*Original Research*  **Feature article**

# Regulation of expression quantitative trait loci by SVA retrotransposons within the major histocompatibility complex

**Jerzy K Kulski[1,2], Abigail L Pfaff[3,4], Luke D Marney[5], Alexander Fröhlich[5] (iD), Vivien J Bubb[5] (iD), John P Quinn[5] and Sulev Koks[3,4] (iD)**

[1]Department of Molecular Life Sciences, School of Medicine, Tokai University, Isehara, Kanagawa 259–1193, Japan; [2]Health and Medical Science. Division of Immunology and Microbiology, School of Biomedical Sciences, The University of Western Australia, Nedlands, WA 6009, Australia; [3]Perron Institute for Neurological and Translational Science, Perth, WA 6009, Australia; [4]Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Perth, WA 6150, Australia; [5]Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, L69 3BX, UK
Corresponding author: Sulev Koks. Email: sulev.koks@perron.uwa.edu.au

## Impact Statement

The genomic region of the major histocompatibility complex (MHC) harbors more than 160 genes involved in innate and adaptive immunity, inflammatory responses and complement functions. There are more than 18 SINE-VNTR-Alu (SVA) retrotransposons within the MHC region, but little or nothing is known about their function or influence on MHC gene expression. Our study shows that eight of these SVA have regulatory effects on gene expression. The regulatory properties of SVAs located within the class I and class II MHC regions were associated with differential co-expression of 71 different genes within and 75 genes outside the MHC region as multilocus expression units in what appears to be a highly coordinated network of transcriptional activity. Thus, future genetic engineering of these SVA elements could be a way to manage various genes in the MHC and the immune response in a coordinated and systematic way not previously envisioned.

## Abstract

Genomic and transcriptomic studies of expression quantitative trait loci (eQTL) revealed that SINE-VNTR-Alu (SVA) retrotransposon insertion polymorphisms (RIPs) within human genomes markedly affect the co-expression of many coding and noncoding genes by coordinated regulatory processes. This study examined the polymorphic SVA modulation of gene co-expression within the major histocompatibility complex (MHC) genomic region where more than 160 coding genes are involved in innate and adaptive immunity. We characterized the modulation of SVA RIPs utilizing the genomic and transcriptomic sequencing data obtained from whole blood of 1266 individuals in the Parkinson's Progression Markers Initiative (PPMI) cohort that included an analysis of human leukocyte antigen (*HLA*)-*A* regulation in a subpopulation of the cohort. The regulatory properties of eight SVAs located within the class I and class II MHC regions were associated with differential co-expression of 71 different genes within and 75 genes outside the MHC region. Some of the same genes were affected by two or more different SVA. Five SVA are annotated in the human genomic reference sequence GRCh38.p14/hg38, whereas the other three were novel insertions within individuals. We also examined and found distinct structural effects (long and short variants and the CT internal variants) for one of the SVA (*R_SVA_24*) insertions on the differential expression of the *HLA-A* gene within a subpopulation (550 individuals) of the PPMI cohort. This is the first time that many HLA and non-HLA genes (multilocus expression units) and splicing mechanisms have been shown to be regulated by eight structurally polymorphic SVA within the MHC genomic region by applying precise statistical analysis of RNA data derived from the blood samples of a human cohort population. This study shows that SVA within the MHC region are important regulators or rheostats of gene co-expression that might have potential roles in diversity, health, and disease.

## Introduction

Human leukocyte antigen (HLA) class I and class II molecules (antigens) are polymorphic cell-membrane-bound glycoproteins that regulate the immune response by presenting peptides of fragmented proteins to circulating cytotoxic and helper T lymphocytes, respectively. The HLA molecules are investigated continuously because of their importance in the regulation of the innate and adaptive immune system, autoimmunity, infectious diseases, and transplantation outcomes.[1–4] The genomic region that encodes these HLA molecules is known as the major histocompatibility complex (MHC) on the short arm of chromosome 6 at 6p21.3 and it encompasses approximately 160 coding genes within ~ 3–4 MB including three distinct structural regions: class I with the classical and nonclassical HLA class I genes and ~39

non-HLA genes; class II with the classical and nonclassical HLA class II genes and some proteasome-processing and peptide antigen transportation non-HLA genes; and class III that harbors more than 60 non-HLA genes including those involved in stress response, complement cascade, immune regulation, inflammation, leukocyte maturation, and regulation of T cell development and differentiation.[3–5] Of the many HLA-like genes, 18 HLA class I genes (six protein-coding genes and 12 pseudogenes) and seven MHC class I chain-related (MIC) genes (two protein-coding genes and five pseudogenes) are in the HLA class I region, and 18 HLA class II genes (13 protein-coding genes and five pseudogenes) are in the HLA class II region. The classical HLA class I genes, *HLA-A, -B*, and *-C*, and the classical HLA class II genes, *HLA-DR, -DQ* and *-DP*, are characterized by their extraordinary large number of polymorphisms, whereas the non-classical HLA class I genes, *HLA-E, -F*, and *-G*, are differentiated by their tissue-specific expression and limited polymorphism.[6,7]

