Thomas et al. Mobile DNA (2018) 9:36
https://doi.org/10.1186/s13100-018-0142-3

Mobile DNA

RESEARCH Open Access

Variation in proviral content among human genomes mediated by LTR recombination

Jainy Thomas[1*], Hervé Perron[2,3] and Cédric Feschotte[4*]

Abstract

Background: Human endogenous retroviruses (HERVs) occupy a substantial fraction of the genome and impact cellular function with both beneficial and deleterious consequences. The vast majority of HERV sequences descend from ancient retroviral families no longer capable of infection or genomic propagation. In fact, most are no longer represented by full-length proviruses but by solitary long terminal repeats (solo LTRs) that arose via non-allelic recombination events between the two LTRs of a proviral insertion. Because LTR-LTR recombination events may occur long after proviral insertion but are challenging to detect in resequencing data, we hypothesize that this mechanism is a source of genomic variation in the human population that remains vastly underestimated.

Results: We developed a computational pipeline specifically designed to capture dimorphic proviral/solo HERV allelic variants from short-read genome sequencing data. When applied to 279 individuals sequenced as part of the Simons Genome Diversity Project, the pipeline retrieves most of the dimorphic loci previously reported for the HERV-K(HML2) subfamily as well as dozens of additional candidates, including members of the HERV-H and HERV-W families previously involved in human development and disease. We experimentally validate several of these newly discovered dimorphisms, including the first reported instance of an unfixed HERV-W provirus and an HERV-H locus driving a transcript (*ESRG*) implicated in the maintenance of embryonic stem cell pluripotency.

Conclusions: Our findings indicate that human proviral content exhibit more extensive interindividual variation than previously recognized, which has important bearings for deciphering the contribution of HERVs to human physiology and disease. Because LTR retroelements and LTR recombination are ubiquitous in eukaryotes, our computational pipeline should facilitate the mapping of this type of genomic variation for a wide range of organisms.

Keywords: Endogenous retrovirus, HERV-H, HERV-W, HERV-K, Transposable elements, Long terminal repeats, Provirus, Solo LTR

Background

Endogenous retroviruses (ERVs) derive from exogenous retroviruses that inserted in the germline of their host and thereby became vertically inheritable. Full-length (proviral) ERV insertions are comprised of two long terminal repeats (LTRs) flanking an internal region encoding the protein-coding genes necessary for retroviral replication and propagation, including *gag* (group antigens); *pol* (polymerase) and *env* (envelope) [1, 2]. ERV sequences are abundant in mammalian genomes, occupying approximately 5 to 10% of the genetic material [3, 4], but virtually each species is unique for its ERV content [5, 6]. Indeed, while a fraction of ERVs descend from ancient infections that occurred prior to the emergence of placental mammals, most are derived from independent waves of invasion from diverse viral progenitors that succeeded throughout mammalian evolution [7–10]. Thus, ERVs represent an important source of genomic variation across and within species, including humans. The accumulation of ERV sequences in mammalian genomes has also provided an abundant raw material, both coding and regulatory, occasionally co-opted to foster the emergence of new cellular functions [2, 11–13].

A considerable amount of work has been invested in investigating the pathogenic impact of ERVs. ERVs are

* Correspondence: jainyt@genetics.utah.edu; cf458@cornell.edu
[1]Department of Human Genetics, University of Utah School of Medicine, 15 North 2030 East, Rm 5100, Salt Lake City, UT 84112, USA
[4]Department of Molecular Biology and Genetics, Cornell University, 107 Biotechnology Building, Ithaca, NY 14853, USA
Full list of author information is available at the end of the article

prominent insertional mutagens in some species, such as in the mouse where many de novo ERV insertions disrupting gene functions have been identified, including tumorigenic insertions [1, 14–16]. In contrast, there remains no direct evidence for de novo ERV insertions in humans, although low-frequency insertions have been reported which may conceivably represent very recent insertions [17]. Nonetheless, overexpression of certain human ERV (HERV) families has been associated with a number of disease states, including a variety of cancers, autoimmune, and neurological diseases [18–23] and there is growing evidence that elevated levels of HERV-derived products, either RNA or proteins, can have pathogenic effects [24, 25]. However, the genomic mechanisms underlying the differential expression of ERV products in diseased individuals remain obscure. Copy number variation represents a potent mechanism to create inter-individual differences in HERV expression [26], but the extent by which HERV genes vary in copy number across humans and how this variation relates to disease susceptibility remains understudied.

Copy number variation in ERV genes may occur through two primary mechanisms: (i) insertion polymorphisms whereby one allele corresponds to the full provirus while the ancestral allele is completely devoid of the element; (ii) ectopic homologous recombination between the LTRs of the provirus, which results in the deletion of the internal coding sequence, leaving behind a solitary (or solo) LTR [2, 27] (Fig. 1a–c). Thus, one can distinguish three allelic states for ERV insertions: empty, proviral, and solo LTR [17, 28]. The process of LTR-LTR recombination has been remarkably efficient in evolution since ~ 90% of all human ERV (HERV) insertions are currently represented by solo LTRs in the reference genome [29]. In theory, the formation of solo LTR from a provirus may occur long after the initial proviral insertion as long as there is sufficient sequence similarity between the two LTRs to promote their recombination. The consequences of this recombination process for the host organism may be significant: not only it removes the entire coding potential of a provirus, but it may also alter the cis-regulatory or transcriptional activity of the LTR [30–35].

Among the diverse assemblage of HERV families in our genome, a single subfamily known as HERV-K(HML2) has been reported to exhibit insertional polymorphism in humans [17, 28, 29, 36–47]. Thus far, approximately 50

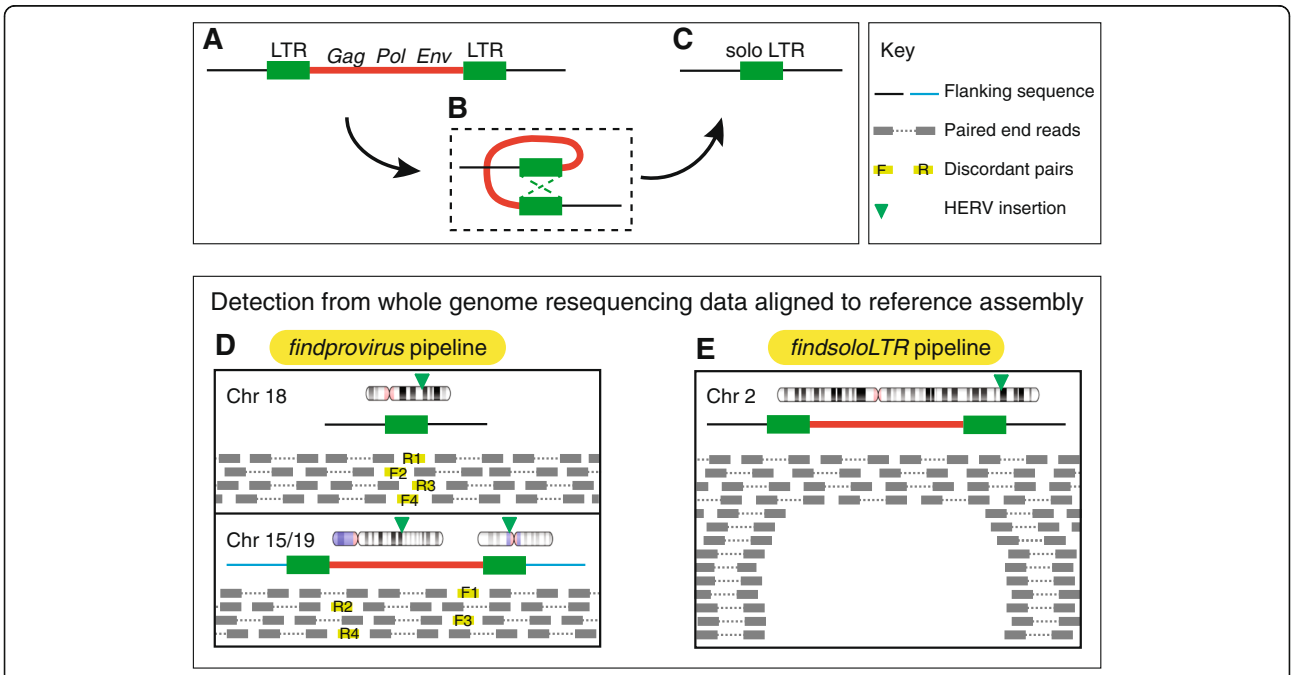

Fig. 1 Structure of a provirus and generation of a solo LTR and their detection from whole genome sequence data. Structure of a typical provirus (**a**) with its internal region (red line) encoding *gag*, *pol* and *env* genes flanked by two long terminal repeats (LTR). Ectopic recombination occurs between the two LTRs of the provirus (**b**) leading to the deletion of the internal region along with one LTR, resulting in the formation of a solo LTR (**c**). Note how the 5′ and 3′ junction sequences between the element and the flanking host DNA (black line), including the target site duplication (not shown), remain the same after recombination. Presence of provirus is identified from whole genome resequencing data aligned to the reference assembly when the reference allele is a solo LTR using the *findprovirus* pipeline (**d**). The *findprovirus* pipeline infer the presence of provirus from the mates of discordant reads with significant homology to the internal region of the respective HERV family. The discordant reads are colored light green and the forward and reverse reads originated from the same fragment are matched by numbers (e.g. F1 and R1). The *findsoloLTR* pipeline identifies the presence of solo LTR when the reference allele is provirus (**e**). It infers the presence of solo LTR based on the deviation of read depth across the provirus and across the flank

