Discovery and characterization of HERV-K (HML-2) integrations in humans and gorillas

A thesis submitted by

Zachary Williams

In partial fulfillment of the requirements

for the degree of

PhD

in

Molecular Microbiology

Tufts University

Sackler School of Graduate Biomedical Sciences

August 2017

Advisor: John M. Coffin, PhD

Abstract

The HERV-K (HML-2) endogenous retrovirus (ERV) clade is the most recently active ERV group in humans, and the only group with human-specific integrations, some of which are insertionally polymorphic. A subset of HML-2 proviruses has some or all ORFs intact, with detectable signals of purifying selection, and expression of these proviruses has been associated with cancer and other diseases. Though no naturally occurring infectious HML-2 provirus is known, these observations raise the possibility that undescribed HML-2 proviruses are still replication competent. If they have been recently active, we predict that rare, intact HML-2 integrations would be detected in surveys of large numbers of human genomes.

We anticipated that mining whole genome sequence data from genetically diverse individuals should reveal rare proviral insertions not present in the reference genome. We searched for conserved HML-2 sequences in both unaligned and discordantly aligned reads from the 1000 Genomes Project and the Human Genome Diversity Project, identifying a total of 36 HML-2 non-reference HML-2 proviruses, including 19 previously unreported loci, with insertion frequencies ranging from <0.0005 to >0.75. Four inserts are present as partial or full length 2-LTR proviruses, one of which has intact ORFs for *gag*, *pro*, *pol*, and *env* –just the second such provirus identified in humans, after HERV-K113, though, like K113, it is not infectious.

Though we did not find an infectious HML-2 provirus in humans, we hypothesized that HML-2 may have maintained infectious activity in other primates. To investigate this possibility, we adapted the tools developed in our first study to search for HML-2 insertions in gorilla whole genome sequence data from the Great Ape Genome Project.

We mined sequence data from 21 individual gorillas, and identified 126 putative gorilla-specific insertions. We confirmed 27 of these by PCR and sequencing, including 6 2-LTR proviruses. We also identified an additional 10 2-LTR proviruses and 92 solo LTRs in the most recent gorilla genome assembly, for a total of 129 confirmed gorilla-specific insertions. Seven proviruses have maintained at least one intact ORF, and one provirus at 9p13.3 has intact ORFs for all viral proteins. Phylogenetic and molecular clock analyses suggest that these elements were active much more recently in gorillas than in humans, and it is possible that HML-2 is still circulating in gorillas as an exogenous virus.

Acknowledgements

First I have to thank John, of course. I would never have heard of Tufts if it hadn't been for my interest in John Coffin's research; considering how much I've enjoyed my time in the Micro department, I'd be grateful to John for that even if I hadn't ended up joining his lab. However, I did join the Coffin lab, and I'm incredibly glad I did; John has given me an alarming amount of freedom to pursue my research interests with minimal interference, while always being willing to provide guidance whenever I ask for it, or mildly worded correction when I make a dumb mistake.

I'm also thankful for my fellow Coffinites, past and present. They've been a joy to be around, both when working and when goofing off. Special mention to my former and current baymates, Neeru Bhardwaj and Farrah Roy, who had to deal with years of me talking to myself and making really dumb jokes. I'm also very grateful to Julia Wildschutte, who first got me working on HERV-K when I joined the lab, and who has been a valued mentor, collaborator, and friend over the years.

As I mentioned, the Micro department has been a wonderful place to work, and I'm grateful to all the people who give it such a friendly, helpful and relaxed atmosphere. Special thanks to my classmate Anne Weeks, who has been a great friend and source of baked goods from the very beginning, and also gave me the opportunity to officiate her wedding, a wonderful and unforgettable experience. There are far too many other people I owe a debt of gratitude to mention them all, but I do want to acknowledge Dr. Todd Wood, who gave me the opportunity to do undergraduate research at a school where such opportunities were rare, and Paul Laskowske, my best friend and fellow ultra-procrastinator.

Lastly, my amazing family: My sister is the only person in the world that I think I might understand somewhat, and I never get tired of talking to her. My parents are both incredibly kind, understanding, and open people, even when I was a bratty kid; on top of that, because I was homeschooled, they had the second job of educating me, and I think they made a pretty good go of it. They filled our house with books and gave me the freedom to spend way too much of my time reading, daydreaming, and puttering around the woods; I suspect those experiences have heavily shaped my approach to science (for better or for worse); they certainly made for a glorious childhood, for which I'll always be grateful.

Table of Contents

List of Tables

List of Figures

List of Copyrighted Materials

Jern P, Sperber GO, Blomberg J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology. 2005;2:50.

Bannert N, Kurth R. The Evolutionary Dynamics of Human Endogenous Retroviral Families. Annu Rev Genomics Hum Genet. 2006;7:149–73.

Hohn O, Hanke K, Bannert N. HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Front Oncol. 2013;3(September):246.

Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011;8(1):90.

1000 Genomes Project Consortium T 1000 GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.

Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci. 2016;201602336.

Jern *et al.*, Hohn *et al.*, Subramanian *et al.,* and The 1000 Genomes Project Consortium paper are open access and do not require formal letters of permission for figure reproduction. Annual Reviews does not require formal permission for figure reproduction in theses and dissertations. Wildschutte *et al.* is my own work, and PNAS does not require authors to obtain permission to use their own figures in later works. Permission statements for each journal are provided in a supplementary file to this document.

List of Abbreviations

1KGP: 1000 Genomes Project

A3: APOBEC3

A3G: APOBEC3G

A: adenine

AAA: poly-adenine tail

ACB: African Caribbean in Barbados

AFR: African ancestry superpopulation

AIDS: acquired immune deficiency syndrome

ALS: amyotrophic lateral sclerosis

ALV: avian leukosis virus

AMR: American ancestry superpopulation

APOBEC: apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

ART: anti-retroviral therapy

ASW: African ancestry in southwest US

BAM: binary SAM

BEB: Bengali in Bangladesh

BED: browser extensible data

BLAST: basic local alignment search tool

BLASTP: protein-protein BLAST

BLAT: BLAST-like alignment tool

BLV: bovine leukemia virus

bp: base pair

C: cytosine

CA: capsid

CDX: Chinese Dai in Xishuangbanna, China

CEU: Utah residents with northern and western European ancestry

CHB: Han Chinese in Bejing, China

chr: chromosome

CHS: southern Han Chinese

CLM: Colombian in Medellin, Colombia

CON: consensus

Chimp: chimpanzee

DNA: deoxyribonucleic acid

De: Denisovan

EAS: east Asian ancestry superpopulation

Env: envelope glycoprotein

ERV: endogenous retrovirus

ESN: Esan in Nigeria

EUR: European ancestry superpopulation

FIN: Finnish in Finland

FeLV: feline leukemia virus

Fv1/4: Friend virus susceptibility 1 and 4

G: guanine

GBR: British in England and Scotland

GIH: Gujarati Indian in Houston, TX

GRCh37: Genome Reference Consortium, human reference genome build 37

GWD: Gambian in Western Division, The Gambia

Gag: group-specific antigen

GATK: Genome Analysis Toolkit

Gor: gorilla

gorGor3/4/5: different gorilla genome assemblies

GAGP: Great Ape Genome Project.

HERV-K: human endogenous retrovirus group K

HERV-K_{CON}: infectious consensus HERV-K proviral construct

HERV-T: human endogenous retrovirus group T

HERV-W: human endogenous retrovirus group W

HERV: human endogenous retrovirus

HERVK-int: HERV-K internal region

hg19: human reference genome build 19 (same as GRCh37)

HIV: human immunodeficiency virus

HML-2: human MMTV-like virus, group 2

HTLV: human T-lymphotropic virus

HGDP: Human Genome Diversity Panel

hg38: the most recent human reference genome build 38 (same as GRCh38)

indel: insertion or deletion

IBS: Iberian populations in Spain

ILS: incomplete lineage sorting

IN: integrase

ITU: Indian Telugu in the UK

JPT: Japanese in Tokyo, Japan

JSRV: Jaagsiekte sheep retrovirus

kb: 1000 base pairs

KHV: Kinh in Ho Chi Minh City, Vietnam

KoRV: Koala retrovirus

LINE-1: long interspersed nuclear element, group 1

lncRNA: long non-coding RNA

LTR: long terminal repeat

LTR5A, B, and Hs: HML-2 subgroups as determined by RepeatMasker

LWK: Luhya in Webuye, Kenya

MA: matrix

MAPQ: mapping quality

MEGA: Molecular Evolutionary Genetics Analysis

MLV : murine leukemia virus

MMTV: mouse mammary tumor virus

MOV10: putative helicase and restriction factor Moloney leukemia virus 10

mRNA: messenger RNA

MSL: Mende in Sierra Leone

MUSCLE: multiple sequence comparison by log-expectation

MXL: Mexican Ancestry in Los Angeles, California

MYA: million years ago

NC: nucleocapsid

NCBI : National Center for Biotechnology Information

Ne: Neanderthal

ng: nanogram

nt: nucleotide

ORF: open reading frame

OWM

panTro3/4/5: chimpanzee genome assemblies

PBS: primer binding site

PCR: polymerase chain reaction

PE: paired-end

PEL: Peruvian in Lima, Peru

PERV: porcine endogenous retrovirus

PJL: Punjabi in Lahore,Pakistan

Pol: retroviral polymerase gene

polyA: polyadenine

ponAbe2: orangutan genome assembly

PPT: polypurine tract

PR: protease

pre/pre-int: pre-integration site

prov: provirus

PUR: Puerto Rican in Puerto Rico

R: repeat region of LTR

RNA: ribonucleic acid

RSV: Rous sarcoma virus

RT: reverse transcriptase

RefSeq: NCBI Reference Sequence Database

SA: splice acceptor

SAM: Sequence Alignment/Map

SAMHD1: SAM domain and HD domain-containing protein 1

SAS: South Asian ancestry superpopulation

SD: splice donor

SINE: short interspersed nuclear elements

SINE-R: SINE of retroviral origin

SIV: simian immunodeficiency virus

SNP: single nucleotide polymorphism

Solo/solo LTR: solitary long terminal repeat

SRA: Sequence Read Archive

STU: Sri Lankan Tamil in the UK

SU: surface unit (env)

SVA: SINE/VNTR/Alu transposable element

T: thymine

TE: transposable element

TM: transmembrane unit (Env)

TRIM5α: Tripartite motif-containing protein 5, isoform alpha

tRNA: transfer RNA

TSD: target site duplication

TSI: Toscani in Italy

U3: retroviral 3’ unique region

U5: retroviral 5’ unique region

UCSC: University of California Santa Cruz

UTR: untranslated region

VCF: variant call format

VLP: viral like particle

WGS: whole genome sequence

XMV: xenotropic MLV

YIDD: amino acid motif sequence at reverse transcriptase catalytic site

YRI: Yoruba in Ibadan, Nigeria

Chapter 1: Introduction

1.1 Retrovirus taxonomy

Retroviruses are a large family of enveloped, reverse transcribing RNA viruses that are found in most vertebrates and virtually all mammals (1), including several viral species that infect humans and can cause severe disease, most notably the human immunodeficiency viruses HIV-1 and 2, which are the cause of the acquired immunodeficiency syndrome (AIDS) pandemic (2).

Retroviruses are members of a larger clade of long terminal repeat (LTR) containing viruses and retrotransposons that infect a wide variety of eukaryotic organisms, including insects, molluscs, worms, fungi, and plants. These more distantly related species are placed within the families *Metaviridae* and *Pseudoviridae,* while the true retroviruses are classified as members of the family *Retroviridae* (1,3,4). *Retroviridae* contains 7 genera: *Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus,* and *Spumavirus*; spumaviruses (also known as foamy viruses) are notably divergent from other retrovirus genera in sequence and have substantial differences in their replication cycle, and so are placed in a separate subfamily, *Spumaretrovirinae,* with the other 6 genera placed within the subfamily *Orthoretrovirinae*. A number of other families of viruses and transposons that utilize reverse transcriptases are known, such as the hepadnaviruses and LINE-1 retrotransposons; however, their replication mechanisms are drastically different from the LTR retroelements, and the evolutionary relationships (if any) between these other groups and retroviruses is unclear (4,5).

There are 4 known human specific retroviruses, three of which can be highly pathogenic. The human immunodeficiency viruses HIV-1 and 2, the causative agents of AIDS, are both lentiviruses, closely related to different species of simian immunodeficiency virus (SIV)(6). There are also two deltaretroviruses, the human T-lymphotropic viruses HTLV-1 and 2; these are asymptomatic in most people, but a small percentage of those infected with HTLV-1 will develop severe leukemia or lymphoma (7,8). HTLV-2 has been associated with myelopathy, but no severe diseases (9). Two more HTLV species have been recently reported, but are very rare, poorly studied, and may be zoonotic, rather than established human viruses (10,11).

1.2 Retrovirus structure and replication cycle

Though very divergent at the sequence level, all retroviruses share a set of core viral genes, their genome and virion structure is highly conserved, and they all follow the same unique life cycle (2). Retroviruses have single stranded, positive sense RNA genomes that code for four characteristic genes: *gag*, *pro*, *pol*, and *env*, always found in that order (Fig. 1-1A). *Gag* encodes the 3 retroviral structural proteins, nucleocapsid (NC), capsid (CA), and matrix (MA); *pro* encodes protease, necessary for virion maturation, *pol* produces two proteins, reverse transcriptase (RT) and integrase (IN), required for replication of the viral genome and integration of the genome into host cellular DNA, respectively; lastly, *env* produces the envelope glycoprotein (Env), responsible for mediating entry into the host cell. Env is processed into two tightly associated subunits, the surface subunit (SU) which is responsible for binding a host cell receptor, and the transmembrane subunit (TM), which anchors Env to the viral membrane and mediates fusion of the viral membrane with the host cell membrane (Fig. 1-1B).

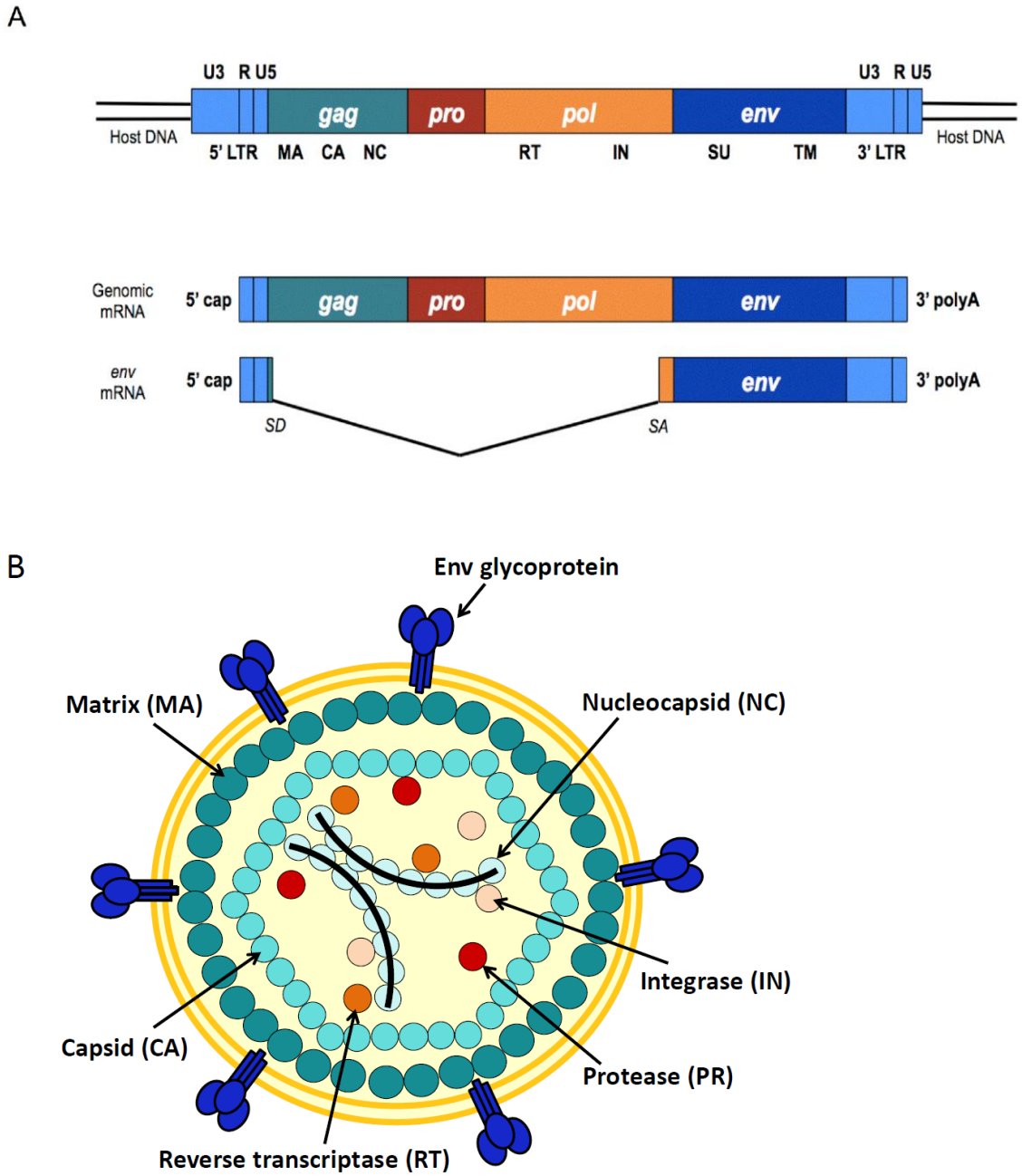


Figure 1-1. Retrovirus genome and virion structure.
(A) Organization of retroviral DNA and RNA genomes, and spliced envelope mRNA.
(B) Schematic structure of mature retrovirus virion.
Figure 1-1A courtesy of Neeru Bhardwaj.

Though *gag*, *pro*, *pol* and *env* are the only genes found in all retroviruses, some species, termed 'complex' retroviruses, also encode additional genes, known as accessory genes, which often have important functions in modulating and antagonizing the host immune response, or in co-opting or antagonizing cellular genes for the virus' benefit (12). Lentiviruses like HIV are probably the most famous for this, encoding at least 7 accessory proteins necessary for full virulence. Other complex retroviruses include spumaviruses, deltaretroviruses, and some betaretroviruses (12–15).

The coding sequences listed above are flanked by two unique non-coding regions, U5 at the 5' end and U3 at the 3' end, which are themselves flanked by 5' and 3' repeats, known as the R sequences (Fig. 1-1A). Each virion contains two copies of this RNA genome in a non-covalently bound dimer. The RNA molecules are coated in nucleocapsid and are packaged within the viral capsid, along with the viral enzymes necessary for replication, tRNAs for priming genome replication, and, for some complex retroviruses, certain accessory proteins. The capsid is embedded in a layer of matrix protein; this is surrounded by a host derived lipid bilayer, which is studded with copies of the Env glycoprotein (Fig. 1-1B)(16,17).

Retroviruses have a life cycle distinguished by two unique steps, reverse transcription and integration, that are responsible for many of their remarkable properties and have major implications for their pathogenesis and evolution. Like all enveloped viruses, retroviruses begin a new infectious cycle by binding to a specific host receptor and fusing their lipid membrane with the cellular membrane, thus allowing entry of the capsid into the host cell (Fig. 1-2).

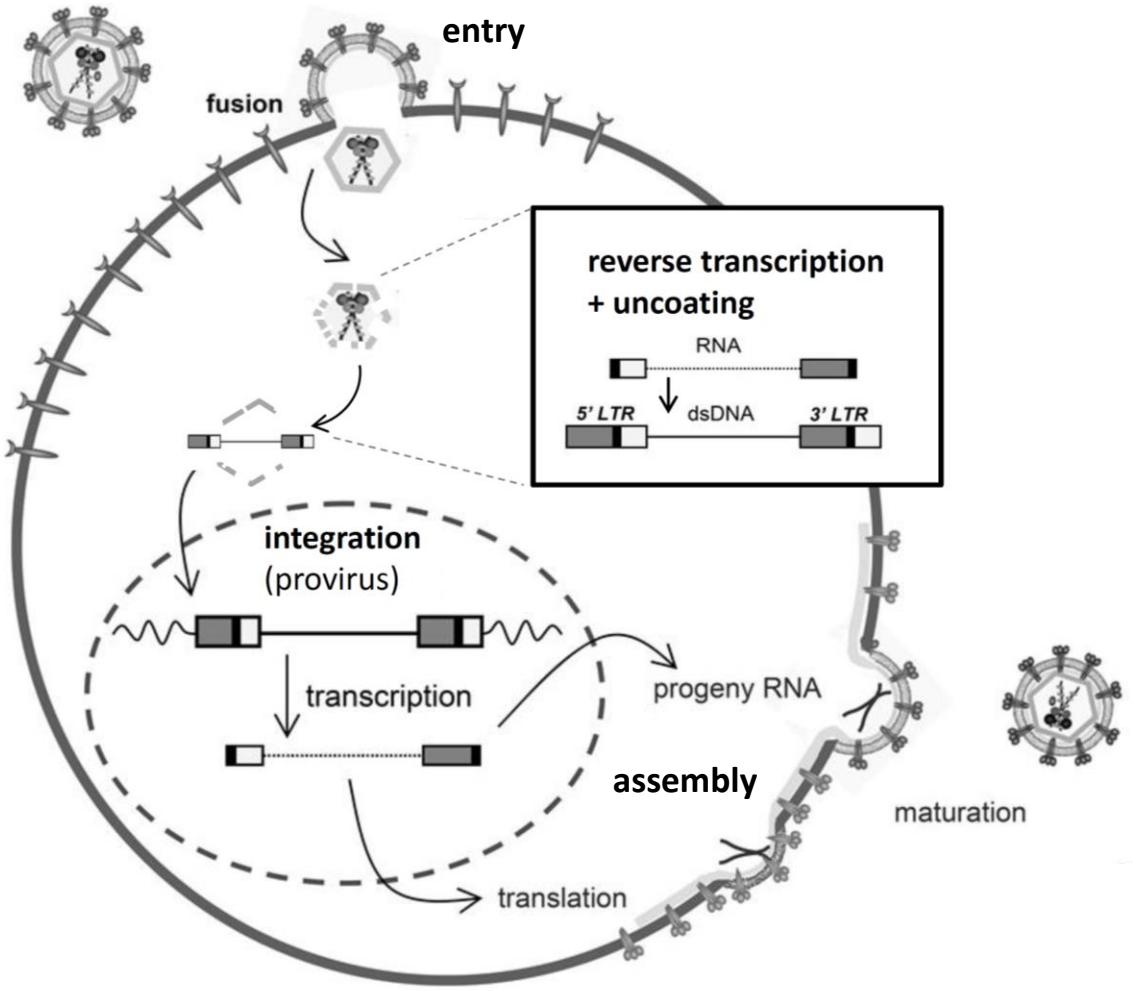


Figure 1-2. Retroviral replication cycle.
Steps of retrovirus replication cycle in a host cell.
Figure courtesy of Julia Wildschutte.

Following entry, reverse transcription of the virus' single stranded RNA genome into double stranded DNA begins, using host tRNA molecules packaged in the virion to prime the initial elongation step by binding to a complementary sequence in the viral genome known as the primer binding site (PBS); the specific tRNA used for priming depends on the sequence of the PBS, and differs from species to species. Spumaviruses, uniquely, sometimes initiate reverse transcription in the producer cell, rather than post-entry.

Though each virion contains two RNA molecules, only one DNA molecule is produced, thus retroviruses are called 'pseudodiploid.' However, during reverse transcription, the polymerase frequently switches between the two RNA templates, producing a recombinant DNA genome. In addition to frequent, relatively random template switching between the two RNA molecules, two specific template switches, the first and second strand transfer reactions, are absolutely required for complete genome replication. The strand transfers also append copies of the U5 and U3 terminal regions to the opposite ends of the genome, generating long terminal repeats, or LTRs, at each end of the DNA genome, each one composed of three regions, U3, R, and U5, in that order; importantly, the LTRs are identical in sequence at the time they are produced (1,17–20).

Following (or during) reverse transcription, the capsid is disassembled, and the viral DNA genome forms a complex with integrase called the intasome; the intasome is trafficked to the host cell's nuclear DNA, where integrase mediates integration of the viral DNA into the host chromosome. The integrated viral genome, known as a provirus, is flanked by short, 4-6 bp duplications of the host genome, known as target site duplications or TSDs (1,17). The integration process is similar to the generation of prophages by temperate bacteriophages, but has two important differences. First, retroviral integration is not site-specific, though it is not entirely random; many retroviruses recruit cellular cofactors to enhance integration efficiency, which leads to biases towards specific genomic regions (e.g. murine leukemia viruses preferentially integrate into promoter regions), and integrase itself has weak preferences for certain sequence contexts. Secondly, integration is irreversible: once integrated, there is no

mechanism for specifically excising the provirus, unlike prophages and DNA transposons (1,20,21).

Following integration, the provirus is transcribed using the host's RNA polymerase and other transcription machinery. The LTRs contain elements required to initiate and regulate transcription. The upstream U3 region functions as a promoter, and often contains enhancer elements as well; transcription initiates at the start of the 5' R region, and polyadenylation begins at the end of the 3' R region, again using the host machinery to generate the poly-A tail. Though retroviruses mostly use host machinery for transcription and mRNA processing, some complex retroviruses do require the activity of accessory genes for full transcriptional activity. The Tat proteins of HIV and other lentiviruses strongly enhance viral transcription levels, and genes with similar transcription-enhancing functions are found in deltaretroviruses (*tax*) and spumaviruses (*tas*), though the genes are not homologous to *tat*. Lentiviruses also require a second accessory protein, Rev, that helps traffic unspliced viral mRNAs out of the nucleus; again, similar, but non-homologous genes are found in some other retroviruses, including deltaretroviruses (*rex*) and some betaretroviruses (*rem*/*rec*)(14,15,22,23).

Two major transcripts are generated in all retroviruses except the spumaviruses; full length transcripts, some of which are packaged into progeny virions as new RNA genomes and some which serve as the mRNA for *gag*, *pro*, and *pol*, and a spliced transcript that encodes only *env* (Fig. 1-1A, Fig. 1-2). Spumaviruses generate an additional spliced transcript that codes for *pro* and *pol*, with only *gag* being produced from the full length viral RNA. In addition to these major transcripts, the complex

retroviruses must also produce transcripts for their accessory genes; these are typically generated through alternative splicing or alternative transcriptional start sites (22).

As Pro and Pol are translated from the same start codon as Gag (again with the exception of spumaviruses), they are initially produced as polyproteins with Gag. However, most retroviruses contain a stop codon between *gag* and *pro,* between *pro* and *pol,* or both; thus, production of the full length polyprotein depends on suppression of these stop codons, either by read-through of a leaky stop codon, or by ribosomal frameshift upstream of the stop codon. As these stop codon suppression mechanisms are inefficient, much more Gag (or, in some species, Gag-Pro) is produced than the full length Gag-Pro-Pol polyprotein, allowing for high numbers of the viral structural proteins without overproducing the viral enzymes, which are needed in much smaller quantities per virion. As previously mentioned, Env is translated separately from the spliced *env* transcript; it is then trafficked to the outer cell surface (1,16,17,20,24).

After translation, Gag and Gag-Pro-Pol spontaneously assemble into virion cores, simultaneously binding and packaging the viral genomes; this can either take place in the cytoplasm (betaretroviruses and spumaretroviruses) or at the cell surface (all other retroviral genera). Core assembly is followed by (or in some cases occurs in concert with) recruitment of Env proteins, envelopment, and budding of the viral particles from the cell surface. The newly produced virions are not infectious until they go through a final maturation step mediated by the viral protease, wherein protease cleaves the Gag and Gag-Pro-Pol polyproteins into the functional proteins necessary for infection; the mature virions can then go on to infect new cells, repeating the cycle (1,20,24).

1.3 Retroviral pathogenesis

The two unique steps of the retroviral lifecycle, reverse transcription and integration, have a number of important consequences for retroviral pathogenesis and evolution. Reverse transcriptases have a high error rate, and have no proofreading activity, leading to high mutation rates in many retrovirus species. Also, as previously mentioned, reverse transcriptase frequently switches between the two RNA templates, allowing the virus to generate a replication competent DNA copy even in the presence of breakages or other lesions in the parental RNA strands, which are quite frequent due to the high mutation rate and fragile nature of RNA. However, the template switching process can also introduce errors by jumping between non-homologous regions of the RNA strands, leading to deletions, duplications, and rearrangements in the DNA copy. If genomes from two different viral strains or species are packaged in the same virion, recombination will produce chimeric progeny, in a form of viral sexual reproduction (18). The high mutation and recombination rates of retroviruses give them a remarkable capacity for rapid evolution, both within a single host and between hosts. In HIV and other lentiviruses, these properties are enhanced by the large effective population size and high replication rate of the virus; this rapid evolution allows the virus to evade the host immune system, and also leads to rapid development of resistance to antiretroviral drugs, unless multiple drugs with different inhibitory mechanisms are used (18,25,26).

Because retroviruses can integrate virtually anywhere within the host genome, they are inherently mutagenic and oncogenic; in fact, one of the earliest retroviruses to be discovered, Rous sarcoma virus (RSV), was isolated from a chicken tumor, and retroviruses were originally called ‘RNA tumor viruses’ because of their propensity to

cause cancer. Retroviruses can cause cancer through simple insertional mutagenesis, knocking out a tumor suppressor for example, however this is fairly rare, and a number of other oncogenic mechanisms exist (27). Most notably, when integrated near a proto-oncogene, the proviral LTRs can function as promoters or enhancers, leading to overexpression of the oncogene. Additionally, retroviruses can code for oncogenes themselves; often, these oncogenes were originally derived from host proto-oncogenes that were accidentally incorporated into a viral genome (23,27,28).

