5–8%, respectively; (iv) that a variable hazard model is much more appropriate than the constant hazard model according to likelihood ratio tests; (v) that simulation shows a variable hazard model is an adequate one; (vi) that MLE estimates of $\omega < 1$ demonstrate that the risk of HERV deletion decreases with HERV age, so that HERVs do "die young;" and (vii) that bootstrap replicates show that only in the case of HERV-K is there any uncertainty over the qualitative dynamics of solo-LTR formation.

In Chapter 4, six conclusions are reached. The first conclusion is that the results on HERV deletion in Chapter 4 are more detailed and generalizable than those of the previous keystone paper (Belshaw et al., 2007) in the area.

Second, an apparent speedup of HERV-K insertions is not an artefact of only considering full-length ERVs, as is claimed by Magiorkinis et al. (2015). The speedup is evident in Chapter 4 also, where solo-LTRs are considered. Therein, Figure 4.3 shows that HERV-K activity in branches specific to human or chimpanzee is higher than in the longer branch common to the human, chimpanzee and gorilla. This finding implies that taking account of solo-LTR insertions is important when interpreting the long-term activity of ERV families as doing so may change our perception of the timing of major bursts of activity. Given it has been proposed that HERV-K may restrict exogenous retroviruses (Grow et al., 2015) it is important to know at what times HERV-K has been most active in order that evolution-guided host-virus analyses (Malik, 2015) may be performed.

Third, the common application of an exponential decay process (e.g. Pereira, 2004; Lynch, 2007) is not an especially useful way to describe the preservation of ERVs

in a full-length state. This is because it gives a misleading overestimate of the duration for which an ERV can be expected to survive.

Fourth, the role of background recombination may account for the dynamics of ERV deletion to a limited extent but it seems likely that mutational divergence is a key factor. Simulation in Chapter 4 shows that if solo-LTR formation did not happen quickly then 67% of alleles would be protected by mutations upon fixation. As experimentalists have shown that a single mutation into DNA has a strong effect (Datta et al., 1997; Opperman et al., 2004) it seems as if solo-LTR formation must occur quickly if it is to occur at all.

Fifth, the markedly lower deletion probabilities obtained for HERV-H suggest that the family has been subject to long-term co-option. However, the persistence of ERVs in a full-length form is not necessarily evidence that the internal regions of the provirus are important to the host. This is because if co-opted ERVs diverge rapidly (as is demonstrated in Chapter 3) then this should reduce the probability that their LTRs undergo homologous recombination.

Sixth, it is not clear how many endogenization events involve ERVs that have already been converted to solo-LTR form, though this does not appear to be a question that is best tackled with a phylogenetic model that operates on genomic data.

6.4 The relationship between HERV-H structure and transcription

The unusual nature of HERV-H has been established in this thesis and elsewhere. However, the most exciting recent research (Lu et al., 2014; Wang et al., 2014) has failed to capitalize on several decades of fairly detailed work on the HERV-H family, including a consensus (Jern et al., 2005) and a characterization of the repeat types of HERV-H LTRs (Mager, 1989; Goodchild et al., 1993). For this reason, in Chapter 5, I asked the question: what features are characteristic of highly transcribed HERV-H loci?

The four results of this investigation were: (i) that non-phylogenetic analyses suggested LTRs have a role in determining the magnitude and specificity of transcription but that, with the exception of the 3′ end of *gag*, the presence of HERV-H genes is negatively correlated with transcription; (ii) that it was not possible to produce an informative phylogenetic regression for HERV-H transcription in somatic cells; (iii) that a phylogenetic regression of transcription in embryonic cells was informative, and suggested that the integrity of *pol* is significantly negatively correlated with HERV-H transcription, while the integrity of the 5′ end of the HERV-H LTR, a type-II repeat, and the 3′ end of *gag* are significantly positively correlated with transcription; (iv) that a phylogenetic regression of transcription in stem cells was also informative, and suggested that the integrity of *pol* is significantly negatively correlated with HERV-H transcription, while youth, the presence of the 5′ end of the HERV-H LTR, a type-I LTR, and the 3′ end of *gag* are significantly positively correlated with transcription.

These results lead to five conclusions, the first of which is that the failure to obtain a useful phylogenetic regression for somatic transcription seems to indicate the somatic transcription of HERV-H is noise or is idiosyncratic, for there is no evidence or a priori reason to believe that HERV-H has a role in somatic tissue. This position is supported by the fact that somatic expression is very much lower than HERV-H expression in embryonic cells or stem cells (e.g. Wang et al., 2014; Göke et al., 2015).

Second, the positive correlation between HERV-H transcription and the 5′ region of its LTRs makes sense given prior knowledge of the importance of LTR integrity for ERV transcription and the fact that this region contains binding sites that have been shown to be important to transcription in previous assays by experimentalists (Sjøttem et al., 1996; Anderssen et al., 1997).

Third, the HERV-H LTR is important in establishing where a HERV-H is transcribed. In particular, a type-II repeat doubles transcription in embryonic cells, all other factors being equal. This is contrary to previous suggestions (Anderssen et al., 1997) that type-II loci are inactive, as the increase is not specific to type-Ia LTRs.

Fourth, knowledge that type-II repeat containing HERV-H are especially active in embryonic cells may be important to researchers in the stem cell community. This is because totipotent cells in mice resemble early stage embryos (Macfarlan et al., 2012), yet type-II loci appear to be overlooked in recent HERV-H research because loci of this kind do not usually cluster into the highly transcribed category (Wang et al., 2014).

