

# SCIENTIFIC REPORTS



OPEN

## Transcriptional landscape of human endogenous retroviruses (HERVs) and other repetitive elements in psoriatic skin

Freddy Lättekivi<sup>1</sup>, Sulev Kõks<sup>1,4</sup>, Maris Keermann<sup>2,3</sup>, Ene Reimann<sup>1</sup>, Ele Prans<sup>1</sup>, Kristi Abram<sup>2,3</sup>, Helgi Silm<sup>2,3</sup>, Gea Kõks<sup>1</sup> & Külli Kingo<sup>2,3</sup>

Human endogenous retrovirus (HERV) sequences make up at least 8% of the human genome. Transcripts originating from these loci as well as proteins encoded by them have been detected in various tissues. HERVs are believed to be implicated in autoimmune diseases, however the extent to which, has remained unclear. Differential expression studies have so far been limited to certain HERV subfamilies with conserved sequences. No studies have been published describing the genome-wide expression pattern of HERVs and repetitive elements in the context of psoriasis. In the present study, we analysed total RNA sequencing data from skin samples of 12 psoriasis patients and 12 healthy controls, which enabled us to describe the entire transcriptional landscape of repetitive elements. We report high levels of repetitive element expression in the skin of psoriasis patients as well as healthy controls. The majority of differentially expressed elements were downregulated in lesional and non-lesional skin, suggesting active HERV suppression in the pro-inflammatory environment of psoriatic skin. However, we also report upregulation of a small subset of HERVs previously described in the context of autoimmune diseases, such as members of the HERV-K and W families, with the potential to affect the immunopathogenesis of psoriasis.

Human endogenous retroviruses (HERVs), members of the long terminal repeat (LTR) retrotransposons repetitive element class, make up at least 8–10% of the human genome<sup>1</sup>. Having acquired numerous mutations over time, human endogenous retroviruses (HERVs) are believed to be incapable of replicating or forming infectious particles<sup>2</sup>. While it has been shown that LINEs and SINEs (non-LTR retrotransposons) can be the cause of several human genetic disorders<sup>3</sup>, the view on HERVs is not yet as clear. Regulatory loci in HERV sequences can act as enhancers, leading to altered expression of nearby genes<sup>4</sup>, or as extra transcription start sites, causing the formation of chimeric transcripts<sup>5</sup>. Normally, HERVs and other retrotransposons are repressed by DNA methylation<sup>6</sup> or intrinsic antiviral factors<sup>7</sup>. As exogenous retroviruses are capable of inducing immune response, the expressed HERV sequences and proteins encoded by them can also be recognized by the host immune system as foreign and potentially hostile elements<sup>8</sup>. For example, members of the HERV-W family have been found to activate the innate immune system via CD14/TLR4 signaling<sup>9</sup>.

Psoriasis is an immune-mediated chronic inflammatory disease of the skin. Implication of viral elements in the immune responses of psoriasis is not a recent idea<sup>10,11</sup>. More recently, expression of retroviral polymerase (pol) encoding members of the HERV-W, K and E families have been reported in psoriatic skin, while rarely detected in non-lesional skin<sup>12</sup>. Transcripts originating from a copy of HERV-K deoxyuridine triphosphate nucleotidohydrolase (dUTPase), located within the psoriasis susceptibility 1 (PSORS1) locus, have been found in both peripheral blood mononuclear cells (PBMCs) and skin tissue<sup>13</sup>. It has been demonstrated that HERV-K dUTPase proteins can induce the secretion of Th1 and Th17 cytokines in both dendritic cells and keratinocytes<sup>14</sup>. Conversely, Gupta and colleagues demonstrated the downregulation of transcripts containing HERV-K env, gag

<sup>1</sup>Department of Pathophysiology, University of Tartu, Tartu, Estonia. <sup>2</sup>Department of Dermatology, University of Tartu, Tartu, Estonia. <sup>3</sup>Clinic of Dermatology, Tartu University Hospital, Tartu, Estonia. <sup>4</sup>Department of Reproductive Biology, Estonian University of Life Sciences, Tartu, Estonia. Correspondence and requests for materials should be addressed to S.K. (email: [sulev.koks@ut.ee](mailto:sulev.koks@ut.ee))

Element Class	LP	NLP	C	RepeatMasker
SINE	31.72	30.84	31.43	11.33
Satellite	0.09	0.10	0.14	0.35
scRNA	0.03	0.03	0.04	≤0.01
LTR	15.96	16.39	16.01	7.46
srpRNA	0.03	0.04	0.03	≤0.01
DNA	7.54	7.67	7.63	3.15
snRNA	0.05	0.05	0.05	≤0.01
RNA	0.06	0.08	0.05	≤0.01
LINE	44.48	44.75	44.56	77.68
RC	0.04	0.04	0.04	≤0.01

**Table 1.** Relative abundance of repetitive element classes. Columns “LP”, “NLP”, and “C” represent relative abundance of repetitive element classes based on read counts in the corresponding study groups and column “RepeatMasker” describes the relative amount of basepairs annotated per element class in RepeatMasker hg19 Library. All values presented as percentages. LP – lesional skin; NLP – non-lesional skin; C – control group.

or pol sequences in the lesional skin of psoriasis patients compared to non-lesional skin and healthy skin from control group subjects<sup>15</sup>.