As mentioned above, proviral integration is irreversible. Proviruses can be lost from the host genome, but only through non-specific deletions that either leave part of the provirus in the genome, remove part of the host genomic flanking sequence, or both. The 5' and 3' LTRs can undergo homologous recombination, however, which removes the internal coding region of the provirus, but leaves a single integrated LTR or 'solo LTR' in the genome (Fig. 1-3). Importantly, productive retroviral infection is usually non-lytic; though retroviruses can be highly cytopathic, cell death is not required for successful viral replication (29,30). If a provirus integrates into a host cell genome, and that cell survives and gives rise to progeny, every daughter cell will carry that provirus in its genome. In HIV, it is likely that this mechanism plays a role in maintaining the viral reservoir that makes complete cure of HIV infections so difficult to achieve (1,20,21,31–36).

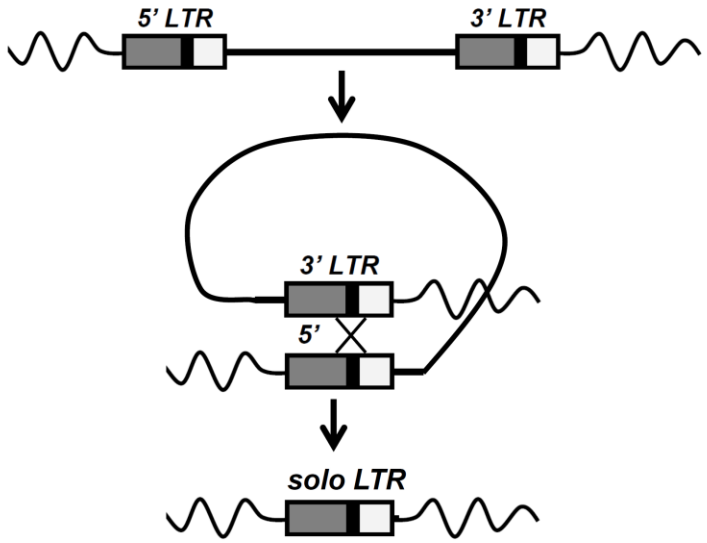


Figure 1-3. Mechanism of solo LTR formation.
After integration, homologous recombination between the identical or nearly identical 5' and 3' LTRs of a provirus will lead to formation of a 'solo LTR,' with complete loss of all the internal sequences that code for viral genes.

1.4 Endogenous retroviruses

A second consequence of irreversible integration is that, if integration occurs in a germline cell, the provirus will be passed to the offspring derived from that cell just like a normal host gene. Any proviruses inherited in this manner will thus be present in every cell of the organism. This phenomenon, known as retroviral endogenization, is thought to be fairly rare, but over the course of evolutionary history it has occurred frequently enough that significant fractions of most vertebrate genomes are derived from such endogenous retroviruses, or ERVs; for example, human ERVs (HERVs) make up 8% of the human genome (1,5,19,20,37). After the initial endogenization event, a provirus will only be present in one or a very few individuals and is thus classified as insertionally polymorphic; over many generations, however, a small fraction of these insertions will spread through the population through random genetic drift, just like any other low

frequency genetic variant, and over long periods of evolution, some may reach fixation in the population, with no individuals carrying the original pre-integration or 'empty' site. As mentioned above, ERVs can also form solo LTRs; a single ERV locus may thus have two different allelic variants, the 2-LTR, full length provirus, or a solo LTR. Insertionally polymorphic ERVs will also have a pre-integration allele, along with one or both of the two insertion alleles (1,5,20,23,38,39).

It is important to note that, though some ERVs (but no known HERVs) can produce fully replication competent virus, most are defective, either due to mutations and deletions present in the provirus at the time of integration and/or to accumulation of deleterious mutations post-integration, and it is likely that many ERVs are the incidental byproduct of exogenous viral replication. Following integration, there is no selective pressure to maintain infectivity, and if the provirus has any pathogenic effect there may be selection for inactivating mutations; thus, the provirus will accumulate mutations as it is passed down over multiple generations (20,40–42). On the other hand, LTR retrotransposons, which share the same basic replication mechanisms as retroviruses, but do not have an *env* gene and thus only replicate via intracellular retrotransposition are typically included in estimates of ERV content and, unlike true retroviruses, are completely dependent on germline transmission, as they are by definition incapable of replicating exogenously (3). The line between retrovirus and retrotransposon is somewhat blurry, however, as many retrotransposon families appear to have evolved from closely related retroviruses by loss of their *env* gene, and conversely, retrotransposons can evolve into genuine transmissible viruses by acquisition of *env* genes or other viral fusogen proteins; this has occurred with *gypsy* elements in *Drosophila*, which have co-opted a

baculovirus envelope protein, and it has been hypothesized that all vertebrate retroviruses derive from a similar *env* acquisition event (3,5,43,44).

A number of classification schemes for ERVs have been developed, none of which is entirely satisfactory. ERVs are typically separated into three large clades or 'classes;' class I includes gammaretroviruses and gamma-like ERVs and class II includes betaretroviruses and beta-like ERVs (Fig. 1-4). Class III ERVs have some affinity with spumaviruses, though most members of this family are highly divergent from exogenous spumaviruses. Somewhat confusingly, these 'classes' are not equivalent to taxonomic classes, as they all fall within the family Retroviridae, and there are a number of ERVs related to alpharetroviruses and lentiviruses that are not included in this scheme (1,45). A second approach to ERV nomenclature has been to define ERV groups or 'families' (again, not equivalent to the taxonomic family rank) based on the tRNA used to prime reverse transcription, e.g., HERVs that use a lysine tRNA are placed in the HERV-K family (K for lysine), HERV-W uses tryptophan, HERV-T uses threonine, and so on.

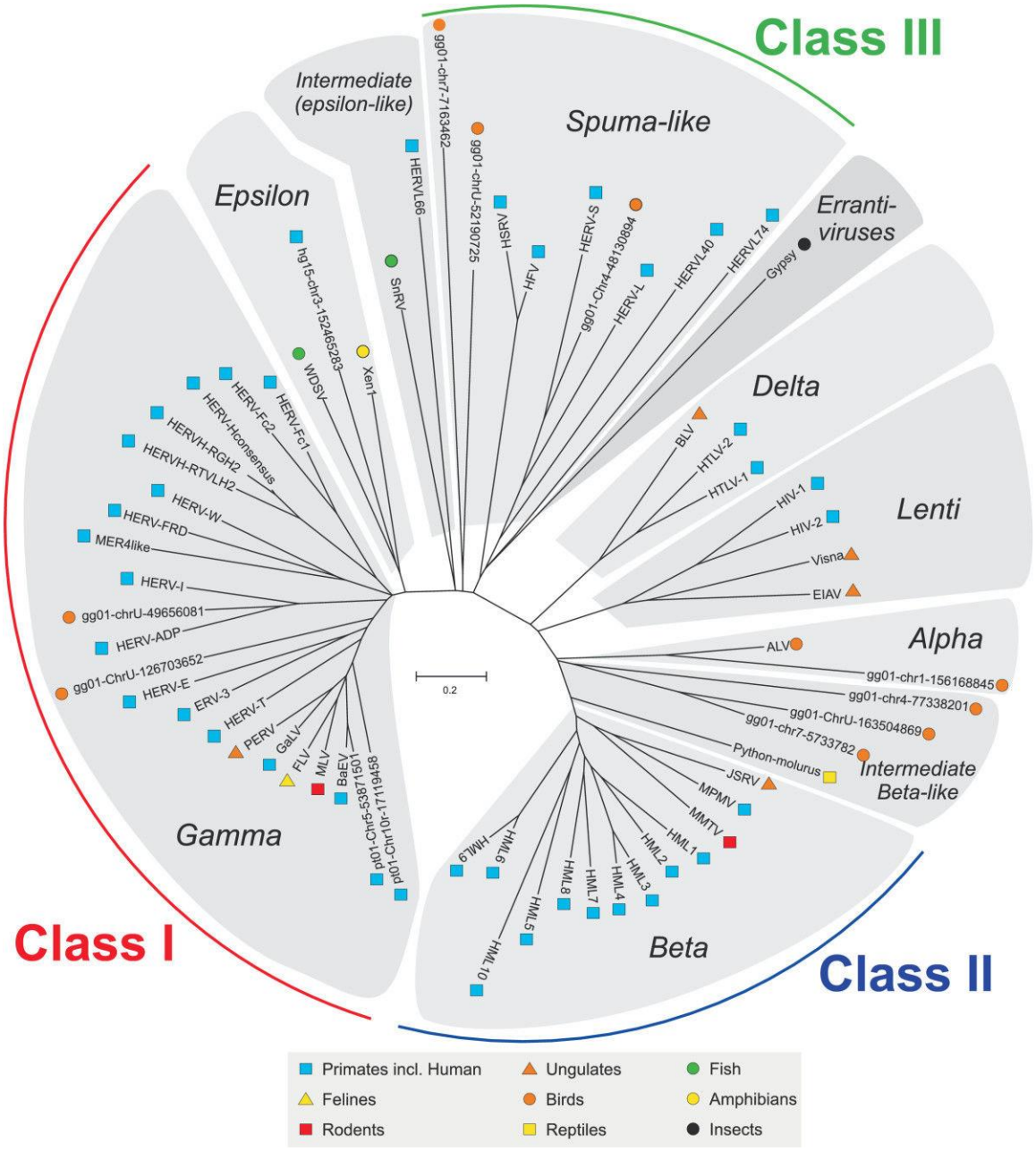


Figure 1-4. Exogenous and endogenous retrovirus taxonomy.
Unrooted phylogeny of the *pol* genes from representative retroviruses from each retroviral genus and from related endogenous retroviruses from Class I (gammaretrovirus-like), Class II (betaretrovirus-like), and Class III (spumaretrovirus-like). ERVs that do not fit into these classes are also included, e.g. endogenous alpharetroviruses. Host organisms for each virus are indicated by the symbols as shown. Republished with the permission of BioMedCentral under the Creative Commons Attribution License 4.0 (45)

This approach has the disadvantage that highly divergent retroviruses can use the same tRNA; for example, betaretroviruses and spumaretroviruses both often use lysine, though the HERV-K classification is traditionally limited to beta-like sequences; additionally, closely related viruses do not necessarily use the same tRNA. Nomenclature for the loci of individual proviruses is even more fractured, with virtually every lab using its own naming system; our lab prefers to identify individual proviruses by their specific genomic location, naming each provirus after its chromosomal band locus (46–50).

Endogenous retroviruses have a number of features that make them very useful for evolutionary research. As integration is not site specific, the likelihood of two different proviruses integrating at the exact same site in a host genome is vanishingly small; thus, ERV insertions (as well as other retrotransposon insertions) can serve as virtually homoplasy free markers of common ancestry, and have proven to be very useful in resolving the evolutionary relationships of mammals, birds, and other organisms (5,51–58). Additionally, since the 5' and 3' LTRs of any given provirus are identical at the time of evolution, the sequence divergence between the two LTRs of a provirus can be used as a molecular clock to estimate the time since integration; this and other methods of ERV age estimation and their limitations are discussed in more detail in section 1.4(51,59).

ERVs can also be used to investigate host-viral coevolution –the accumulation of endogenous viral sequences over time serves as a sort of 'fossil record' of the viruses that infected an organism's ancestors, and thus can show how viruses and their hosts have adapted to each other time (12,20). For example, the discovery of ancient endogenous lentiviruses played an important role in showing that molecular clock studies had profoundly underestimated the age of the *Lentivirus* genus, and that the evolutionary rates

of many viruses are much lower over long periods of time than had been estimated based on short term measurements of viral mutation rates (60–62). ERVs can also play an active role in host evolution; for example, retroviral *env* and *gag* genes have been co-opted by host immune systems to serve as 'restriction factors' that can inhibit viral replication. Endogenously expressed Env proteins can block entry of closely related viruses that use the same receptor by binding to their cognate receptor on the cell surface and thus preventing virus binding. *Fv1* is a mouse gene derived from a retroviral *gag* gene that restricts MLV replication by destabilizing viral capsids (20,63–68). The LTRs of ERVs are also frequently co-opted to promote expression of host genes (69–71). Another example of ERV co-option is the syncytins, host encoded, retrovirus derived envelope genes that help form the trophoblast syncytium by promoting cell-cell fusion. Two such genes, *syncytin-1* and *syncytin-2*, are known in humans, derived from two unrelated HERVs. Similar but independent co-option events have occurred in at least 6 times in other mammalian taxa, a rather remarkable example of convergent evolution (72–78).

Not all ERV-host interactions are beneficial, however. As previously mentioned, some ERVs are capable of producing infectious virus. In some cases, such viruses are derived from single replication competent proviruses; however, multiple partially defective proviruses can also give rise to infectious virus through complementation and recombination (23,38,79). ERV-derived infectious viruses have been found in a number of different animals, including mice, cats, chickens and pigs, and on occasion can cause severe disease (5,19,27,80). The AKR mouse strain, for example, has a very high incidence of lymphoma, which is caused by highly pathogenic murine leukemia viruses

(MLVs) that arise through recombination between endogenous MLV sequences. A similar mechanism leads to MLV-induced lymphomas in antibody-deficient C57/BL6 mice and other mouse strains, and in cats, recombination between exogenous and endogenous feline leukemia virus (FeLV) sequences may play a role in tumorigenesis as well (5,20,81,82).

The above observations in animals naturally lead to the possibility that endogenous retroviruses could cause disease in humans as well. Though HERVs compose a significant fraction of the human genome, the vast majority of these sequences are quite old, and have many mutations and deletions making them incapable of replication and unlikely to have significant biological activity. In fact, most HERV loci are found in other primate species as well, indicating that they formed prior to the split between humans and our closest relatives, the chimpanzees, approximately 6 million years ago, and most appear to be much older than that (Fig. 1-3)(1,23,38,39). A few of these more ancient HERVs have been associated with disease, most notably the HERV-W/*syncytin 1* group, which has been investigated as a possible cause of multiple sclerosis, but the majority appear largely biologically inert (83–85).

1.5 HERV-K HML-2

Though most HERVs are ancient and heavily degraded by mutation, the one exception to this is the HERV-K HML-2 (Human MMTV Like 2) group of betaretrovirus-like HERVs, named for their similarity to the exogenous mouse mammary tumor virus (MMTV). This is the only HERV group that maintained infectious activity in the human lineage after it split from the chimpanzee lineage (Fig. 1-5), and thus is the only group with human-specific insertions, some of which are unfixed in human populations ("insertionally polymorphic"), i.e. the proviral insertion is present in some individuals but the pre-integration site is present in others (38,39,46,86,87). Additionally, many HML-2 proviruses contain intact open reading frames for viral genes, including one provirus, K113, that has open reading frames (ORFs) for all 4 retroviral genes (46,88,89). This provirus has been shown to be non-infectious; however, two separate groups have shown that HML-2 genomes built from a consensus sequence of the youngest, most intact proviruses are weakly infectious in some human cell lines, and one of these groups also showed that a chimeric provirus composed of sequence from just 3 known HML-2 proviruses maintained some infectivity (90,91).

In addition to *gag*, *pro*, *pol* and *env*, HML-2 encodes a single accessory protein, *rec*, produced by alternative splicing of the *env* mRNA, which is important for trafficking unspliced viral mRNA out of the nucleus, similarly to MMTV's *rem* and HIV's *rev*, though these genes do not appear to be homologous to each other (92,93) (Fig. 1-6).

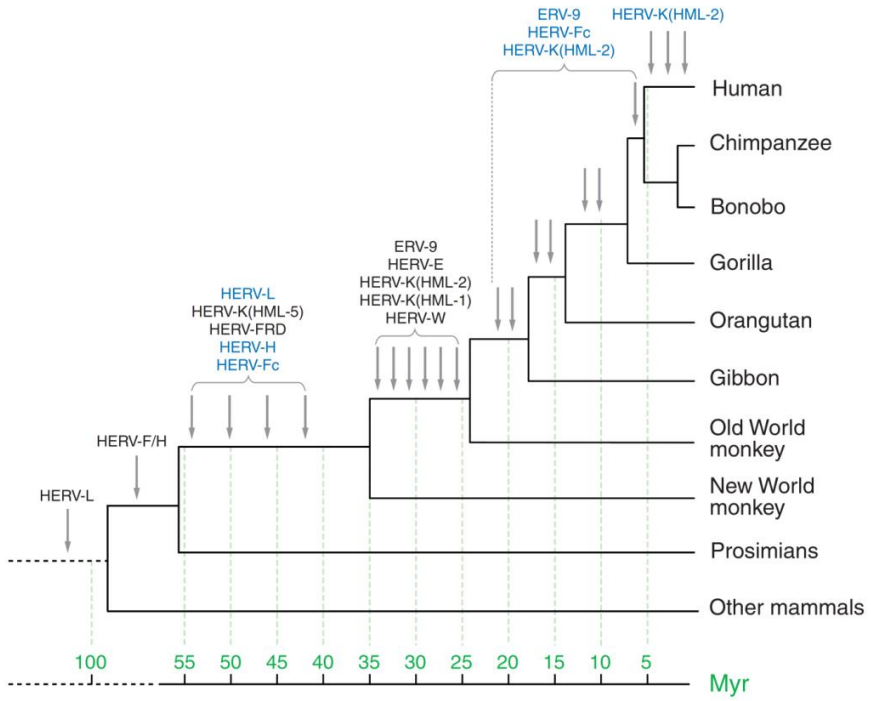


Figure 1-5. Timeline of HERV activity in primates.
Primate tree showing insertion activity of different HERV families found in primates at different points of primate evolution. Timeline at bottom shows timing of evolution and integration events in millions of years before the present. Republished with permission of Annual Reviews; permission conveyed through Copyright Clearance Center, Inc (39).

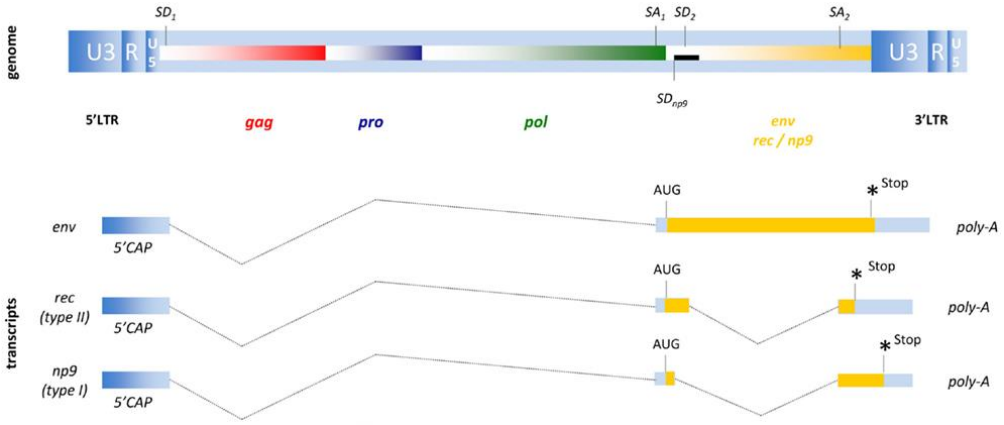


Figure 1-6. Schematic of HERV-K (HML-2) provirus and spliced RNAs
Singly spliced *env* and the doubly spliced Type 2 *rec* and Type 1 *np9* mRNAs are shown. Full length genomic mRNA not shown. SA=splice acceptor, SD=splice donor. Adapted with the permission of Frontiers in Oncology (23) under the Creative Common Attribution license. Scale bar and one transcript figure removed.

About half of the HML-2 proviruses in the human genome, termed type 1 proviruses, contain a 292 bp deletion near the start of *env*, rendering the protein non-functional; full length proviruses are classified as type 2. This deletion also removes the splice donor required to produce the *rec* mRNA; instead, another mRNA is produced using a cryptic splice donor site, leading to production of a different protein called Np9; it is unclear what the function of this protein is, if any (94).

At the time of this study, there were 89 known partial or full length 2-LTR HML-2 proviruses in humans, and about 10 times this many solo LTRs. Approximately 20 of the 2-LTR proviruses are human-specific, and about half of those are insertionally polymorphic (46). The HML-2 group is further subdivided into three subgroups, LTR5Hs, LTR5A, and LTR5B. LTR5Hs and LTR5A are monophyletic clades, while LTR5B is a group of older proviruses basal to both A and Hs (Fig. 1-7, Table 1-1). The LTR5Hs group contains all known human-specific HML-2 insertions; however, many LTR5Hs proviruses are not human-specific, with many proviruses found in chimps, gorillas and orangutans, and a few of the oldest showing up in gibbons as well (23,38,39,95).

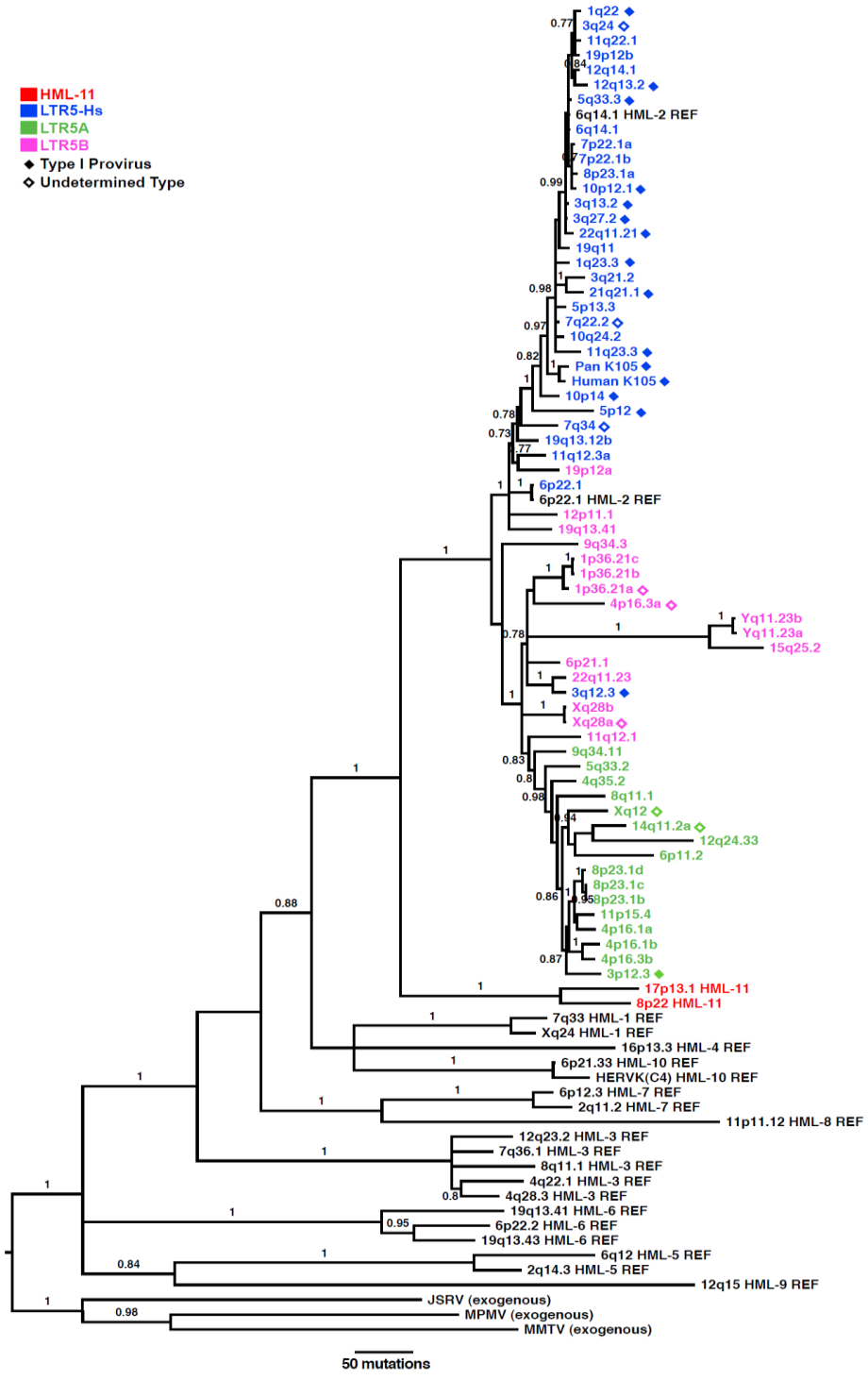


Figure 1-7. Phylogenetic relationships of HML-2 and other betaretroviruses. *Pol* tree of the three main HML-2 subgroups, LTR5Hs (blue), LTR5A (green), and LTR5B (pink) along with other reference sequences of other HERV-K groups (labeled REF) and selected exogenous betaretroviruses. Republished with the permission of BioMedCentral under the Creative Commons Attribution License 4.0(46).

Table 1-1. Previously characterized HML-2 proviruses

Locus	Alias	Coordinates (hg19)	Subgroup	Alleles
1p36.21a		chr1:12840260	B	prov
1p36.21b	K(OLDAL023753), K6, K76	chr1:13458305	B	prov
1p36.21c		chr1:13678850	B	prov
1p34.3		chr1:36955490		prov
1p31.1	K4, K116, ERVK-1	chr1:75842771	Hs	prov, pre
1q21.3		chr1:150605284	Hs	prov
1q22	K102, K(C1b), K50a, ERVK-7	chr1:155596457	Hs	prov
1q23.3	K110, K18, K(C1a), ERVK-18	chr1:160660575	Hs	prov
1q24.1	K12	chr1:166574603	B	prov
1q32.2		chr1:207808457	Hs	prov
1q43		chr1:238925595	B	prov
2q21.1		chr2:130719538	Hs	prov
3p25.3	K11, ERVK-2	chr3:9889346	Hs	prov
3p12.3		chr3:75600465	A	prov
3q12.3	K(II), ERVK-5	chr3:101410737	Hs	prov
3q13.2	K106, K(C3), K68, ERVK-3	chr3:112743479	Hs	prov, pre, solo
3q21.2	K(I), ERVK-4	chr3:125609302	Hs	prov
3q24	ERVK-13	chr3:148281477	Hs	prov
3q27.2	K50b, K117, ERVK-11	chr3:185280336	Hs	prov
4p16.3a		chr4:234989	B	prov
4p16.3b		chr4:3980069	A	prov
4p16.1a	K17b	chr4:9123515	A	prov
4p16.1b		chr4:9659588	A	prov
4q13.2		chr4:69463709	A	prov
4q32.1		chr4:161579938	Hs	prov
4q32.3	K5, ERVK-12	chr4:165916840	Hs	prov
4q35.2		chr4:191027414	A	prov
5p13.3	K104, K50d	chr5:30487114	Hs	prov
5p12		chr5:46000159	Hs	prov
5q33.2		chr5:154016502	A	prov
5q33.3	K107/K10, K(C5), ERVK-10	chr5:156084717	Hs	prov

Adapted with the permission of BioMedCentral under the Creative Commons Attribution License 4.0(46). Columns 4, 6, 7, 8, 10, and 11 removed, 'subgroup' column added.

Table 1-1, cont. Previously characterized HML-2 proviruses

Locus	Alias	Coordinates (hg19)	Type	Alleles
6p11.2		chr6:57623896	A	prov
6p22.1	K(OLDAL121932), K69, K20	chr6:28650367	Hs	prov
6p21.1	K(OLDAL035587), KOLD35587	chr6:42861409	B	prov
6q14.1	K109, K(C6), ERVK-9	chr6:78427019	Hs	prov, pre, solo
6q25.1		chr6:151180749	B	prov
7q11.21		chr7:65469689	B	prov
7p22.1a	K108L, K(HML.2-HOM), K(C7), ERVK-6	chr7:4622057	Hs	prov, pre, solo
7p22.1b	K108R, ERVK-6	chr7:4630561	Hs	prov, pre, solo
7q22.2	ERVK-14	chr7:104388369	Hs	prov
7q34	K(OLDAC004979), ERVK-15	chr7:141450926	Hs	prov
8p23.1a	K115, ERVK-8	chr8:7355397	Hs	prov, pre
8p23.1b	K27	chr8:8054700	A	prov
8p23.1c		chr8:12073970	A	prov
8p23.1d	KOLD130352	chr8:12316492	A	prov
8q11.1	K70, K43	chr8:47175650	A	prov
8q24.3a		chr8:140472149	Hs	prov
9q34.11	K31	chr9:131612515	A	prov
9q34.3	K30	chr9:139674766	B	prov
10p14	K(C11a), K33, ERVK-16	chr10:6867109	Hs	prov
10p12.1	K103, K(C10)	chr10:27182399	Hs	prov, pre, solo
10q24.2	ERVK-17, c10_B	chr10:101580569	Hs	prov
11p15.4	K7	chr11:3468656	A	prov
11q12.1		chr11:58767448	B	prov
11q12.3	K(OLDAC004127)	chr11:62135963	Hs	prov
11q22.1	K(C11c), K36, K118, ERVK-25	chr11:101565794	Hs	prov, pre, solo

Adapted with the permission of BioMedCentral under the Creative Commons Attribution License 4.0(46). Columns 4, 6, 7, 8, 10, and 11 removed, 'subgroup' column added.