This suggestion is circumstantially supported by the fact that although the most highly transcribed HERV-H are younger type-I, it seems a priori unlikely that the co-option of HERV-H is a very recent event, considering the tens of Myr over which many HERV-H have been maintained in a full-length form. If HERV-H co-option is not a very recent event, then some older and less transcribed loci must be important, indicating the absolute transcription level of a locus is not the sole factor that should be used to judge whether it is important to the host or not.

Fifth, the positive correlation between the 3' region of *gag* and transcription may be due to the presence of two zinc finger protein binding sites in this region of the gene (Jern et al., 2005). These binding sites might act as "handles" that allow a host to more easily control transcription where appropriate and more completely silence it elsewhere. This may contribute to an explanation of why HERV-H is more often retained in full-length form when compared to loci from other HERV families.

6.5 Closing thoughts and future research

In this thesis I have argued that the chromosomal distribution of ERVs suggests that there are female specific risk factors to endogenization or that ERVs are under selection. I have shown that selection has acted on highly transcribed HERV-H. I have also shown that HERVs tend to be deleted very quickly, and that HERV-H has unusual dynamics among the larger HERV families. Finally, I have shown that the repeat structure of HERV-H affects the cell types it is transcribed in, and have argued that a region of *gag* may be important to the preservation of HERV-H in a full-length state.

Significant HERV research is accumulating rapidly and I expect this to remain the case for several years. When beginning my D.Phil. it seemed natural to assume that most ERVs drifted to fixation, yet my research, and that of others, suggests otherwise: given the fact that many loci from a large family of ERVs in the human genome have turned out to be important to host biology it would be reckless to assume that similar situations do not occur in many other species.

We are quickly accumulating knowledge about the relationship between ERVs and early human cell types. For example, this year researchers found that HERV-K loci were expressed from the 8-cell to the blastocyst stage, resulting in the accumulation of viral particles, Gag and Rec, the overexpression of which was shown to inhibit viral infection (Grow et al., 2015). We should not be surprised that the immune system responds to the expression of viral genes, nor that the youngest ERVs in the human genome are imperfectly silenced, yet such findings do suggest that there are further discoveries to be made about even the most scrutinized families of ERVs.

Similarly, research on HERV-H is also progressing quickly. For example, a recent meta-analysis of single-cell RNA-seq data was conducted and researchers were able to show that different HERV-H loci were clearly active at different stages of development (Göke et al., 2015). The conclusion of the authors was that ERVs are stage-specific regulatory elements in humans. Given the medical importance of both retroviral and stem cell research, it seems safe to assume that HERVs will be continually studied for some time to come. Speculatively, I would suggest that determining the specific importance of retrovirally derived regulatory elements in humans will eventually lead to insights into the origin of evolutionary innovation that will apply to biology as a whole.

As well as learning about ERVs themselves, we are also learning more about the mechanisms that control them in primates. Mammalian genomes contain hundreds of genes coding for KRAB zinc finger proteins (KZNFs) that have duplicated since their origin in early tetrapod vertebrates (Birtle and Ponting, 2006). Bioinformatics studies

have suggested that the diversity of these KZNF proteins is due to their role as a host defence mechanism against retroelements (Thomas and Schneider, 2011; Lukic et al., 2014). A recent study provided a wonderful example of how such diversity could be generated by showing that, in fact, changes that had occurred roughly 8–12 Ma enabled the repression of SVA elements by ZNF91, and further, that ZNF93 was able to repress L1 LINE elements until roughly 12.1 Ma, whereupon the L1PA3 subfamily managed to escape repression after a deletion removed its KZNF binding site (Jacobs et al., 2014). Additionally, recent experimental research demonstrated that KZNFs silence a diversity of retroelements, such as HERV-K, in human embryonic stem cells (Turelli et al., 2014). In future I think it will be possible to determine the functional role of HERV-H *gag* in the control of HERV-H loci by using a similar method to that deployed by Jacobs et al. (2014), so that G4 variant HERV-H loci (Chapter 5) and a reporter could be co-expressed with appropriate KZNFs in transgenic mouse embryonic stem cells. This kind of experiment could demonstrate an empirical link between host retroviral defences and co-opted retroviral sequences. The fact that different HERV-H loci are expressed at distinct developmental stages (Göke et al., 2015) would only aid in identifying the appropriate KZNF/HERV-H combinations to examine.

While experiments are important, ultimately I think it is population data that will more often be used to answer questions about ERV dynamics. A problem with such an approach is identifying species that are currently undergoing ERV invasion. One potential model species is koala, which is currently experiencing endogenizations of Koala Retrovirus or KoRV (Tarlinton et al., 2006). There is geographic structure to the infections—KoRV is spreading from north to south through the population—and in addition the infection appears to be recent, as an artificially introduced but isolated population of koalas on Kangaroo Island do not possess the virus (Stoye, 2006). This year Ishida et al. (2015) studied polymorphic KoRV loci in ten individuals and found 7 LTR haplotypes originating from 10 insertions. Each insertion had identical paired LTRs and all insertions were heterozygous (Ishida et al., 2015). Together these findings

are consistent with the notion that KoRV is recent. However, in some ways this study is disappointing, because with additional analyses so much more could be learned about the early stages of ERV invasion. Indeed, population data could certainly contribute to several questions tackled in this thesis. For example, low frequency ERVs could provide information on sex-specific risk factors (Chapter 2) while the frequency distribution of ERVs would contribute to an assessment of the fitness cost of an individual insertion (Chapter 3). Information on how often novel ERV integrations are found in solo-LTR form would help clarify our understanding of the origin of many HERV loci and also permit refinements of the results of Chapter 4. It would be especially interesting to know whether ERV insertions are most harmful in heterozygous form, something that might be expected if ectopic recombination often determines the fate of polymorphic proviruses. Any findings could be generalized to other mammals and might therefore provide the best insight yet into retroelement population dynamics in animals beyond insects.