With the exception of the study by Molès and colleagues<sup>12</sup>, limited to pol-encoding HERVs, no studies have been published describing the genome-wide expression pattern of HERVs and repetitive elements in the context of psoriasis. In the present study, we re-analysed our previously published RNA-seq data<sup>16,17</sup>, utilizing the RepEnrich pipeline<sup>18</sup> in order to investigate the genome-wide transcriptional landscape of HERVs and repetitive elements in the lesional and non-lesional skin of psoriasis patients.

## Results

**RNA sequencing and read alignment.** RNA sequencing of the 12 lesional skin (LP), 12 non-lesional skin (NLP) and 12 healthy control (C) samples investigated in this study produced on average 27 million reads per sample (Fig. S1). With no mapping quality cut-off, on average 68.6% of the sequenced reads mapped to the hg19 reference genome. 27.5% of reads mapping to the hg19 reference aligned to annotated repetitive elements loci in the RepeatMasker<sup>19</sup> hg19 Library. Out of 1116 repetitive element subfamilies queried, we detected the expression of 1033 subfamilies with a cut-off of one or more aligned reads per kilobase of transcript per million mapped reads (RPKM ≥ 1) as mean of all samples. Each of the subfamilies represent all known genomic instances of a given repetitive element classified by RepeatMasker, and are hereinafter referred to as *repetitive elements* or simply as *elements*. RPKM values and raw counts of all expressed elements can be found in Supplementary Materials.

The expression of some repetitive elements is characterized by very large RPKM values (Fig. S2A and B). The mean RPKM values of all samples per element are in correlation (Pearson’s  $r = 1$ ;  $p \leq 2.2 \times 10^{-16}$ ) with the number of annotated genomic instances for that element (Fig. S3). For example, 142 003 and 48 914 separate genomic loci are annotated for AluSx and AluSP elements with RPKM values amounting to 40 768 and 14 565, respectively. The subsequently detected differentially expressed (DE) elements were not more likely to have higher expression levels than not differentially expressed elements ( $p$ -value > 0.05, one-sided Mann-Whitney U test). No differences in the relative abundance of reads originating from different repetitive element classes were observed between study groups in our dataset (Table 1). We note that the variability of expression levels is larger in the control group compared to the LP and NLP groups (Fig. S4A–C).

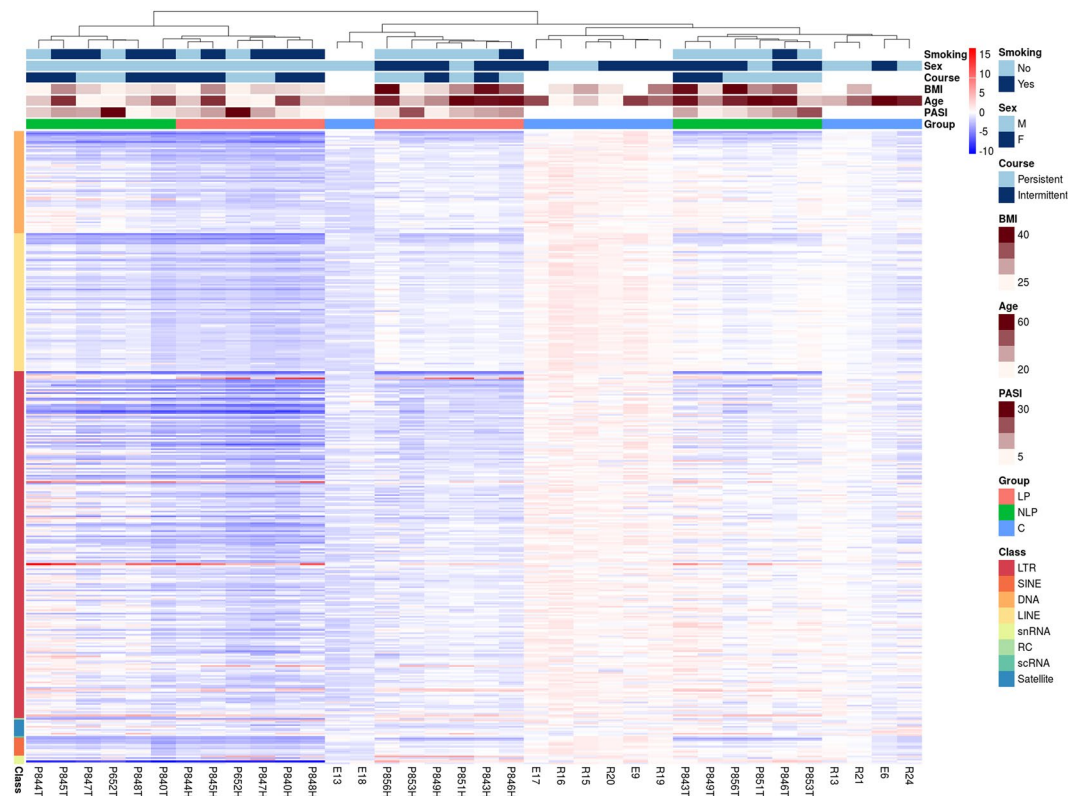
**Differentially expressed elements: LP vs C and NLP vs C.** Comparisons between LP and C, NLP and C groups resulted in 414 and 91 DE elements at  $FDR \leq 0.01$ , respectively. Compared to the control group, most of the repetitive elements were less expressed in both the LP (98%) and NLP (95%) groups (Fig. S5). The few upregulated elements can be found in Table 2. In addition to HERV sequences, small nuclear RNAs (snRNA) U17 and U14, involved in ribosomal RNA processing<sup>20</sup>, and satellite repeat MSR1 were also upregulated in the LP group compared to the control group. In the NLP vs C comparison, the only upregulated non-LTR sequence was satellite repeat HSAT5. Elements differentially expressed at a level of  $\log_2FC \leq -0.5$  or  $\log_2FC \geq 0.5$  (translating into foldchange difference greater than 0.71 or 1.41) are listed in Tables S1 and S2. The full listing of differentially expressed elements can be found in Tables S3 and S4. At the repeat family level, we observed 17 and 3 downregulated families at  $FDR \leq 0.01$  in the LP vs C and NLP vs C comparisons, respectively (Tables S5 and S6).

