PLOS BIOLOGY

RESEARCH ARTICLE

Widespread expression of the ancient HERV-K (HML-2) provirus group in normal human tissues

Aidan Burn[1], Farrah Roy[2], Michael Freeman[1], John M. Coffin[1,3]*

1 Program in Genetics, Graduate School of Biomedical Sciences, Tufts University, Boston, Massachusetts, United States of America, **2** Immuneering Corporation, Cambridge, Massachusetts, United States of America, **3** Department of Molecular Biology and Microbiology, Tufts University, Boston, Massachusetts, United States of America

* john.coffin@tufts.edu

Abstract

Human endogenous retrovirus (HERV) transcripts are known to be highly expressed in cancers, yet their activity in nondiseased tissue is largely unknown. Using the GTEx RNA-seq dataset from normal tissue sampled at autopsy, we characterized individual expression of the recent HERV-K (HML-2) provirus group across 13,000 different samples of 54 different tissues from 948 individuals. HML-2 transcripts could be identified in every tissue sampled and were elevated in the cerebellum, pituitary, testis, and thyroid. A total of 37 different individual proviruses were expressed in 1 or more tissues, representing all 3 LTR5 subgroups. Nine proviruses were identified as having long terminal repeat (LTR)-driven transcription, 7 of which belonged to the most recent LTR5HS subgroup. Proviruses of different subgroups displayed a bias in tissue expression, which may be associated with differences in transcription factor binding sites in their LTRs. Provirus expression was greater in evolutionarily older proviruses with an earliest shared ancestor of gorilla or older. HML-2 expression was significantly affected by biological sex in 1 tissue, while age and timing of death (Hardy score) had little effect. Proviruses containing intact *gag*, *pro*, and *env* open reading frames (ORFs) were expressed in the dataset, with almost every tissue measured potentially expressing at least 1 intact ORF (*gag*).

Citation: Burn A, Roy F, Freeman M, Coffin JM (2022) Widespread expression of the ancient HERV-K (HML-2) provirus group in normal human tissues. PLoS Biol 20(10): e3001826. https://doi.org/10.1371/journal.pbio.3001826

Academic Editor: Frank Kirchhoff, Ulm University Medical Center, GERMANY

Received: June 30, 2022

Accepted: September 13, 2022

Published: October 18, 2022

Data Availability Statement: All RNA-seq files are available from the GTEx project (accession numbers phs000424.v7.p2,phs000424.v8.p2) hosted in the AnVIL repository. Authorized access is required through dbGAP to work with these protected data.

Funding: JMC was supported by Research Grant R35 CA 200421 from the National Cancer Institute (https://www.cancer.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Retroviruses have been infecting mammals and other vertebrates for at least 100 million years, invading somatic and germ cells of their host species [1]. Proviral remnants of ancient retroviral infection of germ cells now make up about 8% of the human genome in the form of human endogenous retroviruses (HERVs) [2,3]. These ancient retroviruses infected the germline of ancestral primates, and the resulting proviruses are inherited in a mendelian fashion and subject to fixation or loss over time. Many of them have become significantly degraded, leaving behind remnants in the form of solo **l**ong **t**erminal **r**epeats (LTRs) and fragments of viral **o**pen **r**eading **f**rames (ORFs). Regardless, their presence in the primate genome allows one to piece

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: JMC was a member of the Scientific Advisory Board and a shareholder of ROME Therapeutics, Inc. The other authors declare no competing interests.

Abbreviations: ESA, earliest shared ancestor; GTEx, Genotype Tissue and Expression; HERV, human endogenous retrovirus; HML, human mouse mammary tumor virus-like; LTR, long terminal repeat; ORF, open reading frame; TF, transcription factor; TPM, transcripts per million.

together the history of retroviral infection in primates, a process analogous to the use of partial fossil remains of multiple individuals of a species to understand its evolutionary history.

There are at least 30 HERV subclades. Many are distantly related to exogenous viruses extant today in other animals; all represent apparently extinct retroviral lineages. Of all the HERV groups, the only subclade containing human-specific—and therefore the most recently integrated—proviruses can be found within the **h**uman **m**ouse mammary tumor virus-**l**ike (HML)-2 subgroup within the HERV-K group belonging to the *Betaretrovirus* genus. The first record of its integration dates to approximately 35 million years ago, continuing, in the human lineage, until as recently as 200,000 years ago [4–6]. HERV-K (HML-2), hereafter referred to as "HML-2," has undergone multiple bursts of integration into the human ancestral germline, leaving a number of proviruses that contain intact ORFS and polymorphic insertions across the human population (Fig 1) [4]. The HERV-K group is named for its inferred lysine tRNA primer utilized for initiating reverse transcription [7]. The LTR regions of each HML-2 provirus can be used to group the provirus phylogenetically into 3 main subgroups: LTR5A, LTR5B, and LTR5HS. The LTR5HS subgroup generally consists of more recent integrations, while the LTR5A and LTR5B subgroups are thought to have been active before human speciation [8,9].

HML-2 proviruses have accumulated insertions, deletions, and internal recombination events over time, leaving defective remnants to be enriched by selection against their potentially pathogenic ancestors. Consequently, all annotated HML-2 proviruses in the human genome are defective for replication. There are currently 94 HML-2 proviruses that retain some internal sequence—sometimes referred to as "full length"—within the hg38 human reference genome along with at least 944 solo LTRs [8,10]. We will refer to the former group as "2-LTR" proviruses. Most human HML-2 proviruses are also found at the orthologous site in chimpanzees, implying that they are >5 million years old but around 35 are human specific (implying a younger age) and at least 14 of them remain unfixed in the human population [4,10]. These relatively young proviruses have been subject to comparatively little evolutionary pressure, unlike the older proviruses, and therefore some retain at least 1 intact ORF [11].

As illustrated in Fig 1A, HML-2 proviruses can be found throughout the human genome and are enriched in regions of high gene density. This insertion pattern reflects the preference of ancestral exogenous virus integrase for integration into areas of high gene expression [12], but insertions within genes are often subject to negative selection due to deleterious effects on their host [13]. The resulting variety of integration sites leads to dramatic differences in the surrounding genomic neighborhood from one provirus to the next. Approximately 60% of HML-2 insertions are within 30 kB of genes and 20% of all HML-2 proviruses are within introns, 80% of which are antisense to the surrounding gene [8].

While no HERV provirus has been shown to retain infectivity, an integrated provirus can affect the host in multiple ways [8]. For example, genes encoding Env proteins of some ERV groups have been co-opted in mammals multiple times. The most famous examples are ERV-encoded Env proteins, now referred to as *syncytins* [14]. The cell-to-cell fusogenic ability of syncytins plays an important role in the formation of the syncytiotrophoblast during pregnancy in placental mammals [15]. In a remarkable example of convergent evolution, different ERVs play the same role in most or all other placental mammalian orders [14]. Perhaps the most important evolutionary effect is that, in some cases, expression of modified ERV Env or Gag proteins can interfere with the replication of related exogenous viruses by blocking access to a receptor or interfering with capsid assembly or intracellular trafficking. Many well-studied examples of ERV-mediated interference have been found in chickens, mice, cats, sheep, and other species [16–20]. Notably, a co-opted HERV-T *env* gene is thought to have led to the extinction of the cognate exogenous virus in our primate ancestors [21].

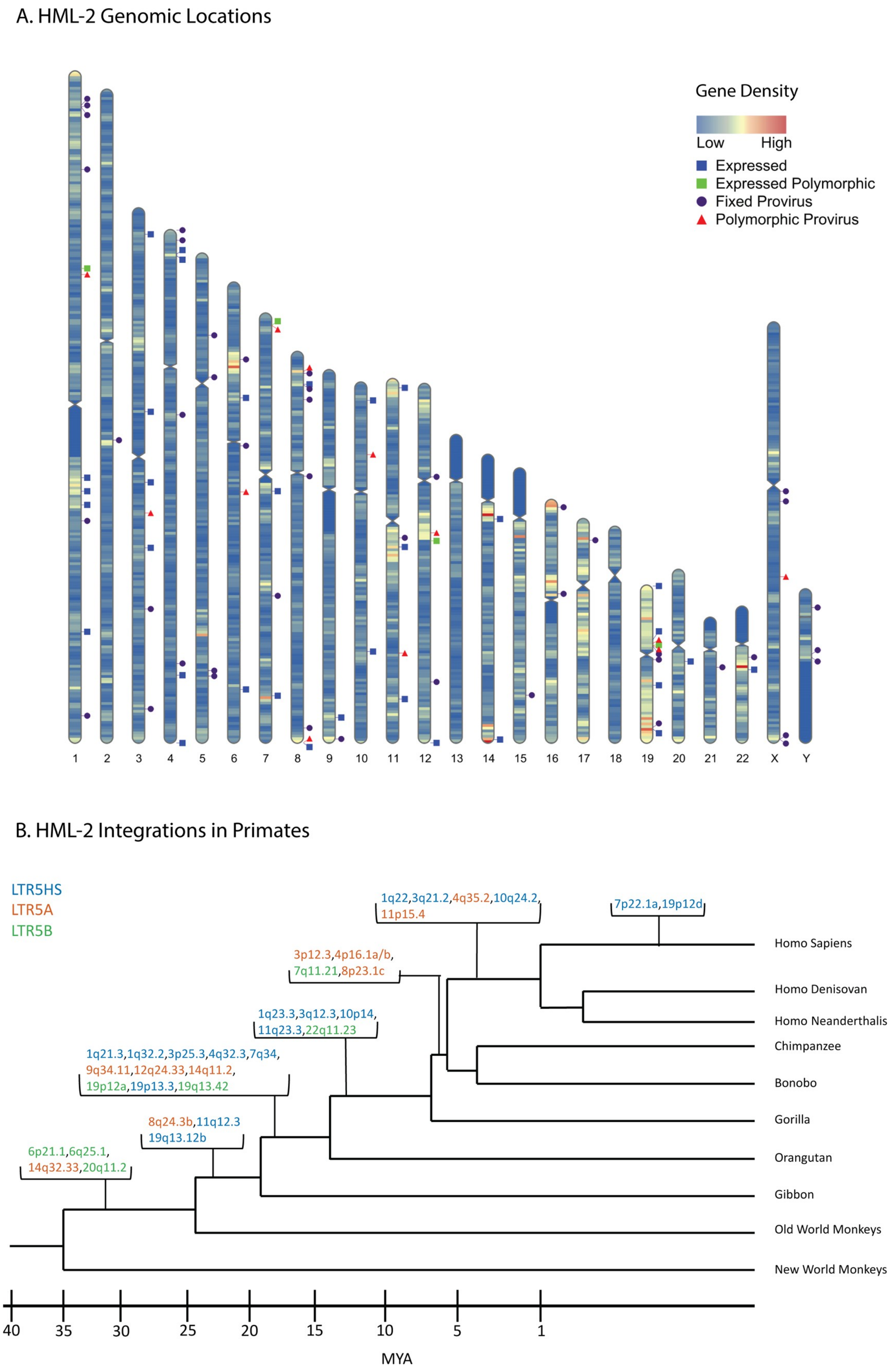

A. HML-2 Genomic Locations
Gene Density
Low
High
Expressed
Expressed Polymorphic
Fixed Provirus
Polymorphic Provirus
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
B. HML-2 Integrations in Primates
LTR5HS
LTR5A
LTR5B
1q22,3q21.2,4q35.2,10q24.2,
11p15.4
7p22.1a,19p12d
3p12.3,4p16.1a/b,
7q11.21,8p23.1c
1q23.3,3q12.3,10p14,
11q23.3,22q11.23
1q21.3,1q32.2,3p25.3,4q32.3,7q34,
9q34.11,12q24.33,14q11.2,
19p12a,19p13.3,19q13.42
8q24.3b,11q12.3
19q13.12b
6p21.1,6q25.1,
14q32.33,20q11.2
Homo Sapiens
Homo Denisovan
Homo Neanderthalis
Chimpanzee
Bonobo
Gorilla
Orangutan
Gibbon
Old World Monkeys
New World Monkeys
40
35
30
25
20
15
10
5
1
MYA

Fig 1. HERV-K HML-2 proviruses vary in chromosomal location and age of integration. (**A**) Human chromosome map showing gene density and location of 2-LTR proviruses in the genome. Purple circles represent proviruses that are fixed in the human population; red triangles represent known polymorphic proviruses; blue squares represent GTEx-expressed (TPM ≥1) proviruses. Note that the expression status of polymorphic proviruses could not be determined. Gene density is represented in a heatmap of low = blue to high = red. Prepared using RIdeogram [70]. (**B**) Primate phylogeny illustrating the ESA of HML-2 proviruses expressed in the GTEx dataset at a TPM of ≥1. As in this and subsequent figures, provirus color (blue, red, green) corresponds to LTR subtype (LTR5HS, LTR5A, and LTR 5B, respectively). The ESA groups for each provirus listed in 1B are presented in S1 Table. These ESA groups were determined by Subramanian and colleagues [8]. ESA, earliest shared ancestor; GTEx, Genotype Tissue and Expression; HERV, human endogenous retrovirus; HML, human mouse mammary tumor virus-like; LTR, long terminal repeat; MYA, million years ago; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g001

Proviruses with intact protein coding genes may be capable of expressing these genes upon activation in pathological states such as cancer. Some cancer cells and cell lines express the HML-2 structural protein Gag and can form mature viral particles [22–24], and anti-Gag antibodies have been reported in individuals suffering from multiple malignancies such as breast cancer, melanoma, and teratocarcinoma [25,26]. Another viral structural protein of disease-relevant interest is Env, a functional gene for which has been retained by at least 1 HML-2 provirus in the human genome, known as 7p22.1a or K108L. This protein retains its original fusogenic capacity, possibly allowing for formation of syncytia upon expression and has been found to be expressed in cancerous tissue and cell lines [11,27,28]. The accessory protein Rec, which is required for unspliced viral RNA transport, has been shown to interact with a tumor suppressor in germ cell tumors and its expression is linked to a number of malignancies, although no causal relationship has yet been defined [25]. A subset comprising about half of the 2-LTR HML-2 proviruses, termed "Type 1," shares an identical 292 nucleotide deletion affecting the pol-env border. This deletion removes a splice donor for Rec, creating the transcript for an ORF called NP9. When translated, the NP9 protein has also been shown to interact with and interfere with a protein in the Numb/Notch pathway in germ cell tumors [25]. The type 1 deletion is found in proviruses shared by humans, chimpanzees, gorillas and orangutans, implying that it arose more than 10 million years ago and is found only in LTR5Hs proviruses.