Among the TEs the ERVs stand out as being closest in structure to exogenous viruses, and I think acknowledging this fact is important going forward. Here population data is again useful as it can be used to investigate the relationship between the loss of exogenous retroviral functionality and ERV fixation and proliferation. It is my suspicion that sequence that is completely representative of exogenous viruses may never fix in full-length form and that the vast majority of full-length ERVs are descendants of lineages which quickly became less viral-like. Of course, this is speculation, but there is some evidence that many proviruses will be defective. For example, a recent study of the latent reservoir of 8 HIV patients showed that of 213 proviruses, one-third contained hypermutations that induced codon changes while nearly half had large internal deletions (Ho et al., 2013). Therefore, even if there was no selective cost associated with proviruses one might expect roughly half of all ERVs to have a large internal deletion due to chance alone. In reality, many of the deletions described by Ho et al. (2013) would actually leave proviruses in a more benign state and so it

is plausible that such proviruses would be more likely to be transmitted in an endogenous form. Population data is necessary to address these hypotheses further and perhaps in future it will be possible to test these ideas via capture and release studies on wild koalas.

By now there is considerable evidence that TEs, including ERVs, are responsible for providing important regulatory elements on a large scale. Beyond the work previously mentioned in this thesis, examples include the findings of Wang et al. (2007) who have shown that one-third of all p53 binding sites in human are due to LTR10 and MER61 TEs; Lynch et al. (2011) who found that perhaps 13% of genes recruited to endometrial cells in placental mammals were recruited by the MER20 TE family; and Kapusta et al. (2013) who found that TEs have been major contributors to the diversity of long non-coding RNAs in a range of species. A recent review paper described the situation with respect to ERVs well: retroviruses are "evolving regulatory elements" that are active only when certain cell specific or other biological conditions are met (Schlesinger and Goff, 2015).

The difficulty, of course, is the lack of a sufficiently precise way to describe the costs, benefits and macro-coevolution of host and regulatory elements. Without a mathematical framework hypotheses such as those of Chuong (2013)—who suggested that ERVs and the placenta may be in symbiosis—run the risk of sounding woolly and ill-conceived. Clearly the same clarity of thought and rigour that has been applied when explaining the existence and proliferation of selfish DNA must now be applied to hypotheses on the relationship between TEs and novel regulatory networks. Though population genetics does not seem to be able to help, mathematics in general can, and there is at least one promising direction for future research in the form of spin-offs of computational complexity theory (Valiant, 2009; Livnat et al., 2014).

In Chapter 3, I proposed that HERV-H be used as a model system for studying ERV exaptation in mammals. I stand by this proposal, and suggest that I have started to address it with the work in chapters 4 and 5. I hope and expect that other researchers

will do so too. In hindsight, I would like to suggest that the signs that HERV-H has been exapted are obvious. In fact, despite the tremendous importance of experimental advances, I think this is clear from bioinformatics evidence alone: as a subset of highly transcribed HERV-H have diverged quickly, and as HERV-H is deleted slowly, it is clear that the family is unusual. *Syncytin* research progressed rapidly after human specific evidence encouraged a broader search for envelope genes with open reading frames. Consider that in the year 2000 the *syncytins* were largely hypothetical yet by 2011 their functionality had been demonstrated in mice, and by 2014 co-options were found to have occurred in eight different lineages. This progress was possible because bioinformatics helped identify candidate genes to examine in experimental assays. By analogy, I suggest it would be beneficial to search genomes that are distant from primates for further HERV-H like exaptations. In particular, I think that if an assessment of deletion rates identifies a preferentially retained ERV family that is also particularly conserved or rapidly evolving, then this would be good evidence that ERVs from that family have been subject to exaptation also.

Chapter 7

Bibliography

G. Abrusán and H.J. Krambeck. Competition may determine the diversity of transposable elements. *Theoretical Population Biology*, 70(3):364–375, 2006.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, David J Lipman, et al. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Sølvi Anderssen, Eva Sjøttem, Gunbjørg Svineng, and Terje Johansen. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. *Virology*, 234(1):14–30, 1997.

N. Bannert and R. Kurth. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.*, 7:149–173, 2006.

Birke Bartosch, Dimitrios Stefanidis, Richard Myers, Robin Weiss, Clive Patience, and Yasuhiro Takeuchi. Evidence and consequence of porcine endogenous retrovirus recombination. *Journal of Virology*, 78(24):13880–13890, 2004.

R. Belshaw, J. Watson, A. Katzourakis, A. Howe, J. Woolven-Allen, A. Burt, and M. Tristem. Rate of recombinational deletion among human endogenous retroviruses. *Journal of Virology*, 81(17):9437–9442, 2007.

Robert Belshaw, Vini Pereira, Aris Katzourakis, Gillian Talbot, Jan Pačes, Austin Burt, and Michael Tristem. Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences*, 101(14):4894–4899, 2004.

Robert Belshaw, Anna LA Dawson, John Woolven-Allen, Joanna Redding, Austin Burt, and Michael Tristem. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): implications for present-day activity. *Journal of virology*, 79(19):12507–12514, 2005a.

Robert Belshaw, Aris Katzourakis, Jan Pačes, Austin Burt, and Michael Tristem. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular biology and evolution*, 22(4):814–817, 2005b.

Farid Benachenhou, Göran O Sperber, Erik Bongcam-Rudloff, Göran Andersson, Jef D Boeke, and Jonas Blomberg. Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). *Mobile DNA*, 4(1):1–16, 2013.

Laurence Bénit, Philippe Dessen, and Thierry Heidmann. Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. *Journal of Virology*, 75(23):11709–11719, 2001.