The larger part of elements downregulated in NLP compared to C overlapped with elements downregulated in LP compared to C, with the number of elements downregulated in LP being more than four-fold greater than in NLP. This trend was also observed for the upregulated elements (Fig. S6). The graphical representation of the first two principal components and clustering analysis of all elements differentially expressed at  $FDR \leq 0.01$  in LP vs C or NLP vs C analyses show distinguishable clustering of the expression patterns of the three groups (Figs S7 and S8). Clustering based on z-scores resulted in all three groups being divided into subgroups, while clustering together individuals with greater age or higher body mass index (Fig. 1).

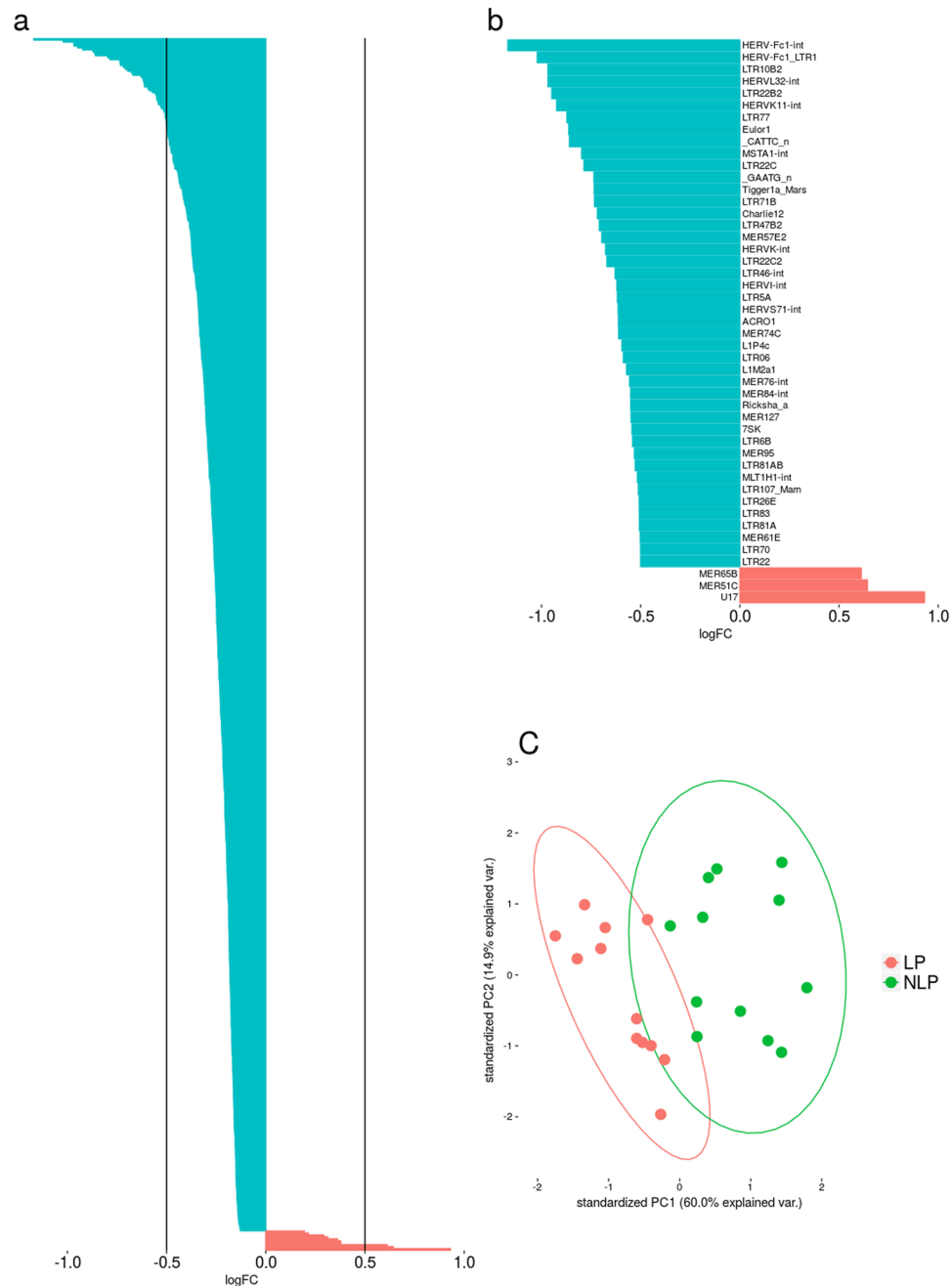
**Differentially expressed elements: pairwise comparison.** Pairwise analysis of LP and NLP groups with the edgeR package resulted in 621 DE elements at  $FDR \leq 0.01$ . Elements differentially expressed at a level