A provirus can also have significant effects on the host without viral gene expression. Proviruses introduced into the genome carry along with them promoters and poly(A) addition sites, as well as multiple splice acceptors. These sites can create alternative or intergenic splicing among host genes, altering transcription and affecting the integrity of the final protein product [25]. Binding sites for different transcription factors (TFs) within proviral LTR promoters can greatly affect surrounding gene expression. These LTRs can act as promoters or enhancers both while flanking or within the integrated provirus or as solo LTRs near transcription start sites. The variety of binding sites in an LTR can affect neighboring gene expression in many ways, including direct changes to tissue specificity or by enhancing the activity of an existing promoter, depending on the location of integration and the surrounding genes [29–31].

Despite the potential importance of HML-2-modified expression of viral or host genes, to date, HML-2 expression in a nondiseased human body has not been well characterized. Data on HML-2 expression exist in various disease contexts, along with evidence of expression in some nondiseased tissues [26,32], but examination of all tissues in a host has yet to be performed. Furthermore, most studies analyzing provirus expression in disease and nondisease contexts looked at total HML-2 expression rather than provirus-specific expression. Reports of total HML-2 expression fail to capture the diversity of HML-2 proviruses in the human genome and the variety of mechanisms controlling their individual expression. To create a detailed characterization, we turned to the **G**enotype **T**issue and **Ex**pression (GTEx) Project [33]. The GTEx project is a database of tissue-specific gene expression collected from 54 nondiseased tissues across 948 donors. The wealth of RNA-seq data contained in this database allows us to characterize HML-2 expression across the entire human body. By analyzing these data, we were able to discover evidence of HML-2 provirus expression in every tissue analyzed, with numerous proviruses showing significant expression across the body. The cerebellum, pituitary, testis, and thyroid showed the highest level of HML-2 expression in nondiseased tissue. A number of proviruses with intact ORFS were found to be expressed, although the functional consequences of their expression remain unclear. Common covariates such as biological sex showed significant differences between individual provirus expression on the basis of biological sex but not in other covariate groups. Furthermore, some proviruses appear to be self-driven by their 5′ LTRs, while many appear to be transcribed through other means, such as read-through transcription, either from a nearby gene or from an unannotated feature within the genome. Our study also revealed interesting patterns of expression, in that the oldest proviruses are the most expressed and most frequent provirus expression was seen in neuronal, endocrine, and reproductive tissue.

Results

Specific tissues have higher abundance of HERV-K (HML-2) transcripts

Aberrant HML-2 expression in human cancer and a number of other disease states was recognized many years ago [22,24–26], and its potential use as a biomarker for detection of illness or as therapeutic targets has long been discussed [34–36]. In addition to their clinical use, there is a great deal of evolutionary biology one can learn from studying the genetic structure and expression of these ancient viral proviruses. For example, the tissue specificity of expression of individual proviruses can shed light on the mechanisms of replication, transmission, and pathogenesis of the ancestral virus. Taking advantage of the recently released human GTEx V8 dataset [33], we sought to profile HML-2 provirus expression in healthy tissues (Fig 2). This dataset is the most comprehensive listing of RNA-seq data for human tissues, consisting of 13,851 samples across 54 different tissues in the body acquired from 948 different postmortem donors. Reads were quality trimmed (Phred score >30, min length of 75) before alignment to hg38 using HISAT2 [37]. Aligned reads were then counted using the Bayesian analysis routine of the Telescope program [38], which aligns multimapping reads to the most likely location based on a statistical model. Ambiguous reads were mapped through an iterative process that compares initial weights of read alignments to expected transcript levels. This procedure allows for more confident alignments of multimapping reads for elements, like endogenous retroviruses, which are composed of repeated sequences that differ only slightly from one another. Expression was measured as transcripts per million (TPM) [39], which normalizes read counts for library size and gene length.

Our approach detected HML-2 expression throughout the body, with identifiable transcript levels (threshold ≥1 TPM) found in each body site analyzed (Fig 3A). The influence of the site sampled was immediately clear, with total HML-2 expression levels varying considerably across the body. Four tissues—cerebellum, pituitary, testis, and thyroid—had relatively high levels of HML-2 expression in comparison to other measured tissues within the GTEx cohort: These 4 tissues expressed combined total HML-2 transcripts at an average of at least 70 TPM across all samples. In contrast, samples from pancreas and whole blood showed lower detectable total HML-2 expression of 5 and 8 TPM/sample, respectively. For comparison, the host gene RAB5A is expressed at around 37 TPM across GTEx and GAPDH is expressed at around 1,300 TPM [33]. In addition to the cerebellum, the rest of the brain and other nervous tissues appeared to be sites of relatively high levels of provirus expression, while tissues related to the circulatory or digestive systems showed lower levels of expression (apart from the spleen).

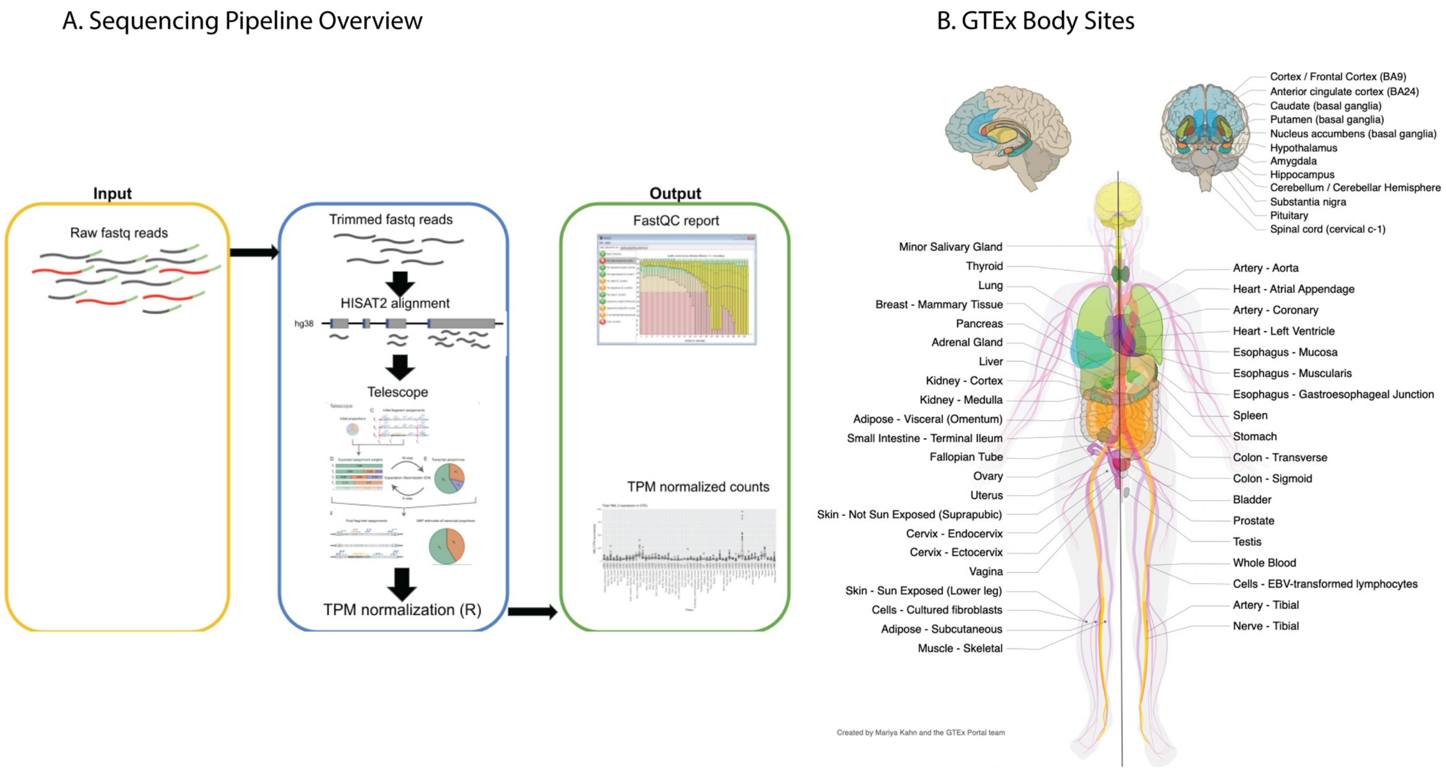


Fig 2. Analysis of the GTEx dataset. (**A**) Schematic overview of the sequencing pipeline used in this study. See Materials and methods for further information. (**B**) Diagram of all body sites samples in the GTEx project, copied from the GTEx portal, with permission. Not listed is 1 cell culture, a chronic myelogenous leukemia line derived from a GTEx donor. EBV, Epstein–Barr virus; GTEx, Genotype Tissue and Expression; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g002

Reproductive tissue such as the testis and ovaries also showed elevated expression compared to other tissues. The variation in HML-2 expression observed among these tissues demonstrates that the intensity of total HML-2 expression is tissue specific with some tissues and organ systems promoting expression more than others. Variation was also observed on an interdonor level. While these figures represent data from all 948 donors, a comparison of 2 individual donors shows another level of variability (S1 Fig). Provirus expression levels, specific provirus allele, and the number of proviruses expressed can change from donor to donor. This variability may result from a number of experimental factors, such as variable collection of tissue samples or could represent biological differences between the donors themselves, such as different health conditions, age, biological sex, or tissue condition at the time of collection. We examined these variables in subsequent analyses.

While the total transcriptional level of HML-2s was of interest, it is only a small part of the picture. Although the HML-2 group comprises a set of closely related proviruses, each provirus is unique and is defined by individual mutational variation along with a unique genomic neighborhood that may influence or be influenced by its transcription. Therefore, exploring HML-2 expression on the individual provirus level is critical for understanding its evolution, functional consequences, and potential therapeutic impact. To properly investigate HML-2 expression in nondiseased tissue, the total HML-2 expression was broken down to the expression of each HML-2 provirus within each tissue with the aid of Telescope, allowing us to use partial expression data to accurately reconstruct the expression level, structure, and genomic location of each provirus [38].