Sheng Bi, Oksana Gavrilova, Da-Wei Gong, Mark M Mason, and Marc Reitman. Identification of a placental enhancer for the human leptin gene. *Journal of Biological Chemistry*, 272(48):30583–30588, 1997.

Zoë Birtle and Chris P Ponting. Meisetz and the birth of the KRAB motif. *Bioinformatics*, 22(23):2841–2845, 2006.

Sandra Blaise, Nathalie de Parseval, Laurence Bénit, and Thierry Heidmann. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proceedings of the National Academy of Sciences*, 100(22):13013–13018, 2003.

Jonas Blomberg, Farid Benachenhou, Vidar Blikstad, Göran Sperber, and Jens Mayer. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene*, 448(2):115–123, 2009.

Jean-Luc Blond, Frédéric Besème, Laurent Duret, Olivier Bouton, Frédéric Bedin, Hervé Perron, Bernard Mandrand, and François Mallet. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. *Journal of Virology*, 73(2):1175–1185, 1999.

Jean-Luc Blond, Dimitri Lavillette, Valérie Cheynet, Olivier Bouton, Guy Oriol, Sylvie Chapel-Fernandes, Bernard Mandrand, François Mallet, and François-Loïc Cosset. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *Journal of Virology*, 74(7):3321–3329, 2000.

Stephane Boissinot, Jerel Davis, Ali Entezam, Dimitri Petrov, and Anthony V Furano. Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences*, 103(25):9590–9594, 2006.

Mohan Bolisetty, Jonas Blomberg, Farid Benachenhou, Göran Sperber, and Karen Beemon. Unexpected diversity and expression of avian endogenous retroviruses. *MBio*, 3(5):e00344–12, 2012.

John FY Brookfield. The ecology of the genome—mobile DNA elements and their hosts. *Nature Reviews Genetics*, 6(2):128–136, 2005.

John FY Brookfield and Richard M Badge. Population genetics models of transposable elements. In *Evolution and Impact of Transposable Elements*, pages 281–294. Springer, 1997.

Terence A Brown. *Genomes*. Garland Science, 2006.

Ian M Campbell, Tomasz Gambin, Piotr Dittwald, Christine R Beck, Andrey Shuvarikov, Patricia Hixson, Ankita Patel, Anna Gambin, Chad A Shaw, Jill A Rosenfeld, et al. Human endogenous retroviral elements promote genome instability via nonallelic homologous recombination. *BMC Biology*, 12(1): 74, 2014.

B Charlesworth and D Charlesworth. The population dynamics of transposable elements. *Genetical Research*, 42:1–27, 1983.

B Charlesworth and CH Langley. The population genetics of *Drosophila* transposable elements. *Annual Review of Genetics*, 23(1):251–287, 1989.

B Charlesworth, JA Coyne, and NH Barton. The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist*, pages 113–146, 1987.

B Charlesworth, P Sniegowski, and W Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371(6494):215–220, 1994.

Brian Charlesworth and Deborah Charlesworth. *Elements of evolutionary genetics*. Roberts and Company Publishers, Greenwood Village, 2010.

Amanda Y Chong, Kenji K Kojima, Jerzy Jurka, David A Ray, Arian FA Smit, Sally R Isberg, and Jaime Gongora. Evolution and gene capture in ancient endogenous retroviruses—insights from the crocodilian genomes. *Retrovirology*, 11(1):71–71, 2014.

Edward B Chuong. Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays*, 35 (10):853–861, 2013.

Guillaume Cornelis, Odile Heidmann, Sibylle Bernard-Stoecklin, Karine Reynaud, Géraldine Véron, Baptiste Mulot, Anne Dupressoir, and Thierry Heidmann. Ancestral capture of *syncytin-Car1*, a fusogenic endogenous retroviral *envelope* gene involved in placentation and conserved in Carnivora. *Proceedings of the National Academy of Sciences*, 109(7):E432–E441, 2012.

Guillaume Cornelis, Odile Heidmann, Séverine A Degrelle, Cécile Vernochet, Christian Lavialle, Claire Letzelter, Sibylle Bernard-Stoecklin, Alexandre Hassanin, Baptiste Mulot, Michel Guillomot, et al. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proceedings of the National Academy of Sciences*, 110(9):E828–E837, 2013.

Guillaume Cornelis, Cécile Vernochet, Quentin Carradec, Sylvie Souquere, Baptiste Mulot, François Catzeflis, Maria A Nilsson, Brandon R Menzies, Marilyn B Renfree, Gérard Pierron, et al. Retroviral envelope gene captures and *syncytin* exaptation for placentation in marsupials. *Proceedings of the National Academy of Sciences*, 112(5):E487–E496, 2015.

Abhijit Datta, Miyono Hendrix, Marc Lipsitch, and Sue Jinks-Robertson. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. *Proceedings of the National Academy of Sciences*, 94(18):9757–9762, 1997.

Matthew D Daugherty and Harmit S Malik. Rules of engagement: molecular insights from host-virus arms races. *Annual Review of Genetics*, 46:677–700, 2012.

WF Doolittle and C Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603, 1980.

Anne Dupressoir, Geoffroy Marceau, Cécile Vernochet, Laurence Bénit, Colette Kanellopoulos, Vincent Sapin, and Thierry Heidmann. Syncytin-a and syncytin-b, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proceedings of the National Academy of Sciences*, 102(3):725–730, 2005.

Anne Dupressoir, Cécile Vernochet, Olivia Bawa, Francis Harper, Gérard Pierron, Paule Opolon, and Thierry Heidmann. Syncytin-a knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences*, 106(29):12127–12132, 2009.