HERV-K(HML2) proviral loci are known to occur as empty (pre-integration) and/or solo LTR alleles segregating in the human population [17, 43, 45, 46], but more may be expected to segregate at low frequency [39, 48]. These observations are consistent with the notion that HERV-K(HML2) is the most recently active HERV subfamily in the human genome [49–53]. To our knowledge, there has been only a single report of another HERV family exhibiting a dimorphic locus: an HERV-H element on chromosome 1 (1q25.3_H3) was shown to exist as proviral and solo LTR alleles in two related individuals [27]. Because LTR recombination may in principle take place long after a proviral insertion has reached fixation [54] and possibly recur in multiple individuals, we hypothesized that many more proviral-to-solo HERV variants occur in the human population. We also surmised that this type of dimorphic variants could easily escape detection with current computational pipelines. Indeed, these tools are, by design, geared toward the identification of structural breakpoints distinguishing empty and insertion alleles [17, 55–57]. By contrast, proviral and solo LTR allelic variants share the same exact junctions with flanking host DNA, thus making them recalcitrant to detection with tools tailored to map insertional polymorphisms.

Here we introduce a novel computational pipeline specifically geared toward the identification of proviral deletion resulting from LTR recombination events. We apply the pipeline to the analysis of genome sequences from 279 individuals from worldwide populations generated as part of the Simons Genome Diversity Project (SGDP) [58]. Our approach identifies most dimorphic HERV-K(HML2) loci previously recognized in other population datasets as well as multiple candidate dimorphic HERV-H and HERV-W loci, several of which we validate experimentally. Our results suggest that LTR recombination is an underappreciated source of structural variation in human genomes generating potentially physiologically significant differences in proviral gene copy numbers between individuals.

Results

Strategy for identification of proviral allele when the reference allele is a solo LTR

We developed a pipeline called *findprovirus* to mine whole genome resequencing data to detect a proviral allele of a locus annotated as a solo LTR in the reference genome (Figs. 1d and 2). The prediction is that a fraction of the read mates to the reads mapping to the annotated solo LTR should be derived from internal sequences of the provirus allele. When mapped to the reference

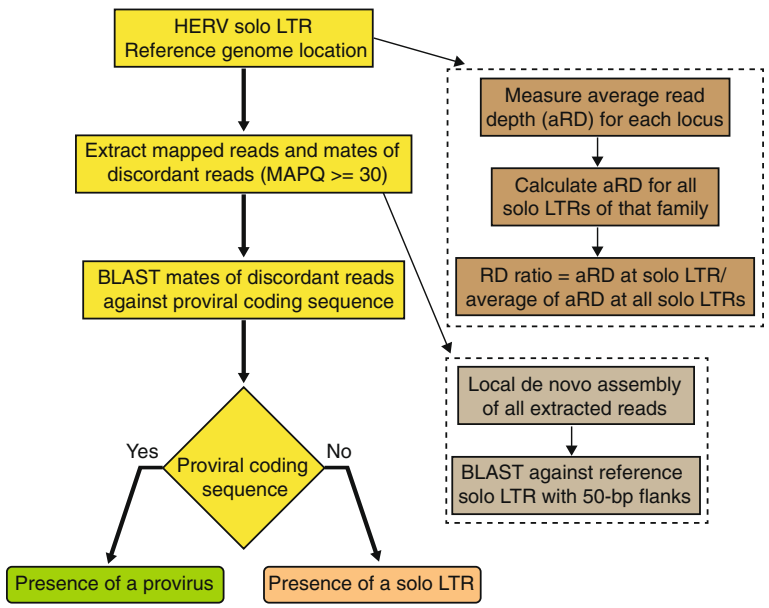

Fig. 2 Flowchart of *findprovirus* pipeline. The first step indexes the coordinates of solo LTRs of a HERV family in the reference genome. Mapped reads (of mapping quality score (MAPQ) equal or greater than 30) and mates of discordant reads are extracted in a window extending ±100-bp from each LTR. Homology based searches are performed with mates of discordant reads against the respective consensus of internal sequence of HERV to infer the presence of a provirus allele at the locus. The read depth for each locus is calculated and compared to the average of read depths for all solo LTRs of that family in an individual. Increased read depth may be observed for some candidate loci reflecting the presence of a provirus allele. A local de novo assembly of the reads is also performed to infer the presence or absence of a solo LTR allele at the locus. These two additional approaches (enclosed by dashed lines) are performed by the pipeline but are not primarily used to infer the presence of a provirus

genome, these events should be identified as discordant read mates mapping elsewhere in the reference genome as they may frequently map to the internal region of non-allelic proviral copies. The pipeline extracts reads mapped to the solo LTR and mates of discordant reads to conduct homology-based searches using the discordant read mates as queries against the consensus sequence of the internal region of the respective provirus as defined in the Repbase database [59] (see also Methods). Presence of at least four reads with significant homology to the internal sequence indicates the presence of a potential allele containing a provirus.

In addition to the principal approach described above, the pipeline employs two alternate methods to detect the presence of a provirus at a locus (Fig. 2). First, average read depth at the solo LTR is compared to the average of read depth of all solo LTRs in the same individual genome. If the sequenced individual has at least one provirus allele instead of a solo LTR (as in the reference genome), we predict to see an increase in the number of uniquely mapping reads mapping to the solo LTR. Indeed, reads derived to the 5′ and 3′ LTR of the proviral allele remain more likely to map uniquely to the solo LTR than to other LTRs located elsewhere in the reference genome. This is because gene conversion events frequently homogenize the sequence of proviral LTRs [60, 61]. Hence the reads derived from the two LTRs of the provirus will preferentially map to the solo LTR annotated in the reference genome, resulting in an increase in read depth at this LTR relative to other solo LTRs in the genome (Additional file 1). Second, a local de novo assembly of all reads including mates is performed and failure to assemble a solo LTR allele is interpreted as an indicator of the presence of two proviral alleles at the locus (Fig. 2, see Methods). Overall the *findprovirus* pipeline predicts the presence of a proviral allele based primarily on the first approach with results from the two alternate approaches used as secondary indicators.

Known and new dimorphic HERVs predicted through the *findprovirus* pipeline

The *findprovirus* pipeline was used to identity dimorphic candidates for HERV-K(HML2), (hereafter simply noted as HERV-K), HERV-H, and HERV-W families in a dataset consisting of whole genome sequence data for 279 individuals from the SGDP [58]. Solo LTRs annotated in the hg38 reference genome for HERV-K (LTR5_Hs) (n = 553), HERV-H (LTR7) (n = 689) and HERV-W (LTR17) (n = 476) were used as initial queries (see Methods). The pipeline reports the following results: (i) number of discordant reads mapping to the region; (ii) number of informative discordant reads (i.e. their mates have a significant hit with the respective HERV coding sequence); (iii) percentage of reference solo LTR allele aligned to de novo assembled contigs from the reads; (iv) ratio of average read depth of the element to the average read depth at all solo LTRs of that individual; (v) average mappability of regions where informative discordant reads are mapped; and (vi) prediction on the presence or absence of the provirus allele. The candidates are then visually inspected using Integrative Genomics Viewer (IGV) for the presence of nested polymorphic transposable element (TE) insertion or presence of internal region of same HERV nearby that could result in false positives. After in silico inspection, we identify three strong candidate loci for HERV-K, two for HERV-H, and one for HERV-W (Additional file 2). Two of the three HERV-K candidates have been previously identified and experimentally validated as dimorphic in prior studies [29, 44, 46] (Table 1). For these two loci, we also identified genomic sequences of the corresponding proviral alleles from the Nucleotide collection (nr/nt) database at the National Centre for Biotechnology Information (NCBI) through homology-based searches (see methods) (Additional file 2). The novel dimorphic candidate that we identified for HERV-K (5q11.2_K3) is predicted to be a provirus in 164 individuals and a maximum of six informative discordant reads are mapped to that locus in an individual (Additional file 2). However, the low average mappability scores for the solo LTR region where the informative discordant reads are mapped suggests that it is a region prone to ambiguous mapping (Additional file 2). Further experimental validations will be necessary to confirm this dimorphism. Nonetheless, these results show that our pipeline efficiently retrieves known dimorphic HERV-K elements.