Class	Family	Element	LP vs C	NLP vs C	LP vs NLP
LTR	ERVK	LTR5	0.76	1.00	<b>-0.23 (0.127)</b>
LTR	ERVK	HERVK11D-int	0.18 (0.207)	<b>-0.20 (0.406)</b>	0.38
LTR	ERVK	HERVK14C-int	0.14 (0.338)	<b>-0.2 (0.190)</b>	0.40
LTR	ERV1	MER34B-int	0.13 (0.246)	<b>-0.17 (0.333)</b>	0.29
LTR	ERV1	MER51C	0.85	0.17 (0.381)	0.64
LTR	ERV1	MER65B	0.10 (0.397)	<b>-0.49</b>	0.61
LTR	ERV1	MER65D	0.28 (0.023)	0.08 (0.777)	0.19
LTR	ERV1	MER66A	0.34	0.29 (0.090)	0.05 (0.497)
LTR	ERV1	MER84	<b>-0.15 (0.199)</b>	<b>-0.36</b>	0.21
LTR	ERV1	ERV24B_Prim-int	0.25 (0.036)	0.45	<b>-0.20</b>
LTR	ERV1	LTR10B2	<b>-0.59</b>	0.40	<b>-0.97</b>
LTR	ERVL	LTR57	0.34	<b>-0.04 (0.898)</b>	0.36
LTR	ERVL-MaLR	THE1C-int	0.36	0.57	<b>-0.19</b>
Satellite	Satellite	HSAT5	0.29 (0.035)	0.67	<b>-0.36</b>
Satellite	Satellite	MSR1	0.61	0.28 (0.160)	0.31
snRNA	snRNA	U14	0.82	0.41 (0.204)	0.36 (0.027)
snRNA	snRNA	U17	1.24	0.27 (0.439)	0.93

**Table 2.** Upregulated repetitive elements amongst the differentially expressed elements. Repetitive elements upregulated in at least one of the comparisons (LP vs C, NLP vs C or LP vs NLP) accompanied by  $\log_2$  foldchange values from all comparisons are presented in this table. Repetitive element annotations based on RepeatMasker hg19 Library are also listed. Negative  $\log_2$  foldchange values, which denoting downregulation, are written in bold. FDR values are presented in parentheses for  $\log_2$  foldchange values which did not exceed the  $FDR \leq 0.01$  cut-off. LP – lesional skin; NLP – non-lesional skin; C – control group.

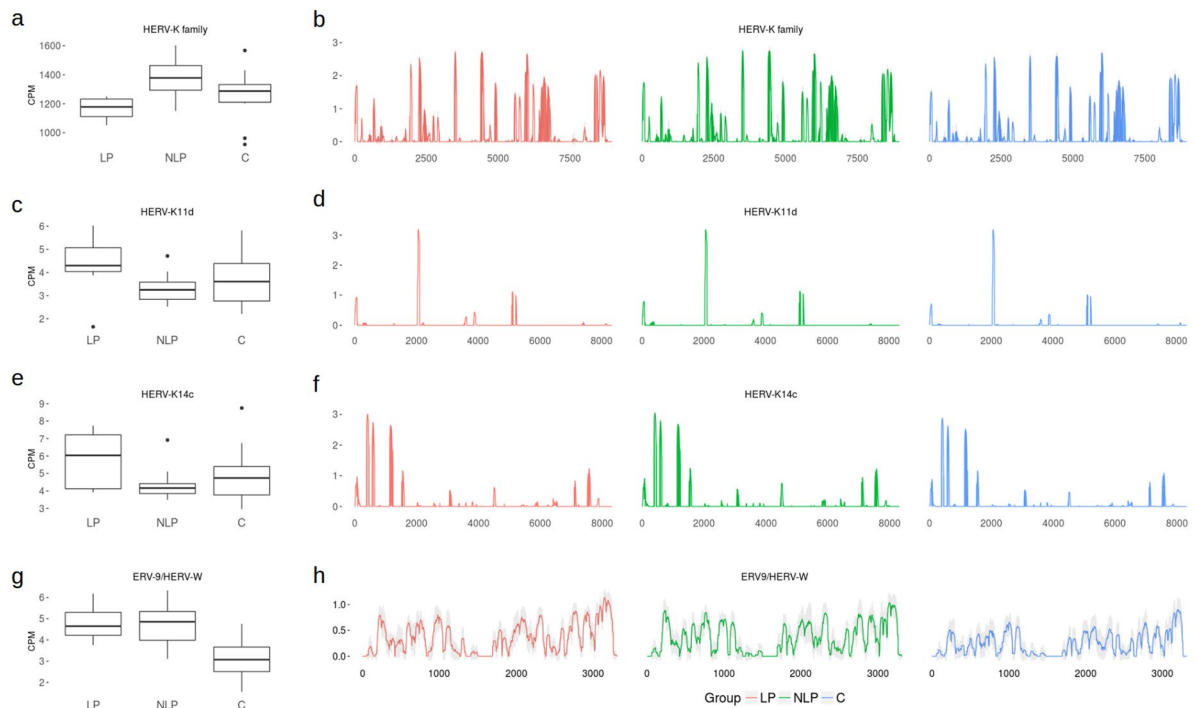


**Figure 1.** Differentially expressed elements in the lesional, non-lesional and healthy control skin. Expression levels of differentially expressed elements at  $FDR \leq 0.01$  in lesional skin (LP) vs control group or non-lesional skin (NLP) vs control group (C) comparisons are presented as a heatmap. Samples are clustered based on Euclidean distance calculated from z-score values. For every repetitive element, the mean and standard deviation based on control group CPM values were used for z-score calculation. Selection of patient traits are also presented: Psoriasis Area Severity Index (PASI) score, age, body mass index (BMI), course of the disease, sex and smoking status. Repetitive element classes are color coded. This information was not available for all control group individuals.