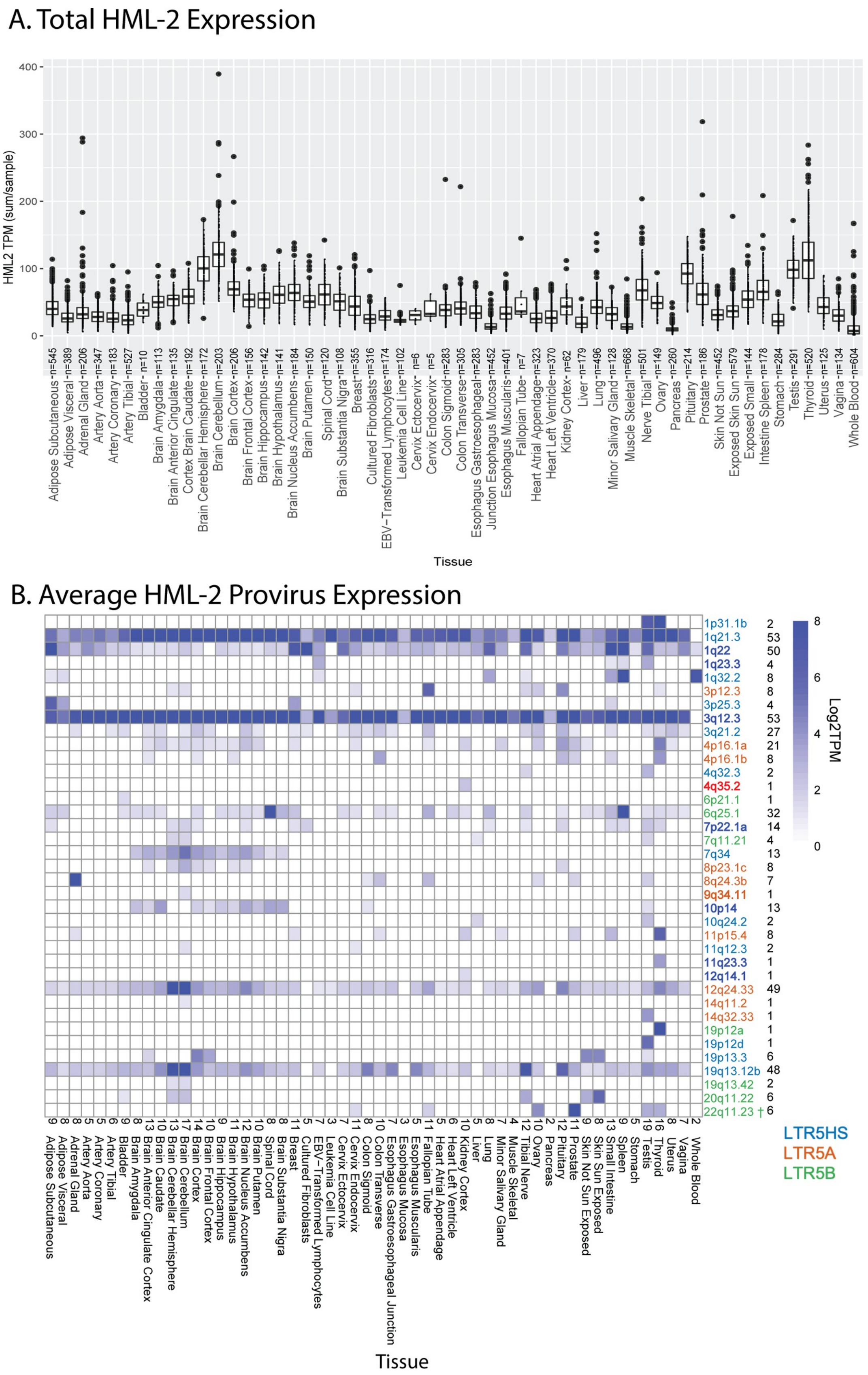

A. Total HML-2 Expression
HML2 TPM (sum/sample)
400
300
200
100
0
Adipose Subcutaneous - n=545
Adipose Visceral - n=389
Adrenal Gland - n=206
Artery Aorta - n=347
Artery Coronary - n=183
Artery Tibial - n=527
Bladder - n=10
Brain Amygdala - n=113
Brain Anterior Cingulate - n=135
Cortex Brain Caudate - n=192
Brain Cerebellar Hemisphere - n=172
Brain Cerebellum - n=203
Brain Cortex - n=206
Brain Frontal Cortex - n=156
Brain Hippocampus - n=142
Brain Hypothalamus - n=141
Brain Nucleus Accumbens - n=184
Brain Putamen - n=150
Spinal Cord - n=120
Brain Substantia Nigra - n=108
Breast - n=355
Cultured Fibroblasts - n=316
EBV−Transformed Lymphocytes - n=174
Leukemia Cell Line - n=102
Cervix Ectocervix - n=6
Cervix Endocervix - n=5
Colon Sigmoid - n=283
Colon Transverse - n=305
Esophagus Gastroesophageal - n=283
Junction Esophagus Mucosa - n=452
Esophagus Muscularis - n=401
Fallopian Tube - n=7
Heart Atrial Appendage - n=323
Heart Left Ventricle - n=370
Kidney Cortex - n=62
Liver - n=179
Lung - n=496
Minor Salivary Gland - n=128
Muscle Skeletal - n=668
Nerve Tibial - n=501
Ovary - n=149
Pancreas - n=260
Pituitary - n=214
Prostate - n=186
Skin Not Sun - n=452
Exposed Skin Sun - n=579
Exposed Small - n=144
Intestine Spleen - n=178
Stomach - n=284
Testis - n=291
Thyroid - n=520
Uterus - n=125
Vagina - n=134
Whole Blood - n=604
Tissue
B. Average HML-2 Provirus Expression
1p31.1b 2
1q21.3 53
1q22 50
1q23.3 4
1q32.2 8
3p12.3 8
3p25.3 4
3q12.3 53
3q21.2 27
4p16.1a 21
4p16.1b 8
4q32.3 2
4q35.2 1
6p21.1 1
6q25.1 32
7p22.1a 14
7q11.21 4
7q34 13
8p23.1c 8
8q24.3b 7
9q34.11 1
10p14 13
10q24.2 2
11p15.4 8
11q12.3 2
11q23.3 1
12q14.1 1
12q24.33 49
14q11.2 1
14q32.33 1
19p12a 1
19p12d 1
19p13.3 6
19q13.12b 48
19q13.42 2
20q11.22 6
22q11.23 † 6
8
6
4
2
0
Log2TPM
LTR5HS
LTR5A
LTR5B
9 Adipose Subcutaneous
8 Adipose Visceral
8 Adrenal Gland
5 Artery Aorta
5 Artery Coronary
6 Artery Tibial
9 Bladder
8 Brain Amygdala
13 Brain Anterior Cingulate Cortex
10 Brain Caudate
13 Brain Cerebellar Hemisphere
17 Brain Cerebellum
14 Brain Cortex
10 Brain Frontal Cortex
9 Brain Hippocampus
11 Brain Hypothalamus
12 Brain Nucleus Accumbens
10 Brain Putamen
8 Spinal Cord
8 Brain Substantia Nigra
11 Breast
5 Cultured Fibroblasts
7 EBV−Transformed Lymphocytes
3 Leukemia Cell Line
7 Cervix Ectocervix
11 Cervix Endocervix
8 Colon Sigmoid
10 Colon Transverse
7 Esophagus Gastroesophageal Junction
3 Esophagus Mucosa
5 Esophagus Muscularis
11 Fallopian Tube
5 Heart Atrial Appendage
6 Heart Left Ventricle
10 Kidney Cortex
5 Liver
8 Lung
7 Minor Salivary Gland
4 Muscle Skeletal
12 Tibial Nerve
10 Ovary
2 Pancreas
12 Pituitary
11 Prostate
9 Skin Not Sun Exposed
8 Skin Sun Exposed
13 Small Intestine
9 Spleen
5 Stomach
19 Testis
16 Thyroid
8 Uterus
7 Vagina
2 Whole Blood
Tissue

Fig 3. HML-2 expression in GTEx. (**A**) Combined expression of all HML-2 proviruses per individual sample per body site. The box and whisker plots show the mean as a line in the middle of a box bounded by the first and third quartile values, with the whiskers extending to 1.5 times the IQR and with outlier values shown as individual dots. All proviruses with an average expression less than 1 TPM are excluded. (**B**) Overall expression in $\log_2$ TPM for HML-2 proviruses expressed at $\geq$1 TPM is displayed for each body site. The specific proviruses, color coded by LTR type as in Fig 2, are shown on the right, followed by the number of tissues where each provirus was expressed. The number of proviruses in each tissue is displayed above the tissues on the horizontal axis. LTR-driven proviruses are in **boldface**. † denotes the provirus 22q11.23, which is driven by a separate upstream 5HS LTR and was therefore not bolded in this figure. TPM counts used to generate each figure can be found in S1 Data with each body site included as an individual sheet. GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; IQR, interquartile range; LTR, long terminal repeat; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g003

We detected 37 HML-2 proviruses expressed at levels above 1 TPM in at least 1 tissue. Individual provirus transcription could be classified into 1 of 3 different patterns of expression (Fig 3B): (1) those that were expressed in almost all tissues sampled; (2) those for which transcripts were only seen in tissues of a specific type; and (3) those that were transcribed in various tissues from multiple types. There were 5 proviruses that demonstrated the first pattern: 1q21.3, 1q22, 3q12.3, 12q24.33 and 19q13.12b. 1q21.3 and 3q12.3 were the highest expressed with levels regularly exceeding 8 TPM. The second pattern included proviruses, such as 10p14, 4p16.1a, and 7q34, which were expressed almost exclusively in the brain samples. The remainder of the expressed proviruses seen in the GTEx dataset, following the third pattern, were only expressed in a small number of tissues and did not appear to be broadly associated with specific tissue types or organ systems. We observed high levels of HML-2 activity in prostate tissue, largely consisting of transcripts from 22q11.23, with lesser contribution from 1q21.3 and 3q12.3. Overall, it is clear that each tissue in the human body has a unique provirus expression pattern, and apart from the 2 proviruses 1q21.3 and 3q12.3, which are expressed in 53/54 tissues, each provirus exhibited a unique pattern of expression.

Mechanisms of HML-2 expression

As described in the introduction, proviruses are canonically expressed using enhancer and promoter signals in the 5′ LTR to initiate transcription at the U3-R border. However, as we reported previously [23,24], other mechanisms, including read-through from nearby or surrounding host genes, can also be used, particularly in cases where the 5′ LTR has been damaged by mutation. Examination of the alignments of RNA-seq reads from the GTEx data using Integrated Genomics Viewer [40] can aid in elucidating such mechanisms. Using this method, the transcriptional mechanism of each provirus was classified (S1 Table). It is expected that, for an LTR-driven provirus, the aligned reads would map to the presumptive transcription start and poly(A) sites at the U3-R and R-U5 borders in the LTRs. Fig 4A, for example, shows the data for the provirus at 3q12.3, one of the highest and most broadly expressed proviruses in the dataset. These aligned read clusters appear to start at the immediate 5′ end of the provirus, rather than the expected U3-R border, but this is likely an artifact resulting from double mapping of reads from the 3′ LTR. Nine HML-2 proviruses expressed in GTEx displayed an LTR-driven mechanism. Reads were clearly observed to align at the TSS without aligned reads preceding the 5′ LTR. Each of these proviruses contain an intact 5′ LTR capable of driving transcription. All but 2 of these proviruses, those at 22q11.23 and 4q35.2, are LTR5HS proviruses. The provirus at 22q11.23 is an LTR5B provirus, yet transcription is driven by an LTR5HS promoter 551 bp upstream. This result was previously observed in the Tera-1 Teratocarcinoma cell line [24] and was seen again in our GTEx data (S2 Fig). 4q35.2 is an LTR5A subtype provirus and the only one apparently expressed from its own LTR.

Unlike these 9 LTR-driven proviruses, there were 24 expressed proviruses that either did not demonstrate clear LTR-driven transcription starting at the 5′ LTR or did not contain a 5′

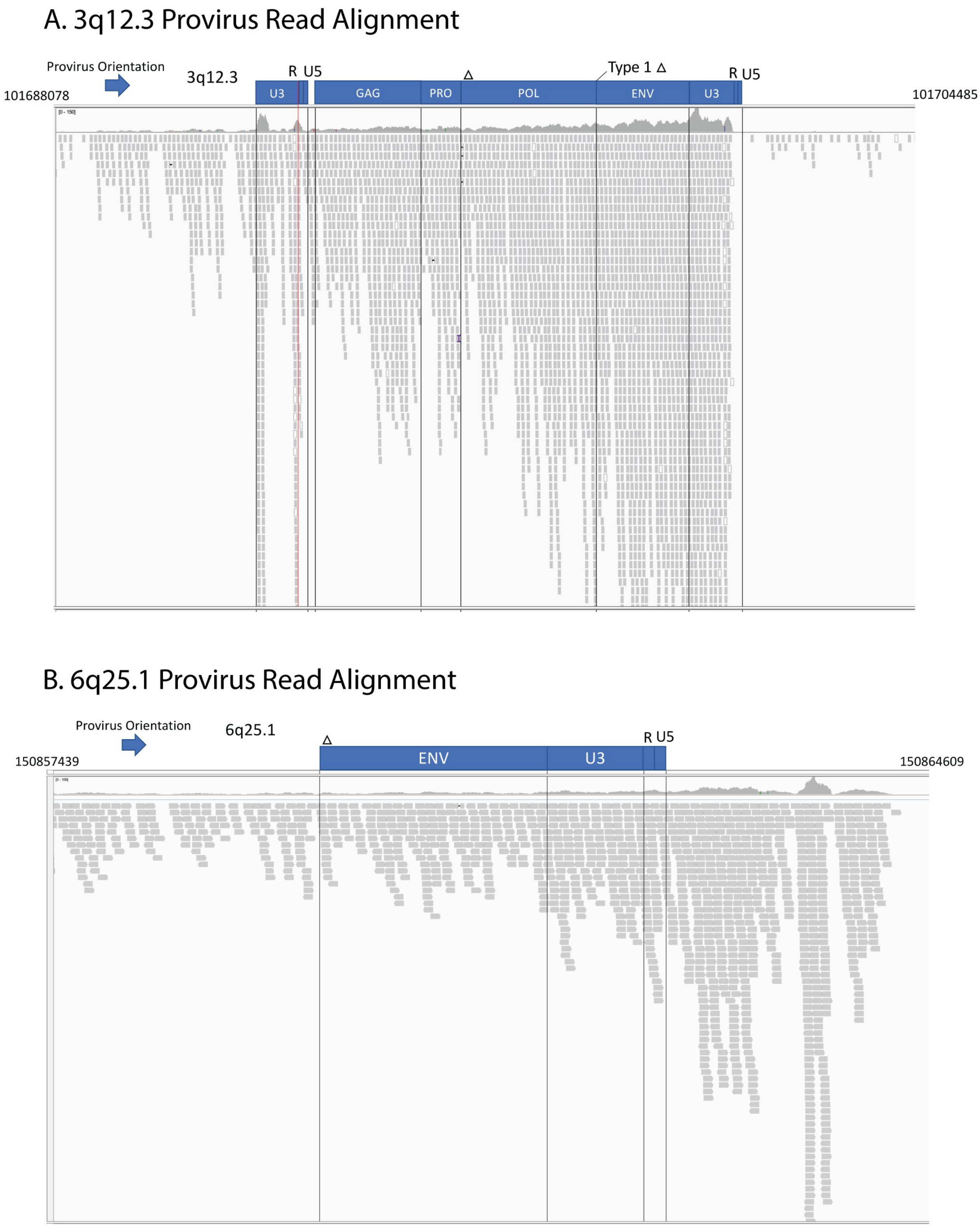


Fig 4. Expression of HML-2 proviruses can be self-driven or a result of the surrounding genomic neighborhood. These Integrated Gene Viewer snapshots show the alignment of GTEX RNA-seq reads to 2 proviruses. (**A**) 3q12.3. This representative alignment is from the cerebellum of 1 GTEx donor. The red line marks the location of the transcription start site. (**B**) Alignment of 6p25.1 from the spleen of another GTEX donor. Vertical lines demarcate divisions between the viral ORFs. Blue boxes denote provirus and gene structure. Genomic coordinates are displayed on the top left and

right corners. The arrow denotes the orientation of each provirus. Bam and Index files of the proviruses used to generate this figure can be found in S2 Data. GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; ORF, open reading frame.

https://doi.org/10.1371/journal.pbio.3001826.g004

LTR at all. The provirus at 6q25.1 provides an example of such read-through transcription (Fig 4B). This provirus completely lacks a 5′ LTR and adjacent *gag-pro-pol sequence*, yet the entirety of the remaining approximately 2,800 bp sequence is expressed. Aligned reads can be seen stretching from upstream of the provirus into the provirus sequence with no changes or characteristic gaps suggesting a new TSS. This even alignment of reads combined with the lack of a required LTR sequence suggests that expression of this provirus sequence is driven by an outside element. Interrogating the source of RNA-seq reads is complicated by technical limitations of the GTEx dataset. Longer read lengths and stranded sequencing would provide a clearer picture of the direction of expression, as well as the clear start and end and splicing pattern of provirus transcripts. Yet it is clear that there are multiple mechanisms driving provirus expression in nondiseased human tissue.

To evaluate the impact of surrounding genes on provirus expression, we turned to WGCNA, which reveals connections between genes that have correlated expression [41]. This tool was used to create gene networks from the RNA-seq dataset of each body site. All expressed proviruses were analyzed alongside all expressed genes within 10 kb of each provirus. We hypothesized that proviruses driven by read-through transcription would be strongly correlated with the surrounding gene that was driving this expression. Only 1 gene–provirus connection was significant, with the provirus 11q12.3 showing a significant connection with the gene ASRGL1 (S3 Fig). This provirus sits inside an intron of ASRGL1, which makes it a likely candidate for read-through transcription. Yet, even though 15 of the 38 expressed proviruses are similarly genic, this was the only provirus whose expression was significantly correlated with that of a surrounding or neighboring gene. This result is likely due to the intronic location of these proviruses. Since no HML-2 proviruses remain in exons, all would therefore commonly be spliced out when the surrounding gene is transcribed and its transcript processed. The variation of expression across the 13,000+ samples could also have affected the ability to find a significant correlation between these proviruses and the host gene.