Table 1-1, cont. Previously characterized HML-2 proviruses

Locus	Alias	Coordinates (hg19)	Type	Alleles
11q23.3	K(C11b), K37, ERVK-20	chr11:118591724	Hs	prov
12p11.1	K50e	chr12:34772555	A	prov
12q13.2		chr12:55727215	Hs	prov,pre,solo
12q14.1	K(C12), K41, K119, ERVK-21	chr12:58721242	Hs	prov, solo
12q24.11		chr12:111007843	Hs	prov
12q24.33		chr12:133667120	A	prov
14q11.2		chr14:24480625	A	prov
14q32.33		chr14:106139659	A	prov
15q25.2		chr15:84829020	B	prov
16p11.2		chr16:34231474	Hs	prov
16p13.3		chr16:2976160	B	prov
19p13.3	ERVK-22	chr19:385095	Hs	prov
19p12a	K52	chr19:20387400	B	prov
19p12b	K113	chr19:21841536	Hs	prov, pre
19p12c	K51	chr19:22757824	Hs	prov
19q11	K(C19), ERVK-19	chr19:28128498	Hs	prov
19q13.12a		chr19:36063207	Hs	prov
19q13.12b	K(OLDAC012309), KOLD12309	chr19:37597549	Hs	prov
19q13.41		chr19:53248274	B	prov
19q13.42	LTR13	chr19:53862348	B	prov
20q11.22	K(OLDAL136419), K59	chr20:32714750	B	prov
21q21.1	K60, ERVK-23	chr21:19933916	Hs	prov
22q11.21	K101, K(C22), ERVK-24	chr22:18926187	Hs	prov
22q11.23		chr22:23879930	B	prov
Xq11.1		chrX:61959549	B	prov
Xq28a		chrX:153817163	B	prov
Xq28b		chrX:153836675	B	prov
Xq12		chrX:65684132	A	prov
Yp11.2		chrY:6826441	A	prov
Yq11.23a		chrY:26397837	B	prov
Yq11.23b		chrY:27561402	B	prov
U219	K105, K111	satellite repeat	Hs	prov,pre,solo

Adapted with the permission of BioMedCentral under the Creative Commons Attribution License 4.0(46). Columns 4, 6, 7, 8, 10, and 11 removed, 'subgroup' column added.

The high numbers of relatively intact, insertionally polymorphic HML-2 proviruses make this group a clear candidate to investigate for links to disease, and indeed HML-2s have been associated with many diseases, including a number of different cancers, autoimmune diseases, and neurodegenerative disorders. Like most ERVs, HML-2 proviruses are typically transcriptionally silenced in healthy tissues, but transcription has been observed in a number of disease states, as well as the healthy placenta and in embryos (96,97). Production of HML-2 RNA, proteins, and in some cases even viral like particles (VLPs) have been reported in breast cancer (98,99), melanoma (100), teratocarcinoma (101,102), amyotrophic lateral sclerosis (ALS)(103), and multiple sclerosis (104,105); additionally, antibodies to HML-2 proteins have been reported in HIV infection, melanoma (106), and germ cell tumors (102,107), and have been proposed as a biomarker for the latter disease. It is important to note that no causative link has been determined for any of these diseases, and it is possible, even likely, that disease-linked HML-2 expression is a result of the disease, rather than a cause. Cancers in particular tend to upregulate expression from HERVs and other elements that are normally epigenetically silenced (23,38,79).

Even if HML-2 plays an important role in disease, proving causation would be quite difficult. The high copy number of very similar elements, some of which are insertionally polymorphic, makes it very difficult to demonstrate that one proviral locus in particular has pathogenic effects, and indeed few studies have tried to tease out which of the 90 or so known 2-LTR proviruses are expressed during disease. These difficulties are exacerbated by our incomplete catalog of HML-2 insertions in humans. It is likely that there are undescribed low frequency proviruses present in human populations, and very

little is known about the frequency of known proviruses, or how their frequency varies in different human ethnicities. A few attempts have been made to link the presence of specific polymorphic proviruses to a specific disease, including schizophrenia and breast cancer (108,109); their results have been largely negative, but these studies are again limited by our lack of a complete or nearly complete catalogue of polymorphic HML-2 loci. It is also possible that disease is a result of interactions between multiple proviruses, or even between endogenous proviruses and some exogenous factor like another virus; as mentioned above, recombination between multiple proviruses is needed for disease in mice and other animals.

The mechanisms by which HML-2 might contribute to disease are also unclear. Two accessory proteins encoded by some HML-2 loci, Rec and Np9, have been implicated as potential oncogenes; in vivo studies have shown specific expression in tumors, and in vitro their expression appears to enhance growth of tumor cells (94,110,111). HML-2 Env has also been shown to have neurotoxic effects when expressed in mice and in human nerve cells, producing an ALS-like phenotype in mice (103). A superantigen encoded by the HML-2 provirus HERV-K18 has also been studied in connection with several autoimmune diseases (112,113).

The most straightforward way HML-2 proviruses might cause disease is, of course, by active replication. Though no infectious HML-2 has ever been found in nature, it is still possible that infectious viruses could be found. There are a number of different potential sources of infectious HML-2. An exogenous population of virus could still be circulating at low frequencies, although it would be quite difficult to detect. Endogenous retroviruses can maintain infectivity for tens or hundreds of thousands of years, although

such infectious proviruses are likely to be quite rare, especially if pathogenic and thus subject to counter-selection, though they might be transcriptionally silenced, preventing such pathogenic effects (19). Diseases that cause reactivation of HERVs could allow production of such viruses; alternately, they could allow generation of chimeric infectious genomes by recombination of non-infectious proviruses, which often occurs in mice and other animals as mentioned above. As previously mentioned, a recombinant HML-2 genome constructed from just three known loci has been shown to be infectious, demonstrating that such a recombinant infectious virus could in theory be generated if the right proviruses are expressed together (90).

Although it is possible that HML-2 viruses are still actively infecting humans, it may be that HML-2 infectious activity has stopped or is at least greatly curtailed in modern humans. This leads to the question of what factors could have contributed to this loss in activity. One possibility is that HML-2 has simply lost one or more of the many co-evolutionary battles it must fight with its host. Many of these battles involve interactions between viral envelope proteins and their cognate host receptor molecules. Avian leukosis viruses (ALVs) and murine leukemia viruses (MLVs) are the best understood examples of this; in both situations, selective pressure on the virus and the host has led to rapid diversification of envelope and receptor genes, leading to a complex of multiple viral strains and host receptor alleles (20). In the case of the xenotropic MLVs (XMVs) the mice seem to have won, as the strains carrying these viruses are typically completely resistant to infection by those viruses (114,115). Unfortunately, the receptor for HML-2 Env is not yet known, although this is a subject of research in our lab and others (116).

Host antiviral restriction factors could also play a role in blocking HML-2 infection. The APOBEC3 (A3) family of proteins inactivate retroviruses and retrotransposons by hypermutation of reverse transcribed single stranded DNA (117). Several HML-2 proviruses show signs of A3 mediated G to A hypermutation, and A3s have been shown to reduce the titer of reconstituted HML-2 viruses in vitro (91,118–120). It is difficult to say how much of an effect these proteins would have in vivo, however, as we do not know what cells are normally targeted by HML-2, or whether those cells express A3 proteins. Another restriction factor that has been shown to have an effect on HML-2 is tetherin, a cell surface protein that can prevent release of newly produced virions, but the same caveats apply to these studies as well (121,122). It would be interesting to investigate the expression levels of these proteins in different cell types, or whether HML-2 infection can induce expression of these factors. TRIM5α, a restriction factor which binds to retroviral capsid proteins and interferes with capsid disassembly and reverse transcription, has also been investigated to see if it inhibited HML-2 infectivity; no effect was seen (91). It might be worthwhile to look at other retroviral restriction factors, such as SAMHD1 and MOV10. The latter is especially interesting, as endogenous expression levels are known to inhibit retrotransposons, but not exogenous retroviruses (123,124).

As mentioned previously, restriction factors can also be viral in origin, such as the mouse *Fv1* gene, derived from an endogenous retrovirus *gag* gene. Fv1 restriction is mediated by binding of Fv1 to viral capsids, and seems to be similar in mechanism to TRIM5α in humans (20,64). Similarly, expression of certain ERV encoded envelope genes confers infection resistance in mice and chickens by binding to receptors, blocking

viral entry (20,66,125). More recently, a HERV-T envelope gene was found in humans that appears to function in the same way, but against a virus, HERV-T, that is no longer actively infecting humans, potentially in part due to the protective effect of the coopted HERV-T gene (68). It is conceivable that something similar could be occurring with endogenous HML-2 *env* genes.

1.6 Endogenous retrovirus age estimation methods

Most of our knowledge of how ERVs cause disease is derived from studies of endogenous retroviruses in other animals, yet without an understanding of the dynamics of endogenization in humans, it is hard to know how relevant such studies are to understanding HML-2 pathogenesis. One proposed model for production of HML-2 loci is that there has been a slow, constant rate of integration and fixation over millions of years, perhaps continuing up until the present day (126–128); alternatively, integration could have taken place in a brief, epidemic-like expansion (23). Such an epidemic is currently taking place in koalas, with concurrent production of endogenous gammaretroviruses, as well as causing a significant disease burden in koalas (129–133). As previously mentioned, there are a number of ways to estimate the age of individual ERV insertions and/or ERV families. These methods fall into two major types: long terminal repeat (LTR) comparison methods, which use LTR sequence divergence as a molecular clock, and comparative genomics methods, which look for the presence or absence of insertions at orthologous sites. Importantly, though none of the available methods of age estimation is entirely reliable, the concordance of multiple lines of evidence allows us to confidently assign approximate ages to ERV loci.

The 5’ and 3’ LTRs of a given provirus are identical at the time of insertion, providing a convenient molecular clock for each integration event (47,51). Assuming a neutral mutation rate post-integration, the time since integration can be calculated from the percent sequence divergence between the two LTRs. Subramanian *et al.* used this method to estimate the average age of the human specific HML-2s as 2.7 (±1.1) million years (46). One limitation is that the variance of the estimates will increase significantly for relatively recent insertions with very little divergence; for the youngest of all, the LTRs will be identical and the method is limited to estimating a maximum age; given the observed divergence rates for HML-2 LTRs, this prevents us from estimating the age of any HML-2 integration younger than ~300,000 years. A second problem with this approach is that gene conversion between the two LTRs can increase their similarity, producing integration time estimates much younger than their true age. For proviruses present in more than one species, such gene conversion events can be detected by the anomalous clustering of orthologous LTR sequences in phylogenetic trees, but for species specific insertions such an analysis is not possible (48,59,134). Lastly, the LTR comparison method is limited to proviruses with two intact LTRs; many proviruses have LTR truncations (or gene conversion as mentioned above), and solo LTRs often greatly outnumber full length proviruses, approximately 10:1 for HML-2(46).

A related algorithm has been developed to date solo LTRs, in which each LTR is aligned to a consensus LTR sequence; the divergence between the two is used to calculate the integration age. This approach will be subject to greater error than 5’-3’ LTR comparisons, as it assumes that all proviruses from a given group are close to identical; since retroviruses exist as quasispecies with significant sequence variation

within a population and the number of viral replication cycles between each endogenization event is unknown, this assumption will almost certainly overestimate insertion ages, on average (46).

A second alternative approach is to identify single nucleotide polymorphisms (SNPs) within a single LTR of a given provirus in different individuals, and use coalescent modeling to estimate the time since the most recent common ancestor of the different LTR alleles; this approach allows significantly more precise dating than either of the above approaches, and does not require both LTRs to be intact, though it is unclear how accurate this approach is (135). Using this method, Jha et al. (136) calculated the age of 3q13.2/K106, a provirus that has no differences between its LTRs in the reference human genome but multiple SNPs within its 3' LTR, as between 91,000 and 154,000 years. This analysis suggests that HML-2s were still active after the emergence of anatomically modern humans.

The above methods all depend on use of LTR sequence divergence as a molecular clock. Another approach is comparative genomics, which looks at orthologous loci in different species; if the divergence time between the species is known, maximum and/or minimum ages can be calculated for individual integrations based on their presence or absence in closely related species (51). The nearest extant relatives of humans are chimpanzees and bonobos, who are believed to have diverged from our common ancestor approximately 6 million years ago (MYA), though much uncertainty about the divergence date remains (137,138). This divergence time places a rough upper limit on the ages of human specific proviruses, although incomplete lineage sorting (ILS) leads to some incongruities. Proviruses that are unfixed in a population during a speciation event

may be lost from one of the diverging populations through genetic drift, but remain in the other population, producing a species specific insertion that, in fact, integrated prior to speciation. Furthermore, though chimpanzees are our closest relatives, the gorilla and chimp divergences occurred close enough in time for ILS to result in significant portions of the human genome being more closely related to their gorilla orthologues than the chimp orthologues, and a similar fraction of the chimp and gorilla genomes are more similar to each other than they are to humans (139–141). Similarly, one HML-2 insertion has been observed in chimps and gorillas, but not humans, and this is presumably due to ILS as well (95). Additionally, Agoni *et al.* identified insertions in the recently published genomes of the Neanderthal and Denisovan archaic humans that may be specific to those populations (142). Again, ILS implies that some species specific insertions may have integrated, but not reached fixation, prior to the human-Neanderthal split, approximately 800-500,000 years ago; however, a high number of species specific insertions would suggest HML-2s were still integrating after the speciation event.

Lastly, the high level of insertional polymorphism within the HML-2 group is indicative of its recent activity. Belshaw *et al.* used simple population genetic modeling to predict what percentage of insertions would be expected to be insertionally polymorphic under the assumption of a constant integration rate from the human-chimpanzee split up to the present day. This percentage was found to not be significantly different from the percentage of insertions experimentally found to be polymorphic, leading the authors to conclude that there may be ongoing infection and integration (127). This study, however, was based on a screen of only 19 individuals. It could be informative to repeat this study using more extensive datasets on HML-2 insertional

polymorphism, and/or with different assumptions about the dynamics of HML-2 integration in the human lineage.

1.7 Limitations of deep sequencing for identifying ERVs

High throughput short read sequencing technologies like Illumina have revolutionized the field of genomics, allowing a massive increase in resequencing and de novo genome sequencing. However, the short read lengths of most current technologies present major difficulties for genome assembly and alignment algorithms. Among the most intractable problems are repetitive DNA and large insertion elements, the first because of difficulties in getting unambiguous alignments, and the second because of difficulties in getting them to align at all. ERVs and other transposon insertions share the problems of both repeats and insertion elements (143). These difficulties often lead alignment and assembly software to simply throw out reads belonging to novel insertion elements; if the insertion is homozygous, the assembler will likely leave a gap in the assembled genome, whereas if it is heterozygous, the resulting alignment will simply look like it is homozygous for the preinsertion allele. Alternately, the software may align such reads to another locus in the reference genome containing a similar element, even if that insertion is not present in the genome being sequenced, leading to a spurious insertion sequence in the new genome sequence. A number of programs have been developed to help with this problem, such as RetroSeq, T-Lex, and BreakDancer (143–145). These programs typically take advantage of paired end sequencing, looking for read pairs that do not align close to each other, as a well as other signs of improper alignment.

The other common approach is to perform targeted re-sequencing; rather than sequencing the whole genome of an organism, it is possible to specifically sequence a set

of insertion elements of interest, and then map them to a published genome. Enrichment strategies include pull downs with specific oligonucleotide probes and PCR amplification with primers specific to the targeted insertion elements. Such libraries have been used to map HIV and ALV proviral insertions in experimental infections, as well as transposons such as Alu elements and LTR retrotransposons (36,146–148).

A number of projects to sequence large numbers of human genomes and catalogue human genetic variation have been started, including the 1000 Genomes Project (1KGP), Human Genome Diversity Project (HGDP), Personal Genome Project, and others, and much of these data are available to the public (149–152). Of these projects, the 1000 Genomes Project is the largest completed study. The 1KGP group has released low coverage (4-6x) genomes of over 2500 individuals from genetically diverse groups across the globe. The subjects were sampled from 26 populations, grouped into 5 superpopulations by ancestry: 5 East Asian (EAS) populations, 5 South Asian populations (SAS), 7 populations of sub-Saharan African ancestry (AFR), 5 European populations (EUR), and 4 indigenous American populations (AMR)(Fig. 1-8, Table 1-2) (150).

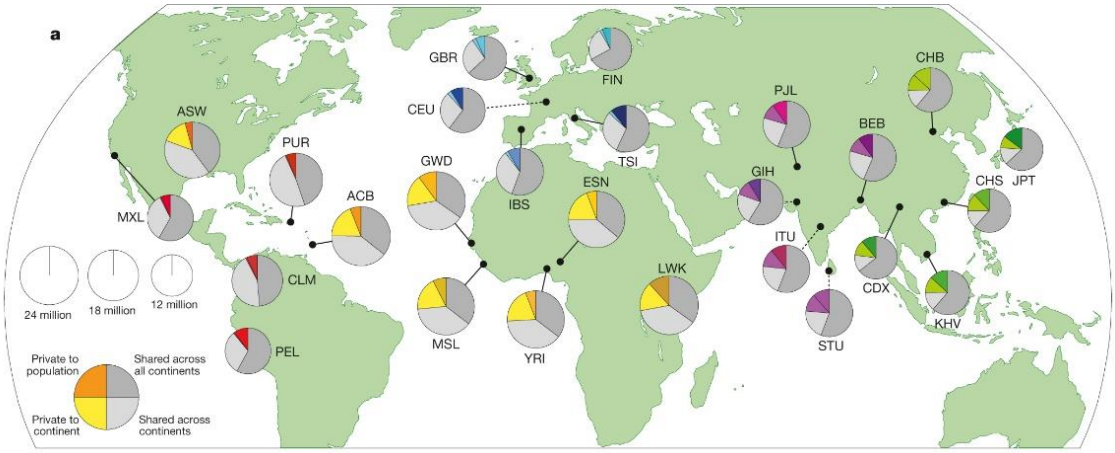


Figure 1-8. Geographic locations of 1000 Genomes Project sample populations
Map showing locations of the 26 populations sampled by the 1000 Genomes Project. Each population is accompanied by a pie chart showing proportions of sequence variants in each population that are either unique to that population (dark color), unique to a superpopulation (light colors), shared between at least two superpopulations (light grey) or shared across all populations (dark grey).Republished with the permission of Nature Publishing Group (150); permission granted through the Copyright Clearance Center, Inc.

Table 1-2. 1KGP and HGDP sample populations and abbreviations

1000 Genomes Samples	Abbreviation
East Asian Ancestry	**EAS**
Chinese Dai in Xishuangbanna, China	CDX
Han Chinese in Bejing, China	CHB
Japanese in Tokyo, Japan	JPT
Kinh in Ho Chi Minh City, Vietnam	KHV
Southern Han Chinese, China	CHS
South Asian Ancestry	**SAS**
Bengali in Bangladesh	BEB
Gujarati Indian in Houston,TX	GIH
Indian Telugu in the UK	ITU
Punjabi in Lahore,Pakistan	PJL
Sri Lankan Tamil in the UK	STU
African Ancestry	**AFR**
African Ancestry in Southwest US	ASW
African Caribbean in Barbados	ACB
Esan in Nigeria	ESN
Gambian in Western Division, The Gambia	GWD
Luhya in Webuye, Kenya	LWK
Mende in Sierra Leone	MSL
Yoruba in Ibadan, Nigeria	YRI

Table 1-2, cont. 1KGP and HGDP sample populations and abbreviations

1000 Genomes Samples	Abbreviation
European Ancestry	**EUR**
British in England and Scotland	GBR
Finnish in Finland	FIN
Iberian populations in Spain	IBS
Toscani in Italia	TSI
Utah residents with Northern and Western European ancestry	CEU
American Ancestry	**AMR**
Colombian in Medellin, Colombia	CLM
Mexican Ancestry in Los Angeles, California	MXL
Peruvian in Lima, Peru	PEL
Puerto Rican in Puerto Rico	PUR
Human Genome Diversity Project (HGDP) Samples	
Population	
African Ancestry	
Mbuti_Pygmy	
Mozabite	
Pathan	
San	
Asian Ancestry	
Cambodian	
Yakut	
American Ancestry	
Maya	

Republished with permission from PNAS (153).

The Human Genome Diversity Project (HGDP) has somewhat similar aims to the 1KGP, but has focused on maximizing the genetic diversity of its sample population rather than aiming for sheer sample size like the 1KGP, and thus has specialized in obtaining samples from indigenous populations with low admixture from other people groups, especially highly divergent populations such as the San of southern Africa and Mbuti Pygmies of central Africa (151,154). Though many of the samples in the HGDP panel have not yet been fully sequenced, a subset that captures much of the panel's genetic diversity has been sequenced and the data is available through the NCBI Short Read Archive (SRA)(Table 1-2)(155).

We are also interested in investigating the activity of HML-2 in other primates, especially those closely related to humans; as mentioned above, comparative genomics can be very helpful in estimating the age of particular ERV insertions and in understanding the overall evolutionary history of past retroviral infections. Though HML-2 proviruses are found in all Old World monkeys and apes, very little is known about HML-2s in other primates. Some chimpanzee specific HML-2 proviruses are known, but it is not known whether they are still active in chimpanzees, and research into HML-2s in other primates has been very limited (95,156–159). The amount of genome information for primates is fairly limited when compared to humans, but genome assemblies of varying quality have been generated for all the great ape species, and some relatively small scale projects to more fully explore genetic variation in primates have begun (160–164). Most notably, the Great Ape Genome Project recently released high coverage whole genome sequences for a panel of 79 individuals covering all 6 non-human great ape species: the common chimpanzee and bonobo (*Pan trogolodytes* and *P.*

paniscus), the Western and Eastern gorillas (*Gorilla gorilla* and *G. beringei*), and the Bornean and Sumatran orangutans (*Pongo pygmaeus* and *P. abelii*)(165).

1.8 Rationale for study

As can be seen in the above review, many of the questions about HML-2 remain unanswered because of our imperfect knowledge of the total extent of HML-2 integration in human genomes. Recent advances in genome sequencing technology and the growing store of publicly available genome data have opened up many new avenues to expand our knowledge. The goal of this project was to make use of these advances to identify HML-2 integrations in humans that are not present in the reference genome, which we refer to as 'non-reference' integrations, sequence and characterize any previously undiscovered insertions, and measure the frequency distributions of polymorphic HML-2s in human populations, including previously both previously described sites and sites newly discovered in this study. These data should help determine whether replication competent HML-2 sequences still exist, either as proviruses, circulating exogenous viruses, recombinant viruses, or in other species. Identifying such undescribed proviruses should also improve our understanding of the potential for HML-2s to cause disease, and may provide insights into human and viral evolution, both as unambiguous markers of ancestry, and as a record of past host-virus molecular interactions.

We expect that most non-reference HML-2 insertions will be quite rare, and will not be picked up by standard whole genome sequencing methods because of the alignment algorithm limitations mentioned above. To solve this problem, we developed two methods to identify unmapped HML-2 integrations from whole genome sequencing data, and have used these methods to mine WGS data from the 1000 Genomes Project and the

Human Genome Diversity Project. Additionally, we developed an *in* silico genotyping method to screen these samples for the presence or absence of known HML-2 insertions, which should provide us with good population level frequency estimates for these insertions.

Though HML-2s may no longer be replicating in humans, it seemed likely to us that they may have maintained activity in other primate species, perhaps even in our closest relatives, the chimpanzees and gorillas. Some chimpanzee specific proviruses have been reported (156–158), but our initial investigations of the chimpanzee reference genome indicated low levels of chimpanzee specific HML-2 integrations compared to humans. In contrast, very little has been published on HML-2s in gorillas; some earlier unpublished work in our lab suggested that there might be some gorilla-specific HML-2 insertions to be found, so we decided to apply the methodologies we developed for human genomes to gorilla genomic data as well. We searched gorilla genome sequences from the Great Ape Genome Project for HML-2 insertions, and also searched for HML-2s in the recently published long read gorilla genome assembly. The sequences of these gorilla-specific HML-2 insertions were analyzed along with the human-specific insertions discovered above as well as previously identified HML-2 insertions to understand the timing and extent of HML-2 activity in humans and other primates, the prospects for ongoing HML-2 replication in primates, and other aspects of HML-2 evolution.

Chapter 2: Materials and Methods

2.1 Data analyzed

Illumina whole genome sequence (WGS) data were obtained from 1000 Genomes Project (1KGP) samples, including a total of 2,484 individuals from 26 populations (149), and 53 individuals in 7 populations from the Human Genome Diversity Project (HGDP)(154). 1KGP data were downloaded in aligned SAM (Sequence Alignment/Map) and/or BAM (Binary SAM) format (ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/data/)(166). HGDP data were processed as described (154), and are available at the NCBI Sequence Read Archive (SRA) under accession SRP036155. Individual BAMs were merged using the Genome Analysis Toolkit (167) by population (1KGP) or dataset (HGDP). The 1KGP populations ranged from 66 to 113 individuals each and had an effective coverage of ~1,067x ±207.4x per pooled BAM; 53 HGDP samples were pooled to a single BAM of ~429x. Gorilla WGS data were obtained from Great Ape Genome Project (GAGP) samples (168), including 21 individuals. 3 subspecies were sampled: 18 Western lowland gorilla samples (*Gorilla gorilla gorilla*), 1 Cross river (*G. gorilla diehli*), and 2 Eastern lowland (*G. beringei graueri*). GAGP data were downloaded from the NCBI SRA website, accession SRP018689, in unaligned FASTQ format. The gorilla reference genomes gorGor4 and gorGor5 were downloaded from the UCSC Genome Browser site (169).

2.2 Gorilla sequence alignment

Gorilla FASTQ files were trimmed to remove low quality sequence, short reads (<50 bp) and unpaired reads using Trimmomatic (170). Trimmed sequences were aligned to the gorGor5 build of the gorilla reference genome using Bowtie2(171). Default- settings

for Bowtie2 were used with the exception of using the 'very-fast' option. Alignments were output in SAM format.

2.3 HML-2 discovery from LTR junctions

Unmapped reads were retrieved from BAM files using SAMtools (166) from all 53 HGDP samples, 825 1KGP samples (≥10 samples per 1KGP population) and 21 gorilla samples from the GAGP, and searched for sequence that precisely matched the HML-2 LTR edges using custom scripts. For the 5' LTR edge, the following sequences and their reverse complements were searched for: TGTGGGGAAAAGCAAGAGA, TGTGGGGAAAAGAAAGAGA, and TGTGGGGAGAAGCAAGAGA. For the 3' LTR edge, the following sequences and their reverse complements were searched: GGGGCAACCCACCCCTACA, GGGGCAACCCACCCCTTCA, and GGGGCAAGCCATCCCTTCA. Sequences with <10bp of non-LTR flanking sequence were excluded. Reads matching HML-2 junctions present in the reference genome (hg19 for human data, gorGor5 for gorilla data) were removed. Candidate reads were then aligned to a reference genome (again hg19 for humans, gorGor5 for gorillas) using the UCSC BLAST-Like Alignment Tool (BLAT) to identify their genomic position (172). For the human samples, flanking sequences with no match to the reference genome, <90% identity, or that aligned to gaps or multiple genomic positions were searched against the chimpanzee (panTro4) and gorilla (gorGor3) references, and available human WGS data from the NCBI Trace Archive to identify insertions in structurally variable or unassembled regions of the human genome. This analysis was not performed for the gorilla samples.

2.4 HML-2 discovery from read pair data

The analysis described below was performed by Julia Wildschutte, a collaborator in the Jeffrey Kidd lab at University of Michigan.

Candidate non-reference HML-2 LTRs were identified using RetroSeq (143). LTR-supporting read pairs were identified by running 'discover' on individual BAM files, with read alignment to the HML-2 LTR5Hs consensus elements from RepBase (173) and previous reports (90,91). A BED file of RepeatMasker (174) HERV coordinates from the GRCh37/hg19 reference was used for exclusion of previously annotated sites. RetroSeq 'call' was applied to the merged BAMs (above), requiring a read support of ≥2 for a call. A maximum read depth per call of 10,000 was applied for the increased coverage of the BAMs. To capture only novel insertions, calls within 500 bp of an annotated HML-2 LTR were excluded. Other RetroSeq options were kept at default values.

For each RetroSeq candidate call, supporting read pairs and split reads within 200 bp of the assigned break were extracted from each sample and subjected to de novo assembly using CAP3 (175,176). Assembled contigs were subjected to RepeatMasker analysis to confirm the LTR presence and type (i.e., LTR5Hs)(174), then filtered to identify the most likely candidates, requiring separate contigs that contained the respective 5' and 3' HML-2 LTR edges, and the presence of ≥30bp of both LTR-derived and genomic sequence at each breakpoint. We examined contig pairs for the presence of 4-6 bp putative TSDs, but did not require their presence for a call. Output assemblies were aligned to the hg19 reference to confirm the position of the 'empty' site per call.

2.5 Validation and sequencing of HML-2 insertions

Genomic DNA from human samples yielding positive reads was obtained from Coriell or Foundation Jean Dausset-CEPH for PCR validation. Michael Jensen-Seaman generously donated genomic DNA from three captive Western lowland gorillas for PCR validation of gorilla-specific insertions: PR00301 ("Shango," studbook #1123), PR00622 ("Chipua," studbook #1419), and PR00671 ("Billy," studbook #1148).(177) Coordinates for each insertion were based on mapping of assembled contigs or read-captured flanking sequence to the hg19 reference. PCR was performed with 100ng of DNA using primers flanking each site to detect either the empty site or solo LTR alleles. A separate PCR was run to infer a 2-LTR allele with a primer situated in the HML-2 5' UTR paired with a flanking primer (46,47). Capillary sequencing was performed on at least one positive sample. 2-LTR provirus alleles were amplified in overlapping fragments from a single sample and sequenced to ≥4x (46,87), and a consensus then constructed with the read traces from each site.