**Figure 2.** Differentially expressed elements in the lesional and non-lesional skin. **(a)** Distribution of foldchange values as logarithms to the base 2 ( $\log_2FC$ ) of differentially expressed elements in the pairwise comparison between lesional (LP) and non-lesional (NLP) skin of psoriasis patients. Only differentially expressed elements at  $FDR \leq 0.01$  are presented. **(b)**  $\log_2FC$  values of differentially expressed elements at  $FDR \leq 0.01$  in the pairwise comparison limited to  $\log_2FC \leq -0.5$  or  $\log_2FC \geq 0.5$  (translating into foldchange greater than 0.71 or 1.41). **(c)** PCA analysis based on counts per million (CPM) values of elements differentially expressed at  $FDR \leq 0.01$  in the LP vs C and NLP vs C comparisons. Plotted ovals represent the 95% confidence interval.

of  $\log_2FC \leq -0.5$  or  $\log_2FC \geq 0.5$  are listed in Table S7. The full listing of DE elements can be found in Table S8. Unlike in LP vs C or NLP vs C comparisons, the number of DE elements from the LTR class was larger than expected (388 out of 621). This is a significant enrichment of LTR elements ( $p$ -value = 0.002, Chi-squared test for given probabilities), given the fact that 55% (566 out of 1033) of the expressed elements were LTR elements. In line with trends emerging from the LP vs C and NLP vs C comparisons, 98% of the DE elements were down-regulated in the lesional skin tissue compared to visually healthy skin from the same patients (Fig. 2a,b). The few upregulated elements were mostly from the ERV1 family (Table 2). snRNA U17 and satellite repeat MSR1 were also upregulated in the LP group compared to the NLP group. Plotting the first two principal components (Fig. 2c) and clustering based on z-score values (Fig. S7) highlight the differences between LP and NLP groups.



**Figure 3.** Coverage of HERVs relevant in psoriasis. Comparison of counts per million (CPM) values and sequence coverage of elements regarded relevant in psoriasis. **(a)** CPM values of HERV-K elements summarized at the family level. **(b)** Coverage data from all sequence loci of HERV-K elements displayed along the length of aligned sequences (x-axis) and on y-axis as  $\log_{10}$  of read depth. **(c)** CPM values of HERV-K11D. **(d)** Coverage data from all HERV-K11D loci displayed along the length of aligned sequences (x-axis) and on y-axis as  $\log_{10}$  of read depth. **(e)** CPM values of HERV-K14C **(f)** Coverage data from all HERV-K14C loci displayed along the length of aligned sequences (x-axis) and on y-axis as  $\log_{10}$  of read depth. **(g)** CPM values of the ERV9/HERV-W element<sup>12</sup> and **(h)** coverage data displayed along the length of the sequences (x-axis) and on y-axis as  $\log_{10}$  of read depth. The study groups are marked “LP” for lesional skin on psoriasis patients, “NLP” for non-lesional skin of psoriasis patients and “C” for healthy controls.

The element family level comparison resulted in 36 downregulated families with mostly modest, but statistically significant ( $FDR \leq 0.01$ ) fold change values (Table S9).

**Coverage analysis.** We performed coverage analysis of the more relevant elements in the context of psoriasis, such as members of the HERV-K family and the novel HERV-W sequence reported by Molès and colleagues<sup>12</sup>. While the HERV-W element and HERV-K sequences showed more uniform coverage (Fig. 3a and d), the specific loci of HERV-K11D and HERV-K14C elements were erratically covered (Fig. 3b,c). This suggests that the latter sequences are not likely expressed in full length.

## Discussion

With on average 27.5% of reads aligning to repetitive element loci, the expression of repetitive elements was prominent in all the 12 lesional skin (LP), 12 non-lesional skin (NLP) and 12 healthy control (C) samples investigated. The repetitive element component of RNA-Seq data has been investigated before, as in a study of post-mortem tissue of human frontal cortex, the authors report that 8% of mappable reads originated from repetitive elements<sup>21</sup>. However, the proportion of reads aligning to repetitive elements loci in our data is notably larger, although still in line with ranges previously obtained by Cap Analysis Gene Expression (CAGE) tagging methods<sup>22,23</sup>. Similar to the results obtained by Tyekucheva and colleagues, LINE, SINE, and LTR elements were most abundantly represented in our RNA-sequencing dataset, which was as expected, since these three classes also comprise the larger part of RepeatMasker Library annotations.

The downregulation of more than half of detected LTR elements in the LP and NLP groups hints at active suppression of these elements. This is further supported by the observation that the overall variability of expression levels of repetitive elements is smaller in the LP and NLP groups. The pro-inflammatory environment of psoriasis results in an antiviral state and the increased expression of antiviral proteins, such as the cytidine deaminase APOBEC3G and phosphohydrolase SAMHD1<sup>24</sup>. In our previously published results of differential gene expression analysis based on the same dataset, we observed 1.2- to 3.4-fold upregulation of the aforementioned genes in both LP and NLP samples<sup>16</sup>.