Anne Dupressoir, Cécile Vernochet, Francis Harper, Justine Guégan, Philippe Dessen, Gérard Pierron, and Thierry Heidmann. A pair of co-opted retroviral envelope *syncytin* genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proceedings of the National Academy of Sciences*, 108(46):E1164–E1173, 2011.

Laurent Duret, Gabriel Marais, and Christian Biémont. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics*, 156(4): 1661–1669, 2000.

Sean R Eddy. The C-value paradox, junk DNA and ENCODE. *Current Biology*, 22(21):R898–R899, 2012.

Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

David Ellinghaus, Stefan Kurtz, and Ute Willhoeft. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1):18, 2008.

Cécile Esnault, Stéphane Priet, David Ribet, Cécile Vernochet, Thomas Bruls, Christian Lavialle, Jean Weissenbach, and Thierry Heidmann. A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. *Proceedings of the National Academy of Sciences*, 105(45):17532–17537, 2008.

Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, et al. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012.

Alexandre Fort, Kosuke Hashimoto, Daisuke Yamada, Md Salimullah, Chaman A Keya, Alka Saxena, Alessandro Bonetti, Irina Voineagu, Nicolas Bertin, Anton Kratz, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, 46(6):558–566, 2014.

Alevtina Gall, Piper Treuting, Keith B Elkon, Yueh-Ming Loo, Michael Gale, Glen N Barber, and Daniel B Stetson. Autoimmunity initiates in nonhematopoietic cells and progresses via lymphocytes in an interferon-dependent autoimmune disease. *Immunity*, 36(1):120–131, 2012.

László Zsolt Garamszegi. *Modern phylogenetic comparative methods and their application in evolutionary biology*. Springer, 2014.

Patrick Gemmell, Jotun Hein, and Aris Katzourakis. Sex-specific aspects of endogenous retroviral insertion and deletion. *BMC Evolutionary Biology*, 13(1):243, 2013.

Patrick Gemmell, Jotun Hein, and Aris Katzourakis. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. *Retrovirology*, 40320(5040):720, 2015.

Robert Gifford and Michael Tristem. The evolution, distribution and diversity of endogenous retroviruses. *Virus genes*, 26(3):291–315, 2003.

Jonathan Göke, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, Lam-Ha Ly, Friedrich Sachs, and Iwona Szczerbinska. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell stem cell*, 16(2):135–141, 2015.

Michael C Golding, Liyue Zhang, and Mellissa RW Mann. Multiple epigenetic modifiers induce aggressive viral extinction in extraembryonic endoderm stem cells. *Cell Stem Cell*, 6(5):457–467, 2010.

Josefa Gonzalez and Dmitri A Petrov. Evolution of genome content: population dynamics of transposable elements in flies and humans. In *Evolutionary Genomics*, pages 361–383. Springer, 2012.

Nancy L Goodchild, David A Wilkinson, and Dixie L Mager. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology*, 196(2):778–788, 1993.

Alan Grafen. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, pages 119–157, 1989.

Dan Graur, Yichen Zheng, and Ricardo BR Azevedo. An evolutionary classification of genomic function. *Genome Biology and Evolution*, 7(3):642–645, 2015.

Edward J Grow, Ryan A Flynn, Shawn L Chavez, Nicholas L Bayless, Mark Wossidlo, Daniel J Wesche, Lance Martin, Carol B Ware, Catherine A Blish, Howard Y Chang, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, 522(7555):221–225, 2015.

David Haig. Genetic conflicts in human pregnancy. *Quarterly Review of Biology*, pages 495–532, 1993.

David Haig. Retroviruses and the placenta. *Current Biology*, 22(15):R609–R613, 2012.

David Haig. Genomic vagabonds: Endogenous retroviruses and placental evolution (comment on doi 10.1002/bies.201300059). *BioEssays*, 35(10):845–846, 2013.

Alexander Hayward, Manfred Grabherr, and Patric Jern. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proceedings of the National Academy of Sciences*, 110(50):20146–20151, 2013.

Alexander Hayward, Charlie K Cornwallis, and Patric Jern. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences*, 112(2):464–469, 2015.

Odile Heidmann, Cécile Vernochet, Anne Dupressoir, and Thierry Heidmann. 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):107, 2009.

D.A. Hickey. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101(3-4):519–531, 1982.

Ya-Chi Ho, Liang Shan, Nina N Hosmane, Jeffrey Wang, Sarah B Laskey, Daniel IS Rosenbloom, Jun Lai, Joel N Blankson, Janet D Siliciano, and Robert F Siliciano. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell*, 155(3):540–551, 2013.

Jennifer F Hughes and John M Coffin. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature Genetics*, 29(4):487–489, 2001.

Jennifer F Hughes and John M Coffin. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics*, 171(3):1183–1194, 2005.

Enrique Ibarra-Laclette, Eric Lyons, Gustavo Hernández-Guzmán, Claudia Anahí Pérez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, Andreanna J Welch, María Jazmín Abraham Juárez, June Simpson, et al. Architecture and evolution of a minute plant genome. *Nature*, 498 (7452):94–98, 2013.

Luke Isbel and Emma Whitelaw. Endogenous retroviruses in mammals: an emerging picture of how ERVs modify expression of adjacent genes. *Bioessays*, 34(9):734–738, 2012.

Yasuko Ishida, Kai Zhao, Alex D Greenwood, and Alfred L Roca. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. *Molecular Biology and Evolution*, 32(1):109–120, 2015.

FM Jacobs, D Greenberg, N Nguyen, M Haeussler, AD Ewing, S Katzman, B Paten, SR Salama, and D Haussler. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, 516(7530):242, 2014.