To the best of our knowledge, none of the dimorphic HERV-H and HERV-W candidates identified herein have been reported in the literature. The two HERV-H candidates were flagged by up to 23 and 6 discordant mate reads aligned to the internal sequence of HERV-H in an individual (Additional file 2). The HERV-W candidate, 18q21.1_W2 displayed up to 33 discordant mates aligned to HERV-W internal sequence in a given individual (Additional file 1). The *findprovirus* pipeline predicted that 194 of 279 individuals had at least one proviral allele of 18q21.1_W2, suggesting that this is a common allele in the human population (Additional file 2). To experimentally validate these three candidates (Additional file 2), we used Polymerase Chain Reaction (PCR) to genotype a panel of individuals from the SGDP predicted to include a mixture of genotypes. Primers were designed in the flanking regions and used as a pair to detect the solo LTR allele or in combination with an internal primer (located in *gag* and/or *env* region) to detect the proviral allele (see Methods). The PCR products were analyzed by gel electrophoresis and their identity

Table 1 Dimorphic HERV-K, HERV-H and HERV-W candidates

HERV name	Coordinate (GRCh38/hg38)	Reference allele	Previously reported
1p31.1_K2[b]	chr1:73129298–73,130,265	Solo LTR	[46]
K111/K105[b]	chrUn_GL000219v1:175210–176,178	Solo LTR	[29, 44]
1p31.1_K3[c]	chr1:75377086–75,383,458	Provirus	[28]
11q22.1_K1[c]	chr11:101695063–101,704,528	Provirus	[28]
12q14.1_K1[c]	chr12:58327459–58,336,915	Provirus	[28]
3q13.2_K2[c]	chr3:113024277–113,033,435	Provirus	[38]
3q27.2_K1	chr3:185562548–185,571,727	Provirus	[43]
5q33.3_K1[c]	chr5:156657706–156,666,885	Provirus	[43]
6q14.1_K1[c]	chr6:77716945–77,726,366	Provirus	[28]
7p22.1_K1	chr7:4582426–4,591,897	Provirus	[28, 38]
5p13.3_K2[c,a]	chr5:30486653–30,496,098	Provirus	This study
4q22.1_H8[b,a]	chr4:91045790–91,046,151	Solo LTR	This study
5p15.31_H2[b,a]	chr5:7262337–7,262,742	Solo LTR	This study
11q13.2_H5[c]	chr11:68633778–68,639,439	Provirus	This study
2q34_H4[a,c]	chr2:209078020–209,084,376	Provirus	This study
2p14_H2[c]	chr2:64252414–64,257,646	Provirus	This study
13q21.32_H1[c]	chr13:66141332–66,147,036	Provirus	This study
1q32.2_H3[c]	chr1:210111090–210,116,207	Provirus	This study
1p32.3_H6[c]	chr1:54897904–54,903,584	Provirus	This study
11q24.3_H2[c]	chr11:130753498–130,759,137	Provirus	This study
11p14.3_H1[c]	chr11:23183934–23,189,744	Provirus	This study
12p12.1_H2[c]	chr12:25163213–25,169,508	Provirus	This study
13q21.1_H1[c]	chr13:55578228–55,584,087	Provirus	This study
13q22.3_H1[c]	chr13:77933223–77,939,379	Provirus	This study
2q36.1_H5[c]	chr2:224296633–224,302,363	Provirus	This study
2p12_H2[c]	chr2:75213731–75,219,537	Provirus	This study
3q22.3_H2[c]	chr3:137595601–137,601,190	Provirus	This study
3p14.3_H1[a,c]	chr3:54634484–54,640,204	Provirus	This study
4q32.3_H5[c]	chr4:166716125–166,722,054	Provirus	This study
6q23.2_H3[c]	chr6:131338800–131,344,564	Provirus	This study
6p22.3_H3[c]	chr6:18754144–18,759,870	Provirus	This study
6p12.2_H1[c]	chr6:51938241–51,944,426	Provirus	This study
6q16.1_H1[c]	chr6:93830156–93,835,749	Provirus	This study
18q21.1_W2[b,a]	chr18:50449151–50,449,914	Solo LTR	This study

Footnotes: This table only lists candidates identified by our pipeline and supported by at least one additional piece of evidence: [a] PCR validation, [b] alternative allele genomic sequence, [c] annotation in the Database of Genomic Variants. Other candidates not listed here are in Additional file 2 and Additional file 9. The notations 'K' 'H' and 'W' in the HERV name represent HERV-K(HML2), HERV-H and HERV-W families respectively

was confirmed by Sanger sequencing (Additional file 3). The results validated that each of the three loci exist as proviral and solo LTR alleles in the human population (Fig. 3a–c, Table 1, Additional file 4). In addition, we also identified seven FOSMID clones in the nr/nt database at NCBI supporting the presence of proviral alleles (Additional files 2, 5, 6 and 7). Altogether these data strongly support the dimorphic HERV-H and HERV-W calls made through our *findprovirus* pipeline.

Strategy for identification of solo LTR allele when the reference allele is a provirus

We developed a complementary pipeline called *findsoloLTR* to mine whole genome resequencing data to detect a solo LTR allele of a locus annotated as a provirus in the reference genome (Figs. 1e and 4). Here the prediction is that an individual with one copy of a proviral allele instead of two will have a decreased number of reads mapping uniquely (mapping quality > = 30) to the internal region

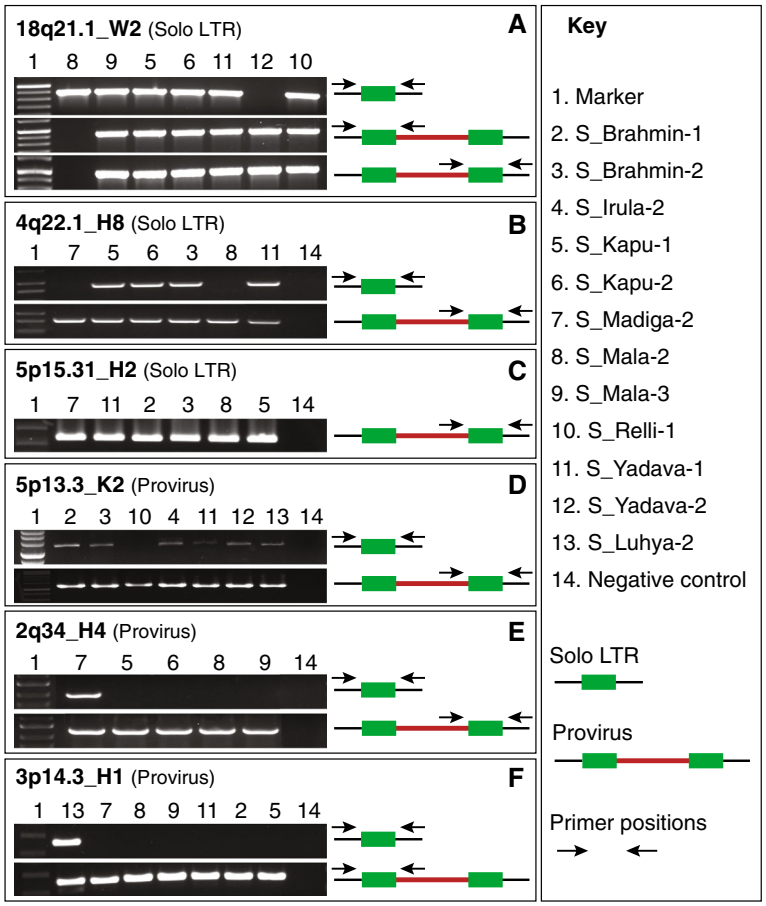

Fig. 3 Experimental validation of dimorphic HERV loci. Type of HERV allele in the reference assembly is shown within brackets after the name of the element. **a** PCR amplification of HERV-W solo LTR at the 18q21.1 locus in the human reference assembly. Primers were designed flanking the solo LTR. PCR amplification of the 18q21.1_W2 provirus with primers designed to the flank and internal *gag* sequence and with primers to the *env* sequence and flank. **b** PCR amplification of HERV-H solo LTR at the 4q22.1 locus in the reference assembly with primers flanking the solo LTR. PCR amplification of the 4q22.1_H8 provirus with primers designed to the internal *env* sequence and flank. **c** PCR amplification of HERV-H provirus at the 5p15.31 locus with primers designed to the internal *env* sequence and flank. The reference allele is solo LTR. **d** PCR amplification of HERV-K solo LTR at the 5p13.3 locus with primers flanking the solo LTR. PCR amplification of the the reference allele 5p13.3_K2 provirus with primers designed to the internal *env* sequence and flank. **e** PCR amplification of HERV-H solo LTR at 2q34 locus with primers flanking the solo LTR. PCR amplification of the reference provirus 2q34_H4 with primers designed to the internal *env* sequence and flank. **f** PCR amplification of HERV-H solo LTR at 3p14.3 locus with primers flanking the solo LTR. PCR amplification of the reference provirus 3p14.3_H1 with primers designed to the internal *gag* sequence and flank. The DNA samples of various South Asian populations and an African individual used for validation are listed in the key. LTRs are in shown as green boxes, the internal region as a red line, the flanking region as a black line. The primer positions are shown as black arrows

and an individual with two solo LTR alleles will have even fewer or no reads mapping uniquely to the internal region of the provirus. The *findsoloLTR* pipeline systematically measures the read depth across the provirus and in the flanking 250-bp regions of the provirus. The pipeline then expresses the average read depth across the provirus as the percentage of the average read depth across its flanking genomic regions (Fig. 4). The candidate locus is considered harboring a solo LTR allele when the calculated read depth ratio across the provirus is lower than 50%. The presence of two solo LTRs alleles is inferred when read depth gets lower than 10% in comparison with the average read depth of the flanking regions (Additional file 8).