HML-2 provirus transcription is affected by LTR sequence

The HML-2 proviruses in the human genome comprise 3 subtypes, named for LTR differences, but with a variety of defining mutations across the whole sequence that have accumulated over evolutionary time. For each transcribed provirus, it is important to understand both the drivers and the downstream effects of expression. The first characteristic we used to identify and group the proviruses expressed in the GTEx dataset was LTR subtype, which is a useful proxy for understanding sequence diversity and provirus integration age [8]. The expression patterns associated with the different subtypes (LTR5A, LTR5B, and LTR5HS) are shown by the color coding (red, green, and blue, respectively) in Fig 3A. A neighbor-joining tree was generated using the 5′ and 3′ LTRs, when present, from all expressed HML-2 proviruses in the GTEx dataset (Fig 5A). All 3 LTR subtypes are represented in this group, which includes 18 LTR5HS proviruses, 7 LTR5A proviruses, and 13 LTR5B proviruses, comprising 40%, 33%, and 56% of each 2-LTR HML-2 subtype, respectively. As expected [42], in most cases, LTRs flanking the same provirus are nearest neighbors on the tree, with the branch lengths separating them reflecting the time since integration. The few exceptions to this pattern, such as the proviruses at 3p25.3 and 19q13.12 reflect ancestral rearrangement mediated by recombination between proviruses at different chromosomal locations [43].

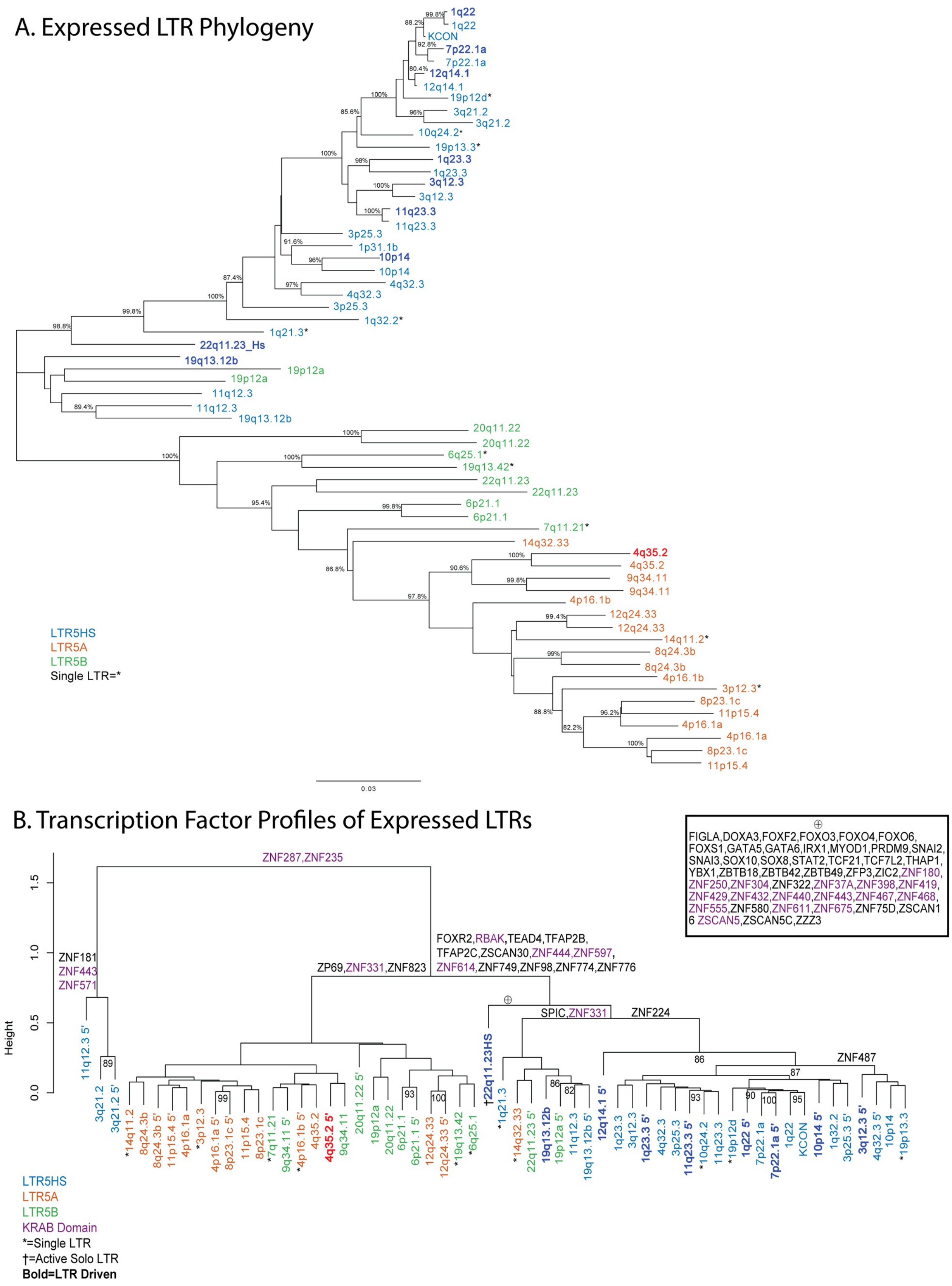

A. Expressed LTR Phylogeny
1q22
1q22
KCON
7p22.1a
7p22.1a
12q14.1
12q14.1
19p12d*
3q21.2
3q21.2
10q24.2*
19p13.3*
1q23.3
1q23.3
3q12.3
3q12.3
11q23.3
11q23.3
3p25.3
1p31.1b
10p14
10p14
4q32.3
4q32.3
3p25.3
1q32.2*
1q21.3*
22q11.23_Hs
19q13.12b
19p12a
19p12a
11q12.3
11q12.3
19q13.12b
20q11.22
20q11.22
6q25.1*
19q13.42*
22q11.23
22q11.23
6p21.1
6p21.1
7q11.21*
14q32.33
4q35.2
4q35.2
9q34.11
9q34.11
4p16.1b
12q24.33
12q24.33
14q11.2*
8q24.3b
8q24.3b
4p16.1b
3p12.3*
8p23.1c
11p15.4
4p16.1a
4p16.1a
8p23.1c
11p15.4
99.8%
88.2%
92.8%
80.4%
100%
85.6%
96%
100%
98%
100%
100%
91.6%
96%
97%
87.4%
100%
99.8%
98.8%
89.4%
100%
100%
100%
95.4%
99.8%
86.8%
100%
90.6%
99.8%
97.8%
99.4%
99%
88.8%
96.2%
82.2%
100%
0.03
LTR5HS
LTR5A
LTR5B
Single LTR=*
B. Transcription Factor Profiles of Expressed LTRs
⊕
FIGLA,DOXA3,FOXF2,FOXO3,FOXO4,FOXO6,
FOXS1,GATA5,GATA6,IRX1,MYOD1,PRDM9,SNAI2,
SNAI3,SOX10,SOX8,STAT2,TCF21,TCF7L2,THAP1,
YBX1,ZBTB18,ZBTB42,ZBTB49,ZFP3,ZIC2,ZNF180,
ZNF250,ZNF304,ZNF322,ZNF37A,ZNF398,ZNF419,
ZNF429,ZNF432,ZNF440,ZNF443,ZNF467,ZNF468,
ZNF555,ZNF580,ZNF611,ZNF675,ZNF75D,ZSCAN1
6 ZSCAN5,ZSCAN5C,ZZZ3
Height
0.0
0.5
1.0
1.5
ZNF287,ZNF235
ZNF181
ZNF443
ZNF571
FOXR2,RBAK,TEAD4,TFAP2B,
TFAP2C,ZSCAN30,ZNF444,ZNF597,
ZNF614,ZNF749,ZNF98,ZNF774,ZNF776
ZP69,ZNF331,ZNF823
⊕
SPIC,ZNF331
ZNF224
86
ZNF487
87
89
99
93
100
86
82
93
90
100
95
11q12.3 5'
3q21.2
3q21.2 5'
*14q11.2
8q24.3b
8q24.3b 5'
11p15.4 5'
4p16.1a
*3p12.3
4p16.1a 5'
8p23.1c 5'
11p15.4
8p23.1c
*7q11.21
9q34.11 5'
*4p16.1b 5'
4q35.2
4q35.2 5'
9q34.11
20q11.22 5'
19p12a
20q11.22
6p21.1
6p21.1 5'
12q24.33
12q24.33 5'
*19q13.42
*6q25.1
†22q11.23HS
*1q21.3
*14q32.33
22q11.23 5'
19q13.12b
19p12a 5'
11q12.3
19q13.12b 5'
12q14.1 5'
1q23.3
3q12.3
1q23.3 5'
4q32.3
3p25.3
11q23.3 5'
*10q24.2
11q23.3
*19p12d
1q22 5'
7p22.1a
7p22.1a 5'
1q22
KCON
10p14 5'
1q32.2
3p25.3 5'
3q12.3 5'
4q32.3 5'
10p14
*19p13.3
LTR5HS
LTR5A
LTR5B
KRAB Domain
*=Single LTR
†=Active Solo LTR
Bold=LTR Driven

Fig 5. Relationships among multiple HML-2 proviruses of different LTR subtype. (**A**) Neighbor-joining tree of the LTRs of each expressed HML-2 provirus in the GTEx dataset. Both LTRs are included when present, if only 1 LTR is present it is labeled with *. Color defines the subtype of each LTR as in Figs 1B and 3B. Darker color indicates proviruses whose expression is driven by the 5′ LTR. In both panels, KCON refers to the LTR of HERV-KCON, the infectious consensus HERV-K (HML-2) sequence [66]. (**B**) Cluster dendrogram of expressed HML-2 provirus LTRs based on TF binding profile as determined by FIMO (See Materials and methods). Solo LTRs are denoted by *. LTRs are colored by LTR5 subtype. A darker color and **boldface** signify an LTR observed to drive provirus expression. Defining TFs are shown above the branches of the dendrogram. Purple TFs are those that are known to have a KRAB domain [51]. The large number of TF motifs that define the branch containing the 22q11.23 LTR5HS solo LTR, which drives the expression of the adjacent LTR5B provirus [24], are shown in the inset under the ⊕ symbol. The sequences of HML-2 LTRs were collected from the HG38 human genomes using the coordinates included in S1 Table [8,10]. The matrix of TF sites is found in S3 Data. GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; LTR, long terminal repeat; TF, transcription factor.

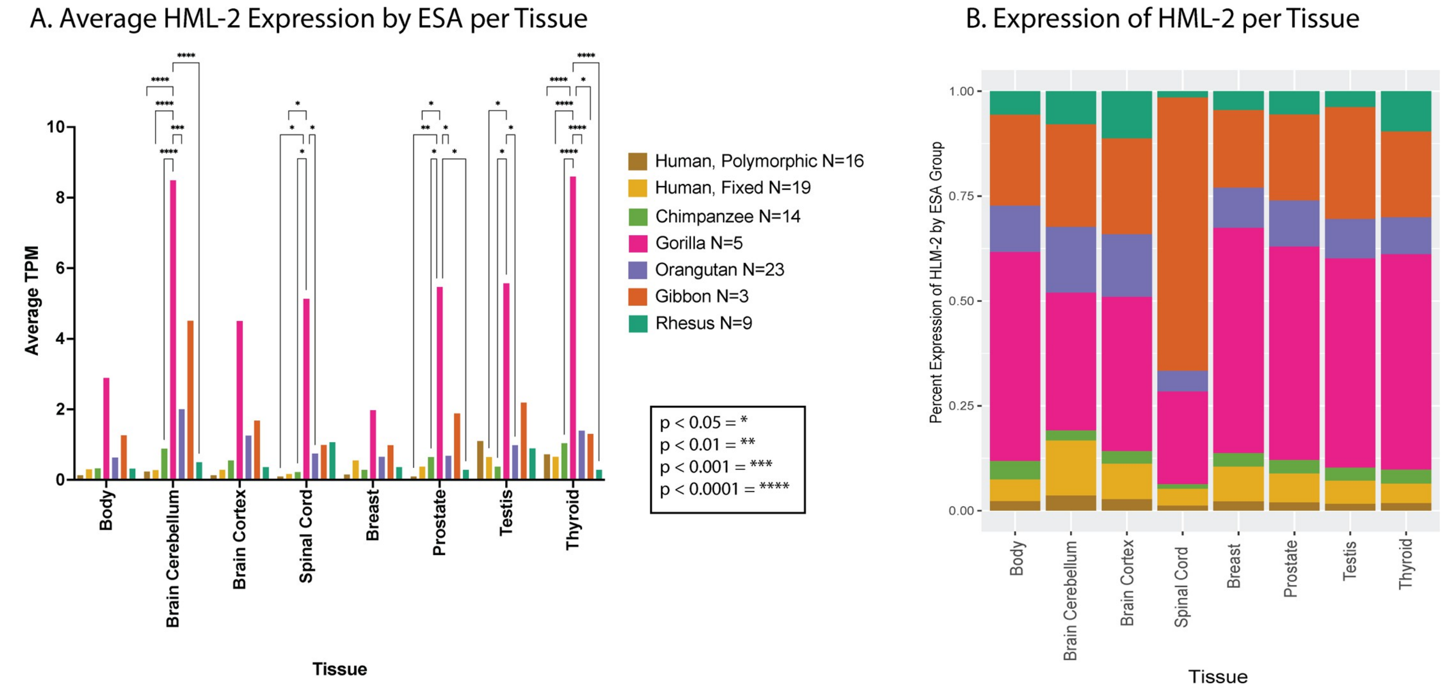
https://doi.org/10.1371/journal.pbio.3001826.g005

As indicated by the greater p distance between their LTRs, LTR5A and 5B proviruses are the older groups, thought to have been active about approximately 15 to 21 million years ago and to have gone extinct prior to the hominid divergence [8]. None of the proviruses in these 2 groups contain an intact ORF for any viral gene (6 contain partial *gag* ORFs and 1 contains a polymorphic *gag* ORF [8].