The most prominent mechanisms responsible for suppressing LTRs and other repetitive elements are believed to be methylation and histone modifications<sup>6,25</sup>. It has been reported that the number of hypermethylated regions is increased in the lesional skin of psoriasis patients<sup>26</sup>. In our previously published gene DE analysis results<sup>16</sup>, we

observed the increased expression of DNA methyltransferases *DNMT1* and *DNMT3B*, and histone deacetylase *HDAC1*, as well as the decreased expression of histone acetyltransferase encoding genes (*EP300*, *CREBBP* and *SIRT1*) in the LP and NLP samples compared to healthy controls. The changes in expression levels of the aforementioned antiviral elements and chromatin modifying enzymes are in concordance with the gradually reduced expression of downregulated repetitive elements in the NLP and LP samples, supporting the involvement of these mechanisms in the observed downregulation of repetitive elements.

HERV-K, as the most recently active HERV family in the human genome<sup>27</sup>, has lately been actively investigated in terms of autoimmune diseases<sup>28,29</sup>. We observed that the DE elements from the HERV-K family were downregulated in the LP and NLP tissues compared to controls, as previously reported by Gupta and colleagues<sup>15</sup>. In the LP group, the downregulation was statistically significant at the element family level as well.

The background of suppressed repetitive elements raises the intriguing question of why some of the elements are upregulated in LP and NLP tissues. The activation of LTR elements can be caused by *de novo* loss of epigenetic silencing at a given locus<sup>30</sup>. Interestingly, Roberson and colleagues have reported that the loci of genes often upregulated in psoriatic lesion are hypomethylated compared to healthy skin<sup>31</sup>. This could potentially affect the transcription of nearby LTR elements as well, leading us to consider the possibility of repetitive element expression being induced by changes in the gene expression profile inherent to the pro-inflammatory environment in psoriasis. Once expressed, cytosolic retroviral nucleic acids and proteins can be detected by TLRs as PAMPs or DAMPs<sup>9,14</sup>. Thus, the increased expression of HERVs could, in turn, contribute to the already established pro-inflammatory environment.

We detected minor upregulation (1.20 to 1.34-fold) of HERV-W (HERV-17) elements in the LP and NLP tissues compared to controls. Pro-inflammatory transcription factors have been shown to have an inducing effect on HERV-W expression<sup>32</sup>. Increased expression of HERV-W elements and a novel HERV variant in psoriatic skin has been previously reported by Molès and colleagues<sup>12</sup>. In line with these results, we report the increased expression of the aforementioned HERV variant in the LP and NLP tissues. However, differing from the findings of Molès and colleagues, who reported the novel HERV variant being mostly undetectable in healthy skin, using the consensus sequence provided, we observed the expression of the HERV-W/HERV9 variant in the healthy skin of control group subjects as well.

In recent years, it has been demonstrated that the cytosolic sensing of excess amounts of extracellular DNA (cyDNA) can be mediated by antimicrobial peptides such as LL37. This can lead to a pro-inflammatory state, involving the activation of TLR pathways<sup>33,34</sup>. A recent study reports the presence cytosolic RNA:DNA duplexes in psoriatic lesions generated by endogenous reverse transcriptases, which can contribute to exceeding the tolerance for cyDNAs in keratinocytes and therefore play a role in the disease initiation<sup>35</sup>. We report a ~1.4-fold upregulation of two members of HERV-K family (HERV-K11d and HERV-K14c) encoding reverse transcriptase genes in lesional skin compared to unlesional skin of psoriasis patients. However, the results of this study suggest that these sequences are not likely expressed in full length.

In conclusion, the pro-inflammatory environment seems to result in a general suppression of repetitive elements. Conversely, against this background of suppression, a small subset of HERVs, some of which previously described in the context of psoriasis and other autoimmune diseases, are upregulated. The described differential expression of repetitive elements adds to the evidence of their potential role in the immunopathogenesis of psoriasis. This study, as well as the previous investigations, suggest a complex interplay between autoimmune processes and repetitive element expression which can not be determined from expression data alone, but calls for an integrated approach involving also inquiries into chromatin modifications, locus-specific effects of *cis*- and *trans*-regulatory elements and immunoreactivity of potentially functional endogenous retroviral proteins.

## Materials and Methods

**Patients and Controls.** The protocols and informed consent forms used in this study were approved by the Ethics Review Committee on Human Research of the University of Tartu. The participants signed an informed consent form and all the study procedures conformed to the relevant regulatory standards. The patients ( $N = 12$ ) were affected by plaque psoriasis, with PASI scores ranging from 4 to 32, and recruited in the Department of Dermatology of the Tartu University Hospital. The age- and sex-matched control group subjects ( $N = 12$ ) did not show any inflammatory skin conditions and were without a positive family history of psoriasis. Both groups consisted of unrelated Caucasians living in Estonia. For a more detailed description of individuals enrolled in this study, please refer to Keerman *et al.*<sup>16</sup>.

**RNA sequencing.** Total RNA was extracted from skin samples of patients and control group subjects obtained by 4 mm punch biopsy. From patients, both lesional (LP) and nonlesional (NLP) tissue were sampled. This resulted in a total of 36 skin biopsy samples. The cDNA libraries were sequenced on the SOLiD 5500 W platform using 75 bp single-end chemistry (Life Technologies Corp., Carlsbad, CA, USA). For the in-depth description of RNA extraction, library preparation and sequencing, please see Keerman *et al.*<sup>16</sup>.