Table 2-1. Human allele specific primers and DNA samples used for sequencing

Locus	Primer Direction	Sequence	Sample Sequenced
1p13.2	F	CTAGCTGAATTGCTGCGTGA	NA19240
1p13.2	R	GAAAGCACATAGTGCCGTGA	
1p21.1	F	AATAGCCATCAGATCACACTTT	HGDP00855
1p21.1	R	GTAAACAAAGGCTCCAGCAA	
1p31.1c	F	AGGCATTCTGATAGGTATAGGGA	NA18867
1p31.1c	R	ACCCAATTCACTAGCCAATTAGA	
1q41	F	CATGCTACTTGAAGCCAGGA	HG03159
1q41	R	CACGCCCAGCTACAGTTCTA	
3q11.2	F	CCATGTGGCACTGTGAAGAA	HGDP01036
3q11.2	R	CAGCAGCCAATCAAATCACA	
4p16a	F	TGTTGTTACAAGAACAAGTGTGAAAA	NA06984
4p16a	R	GCAGTACGGGCTTACGTTTC	
4p16b	F	AGGGTGAAGTCTGTGGTGG	HG03378
4p16b	R	TTTGGCTTGCTGTTTTGGGA	
5p15.32	F	TCCACCCAAATGTTGCCTTT	HGDP1032
5p15.32	R	CAGTCACCCTCTCTGATGGG	
5q12.3	F	TTCATGGAGCAGTCCATTTAT	NA15885
5q12.3	R	GAGCCTCTCTCCATGCAAC	
5q14.1	F	GCTGTTTGAGAGCCTAGAGCA	HGDP00449
5q14.1	R	GAGCCTTTCCAACAGAGCAG	
6p21.32	F	GGGCTCACAAGAACATCTCC	NA19704
6p21.32	R	TGGGTGGAATTGTTGGTCTT	
6p22.3	F	TGGCACTGTAATTAGGCACA	HG02879
6p22.3	R	ACTTTCATTCATGCACACAGGT	
6q26	F	GACAGCGGTAAACCCAGAAG	HGDP00213
6q26	R	GCCTTTCCTTGGTCTCACTG	
7q36.3	F	CTTGTACCCCACCTA/GGATGT	HG03401
7q36.3	R	CACTGCCCTACGTGAATGTG	
8q24.3	F	ATAAGGCAGCTGAGGCTGAT	HGDP01029
8q24.3	R	TGCAGAAGTTGCCTATCCAG	
10q24.2b	F	TCCTCCACCACTGAAGATCA	*
10q24.2b	R	GCTTCTTAGTAGTGTCCATCAGA	
10q26.3	F	GGGAAGGCATCAGTGAAGTG	HG03539
10q26.3	R	ACGGGCAGAGGTGGAAAC	

* Sequence taken from assembled reads; Insertion presence inferred by PCR.
Reprinted with permission from PNAS (153).

Table 2-1, cont. Human allele specific primers and DNA samples used for sequencing

Locus	Primer Direction	Sequence	Sample Sequenced
11q12.2	F	CCCAACTTTCCCCCTTAAAA	HG00117
11q12.2	R	TGGGAAAAAGTTTGTGAATAACTA	
12q12	F	GGCTGTTGAAACTGCAGTGAA	NA10494
12q12	R	AGGAGAGAGAGGTCAGAGAGCA	
12q24.31	F	GCCAAAGCGGGTAGATCA	NA10969
12q24.31	R	TGACAGTGTACAGTTGACCTTGA	
12q24.32	F	TTATAACACCAGCCCTCTTGC	†
12q24.32	R	ATGGGATTTCTAAAATTCTAATGTCT	
13q31.3	F	TGAAGGTATTAATCATAGAAACAGCAA	NA12340
13q31.3	R	CAGAAATACTGACCCCCAAAA	
15q13.1	F	CGTGCACCTGTAGTCCCA	*
15q13.1	R	TAGGGCTCTGTGATCTTGGC	
15q22.2	F	GCCTGCGCCTTATAGTTGTG	NA10540
15q22.2	R	GGCCTCAGACATGTTTTACCTT	
19p12b	F	TGCATGGGGAGATTCAGAACC	‡
19p12b	R	ATCCATACATTTCTGAGTCCTGA	
19p12d	F	CGACACAAAGGAAGACACAGAG	§
19p12d	R	AAAATTAGCTGGCCATGGTG	
19p12e	F	TCTCTGGGATGCAGCTACAG	HG03378
19p12e	R	AGGCCGAGGGTGCAGTGA	
19q12	F	CAAGTTCAAACAGCCACAGG	HGDP01036
19q12	R	TCAGCTTATCTTGGGGAGAAA	
19q13.43	F	CATCTCCATGGAAGTCCAGA	HGDP01029
19q13.43	R	CTGGTTGTTTCACAGATGTGGT	
20p12.1	F	AGTATGCTACCCAATCCTCCA	NA10969
20p12.1	R	GGGAAACTTGCGCCTCTTTG	
22q11.32	F	AAGACCTGCCTCAGCCACT	NA12878
22q11.32	R	CCATCTATGTGCCAGCTCCT	
Xq21.33	F	TCAGTCTGAAGACAATGACATACTT	HG02879
Xq21.33	R	TCTTACACTGTTTTGTCCCTTTGA	

* Sequence taken from assembled reads; Insertion presence inferred by PCR
† Sequence from Kidd JM, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56-64. Genbank accession AC195745.1
‡Sequence from Genbank accession AY037928
§Sequence from fosmid, 'Homo sapiens FOSMID clone COR2A-DD0002RIQNU_L18 from chromosome 19, complete sequence,' Genbank accession AC245253.1
Reprinted with permission from PNAS (153).

Table 2-1, cont. Human allele specific primers and DNA samples used for sequencing

Locus	Primer Direction	Sequence	Sample Sequenced
dup1	F	ACCTTACGAGAAACACCCAC*	†
dup1	R	TCCTTCACCTTCAAGAGCAGT	
dup2	F	GCAGGCTGTAGTTTGTGATCC	HG03378
dup2	R	TACCTCATTGCCTGTTCCCA	
dup3	F	TGTCAAAGCAAGGCCTCAAC	HG03372
dup3	R	AATGGGCAGCAAATCCAGTG	
dup4	F	TGAGAGAATACAGTGCAGTAGGG	NA19443
dup4	R	CCTCAAGCAATCCAACCACC	

*Primer within LTR sequence, not flanking sequence
†Sequence reconstructed from NCBI Trace Archive reads, TI numbers 1468770310, 1737141807, 1743912617
Reprinted with permission from PNAS (153).

Table 2-2. Gorilla allele specific primers and DNA samples used for sequencing

Locus	Primer Direction	Sequence	Sample Sequenced
1p34.2	F	TGAAAGCAGCCACAGACAAT	PR00622
1p34.2	R	CTGAGGCTTGGAAGGATTAAGT	
1p36.13	F	CTGTGACAAGCTCCAGAGGA	PR00301
1p36.13	R	AGGGATGGATTGGAAGAGCT	
1q43	F	GACCCCTCCAAACTTTCAGC	PR00301
1q43	R	AGGAGGAATCTGGTAATGGGT	
2p13.3	F	GCAAGGATGTGGGGAAACTG	PR00671
2p13.3	R	TTTCTTTGGCTTCTCTCTCTGA	
3p12.2	F	TGGATGGTAGCAGGAGAGAA	PR00671
3p12.2	R	CAGCCTAGACTTTTGATCGCC	
3p22.3	F	GTAGCATCCTCTCCCACCAA	PR00301
3p22.3	R	GGACAGAAGGACAGATGGGA	
3q23	F	TTGGGAGTTTCAAGCCCTCT	PR00622
3q23	R	ACTTTACATTTCCAGTTTTGCCT	
4q28.1	F	TCCATCATGTTCTCCCCTCAC	PR00301
4q28.1	R	GGAAACCACTGAAAGGCTTTAAG	
5q23.2	F	TCTCTAATCCATTGCTACCCCT	PR00671
5q23.2	R	GAGATCGCACCCCTGTACTC	
6p12.3	F	AGCCCAACCAAGTAATAAGAGAA	PR00301
6p12.3	R	GGACTCAAGCTATCCTCCCG	

Table 2-2, cont. Gorilla allele specific primers and DNA samples used for sequencing

Locus	Primer Direction	Sequence	Sample Sequenced
6p22.1	F	AAGCCAGTCACAGCACAAAT	PR00301*
6p22.1	R	TGATCCATCACCACAATCAAGG	
7p13	F	GTCTCGAACTCCCAACCTCA	PR00622
7p13	R	ATTTGTTGGCTTGCCAGACC	
7p21.2	F	CCTGCCCCTGACTCTGTTAA	PR00301
7p21.2	R	TCAGTTTTGACAGGCACGTT	
8p22	F	CAGTGGCAAGGTTTGAGAGA	PR00671
8p22	R	AATATTGCAGGTCTGGTTCCA	
8q22.2	R	AGGTACTCTCTCACAGGCTT	PR00622*
8q22.2	F	CAAATTTAAACATGCCGGCCA	
9p13.3	F	TCCCCTTTTGCTTGACTCTCA	PR00301
9p13.3	R	TCCATTTGTTCATTGCTGGCA	
10p11.23	F	TCCTCCCTTTTAATGGCTGAGT	PR00301
10p11.23	R	AACTCAACGCATCCACAGTG	
10q11.21	F	GGGGTCCTGATAACTGCACT	PR00671
10q11.21	R	TCTCCTGACCTCCTGATCCA	
11q21	F	TCGTGTACAAGTTGAGTATCCCT	PR00301
11q21	R	CCCCAGCACCTTTTAGCATA	
11q22.1	F	AGAGCCTCGTGATTTAACTGT	PR00301
11q22.1	R	CACTCGGCCTTAATCGTTTATTG	
12q12	F	GACCTCACTTCTCACAACACA	PR00301
12q12	R	TCCCAGGACCAATCTTTCACT	
14q23.1	F	TGAATGACACAATACCCAGCT	PR00622
14q23.1	R	TACCCAAAAGCCATTCCCCT	
18q22.1	F	TCCAATCCATGAGCAGGAGA	PR00622
18q22.1	R	CCAGATAAACCAGGCAAGGA	
19q13.2	F	CGAGAATACCCACAAAGTCAGC	PR00622
19q13.2	R	GGTCTCACAAGCCCTCAGTA	
20p13	F	CACAGTGGAGATGGTGTTCC	PR00622
20p13	R	CAAGGGCAATAAAGGGCTGG	
20q13.33	F	TGGTGGTGTTATCTGGGACA	PR00671
20q13.33	R	CCTCCCAGTTTGTACCTGGA	
21q22.3	F	AGGGTGACAGATGAGCCTTT	PR00671
21q22.3	R	TGTCTTCTTGAGGGCCACAT	

*Validated by PCR and partial sequencing. Full sequence obtained from gorGor5 genome assembly

2.6 In silico genotyping of human proviruses

The analysis described below was performed by Julia Wildschutte, a collaborator in the Jeffrey Kidd lab at University of Michigan.

Genotyping of human proviruses was performed for both reference (hg19) and non-reference unfixed insertions. Reference and alternate alleles were recreated based on ±600bp flanking each insertion point. The reference insertion at 7p22.1 is present as a tandem duplication of two proviruses that share a central LTR (47,93), and was treated as a single insertion (chr7:4,622,057-4,640,031). Nine of the 36 non-reference loci could not be aligned to the hg19 reference and were excluded from genotyping: insertions within duplicated segments ('dup1-4'); of unusual assembled structure (10q24.2b and 15q13.1); or that could not be mapped to the hg19 assembly (10q26.3, 12q24.32, and 22q11.23b). For each sample, genotype likelihoods were then determined based on re-mapping of those reads to either allele, with error probabilities based on read mapping quality (178). Samples without remapped reads were not genotyped at that site. Insertion allele frequencies were calculated for all genotyped samples as the total number of insertion alleles divided by total alleles.

2.7 Identification of HML-2 LTR5Hs insertions from reference genome assemblies

Gorilla specific 2-LTR proviruses were identified by searching the most recent gorilla genome build (gorGor5) with the UCSC BLAT alignment program, using as a query the sequence of 3q23, a full length gorilla-specific provirus we identified in the GAGP data and fully sequenced (see above). The proviruses identified were confirmed to be gorilla-specific by searching for the orthologous preintegration sites in the human and chimpanzee reference genomes using BLAT.

To identify human, chimpanzee and gorilla specific solo LTRs from their respective genome assemblies, as well as solo LTRs specific to the human-chimpanzee clade, and solo LTRs with orthologues in all three species, coordinates of all annotated HML-2 LTR5Hs insertions were downloaded from the RepeatMasker track of the human (hg38 build), gorilla (gorGor5 build), chimpanzee (panTro5 build) and orangutan (ponAbe2) reference genomes on the UCSC Genome Browser. Sequences <900 bp or corresponding to known 2-LTR proviruses were excluded. Species-specific loci were identified using BEDtools (179) to filter out sites present in at least one of the other three species; coordinates were converted from one genome to another as necessary using the UCSC Genome Browser's liftOver genome conversion tool (169). The same approach was used to identify loci present in both humans and chimpanzees, but not gorillas. Lastly, the remaining sites from the hg38, gorGor5, and panTro5 genomes were filtered with BEDtools to remove any sites not present in all three genomes to generate a set of shared orthologous solo LTRs (179).

2.8 Phylogenetic analysis

Sequences of both solo LTRs and proviral 5' and 3' LTRs were aligned to the consensus HML-2 using ClustalX (180) or MUSCLE (181), and the alignment then edited and truncated LTRs removed (46). Neighbor-joining trees were generated from the remaining insertions using MEGA6 or MEGA7(182). The Kimura 2 parameter model was used for branch length estimation, with α of 2.5 and deletions treated pairwise (183). Support for trees was assessed using 1000 bootstrap replicates.

2.9 Molecular clock age estimation

LTR divergence was used to infer the time since insertion, normalized to a neutral mutation rate of 0.24 to 0.45% per million years (46,47). Briefly, the nucleotide differences were totaled between proviral 5' and 3' LTRs, and the total divided by the LTR length, treating gaps >2bp as single changes. The percent divergence was then divided by the upper and lower bound mutation rates for age range estimation in my. For solo LTRs, neighbor-joining trees of solo LTRs for each species were separately created using MEGA6, and the divergence from the nearest neighboring sequence was used as a proxy for 5'-3' LTR divergence; insertion age was calculated from this divergence as above. P-values were calculated with both Mann-Whitney and Kolmogorov-Smirnov tests, as it seemed likely that the shape of the underlying frequency distributions would be different, which might cause problems with the Mann-Whitney test, as it assumes all samples have the same frequency distribution shape.

2.10 ORF identification

Full length and nearly full length open reading frames were identified for all 2-LTR proviruses using the NCBI ORF Finder tool (184). The presence or absence of g*ag, pro, pol, env* and *rec* open reading frames was determined for each provirus, using sequences from the HERV-Kcon consensus genome as a reference (91). Genes were classified as intact if >95% of the ORF remained free of nonsense or frameshift mutations. Peptide sequences of predicted proteins were queried against the UniProtKB/SwissProt protein database using BLASTP to confirm their identity as HERV-K *gag, pro, pol* or *env* (185,186).

Chapter 3: Mining human genomic data for undiscovered HML-2 insertions

3.1 HERV-K(HML-2) insertions discovered from WGS data

At the start of this study, the number of human genomes analyzed for unfixed HML-2 proviruses was fairly small, limiting discovery of elements not present in the human reference genome, or 'non-reference' elements, to those with relatively high allele population frequencies. Our goal was to use the extensive available WGS data in the 1KGP and HGDP collections to identify relatively rare polymorphic non-reference HML-2 insertions. In principle, the raw data for these projects contain all the required information to extract these sequences; however available algorithms for generating such calls did so by aligning short Illumina reads to a reference genome, and therefore will generally fail to directly detect large insertions or insertions located within genomic regions not present in the reference genome used. We therefore applied two approaches to identify candidate non-reference HML-2 insertions in the raw read data for these collections (Fig. 3-1).

First, we mined unmapped reads for evidence of LTR-genome junctions captured in reads that failed to map properly when aligned to the human reference genome. While single nucleotide polymorphisms, and small insertions and deletions are efficiently detected by high throughput, short read sequencing approaches such as that used by Illumina sequencers, large, 1-10 kb insertions like ERVs are far longer than the 100-150 base pair reads generated by these methods; thus, reads containing sequence from such insertions will typically either fail to map to the reference genome, or will be aligned to similar sequences elsewhere in the reference genome.

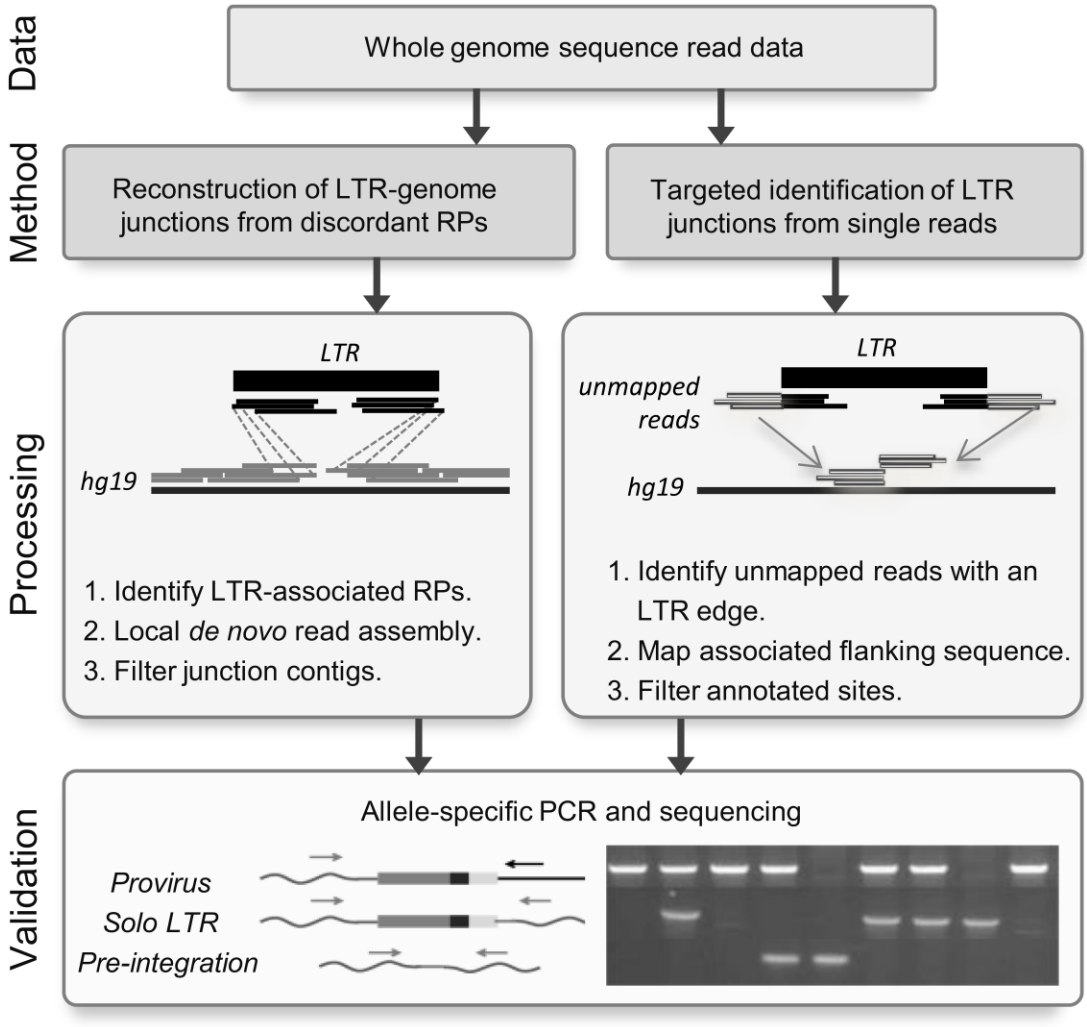


Figure 3-1. Approaches for the detection of non-reference HML-2 insertions from WGS read data.

Illumina reads were processed by one of two methods. (Left) Read pairs (RPs) were identified that have one read mapped to the genome (gray) and mate to reads that map to the sequence matching the HML-2 LTR consensus (black). Supporting reads from each site were extracted and subjected to local assembly, and the resulting contigs were analyzed for the presence of LTR–genome junctions. (Right) Unmapped reads from each sample were identified that contained a sequence corresponding to the LTR edge, and the cognate sequence was then used to determine candidate integration positions from genomic data. (Bottom) PCR and capillary sequencing were used to validate candidate insertions in reactions that used flanking primers (gray arrows) to detect the presence of a solo-LTR or empty site, or a flanking primer paired with an internal proviral primer (black arrow) to infer the presence of a full-length allele. Representative products are shown in a genotyping gel to the right. Republished with permission from PNAS (153).

Reads crossing the junction between the proviral LTR and the flanking genomic DNA are especially likely to fail to map properly, as these junctions are unique to each provirus. We took advantage of this property by selecting only the reads that failed to map in 825 aligned genomes from the 1KGP collection (at least 10 samples per 1KGP population, see table 1-2) and all 53 samples from the HGDP collection, and searching for reads containing such LTR-genome junctions. After excluding reads that could be aligned to annotated HML-2 junctions, we identified 36 loci with apparent non-reference HML-2 insertions, including the previously described K113 provirus (19p12b in our nomenclature, Table 1-1).

A second approach, developed by our collaborators Julia Wildschutte and Jeffrey Kidd, identified insertions based on read pair signatures using the program RetroSeq (143) (Fig. 1, *left*). This approach again depends on the anomalous mapping of reads from non-reference insertions; however, rather than searching for unmapped single reads containing LTR-genome junctions, it looks for discordantly mapped paired reads (e.g. two sequences reading from either end of a single DNA molecule) where one read aligns to a site in the reference genome with no known insertion of the insertion element of interest (in this case, HML-2), and the other read aligns to a reference HML-2 sequence. To improve the sensitivity of detection of insertions present in multiple samples, we combined read data within a population (1KGP) or study (HGDP)(176). Excluding calls within ±500bp of a reference HML-2 sequence, we obtained 140.3 ± 56.1 candidate calls per pool. Next, we applied a *de novo* assembly approach to insertion-supporting reads to reconstruct the LTR-genome junction for as many sites as possible (176). Given the size of HML-2 LTRs (~968 bp per LTR), we inferred the presence of an insertion based on

the presence of separately assembled 5' and 3' breakpoints. This requirement reduced false positives, for example as caused by SVA (SINE-R) elements which contain a fragment derived from an ancient HML-2 provirus, and thus have high identity to bases 1-329 of the HML-2 LTR. This approach confirmed 29 the 36 candidate HML-2 insertions identified with the junction discovery approach; of the remaining 7, 6 were within large structural variants not present in the hg19 reference genome, and one (1p31.1c) was only found in a single 1KGP sample not present in the dataset used for the RetroSeq analysis. Our final call set included 17 insertions identified in recent reports from Marchi *et al*. and Lee *et al* (128,187). The nomenclature for all sites is as maintained in those studies and as in other previous reports (see Table 1-1)(46).

These two approaches have complementary strengths and weaknesses. The LTR junction discovery method is highly sensitive, and capable of identifying integrations in genomic regions excluded from the reference assembly (for example large deletions, centromeric and other repetitive regions, or large segmental duplications), which will not show up in RetroSeq, as it depends on reference-mapped reads to determine the location of each integration.

However, RetroSeq is much faster and less labor intensive than the junction discovery approach, and should in principle be able to identify more divergent sequences, as the junction discovery approach is dependent on finding reads with exact matches to the conserved LTR edge sequences, whereas RetroSeq can find reads that align across the entire LTR region, including reads with mismatches near the ends. In practice, no such divergent insertions appeared in our study, but it could be important in studying elements with less sequence conservation than HML-2.

As the high throughput RetroSeq pipeline ensured that we would have at least some coverage of all 2484 individuals within the 1000 Genomes dataset, we decided to screen a more targeted subset of the samples with the junction discovery pipeline. We screened 825 1KGP samples, with at least 10 individuals from each of the 26 subpopulations (Table 1-2), ensuring good coverage of all the major geographic regions. We looked at a much higher number of individuals from the populations with sub-Saharan African ancestry, as previous studies of genomic variation indicate much greater sequence diversity within African genomes relative to other populations. We also mined all samples from the HGDP collection, which includes a very high level of genetic diversity, including samples from extremely divergent populations such as the San people of southern Africa.

3.2 Validation and sequencing

We validated the presence of 34 of the 36 candidate insertions by PCR and sequencing of DNA from one or more individuals predicted to have the insertion, obtained either from Coriell's 1000 Genomes DNA panel or from the Foundation Jean Dausset-CEPH for HGDP samples (Tables 2-1 and 3-1). The remaining two sites (at 10q24.2 and 15q13.1) were predicted to have an unusual inverted repeat structure based on assemblies of supporting reads at either site, and could not be conclusively confirmed by sequencing, possibly due to hairpin formation (Fig. 3-2). For the 34 validated non-reference sites, we confirmed 29 sites as having solo LTRs and 5 sites with 2-LTR proviruses, including the previously described K113 provirus at 19p12b (Table 3-1).

Table 3-1. Non-reference HML-2 insertions in human genomes

Locus	Coordinates (hg19, starting)	Alias*	Alleles†	ORFs and other properties	First report in humans
1p13.2‡	chr1:111,802,592	De5;K1	LTR, pre		Lee (187)
1p21.1‡	chr1:106,015,875		LTR, pre		Marchi (128)
1p31.1c	chr1:79,792,629		LTR, pre		This study
1q41	chr1:223,578,304	K2	LTR, pre		Marchi
3q11.2	chr3:94,943,488		LTR, pre		This study
4p16c	chr4:9,603,240	K6	LTR, pre		Marchi
4p16d	chr4:9,981,605		LTR, pre		This study
5p15.32	chr5:4,537,604		LTR, pre		This study
5q12.3	chr5:64,388,440	Ne7;K12	LTR, pre		Marchi
5q14.1	chr5:80,442,266	De6/Ne1; K10	LTR, pre		Marchi
6p21.32	chr6:32,648,036		LTR, pre		Marchi
6p22.3	chr6:16,004,859		LTR, pre		This study
6q26	chr6:161,270,899	De2;K12	LTR, pre		Marchi
7q36.3	chr7:158,773,385		LTR, pre		This study
8q24.3c	chr8:146,086,169		prov, pre	*gag* and *pro*	This study
10q24.2b	chr10:101,016,122	De12			This study
10q26.3§	chr10:134,444,012		LTR, pre		This study
11q12.2	chr11:60,449,890	De4;K18	LTR, pre		Marchi
12q12	chr12:44,313,657	Ne6;K20	LTR, pre		Marchi
12q24.31	chr12:124,066,477	K21	LTR, pre		Marchi
12q24.32§	chr12:127,638,080 -127,639,871		LTR, pre	deleted in hg19; from fosmid CloneDB: AC195745.1	Kidd (188)
13q31.3	chr13:90,743,183	Ne2;K22	LTR, pre		Marchi
15q13.1	chr15:28,430,088			Inverted repeat structure	This study
15q22.2	chr15:63,374,594	K24	LTR, pre		Marchi
19p12b	chr19:21,841,536	De1;K113	prov, pre		Turner (87)

*Reported originally in the sequenced Neandertal (Ne) or Denisovan (De) by Agoni et al.(142) or Lee et al.(189), or in modern humans (K) by Marchi et al.(128) or Lee et al.(187)

†Alleles detected. LTR, solo LTR; pre, preinsertion site; pro, 2-LTR provirus.

‡Previously PCR validated as solo-LTR by Lee et al (187).

§Insertion is located within an encompassing structural variant not present in the hg19 reference.

Adapted with permission from PNAS (153). Column 5 removed, 'ORFs' column added.