One way to analyze mechanisms of provirus expression is to investigate the TF binding motifs present in each LTR. The U3 region of the LTR includes a number of elements required for provirus expression including promoter and enhancer regions, which contain binding sites for TFs and other regulatory proteins [44]. The ability of TFs to bind to their cognate motifs in these regions can be modified by sequence variation, and, therefore, this variation can have significant effects on provirus expression in vitro [45]. To test whether the LTR sequence variation affected potential TF binding and was therefore likely to influence the pattern of HML-2 expression seen in the GTEx dataset, an alignment of expressed provirus LTR sequences was analyzed using the bioinformatics software FIMO to scan the DNA sequence of each provirus LTR for confirmed binding motifs of known human TFs [46]. The LTRs were then clustered based on a matrix of TF binding motif presence (Fig 5B). This analysis separated the proviruses into 2 major clades, based on the presence/absence of approximately 16 different TF motifs with LTR5HS proviruses largely separating into 1 clade and most of the LTR5A and LTR5B proviruses combined in the other; the 5′ LTR of 11q12.3 along with both LTRs of 3q21.2 separated from the rest of the proviruses due to the absence of binding sites for ZNF287 and ZNF235, 2 factors with a currently unknown function. Notably, the 3′ LTR of 11q12.3 did contain the 2 ZNF TF binding sites and clustered with other LTR5HS and LTR5B proviruses. The LTR5HS containing clade was defined by the presence of 13 different TF motifs not found in the LTR5A/B clade. A number of these TFs were either involved in development (TEAD4, TFAP2B, TFAP2C) [47–49] or known to be repressive (RBAK, ZNF597, ZNF614) [50,51]. Interestingly, the 22q11.23HS solo LTR that drives the expression of the 22q11.23 provirus contains binding sites for 48 TFs that were not seen in any other LTR analyzed (Fig 5B, inset). This LTR is quite diverged from the other LTR5HS LTRs, sitting basal to nearly every other LTR5HS LTR besides 19q13.12b. The high-level clustering by TF binding so closely replicating the clustering by LTR subtype suggests that the defining mutations of LTR5HS provirus LTRs have (or had) a significant effect on the ability of these proviruses to bind TFs and the consequent tissue specificity of expression. For example, the LTR 5A cluster on the left, from 14q11.2 to 8p21.3c, has 5 proviruses with 5′ LTRs with CNS (brain and cerebellum)-specific expression, and 12 of the proviruses in the 5Hs cluster from 1q23.3 to 11q 23.3 with 5′ LTRs show reproductive tissue–specific expression. Interestingly, LTR5A/B LTRS appear to drive only 1 of the LTR-expressed proviruses while LTR5HS LTRS drive the other 8. This observation suggests that the 3 development-associated TFs present in the LTR5HS branch could play a role in driving this reproductive tissue expression.

HML-2 provirus transcription is biased toward older proviruses

HML-2 proviruses began accumulating in the primate ancestral genome around 35 million years ago and continued to accumulate in human ancestors in waves until less than 500,000 years ago (Fig 1B). This wide range of integration timing has created a diverse population of HML-2 proviruses in the human genome, each of which has been subject to unique selection pressure since integration. To understand the influence of proviral age on expression of each provirus in the GTEx dataset, the proviruses were grouped by the **e**arliest **s**hared **a**ncestor (**ESA**) of humans and modern primate species, creating 7 different age groups of proviruses (Rhesus Macaque, Gibbon, Orangutan, Gorilla, Chimpanzee, Human Specific-Fixed, Human Specific-Polymorphic). The expression of these groups was then broken down in 2 different ways: by the average expression of each group across all tissues tested (Fig 6A) and by the percentage of proviruses in each group that were expressed >1 TPM (Fig 6B). Both methods revealed that, among the Hominoid-specific proviruses, the "older" ones (ESA from Gibbon to



C. Number of HML-2 Proviruses Expressed by ESA

	Rhesus Macaque (25-30)	Gibbon (16-20)	Orangutan (14-18)	Gorilla (8-10)	Chimpanzee (6-7)	Human, Fixed (>0.3)	Human, Polymorphic(<0.3)	Total
Total HML-2	9	3	23	5	14	19	15	88
Brain Cerebellum	1(11%)	2(67%)	6(26%)	1(20%)	4(29%)	2(11%)	1(7%)	17
Brain Cortex	1(11%)	1(33%)	4(17%)	2(40%)	3(21%)	2(11%)	1(7%)	12
Spinal Cord	1(11%)	1(33%)	3(13%)	2(40%)	0(0%)	1(5%)	0(0%)	8
Breast Mammary Tissue	1(11%)	1(33%)	3(13%)	1(20%)	1(7%)	3(16%)	1(7%)	11
Prostate	0(0%)	2(67%)	2(9%)	2(40%)	2(14%)	3(16%)	0(0%)	11
Testis	2(22%)	2(67%)	4(17%)	4(80%)	1(7%)	3(16%)	3(20%)	19
Thyroid	1(11%)	1(33%)	3(13%)	3(60%)	3(21%)	3(16%)	2(13%)	16
Entire Body	4(44%)	3(100%)	11(48%)	5(100%)	4(29%)	6(32%)	4(27%)	37

Fig 6. Expression of HML-2 proviruses in the GTEx dataset as a function of provirus age. (**A**) The plot displays the average TPM of proviruses in the GTEx dataset grouped by the ESA. Proviruses are sorted by identification of orthologous insertions in species related to humans [8], and their average expression is displayed for the body sites of interest. *P* values (when ≤0.05) of the comparisons indicated in brackets above the plots are shown by asterisks as described in the key. (**B**) The average expression of HML-2 proviruses in each ESA group normalized for the number of proviruses in that group is displayed as the percentage of total HML-2 expression at each body site. (**C**) HML-2 expression >1 TPM broken down by ESA group. Numbers in parentheses next to the species names indicate estimated time, in millions of years, to their last common ancestor with humans. This figure was generated using the TPM counts in S1 Data and the ESA group listed in S1 Table. ESA, earliest shared ancestor; GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g006

Gorilla) were expressed to a higher level than the recently integrated ones shared with Chimpanzee or Human specific. Expression from proviruses that are shared with Gorillas makes up the majority of HML-2 transcripts in each tissue, along with proviruses that are shared with Gibbon (Fig 6B). The Gorilla group, while only consisting of 5 proviruses, is expressed at a much higher level than any other in each tissue observed, with significant differences between Gorilla and Human polymorphic proviruses measured in 5 of the 7 body sites highlighted. The Orangutan group proviruses were the largest ESA group [11], expressed to the third highest levels after Gibbon in Fig 6A, and displayed a lower average when normalized to ESA group size in Fig 6B. The "younger" proviruses in the Chimpanzee and Human specific groups were more highly expressed in certain tissues, including thyroid and testis. The level of expression of Human specific-polymorphic proviruses was also relatively higher in the thyroid and testis.

Along with higher expression levels in specific tissues, the percentage of older proviruses widely expressed in most tissues was higher than that of the younger proviruses (Fig 6C). The cerebellum, testis, and thyroid are the only tissues that contained a higher number of expressed Chimpanzee group and Human Specific group proviruses. These tissues appeared to promote expression from more of the younger proviruses as compared to other tissues. These data suggest that proviruses of different ESA groups are under different levels of control in the human genome where older proviruses that may have undergone longer periods of selection are more likely to be expressed. It is therefore possible that certain expressed proviruses might have provided a selective advantage that contributed to their fixation and protected them from being lost by drift or solo LTR formation.

Covariates, including age and sex of donor, only slightly affect HERV-K (HML-2) transcription

The individual variation in expression of distinct proviruses across different tissues implies that not all nondiseased individuals express HML-2 proviruses in the same way. To address possible underlying causes of this variability, we utilized the metadata provided with the GTEx cohort to break down the donors according to different covariates. We hypothesize that interdonor variation may be partially explained by the sex, age, and/or timing of death of the individual donors.

Biological sex can have an important impact on gene expression [52]. We calculated total HML-2 expression in tissues of interest depending on sex as determined by the expression of Y chromosome genes (Fig 7A). At the total HML-2 level, no significant differences can be seen between the 2 groups in any tissue tested, with few exceptions such as breast and cerebellum, which displayed slight, but not statistically significant, differences in total HML-2 expression. Focusing specifically on these tissues, and the individual provirus expression in each of them, can offer more insight. In breast tissue samples, including 218 male and 134 female donors, 6 proviruses showed significantly higher expression in biological females than males (Fig 7D, $p < 0.05$). Thus, the presence or absence of Y chromosome gene expression may not correlate with overall HML-2 expression but can do so in certain body sites in a more specific and targeted way, although we cannot exclude sampling issues, such as distribution of specific cell types, which may affect the apparent differences in expression levels between the sexes.

The GTEx cohort includes donors ranging in age from 20 to 70 years, and, therefore, age-related characteristics could have affected HML-2 expression in donors to this dataset. To examine age-related changes in expression, samples for each tissue were sorted into 3 donor groups: ages 20 to 35, 36 to 51, and 52 to 70. Total HML-2 expression among the age groups was variable in certain body sites, particularly in the nervous system (Fig 7B). The cortex and spinal cord showed altered expression of HML-2 proviruses in the 20 to 35 age group (Cortex,

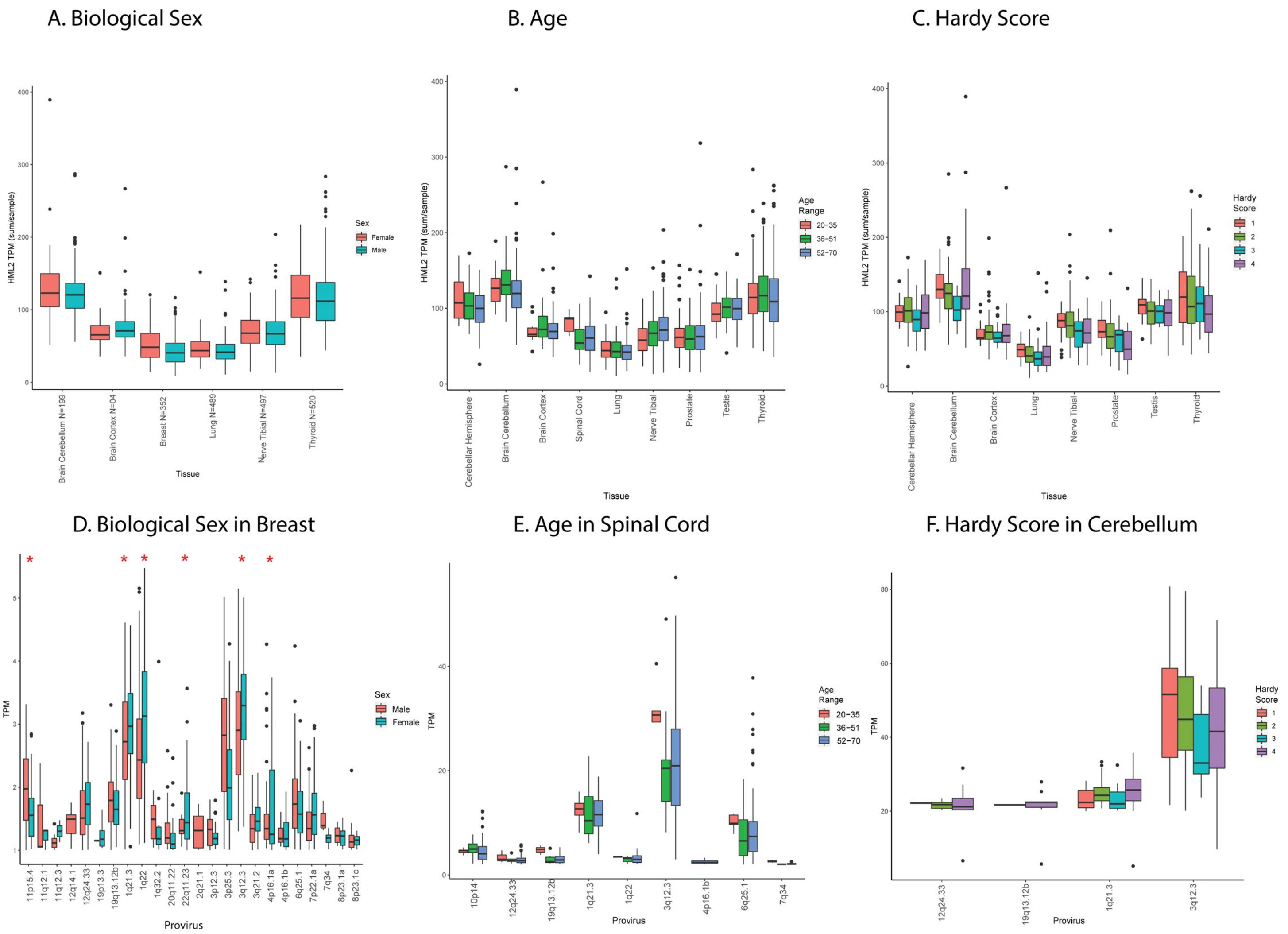


Fig 7. Effects of covariates and morbidities on HML-2 expression. (**A**, **B**, **C**) Average TPM per tissue of interest of total HML-2 expression for donors separated by biological sex, age, and Hardy score plotted as in Fig 3A. Biological sex was identified by Y chromosome expression: Age was divided into 3 groups (20–35, 35–51, and 52–70); Hardy score is broken down into 4 categories. 1 is the fastest death; 4 is the slowest. (**D**, **E**, **F**) Expression of individual proviruses in the indicated body sites (breast, spinal cord, and cerebellum). Average TPM of all proviruses expressed greater than or equal to a TPM of 1 is displayed from all samples of each body site. Red asterisks indicate statistically significant ($p < 0.5$) differences (*t* test). All TPM data are from the relevant sheets of S1 Data. Covariate information for each donor can be found in S1 Data on the SUBJID_Pheno sheet. HML, human mouse mammary tumor virus-like; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g007

N = 10 samples, spinal cord, *N* = 55) as compared to older donors (cortex: 36 to 51 years, *N* = 30; 52 to 70 years, *N* = 164; spinal cord: 36 to 51 years, *N* = 23; 52 to 70 years, *N* = 92), yet no differences were statistically significant. The lack of significance could be in part due to the small number of samples from the 20 to 35 age group (*N* = 5). Alternatively, the cerebellum had lower expression of total HML-2 proviruses in the 52 to 70 age group (*N* = 157) as compared to the younger groups (20 to 35 years, *N* = 12, 52 to 70 years, *N* = 30). Much of this variation was the result of expression of the provirus at 3q12.3, which was elevated in the 20 to 35 age group in the spinal cord and had a large variation in expression in both the cortex and cerebellum (Fig 7E), resulting in the total HML-2 levels previously mentioned. The cerebellum also displayed a lower level of 1q21.3 expression in the 20 to 35 age group as compared to the 52 to 70 age group, a difference that was not observed in other body types. Similar to the effects of biological sex, the age of the donors appears to affect specific proviruses in individual body sites, and the provirus at 3q12.3 was the one whose expression was most affected by changes in age in multiple body sites measured.