**Read alignment and quantification.** Raw color-space reads were filtered for rRNA, active tRNA, and SOLiD adaptor sequences. For quantification purposes, the remaining reads were aligned to the hg19 reference genome while allowing multi-mapping in order to detect reads aligning to possible repeat sequences. No mapping quality cut-off was set in this stage to ensure that secondary alignments were reported in the BAM files. LifeScope software (Life Technologies, Ltd) with recommended settings, designed for color-space read alignment and analysis, was used to perform both mapping steps.

In order to connect color-space mapping with the RepEnrich pipeline<sup>18</sup>, BAM files resulting from mapping to the hg19 were parsed using samtools<sup>36</sup> and in-house perl scripts to separate uniquely and non-uniquely mapping reads. For uniquely mapping reads, only alignments with MAPQ  $\geq 10$  (in Phred scale) were retained. For

multi-mapping reads, the base-space sequence was inferred from the longest alignment and these reads were converted to FASTQ format.

RepEnrich pipeline with default parameters was applied to obtain read counts of repetitive elements, excluding simple repeats. The alignments were quantified at repeat class, family and subfamily level. Repeat subfamilies are a collection of highly similar sequences representing all known instances of a given repetitive element copies in the hg19 genome build annotated in the RepeatMasker Library. The repetitive element subfamilies as quantified units are referred to as *repetitive elements* or simply as *elements* by the authors. RepEnrich pipeline applies different quantification strategies to uniquely and multi-mapping reads in order to more accurately infer read counts, which reflect the true abundance of expressed repetitive elements.

In order to investigate the coverage of individual elements, we extracted the alignment data for all genomic instances of a given repetitive element annotated in the RepeatMasker hg19 Library. Maintaining the read alignment information, all the genomic instances were aligned using a local version of Clustal Omega v1.2.4<sup>37</sup> with default parameters. The contribution of multi-mapping reads to the coverage of a given position in the sequence was normalized against the number of alternative alignments for the read ( $1/n$ , where  $n$  is the number of alternative alignments), similar to the approach used in the RepEnrich pipeline.

**Analysis of differentially expressed repetitive elements.** The edgeR package<sup>38</sup> in R was used for the analysis of differentially expressed (DE) elements at both element and element family level. The normalization factors, which edgeR uses to normalize against RNA composition were calculated based on both repetitive element counts and counts of the coding genes to account for possible composition biases deriving from the coding part of the library as well. We used the paired sample model design for comparing LP and NLP samples as they originated from the same patient. In the groupwise comparisons of LP and NLP groups with the control group, age was used as a covariate. The edgeR quasi-likelihood pipeline was used for all analyses.

**Statistical calculations and data visualization.** The Chi-squared test and the Mann-Whitney U test were conducted using R core functions<sup>39</sup>. Ggpubr package<sup>40</sup> was used for correlation coefficient calculation and visualization. The mean and standard deviation for z-score calculations were derived from the mean control group counts per million (CPM) values for every repetitive element subfamily. The total number of mapped reads in millions was used as the denominator for calculating CPM values. For reads per kilobase of transcript per million mapped reads (RPKM) calculations, the mean locus length over all genomic instances for a given repetitive element subfamily was used as the transcript length. Principle component analysis (PCA) was conducted with prcomp method of the stats core package in R<sup>39</sup> and the results were visualized using ggbiplot package<sup>41</sup>. The pheatmap package<sup>42</sup> was used for heatmap visualization and unsupervised hierarchical clustering based on Euclidean distance. The barplots of differentially expressed elements were constructed using ggplot2<sup>43</sup>.

**Data availability.** The datasets analysed during the current study are available in the NCBI Gene Expression Omnibus (NCBI-GEO) repository. The accession number for dataset is GSE66511 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66511>). All data generated during this study are included in this published article (and its Supplementary Information files).