The MHC genomic region contains numerous coding and noncoding gene sequence duplications, insertions, and deletions and considerable sequence diversity or polymorphisms[8,9] that have accumulated by ancestral descent and recombined into thousands of distinct multilocus haplotypes at variable worldwide population frequencies.[5,10,11] The two most frequent (>1%) European HLA ancestral haplotypes 8.1AH (*HLA-A\*01:C\*07:B\*08:DRB1\*03:DQB1\*02*) and 7.1AH (*HLA-A\*03:C\*07:B\*07:DRB1\*15:DQB1\*06*)[12,13] were estimated to have diverged from their common ancestors at least 23,500 years ago.[14] Polymorphic MHC multilocus haplotypes have been associated with Parkinson's disease (PD)[15–17] and autoimmune diseases such as ankylosing spondylitis, type 1 diabetes mellitus, multiple sclerosis, and rheumatoid arthritis.[18] Polymorphic frozen blocks[1] or single nucleotide polymorphism-linkage disequilibrium (SNP-LD) blocks that were identified by LD or long-range haplotype and extended haplotype homozygosity tests[19] of the MHC genomic regions include a variety of genotyped microsatellites, SNPs, and indels.[2] The absence and presence of a variety of structurally dimorphic retroelements within the MHC, such as Alu, SVA, long terminal repeats (LTR), and human endogenous viruses (HERVs), have been used as haplotypic markers and for phylogenetic analyses of population ancestral relationships and diversity.[20–24]

The investigation of MHC multilocus haplotypes and disease associations is complicated by the regulatory network of *cis*-acting multilocus expression units known as haplotype-specific expression quantitative trait loci (eQTL).[2,25,26] These eQTL are controlled by an array of retroelements and DNA transposons,[27] both as binding sites for transcription factors[28] and as sources of regulatory noncoding RNAs.[29,30] In this regard, eQTL studies associate genomic and transcriptomic data together from the same individuals to identify loci that affect mRNA expression by linking SNPs to changes in gene expression, which, in turn can be applied as a useful procedure for uncovering important patterns of gene regulation and genome-wide association study variants.[31,32] Studies using homozygous cell-lines and/or biological samples have demonstrated that the expression of various clusters of genes inside or outside the MHC genomic region can be affected

differentially by different haplotypes.[2,33] Lam *et al.*[26] used haplotypic RNA and DNA-sequencing data to show that haplotype sequence variations represented by eQTL SNP alleles can function as *cis*-acting regulatory variants for multiple MHC genes, and that these eQTL SNPs were localized especially within four segmental regions containing *HLA-A* (alpha block), *HLA-C* (beta block), *C4A* (gamma block), and *HLA-DRB1* (delta block). In a study of eQTL within the genomic sequences of lymphoblastoid cell lines, Spirito *et al.*[34] found that the chromosomal location *6p21.32*, which includes the MHC class I and II regions, was a particularly enriched genomic region where structurally polymorphic transposable elements (TEs) influenced gene expression. Taken together, these pioneering eQTL studies incorporating either HLA haplotypes or SNPs are a powerful new approach to identifying causal genetic mechanisms underlying disease associations both inside and outside the MHC region.

Apart from the protein coding genes, pseudogenes, non-coding transcribed RNA, and small nucleolar transcribed RNAs loci, there are at least 8604 repeat elements, including those known as TEs and/or retroelements, and 723 simple repeats (microsatellites) within the MHC of the human genomic reference sequence GRCh38.p14/hg38.[5] The MHC of this genomic reference sequence is represented by the 7.1AH derived mostly from the DNA sequencing data of the immortalized PGF homozygous lymphocytic cell-line.[8] Retroelements, which are a class I subcategory of TEs, consist of various subfamilies including HERVs, LTRs, long interspersed elements (LINE 1, LINE 2), short interspersed elements (SINE 1, SINE 2), and SINE-R-VNTR-*Alu* (SVA).[35] Furthermore, retrotransposon insertion polymorphisms (RIPs)[36] are either absent or present in the genome as insertions or deletions (indels) depending on their state of origin or ancestral state.[37] RIPs or indels have been associated strongly with the regulation of gene expression in health and disease.[27,38] In addition, eQTL of non-LTR retroelements such as SINEs, LINEs, and SVAs and their affect on the regulation of gene expression were identified and described for a Parkinson's Progression Markers Initiative (PPMI) cohort using whole genome sequence and transcriptome data obtained from the blood of hundreds of individuals.[27]

At least 18 SVA polymorphic insertions were described previously within the MHC class I, II, and III regions; some were haplotypic or haplospecific for particular HLA gene alleles that varied in frequency between European, Japanese, and African American populations.[39] For example, the *SVA-HF*, *SVA-HA*, and *SVA-HC* were inserted at a relatively low frequency (<0.2) in European (Caucasian) populations and strongly associated with 7.1AH, but not with the 8.1 haplotype.[20,21]