Provirus expression could also have been artifactually affected by stress associated with the process of death. The Hardy score is a medical classification assigned by the GTEx consortium that describes how quickly an individual died, allowing one to take into account stress-related effects on gene transcription from postmortem donors [53]. Hardy scores are on a scale from 1 to 4, where 1 is assigned to death from accident or blunt force trauma that lasts less than 10 minutes, 2 is assigned sudden death of a previously healthy individual (e.g., myocardial infarction), 3 is assigned to a death over the course of 1 to 24 hours, and 4 is assigned to a death from a long-term illness. A quick violent death as compared to a longer one can have different effects on gene expression in sites across the body and was therefore an important effect to look at [53]. On the level of total HML-2 expression, the cerebellum showed unique patterns across the 4 Hardy scores (Fig 7C). In the cerebellum, HML-2 expression also decreased as the Hardy score increased until the fourth category where it increased again. This change was the result of increased expression in the fourth category of donors by the proviruses 3q12.3 and 1q21.3, which decreased over the initial 3 categories in the cerebellum (Fig 7F). The change in HML-2 expression as a result of cause of death was more uniform than the other covariates analyzed, with HML-2 expression the highest in the violent death category and decreasing as the length of the terminal phase increased. Yet it appears to rebound in the cerebellum in the fourth category. This observation could suggest retained activity in the brain for the long-term deaths that is not retained in other body sites. Overall, the covariates of biological sex, age, and Hardy score can partially explain the variation in expression of certain proviruses in specific tissues. However, even among donors with similar covariates, the provirus expression can vary dramatically, suggesting that more complex correlates of expression remain to be found.

Potential for HML-2 protein expression

The HERV-K (HML-2) group is unique among HERVs in the number of proviruses that retain intact ORFs for viral genes. Products of these viral genes have been associated with a number of diseases and also could be providing beneficial effects, for example, through viral restriction or immune modulation. While specific increases in some of these viral proteins have been observed in certain disease states, leading to suggestions that they might be useful as diagnostic indicators or therapeutic targets [54], the extent to which they are expressed across a nondiseased human body is largely unknown. In order to answer this question, we examined the sequence of each expressed HML-2 provirus, analyzing each viral gene (i.e., *gag*, *pro*, *pol*, *env*, *rec*, and the product of the type 1 mutant transcript *np9*) for the presence of deletions, stop codons, or frame shifts that would disrupt the ORF. Although it was not possible to evaluate the functionality of each gene, its ability to express a viral antigen could be inferred (Fig 8A). Across all HML-2 proviruses, intact ORFs are maintained for all viral genes. Eleven proviruses contain an intact *gag* gene; 19 others can only encode 1 subunit of *gag*. Eleven proviruses contain an intact *pro*, yet only 6 of these proviruses have an intact *gag* upstream for proper translation. Although only 3 proviruses possess an intact *gag* and *pro* upstream of the pol gene, 9 carry an intact *pol* gene. Additionally, there are 8 proviruses potentially capable of expressing an intact *env* gene.

Our large dataset of HML-2 expression data provided an opportunity to predict the pattern of potentially expressed viral proteins in different tissues. While the RNA-seq read length limits the identification of gene transcripts themselves, we could associate expression of each specific provirus with its intact gene content inferred from its DNA sequence (Fig 8A). These data were used to create the map seen in Fig 8B. This analysis reveals that very few of the expressed

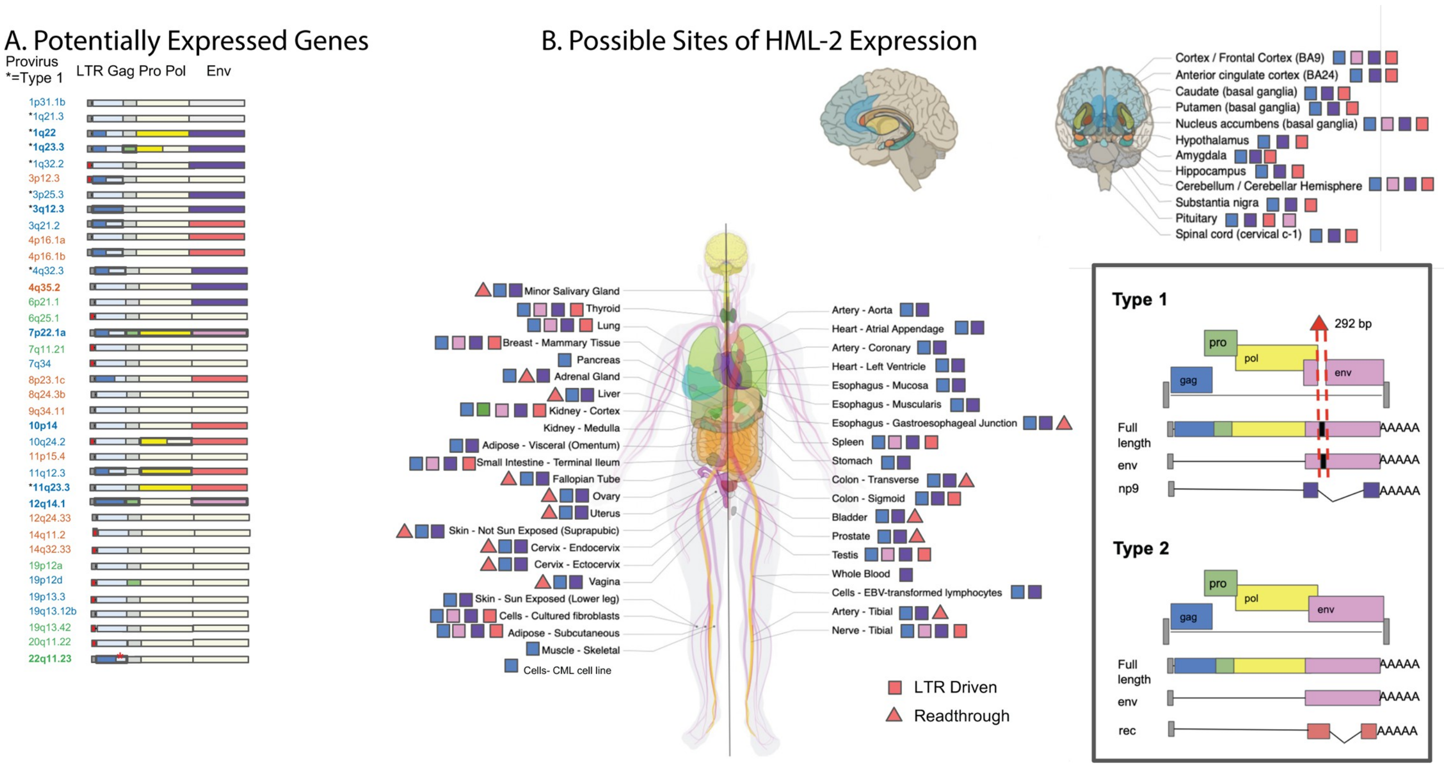


Fig 8. Potential consequences of HML-2 expression in the GTEx dataset. (**A**) Schematic of each provirus expressed at TPM >1 in the GTEx dataset. Each ORF and LTR is designated with a colored box; a filled box represents an intact ORF. A partial LTR is represented by a red box. Colors are explained in the legend. Asterisks next to the provirus locations indicate type 1 proviruses, which are incapable of expressing *env* or *rec*, but rather express *np9*. Bold text represents an LTR-driven provirus. (**B**) Each body site included in the GTEx dataset is represented in this diagram, copied from the GTEx portal, with permission. All nonbrain tissues are labeled on the body; a zoom out of the brain labels each of its tissues individually. A colored symbol next to each body site denotes which intact ORFs could be expressed in that body site based on provirus sequence and transcriptome analysis. A square indicates LTR-expressed genes; a triangle indicates other potentially expressed genes. The box in the lower right displays the colors of each ORF and shows the differences between type 1 and type 2 proviruses. The * marks a polymorphic insertion in the provirus at 22q11.23 that breaks the Gag ORF in 43% of the population and is found in the hg38 reference genome. This figure used expression data from S1 Data along with ORF data from S1 Table and sequence alignments using the coordinates from S1 Table. GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; LTR, long terminal repeat; ORF, open reading frame; TPM, transcripts per million.

https://doi.org/10.1371/journal.pbio.3001826.g008

proviruses contain intact ORFs. There are 2 intact *gag* genes, coming from the proviruses 3q12.3 and 12q14.1. The provirus 3q12.3 is expressed in nearly every GTEx tissue, while 12q14.1 is expressed only in the kidney cortex. None of the 3 known proviruses with intact *gag-pro-pol* sequences are expressed in GTEx tissue samples, although the kidney-expressed provirus at 12q14.1 contains an intact *gag-pro* sequence. Additionally, 4 different expressed proviruses contain individual intact *pro* or *pol* genes without a fully intact Gag polyprotein. Two proviruses, 7p22.1a and 12q14.1, express an intact env ORF, yet 7p22.1a is the only one with a fusogenic *env* gene [11]. This provirus is expressed in 12 different tissues from numerous different organ systems. Despite there being only 2 intact Env ORFs expressed in the GTEx cohort, 6 different proviruses with an intact Rec ORF are expressed. These proviruses are expressed in body sites from multiple organ systems, suggesting that HML-2 Rec expression may be a common occurrence in nondiseased tissue. These data suggest that, even in nondiseased individuals, HML-2 Gag can be present in as many as 53/54 body sites sampled and HML-2 Env could be present in 15/54 body sites. Therefore, the presence of HML-2 transcripts, and possibly proteins, can be viewed as a normal part of the proteome and transcriptome of nondiseased tissue.

Discussion

The role of HERVs in human biology is still largely a mystery. While the activity of recently integrated HML-2 proviruses has been previously studied in a number of disease contexts, their activity in nondiseased tissue has been largely unexplored. In this study, we leveraged the scale of the NIH GTEx project to analyze HML-2 transcription in 54 different body sites from over 948 different donors. We detected HML-2 expression across the nondiseased human body, finding that 37 different proviruses are expressed and that every tissue site shows some level of expression. The pattern of expression is heterogenous: With many proviruses, it is likely affected by LTR sequence, at least partly through the unique set of TF binding motifs contained in each LTR; with others, particularly those lacking 5′ LTR sequences, it is related to expression of nearby or surrounding genes. The considerable heterogeneity of expression of individual proviruses among the donors sampled, for the most part, remains unexplained, although biological sex also appears to significantly affect the expression of some proviruses in specific tissues. The overall HML-2 expression profile is largely made up of transcripts of evolutionarily older proviruses, yet some younger proviruses with intact ORFs are also expressed.

When analyzing the RNA-seq data from GTEx, it was important to use bioinformatics software that correctly aligns and counts the reads coming from HML-2 proviruses. Correct read assignment has long been an issue in the field due to the sequence similarity among many of the proviruses, especially the younger insertions. This similarity can create many multimapping reads that are difficult to assign to an individual provirus [24]. To address this issue, we employed the Telescope software, which was specifically designed to align multimapping reads from retroelements like HERVs. This software takes an alignment file (such as that produced by HISAT2) and uses a Bayesian mixture model to align the multimapping reads to the most likely source based on the proportion of other, confidently aligned, reads to each provirus. The efficacy of Telescope has been compared to 6 other alignment methods and found to be the most precise in read assignment, avoiding issues of false detection while reducing the amount of unused ambiguously aligned reads [38].

The highest total levels of HML-2 expression were found in the cerebellum, testis, thyroid, and pituitary gland. The cerebellum and testis also supported the widest range of provirus expression of any tissue, with 17 and 19 proviruses expressed, respectively. The varied pattern of HML-2 expression across tissues suggests the existence of tissue-specific factors that could drive this expression or certain tissues that exhibit a reduced restriction of provirus expression. In at least some cases, the HML-2 expression observed could be the result of cellular environments that promote widespread gene expression including many HML-2 proviruses. It is known that both cerebellum and testis express large numbers of tissue-enriched genes [13]. In other hotspot tissues like the thyroid and pituitary gland, the proviral expression observed could be a result of HML-2 response to different signaling hormones. HERV-K LTRs are known to contain binding motifs for signaling hormones such as androgen, estrogen, and progesterone, which could have an activating role in these tissues [44]. Of course, the function of these predicted sites also depends on the expression of TFs in the relevant cell type, as well as other epigenetic features, including DNA, histone, and TF modifications, presence of negative TF elements, and more. More studies, such as ChIP-seq and functional analyses, will be required to identify the relevant factors and to confirm the activities predicted by sequence alone. Published data report that the factors SP1, SP3, and YY1 are involved in HML-2 LTR expression [55,56], yet no motif for these factors was identified in this analysis. Additionally, it is unclear from our data what role repressive factors play in regulating HERV expression. Zinc finger proteins are known regulators of HERVs, via their ability to identify and bind motifs in the LTR and the PBS [57–59]. The KRAB-ZFPs have been reported to regulate DNA methylation and histone modification of HERVs through the recruitment of proteins such as KAP1/TRIM28 [60]. A total of 86 such proteins were detected in our FIMO sequence analysis with 26 of them possessing KRAB domains [51]. It is likely that the activity of these proteins helps to regulate HERV expression in nondiseased tissue, and the differential expression of these multitude of factors could result in the pattern of expression we observed.