## References

- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Bannert, N. & Kurth, R. The Evolutionary Dynamics of Human Endogenous Retroviral Families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149–173 (2006).
- Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* **22**, 191–203 (2012).
- Ling, J. *et al.* The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. *J. Virol.* **76**, 2410–2423 (2002).
- Sokol, M., Jessen, K. M. & Pedersen, F. S. Utility of next-generation RNA-sequencing in identifying chimeric transcription involving human endogenous retroviruses. *APMIS Acta Pathol. Microbiol. Immunol. Scand.* **124**, 127–139 (2016).
- Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
- Esnault, C. *et al.* APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature* **433**, 430–433 (2005).
- Trela, M., Nelson, P. N. & Rylance, P. B. The role of molecular mimicry and other factors in the association of Human Endogenous Retroviruses and autoimmunity. *APMIS* **124**, 88–104 (2016).
- Rolland, A. *et al.* The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses. *J. Immunol. Baltim. Md 1950* **176**, 7636–7644 (2006).
- Dalen, A. B., Hellgren, L., Iversen, O.-J. & Vincent, J. A Virus-Like Particle Associated with Psoriasis. *Acta Pathol. Microbiol. Scand. Ser. B Microbiol.* **91B**, 221–229 (1983).
- Guilhous, J. J., Meynadier, J. & Clot, J. New concepts in the pathogenesis of psoriasis. *Br. J. Dermatol.* **98**, 585–592 (1978).
- Molès, J.-P., Tesniere, A. & Guilhou, J.-J. A new endogenous retroviral sequence is expressed in skin of patients with psoriasis. *Br. J. Dermatol.* **153**, 83–89 (2005).
- Foerster, J. *et al.* Haplotype sharing analysis identifies a retroviral dUTPase as candidate susceptibility gene for psoriasis. *J. Invest. Dermatol.* **124**, 99–102 (2005).
- Ariza, M.-E. & Williams, M. V. A Human Endogenous Retrovirus K dUTPase Triggers a TH1, TH17 Cytokine Response: Does It Have a Role in Psoriasis? *J. Invest. Dermatol.* **131**, 2419–2427 (2011).
- Gupta, R. *et al.* Diminished humoral responses against and reduced gene expression levels of human endogenous retrovirus-K (HERV-K) in psoriasis. *J. Transl. Med.* **12**, 256 (2014).
- Keermann, M. *et al.* Transcriptional landscape of psoriasis identifies the involvement of IL36 and IL36RN. *BMC Genomics* **16**, 322 (2015).
- Köks, S. *et al.* Psoriasis-Specific RNA Isoforms Identified by RNA-Seq Analysis of 173,446 Transcripts. *Front. Med.* **3**, 46 (2016).
- Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 583 (2014).
- Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* Chapter 4, Unit 4.10 (2009).

20. Enright, C. A., Maxwell, E. S., Eliceiri, G. L. & Sollner-Webb, B. 5'ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3. *RNA* **2**, 1094–1099 (1996).
21. Tyekucheva, S. *et al.* Establishing the baseline level of repetitive element expression in the human cortex. *BMC Genomics* **12**, 495 (2011).
22. Fadloun, A. *et al.* Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat. Struct. Mol. Biol.* **20**, 332–338 (2013).
23. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).
24. Raposo, R. A. *et al.* Antiviral gene expression in psoriasis. *J. Eur. Acad. Dermatol. Venereol. JEADV* **29**, 1951–1957 (2015).
25. Maksakova, I. A., Mager, D. L. & Reiss, D. Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. *Cell. Mol. Life Sci. CMLS* **65**, 3329–3347 (2008).
26. Zhang, P. *et al.* Whole-genome DNA methylation in skin lesions from patients with psoriasis vulgaris. *J. Autoimmun.* **41**, 17–24 (2013).
27. Marchi, E., Kanapin, A., Magiorkinis, G. & Belshaw, R. Unfixed Endogenous Retroviral Insertions in the Human Population. *J. Virol.* **88**, 9529–9537 (2014).
28. Sicat, J., Sutkowski, N. & Huber, B. T. Expression of human endogenous retrovirus HERV-K18 superantigen is elevated in juvenile rheumatoid arthritis. *J. Rheumatol.* **32**, 1821–1831 (2005).
29. Wildschutte, J. H. *et al.* Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl. Acad. Sci.* **113**, E2326–E2334 (2016).
30. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579, 1p following 579 (2010).
31. Roberson, E. D. O. *et al.* A subset of methylated CpG sites differentiate psoriatic from normal skin. *J. Invest. Dermatol.* **132**, 583–592 (2012).
32. Mameli, G. *et al.* Regulation of the syncytin-1 promoter in human astrocytes by multiple sclerosis-related cytokines. *Virology* **362**, 120–130 (2007).
33. Chiliveru, S. *et al.* Inflammatory cytokines break down intrinsic immunological tolerance of human primary keratinocytes to cytosolic DNA. *J. Immunol. Baltim. Md 1950* **192**, 2395–2404 (2014).
34. Lande, R. *et al.* Cationic antimicrobial peptides in psoriatic skin cooperate to break innate tolerance to self-DNA. *Eur. J. Immunol.* **45**, 203–213 (2015).
35. Molès, J.-P. *et al.* Cytosolic RNA:DNA Duplexes Generated by Endogenous Reverse Transcriptase Activity as Autonomous Inducers of Skin Inflammation in Psoriasis. *PLOS ONE* **12**, e0169879 (2017).
36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
37. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
38. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
39. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [www.R-project.org/](http://www.R-project.org/) (2013).
40. Kassambara, A. ggpvr: 'ggplot2' Based Publication Ready Plots. CRAN.R-project.org/package=ggpvr (2017).
41. Vu, V. Q. ggbiplot: A ggplot2 based biplot. [www.github.com/vqv/ggbiplot](http://www.github.com/vqv/ggbiplot) (2011).
42. Kolde, R. pheatmap: Pretty Heatmaps. CRAN.R-project.org/package=pheatmap (2015).
43. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, New York, United States of America. [www.ggplot2.org](http://www.ggplot2.org) (2009).

## Acknowledgements

This work was supported by personal research funding PUT1465 (Küllü Kingo), institutional research funding IUT20-46 of the Estonian Ministry of Education and Research, and the ERA Chair for Translational Genomics and Personalized Medicine at the University of Tartu.

## Author Contributions

FL. – data analysis and manuscript writing. S.K. – conceived study, data analysis, and manuscript writing. M.K. – clinical sample collection and clinical expertise. E.R. and E.P. – library preparation and sequencing. K.A. and H.S. – clinical experts. G.K. – data analysis. K.K. – clinical expert, lead of the clinical group, and conceived the study.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-22734-7>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018