Patric Jern and John M Coffin. Effects of retroviruses on host genome function. *Annual Review of Genetics*, 42:709–732, 2008.

Patric Jern, Göran O Sperber, and Jonas Blomberg. Definition and variation of human endogenous retrovirus H. *Virology*, 327(1):93–110, 2004.

Patric Jern, Göran O Sperber, Göran Ahlsén, and Jonas Blomberg. Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H. *Journal of Virology*, 79(10):6325–6337, 2005.

Aashish R Jha, Satish K Pillai, Vanessa A York, Elizabeth R Sharp, Emily C Storm, Douglas J Wachter, Jeffrey N Martin, Steven G Deeks, Michael G Rosenberg, Douglas F Nixon, et al. Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before *Homo sapiens*. *Molecular Biology and Evolution*, 26(11):2617–2626, 2009.

LJ Johnson. The genome strikes back: the evolutionary importance of defence against mobile elements. *Evolutionary Biology*, 34(3):121–129, 2007.

I King Jordan, Lilya V Matyunina, and John F McDonald. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. *Proceedings of the National Academy of Sciences*, 96(22):12621–12625, 1999.

Christine Kamp, Peter Hirschmann, Hartmut Voss, Karin Huellen, and Peter H Vogt. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. *Human Molecular Genetics*, 9(17):2563–2572, 2000.

Aurélie Kapusta, Zev Kronenberg, Vincent J Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cédric Feschotte. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding rnas. *PLoS Genetics*, 9(4), 2013.

A Katzourakis and M Tristem. Phylogeny of human endogenous and exogenous retroviruses. In *Retroviruses and Primate Genome Evolution*, pages 186–203. Landes Bioscience, Georgetown, 2005.

A Katzourakis, A Rambaut, and OG Pybus. The evolutionary dynamics of endogenous retroviruses. *Trends in Microbiology*, 13(10):463–468, 2005.

A Katzourakis, V Pereira, and M Tristem. Effects of recombination rate on human endogenous retrovirus fixation and persistence. *Journal of Virology*, 81(19):10712–10717, 2007a.

A Katzourakis, M Tristem, OG Pybus, and RJ Gifford. Discovery and analysis of the first endogenous lentivirus. *Proceedings of the National Academy of Sciences*, 104(15):6261–6265, 2007b.

Aris Katzourakis. Paleovirology: inferring viral evolution from host genome sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1626):20120493, 2013.

Aris Katzourakis, Gkikas Magiorkinis, Aaron G Lim, Sunetra Gupta, Robert Belshaw, and Robert Gifford. Larger mammalian body size leads to lower retroviral activity. *PLoS Pathogens*, 10(7), 2014.

Galih Kunarso, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics*, 42(7):631–634, 2010.

Björn Lamprecht, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido Lenze, Karl Köchert, Mohamed Amine Bouhlel, Julia Richter, Eric Soler, et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature medicine*, 16 (5):571–579, 2010.

Charles H Langley, John FY Brookfield, and Norman Kaplan. Transposable elements in Mendelian populations. I. A theory. *Genetics*, 104(3):457–471, 1983.

Christian Lavialle, Guillaume Cornelis, Anne Dupressoir, Cécile Esnault, Odile Heidmann, Cécile Vernochet, and Thierry Heidmann. Paleovirology of '*syncytins*', retroviral *env* genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1626):20120507, 2013.

A Le Rouzic and P Capy. The first steps of transposable elements invasion parasitic strategy vs. genetic drift. *Genetics*, 169(2):1033–1043, 2005.

A Le Rouzic and P Capy. Population genetics models of competition between transposable element subfamilies. *Genetics*, 174(2):785–793, 2006.

A Le Rouzic, S Dupas, and P Capy. Genome ecosystem and transposable elements species. *Gene*, 390 (1):214–220, 2007.

Wen-Hsiung Li, Soojin Yi, and Kateryna Makova. Male-driven evolution. *Current Opinion in Genetics & Development*, 12(6):650–656, 2002.

Stefan Linquist, Brent Saylor, Karl Cottenie, Tyler A Elliott, Stefan C Kremer, and T Ryan Gregory. Distinguishing ecological from evolutionary approaches to transposable elements. *Biological Reviews*, 88(3):573–584, 2013.

Adi Livnat, Christos Papadimitriou, Aviad Rubinstein, Gregory Valiant, and Andrew Wan. Satisfiability and evolution. In *Foundations of Computer Science (FOCS)*, pages 524–530. IEEE, 2014.

Xinyi Lu, Friedrich Sachs, LeeAnn Ramsay, Pierre-Étienne Jacques, Jonathan Göke, Guillaume Bourque, and Huck-Hui Ng. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, 21(4):423–425, 2014.

S Lukic, JC Nicolas, and AJ Levine. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death & Differentiation*, 21(3):381–387, 2014.

Michael Lynch. *The origins of genome architecture*. Sinauer Associates Sunderland, 2007.

Vincent J Lynch, Robert D Leclerc, Gemma May, and Günter P Wagner. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, 43(11):1154–1159, 2011.

Todd S Macfarlan, Wesley D Gifford, Shawn Driscoll, Karen Lettieri, Helen M Rowe, Dario Bonanomi, Amy Firth, Oded Singer, Didier Trono, and Samuel L Pfaff. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405):57–63, 2012.

Dixie L Mager. Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. *Virology*, 173(2):591–599, 1989.

Dixie L Mager and J Douglas Freeman. HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage. *Virology*, 213(2):395–404, 1995.

G Magiorkinis, RJ Gifford, A Katzourakis, J De Ranter, and R Belshaw. *Env*-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences*, 109(19):7385–7390, 2012.