It has been previously proposed that the expression of HERV sequences in the testis is a result of HERV involvement in development and reproduction [61]. According to our analysis, the LTR5HS subtype binds factors associated with the regulation of early development such as FOXR2 and TFAP2B, suggesting that the expression in reproductive tissues like the testis could indeed have a regulatory role. We consider it more likely, however, that the HML-2 activity in reproductive tissue and the testis, in particular, could be a relic of ancient viral infection [62]. Although alternative mechanisms have been proposed, most evidence supports the interpretation that ERVs, in general, and HML-2, in particular, arose from infection of germ line cells with the corresponding exogenous virus, likely during epi- (or pan-) zootic infection in ancestral species [1]. According to this view, at the time of integration, the LTR of each newly minted provirus must have directed an expression pattern adapted to support a virus lifestyle, including replication in sites that promote transmission from one individual to another, be it sexual, blood contact, or vertical (mother to infant) [62]. With evolutionary time, these patterns can be altered by forces of mutation, selection, and drift acting on the proviruses or the individuals containing them, leading to loss, or inactivation of the proviruses whose expression is deleterious or even neutral and over long periods of evolution to exaptation to a new function. An example of the latter process may be the provirus at 3q12.3, an old LTR5HS type 1 provirus, whose ESA is gorilla, and whose LTR-driven expression is found in nearly all body sites examined (Figs 3A and 4A). The 5′ LTR of this provirus has a 4-bp duplication, which is fixed in the human population, but not present in gorillas or chimpanzees, and which creates a binding site for HOX-PBX family of TFs, likely responsible for the widespread expression [45]. The predominant expression of younger LTR5HS proviruses in reproductive tissue may reflect persistence of the original viral expression pattern, while expression of the older LTR5A/B proviruses in CNS tissue, with no obvious relationship to viral replication or transmission, may reflect their postintegration evolution to completely different functions, which remain to be discovered. Thus, the greater frequency of expression of older than younger proviruses might be due to selective protection of these proviruses from loss due to drift through (for example) solo LTR formation.

While many HML-2 proviruses have been damaged by mutation and deletion, 9 of the expressed proviruses contain an intact 5′ LTR exhibit and LTR-driven expression, placing an increased importance on the binding motifs in their 5′ LTRs. It is therefore interesting that all but one is an LTR5HS LTR. The observation that the regulatory and developmental motifs in the LTR5HS sequence directly regulate transcription of these proviruses and the role of LTR5HS expression in relation to development require further study.

The presence of intact LTR5HS LTRs that are actively driving transcription would seem to imply that the comparatively younger HML-2 proviruses are expressed to a higher degree than the older proviruses. Yet, HML-2 expression across the body is largely made up of older proviruses, mostly expressed via non-LTR mechanisms, to a higher level than younger proviruses. This selective expression is made more interesting when considering which proviruses contain intact viral ORFs. Only 2 of the 11 proviruses with an intact Gag are expressed in nondiseased tissues, as are 2 of the 8 proviruses with an intact Env, along with none of the proviruses with a fully intact Gag-Pro-Pol polypeptide found in the GTEx data. The proviruses with retained intact ORFs are generally younger than the more defective ones, reflecting the lesser accumulation of damaging mutations over time. Therefore, it appears that the bias of expression of older provirus genes is due to the lower concentration of intact ORFs in these insertions. They could be expressed due to adaptive mutations that drive some beneficial activity such as enhancing host gene expression, as proposed by Xiang and colleagues [31]. Alternatively, the expression of these proviruses could be retained merely due to the lack of viral protein expression driving any purifying selection against them.

If the majority of HML-2 proviruses that contain intact ORFs are repressed to prevent the possible deleterious expression of viral proteins, the viral ORFs that are still expressed may well have a higher chance of having a beneficial effect on the host. As previously discussed, there are numerous examples of ERV *gag* and *env* genes being co-opted by their hosts for beneficial purposes, with defense against exogenous viruses being the most common function [16,18,21,63]. Viral proteins could also be co-opted for another mechanistic benefit such as transport of signaling molecules in a Gag-like protein shell [64]. All expressed proviruses observed in GTEx that contain intact ORFs have had to endure hundreds of thousands to millions of years of selection in the human ancestral genome and still retain the ability to be expressed and contain specific intact ORFs. Three of these proviruses (3q12.3,7p22.1a, and 12q14.1) are also LTR driven, which leaves a viral ORF under direct LTR control. The provirus at 3q12.3 is expressed throughout the body at high levels and contains an intact *gag* gene, making it a prime candidate for co-option and worthy of future study. 12q14.1 is a provirus that retains both an intact *gag* and *env* ORF, and yet it is only expressed in the kidney, possibly pointing to a much more targeted function inside that organ. 7p22.1a carries a well-known fully functional *env* gene that has been studied previously in the literature [11]. Together, these ORF-containing proviruses are highly expressed and LTR driven, likely creating translatable viral transcripts in tissue throughout the body (as shown by squares in Fig 8B). By contrast, translation of the predicted ORFs of the non-LTR-expressed proviruses depends on the structure of the individual transcripts (as shown by triangles in Fig 8B). Further analysis of the protein products of each of these proviruses could reveal a role for an HML-2 provirus in normal human biology, as has been previously uncovered for other HERV families [15].

The presence of proviral transcripts and the likely expression of viral proteins in tissues through the nondiseased human body reveal a constant expression of HML-2 proviruses, with potential implications for the clinic. Up-regulated expression of HML-2 is characteristic of some cancers and is being studied as both a phenotypic marker and an immunotherapeutic target for cancer detection and treatment [54]. For example, the provirus at 12q14.1 is expressed uniquely in the kidney cortex (Fig 3B) and encodes an intact, albeit nonfunctional, Env protein [11]. Therefore, clinical studies of antibodies or cytotoxic T cells directed against HML-2 Env epitopes will need to take potential off target effects due to expression in normal tissue into account into consideration in their design. It will also be important to keep in mind that large differences in the nature, distribution, and expression patterns of ERVs from one species to another will make it impossible to test such safety issues in standard animal models.

Our findings suggest that more work will be required to understand HML-2 transcription in disease states. The significant differences in expression among individual proviruses in healthy tissue demonstrate that the measurement of a total family or subtype of HERVs obscures much of the biology taking place with respect to individual proviruses. Additionally, particularly in the case of cancer, the apparent up-regulation of HERVs seen in a disease state may be a reflection of the epigenetic state of the specific progenitor cell type. It is likely that a more fine-grained analysis of normal tissue will reveal considerable additional heterogeneity in HERV expression not readily visible in the bulk tissues, and this possibility needs to be further examined. Nonetheless, it is not unreasonable to think—and is supported by in vitro oncogenesis models (23)—that a specific cancer or disease state would promote a unique pattern of HML-2 transcription, much like the 54 tissues analyzed here and that in certain contexts HML-2 proviruses will be helpful disease markers and, possibly, play a pathogenic role.

While more work will be required to understand the mechanisms driving their expression and the differences in expression between diseased and nondiseased tissue, it will be important to remember that HML-2 transcripts can always be found in the human body. Therefore, using the level of provirus expression or the presence of viral protein as a phenotypic marker will require controlling for this activity of nondiseased expression of 37 HML-2 proviruses across the body. Our characterization of HML-2 expression in the GTEx database should therefore serve as a useful resource for the clinical application of HML-2 moving forward.

Materials and methods

RNA-seq analysis of the GTEx cohort

The entire RNA-seq portion of the GTEx cohort V8 was downloaded from dbGAP and later AnVIL under phs000424.v8.p2 (https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V8_hg38) between Fall 2019 and Spring 2021. The analytical workflow is outlined in Fig 2A. To start, RNA-seq data were downloaded in SRA format from dbGAP, and, subsequently, paired-end fastq format were extracted using the SRA Toolkit [65]. V8 files were later downloaded from AnVIL as aligned bam files and converted to fastq using a python script that calls Picard samtofastq provided by GTEx [33]. Resulting fastqs were passed to FastQC to check initial read quality, and files were passed to Trimmomatic to remove reads that fell below our threshold (phred score <30, length <75 bp, adapter sequence presence). Reads were aligned to hg38 (UCSC version, 4) using HISAT2 [37]. Alignment files were passed to Telescope [38] to generate provirus-corrected counts files. Individual Telescope files are grouped by tissue using the “Telescope_Merge_Counts” script. SRA and AnVIL files had to be merged separately, due to the differences between the usage of SRA ID by dbGAP and SUBJID by AnVIL. Separate R scripts were utilized to address this matter. Raw counts were TPM normalized in R prior to any further analysis. All code is available at (https://github.com/Coffinlab/GTEx_HML2).

Expression plotting

The boxplot of total HML-2 expression was generated using the script “HML2 expression-HML2 per tissue.” This script converts raw counts to TPM and then groups provirus counts per tissue. These counts are then averaged per sample number and plotted in the boxplot. The script “graph heatmap of individual HML2 proviruses” was used to generate the provirus heatmap. This script took the previously generated TPM counts and calculated the average TPM in each tissue for each provirus. These averages were then fed into pheatmap to generate the figure.

Covariate statistical analysis

The statistical analysis completed on HML-2 counts for biological sex, age, and Hardy score were completed in R using Limma and Voom. This analysis was completed using the script “individual provirus covariate plotting.” Each tissue’s raw counts were filtered for low-expressed genes, and a multidimensional scaling plot was generated to check for proper clustering of samples. The raw counts were then fitted to a curve using Voom before limma was used to fit a linear model to the data and calculate contrasts and significance between each covariate group.

LTR phylogeny

2-LTR proviral HML-2 sequences were downloaded from the hg38 human sequence database on the UCSC Genome browser and loaded into BioEdit. This file was aligned to the KCON consensus sequence [66] using the online MUSCLE interface through EMBL-EBI [67,68]. The 5′ LTR: (corresponding to bp 1–968 of the consensus) of each provirus, and 3′ LTR: (bp 8505–9472), or 3′ LTR alone, when a 5′ LTR did not exist, were extracted from the total provirus genomes and realigned using MUSCLE. This alignment of LTR sequences was then loaded into MEGA X, and a neighbor joining tree was constructed with 500 bootstrap replications. The labels and colors of the tree were then edited in Figtree.

Transcription factor binding motif identification

The HML-2 LTR alignment, curated for the phylogeny, was run through the FIMO software [46]. This software searches each sequence for individual matches of provided motifs. We provided FIMO the binding motifs of all human TFs from CIS-BP [69] along with an alignment of each 5′ and 3′ HML-2 LTR. The *p*-value threshold used was 10^{-4}. The R script "TF Motif dendro solo" was then used. This script loads the .tsv file of identified motifs, filters out unused LTRs, and creates a matrix of motifs per LTR. This matrix is clustered using pvclust (ward.D2 method and 1,000 bootstrap value).This clustering is then plotted before labels and colors are added in Adobe Acrobat.

Provirus age analysis

HML-2 proviruses were grouped by the ESA for each provirus by comparing orthologous insertions in related species [8]. These groups were then used to calculate average expression in each tissue for Fig 6A. Each group average was then scaled by the number of proviruses in each group and plotted as a percentage of HML-2 expression in Fig 6B. These data were then used to generate the 2 bar plots shown. The analysis was completed using the script "Provirus Age Expression." The bar graph in Fig 6A was generated using Prism in order to provide statistics. A Tukey's multiple comparison's test was completed between each ESA group for each body site. The totals of expressed proviruses in each ESA group are displayed in Fig 6C.

ORF determination

HML-2 proviruses were aligned to the full-length HERV-K (HML-2) provirus, 19p12b, with MUSCLE. Proviral segments aligned to 19p12b *gag* (nt 1112–3112), *pro* (nt 2938–3918), *pol* (nt 3915–6749), *env* (nt 6451–8550), *rec* (nt 6451-6711/8411-8467), and *np9* (nt 6451-6494/8411-8591) were translated with ExPASY (https://web.expasy.org/translate). Viral ORFs, which lacked nonsense mutations and frameshifting indels, were run through Motif Scan (https://myhits.sib.swiss/cgi-bin/motif_scan). Only ORFs that retained all predicted Pfam domains were determined to be intact. Additionally, *pro* and *pol* genes were only considered intact if they remained translatable in the context of the GagProPol polypeptide, as a deleterious N-terminal mutation should ablate translation of downstream ORFs.

Supporting information

S1 Fig. Comparison of individual donors. This figure displays a heatmap of individual HML-2 expression from 2 female GTEx donors. Provirus expression is given in TPM for each provirus detected (10 in Donor 1 and 8 in Donor 2) for 22 body sites. Each provirus is labeled on the side with which donor it was measured in. This heatmap was made using data from S1

Data. GTEx, Genotype Tissue and Expression; HML, human mouse mammary tumor virus-like; TPM, transcripts per million.
(PDF)

S2 Fig. 22q11.23 IGV screenshot. This figure displays the alignment of RNA-seq reads to the provirus 22q11.23 in a prostate sample visualized in Integrated Genomics Viewer. The sequence of 22q11.23 is defined by the vertical black lines. The LTR5HS LTR is displayed upstream to the left of the black line. The blue bars shown above the image indicate repeat elements as defined by the Repeatmasker track for HG38 downloaded from UCSC. The bam file and index file for this screenshot can be found in S4 Data.
(PDF)

S3 Fig. Scatterplot of ASRGL1 and 11q12.3 in prostate tissue. This figure displays a scatterplot of expression of both 11q12.3 and ASRGL1 expression in 166 prostate samples in the GTEx dataset. Outliers more than 1.5 times beyond the upper and lower quartiles were removed. Smoothing line was added using linear model method in R ggplot2. A 0.95 confidence interval is displayed around the line. This figure was generated using data from S1 Data.
(PDF)

S1 Table. Expressed HML-2 provirus reference. It lists all expressed HML-2 proviruses and necessary information for each provirus including chromosomal coordinates, intact ORFs, and earliest shared ancestor group. These data were assembled from Subramanian and colleagues [8] and Wildschutte and collegues [10], along with our own work on the intact ORFs.
(XLSX)

S1 Data. TPM counts of RNA-seq results. This Excel workbook contains 1 sheet for HML-2 counts for each of the 54 body sites. Each sheet lists the TPM counts for each donor that provided a sample of that body site, with the provirus name found to the left and right of the data. The donor IDs have been deidentified due to GTEx data restrictions. This file also contains the sheet "SUBJID_Pheno," which lists covariate information for each donor, listing age, sex, and Hardy score when available. These data were used for Figs 3A and 3B, 6A–6C, 7A–7F and 8.
(XLSX)

S2 Data. Bam files of example proviruses. This directory includes the bam files for the proviruses 3q12.3 (Fig 4A) and 6q25.1 (Fig 4B) from an example sequencing run in GTEx for the cerebellum and spleen, respectively. Each bam file also has an index file for viewing in IGV. To satisfy size requirements, the bam files were trimmed to the provirus with 1 kb on each side. These are smaller versions of the files used for Fig 4AB.
(RAR)

S3 Data. TF motifs in HML-2 LTRs. This data sheet includes that matrix of TF sites identified by FIMO for each HML-2 LTR analyzed. If a TF motif was found somewhere in the LTR by FIMO, it was counted and included in this matrix. This matrix was used to generate the dendrogram in Fig 5B. HML, human mouse mammary tumor virus-like; LTR, long terminal repeat; TF, transcription factor
(XLSX)

S4 Data. Bam files for 22q11.23HS. This directory includes the bam file and index file for the 22q11.23HS LTR presented in S3 Fig.
(RAR)

Acknowledgments

We thank Rebecca Batorsky for assistance with the bioinformatics analysis and William Johnson for help with the FIMO analysis.