Table 3-1, continued. Non-reference HML-2 insertions in human genomes

Locus	Coordinates (hg19, starting)	Alias*	Alleles†	ORFs and other properties	First report in humans
19p12d	chr19:22,414,379		prov, pre	deletion in 5' LTR; *pro*	Lee (187)
19p12e	chr19:22,457,244	De11	prov, pre		This study
19q12	chr19:29,855,781	De3;K28	LTR, pre		Marchi
19q13.43	chr19:57,996,939	Ne5	LTR, pre		This study
20p12.1	chr20:12,402,387	De14*;K30	LTR, pre		Marchi
22q11.23b§	chr22:23852639-23852640	De7;K16	LTR, pre	maps to Hg38 alt locus scaffold 22_KI270878v1_alt	This study
Xq21.33	chrX:93,606,603	De9	pro,pre	*gag, pro* *, pol, env*	This study
Dup 1§	Not determined		LTR	Maps to centromere associated duplications	This study
Dup 2§	Not determined		LTR, pre	Maps to duplicated regions within FAM86 and ALG1L2	This study
Dup 3§	Not determined		LTR, pre	Maps to segmental duplications on chr1	This study
Dup 4§	Not determined		LTR, pre	Deletion in hg19; empty site within fosmid CloneDB: AC232224.2	This study

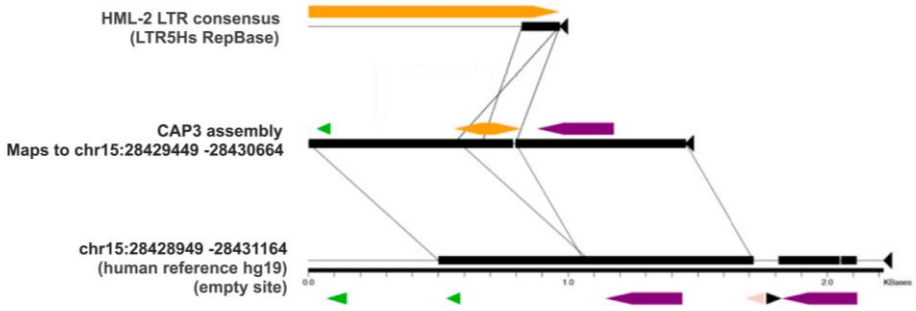
*Reported originally in the sequenced Neandertal (Ne) or Denisovan (De) by Agoni et al.(142) or Lee et al.(189), or in modern humans (K) by Marchi et al.(128) or Lee et al.(187)
†Alleles detected. LTR, solo LTR; pre, preinsertion site; pro, 2-LTR provirus.
‡Previously PCR validated as solo-LTR by Lee et al (187).
§Insertion is located within an encompassing structural variant not present in the hg19 reference.
Adapted with permission from PNAS (153). Column 5 removed, 'ORFs' column added.

A.

>HML-2_LTR_consensus
tgtggggaaaagcaagagagatcagattgttactgtgtctgtgtagaaagaagtagacataggagactccatttt
gttatgtactaagaaaaattcttctgccttgagattctgttaatctatgaccttacccccaaccccgtgctctct
gaaacgtgtgctgtgtcaactcagggttaaatggattaagggcggtgcaggatgtgctttgttaaacagatgctt
gaaggcagcatgctccttaagagtcatcaccactccctaatctcaagtacccagggacacaaaaactgcggaagg
ccgcagggacctctgcctaggaaagccaggttattgtccaaggtttctccccatgtgatagtctgaaatatggcct
cgtgggaagggaaagacctgaccgtcccccagcccgacacccgtaaagggtctgtgctgaggaggattagtaaaa
gaggaaggaatgcctcttgcagttgagacaagaggaaggcatctgtctcctgcccgtccctgggcaatggaatgt
ctcggtataaaacccgattgtatgctccatctactgagatagggaaaaaccgccttagggctggaggtgggacct
gcgggcagcaatactgctttgtaaagcattgagatgtttatgtgtatgcatatctaaaagcacagcacttaatcc
tttacattgtctatgatgcaaagacctttgttcacgtgtttgtctgctgaccctctccccacaattgtcttgtga
ccctgacacatccccctcttcgagaaacacccacagatgatcaataaatactaagggaactcagaggctggc***ggg***
atcctccatatgctgaacgctggttccccgggtcccctttatttctttctctatactttgtctctgtgtctttttc
ttttccaaatctctcgtcccaccttacgagaaacacccacaggtgTGTAGGGGCAACCCACCCCTACA

B.

>15q13.1_LTR_assembly
cagaaactttgtgtagaagttgaaaagaaaatagtatctataacattaccttatctagaatatatgttatactat
atgctgtagttcaaactgattgcatctgtagatattaattaaatagggacaataccagttcacttaaataaggaa
gtcattgggatttttaataaaaggattatagattgtaggggtgggttgcccctacacacctgtgggtgtttctcg
taaggtgggacgagagatttggaaaagaaaaagacacagagacaaagtatagagaaagaaataa***ggggatcctcc***
atatgctgaacgctggttccccgggtcccctttatttctttctctatactttgtctctgtgtctttttctttcca
aatctctcgtcccaccttacgagaaacacccacaggggcaccagaTGTAGGGGTGGGTTGCCCCTACAatagatt
agcattttcttactcagtagctactaaaatgaattggacaaataaaattattatttatttatttttgagacgga
gtcttgctccgtcacctaggctggagtgcagtggtgcaatctcggctcactgcaagctccgcttcctgggttcac



>10q24.2_LTR_assembly
atagtgttgggatgtgcagcctaatgggaggtatttaggtcatgcaggctccaccctcatgaatggattaatgcc
atttataaaaaggcttgagcctgcaagtttgatctcttgctctttcttgccctctctttgcccttctgccatgag
atgtgtaggggtgggttgcccctacacacctgtgggtgtttctcgtaaggtgggacgagagatttggaaaagaaa
aagacacagagacaaagtatagagaaagaaataagggggacccg***gggaaccagcgttcagcatatgga***ggatcccc
ttatttctttctctatactttgtctctgtgtctttttctttccaaatctctcgtcccaccttacgagaaacacc
cacaggtgTGTAGGGGCAACCCACCCCTACAagatgactcaccaaacaggtccttatgatatgccagcccttcaa
tcttagacttcccagcgtctcgaaccttgagccaacacacttctgttcattataaattacccagtctgatattcc

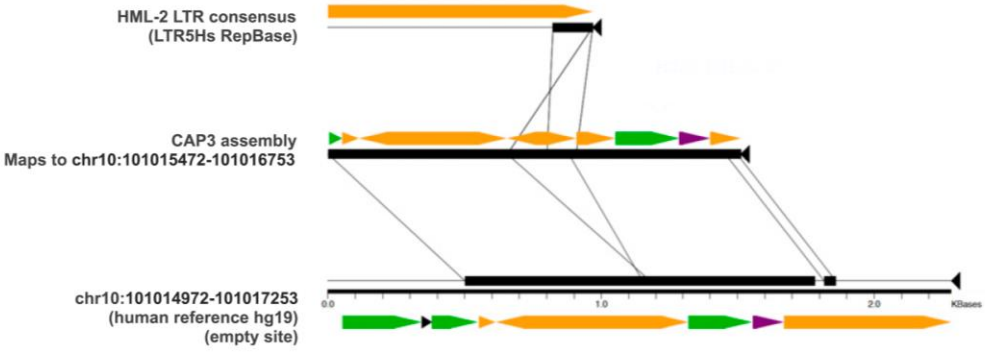
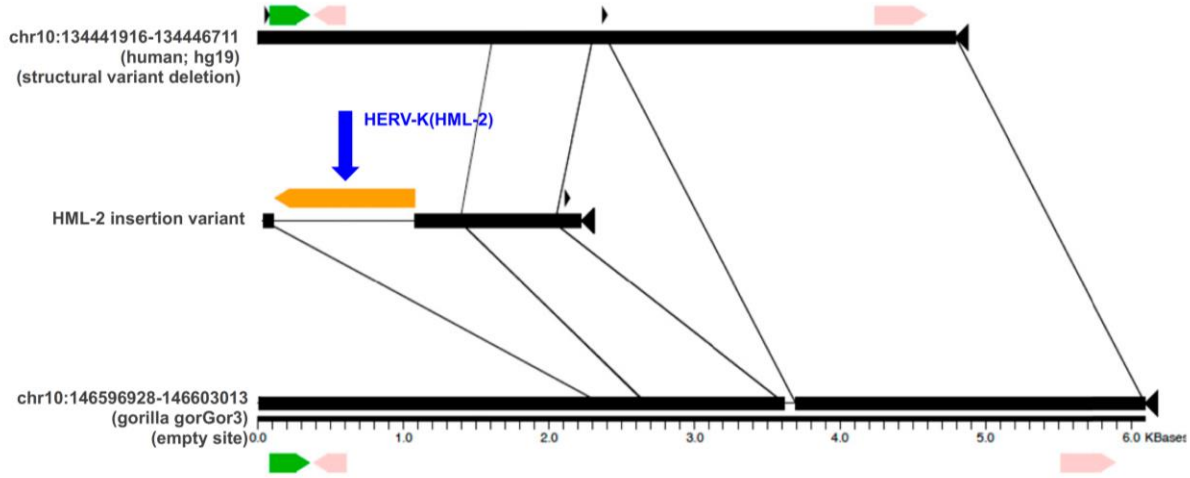


Figure 3-2. Assembled loci with unusual structure.
Detailed nucleotide structure of the assembled contigs at 15q13.1 and 10q24.2. (A) Consensus LTR from the most recent HERV-K HML-2 insertions corresponds to the 968-bp LTR5_Hs consensus (RepBase release 1.03). The last 146 bp (yellow shading) are detected in two of our assembled structures; specific regions are labeled to indicate structural similarities with the two assembled elements as follows. (B) Nucleotide sequences and Miropeats (190) alignments derived in local read assemblies at the 15q13.1 (Upper) and 10q24.2 (Lower) loci. Shading is used to discriminate LTR derived portions from flanking genomic sequence. The yellow shading indicates an LTR5_Hs-matching sequence that assembles as an inverted repeat (red arrows) with a central unique portion (bolded and italicized in all three sequences) at those sites. The presence of a short stretch of the LTR-derived portion that is present in the hg19 reference is shown in blue, and the putative TSDs of 6 bp and 5 bp, respectively, are shaded in green. Block arrows indicate RepeatMasker annotated repeats within each aligned segment. DNA repeats are shown in pink, LTRs are shown in orange, long interspersed elements (LINEs) are shown in green, short interspersed elements (SINEs) are shown in purple. Republished with permission from PNAS (153).

Four of the solo LTRs were situated within duplicated segments and could not be mapped to unique positions in the hg19 reference (‘dup 1’-‘dup 4’); two insertions, at 12q24.32 and 10q26.3, were located within structurally variable regions that are absent from the hg19 reference (Fig. 3-3). One insertion was initially mapped to the reported 9q34.11 locus (128,187); however, comparison of the Sanger reads from its validated LTR-genome junctions revealed unexpectedly low identity in the extended flanking sequence. Our reexamination of this site indicates it maps instead to a region that is not in hg19 but is present in an alternate scaffold in the hg38/GRCh38 assembly at 22q11.23 (Table 3-1). This discrepancy may explain why this particular site has only been previously inferred by reads supporting only the 5’ breakpoint of the integration (128,187).

A. 10q26.3



B. 12q24.32

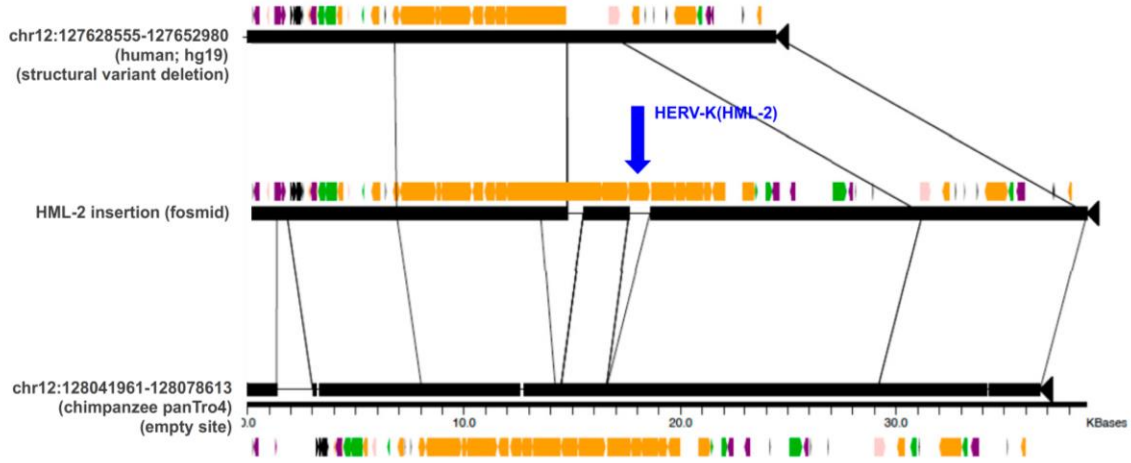


Figure 3-3. Insertions located within genomic structural variants.
Two HML-2 insertions within regions deleted in the reference genome, located within 10q26.3 (A) and 12q24.32 (B) relative to the hg19 reference. Junction sequences corresponding to these insertions were recovered in BLAST searches of the NCBI Trace Archive; the adjacent flanking sequence maps to putative preinsertion sites within nonhuman primate reference genomes (gorilla or chimp as indicated). The alignments for each insertion are arranged to indicate the absence of the insertion in the hg19 segment (Top), the presence of the insertion within the validated variant region (Middle), and the homologous sequence from the reference to which the preinsertion site could be mapped (Bottom). Lines are used to indicate segments of homology between the aligned segments. The LTR at 10q26.3 was sequenced in this study, and the LTR at 12q24.32 was taken from a fosmid clone from Kidd et al.(188) The LTRs corresponding to the identified HML-2 insertions at 10q26.3 and 12q23.32 are labeled in blue with arrows. DNA repeats are in pink, LTRs are orange, LINE elements are green, SINEs are purple. Republished with permission from PNAS (153).

We obtained full sequences for 30 of the 36 candidate insertions in at least one individual predicted to have the insertion (Table 2-1). The remaining insertions were extracted or reconstructed from public sequence databases for subsequent analysis as follows. The full sequence from one locus identified within a duplicated segment was reconstructed from Sanger reads corresponding to that site from the NCBI Trace Archive ('dup1'). The sequence flanking the insertion at 12q24.32 could be mapped to a previously sequenced fosmid clone in a region corresponding to an encompassing deletion of ~14.3 kb in the hg19 reference (188) (Fig. 3-3). Another insertion, corresponding to a 2-LTR provirus, was also from a sequenced fosmid clone (19p12d) as reported (191). The complete sequence of the K113 provirus (19p12b) was from Genbank (AY037928). One solo LTR, 1p31.1c, was detected and validated as a solo LTR in a single individual of the 1KGP Yoruba population from Nigeria. We searched for, but did not find evidence of this rare site in subsequent PCR screens of other samples, including relatives of the initial individual.

3.3 Inferred frequencies of unfixed HML-2 loci

This analysis was performed by Julia Wildschutte in collaboration with Zachary Williams.

In order to better understand the frequency and distribution of insertionally polymorphic HML-2 proviruses in human populations, we performed *in silico* read-based genotyping of 27 non-reference insertions with clear integration coordinates across the 1KGP and HGDP collections, and extended the analysis to include 13 annotated polymorphic HML-2 loci from the hg19 human reference (46,127). 9 insertions in sites

that could not be properly mapped to the hg19 reference genome were excluded. For the analysis, reference (hg19) and alternate alleles representing each HML-2 locus were recreated and then individual genotypes inferred based on the re-mapping of proximal Illumina reads to the reconstructed alleles per site per sample. Given the larger size of the HML-2 LTR (~968 bp) and relatively short reads in these data, 2-LTR and solo LTR insertions were indistinguishable in read-based genotyping alone. Consequently, genotypes were inferred based on the presence or absence of the HML-2 insertion at each locus, regardless of solo LTR or 2-LTR presences at each site.

Allele frequencies of the variable HML-2 insertions present in the reference genome were inferred to be from ~0.25 to >0.99 in genotyped samples (Fig. 3-4, *upper*). Sites with the highest frequencies corresponded to those loci previously reported with a solo LTR or provirus present, but not a pre-insertion site, based on limited PCR screens of those sites (at 1p31.1, 3q13.2, 7p22.1, 12q14.1, and 6q14.1 in Fig. 3-4)(47). This pattern is consistent with variability at these sites based predominantly on the 2-LTR and solo-LTR states. Genotyping of the insertions at 11q22.1 and 8p23.1a (K115) implied the presence of both insertion and pre-insertion alleles, also consistent with PCR screens in other reports (47,87), noting the higher frequency of 8p23.1a/K115 within our samples (~53%) than in those reports (up to ~34% depending on ancestry). Four unfixed reference solo LTRs ranged in frequencies from ~0.25 to as high as ~0.93, also consistent with previous analysis of these sites (127).

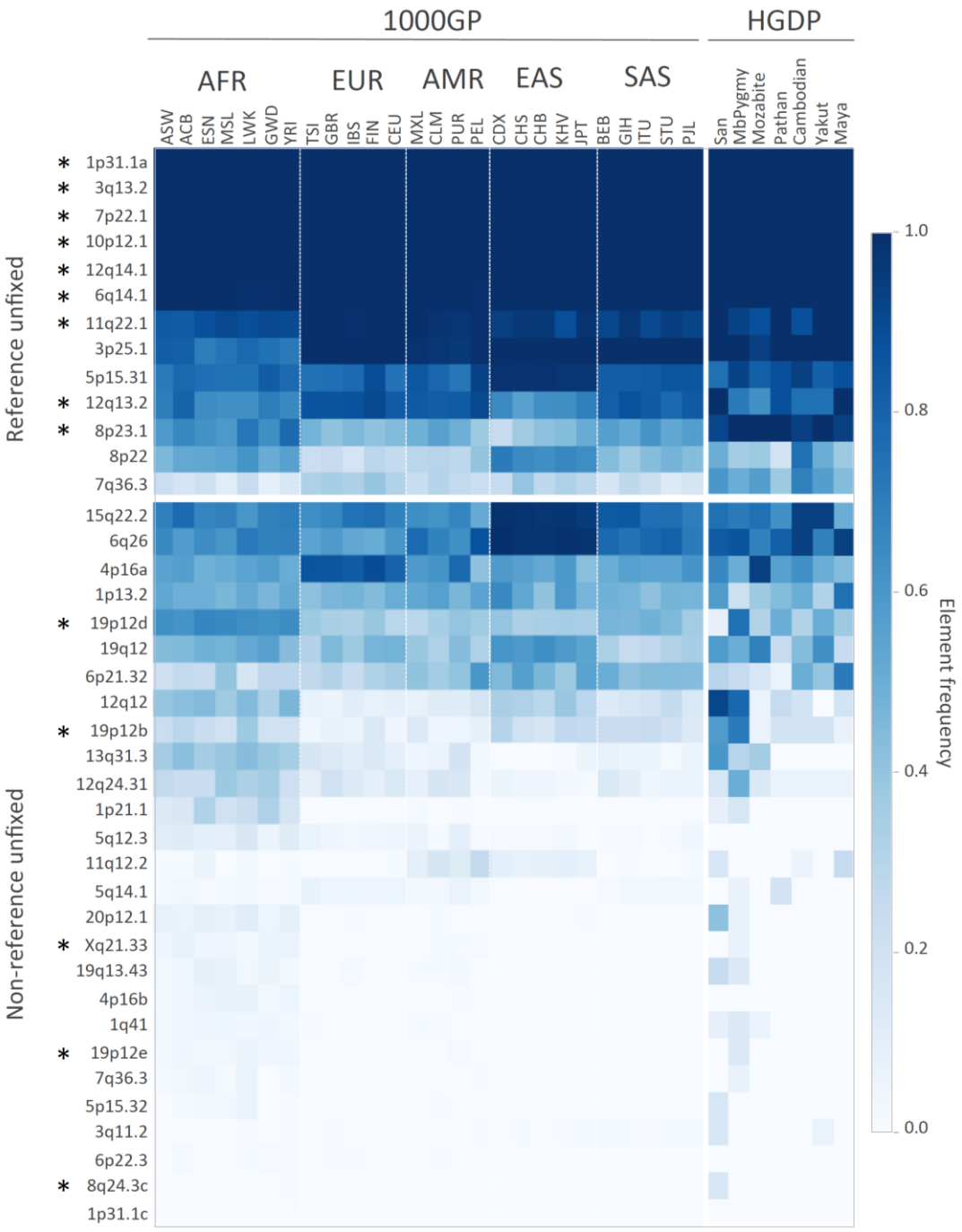


Figure 3-4. Estimated insertion allele frequencies of unfixed HML-2 insertions in humans.

A total of 40 HML-2 loci were subjected to in silico genotyping: 13 sites represented the unfixed HML-2 loci from the hg19 reference, and 27 sites corresponded to nonreference polymorphic HML-2 reported here. Genotypes were inferred for each unfixed HML-2 locus across samples based on remapping of Illumina reads to reconstructed insertion or empty alleles corresponding to each site. Samples lacking remapped reads at a particular site were excluded from genotyping at that site. *(continued on next page)*

Figure 3-4, continued. Estimated insertion allele frequencies of unfixed HML-2 insertions in humans.
Allele frequencies were then calculated for each population as the total number of insertion alleles divided by total alleles. Allele frequencies are depicted as a heat map according to the color legend to the right. The 1KGP (1000GP) and HGDP populations are labeled above (also refer to Dataset S1 for population descriptors and other information). The locus of each of the unfixed HML-2 proviruses is labeled to the left according to its cytoband position. An asterisk is used to indicate insertions that have confirmed full length copies. (Upper) Estimated distribution of reference unfixed HML-2 [from loci reported by Subramanian et al.(46) and Belshaw et al.(127)]. (Lower) Estimated distribution of nonreference HML-2 insertions. AFR, African; AMR, Admixed American; EAS, East Asian; EUR, European; SAS, South Asian. Reprinted with permission from PNAS (153).

Frequencies of the non-reference HML-2 insertions ranged from <0.0005 (insertion in 1 or few individuals) to >0.75 of genotyped samples (Fig. 3-4, *lower*). Nine of the 10 common insertions (detected in >10% of all samples), including the K113 provirus, have been previously reported in searches of WGS data (128,187). The inferred frequencies for the majority of these loci were generally in agreement with those reports (Fig. 3-5). We observed higher levels of K113 in our data than in previous reports, in ~18% of all samples and as high as ~38% in African populations, consistent with the prevalence of this insertion varying with ancestry (192). More than half of the non-reference HML-2 insertions were rare, with 15 insertions in <5%, and 6 insertions in <1% of all samples; just 4 of these loci have been previously reported (128). The lowest frequency sites were predominantly in individuals of African ancestry, with 9 of 13 loci detected in <5% of all samples but mostly limited to African populations, although insertions were also detected in non-African samples at ~0.005 to ~0.016 in those populations (for example, at 5q14.1 and Xq21.33 in Fig. 3-4). The solo LTR insertion at 1p31.1c was only identified in a single sample and was not detected in any other sample by genotyping, however this observation does not exclude the possibility of its presence

in some individuals, given the variability in read coverage between samples (also see Discussion).

Eight of our validated loci were recently reported in the genomes of archaic humans, both Neanderthals and Denisovans (142,189), in addition to modern humans (128). We confirmed an additional 3 reported 'archaic' sites in our data (19p12e and 10q24.2b, aka 'De11' and 'De12' from Agoni *et al.,* and 19q13.43, 'Ne5' from Lee *et al.*), but found no evidence of the remaining 8 reported archaic events (Table 3-1). Properties of these 11 HML-2 loci are more consistent with insertion in a common ancestor of archaic and modern humans prior to their divergence ~0.6 to 0.8 MYA (193), than with introgression of DNA from archaic species into modern humans, as has been reported to have occurred ~100,000-50,000 years ago (194–196). For example, the 2-LTR insertions at 19p12e and Xq21.33 are most prevalent in samples of African ancestry, and LTR divergences indicate their respective insertions to have been ~1.8–3.3 MYA, and ~0.67–1.3 MYA, consistent with this timeframe. In contrast, introgressed sequences from Neanderthals and Denisovans primarily appear in non-African populations, and introgression is thought to have occurred within the last 100,000 years (194–197). Both sites are rare, with sample-wide frequencies estimated at 0.0103 and 0.0157 (~0.026 to 0.069 in African sample) in our data. Of the remaining genotyped loci also in archaic genomes, each was also most represented in African ancestry, with exception of the insertions at 11q12.2 and 5q14.1 (sample-wide frequencies of ~0.046 and 0.026) that appeared most frequently in populations from the Americas or of East Asian ancestries but are also present in African populations, again implying ancient events (198). Given their overall distribution it is

likely these insertions are also older, though our ability to estimate insertion times is limited given their presence as solo LTRs.

3.4 Phylogenetic analysis of unfixed HML-2 proviruses

Utilizing sequence information obtained for each insertion, we performed an LTR-based phylogenetic analysis (Fig. 3-5). Because proviral LTRs are identical at integration, the two LTRs on the same provirus will always pair in a phylogenetic tree, barring recombination between different proviruses (48). Recombination between proviruses can also sometimes be detected by loss of the target site duplications generated during the integration process, which are preserved during solo LTR formation but will be lost during inter-element recombination (47,48,51). To create the most informative tree, we added the LTRs from 21 human-specific proviruses, including 11 polymorphic 2-LTR insertions as reported in Subramanian *et al.*, and four solo LTRs reported in Belshaw *et al*.(46,127), to our validated set of LTRs from three 2-LTR proviruses (insertions at 8q24.3c, 19p12e, Xq21.33), the 3' LTR of a truncated provirus (19p12d), and 30 solo LTRs.

The analysis revealed a major lineage leading to a well-supported clade that contained all human-specific and polymorphic HML-2 sequences (Fig. 3-5A, *boxed*); a minor lineage included HML-2 elements that are fixed in humans (46) and did not contain any newly identified insertions. This phylogeny is consistent with previous analyses (46,48,126); however the addition of our 34 unfixed loci permitted a more detailed examination of variable insertions (Fig. 3-5B).

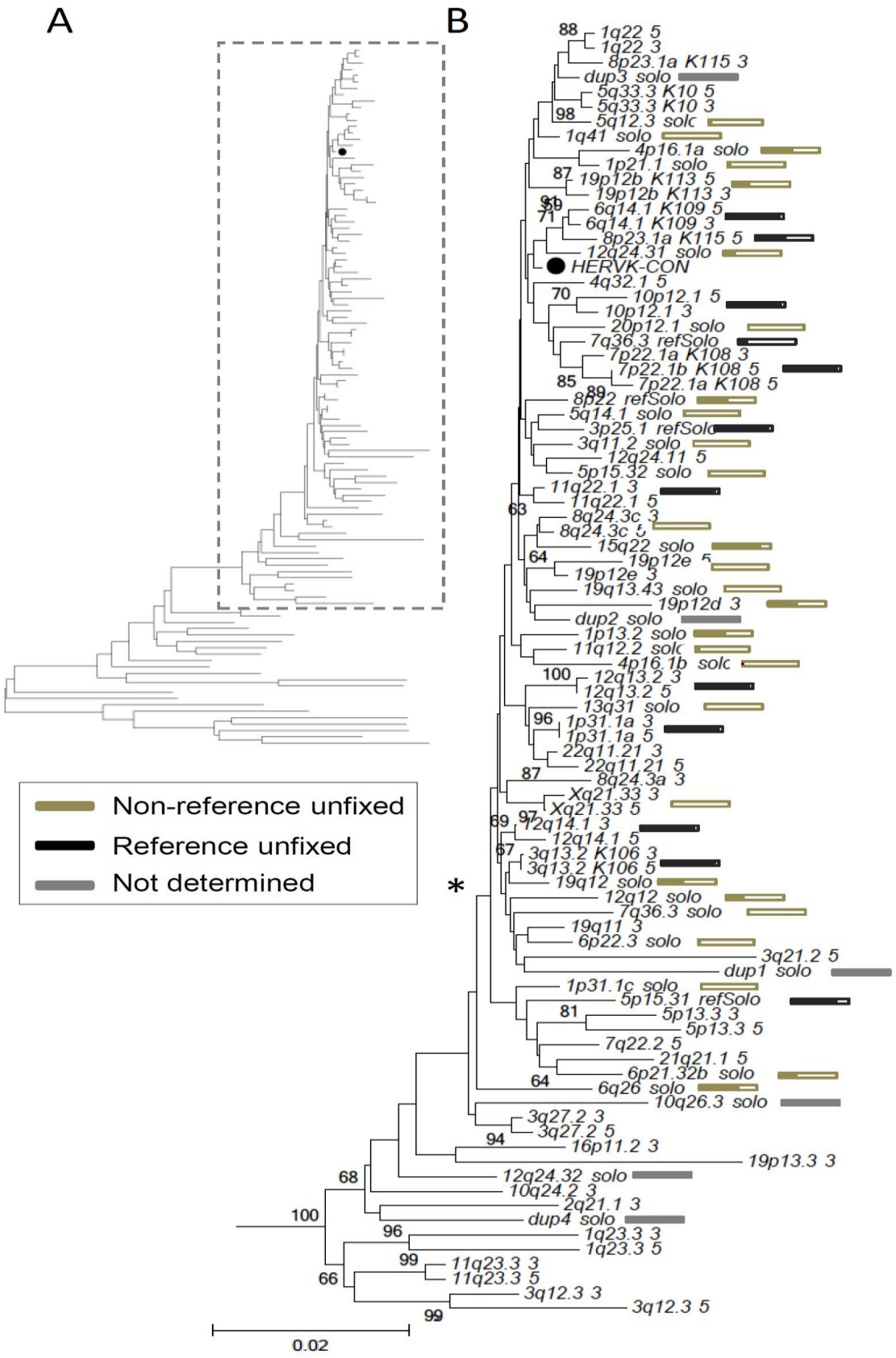

A
B
Non-reference unfixed
Reference unfixed
Not determined
*
1q22 5
1q22 3
8p23.1a K115 3
dup3 solo
5q33.3 K10 5
5q33.3 K10 3
5q12.3 solo
1q41 solo
4p16.1a solo
1p21.1 solo
19p12b K113 5
19p12b K113 3
6q14.1 K109 5
6q14.1 K109 3
8p23.1a K115 5
12q24.31 solo
HERVK-CON
4q32.1 5
10p12.1 5
10p12.1 3
20p12.1 solo
7q36.3 refSolo
7p22.1a K108 3
7p22.1b K108 5
7p22.1a K108 5
8p22 refSolo
5q14.1 solo
3p25.1 refSolo
3q11.2 solo
12q24.11 5
5p15.32 solo
11q22.1 3
11q22.1 5
8q24.3c 3
8q24.3c 5
15q22 solo
19p12e 5
19p12e 3
19q13.43 solo
19p12d 3
dup2 solo
1p13.2 solo
11q12.2 solo
4p16.1b solo
12q13.2 3
12q13.2 5
13q31 solo
1p31.1a 3
1p31.1a 5
22q11.21 3
22q11.21 5
8q24.3a 3
Xq21.33 3
Xq21.33 5
12q14.1 3
12q14.1 5
3q13.2 K106 3
3q13.2 K106 5
19q12 solo
12q12 solo
7q36.3 solo
19q11 3
6p22.3 solo
3q21.2 5
dup1 solo
1p31.1c solo
5p15.31 refSolo
5p13.3 3
5p13.3 5
7q22.2 5
21q21.1 5
6p21.32b solo
6q26 solo
10q26.3 solo
3q27.2 3
3q27.2 5
16p11.2 3
19p13.3 3
12q24.32 solo
10q24.2 3
2q21.1 3
dup4 solo
1q23.3 3
1q23.3 5
11q23.3 3
11q23.3 5
3q12.3 3
3q12.3 5
0.02

Figure 3-5. Phylogeny of HML-2 LTRs in humans.
(A) A neighbor-joining tree was constructed using HML-2 LTRs from the LTR5Hs sub-group, including all known human-specific and polymorphic loci. The LTRs were extracted from (i) all reference LTR5Hs HML-2 proviruses [as reported by Subramanian *et al.*(46)], (ii) unfixed reference solo-LTRs [as reported by Belshaw *et al.*(127)], and (iii) unfixed non-reference insertions as reported here. The closed circle (●) indicates the HERV-K_{CON} consensus LTR. Classic nomenclature has been included in taxon names for some insertions (see table 1-1 for provirus synonyms).
(B) Detailed view of clade containing unfixed elements. Individual HML-2 loci are indicated as follows: the cytoband followed by a 5′ or 3′ for the 5′ or 3′ LTRs from full-length insertions, solo for non-reference unfixed solo LTR insertions, or refSolo for reference unfixed loci. The asterisk indicates the clade containing most unfixed insertions. Boxes indicate estimated allele frequencies for each unfixed insertion at the end of each respective branch. The filled area is proportional to the estimated frequency of the insertion in all samples; gold and black coloring indicates non-reference and reference unfixed insertions, respectively, and gray bars indicate the elements for which the frequency could not be determined. 1000 bootstrap replicates were performed; nodes with greater than 60% support are labeled with their bootstrap value. Reprinted with permission from PNAS (153).