Gkikas Magiorkinis, Daniel Blanco-Melo, and Robert Belshaw. The decline of human endogenous retroviruses: extinction and survival. *Retrovirology*, 12(1):8, 2015.

Harmit Malik. Id: 253: Evolution of antiviral factors in primates. *Cytokine*, 76(1):59, 2015.

Harmit S Malik, Steve Henikoff, and Thomas H Eickbush. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research*, 10(9):1307–1318, 2000.

Harmit Singh Malik. Retroviruses push the envelope for mammalian placentation. *Proceedings of the National Academy of Sciences*, 109(7):2184–2185, 2012.

François Mallet, Olivier Bouton, Sarah Prudhomme, Valérie Cheynet, Guy Oriol, Bertrand Bonnaud, Gérard Lucotte, Laurent Duret, and Bernard Mandrand. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proceedings of the National Academy of Sciences*, 101(6):1731–1736, 2004.

Emanuele Marchi, Alex Kanapin, Gkikas Magiorkinis, and Robert Belshaw. Unfixed endogenous retroviral insertions in the human population. *Journal of Virology*, 88(17):9529–9537, 2014.

Helena Mata, Jaime Gongora, Eduardo Eizirik, Brunna M Alves, Marcelo A Soares, and Ana P Ravazzolo. Identification and characterization of diverse groups of endogenous retroviruses in felids. *Retrovirology*, 12(1):26, 2015.

Sha Mi, Xinhua Lee, Xiang-Ping Li, Geertruida M Veldman, Heather Finnerty, Lisa Racie, Edward Lavallie, Xiang-Yang Tang, Philippe Edouard, Steve Howes, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771):785–789, 2000.

Tarjei S Mikkelsen, Matthew J Wakefield, Bronwen Aken, Chris T Amemiya, Jean L Chang, Shannon Duke, Manuel Garber, Andrew J Gentles, Leo Goodstadt, Andreas Heger, et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature*, 447(7141):167–177, 2007.

T Miyata, H Hayashida, K Kuma, K Mitsuyasu, and T Yasunaga. Male-driven molecular evolution: a model and nucleotide sequence analysis. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, pages 863–867. Cold Spring Harbor Lab, 1987.

S Myers, L Bottolo, C Freeman, G McVean, and P Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

C. Nellåker, T.M. Keane, B. Yalcin, K. Wong, A. Agam, T.G. Belgard, J. Flint, D.J. Adams, W.N. Frankel, and C.P. Ponting. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology*, 13(6):R45, 2012.

Björn Nystedt, Nathaniel R Street, Anna Wetterbom, Andrea Zuccolo, Yao-Cheng Lin, Douglas G Scofield, Francesco Vezzi, Nicolas Delhomme, Stefania Giacomello, Andrey Alexeyenko, et al. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451):579–584, 2013.

Roy Opperman, Eyal Emmanuel, and Avraham A Levy. The effect of sequence divergence on recombination between direct repeats in arabidopsis. *Genetics*, 168(4):2207–2215, 2004.

LE Orgel and FH Crick. Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604, 1980.

Jan Paces, Adam Pavlícek, and Václav Paces. HERVd: database of human endogenous retroviruses. *Nucleic Acids Research*, 30(1):205–206, 2002.

Mark Pagel. Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, 26(4):331–348, 1997.

Emmanuel Paradis. *Analysis of Phylogenetics and Evolution with R*. Springer Science & Business Media, 2011.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

Vini Pereira. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biology*, 5(10):R79, 2004.

Vini Pereira. Automated paleontology of repetitive DNA with REANNOTATE. *BMC Genomics*, 9(1): 614, 2008.

Dmitri A Petrov, Anna-Sophie Fiston-Lavier, Mikhail Lipatov, Kapa Lenkov, and Josefa González. Population genomics of transposable elements in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 28(5):1633–1644, 2011.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Kristy Red-Horse, Yan Zhou, Olga Genbacev, Akraporn Prakobphol, Russell Foulk, Michael McMaster, and Susan J Fisher. Trophoblast differentiation during embryo implantation and formation of the maternal-fetal interface. *Journal of Clinical Investigation*, 114(6):744, 2004.

François Redelsperger, Guillaume Cornelis, Cécile Vernochet, Bud C Tennant, François Catzeflis, Baptiste Mulot, Odile Heidmann, Thierry Heidmann, and Anne Dupressoir. Capture of *syncytin-Mar1*,

a fusogenic endogenous retroviral envelope gene involved in placentation in the rodentia squirrel-related clade. *Journal of Virology*, 88(14):7915–7928, 2014.

Peter Rice, Ian Longden, Alan Bleasby, et al. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

William R Rice. Sex chromosomes and the evolution of sexual dimorphism. *Evolution*, 38(4):735–742, 1984.

Carène Rizzon, Gabriel Marais, Manolo Gouy, and Christian Biémont. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research*, 12(3): 400–407, 2002.

Luisa Robbez-Masson and Helen M Rowe. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology*, 12(1):45, 2015.

TaiYun Roe, Thomas C Reynolds, G Yu, and PO Brown. Integration of murine leukemia virus DNA depends on mitosis. *The EMBO Journal*, 12(5):2099, 1993.

Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2015.

Helen M Rowe and Didier Trono. Dynamic control of endogenous retroviruses during development. *Virology*, 411(2):273–287, 2011.

Diane J Rowold and Rene J Herrera. Alu elements and the human genome. *Genetica*, 108(1):57–72, 2000.

Janet P Sanford, Verne M Chapman, and Janet Rossant. DNA methylation in extraembryonic lineages of mammals. *Trends in Genetics*, 1:89–93, 1985.

Melissa A Wilson Sayres, Chris Venditti, Mark Pagel, and Kateryna D Makova. Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution*, 65(10):2800–2815, 2011.