Known and new dimorphic HERVs predicted through the *findsoloLTR* pipeline

The *findsoloLTR* pipeline was used to analyze the SGDP data for the presence of solo LTR alleles to a set of sequences annotated as proviruses in the reference genome for HERV-K ($n = 23$), HERV-H ($n = 720$) and HERV-W ($n = 53$). The *findsoloLTR* pipeline reports: (i) mean read depth across the provirus, (ii) mean read depth of the 5′ and 3′ flanks, (iii) percentage of read depth at the provirus to the average of read depth of the flanks and (iv) prediction of the presence of a solo LTR allele. The candidates were visually inspected using IGV to assess whether the decreased read depth ratio was

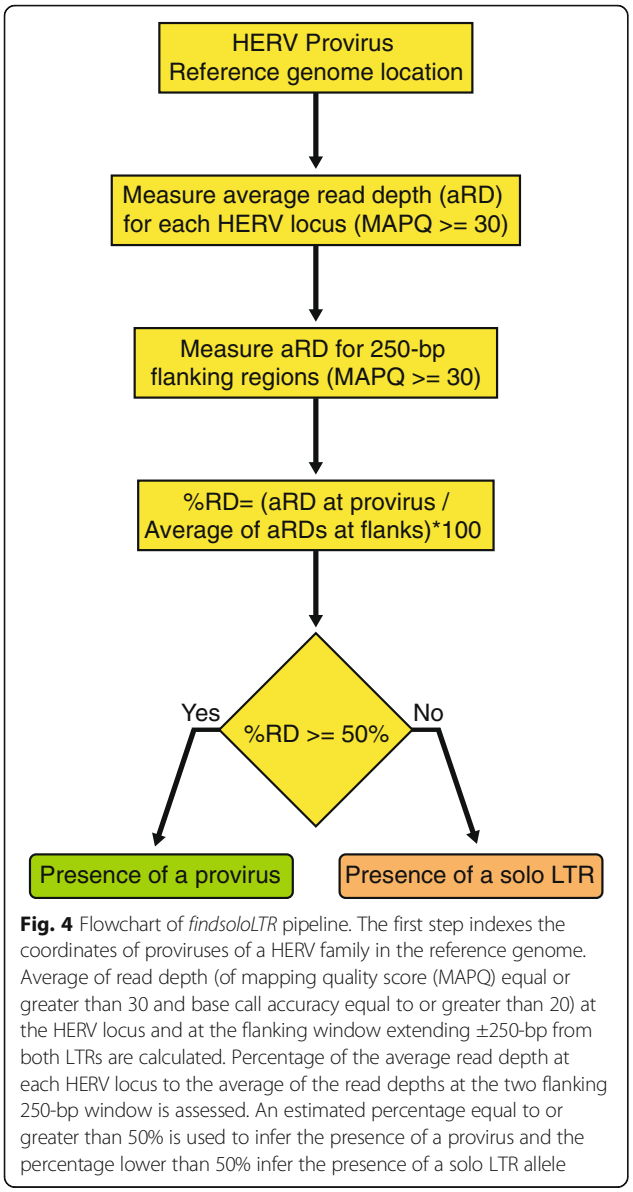

Fig. 4 Flowchart of *findsoloLTR* pipeline. The first step indexes the coordinates of proviruses of a HERV family in the reference genome. Average of read depth (of mapping quality score (MAPQ) equal or greater than 30 and base call accuracy equal to or greater than 20) at the HERV locus and at the flanking window extending ±250-bp from both LTRs are calculated. Percentage of the average read depth at each HERV locus to the average of the read depths at the two flanking 250-bp window is assessed. An estimated percentage equal to or greater than 50% is used to infer the presence of a provirus and the percentage lower than 50% infer the presence of a solo LTR allele

due to a partial deletion instead of the outcome expected for a LTR recombination event which precisely deletes one LTR along with the internal sequence (see Additional file 8 for a legitimate candidate). After in silico inspection, we retained 12 HERV-K candidates, 67 HERV-H candidates, and no HERV-W candidate (Additional file 9).

In the case of HERV-K, eight of the 12 candidate loci were previously reported to be dimorphic, and some were known to be also insertionally polymorphic, i.e. a pre-integration 'empty' allele has also been reported [28, 29, 38, 43, 46] (see Additional file 9). The pipeline predicts four novel HERV-K loci to be dimorphic in the population (Additional file 9). For HERV-H, we observe that many of the predicted solo LTR allele occurs at low frequency in the SGDP dataset, being predicted in only a few individuals (Additional file 9). This might be expected if these alleles arose from relatively recent recombination events. Alternatively, they may represent false positives. To corroborate the *findsoloLTR* results, we interrogated the Database of Genomic Variants (DGV) [62] to assess whether any of the candidate dimorphic HERV-K or HERV-H loci had been previously predicted as copy number variants in the human population. The DGV systematically catalogs structural variants in human genomes reported in prior studies, but importantly it does not yet include data collected from the SDGP [58], thereby potentially serving as independent validation of our predictions from that dataset. We found that two of the four HERV-K candidates and more than half (35 out of 67) of the HERV-H candidates were catalogued in DGV as putative deletion variants (Additional file 9). One of the HERV-K-associated deletions and 20 of the 35 HERV-H-associated deletions were inferred to have breakpoints mapping within the proviral LTRs, consistent with the idea that LTR recombination events caused these deletions (Table 1). The second HERV-K deletion reported in DGV has both breakpoints precisely at the outer boundaries of LTRs, which is consistent with a pre-integration allele previously reported [29]. The remaining 15 HERV-H-associated deletions catalogued in DGV have predicted breakpoints mapping outside of the annotated LTR sequences, which suggests that a different mechanism than LTR recombination could have caused the deletion or that previous breakpoint identification might have been imprecise.

To further validate the *findsoloLTR* results, we selected one HERV-K candidate (5p13.3_K2) and two HERV-H candidates (2q34_H4, 3p14.3_H1) for experimental validation using PCR with primers designed in the flanking regions. In all three cases, the predicted solo LTR alleles were successfully detected by PCR and sequencing (Fig. 3d–f), (Table 1, Additional file 9, Additional file 3). Collectively these data demonstrate that the *findsoloLTR* pipeline efficiently predicts dimorphic HERVs (Additional file 4) and reveal that a surprisingly high fraction (up to ~10%) of HERV-H proviruses occur as solo LTR alleles in the human population, albeit at relatively low frequency.

Potential consequences for transcriptome variation

To begin exploring the functional consequences of these structural variants, we sought to examine whether the candidate dimorphic HERVs were associated with any known protein-coding or non-coding genes (see methods). We found that three HERV-H candidates contribute exonic sequences including transcription start sites or polyadenylation signals to different RefSeq genes and 10 additional HERV-K and HERV-H loci contribute long intergenic non-coding RNA transcripts annotated in the human

reference genome (Additional file 9). Furthermore, 52 of the HERV-H proviruses we predict to occur as solo LTRs in the population have been previously reported as either moderately or highly transcribed in human induced pluripotent stem cells [63]. One of these HERV-H loci, which we validated experimentally (Fig. 3f) corresponds to the RefSeq gene *Embryonic Stem cell Related Gene* (*ESRG*), which has been identified as a marker of pluripotency [63–66]. The *ESRG* transcript initiates within the 5′ LTR of HERV-H and parts of its first and second exons are derived from the internal region of the element [63–65]. Thus, it is likely that recombination to solo LTR would impair *ESRG* transcription and most likely its function. While preliminary, these observations suggest that HERV dimorphisms create structural variation that has the potential to impact the human transcriptome.

Discussion

Sustained efforts have been undertaken to map structural variation across human genomes in the general population or in association with diseases. But relatively sparse attention has been given to the identification of structural variants associated with HERVs, and particularly the type of dimorphism investigated in this study in which the ancestral allele is a provirus and the derived allele is a solo LTR. Such dimorphisms are challenging to identify because the two variants share the exact same junctions with flanking host DNA, which prevents their identification using 'standard' approaches based on split and discordant read mapping (e.g. [17, 55–57]). Here we have developed two pipelines that circumvent these challenges and efficiently identify dimorphic HERVs (Figs. 1d, e, 2 and 4). Both pipelines rely on a priori knowledge of insertion sites in the reference genome and make use of paired-end and read depth information to infer whether a locus annotated as a provirus in the reference genome exist as a solo LTR in a sequenced individual and vice versa (Figs. 2 and 4). Hence our approach differs from but complements previous efforts to identify HERV insertional polymorphisms (presence/absence), which by design cannot typically differentiate proviruses from solo LTRs [17, 55–57].