The majority of unfixed insertions clustered with the HML-2 consensus LTR (closed circle in Fig. 3-5*B*) in a clade ('*') whose members tended to have the shortest branches, consistent with their relatively recent integration and insertionally variable presence within humans (also refer to filled boxes in Fig. 3-5B). The human-specific reference elements were also distributed within this clade, consistent with previous observations (46). The remaining insertions were on branches with longer lengths, reflecting changes that accrued prior to insertion as well as during their longer existence as endogenous elements. We searched for, but did not observe, mis-paired LTRs from the 2-LTR proviruses reported here. Subsequent examination of the TSDs from these proviruses confirmed all were intact, indicating that these elements have not seeded past rearrangements (48). Extending this comparison to the TSDs of the identified solo LTRs also verified their intact state.

3.5 Properties of non-reference 2-LTR HML-2 integrations

As explained in section 1.4, the nucleotide divergence between cognate 5'-3' LTR pairs may be used to estimate the time since integration (51). Utilizing this method we previously estimated the average age of the human-specific 2-LTR insertions to within ~2.7 (±1.1) MYA (46). Applying this method to the 2-LTR proviruses identified here suggests these insertions were formed within ~0.67 to 1.8 MYA (Fig. 3-5). More precision for the youngest elements is limited as the variance for age estimates increases significantly for insertions with very little LTR divergence, and elements with identical LTRs can only be given a maximum age. The 19p12d provirus was excluded from this analysis due to deletion of the majority of its 5' LTR (Fig. 3-5). This truncation has also been observed in a few reference LTR5Hs 2-LTR insertions (8q24.3a, 10q24.2a and 19q11)(46), which all possess unique, intact TSDs and do not share any flanking sequence, supporting their classification as independent integrations. It is likely that this common rearrangement occurred as a result of aberrant strand transfer during reverse transcription, as has been discussed (199).

HML-2 proviruses are separated into two groups based on the presence or absence of a 292 bp deletion at the *pol-env* boundary, designated 'type 1' or 'type 2' proviruses, respectively. Deletion of a splice site in type 1 elements obliterates *env* and *rec* expression and results in mRNA encoding ~9kD protein, Np9, of uncertain function (38). Sequence comparison of the 2-LTR proviruses identified here classifies the 19p12d and 19p12e proviruses as type 1, and 8q24.3c and Xq21.33 as type 2.

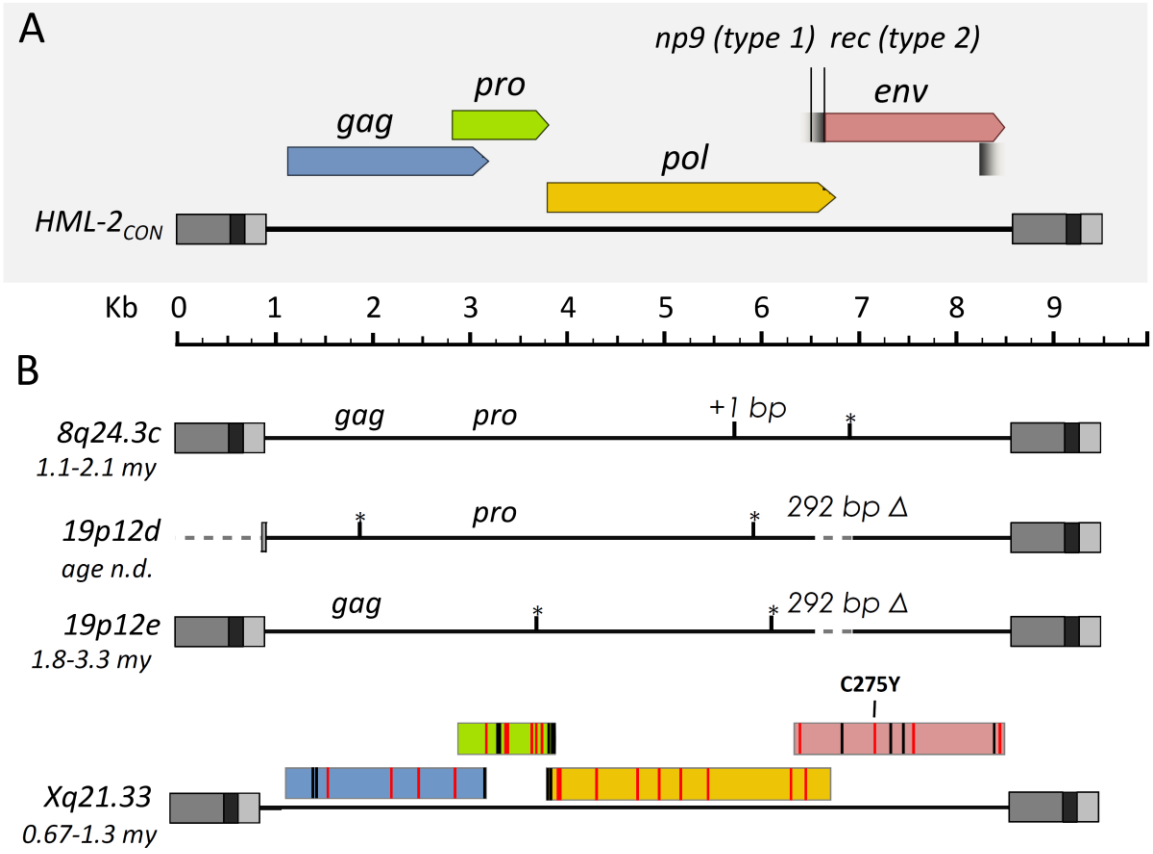


Figure 3-6. Features of newly identified HML-2 proviruses in humans.
(A) Schematic representation of the consensus HML-2 provirus, including the viral gene positions and frames to scale. Splice sites for *np9* (type 1 provirus, 292 bp Δ) and *rec* (type 2) are indicated. Regions within the LTRs are colored in gray: U3, medium; R, dark; U5, light.
(B) Features of non-reference identified proviruses are shown roughly to scale. The region of 292 bp is labeled for type 1 proviruses. Age estimations are shown for each site. n.d., not determined. The black vertical line indicates a frameshift mutation (as indicated "+1 bp"); black lines with asterisks are used to indicate positions of stop codons where present. Reading frames are shown for the Xq21.33 2-LTR provirus colored as in A. Black vertical lines within the frames indicate the positions of amino acid changes that are observed in other full-length HML-2 proviruses. Red vertical lines are used to indicate base changes that are unique to the sequenced Xq21.33 provirus, with the cysteine to tyrosine mutation that renders Xq21.33 Env non-functional pointed out. Adapted with permission from PNAS (153). Annotation of deleterious C275Y mutation added.

The 19p12d and 19p12e insertions were found to have intact *pro* and *gag* ORFs respectively, and the 8q24.3c insertion had both *gag* and *pro* ORFs. The Xq21.33 2-LTR element was found to be intact with ORFs for all HML-2 encoded genes (*i.e.*, *gag*, *pro*, *pol*, *env*, and *rec*) (Fig. 4). Indeed, it differs by only 39 out of a total of 2820 amino acids (98.6% amino acid identity) in all genes from the infectious consensus provirus HERV-K_{CON} (91). We searched for, but did not observe, any substitutions that would alter conserved sequence motifs (including the YIDD motif in reverse transcriptase), which suggested to us that it could potentially be replication competent. This provirus is only the second naturally occurring intact HERV to be described, the other being the noninfectious K113 (19p12b), which shares 98.9% amino acid identity to HERV-K_{CON} (87). Though no mutations in previously known essential sequence motifs were detected, we did note the presence of a cysteine to tyrosine mutation in the SU region of env; although this amino acid is not known to be essential, most envelope proteins have a number of cysteines in their SU regions that form disulfide bridges important for protein stability and participate in rearrangement during the fusion and entry process. Another graduate student in our lab, Michael Freeman, has shown that Xq21.33 *env* can only produce protein if this cysteine is restored; thus, Xq21.33 is almost certainly not fully replication competent. It is as yet undetermined whether reversion of this mutation is sufficient to restore infectivity, but preliminary data suggests the presence of at least one other inactivating mutation in the provirus.

Chapter 4: HML-2 activity in gorillas and other non-human primates

4.1 Proviruses identified in gorillas from the Great Ape Genome Project

Though we were able to identify rare, low-frequency proviruses in humans, including one with intact ORFs, none of the proviruses identified appear to be infectious. However, HML-2s have been found across the Old World monkey family, including chimpanzees and gorillas, and it seemed worthwhile to investigate the possibility that HML-2 has been recently active in one or more of these primates. Though some research has been published on HML-2s in non-human primates, they remained relatively unstudied. Age estimates of HML-2s from humans suggest that they were active during the divergence of gorillas, chimpanzees, and humans. Some chimpanzee specific proviruses have been reported (156–158), but our initial investigations of the chimpanzee reference genome indicated a relative paucity of chimpanzee-specific insertions compared to humans. A previous post-doc in our lab, Ravi Subramanian, began searching the gorilla reference gorGor3 genome for species-specific insertions, however, instead of gorilla-specific insertions, he found that the assembly had mis-placed HML-2 sequences at the loci of known human-specific insertions. These appeared to be orthologous insertions; however, PCR genotyping of the putative insertion sites indicated that no provirus was actually present at these loci, and thus he termed them pseudo-orthologues or 'pseudologues.' However, these sequences did appear to be derived from genuinely novel HML-2 proviruses, as their sequences were divergent from known human HML-2s, and we were able to demonstrate that one of the pseudologues included sequence from a genuine gorilla-specific solo LTR. Although a newer gorilla assembly (gorGor4) had been

released by the time we began this study, it retained the same assembly errors found in gorGor3. While these sequences appeared genuine, they were quite fragmented, and we could not be sure whether they contained sequence from a single provirus or were assembled from multiple different proviruses; thus we decided that the only way forward was through mining raw sequence reads.

Based on these data, it seemed likely that undescribed HML-2 proviruses could be found in gorillas using the same basic sequence mining approach we used with humans. Though gorilla genomes have not been sequenced at anywhere near the scale of the 1000 genomes project, a smaller scale study, the Great Ape Genome Project (GAGP) was finished in 2013, which included high coverage Illumina sequencing of 79 great apes from all 6 great ape species. The study included data from 31 gorillas, primarily Western lowland gorillas, but also included 3 samples from the Eastern lowland gorilla species. We analyzed sequence data from 21 of these gorillas (18 Western and 3 Eastern). Unlike the 1KGP data, these sequences were only available in raw, unaligned FASTQ format. Thus, we added an alignment step. Initially, we used the gorGor4 assembly as a reference genome for our alignments, however, when the gorGor5 long read assembly became available we began using it as a reference instead.

For this study, we only used the LTR-junction discovery method to identify insertions; our results from the human study indicated that this method is as or more sensitive than RetroSeq at identifying non-reference insertions, and the main advantage of RetroSeq is its high throughput, which was less of a concern here, since we were screening 100-fold fewer genomes. We were also concerned that the lower quality of the

gorilla assemblies might cause some problems for RetroSeq, which depends on proper mapping of reads to the reference for validating putative insertions.

As before, we retrieved unmapped reads from each aligned gorilla genome and searched reads with sequence matching the HML-2 5' or 3' LTR edges, filtering out any reads with junctions that matched known HML-2s in humans or gorillas, and any reads with <10 bp of sequence flanking the insertion. From the 21 gorillas screened, we identified 675 putative insertion sites. Strikingly, this is almost 20 times the number of insertions we found in our human study, despite screening over 100 times as many individuals in that study. Far more human-specific insertions were already known at the time of our study, which could partially explain this discrepancy, though there are only ~150 known human-specific insertions, still far less than we found in gorillas. It is possible that some of the gorilla hits are artifacts; however, when we filtered for hits that had reads corresponding to both the 5' and 3' junctions, 126 high confidence hits remained. Though we primarily focused on validating insertions with hits to both junctions, we did validate 6 one-sided hits via PCR and sequencing, so it is likely that many if not most of the 675 total hits are genuine insertions.

Interestingly, many of the insertions mapped to regions with gaps in the gorGor4 assembly, suggesting that, in addition the pseudologous proviruses mentioned above, non-reference insertion elements also interfere with the assembly process at their integration site, leading to gap artifacts in the genome assembly. Though most of these gaps were not present in the gorGor5 assembly, we have found some hits at sites with partial LTRs present. We designed primers to several of these sites, and were able to

amplify and sequence complete solo LTRs at two sites, suggesting that even with long read sequencing some insertion related assembly errors remain.

As with the human data, we validated hits by allele-specific PCR and sequencing. As we were unable to obtain DNA from the gorillas sequenced in the GAGP, we used genomic DNA from three Western lowland gorillas provided to us by Michael Jensen-Seaman of Duquesne University for PCR validation. We designed primers to 66 of our hits, and were able to confirm 27 of these loci by PCR and Sanger sequencing (Table 2-2). Of these, 21 were solo LTRs and 6 were full length or partial 2-LTR proviruses (Table 4-1). Most of the insertions appeared to be insertionally polymorphic.

4.2 Proviruses identified from the gorGor5 long read genome assembly

As mentioned above, the gorilla genome assemblies available when this project began had very unreliable assembly of HML-2 elements, with artifactual 'pseudologue' proviruses placed at the sites of known human specific HML-2 insertions, as well as gaps in the assembly at sites of genuine gorilla specific insertions. However, partway through the project, a new gorilla reference (gorGor5) was published that had been generated using PacBio long read technology (200). We were hopeful that the long reads would improve the assembly of large insertions such as ERVs, perhaps sufficiently to identify more 2-LTR proviruses.

Table 4-1. Gorilla-specific reference and non-reference proviruses, and non-reference solo LTRs

Locus	Coordinates (gorGor5)	Alleles	ORFs	Data source
1p34.2	CYUI01015058v1:8698164	solo, pre		GAGP
1p36.13	segmental duplication	prov	*gag*	GAGP
1q43	CYUI01014905v1:7,379,133	solo,pre		GAGP
2p13.3	CYUI01015158v1:5,977,400	solo,pre		GAGP
3p12.2	CYUI01015020v1:6,155,112	solo,pre		GAGP
3p22.3	CYUI01014916v1:23806527	solo,pre		GAGP
3q13.13	CYUI01014906v1:15912756	prov		gorGor5
3q23	CYUI01014939v1:11491901	prov,pre	*gag, pol, env*	GAGP
3q27.2	CYUI01015043v1:1542506	prov	*gag*	gorGor5
4q28.1	CYUI01015140v1:2,825,207	prov,pre	*gag, pro*	GAGP
5p14.3	CYUI01015420v1:1207415	prov	*gag*	gorGor5
5q23.2	chr5:110,516,789 (gorGor4)*	solo,pre		GAGP
6p11.2	CYUI01015849v1:88674	prov		gorGor5
6p12.3	CYUI01015013v1:7157648	solo		GAGP
6p22.1	CYUI01015250v1:4,896,903	prov		GAGP, gorGor5
6p25.2	CYUI01015393v1:3795101	prov		gorGor5
7p13	CYUI01015540v1:1,190,051	solo,pre		GAGP
7p21.2	CYUI01014991v1:10257639	solo,pre		GAGP
8p22	CYUI01015432v1:2,640,415	solo,pre		GAGP
8q22.2	CYUI01014940v1:14,904,040	solo		GAGP gorGor5
9p13.3	CYUI01015156v1:4,751,052	prov,pre	*gag,pro,pol,env*	GAGP
10p11.23	CYUI01015287v1:1,891,774	prov,pre	*gag, env*	GAGP
10q11.21	CYUI01015448v1:2,752,878	solo		GAGP
11q21	CYUI01015079v1:4,818,452	solo,pre		GAGP
11q22.1	CYUI01015153v1:726965	solo,pre		GAGP
12p13.32	CYUI01015151v1:4533155	prov		gorGor5
12q12	CYUI01015106v1:835,671	solo,pre		GAGP
14q23.1	CYUI01014968v1:9700255	solo,pre		GAGP
18q22.1	CYUI01014964v1:2,006,472	solo,pre		GAGP
19p12	CYUI01015110v1:8706168	prov		gorGor5
19p13.2	CYUI01015260v1:5498638	prov		gorGor5
19q13.2	CYUI01015741v1 :71604	solo		GAGP
20p13	CYUI01015378v1:1074507	solo,pre		GAGP
20q13.33	CYUI01015942v1:80936	solo,pre		GAGP
21q22.3	CYUI01015595v1:902,995	solo,pre		GAGP
22q11.23	CYUI01015765v1:799709	prov	*env*	gorGor5
Xq21.1	CYUI01014915v1:18600100	prov		gorGor5

*gorGor4 coords given for 5q23.2; no orthologous position present in gorGor5

We first checked to see whether the assembly still contained the pseudologous insertions we had found in previous versions of the gorilla genome. We searched the gorGor5 assembly using the flanking sequences from each pseudologue as queries in BLAT (172). Encouragingly, we were able to identify intact preintegration sites for all but one of the pseudologues, with no gaps or HML-2 sequences inserted. The remaining site did not appear to be present in the new assembly at all, either as a pre-integration site or a provirus.

We downloaded all LTR5Hs and LTR5 coordinates annotated by RepeatMasker (174) in the gorGor5 assembly available on the UCSC Genome Browser (169). To identify novel-gorilla specific HML-2s, we cross-referenced these sites against the same data from the human, chimpanzee, and orangutan genomes, filtering out any sites where at least one of those three genomes had an annotated HML-2 insertion within 1 kb. Full length proviruses within this dataset were identified by filtering for sites with HML-2 internal genic regions, classified as HERVK-int by RepeatMasker. We identified 11 gorilla-specific 2-LTR full length or nearly full length proviruses present in the gorGor5 reference (Table 4-1), including one of the sites previously identified in the GAGP data, and 93 gorilla-specific solo LTRs, again including one solo LTR from the GAGP data. As with humans, a roughly 10:1 ratio between solo LTRs and proviruses is seen.

4.3 Evolutionary dynamics of HML-2s in gorillas and other primates

We generated a neighbor-joining phylogeny of gorilla-specific insertions, using the LTR sequences of the non-reference 2-LTR proviruses and solo LTRs we identified from the GAGP, and all gorilla-specific 2-LTR proviruses from the gorGor5 reference genome. For comparison, we also included the sequences of all the 2-LTR proviruses known from humans, the human-specific solo LTRs we identified in our first study, and a small selection of chimpanzee specific LTRs from the chimpanzee reference genome (panTro3).

As expected, most of the species specific LTRs segregated in monophyletic clades, with the human and chimp insertions clustering together and the gorilla specific insertions forming an outgroup to chimpanzees and humans (Figure 4-1A). In addition to the main gorilla-specific clade, a small number of sequences clustered in a separate clade, suggesting the possibility that two separate HML-2 strains infected the gorilla lineage after they split from the human-chimp lineage.

Strikingly, the gorilla-specific clade had noticeably shorter branch lengths than the human-specific or chimp-specific clades, indicating much less sequence divergence in the gorilla-specific viruses; indeed, two groups of insertions have not diverged at all and thus have identical LTRs (Fig. 4-1B). This low level of divergence suggested to us that many of these insertions were very young, and that the HML-2s may have been active in gorillas more recently than in humans or chimpanzees.

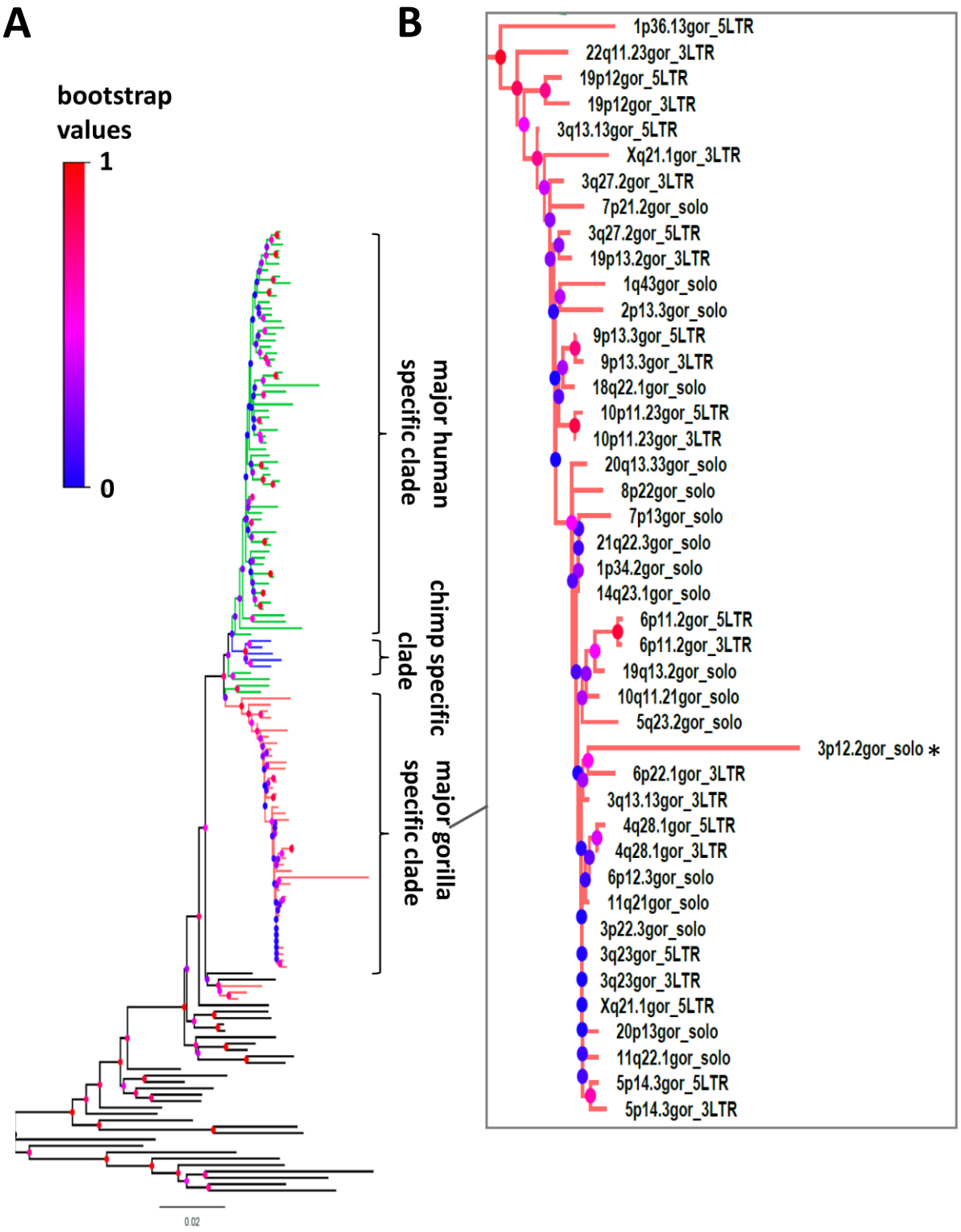

A
bootstrap values
1
0
major human specific clade
chimp specific clade
major gorilla specific clade
0.02
B
1p36.13gor_5LTR
22q11.23gor_3LTR
19p12gor_5LTR
19p12gor_3LTR
3q13.13gor_5LTR
Xq21.1gor_3LTR
3q27.2gor_3LTR
7p21.2gor_solo
3q27.2gor_5LTR
19p13.2gor_3LTR
1q43gor_solo
2p13.3gor_solo
9p13.3gor_5LTR
9p13.3gor_3LTR
18q22.1gor_solo
10p11.23gor_5LTR
10p11.23gor_3LTR
20q13.33gor_solo
8p22gor_solo
7p13gor_solo
21q22.3gor_solo
1p34.2gor_solo
14q23.1gor_solo
6p11.2gor_5LTR
6p11.2gor_3LTR
19q13.2gor_solo
10q11.21gor_solo
5q23.2gor_solo
3p12.2gor_solo *
6p22.1gor_3LTR
3q13.13gor_3LTR
4q28.1gor_5LTR
4q28.1gor_3LTR
6p12.3gor_solo
11q21gor_solo
3p22.3gor_solo
3q23gor_5LTR
3q23gor_3LTR
Xq21.1gor_5LTR
20p13gor_solo
11q22.1gor_solo
5p14.3gor_5LTR
5p14.3gor_3LTR

Figure 4-1. LTR tree of gorilla-specific proviruses and non-reference solo LTRs.
(A) Relationship of gorilla-specific HML-2s to HML-2s in humans and chimps. LTRs of gorilla specific proviruses identified from the gorGor5 reference genome and provirus and solo LTRs identified from the GAGP sequence data were aligned using the MUSCLE alignment program to a set of human-specific and chimp specific LTR5Hs sequences and a set of LTR5Hs sequences with orthologues shared in all three species. This alignment was used to produce a neighbor-joining tree with 1000 bootstrap replicates. Nodes are colored according to their bootstrap value, with low confidence nodes colored blue and high confidence nodes colored red. Gorilla-specific branches are in red, human in green, chimp in blue, and shared in black. Branch length is proportional to sequence divergence. Three monophyletic clades representing the majority of specie-specific insertions are indicated.
(B) Expansion of major gorilla-specific clade to show greater detail. Branch and node coloring as before. Each sequence is labeled with its cytoband, 'gor' indicating gorilla-specific, and either 'solo' for solo LTRs or '5/3LTR' for the 5' and 3' LTRs of 2-LTR proviruses. The asterisk marks a divergent sequence that showed signs of APOBEC3G-mediated hypermutation.

To investigate this possibility more thoroughly we first used the 5'-3' LTR comparison method to date those proviruses with two LTRs. The median age of the insertions was ~312,000 years, plus or minus 100,000 years. Three proviruses had identical LTRs and thus we could only estimate that they integrated at most 300,000 years ago, and possibly much more recently. Thus the limits of this dating method prevent us from determining whether any of these proviruses are younger than 300,000 years old.

Though 5'-3' LTR divergence is the preferred method to date individual proviruses, the limited number of 2-LTR proviruses in the genome limits the usefulness of this method for estimating the actual level of HERV activity over time. There are10 times as many HML-2 solo LTRs as there are 2-LTR proviruses, however, and we wanted to use this abundance of sequence information to get a better sense of the levels of HML-2 activity over time in both gorillas and other great ape species. In order to estimate the insertion times of solo LTRs, we modified our 5'-3' LTR comparison method to use the

divergence between each LTR and its nearest neighbor on a neighbor-joining phylogeny. We previously had used the divergence from a consensus LTR sequence as a clock, however, this method assumes that all LTRs are identical in sequence at the time of integration, which leads to a bias towards older ages. Using the nearest neighbor instead of a consensus should significantly reduce this bias.