Sharon Schlesinger and Stephen P Goff. Retroviral transcriptional regulation and embryonic stem cells: war and peace. *Molecular and Cellular Biology*, 35(5):770–777, 2015.

Anke M Schulte, Shoupeng Lai, Andreas Kurtz, Frank Czubayko, Anna T Riegel, and Anton Wellstein. Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable

to germ-line insertion of an endogenous retrovirus. *Proceedings of the National Academy of Sciences*, 93 (25):14759–14764, 1996.

E Sjøttem, Solvi Anderssen, and Terje Johansen. The promoter activity of long terminal repeats of the HERV-H family of human retrovirus-like elements is critically dependent on Sp1 family proteins interacting with a GC/GT box located immediately 3'to the TATA box. *Journal of Virology*, 70(1): 188–198, 1996.

AFA Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0. 1996–2004. *Institute for Systems Biology*, 2004.

Zachary D Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

Pawel Stankiewicz and James R Lupski. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82, 2002.

Sascha Steinbiss, Ute Willhoeft, Gordon Gremme, and Stefan Kurtz. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, 37(21):7002–7013, 2009.

Michael E Steiper and Nathan M Young. Primate molecular divergence dates. *Molecular Phylogenetics and Evolution*, 41(2):384–394, 2006.

Jonathan P Stoye. Koala retrovirus: a genome invasion in real time. *Genome Biology*, 7(11):241, 2006.

Ravi P Subramanian, Julia H Wildschutte, Crystal Russo, and John M Coffin. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, 8(1):1–22, 2011.

Chao Sun, Helen Skaletsky, Steve Rozen, Jörg Gromoll, Eberhard Nieschlag, Robert Oates, and David C Page. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Human Molecular Genetics*, 9(15):2291–2296, 2000.

Youichi Suzuki and Robert Craigie. The road to chromatin: nuclear entry of retroviruses. *Nature Reviews Microbiology*, 5(3):187–196, 2007.

Eugene D Sverdlov. Retroviruses and primate evolution. *Bioessays*, (22):161–71, 2000.

Rachael E Tarlinton, Joanne Meers, and Paul R Young. Retroviral invasion of the koala genome. *Nature*, 442(7098):79–81, 2006.

Michael Thain, Michael Hickman, et al. *Penguin dictionary of biology*. Penguin Books, 2004.

James H Thomas and Sean Schneider. Coevolution of retroelements and tandem zinc finger genes. *Genome Research*, 21(11):1800–1812, 2011.

Ken Thompson. Programming techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6):419–422, 1968.

Michael Tristem. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of Virology*, 74(8): 3715–3730, 2000.

Robert L Trivers. Parent-offspring conflict. *American Zoologist*, 14(1):249–264, 1974.

Priscilla Turelli, Nathaly Castro-Diaz, Flavia Marzetta, Adamandia Kapopoulou, Charlene Raclot, Julien Duc, Vannary Tieng, Simon Quenneville, and Didier Trono. Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome research*, 24(8):1260–1270, 2014.

Leslie G Valiant. Evolvability. *Journal of the ACM (JACM)*, 56(1):3, 2009.

Marc HV van Regenmortel, Claude M Fauquet, David HL Bishop, EB Carstens, MK Estes, SM Lemon, J Maniloff, MA Mayo, DJ McGeoch, CR Pringle, et al. *Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses.* Academic Press, 2000.

Mariana Varela, Thomas E Spencer, Massimo Palmarini, and Frederick Arnaud. Friendly viruses. *Annals of the New York Academy of Sciences*, 1178(1):157–172, 2009.

S Venner, C Feschotte, and C Biémont. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics*, 25(7):317–323, 2009.

C Vernochet, O Heidmann, A Dupressoir, G Cornelis, P Dessen, F Catzeflis, and T Heidmann. A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. *Placenta*, 32(11):885–892, 2011.

Beatriz Vicoso and Brian Charlesworth. Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8):645–653, 2006.

Cecile Voisset, Antoine Blancher, Herve Perron, Bernard Mandrand, Francois Mallet, and Glaucia Paranhos-Baccala. Phylogeny of a novel family of human endogenous retrovirus sequences, HERV-W, in humans and other primates. *AIDS Research and Human Retroviruses*, 15(17):1529–1533, 1999.

Jichang Wang, Gangcai Xie, Manvendra Singh, Avazeh T Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, Daniel Besser, Alessandro Prigione, Nina V Fuchs, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531):405–409, 2014.

Ting Wang, Jue Zeng, Craig B Lowe, Robert G Sellers, Sofie R Salama, Min Yang, Shawn M Burgess, Rainer K Brachmann, and David Haussler. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences*, 104(47):18613–18618, 2007.

Robin A Weiss. The discovery of endogenous retroviruses. *Retrovirology*, 3(1):67, 2006.

Travis J Wheeler, Jody Clements, Sean R Eddy, Robert Hubley, Thomas A Jones, Jerzy Jurka, Arian FA Smit, and Robert D Finn. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, 41(D1):D70–D82, 2013.

Patricia J Wittkopp and Gizem Kalay. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69, 2012.

Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

George R Young, Jonathan P Stoye, and George Kassiotis. Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. *Bioessays*, 35(9):794–803, 2013.

G.R. Young, U. Eksmond, R. Salcedo, L. Alexopoulou, J.P. Stoye, and G. Kassiotis. Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature*, 491(7426):774–778, 2012.

Xiaoyu Zhuo, Mina Rho, and Cédric Feschotte. Genome-wide characterization of endogenous retroviruses in the bat myotis lucifugus reveals recent and diverse infections. *Journal of Virology*, 87(15): 8493–8501, 2013.