Author Contributions

Conceptualization: Aidan Burn, Farrah Roy, John M. Coffin.

Data curation: Aidan Burn, Michael Freeman.

Formal analysis: Aidan Burn, Michael Freeman, John M. Coffin.

Funding acquisition: John M. Coffin.

Investigation: Aidan Burn, Farrah Roy, Michael Freeman, John M. Coffin.

Methodology: Aidan Burn, Farrah Roy, John M. Coffin.

Project administration: John M. Coffin.

Resources: John M. Coffin.

Software: Aidan Burn, Farrah Roy.

Supervision: John M. Coffin.

Validation: Aidan Burn.

Visualization: Aidan Burn, Michael Freeman, John M. Coffin.

Writing – original draft: Aidan Burn.

Writing – review & editing: Aidan Burn, Farrah Roy, Michael Freeman, John M. Coffin.

References

1. Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet. 2006; 7:149–173. https://doi.org/10.1146/annurev.genom.7.080505.115700 PMID: 16722807
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860–921. https://doi.org/10.1038/35057062 PMID: 11237011
3. Li WH, Gu Z, Wang H, Nekrutenko A. Evolutionary analyses of the human genome. Nature. 2001; 409 (6822):847–849. https://doi.org/10.1038/35057039 PMID: 11237007
4. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in hussmans. Curr Biol. 2001; 11(19):1531–1535.
5. Nelson PN, Carnegie PR, Martin J, Davari Ejtehadi H, Hooley P, Roden D, et al. Demystified. Human endogenous retroviruses. Mol Pathol. 2003; 56(1):11–18. https://doi.org/10.1136/mp.56.1.11 PMID: 12560456
6. Hanke K, Hohn O, Bannert N. HERV-K(HML-2), a seemingly silent subtenant—but still waters run deep. APMIS. 2016; 124(1–2):67–87. https://doi.org/10.1111/apm.12475 PMID: 26818263
7. Larsson E, Kato N, Cohen M. Human endogenous proviruses. Curr Top Microbiol Immunol. 1989; 148:115–132. https://doi.org/10.1007/978-3-642-74700-7_4 PMID: 2684548
8. 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; 8:90. https://doi.org/10.1186/1742-4690-8-90 PMID: 22067224
9. Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, et al. Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. Genomics. 2003; 81(2):149–156. https://doi.org/10.1016/s0888-7543(02)00027-7 PMID: 12620392

10. 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 U S A. 2016; 113 (16):E2326–E2334. https://doi.org/10.1073/pnas.1602336113 PMID: 27001843

11. Dewannieux M, Blaise S, Heidmann T. Identification of a functional envelope protein from the HERV-K family of human endogenous retroviruses. J Virol. 2005; 79(24):15573–15577. https://doi.org/10.1128/JVI.79.24.15573-15577.2005 PMID: 16306628

12. Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, et al. Integration target site selection by a resurrected human endogenous retrovirus. Genes Dev. 2009; 23(5):633–642. https://doi.org/10.1101/gad.1762309 PMID: 19270161

13. van de Lagemaat LN, Medstrand P, Mager DL. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol. 2006; 7(9):R86. https://doi.org/10.1186/gb-2006-7-9-r86 PMID: 17005047

14. Dupressoir A, Lavialle C, Heidmann T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta. 2012; 33(9):663–671. https://doi.org/10.1016/j.placenta.2012.05.005 PMID: 22695103

15. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, et al. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci. 2013; 368(1626):20120507. https://doi.org/10.1098/rstb.2012.0507 PMID: 23938756

16. Best S, Le Tissier P, Towers G, Stoye JP. Positional cloning of the mouse retrovirus restriction gene Fv1. Nature. 1996; 382(6594):826–829. https://doi.org/10.1038/382826a0 PMID: 8752279

17. Yan Y, Buckler-White A, Wollenberg K, Kozak CA. Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. Proc Natl Acad Sci U S A 2009; 106(9):3259–3263. https://doi.org/10.1073/pnas.0900181106 PMID: 19221034

18. Murcia PR, Arnaud F, Palmarini M. The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of Jaagsiekte sheep retrovirus Gag. J Virol. 2007; 81(4):1762–1772. https://doi.org/10.1128/JVI.01859-06 PMID: 17135320

19. Ito J, Baba T, Kawasaki J, Nishigaki K. Ancestral Mutations Acquired in Refrex-1, a Restriction Factor against Feline Retroviruses, during its Cooption and Domestication. J Virol. 2016; 90(3):1470–1485. https://doi.org/10.1128/JVI.01904-15 PMID: 26581999

20. Ito J, Watanabe S, Hiratsuka T, Kuse K, Odahara Y, Ochi H, et al. Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. J Virol. 2013; 87(22):12029–12040. https://doi.org/10.1128/JVI.01267-13 PMID: 23966402

21. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Elife. 2017;6. https://doi.org/10.7554/eLife.22519 PMID: 28397686

22. Löwer R, Löwer J, Frank H, Harzmann R, Kurth R. Human teratocarcinomas cultured in vitro produce unique retrovirus-like viruses. J Gen Virol. 1984; 65(Pt 5):887–898. https://doi.org/10.1099/0022-1317-65-5-887 PMID: 6202829

23. Montesion M, Bhardwaj N, Williams ZH, Kuperwasser C, Coffin JM. Mechanisms of HERV-K (HML-2) Transcription during Human Mammary Epithelial Cell Transformation. J Virol. 2018; 92(1). https://doi.org/10.1128/JVI.01258-17 PMID: 29046454

24. Bhardwaj N, Montesion 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):939–968. https://doi.org/10.3390/v7030939 PMID: 25746218

25. Ruprecht K, Mayer J, Sauter M, Roemer K, Mueller-Lantzsch N. Endogenous retroviruses and cancer. Cell Mol Life Sci. 2008; 65(21):3366–3382. https://doi.org/10.1007/s00018-008-8496-1 PMID: 18818873

26. 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; 3:246. https://doi.org/10.3389/fonc.2013.00246 PMID: 24066280

27. Bhardwaj N, Coffin JM. Endogenous retroviruses and human cancer: is there anything to the rumors? Cell Host Microbe. 2014; 15(3):255–259. https://doi.org/10.1016/j.chom.2014.02.013 PMID: 24629332

28. Huang G, Li Z, Wan X, Wang Y, Dong J. Human endogenous retroviral K element encodes fusogenic activity in melanoma cells. J Carcinog. 2013; 12:5. https://doi.org/10.4103/1477-3163.109032 PMID: 23599687

29. Jern P, Coffin JM. Effects of retroviruses on host genome function. Annu Rev Genet. 2008; 42:709–732. https://doi.org/10.1146/annurev.genet.42.110807.091501 PMID: 18694346

30. Feuchter A, Mager D. Functional heterogeneity of a large family of human LTR-like promoters and enhancers. Nucleic Acids Res. 1990; 18(5):1261–1270. https://doi.org/10.1093/nar/18.5.1261 PMID: 1690875

31. Xiang X, Tao Y, DiRusso J, Hsu FM, Zhang J, Xue Z, et al. Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. Nat Commun. 2022; 13(1):463. https://doi.org/10.1038/s41467-022-28105-1 PMID: 35075135

32. Schmitt K, Heyne K, Roemer K, Meese E, Mayer J. HERV-K(HML-2) rec and np9 transcripts not restricted to disease but present in many normal human tissues. Mob DNA. 2015; 6:4. https://doi.org/10.1186/s13100-015-0035-7 PMID: 25750667

33. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020; 369(6509):1318–1330. https://doi.org/10.1126/science.aaz1776 PMID: 32913098

34. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure JP, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018; 10(470). https://doi.org/10.1126/scitranslmed.aau5516 PMID: 30518613

35. Tavakolian S, Goudarzi H, Moridi A, Faghihloo E. Analysing the HERV-K env, np9, rec and gag expression in cervical tissues. New Microbes New Infect. 2021; 44:100936. https://doi.org/10.1016/j.nmni.2021.100936 PMID: 34621524

36. Gimenez-Orenga K, Oltra E. Human Endogenous Retrovirus as Therapeutic Targets in Neurologic Disease. Pharmaceuticals (Basel). 2021; 14(6). https://doi.org/10.3390/ph14060495 PMID: 34073730

37. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019; 37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807

38. Bendall ML, de Mulder M, Iniguez LP, Lecanda-Sanchez A, Perez-Losada M, Ostrowski MA, et al. Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. PLoS Comput Biol. 2019; 15(9):e1006453. https://doi.org/10.1371/journal.pcbi.1006453 PMID: 31568525

39. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401

40. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14(2):178–192. https://doi.org/10.1093/bib/bbs017 PMID: 22517427

41. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005; 4:Article17. https://doi.org/10.2202/1544-6115.1128 PMID: 16646834

42. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A. 1999; 96(18):10254–10260. https://doi.org/10.1073/pnas.96.18.10254 PMID: 10468595

43. Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet. 2001; 29(4):487–489. https://doi.org/10.1038/ng775 PMID: 11704760

44. Manghera M, Douville RN. Endogenous retrovirus-K promoter: a landing strip for inflammatory transcription factors? Retrovirology. 2013; 10:16. https://doi.org/10.1186/1742-4690-10-16 PMID: 23394165

45. Montesion M, Williams ZH, Subramanian RP, Kuperwasser C, Coffin JM. Promoter expression of HERV-K (HML-2) provirus-derived sequences is related to LTR sequence variation and polymorphic transcription factor binding sites. Retrovirology. 2018; 15(1):57. https://doi.org/10.1186/s12977-018-0441-2 PMID: 30126415

46. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27(7):1017–1018. https://doi.org/10.1093/bioinformatics/btr064 PMID: 21330290

47. Zhao B, Ye X, Yu J, Li L, Li W, Li S, et al. TEAD mediates YAP-dependent gene induction and growth control. Genes Dev. 2008; 22(14):1962–1971. https://doi.org/10.1101/gad.1664408 PMID: 18579750

48. Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR, et al. Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. Nat Genet. 2000; 25(1):42–46. https://doi.org/10.1038/75578 PMID: 10802654

49. Pastor WA, Liu W, Chen D, Ho J, Kim R, Hunt TJ, et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. Nat Cell Biol. 2018; 20(5):553–564. https://doi.org/10.1038/s41556-018-0089-0 PMID: 29695788

50. Skapek SX, Jansen D, Wei TF, McDermott T, Huang W, Olson EN, et al. Cloning and characterization of a novel Kruppel-associated box family transcriptional repressor that interacts with the retinoblastoma gene product. RB J Biol Chem. 2000; 275(10):7212–7223.

51. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, et al. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 2006; 16(5):669–677. https://doi.org/10.1101/gr.4842106 PMID: 16606702

52. Lopes-Ramos CM, Chen CY, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. Cell Rep. 2020; 31(12):107795. https://doi.org/10.1016/j.celrep.2020.107795 PMID: 32579922

53. Ferreira PG, Munoz-Aguirre M, Reverter F, Sa Godinho CP, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nat Commun. 2018; 9(1):490. https://doi.org/10.1038/s41467-017-02772-x PMID: 29440659

54. Curty G, Marston JL, de Mulder RM, Leal FE, Nixon DF, Soares MA. Human Endogenous Retrovirus K in Cancer: A Potential Biomarker and Immunotherapeutic Target. Viruses. 2020; 12(7). https://doi.org/10.3390/v12070726 PMID: 32640516

55. Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R. Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. J Virol. 2011; 85(7):3436–3448. https://doi.org/10.1128/JVI.02539-10 PMID: 21248046

56. Knossl M, Lower R, Lower J. Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. J Virol. 1999; 73(2):1254–1261. https://doi.org/10.1128/JVI.73.2.1254-1261.1999 PMID: 9882329

57. Yang B, Fang L, Gao Q, Xu C, Xu J, Chen ZX, et al. Species-specific KRAB-ZFPs function as repressors of retroviruses by targeting PBS regions. Proc Natl Acad Sci U S A. 2022; 119(11):e2119415119. https://doi.org/10.1073/pnas.2119415119 PMID: 35259018

58. Iouranova A, Grun D, Rossy T, Duc J, Coudray A, Imbeault M, et al. KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. Mob DNA. 2022; 13(1):4. https://doi.org/10.1186/s13100-021-00260-0 PMID: 35042549

59. Wolf G, Yang P, Fuchtbauer AC, Fuchtbauer EM, Silva AM, Park C, et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes Dev. 2015; 29 (5):538–554. https://doi.org/10.1101/gad.252767.114 PMID: 25737282

60. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature. 2010; 463(7278):237–240. https://doi.org/10.1038/nature08674 PMID: 20075919

61. Larsson E, Andersson AC, Nilsson BO. Expression of an endogenous retrovirus (ERV3 HERV-R) in human reproductive and embryonic tissues—evidence for a function for envelope gene products. Ups J Med Sci. 1994; 99(2):113–120. https://doi.org/10.3109/03009739409179354 PMID: 7716822

62. Diehl WE, Patel N, Halm K, Johnson WE. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. Elife. 2016; 5:e12704. https://doi.org/10.7554/eLife.12704 PMID: 26952212

63. Robinson HL, Astrin SM, Senior AM, Salazar FH. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. J Virol. 1981; 40(3):745–751. https://doi.org/10.1128/JVI.40.3.745-751.1981 PMID: 6275116

64. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell. 2018; 172(1–2):275–88 e18.

65. Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011; 39(Database issue):D19–D21.

66. Lee YN, Bieniasz PD. Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog. 2007; 3(1):e10. https://doi.org/10.1371/journal.ppat.0030010 PMID: 17257061

67. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

68. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019; 47(W1):W636–W641. https://doi.org/10.1093/nar/gkz268 PMID: 30976793

69. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158(6):1431–1443. https://doi.org/10.1016/j.cell.2014.08.009 PMID: 25215497

70. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020; 6:e251. https://doi.org/10.7717/peerj-cs.251 PMID: 33816903