We applied our pipeline to discover dimorphic loci from three major HERV families of different ages (HERV-K, HERV-H, HERV-W) using sequence data generated from 279 individuals from diverse populations [58] (Fig. 5). Previously, only a dozen HERV-K insertions have been reported to exist as dimorphic provirus/solo LTR alleles in the human population [17, 28, 29, 38, 39, 43, 44, 46]. Our results yielded 15 strong candidate HERV-K dimorphic loci, including 10 previously recognized as dimorphic in the human population, a subset of which are also known to be insertionally polymorphic (see Table 1, Fig. 5, Additional file 2, Additional file 9) [17, 28, 29, 36–47]. These results indicate that our approach did not yield an extensive set of HERV-K candidates that were not identified previously. This observation suggests that the number of HERV-K loci with dimorphic alleles segregating with relatively high frequency in the human population is rather small and it appears that most of these loci have now been identified. Of course it is possible, and even likely, that many more dimorphic HERV-K loci segregate at low frequency in the population. While the SDGP represents a fairly diverse sampling of the human population compared to those previously surveyed for HERV polymorphisms such as the 1000 Genome Project, it still remains minuscule. As sequencing efforts continue to intensify worldwide, our pipeline brings a valuable addition to the toolbox for cataloguing structural variants.

We were intrigued to discover a dimorphic element for the HERV-W family (18q21.1_W2). This element is represented as a solo LTR in the reference genome, but our data clearly show that it also occurs as a provirus segregating in South Asian populations (Fig. 3a) and likely in other diverse populations (our pipeline predicted a provirus allele in 194 out of 279 individuals surveyed, Additional file 2). To the best of our knowledge, this is the first HERV-W locus reported to show any type of dimorphism. This particular HERV-W insertion must have occurred between 18 and 25 million years ago because a provirus is found at orthologous position in all other ape genomes including gibbon, but is absent in Old and New World monkeys [67]. Our discovery illustrates the potential of LTR recombination to alter genome structure long after a proviral insertion has occurred.

We also identified a relatively large number (~ 69) of candidate HERV-H dimorphisms. We experimentally validated the dimorphic nature of four of these HERV-H loci in South Asian populations and in an African individual (Table 1, Figs. 3 and 5, Additional file 2, Additional file 9). While this is a small validation sample, the results suggest that a substantial number of HERV-H loci occur as dimorphic alleles in the human population, with solo LTR alleles apparently segregating at low frequency relative to proviral elements (Table 1, Additional file 2, Additional file 9). To our knowledge, prior to this study only a single dimorphic HERV-H locus had been documented [27]. We did not identify this particular locus in our analysis. However, we noticed that the 5′ and 3′ LTRs of this provirus are annotated by Repeatmasker as belonging to different subfamilies (LTR7 and LTR7Y respectively), an annotation either erroneous or reflecting an inter-element recombination event [68]. In either case, this discrepancy would have excluded this locus from our analysis because the program we used [69] to assemble the starting set of queries requires 5′

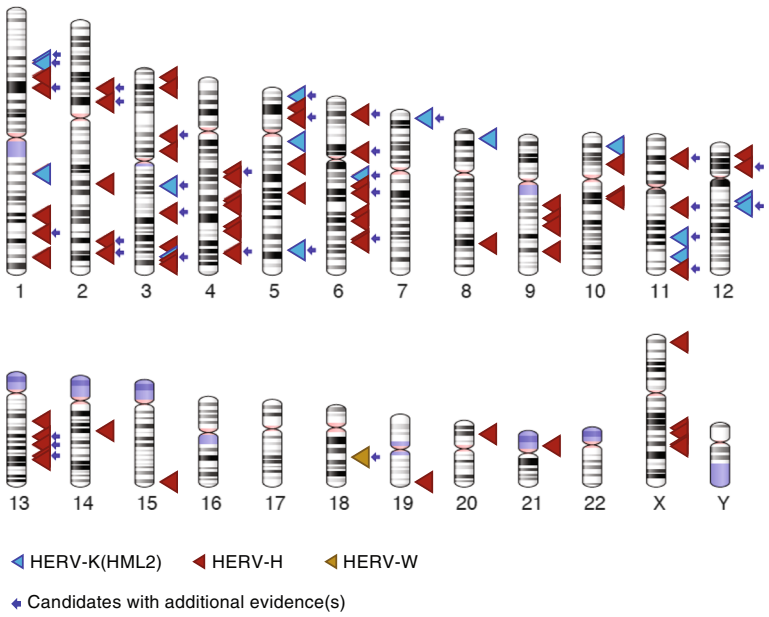

Fig. 5 Karyotypic view of the location of the candidate dimorphic HERVs. The dimorphic candidates of HERV-K (HML2) are shown as blue triangles, HERV-H as red triangles and HERV-W as golden yellow triangle. The candidates that are supported by at least one additional evidence such as PCR validation, alternative allele genomic sequence, annotation in the Database of Genomic Variants are marked with a blue arrow. The genomic coordinates and other details of the candidates are detailed in Additional file 2 and Additional file 9. The ideograms were generated using the genome decoration page at NCBI https://www.ncbi.nlm.nih.gov/genome/tools/gdp

and 3′ LTR names to match in order for a locus to be flagged as a provirus (see Methods). This observation highlights a caveat of our approach: it relies on accurate pre-annotations of the elements in a reference genome in order to correctly identify proviral and solo LTR queries. Clearly, repeat annotation remains an imperfect process even in a 'reference' genome, and HERVs and other LTR elements pose particular challenges for both technical and biological reasons [68, 70, 71]. Efforts are underway to automate and improve repeat annotation [59, 72–75] as well as projects to enhance the quality of genome assemblies and annotations for a wide variety of species. These developments are bound to facilitate and expand the application of our pipeline to many more genomes, both human and non-human.

The large number of dimorphic HERV-H loci we predict to occur in the population may seem surprising given that relatively few HERV-K loci appear to exhibit this type of dimorphism. This difference can be in part explained by the fact that HERV-H is a relatively abundant family with an exceptionally high proportion of proviral insertions relative to solo LTRs maintained in the genome [76, 77]. By our estimates (see Methods) the reference genome includes ~720 HERV-H proviral insertions and 689 solo LTRs. Phylogenetic modeling of the LTR recombination process [76] suggests that HERV-H proviruses have formed solo LTRs at a much lower rate than expected based on their age of residence and the level of sequence divergence of their LTRs. Indeed HERV-K, a younger family, includes 23 proviral copies and 553 solo LTRs (see Methods). The apparent resistance of HERV-H to LTR recombination may be driven by purifying selection to retain proviral HERV-H copies for some sort of cellular function [76]. In fact it has been documented that a subset of HERV-H proviruses are bound by pluripotency transcription factors and are highly expressed in human embryonic stem cells as long noncoding RNAs and chimeric transcripts playing a possible role in the maintenance of pluripotency [63, 78–81]. Our finding that several HERV-H proviruses are reduced to solo LTR alleles in some individuals argues that haploidy for the internal sequences of these elements is sufficient for normal human development. But that is not to say that such structural variation bears no biological consequences. In fact, one of the dimorphic HERV-H loci we validated at 3p14.3 is known to drive *ESRG*, a transcript acting as an early marker of reprogramming of human cells to induced pluripotent stem cells [63–66]. Experimental knockdown of the *ESRG* transcript in human embryonic stem cells leads to a loss of pluripotency and self-renewal [63]. Thus it is intriguing that we identified a solo LTR allele of *ESRG* in two individuals from different African populations (Additional file 9, Fig. 3f). Whether this deletion event impairs *ESRG* transcription and has any functional consequences for human embryonic development awaits further investigation. More generally, our catalog of candidate dimorphic HERVs

provides a valuable resource to assess the regulatory significance of these type of elements [13] and assess whether the process of LTR recombination represents an hitherto 'hidden' source of regulatory divergence in the human population.

These findings also bear important implications for studies that link the coding activities of HERVs to human pathologies. Our results imply that there are more frequent alterations in the copy number of HERV coding sequences than previously appreciated, even for families that apparently have long ceased to be infectious or transpositionally active such as HERV-H and HERV-W [82, 83]. Overexpression of gene products encoded by these families as well as HERV-K has been documented in a number of conditions, including multiple sclerosis (MS) [21], amyotrophic lateral sclerosis (ALS) [25], rheumatoid arthritis [84], systemic lupus erythematosus [85], schizophrenia [86] and type 1 diabetes [87] and several cancers [88–91]. It remains uncertain whether overexpression of HERVs contributes to the etiology or progression of these diseases. But evidence is mounting in the cases of MS and ALS, for which both in vitro studies and mouse models have established that envelope (*env)* proteins expressed by HERV-W and HERV-K respectively, can exert biochemical, cellular and immunological effects that recapitulate the disease symptoms [21]. Conceivably then, variation in the copy number of HERV-encoded genes caused by sporadic LTR recombination events, either in the germline or in somatic cells, could modulate susceptibility to these pathologies. Importantly, three of the dimorphic HERV-K loci predicted herein (Additional file 9) are known to encode full-length env proteins [92]. Thus our results reveal a previously underappreciated source of HERV gene copy number variation with potential pathological ramifications.