Using the same methods as we used to identify species specific insertions in the gorilla reference genome, we downloaded and determined the species specificity of HML-2 LTR5Hs solo LTRs in the latest human and chimpanzee reference genomes (hg38 and panTro3, respectively); to this we added the sequences of all known non-reference human-specific LTR5hs LTRs. For each species, we made a neighbor-joining tree of all the LTR sequences from that species and estimated the age of each LTR based on the branch distance to the nearest node on the tree (Fig. 4-2). As this distance is only half the divergence to the nearest neighboring LTR, we multiplied each distance by 2 before applying the same aging algorithm we use for 5'-3' LTR comparisons.

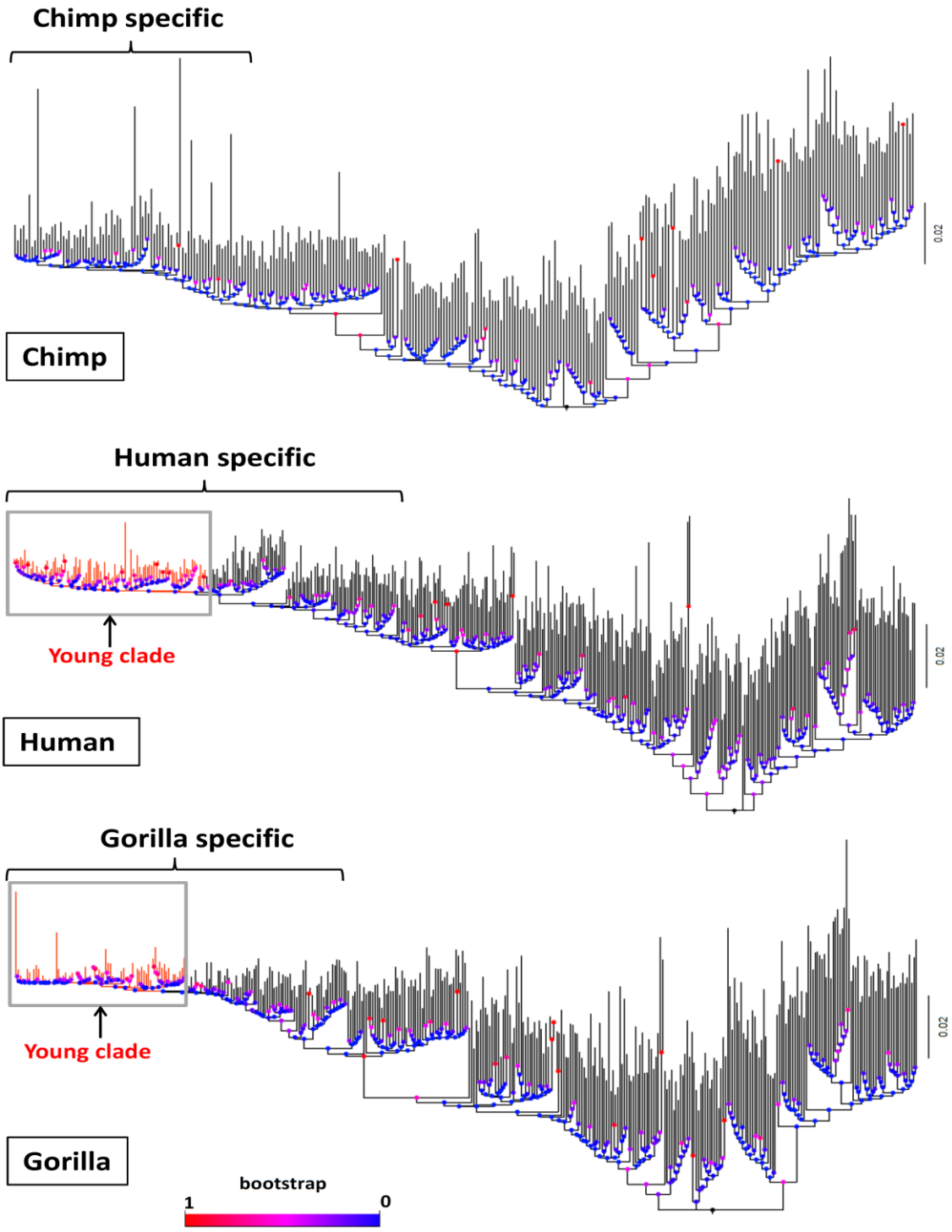
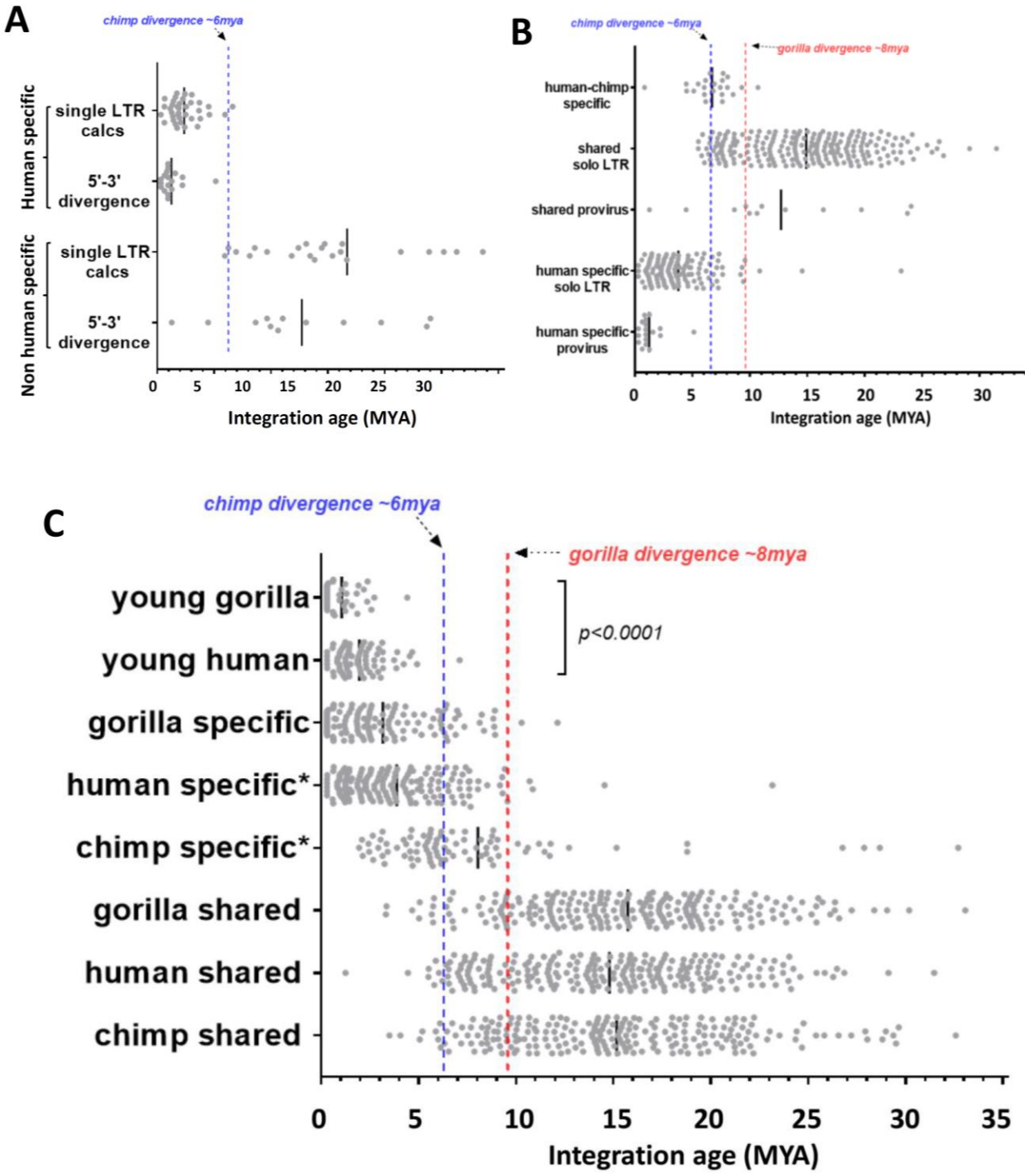


Figure 4-2. HML-2 phylogenies for age estimation.
Neighbor joining trees of LTR5Hs LTRs from gorillas, humans and chimps with 1000 bootstrap replicates. Nodes are colored according to their bootstrap confidence as in Fig.4-1. Branch lengths are proportional to sequence divergence. Species specific clades are marked, and low divergence subclades in gorillas and humans are in red.

In order to test the accuracy of this method as compared to 2-LTR estimates, we used the same approach with the 5' and 3' LTRs of 2-LTR proviruses, using the distance to the nearest LTR not belonging to the same provirus. The ages for proviruses calculated in this manner were indeed older than the ages for the same proviruses calculated by 2-LTR comparison, with a median age of 1.9 million years for human specific loci as compared to 0.9 million years for the standard method, and 14.6 vs 10.8 million years for the older, shared proviruses (Figure 4-3B). The difference was not statistically significant for shared proviruses, but was for human-specific (Mann-Whitney, $p=0.0002$, Kolmogorov-Smirnov, $p<0.0001$). However, a number of the shared proviruses have anomalously young ages when calculated by the standard method, likely due to inter-LTR gene conversion; we did not control for this in our calculations, so the 2-LTR estimation may in fact be biased younger, which may account for some of the difference in addition to the expected bias towards older ages when using the solo-LTR method. The ages calculated in this manner do correspond fairly well with the expected ages based on their species specificity, e.g. all but two of the human specific LTRs end up with ages less than 6 million years, the date of the human-chimpanzee divergence, and all of the shared LTRs are more than 6 million years old (Fig.4-3B).



** includes sequences specific to chimp/human clade*

Figure 4-3. Ages of solo LTRs and 2-LTR proviruses in humans, chimpanzees and gorillas.
Neighbor joining trees of LTR5Hs LTRs from gorillas, humans and chimpanzees were generated (see fig 4-2), and ages were calculated using the distance to the nearest neighboring LTR. Each dot represents the age of one insertion.
(A) Ages of human proviruses as calculated by 5'-3' LTR comparison compared to the same loci ages calculated using the solo LTR method.
(B) Comparison of ages of solo LTRs and proviruses in humans. 'Human-specific' loci are found in only humans, 'human-chimp' specific are found only in humans and chimps, 'shared' are also found in gorillas. The dotted red line and blue lines respectively mark the estimated times of divergence of chimpanzees and gorillas from humans.
(C) Ages of solo LTRs and proviruses from gorillas, chimps, and humans. Human specific and chimp specific groups include orthologues found in both chimps and humans, but not gorillas.'young gorilla' and 'young human' are monophyletic clades containing the majority of the most recently integrated gorilla specific and human specific insertions, respectively, as labeled in Fig. 4-2A/B. P-value calculated with both Mann-Whitney and Kolmogorov-Smirnov tests.

A similar pattern is seen for the human solo LTR ages calculated with the new method, with greater ages than proviruses with the same species distribution, but still within the expected age range based on their species distribution (Fig. 4-3A). We identified a small number of solo LTRs found in chimps and humans but not gorillas, which were presumably integrated in the period between the gorilla-human divergence and the chimpanzee-human divergence, ~6-8 million years ago. The ages of these loci cluster fairly closely to this time period, though slightly younger than expected (Fig. 4-3A). The solo LTR ages thus appear to still have some bias towards older integration times, but seem to be a useful measure of relative integration activity over time.

We therefore applied this dating method to our full dataset of human, gorilla, and chimpanzee-specific solo LTRs, as well as loci shared between these species (Fig. 4-3C). The chimpanzee specific insertions were significantly older than the human specific and gorilla specific loci, with no sites younger than 2 million years, and most significantly older (Mann-Whitney, $p<0.0001$, Kolmogorov-Smirnov, $p<0.0001$). As expected, the

gorilla-specific loci are younger than the human specific loci; when we looked only at the ages of LTRs belonging to the clade containing most of the youngest insertions, they had a median age of ~0.6 million years, approximately 1 million years younger than clade containing the youngest human specific insertions (Mann-Whitney, $p<0.0001$, Kolmogorov-Smirnov, $p=0.0002$).

4.4 Structure and coding capacity of gorilla-specific proviruses

As with the human proviruses in chapter 2, we used the NCBI ORFfinder tool to investigate the coding capacity of the gorilla-specific proviruses we discovered. Seven of the sixteen 2-LTR proviruses discovered had intact ORFs for at least one viral gene, with 6 *gag*, 2 *pro*, 2 *pol*, and 4 *env* ORFs in total (Table 4-1, Fig. 4-4). Interestingly, one of the proviruses, 9p13.3, had intact ORFs for all 4 retroviral genes. A 2 bp frameshift towards the 3' end of 9p13.3 *env* completely changes the sequence of its cytoplasmic tail, in fact to the sequence of *np9*, however the rest of *env* including the entire transmembrane region is intact. At least 2 of the proviruses are type 1, containing the characteristic 292 bp deletion in *env*. Interestingly, 4 of the proviruses shared another deletion which removes a large fraction of *pro* and *pol*, presumably rendering them completely incapable of infection or intracellular retrotransposition, at least without the aid of a replication competent helper virus.

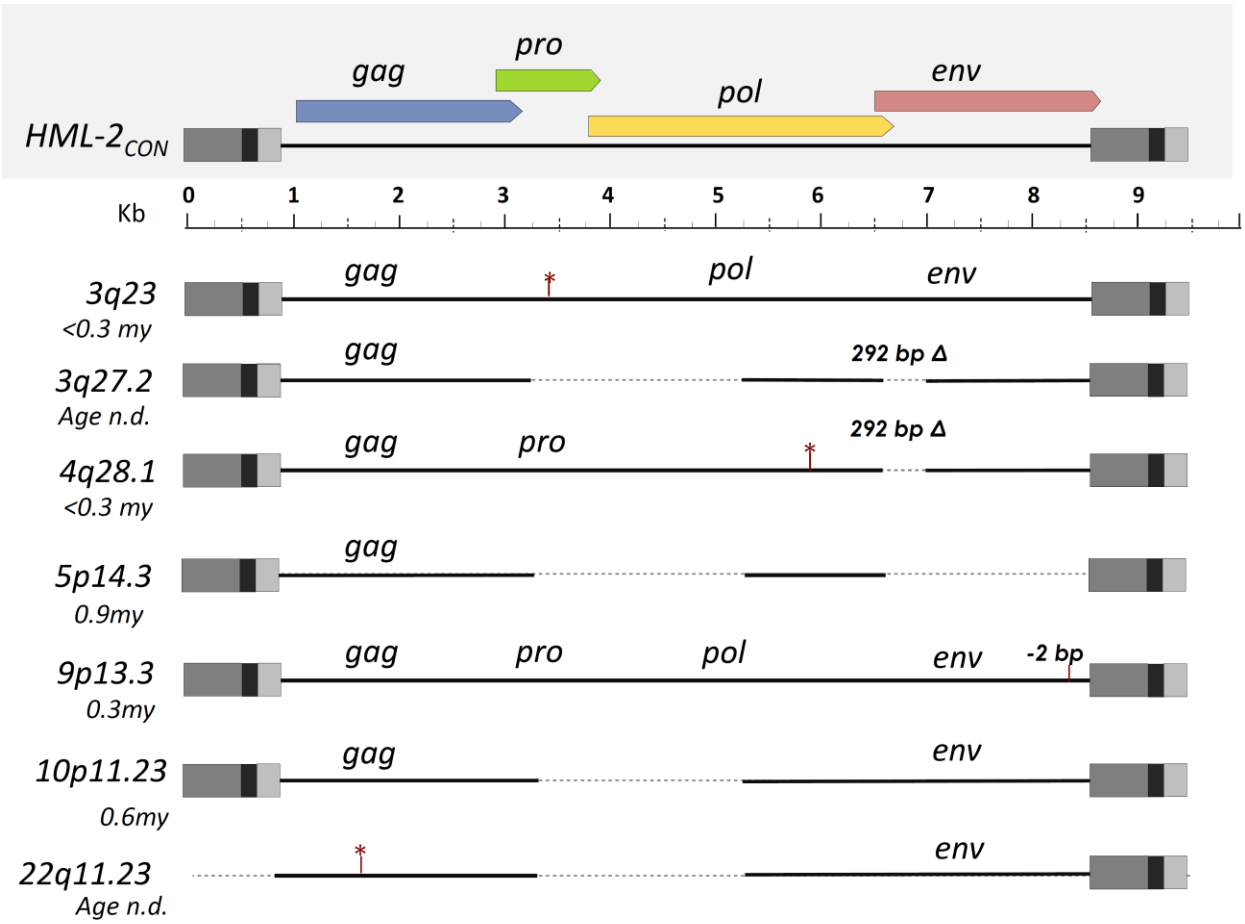


Figure 4-4. Structure and coding capacity of gorilla-specific proviruses.
Schematic of the 7 gorilla-specific proviruses with at least one intact ORF, as compared to a reference intact HML-2 provirus (HML-2_{CON}). Proviruses are named according to their cytoband locus. Ages as estimated by 5'-3' LTR divergence are given below the names of each provirus.

Open reading frames are indicated by gene name, with stop codons that disrupt a reading frame annotated with asterisks. A 2bp deletion in 9p13.3 that causes a frameshift near the 3' end of *env* is annotated as such. Deletions are marked with dotted lines.

Chapter 5: Conclusions and Discussion

We have developed methods to mine high throughput short read sequencing data to identify endogenous retrovirus insertions not present in the reference genomes used to align the data. We used these methods to identify 36 non-reference insertions, including 19 novel loci, in a genetically diverse set of >2500 individuals sequenced by the 1000 Genomes Project and the Human Genome Diversity Project, including 5 2-LTR proviruses, 29 solo LTRs, and 2 integrations with an unusual inverted repeat structure. Seventeen of the 36 sites were recently reported in humans (128,187), though with limited validation or element characterization. We took full advantage of the 1KGP and HGDP WGS read data to identify non-reference viral-genome junctions from assembled anchored read pairs and individual unmapped reads, and utilized these data to estimate the presence of each of these elements within our sampled populations. We validated the presence of 34 of the 36 loci, including five loci with 2-LTR proviruses (including K113) and 29 solo LTRs, and report the complete sequences for 30 of these insertions, including a 2-LTR provirus at Xq21.33 that has retained open reading frames for all viral genes.

We utilized the available reads from each sample for *in silico* genotyping of a subset of sites to infer the population-wide frequencies of unfixed HML-2 elements, impractical on this scale in standard PCR based genotyping screens. The inferred allele frequencies of the non-reference insertions ranged from 0.05% to >75% of genotyped samples and varied between populations, generally with the highest presence in African populations. With the exception of two previously sequenced sites in our set (dup1 and 12q24.32), all non-reference insertions were validated in samples of African ancestry, as has been observed for all HERV-K loci characterized to date, implying their insertion prior to the

human migration out of Africa approximately 45,000 to 60,000 years ago (198). Dup1 and 12q24.32 could not be confidently mapped to the hg19 reference and were therefore excluded from the *in silico* genotyping analysis.

Eleven of the non-reference insertions identified here have previously been reported from the genomes of Neanderthals and/or Denisovans, though the fragmentary nature of those genomes precluded complete sequencing or any detailed analysis (142,189). It is conceivable that some of these insertions could have formed from infections of these archaic human populations after they split from the modern human lineage, and afterwards entered the human genome through interbreeding. Such introgression of archaic human DNA into modern humans is believed to have occurred ~100,000-~50,000 years ago; however, these introgression events occurred only in non-African populations, and thus archaic DNA is not found in sub-Saharan African populations at a significant level. The insertions identified here are widely distributed in the African populations sequenced by the 1000 Genomes Project, suggesting that they in fact inserted in the common ancestor of modern humans, Neanderthals, and Denisovans. The estimated ages of 19p12e/De11 and Xq21.33/De9 are also consistent with insertion prior to this split.

All but one non-reference insertion was identified in more than one individual, the exception being the 1p31.1c solo LTR validated in NA18867. Genotyping of that site failed to reveal its presence in other individuals, but does not rulethat possibility out, given the low sequence coverage of the data analyzed. Analysis of the surrounding region revealed the presence of several SNPs that were unique to NA18867 within the 1KGP panel, suggesting that 1p31.1c may be associated with a very low frequency haplotype, rather than a de novo infection event, in the absence of comprehensive screening. These

observations support the utility of short read data for element discoveries and sequence-based analysis, but also underscore the necessity of additional experimental validation steps and characterization of candidate proviruses.

We confirmed the presence of full-length proviruses at four loci, including the Xq21.33 provirus, which appears to be intact and without obvious defects, which suggested the potential for replication competence. However, recent experiments by Michael Freeman in our lab have shown that a cysteine to tyrosine mutation in Xq21.33 *env* renders it non-functional. Other deleterious mutations have not yet been identified, though it appears likely that at least one more inactivating change is present.

Low levels of ongoing replication in humans have been suggested based on inferred frequencies of unfixed HML-2 loci (126,127). Coalescent analysis of globally sequenced LTRs from the K106 provirus has produced an estimated insertion time of 0.15 MYA based on sequence conservation of that site across sampled individuals (136). Other studies suggest replication until at least 0.25 MYA based on modeling estimations of an expected number of loci given the number and frequencies of observed unfixed sites (128). Our extensive mining of >2500 genomes revealed 36 non-reference insertions; this is well below the reported estimations of observed unfixed sites in the case of continued replication (128). The number of rare insertions in our data (15 insertions in <5%, and 6 in <1% of all samples), including the 1p31.1c LTR detected in a single individual and the 2-LTR provirus at 8q24.3c in just three samples (from the HGDP San and 1KGP Yoruba populations), suggests that additional remaining HML-2 loci are likely to be very rare, specific to groups not yet surveyed, or within low coverage regions of the genome. A model of ongoing low level replication would be expected to produce some level of rare,

newly endogenized proviruses over time; however, this model would also predict that those rare proviruses would cluster together phylogenetically, separately from proviruses that integrated at earlier times. Although we were able to identify very low-frequency proviruses in humans, including one with intact ORFs, none of the proviruses identified were infectious; the sequences we identified, though relatively young, are no younger on average than previously identified human-specific insertions, nor do they cluster together phylogenetically, but rather are scattered throughout a larger clade of human-specific insertions. This suggested to us that the rare proviruses we had found were not the footprint of current HML-2 infectious activity; rather than being very rare because they are very young, they are rare because of the stochastic nature of genetic drift.

Thus, we have found no evidence of ongoing HML-2 replication in humans, despite screening >2500 individuals for the presence of rare integrations. This does not rule out the possibility that infectious HML-2s still exist in humans, but it seems increasingly likely that HML-2 is effectively extinct in humans today. It is still possible that HML-2s play a role in disease, and it is even conceivable that an infectious mechanism could play a role. Though HML-2s and other HERVS are likely to be regulated by silencing and downstream host mechanisms, disease states causing prolonged reactivation of HERVs could drive expression of such proviruses or the generation of recombinant infectious chimeras, as has been shown to occur in antibody deficient mice (81) and as claimed for HML-2-derived transcripts in the blood of HIV-infected individuals (191). Indeed, a ‘recombinant’ HML-2 provirus engineered from just 3 defective reference loci is infectious (derived from portions of K109, K115, and K108 respectively at 6q14.1, 8p23.1a, and 7p22.1), as are the HML-2 consensus genomes (90,91). As endogenization

requires infection of germ cells, it is conceivable that such infectious chimeric viruses could be generated in diseased tissues and lead to pathogenic effects, without ever encountering a germ cell, or establishing any chain of transmission outside the affected host. It is important to distinguish which proviruses could plausibly lead to chimeras: at the very least, it would require expression of one provirus with intact ORFs for *gag*, *pro*, and *pol* (or at least the RT domain), as stop codons upstream of *pro* or *pol* would prevent their expression, even if their ORF remained intact. Only a few proviruses meet this criterion, and most of them are quite rare (46). In addition such an event would require the expression of a provirus that makes functional Env protein, also fairly rare. Most diseases that activate HML-2 infection appear to only activate a subset of HML-2 loci; it is possible that, though the elements necessary to reconstitute infectious HML-2 are in fact present in humans, they are not always found together in the same genome, and even more rarely expressed together; and thus infectious recombination could occur, but be exceedingly rare (100,201,202). The frequency estimations produced in this study should make it possible to investigate how often any given set of proviruses are actually found in the same person; this, in combination with our lab's ongoing projects to profile HML-2 transcription in different diseases and to efficiently identify new HML-2 integration sites with targeted amplification and sequencing, presents for the first time the possibility of a relatively rigorous search for such infectious chimeras.

Our discovery process shares limitations common to all read-based discovery methods. Given the variability in per-sample coverage, we may have missed other sites that may be present in one or few samples, or insertions located in otherwise inaccessible regions of the genome. This is exacerbated by the relatively low coverage (4-6x) of the

1000 Genomes Project data; by comparison, 30x coverage is the standard for clinical applications of whole genome sequencing. Continued improvements in sequencing technologies (longer read lengths and deeper coverage) and costs will ameliorate such issues in the future. Such changes will also increase the feasibility of assembly-based approaches, permitting the direct reconstruction of full insertion, ultimately contributing to a more a more complete picture of all types of genomic variation.

Though we have not completely ruled out the possibility of ongoing replication of HML-2 in humans, it is clear that any such replication would be exceedingly rare, and thus in the vast majority of humans HML-2 is probably functionally extinct. As discussed in the introduction, a number of factors could conceivably have played a role in this, including evolution of host resistance, either passive adaptations such as mutations that abrogate Env-receptor binding, or active adaptations such as restriction factors. As HML-2 was clearly quite infectious in the human lineage after the split from chimpanzees, any such evolutionary adaptations would likely be human-specific. Thus, we thought it could be informative to look at HML-2s in chimpanzees and gorillas to see if perhaps they had a different fate in our sister lineages. Additionally, it is conceivable that infectious HML-2 could be found in one of our relatives. As HML-2s are present in many different Old World primates, it seems likely that HML-2 has survived to the present day in at least one species. Past research in other labs as well as our own preliminary investigations suggested that relatively few chimpanzee-specific HML-2s exist. In contrast, earlier work in our lab on gorilla proviruses seemed quite promising. Though no gorilla-specific insertions were definitively identified due to the problems with the gorilla genome assembly detailed in chapter 4, many HML-2 sequences in the gorilla assembly did not

match any known proviruses, suggesting that they could belong to as yet undiscovered gorilla-specific proviruses.

In order to investigate this, we adapted the data mining methods we used for humans to work with high throughput sequencing data from the Great Ape Genome Project. This mining produced hits to 675 putative insertions, including 126 sites with reads mapping to both LTR junctions; from this list we validated and sequenced 27 previously undescribed gorilla-specific insertions, including 21 solo LTRs and 6 2-LTR proviruses. We also searched the newly released long read assembly gorGor5 reference genome for HML-2 sequences, and identified a further 10 2-LTR proviruses and 92 solo LTRs, for a total of 129 previously unreported, validated gorilla-specific HML-2 insertions. We have only screened 66 of the 675 GAGP hits, and have only screened three gorillas for those hits; thus it is likely that many or most of those hits represent genuine insertions, which suggest a remarkably high level of HML-2 polymorphism in gorillas as compared to humans, which have ~150 human specific HML-2 insertions in total.

We developed a new method to estimate the insertion times of solo LTRs, which produces ages congruent with what would be expected based on their species distribution, though it gives somewhat older ages than the 5'-3' LTR divergence method. Phylogenetic and molecular clock analysis of the LTR sequences from the gorilla-specific insertions showed that almost all of them cluster in a single clade, separate from known human and chimpanzee specific sequences, and include a very low divergence subclade with a median age of ~0.6 million years, 1 million years younger than the most comparable clade of human-specific sequences. Many of the LTRs are in fact identical to each other, and thus could have integrated very recently, at most ~300,000 years ago;

more precise age estimation methods such as coalescence dating could potentially give better age resolution in the future. In contrast, only a few human-specific proviruses have such low sequence divergence, and none of them cluster together phylogenetically as we would expect if they were due to recent infection. Additionally, several of the proviruses discovered have intact ORFs for viral genes, including one provirus at 9p13.3 with ORFs for *gag, pro, pol* and *env.*

Taken together, the high copy number, high levels of polymorphism, low sequence divergence and the presence of well-preserved proviruses strongly suggest that HML-2 has been active more recently in gorillas than in humans, and could potentially still be infecting gorillas today. In contrast, chimpanzees appear to have lost HML-2 activity earlier than humans or gorillas. It is interesting to speculate what may have driven the differential fate of HML-2 in these three lineages. Some possible avenues of research include: searching for signals of positive selection in HML-2 genomes, known restriction factors and the putative HML-2 env receptor, if and when said receptor is discovered; testing the activity of different primate restriction factors on HML-2 replication; and investigating the possibility that some HML-2 loci have been coopted to help restrict exogenous replication, perhaps by searching for signs of purifying selection at particular HML-2 loci.

A number of the sequenced full length proviruses have a shared deletion that eliminates portions of *pro* and *pol*. Like the previously known type 1 proviruses with their shared 292 bp deletion, these proviruses do not appear to have diverged in sequence from the wild type genomes, and thus presumably replicated by copackaging and recombining with replication competent HML-2 viruses. It is curious that these mutant

genomes seem to be maintained in the viral population, despite no obvious benefit to the virus.