Lastly, a growing number of studies have implicated HERV-encoded proteins in beneficial physiological activities, notably in immunity (for review [12]). For instance, overexpression of the HERV-K *gag* protein can interfere with the late phase replication of the HIV-1 retrovirus [93]. Moreover, biochemically active HERV-K proteins appear to be expressed during normal human development where they may confer some form of immunity to the early embryo [94, 95]. For example, endogenous *env* can compete with and effectively restrict the cellular entry of cognate exogenous retroviruses [96, 97], and *env* of the HERV-H and HERV-W families have been shown to have immunosuppressive properties [98, 99]. Thus it is tempting to speculate that some of the genomic variants uncovered herein could contribute to inter-individual immune variation and modulate the risk to develop certain pathologies.

Conclusions

Collectively our results show that we have successfully developed a pipeline to discover dimorphic loci from a variety of HERV families from resequencing data, including two families for which such copy number variation had been scarcely (HERV-H) or never (HERV-W) reported before. Given that there are dozens more HERV families in the human genome, including some substantially younger than HERV-H or HERV-W [68, 71], it is likely that this form of structural variation affect other families and is more common than previously appreciated. Further studies are warranted to investigate the association of such variants with human phenotypes, including disease susceptibility.

Methods

Classification of proviruses and solo LTRs in the reference genome

The repeats annotated as LTR5-Hs and HERV-K-int (HERV-K(HML2 family)), as LTR17 and HERV17-int (HERV-W family) and as LTR7 and HERV-H-int (HERV-H family) are extracted from the RepeatMasker annotation of the human reference (GRCh38/hg38) assembly (RepeatMasker open-4.0.5 - Repeat Library 20140131 available at http://www.repeatmasker.org/). The extracted RepeatMasker data is parsed to identify potentially full-length proviruses and solo LTRs using the tool "One Code to Find Them All" [69]. Using a custom script, (https://github.com/jainy/dimorphicERV) each copy in the parsed output is further classified as a provirus containing (i) 2 LTRs and internal region (ii) 1 LTR and internal region (iii) only internal region or as a solo LTR. The coordinates at the boundaries of each copy is then extracted from the parsed output. Each HERV locus is then given a unique identifier depending on the cytoband it belonged to and based on the total number of copies of that family found in each band. The positions of cytoband for GRCh38/hg38 is downloaded (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBand.txt.gz). The coordinates of HERV copies marked as proviruses with 2LTRs and internal regions and as solo LTRs are used in the subsequent analysis. For HERV-W, the copies that are generated by retrotransposition mediated by LINE-1 machinery have partial LTRs [100] and such copies annotated as pseudogenes [82] were excluded from our analysis.

Identification of provirus allele when the reference allele is a solo LTR

The *findprovirus* pipeline identifies solo LTR to provirus variants in the Binary Alignment/Map (bam) format files where paired end reads from whole genome resequencing data are mapped to reference assembly using Burrows-Wheeler Aligner (BWA) [101] (Figs. 1d and 2) (https://github.com/jainy/dimorphicERV). The pipeline analyses the coordinates of all solo LTRs obtained from One Code to Find Them All (see methods). The *findprovirus* pipeline extracts reads mapped to each solo LTR

and to a flanking 100-bp region using samtools (version 1.4.1) [102]. Only reads that are mapped with a mapping quality of 30 or greater (i.e. mapped with > 99.99% probability) are collected and the reads are processed to fasta format using SeqKit [103]. The discordant reads in the solo LTR and in the flanking 100-bp region are identified using samtools [102] and the mates of discordant reads are extracted using picard tools (version 2.9.2) (http://broadinstitute.github.io/picard/). Sequence homology of mates of discordant reads to the consensus coding sequence of the respective HERV extracted from the Repbase database [59] is tested using BLASTn (version 2.6.0, default parameters) and the number of reads with significant hits (e-value < 0.0001) are counted. Discordant read mates with significant sequence similarity to an internal HERV sequence suggest the presence of a proviral allele in that individual. Two independent, additional approaches are also used to assess the presence of a proviral allele. The first approach measures the average read depth at the solo LTR using samtools and reads with a mapping quality of 20 or more (mapped with > 99% probability) and reads with a base quality of 20 or more (base call accuracy of > 99%) are counted. To get an estimate of the expected coverage at a solo LTR, average of read depths at all solo LTRs of that HERV family for an individual is calculated. This also helps to account for the variability in the coverage between individual genomes. The ratio of average read depth at a solo LTR to the average of read depths observed at all solo LTRs of that HERV family for the individual is determined. An increased read depth pertained to the solo LTR (ratio > 1) is indicative of an increased number of reads mapping to that locus, which is suggestive of the presence of a provirus allele (Fig. 2). As part of the second approach, a local de novo assembly of all extracted reads from a locus (mapped reads and discordant mates) is performed using CAP3 [104] and/or SPAdes (version 3.11.1) [105] to test if the solo LTR allele could be reconstructed. The corresponding reference solo LTR sequence with 50-bp flanking is extracted and sequence similarity of the reference sequence is tested (BLASTn version 2.6.0, default parameters) against assembled contigs. A significant blast hit (e-value < 0.0001) spanning ⩾95% reference genome sequence is indicative of the presence of a solo LTR allele in the individual examined. However, since these two alternate approaches are not always consistent in detecting provirus allele, the results from the two approaches are presented and are not used for the prediction of the provirus allele, but rather as additional indicators.

The performance of the pipeline depends heavily on how accurately reads are mapped to the reference genome. In fact, the mappability across the genome varies remarkably and in order to discern a strong candidate from weak candidate, the mappability of genomic regions [106] where informative discordant reads are mapped is determined for each locus. The regions of low mappability generate ambiguous mapping and regions of high mappability generate unique mapping. The mappability scores are downloaded for the GRCh37/hg19 version of reference assembly (ftp://hgdownload.soe.ucsc.edu/gbdb/hg19/bbi/wgEncodeCrgMapabilityAlign100mer.bw). The downloaded file is processed [107] and is converted to bed format [108] and scores are lifted over [109] to hg38 version. This data is stored in an indexed mysql table. The coordinates of the reference assembly where the informative discordant reads are mapped for each solo LTR are identified using bedtools (version 2.26.0) [110]. The mappability scores for those genomic regions are extracted from the table and the mean of the mappability scores is provided in the output of the pipeline.

Identification of solo LTR allele when the reference allele is a provirus

The *findsoloLTR* pipeline identifies the provirus to solo LTR variants in bam files (Fig. 1e and 4, https://github.com/jainy/dimorphicERV). It first calculates the read depth across the provirus using samtools [102]. Read depth is calculated for reads with a mapping quality of 30 or more and with a base quality score of 20 or more. Similarly, read depth is calculated across 5′ and 3′ flanking 250-bp regions. The pipeline then assesses the percentage of average read depth across the provirus to the average of read depths across the flanks. Presence of two proviral alleles is inferred when the read depth percentage greater than or equal to 50% and read depth percentage lower than 50% is used to infer the presence of solo LTR allele (Fig. 1e). A read depth percentage lower than 10% is arbitrarily used to infer the presence of two solo LTR alleles. The mappability scores [106] of the genomic region spanning the the provirus are extracted (see methods for *findprovirus*) and the mean of the mappability scores is provided in the output of the pipeline.

Dataset analyzed

The two pipelines were run on the publicly available whole genome sequence data generated as part of the SGDP for 279 individuals from 130 populations [58]. The bam files used for the analysis are generated by aligning 100-bp long paired-end reads to the GRch38/hg38 version of the human genome using BWA aligner (version 0.7.12) [101]. The bwa-mem alignment allowed a mismatch penalty of 4 (equivalent to 96% identity) and allowed secondary alignments (multi-mapping).

In silico validation

An in silico validation of the candidates identified by both pipelines is performed to filter out false positives. Each of the candidate loci including their flanking region (1000 bp) was visually inspected using IGV (version

2.3.97) after loading a track with RepeatMasker annotation of hg38 version of the human genome (RepeatMasker open-4.0.5 - Repeat Library 20,140,131). The candidates (identified through *findprovirus* pipeline) having an internal region of the respective HERV family nearby or having a nested polymorphic TE, both hallmarks of false positives, are filtered out. Candidate loci not supported by a minimum of four discordant reads where mates align to the internal coding sequence of HERV in at least one individual are also filtered out. The candidates (identified through *findsoloLTR* pipeline) having deletion restricted to a fragment of internal sequence are removed. After visual inspection, the candidates are then queried in the DGV [62] to identify if any previous studies have reported those loci as a copy number variant (CNV). The CNVs identified in DGV are visually inspected for the concordance of their breakpoints with the two LTRs, which is suggestive of their origin through LTR mediated recombination. The CNVs having one or both breakpoints lie outside the LTRs are also identified. The candidates along with 100-bp flanking sequence are also queried against nr/nt database at NCBI to identify the presence of any BAC/FOSMID clones containing corresponding the solo LTR or provirus variant.

Experimental validation

After in silico validation, PCR primers are designed in the regions flanking the LTR and in the *gag* and/or *env* regions assembled from the mates of the discordant reads for selected candidates. The solo LTR allele is amplified by primer pairs flanking the solo LTR and the proviral allele is amplified with the internal primer located on the *env* region or *gag* region. The primers for validating the dimorphic HERVs are designed using PrimerQuest [111] and the oligos are synthesized from Integrated DNA Technologies (IDT). For PCR validation, genomic DNA samples are selected based on the predicted genotype and availability. The sample ids of 12 individuals in the SGDP data set [58] used for PCR analysis are S_Brahmin-1, S_Brahmin-2, S_Irula-2, S_Kapu-1, S_Kapu-2, S_Madiga-2, S_Mala-2, S_Mala-3, S_Relli-1, S_Yadava-1, S_Yadava-2 and S_Luhya-2. PCR amplifications are performed using GoTaq PCR Master Mix (Promega) or Platinum SuperFi PCR Master Mix (Thermo Fisher Scientific). The primer sequences and PCR conditions used for each reaction are given in Additional file 10. PCR products are visualised using agarose gel electrophoresis and are purified using DNA Clean & Concentrator™-5 (Zymo Research) following manufacturer's instructions. The purified PCR products are Sanger sequenced at the DNA sequencing Core Facility, University of Utah or at Genewiz. The generated sequences are analyzed using Sequencher 5.4.6 (Gene Codes Corporation).

Analysis of contribution of dimorphic candidate HERVs to annotated genes/transcripts

The dimorphic candidate HERV loci are examined individually using the University of California, Santa Cruz (UCSC) genome browser on human GRCh38/hg38 assembly [112] (last accessed June 6 2018) to identify any overlap with known NCBI RefSeq protein-coding or non-coding genes (NM_*, NR_*, and YP_*). In addition, to determine the dimorphic candidates that encode an intact *env* gene, the HERV coordinates are compared with that of intact *env* Open Reading Frames (ORFs) identified by Heidmann et al. [92] in the human genome (hg38). In order to find the candidate dimorphic HERV-Hs that are actively transcribed in human embryonic or induced pluripotent stem cells (iPSCs), coordinates of HERV-Hs, which are known to be moderately or highly expressed in hiPSC lines and single cells [63] are intersected with coordinates of dimorphic HERV candidates using bedtools v2.26.0 [110].

Additional files

Additional file 1: IGV screenshot of the dimorphic HERV-W locus 18q21.1_W2. (PDF 30 kb)

Additional file 2: HERV candidates identified as solo LTR to provirus variants using *findprovirus* pipeline. (PDF 96 kb)

Additional file 3: Sequence verification of non-reference solo LTR or provirus alleles. (PDF 47 kb)

Additional file 4: Comparison of predictions of solo LTR and provirus from findsoloLTR and findprovirus pipelines to PCR based genotypes for solo LTR (S) and provirus (P) polymorphisms. (PDF 168 kb)

Additional file 5: Alignment of 18q21.1_W2 solo LTR and provirus with HERV-W consensus. (PDF 707 kb)

Additional file 6: Alignment of 4q22.1_H8 solo LTR and provirus with HERV-H consensus. (PDF 410 kb)

Additional file 7: Alignment of 5p15.31_H2 solo LTR and provirus with HERV-H consensus. (PDF 350 kb)

Additional file 8: IGV screenshot of dimorphic HERV-H (2q34_H4) locus. (PDF 36 kb)

Additional file 9: HERV candidates identified as provirus to solo LTR variants using *findsoloLTR* pipeline. (PDF 723 kb)

Additional file 10: List of primer sequences used for amplifying solo LTR and provirus alleles shown in Fig. 3. (PDF 124 kb)

Abbreviations

CNV: Copy number variant; DGV: Database of genomic variation; *ESRG*: *Embryonic Stem cell Related Gene*; HERV: Human endogenous retrovirus elements; LTR: Long terminal repeat; PCR: Polymerase Chain Reaction; SGDP: Simons Genome Diversity Project Project

Acknowledgments

We would like to thank W. Scott Watkins for providing the 279 SGDP bam files for the analysis and Lynn B. Jorde for providing the DNA samples as well as equipment and infrastructure for performing this study. We also would like to thank Aurelie Kapusta and Clement Goubert for many helpful discussions.

Funding

This work was supported by a research agreement between GeNeuro Innovation SAS (http://www.geneuro.com/en/about/overview) and the

University of Utah as well as funds from the National Institutes of Health (R35 GM122550, R01 GM059290) to C.F. The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets generated as part of the study are available as supplementary information. The scripts developed as part of the study are available at github (https://github.com/jainy/dimorphicERV).

Authors' contributions
JT designed, wrote the scripts, analyzed, interpreted the data and drafted the manuscript. HP conceived the study and contributed to writing. CF conceived and designed the study, contributed to analysis and interpretation and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
J.T and C.F declare no competing interests. H.P receives compensation for his work by Geneuro and is inventor on patents owned by BioMérieux, INSERM, or Geneuro, but has transferred all his rights to BioMérieux or to Geneuro under applicable laws for employed inventors.

Publisher's Note

Author details
[1]Department of Human Genetics, University of Utah School of Medicine, 15 North 2030 East, Rm 5100, Salt Lake City, UT 84112, USA. [2]GeNeuro, Plan-les-Ouates, Geneva, Switzerland. [3]Université Claude Bernard, Lyon, France. [4]Department of Molecular Biology and Genetics, Cornell University, 107 Biotechnology Building, Ithaca, NY 14853, USA.

Received: 17 September 2018 Accepted: 29 November 2018
Published online: 18 December 2018

References
1. Boeke JD, Stoye JP. Retrotransposons, endogenous retroviruses, and the evolution of Retroelements. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1997. p. 343–436.
2. Jern P, Coffin JM. Effects of retroviruses on host genome function. Annu Rev Genet. 2008;42(1):709–32.
3. 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.
4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520–62.
5. Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M. Retroviral diversity and distribution in vertebrates. J Virol. 1998;72(7):5955–66.
6. Zhuo X, Rho M, Feschotte C. Genome-wide characterization of endogenous retroviruses in the bat Myotis lucifugus reveals recent and diverse infections. J Virol. 2013;87(15):8493–501.
7. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437(7055):69–87.
8. Jern P, Sperber GO, Blomberg J. Divergent patterns of recent retroviral integrations in the human and chimpanzee genomes: probable transmissions between other primates and chimpanzees. J Virol. 2006;80(3):1367–75.
9. Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, et al. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 2012;13(6):R45.
10. Schauer SN, Carreira PE, Shukla R, Gerhardt DJ, Gerdes P, Sanchez-Luque FJ, et al. L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. 2018;28(5):639–53.
11. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21–42.
12. Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. Curr Opin Virol. 2017;25:81–9.
13. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017;18(2):71–86.
14. Jenkins NA, Copeland NG, Taylor BA, Lee BK. Dilute (d) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. Nature. 1981;293(5831):370–4.
15. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2006;2(1):e2.
16. Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene. 2008;27(3):404–8.
17. 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):E2326–34.
18. Engel ME, Hiebert SW. The enemy within: dormant retroviruses awaken. Nat Med. 2010;16(5):517–8.
19. María G-C, Paola I, Niki K, Mariacarmela S, Julià B, Rafael R. Human endogenous retroviruses and cancer. Cancer Biol Med. 2016;13(4):483.
20. Bannert N, Hofmann H, Block A, Hohn O. HERVs New Role in Cancer: From Accused Perpetrators to Cheerful Protectors. Front Microbiol. 2018;9:178. Available from: https://doi.org/10.3389/fmicb.2018.00178
21. Küry P, Nath A, Créange A, Dolei A, Marche P, Gold J, et al. Human endogenous retroviruses in neurological diseases. Trends Mol Med. 2018;24(4):379–94.
22. Gröger V, Cynis H. Human endogenous retroviruses and their putative role in the development of autoimmune disorders such as multiple sclerosis. Front Microbiol. 2018;9:265.
23. Grandi N, Tramontano E. HERV envelope proteins: physiological role and pathogenic potential in Cancer and autoimmunity. Front Microbiol. 2018;9:462.
24. Perron H, Dougier-Reynaud H-L, Lomparski C, Popa I, Firouzi R, Bertrand J-B, et al. Human endogenous retrovirus protein activates innate immunity and promotes experimental allergic encephalomyelitis in mice. PLoS One. 2013;8(12):e80128.
25. Li W, Lee M-H, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med. 2015;7(307):307ra153.
26. Moyes D, Griffiths DJ, Venables PJ. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet. 2007;23(7):326–33.
27. Mager DL, Goodchild NL. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am J Hum Genet. 1989;45(6):848–54.
28. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. Proc Natl Acad Sci U S A. 2004;101(6):1668–72.
29. 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.
30. Seperack PK, Strobel MC, Corrow DJ, Jenkins NA, Copeland NG. Somatic and germ-line reverse mutation rates of the retrovirus-induced dilute coat-color mutation of DBA mice. Proc Natl Acad Sci U S A. 1988;85(1):189–92.
31. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. 1992;6(8):1457–65.
32. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. Science. 2004;304(5673):982.
33. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24(3):1242–55.
34. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49–61.
35. Copeland NG, Hutchison KW, Jenkins NA. Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. Cell. 1983;33(2):379–87.

36. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol. 2001;11(19):1531–5.
37. Mamedov I, Lebedev Y, Hunsmann G, Khusnutdinova E, Sverdlov E. A rare event of insertion polymorphism of a HERV-K LTR in the human genome. Genomics. 2004;84(3):596–9.
38. Macfarlane C, Simmonds P. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. J Mol Evol. 2004;59(5):642–56.
39. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol. 2005;79(19):12507–14.
40. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56–64.
41. Jha AR, Nixon DF, Rosenberg MG, Martin JN, Deeks SG, Hudson RR, et al. Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. PLoS One. 2011;6(5):e20234.
42. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012; 337(6097):967–71.
43. Shin W, Lee J, Son S-Y, Ahn K, Kim H-S, Han K. Human-specific HERV-K insertion causes genomic variations in the human genome. PLoS One. 2013;8(4):e60605.
44. Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Kappes F, et al. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. Genome Res. 2013; 23(9):1505–13.
45. Marchi E, Kanapin A, Magiorkinis G, Belshaw R. Unfixed endogenous retroviral insertions in the human population. J Virol. 2014;88(17):9529–37.
46. Macfarlane CM, Badge RM. Genome-wide amplification of proviral sequences reveals new polymorphic HERV-K(HML-2) proviruses in humans and chimpanzees that are absent from genome assemblies. Retrovirology. 2015;12:35.
47. Kahyo T, Yamada H, Tao H, Kurabe N, Sugimura H. Insertionally polymorphic sites of human endogenous retrovirus-K (HML-2) with long target site duplications. BMC Genomics. 2017;18(1):487.
48. Lee A, Huntley D, Aiewsakun P, Kanda RK, Lynn C, Tristem M. Novel Denisovan and Neanderthal retroviruses. J Virol. 2014;88(21):12907–9.
49. Medstrand P, Mager DL. Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol. 1998;72(12):9782–7.
50. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol. 1999;9(16):861–8.
51. Mayer J, Sauter M, Rácz A, Scherer D, Mueller-Lantzsch N, Meese E. An almost-intact human endogenous retrovirus K on human chromosome 7. Nat Genet. 1999;21(3):257–8.
52. Costas J. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length Proviral genomes. J Mol Evol. 2001;53(3):237–43.
53. Reus K, Mayer J, Sauter M, Zischler H, Müller-Lantzsch N, Meese E. HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J Virol. 2001;75(19):8917–26.
54. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, et al. Rate of Recombinational deletion among human endogenous retroviruses. J Virol. 2007;81(17):9437–42.
55. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29(3):389–90.
56. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27(11):1916–29.
57. Santander CG, Gambron P, Marchi E, Karamitros T, Katzourakis A, Magiorkinis G. STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. Virus Evol. 2017;3(2):vex023.
58. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6.
59. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
60. Kijima TE, Innan H. On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol. 2010;27(4):896–904.
61. Trombetta B, Fantini G, D'Atanasio E, Sellitto D, Cruciani F. Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. Sci Rep. 2016;6:28710.
62. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986–92.
63. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516(7531):405–9.
64. Zhao M, Ren C, Yang H, Feng X, Jiang X, Zhu B, et al. Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies HESRG, a novel stem cell gene. Biochem Biophys Res Commun. 2007;362(4):916–22.
65. Li G, Ren C, Shi J, Huang W, Liu H, Feng X, et al. Identification, expression and subcellular localization of ESRG. Biochem Biophys Res Commun. 2013; 435(1):160–4.
66. Rand TA, Sutou K, Tanabe K, Jeong D, Nomura M, Kitaoka F, et al. MYC releases early reprogrammed human cells from proliferation pause via retinoblastoma protein inhibition. Cell Rep. 2018;23(2):361–75.
67. Grandi N, Cadeddu M, Blomberg J, Mayer J, Tramontano E. HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in Catarrhini and related ERV groups in Platyrrhini. BMC Evol Biol. 2018;18(1):6.
68. Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, et al. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology. 2016;13:7.
69. Bailly-Bechet M, Haudry A, Lerat E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA. 2014;5(1):13.
70. Ou S, Jiang N. LTR_retriever: A highly accurate and sensitive program for identification of LTR retrotransposons [Internet]; 2017. Available from: https://doi.org/10.1101/137141
71. Kojima KK. Human transposable elements in Repbase: genomic footprints from fish to humans. Mob DNA. 2018;9:2.
72. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015;6:13.
73. Arensburger P, Piégu B, Bigot Y. The future of transposable element annotation and their classification in the light of functional genomics - what we can learn from the fables of Jean de la Fontaine? Mob Genet Elem. 2016;6(6):e1256852.
74. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44(D1):D81–9.
75. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017;8:19.
76. Gemmell P, Hein J, Katzourakis A. Phylogenetic analysis reveals that ERVs "die young" but HERV-H is unusually conserved. PLoS Comput Biol. 2016; 12(6):e1004964.
77. Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD. Pluripotency and the endogenous retrovirus HERVH: conflict or serendipity? BioEssays. 2016;38(1):109–17.
78. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13(11):R107.
79. Santoni FA, Guerra J, Luban J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. Retrovirology. 2012;9:111.
80. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9(4): e1003470.
81. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. 2014;21(4):423–5.
82. Grandi N, Cadeddu M, Blomberg J, Tramontano E. Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. Retrovirology. 2016; 13(1):67.
83. Mayer J, Meese E. Human endogenous retroviruses in the primate lineage and their influence on host genomes. Cytogenet Genome Res. 2005;110(1–4):448–56.
84. Freimanis G, Hooley P, Ejtehadi HD, Ali HA, Veitch A, Rylance PB, et al. A role for human endogenous retrovirus-K (HML-2) in rheumatoid arthritis: investigating mechanisms of pathogenesis. Clin Exp Immunol. 2010;160(3):340–7.
85. Nelson P, Rylance P, Roden D, Trela M, Tugnet N. Viruses as potential pathogenic agents in systemic lupus erythematosus. Lupus. 2014;23(6):596–605.

86. Slokar G, Hasler G. Human Endogenous Retroviruses as Pathogenic Factors in the Development of Schizophrenia. Front Psychiatry. 2016;6:183. Available from: https://doi.org/10.3389/fpsyt.2015.00183
87. Levet S, Medina J, Joanou J, Demolder A, Queruel N, Réant K, et al. An ancestral retroviral protein identified as a therapeutic target in type-1 diabetes. JCI Insight. 2017;2(17):e94387.
88. Takahashi Y, Harashima N, Kajigaya S, Yokoyama H, Cherkasova E, McCoy JP, et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. J Clin Invest. 2008;118(3):1099–109.
89. Goering W, Schmitt K, Dostert M, Schaal H, Deenen R, Mayer J, et al. Human endogenous retrovirus HERV-K(HML-2) activity in prostate cancer is dominated by a few loci. Prostate. 2015;75(16):1958–71.
90. Cherkasova E, Scrivani C, Doh S, Weisman Q, Takahashi Y, Harashima N, et al. Detection of an immunogenic HERV-E envelope with selective expression in clear cell kidney Cancer. Cancer Res. 2016;76(8):2177–85.
91. Grandi N, Tramontano E. Type W Human Endogenous Retrovirus (HERV-W) Integrations and Their Mobilization by L1 Machinery: Contribution to the Human Transcriptome and Impact on the Host Physiopathology. Viruses. 2017;9(7):162.
92. Heidmann O, Béguin A, Paternina J, Berthier R, Deloger M, Bawa O, et al. HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. Proc Natl Acad Sci U S A. 2017;114(32):E6642–51.
93. Monde K, Contreras-Galindo R, Kaplan MH, Markovitz DM, Ono A. Human endogenous retrovirus K gag coassembles with HIV-1 gag and reduces the release efficiency and infectivity of HIV-1. J Virol. 2012;86(20):11194–208.
94. Lokossou AG, Toudic C, Barbeau B. Implication of human endogenous retrovirus envelope proteins in placental functions. Viruses. 2014;6(11):4609–27.
95. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522(7555):221–5.
96. Malfavon-Borja R, Feschotte C. Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. J Virol. 2015;89(8):4047–50.
97. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Elife. 2017;6:e22519.
98. Mangeney M, Thomas G, de Parseval N, Heidmann T. The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties. J Gen Virol. 2001;82(10):2515–8.
99. Tolosa JM, Schjenken JE, Clifton VL, Vargas A, Barbeau B, Lowry P, et al. The endogenous retroviral envelope protein syncytin-1 inhibits LPS/PHA-stimulated cytokine responses in human blood and is sorted into placental exosomes. Placenta. 2012;33(11):933–41.
100. Pavlicek A, Paces J, Elleder D, Hejnar J. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res. 2002;12(3):391–9.
101. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60.
102. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
103. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11(10):e0163962.
104. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9(9):868–77.
105. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
106. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PLoS One. 2012;7(1):e30377.
107. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17):2204–7.
108. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28(14):1919–20.
109. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC genome browser database: update 2006. Nucleic Acids Res. 2006;34(Database issue):D590–8.
110. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
111. Owczarzy R, Tataurov AV, Wu Y, Manthey JA, KA MQ, Almabrazi HG, et al. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res. 2008;36(Web Server issue):W163–9.
112. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, et al. The UCSC genome browser database: 2018 update. Nucleic Acids Res. 2018;46(D1):D762–9.