The intact provirus at 9p13.3 is probably worth testing for replication competence, though of course we have been down this path before with both Xq21.33 and prior to that with K113. Further screening of the hits from the GAGP will likely be rewarding, as well as potentially mining other primate genomes. One intriguing possibility is that some of the hits could be derived from somatic integrations, which would strongly indicate ongoing viral replication. It seems likely that we have only scratched the surface of HML-2 diversity in primates, and exploring that diversity could tell us volumes about HML-2's enigmatic role in human biology. In combination with direct investigation into HML-2 expression in disease, population distribution and integration in humans, it may be possible to finally answer questions about HML-2's infectivity, pathogenicity, and evolution with confidence.

Chapter 6: References

1. Jern P, Coffin JM. Effects of Retroviruses on Host Genome Function. Annu Rev Genet. 2008;42(1):709–32.

2. Coffin JM, Hughes SH, Varmus HE. Retroviruses. Retroviruses. 1997.

3. Havecker ER, Gao X, Voytas DF. The diversity of LTR retrotransposons. Genome Biol. 2004;5(6):225.

4. Eickbush TH, Malik HS. Origins and Evolution of Retrotransposons. Mobile DNA II. 2014. 1111-1144 p.

5. Boeke JD, Stoye JP. Retrotransposons, endogenous retroviruses and the evolution of retroviruses. In: In Retroviruses. 1997. p. 343–435.

6. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. 2011;1(1).

7. Matsuoka M. Human T-cell leukemia virus type I (HTLV-I) infection and the onset of adult T-cell leukemia (ATL). Retrovirology. 2005;2(1):27.

8. Ishitsuka K, Tamura K. Human T-cell leukaemia virus type I and adult T-cell leukaemia-lymphoma. Vol. 15, The Lancet Oncology. 2014. p. e517–26.

9. Murphy EL, Fridey J, Smith JW, Engstrom J, Sacher R a, Miller K, Gibble J, Stevens J, Thomson R, Hansma D, Kaplan J, Khabbaz R, Nemo G. HTLV-associated myelopathy in a cohort of HTLV-I and HTLV-II-infected blood donors. The REDS investigators. Neurology. 1997;48(2):315–20.

10. Mahieux R, Gessain A. The human HTLV-3 and HTLV-4 retroviruses: New members of the HTLV family. Vol. 57, Pathologie Biologie. 2009. p. 161–6.

11. Mahieux R, Gessain A. HTLV-3/STLV-3 and HTLV-4 viruses: Discovery, epidemiology, serology and molecular aspects. Vol. 3, Viruses. 2011. p. 1074–90.

12. Coffin JM, Hughes SH, Varmus HE. The Interactions of Retroviruses and their Hosts. Retroviruses. 1997;

13. Mertz JA, Simper MS, Lozano MM, Payne SM, Dudley JP. Mouse Mammary Tumor Virus Encodes a Self-Regulatory RNA Export Protein and Is a Complex Retrovirus. J Virol. 2005;79(23):14737–47.

14. Indik S, Günzburg WH, Salmons B, Rouault F. A novel, mouse mammary tumor virus encoded protein with Rev-like properties. Virology. 2005;337(1):1–6.

15. Mertz JA, Lozano MM, Dudley JP. Rev and Rex proteins of human complex retroviruses function with the MMTV Rem-responsive element. Retrovirology. 2009;6:10.

16. Vogt V. Retroviral Virions and Genomes. In: Retroviruses. 1997. p. 1–5.

17. Coffin JM, Hughes SH, Varmus HE. Historical Introduction to the General Properties of Retroviruses. In: Retroviruses. 1997.

18. Telesnitsky A, Goff SP. Reverse transcriptase and the generation of retroviral DNA. Retroviruses. 1997;121–60.

19. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006;3(1):67.

20. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev Microbiol. 2012;10(6):395–406.

21. Brown PO. Integration. In: Retroviruses. 1997.

22. Rabson AB, Graves BJ. Synthesis and Processing of Viral RNA. Retroviruses. 1997;205–62.

23. Hohn O, Hanke K, Bannert N. HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Front Oncol. 2013;3(September):246.

24. Swanstrom R, Wills JW. Synthesis, Assembly, and Processing of Viral Proteins. In: Retroviruses. 1997. p. 263–334.

25. Coffin J, Swanstrom R. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. Cold Spring Harb Perspect Med. 2013;3(1).

26. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science. 1995;267(5197):483–9.

27. Rosenberg N, Jolicoeur P. Retroviral Pathogenesis. Retroviruses. 1997;475–585.

28. Fan H, Johnson C. Insertional oncogenesis by non-acute retroviruses: Implications for gene therapy. Viruses. 2011;3(4):398–422.

29. Kurth R, Bannert N. Retroviruses : molecular biology, genomics, and pathogenesis. Caister Academic Press; 2010. 454 p.

30. Costin JM. Cytopathic mechanisms of HIV-1. Virol J. 2007;4:100.

31. Palmer S, Josefsson L, Coffin JM. HIV reservoirs and the possibility of a cure for HIV infection. J Intern Med. 2011;270(6):550–60.

32. Churchill MJ, Deeks SG, Margolis DM, Siliciano RF, Swanstrom R. HIV reservoirs: what, where and how to target them. Nat Rev Microbiol. 2016;14(1):1–6.

33. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, Anderson EM, Watters SA, Hill S, Wu X, Wells D, Su L, Luke BT, Halvas EK, Besson G, Penrose KJ, Yang Z, Kwan RW, Van Waes C, Uldrick T, Citrin DE, Kovacs J, Polis MA, Rehm CA, Gorelick R, Piatak M, Keele BF, Kearney MF, Coffin JM, Hughes SH, Mellors JW, Maldarelli F. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. Proc Natl Acad Sci U S A. 2016;113(7):1883–8.

34. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, Kovacs C, Gange SJ, Siliciano RF. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. Nat Med. 2003;9(6):727–8.

35. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, Coffin JM, Hughes SH. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science. 2014;345(6193):179–83.

36. Hughes SH, Coffin JM. What Integration Sites Tell Us about HIV Persistence. Cell Host Microbe. 2016;19(5):588–98.

37. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13(4):283–96.

38. Magiorkinis G, Belshaw R, Katzourakis A. "There and back again": revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. Philos Trans R Soc B Biol Sci. 2013;368(1626):20120504–20120504.

39. Bannert N, Kurth R. The Evolutionary Dynamics of Human Endogenous Retroviral Families. Annu Rev Genomics Hum Genet. 2006;7:149–73.

40. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. Larger Mammalian Body Size Leads to Lower Retroviral Activity. PLoS Pathog. 2014;10(7).

41. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M. Rate of recombinational deletion among human endogenous retroviruses. J Virol. 2007;81(17):9437–42.

42. Gemmell P, Hein J, Katzourakis A. Orthologous endogenous retroviruses exhibit directional selection since the chimp-human split. Retrovirology. 2015;12:52.

43. Malik HS, Henikoff S, Eickbush TH. Poised for Contagion : Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses Poised for Contagion : Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. 2000;1307–18.

44. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A. 2012;109(19):7385–90.

45. Jern P, Sperber GO, Blomberg J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology. 2005;2:50.

46. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011;8(1):90.

47. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A. 2004;101(6):1668–72.

48. Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet. 2001;29(4):487–9.

49. Mayer J, Blomberg J, Seal RL. A revised nomenclature for transcribed human endogenous retroviral loci. Mob DNA. 2011;2(1):7.

50. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J. Classification and nomenclature of endogenous retroviral sequences (ERVs). Problems and recommendations. Gene. 2009;448(2):115–23.

51. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A. 1999;96(18):10254–60.

52. Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J. Retroposon insertion patterns of neoavian birds: Strong evidence for an extensive incomplete lineage sorting era. Mol Biol Evol. 2012;29(6):1497–501.

53. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, Brosius J, Kriegs JO, Schmitz J. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nat Commun. 2011;2(1):443.

54. Doronina L, Churakov G, Kuritzin A, Shi J, Baertsch R, Clawson H, Schmitz J. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome Res. 2017;1–7.

55. Hillis DM. SINEs of the perfect character. Proc Natl Acad Sci. 1999;96(18):9979–81.

56. Ray DA, Xing J, Salem AH, Batzer MA. SINEs of a nearly perfect character. Syst Biol. 2006;55(6):928–35.

57. Chessa B, Pereira F, Arnaud F, Amorim A, Goyache F, Mainland I, Kao RR, Pemberton JM, Beraldi D, Stear MJ, Alberti A, Pittau M, Iannuzzi L, Banabazi MH, Kazwala RR, Zhang Y -p., Arranz JJ, Ali BA, Wang Z, Uzun M, Dione MM, Olsaker I, Holm L-E, Saarma U, Ahmad S, Marzanov N, Eythorsdottir E, Holland MJ, Ajmone-Marsan P, Bruford MW, Kantanen J, Spencer TE, Palmarini M. Revealing the History of Sheep Domestication Using Retrovirus Integrations. Science (80-). 2009;324(5926):532–6.

58. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 2006;4(4):537–44.

59. Hughes JF, Coffin JM. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics. 2005;171(3):1183–94.

60. Gifford RJ. Viral evolution in deep time: Lentiviruses and mammals. Trends Genet. 2012;28(2):89–100.

61. Aiewsakun P, Katzourakis A. Time-dependent rate phenomenon in viruses. J Virol. 2016;90(16):7184–95.

62. Holmes EC. Molecular clocks and the puzzle of RNA virus origins. J Virol. 2003;77(7):3893–7.

63. Nair S, Rein A. Antiretroviral restriction factors in mice. Virus Res. 2014;193:130–4.

64. Stoye JP. Fv1, the mouse retrovirus resistance gene. Rev Sci Tech. 1998;17(1):269–77.

65. Bieniasz PD. Restriction factors: A defense against retroviral infection. Vol. 11, Trends in Microbiology. 2003. p. 286–91.

66. Nihrane A, Lebedeva I, Lyu MS, Fujita K, Silver J. Secretion of a murine retroviral Env associated with resistance to infection. J Gen Virol. 1997;78(4):785–93.

67. Goff SP. Retrovirus restriction factors. Mol Cell. 2004;16(6):849–59.

68. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Elife. 2017;6:1–19.

69. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene. 2009;448(2):105–14.

70. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science (80-). 2016;351(6277):1083–7.

71. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs N V., Schumann GG, Chen W, Lorincz MC, Ivics Z, Hurst LD, Izsvák Z. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516(7531):405–9.

72. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang X-Y, Edouard P, Howes S, Keith JC, McCoy JM. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000;403(6771):785–9.

73. Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Veron G, Mulot B, Dupressoir A, Heidmann T. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc Natl Acad Sci U S A. 2012;109(7):E432-41.

74. Cornelis G, Heidmann O, Degrelle SA, Vernochet C, Lavialle C, Letzelter C, Bernard-Stoecklin S, Hassanin A, Mulot B, Guillomot M, Hue I, Heidmann T, Dupressoir A. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. Proc Natl Acad Sci U S A. 2013;110(9):E828-37.

75. Cáceres M, Thomas JW. The gene of retroviral origin syncytin I is specific to hominoids and is inactive in old world monkeys. J Hered. 2006;97(2):100–6.

76. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation. Philos Trans R Soc B Biol Sci. 2013;368(1626):20120507–20120507.

77. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, Nilsson MA, Menzies BR, Renfree MB, Pierron G, Zeller U, Heidmann O, Dupressoir A, Heidmann T. Retroviral envelope gene captures and *syncytin* exaptation for placentation in marsupials. Proc Natl Acad Sci. 2015;112(5):E487–96.

78. Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T. Syncytin-A knockout mice demonstrate the critical role in placentation of a

fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci. 2009;106(29):12127–32.

79. Young GR, Stoye JP, Kassiotis G. Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. BioEssays. 2013;35(9):794–803.

80. Lenz J, Crowther R, Klimenko S, Haseltine W. Molecular cloning of a highly leukemogenic, ecotropic retrovirus from an AKR mouse. J Virol. 1982;43(3):943–51.

81. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G. Resurrection of endogenous retroviruses in antibody-deficient mice. Nature. 2012;491(7426):774–8.

82. Stoye JP, Moroni C, Coffin JM. Virological events leading to spontaneous AKR thymomas. J Virol. 1991;65(3):1273–85.

83. Morandi E, Tanasescu R, Tarlinton RE, Constantinescu CS, Zhang W, Tench C, Gran B. The association between human endogenous retroviruses and multiple sclerosis: A systematic review and meta-analysis. PLoS One. 2017;12(2):e0172415.

84. Antony JM, Ellestad KK, Hammond R, Imaizumi K, Mallet F, Warren KG, Power C. The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes. J Immunol. 2007;179(2):1210–24.

85. Schmitt K, Richter C, Backes C, Meese E, Ruprecht K, Mayer J. Comprehensive analysis of human endogenous retrovirus group HERV-W locus transcription in multiple sclerosis brain lesions by high-throughput amplicon sequencing. J Virol. 2013;87(24):13837–52.

86. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol. 1999;9(16):861–8.

87. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol. 2001;11(19):1531–5.

88. Beimforde N, Hanke K, Ammar I, Kurth R, Bannert N. Molecular cloning and functional characterization of the human endogenous retrovirus K113. Virology. 2008;371(1):216–25.

89. Boller K, Schönfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tönjes RR. Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. J Gen Virol. 2008;89(2):567–72.

90. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. 2006;16(12):1548–56.

91. Young NL, Bieniasz PD. Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog. 2007;3(1):0119–30.

92. Löwer R, Tönjes RR, Korbmacher C, Kurth R, Löwer J. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J Virol. 1995;69(1):141–9.

93. Reus K, Mayer J, Sauter M, Scherer D, Müller-Lantzsch N, Meese E. Genomic organization of the human endogenous retrovirus HERV-K(HML-2.HOM) (ERVK6) on chromosome 7. Genomics. 2001;72(3):314–20.

94. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N. A novel gene from the human endogenous retrovirus K expressed in transformed cells. Clin Cancer Res. 2002;8(6):1800–7.

95. Barbulescu M, Turner G, Su M, Kim R, Jensen-Seaman MI, Deinard AS, Kidd KK, Lenz J. A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. Curr Biol. 2001;11(10):779–83.

96. Kämmerer U, Germeyer A, Stengel S, Kapp M, Denner J. Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta. J Reprod Immunol. 2011;91(1–2):1–8.

97. Grow EJ, Flynn R a., Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish C a., Chang HY, Pera RAR, Wysocka J. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522(7555):221–5.

98. Wang-Johanning F, Li M, Esteva FJ, Hess KR, Yin B, Rycaj K, Plummer JB, Garza JG, Ambs S, Johanning GL. Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer. Int J Cancer. 2013;134(3):587–95.

99. Franklin GC, Chretien S, Hanson IM, Rochefort H, May FE, Westley BR. Expression of human sequences related to those of mouse mammary tumor virus. J Virol. 1988;62(4):1203–10.

100. Schmitt K, Reichrath J, Roesch A, Meese E, Mayer J. Transcriptional profiling of

human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. Genome Biol Evol. 2013;5(2):307–28.

101. Boller K, König H, Sauter M, Mueller-Lantzsch N, Löwer R, Löwer J, Kurth R. Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Vol. 196, Virology. 1993. p. 349–53.

102. Sauter M, Schommer S, Kremmer E, Remberger K, Dolken G, Lemm I, Buck M, Best B, Neumann-Haefelin D, Mueller-Lantzsch N. Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. J Virol. 1995;69(1):414–21.

103. Li W, Lee M, Henderson L, Tyagi R, Bachani M, Steiner J, Campanac E, Hoffman DA, von Geldern G, Johnson K, Maric D, Morris HD, Lentz M, Pak K, Mammen A, Ostrow L, Rothstein J, Nath A. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med. 2015;7(307):307ra153.

104. Christensen T. Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. Vol. 15, Reviews in Medical Virology. 2005. p. 179–211.

105. Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C. Monocyte activation and differentiation augment human endogenous retrovirus expression: Implications for inflammatory brain diseases. Ann Neurol. 2001;50(4):434–42.

106. Hahn S, Ugurel S, Hanschmann K-M, Strobel H, Tondera C, Schadendorf D, Löwer J, Löwer R. Serological Response to Human Endogenous Retrovirus K in Melanoma Patients Correlates with Survival Probability. AIDS Res Hum Retroviruses. 2008;24(5):717–23.

107. Boller K, Janssen O, Schuldes H, Tonjes RR, Kurth R. Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J Virol. 1997;71(6):4581–8.

108. Otowa T, Tochigi M, Rogers M, Umekage T, Kato N, Sasaki T. Insertional polymorphism of endogenous retrovirus HERV-K115 in schizophrenia. Neurosci Lett. 2006;408(3):226–9.

109. Wildschutte JH, Ram D, Subramanian R, Stevens VL, Coffin JM,. The distribution of insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls. Retrovirology. 2014;11(1):62.

110. Chen T, Meng Z, Gan Y, Wang X, Xu F, Gu Y, Xu X, Tang J, Zhou H, Zhang X, Gan X, Van Ness C, Xu G, Huang L, Zhang X, Fang Y, Wu J, Zheng S, Jin J, Huang W, Xu R. The viral oncogene Np9 acts as a critical molecular switch for

co-activating β-catenin, ERK, Akt and Notch1 and promoting the growth of human leukemia stem/progenitor cells. Leukemia. 2013 Jul 11;27(7):1469–78.

111. Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, Roemer K, Mueller-Lantzsch N. Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene. 2005;24(19):3223–8.

112. Hsiao FC, Lin M, Tai A, Chen G, Huber BT. Cutting edge: Epstein-Barr virus transactivates the HERV-K18 superantigen by docking to the human complement receptor 2 (CD21) on primary B cells. J Immunol. 2006;177(4):2056–60.

113. Sutkowski N, Conrad B, Thorley-Lawson D a, Huber BT. Epstein-Barr virus transactivates the human endogenous retrovirus HERV- K18 that encodes a superantigen. Immunity. 2001;15(4):579–89.

114. Kozak CA. The mouse "xenotropic" gammaretroviruses and their XPR1 receptor. Retrovirology. 2010;7(1):101.

115. Martin C, Buckler-White A, Wollenberg K, Kozak CA. The avian XPR1 gammaretrovirus receptor is under positive selection and is disabled in bird species in contact with virus-infected wild mice. J Virol. 2013;87(18):10094–104.

116. Robinson LR, Whelan SPJ. Infectious entry pathway mediated by the human endogenous retrovirus K envelope protein. J Virol. 2016;90(7):JVI.03136-15.

117. Chiu Y-L, Greene WC. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annu Rev Immunol. 2008;26:317–53.

118. Armitage AE, Katzourakis A, de Oliveira T, Welch JJ, Belshaw R, Bishop KN, Kramer B, McMichael AJ, Rambaut A, Iversen AKN. Conserved footprints of APOBEC3G on Hypermutated human immunodeficiency virus type 1 and human endogenous retrovirus HERV-K(HML2) sequences. J Virol. 2008;82(17):8743–61.

119. Esnault C, Priet S, Ribet D, Heidmann O, Heidmann T. Restriction by APOBEC3 proteins of endogenous retroviruses with an extracellular life cycle: ex vivo effects and in vivo "traces" on the murine IAPE and human HERV-K elements. Retrovirology. 2008;5:75.

120. Lee YN, Malim MH, Bieniasz PD. Hypermutation of an ancient human retrovirus by APOBEC3G. J Virol. 2008;82(17):8762–70.

121. Jouvenet N, Neil SJD, Zhadina M, Zang T, Kratovac Z, Lee Y, McNatt M, Hatziioannou T, Bieniasz PD. Broad-Spectrum Inhibition of Retroviral and

Filoviral Particle Release by Tetherin. J Virol. 2009;83(4):1837–44.

122. Zheng Y-H, Jeang K-T, Tokunaga K. Host restriction factors in retroviral infection: promises in virus-host interaction. Retrovirology. 2012;9(1):112.

123. Arjan-Odedra S, Swanson CM, Sherer NM, Wolinsky SM, Malim MH. Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication of exogenous retroviruses. Retrovirology. 2012;9:53.

124. St Gelais C, Wu L. SAMHD1: a new insight into HIV-1 restriction in myeloid cells. Retrovirology. 2011;8(July):55.

125. Robinson HL, Lamoreux WF. Expression of endogenous ALV antigens and susceptibility to subgroup E ALV in three strains of chickens (endogenous avian C-type virus). Virology. 1976;69(1):50–62.

126. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A. 2004;101(14):4894–9.

127. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Imlications for Present-Day Activity. J Virol. 2005;79(19):973.

128. Marchi E, Kanapin A, Magiorkinis G, Belshaw R. Unfixed endogenous retroviral insertions in the human population. J Virol. 2014;88(17):9529–37.

129. Xu WQ, Stadler CK, Gorman K, Jensen N, Kim D, Zheng HQ, Tang SH, Switzer WM, Pye GW, Eiden M V. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proc Natl Acad Sci U S A. 2013;110(28):11547–52.

130. Shojima T, Yoshikawa R, Hoshino S, Shimode S, Nakagawa S, Ohata T, Nakaoka R, Miyazawaa T. Identification of a Novel Subgroup of Koala Retrovirus from Koalas in Japanese Zoos. J Virol. 2013;87(17):9943–8.

131. Chappell KJ, Brealey JC, Amarilla AA, Watterson D, Hulse L, Palmieri C, Johnston SD, Holmes EC, Meers J, Young PR. Phylogenetic diversity of Koala Retrovirus within a Wild Koala Population. J Virol. 2016;(November):JVI.01820-16.

132. Stoye JP. Koala retrovirus: a genome invasion in real time. Genome Biol. 2006;7(11):241.

133. Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome.

Nature. 2006;442(7098):79–81.

134. Kijima TE, Innan H. On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol. 2010;27(4):896–904.

135. Jha AR, Pillai SK, York VA, Sharp ER, Storm EC, Wachter DJ, Martin JN, Deeks SG, Rosenberg MG, Nixon DF, Garrison KE. Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. Mol Biol Evol. 2009;26(11):2617–26.

136. Jha AR, Nixon DF, Rosenberg MG, Martin JN, Deeks SG, Hudson RR, Garrison KE, Pillai SK. Human endogenous retrovirus K106 (HERV-k106) was infectious after the emergence of anatomically modern humans. PLoS One. 2011;6(5):1–8.

137. Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 2007;3(2):0294–304.

138. Jensen-Seaman MI, Hooper-Boyd K a. Molecular Clocks: Determining the Age of the Human-Chimpanzee Divergence. eLS. 2008;1–6.

139. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. Genetic evidence for complex speciation of humans and chimpanzees. 2006;441(June):1–6.

140. Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 2011;21(3):349–56.

141. Rogers J, Gibbs RA. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet. 2014;15(5):347–59.

142. Agoni L, Golden A, Guha C, Lenz J. Neandertal and Denisovan retroviruses. Curr Biol. 2012;22(11):R437–8.

143. Keane TM, Wong K, Adams DJ. RetroSeq: Transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29(3):389–90.

144. Fiston-Lavier A-S, Carrigan M, Petrov D a, González J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. 2010;39(6):1–10.

145. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677–81.

146. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010;141(7):1253–61.

147. Guo Y, Levin HL. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. Genome Res. 2010;20(2):239–48.

148. Witherspoon DJ, Zhang Y, Xing J, Watkins WS, Ha H, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. Genome Res. 2013;23(7):1170–81.

149. 1000 Genomes Project Consortium T 1000 GP, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.

150. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Vol. 526, Nature. 2015. p. 68–74.

151. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 2005;6(4):333–40.

152. Church GM. The Personal Genome Project. Mol Syst Biol. 2005;1(1):E1–3.

153. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci. 2016;201602336.

154. Cann HM, de Toma C, Cazes L, Legrand MF. A human genome diversity cell line panel. Science (80-). 2002;296(5566):261.

155. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, Grubert F, Cann HM, Snyder M, Montgomery SB, Bustamante CD. Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. PLoS Genet. 2014;10(8).

156. Romano CM, de Melo FL, Corsini MAB, Holmes EC, Zanotto PM de A. Demographic histories of ERV-K in humans, chimpanzees and rhesus monkeys. PLoS One. 2007;2(10).

157. Romano CM, Ramalho RF, Zanotto PMDA. Tempo and mode of ERV-K evolution in human and chimpanzee genomes. Arch Virol. 2006;151(11):2215–28.

158. Macfarlane CM, Badge RM. Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and

chimpanzees that are absent from genome assemblies. Retrovirology. 2015;12(1):35.

159. Wu HL, Léon EJ, Wallace LT, Nimiyongskul FA, Buechler MB, Newman LP, Castrovinci PA, Paul Johnson R, Gifford RJ, Brad Jones R, Sacha JB. Identification and spontaneous immune targeting of an endogenous retrovirus K envelope protein in the Indian rhesus macaque model of human disease. Retrovirology. 2016;13(1):6.

160. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball E V., Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169–75.

161. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437(7055):69–87.

162. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR, Mullikin JC, Meader SJ, Ponting CP, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoç E, Alkan C, Sajjadian S, Catacchio CR, Ventura M, Marques-Bonet T, Eichler EE, André C, Atencia R, Mugisha L, Junhold J, Patterson N, Siebauer M, Good JM, Fischer A, Ptak SE, Lachmann M, Symer DE, Mailund T, Schierup MH, Andrés AM, Kelso J, Pääbo S. The bonobo genome compared with the chimpanzee and human genomes. Nature. 2012;

163. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth L V., Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darré F, Farré D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Valle G Della, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AFA, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordoñez GR, López-Otín C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A,

RoyChoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA, Wilson RK. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469(7331):529–33.

164. Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, Lee C, Turner EH, Chen L, Kidd JM, Archidiacono N, Shendure J, Wilson RK, Eichler EE. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. Genome Res. 2011;21(10):1640–9.

165. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. Great ape genetic diversity and population history. Nature. 2013;499(7459):471–5.

166. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

167. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

168. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. Great ape genetic

diversity and population history. Nature. 2013;499(7459):471–5.

169. Karolchik D, Hinrichs AS, Kent WJ. The UCSC genome browser. Curr Protoc Hum Genet. 2011;(SUPPL. 71).

170. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

171. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

172. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.

173. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):11.

174. Smit A, Hubley R, Green P. RepeatMasker Open-3.0. RepeatMasker Open-3.0. 1996. p. www.repeatmasker.org.

175. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9(9):868–77.

176. Wildschutte JH, Baron A, Diroff NM, Kidd JM. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res. 2015;43(21):10292–307.

177. Wharton D. Gorilla, Western Lowland Studbook 2007. 2007.

178. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.

179. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

180. Larkin M a, Blackshields G, Brown NP. ClustalW2 and ClustalX version 2. …. 2007;1–2.

181. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

182. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 2008;9(4):299–306.

183. Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16(2):111–20.

184. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. Database resources of the national center for biotechnology. Vol. 31, Nucleic Acids Research. 2003. p. 28–33.

185. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schäffer AA, Yu YK. Protein database searches using compositionally adjusted substitution matrices. Vol. 272, FEBS Journal. 2005. p. 5101–9.

186. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Vol. 25, Nucleic Acids Research. 1997. p. 3389–402.

187. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette 3rd LJ, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko P V, Park PJ, Cancer Genome Atlas Research Network. Landscape of somatic retrotransposition in human cancers. Science (80-). 2012;337(6097):967–71.

188. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Graves T, Hansen N, Teague B, Alkan C, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Mckernan K, Chen L, Malig M, Smith JD, M J, Mccarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson M V, Kaul R, R D, Eichler EE. Mapping and sequencing of structural variation from eight human genomes. Evaluation. 2008;453(7191):56–64.

189. Lee A, Huntley D, Aiewsakun P, Kanda RK, Lynn C, Tristem M. Novel Denisovan and Neanderthal retroviruses. J Virol. 2014;88(21):12907–9.

190. Parsons JD. Miropeats: Graphical DNA sequence comparisons. Bioinformatics. 1995;11(6):615–9.

191. Contreras-Galindo R, Kaplan MH, Contreras-Galindo a. C, Gonzalez-Hernandez MJ, Ferlenghi I, Giusti F, Lorenzo E, Gitlin SD, Dosik MH, Yamamura Y, Markovitz DM. Characterization of Human Endogenous Retroviral Elements in the Blood of HIV-1-Infected Individuals. J Virol. 2012;86:262–76.

192. Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, Venables PJ. The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. Genomics. 2005;86(3):337–41.

193. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov M V, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053–60.

194. Patterson N, Reich D, Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507(7492):354–7.

195. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M. Higher levels of Neanderthal ancestry in east Asians than in Europeans. Genetics. 2013;194(1):199–209.

196. Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Slatkin M, Reich D. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012;222(2012):1–14.

197. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova L V, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–22.

198. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. Proc Natl Acad Sci. 2012;109(44):17758–64.

199. Hughes JF, Coffin JM. A novel endogenous retrovirus-related element in the human genome resembles a DNA transposon: Evidence for an evolutionary link? Genomics. 2002;80(5):453–5.

200. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, Dunn C, Baker C, Armstrong J, Diekhans M, Paten B, Shendure J, Wilson RK, Haussler D, Chin C-S, Eichler EE. Long-read sequence assembly of the gorilla genome. Science (80-). 2016;352(6281):aae0344-aae0344.

201. Bhardwaj N, Montesio M, Roy F, Coffin JM. Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line tera-1.

Viruses. 2015;7(3):939–68.

202. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Müller-Lantzsch N, Leib-Mösch C, Meese E, Mayer J. Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics. 2008;9:354.