In this study, we identified the eQTLs associated with various SVA RIPs within the MHC genomic region of 1266 individuals within the PPMI cohort. We re-mined and added new transcriptome data to the previous analyses of SVA[27,40] and focused on the presence and roles of eight distinct SVA RIPs within the MHC genomic region and their eQTL effects on HLA and non-HLA genes within (*cis*) and outside (*trans*) the MHC. We also examined and found distinct structural effects (long and short variants and the CT internal variants) for one of the SVA (*R_SVA_24*) insertions on the differential
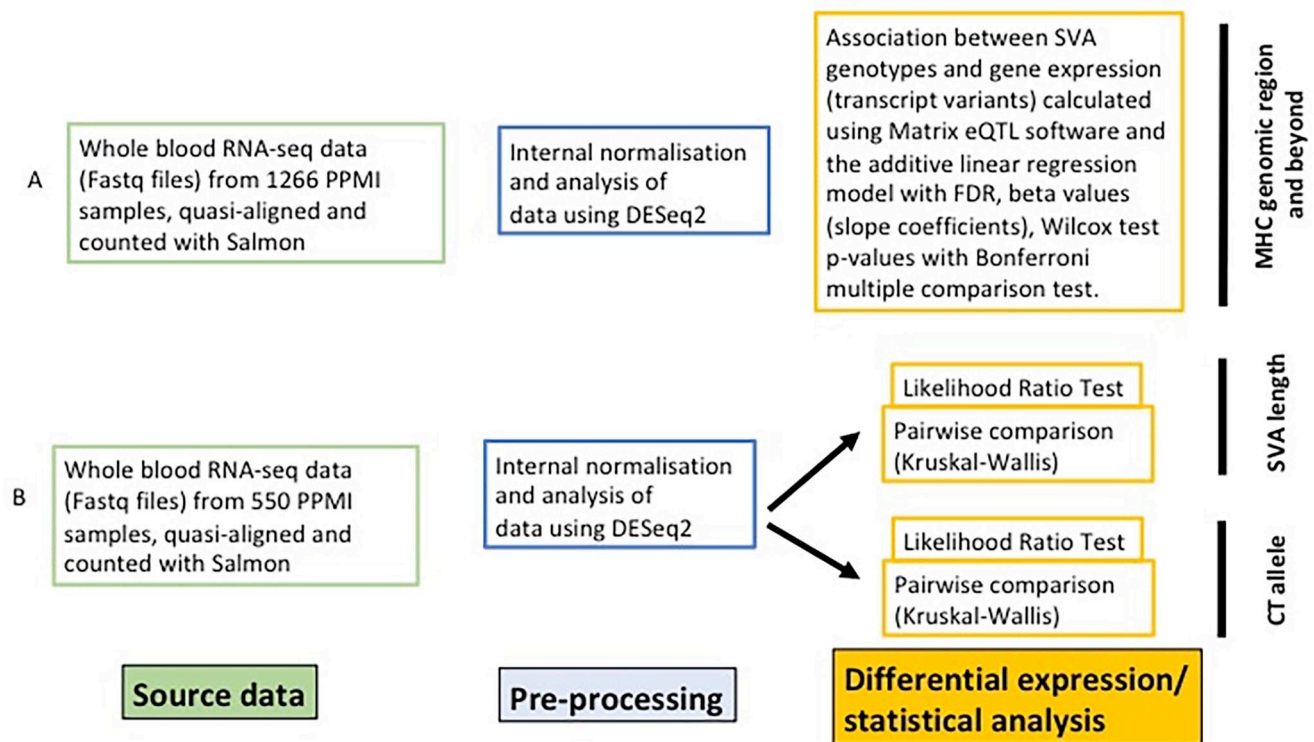
**Figure 1.** Methodology Workflow. (A) Whole blood RNA-seq data obtained from 1266 PPMI individuals were quasi-aligned and counted with Salmon. The counts were normalized and analyzed using DESeq2. The effect of SVA RIP genotypes on the expression of genes inside and outside the MHC genomic region were evaluated using the Matrix eQTL software with inbuilt statistical tests including the false discovery rate (FDR). (B) Whole blood RNA-seq data obtained from 550 individuals of the PPMI cohort were quasi-aligned and counted with Salmon, normalized and analyzed using DESeq2. Differential expression of HLA-A transcripts was analyzed via likelihood ratio tests and multiple comparisons (Kruskal–Wallis), stratified by R_SVA_24 length and CT allele genotype data.

expression of the *HLA-A* gene within a subpopulation (550 individuals) of the PPMI cohort.

## Materials and methods

### Datasets

In this study of SVA regulation of HLA and non-HLA genes within the MHC genomic region, we utilized the data of 1266 individuals within the PPMI cohort without assessing any differences between the cases and controls. PPMI cohort data[41] were downloaded from http://www.ppmi-info.org/data (accessed on 19 January 2021) as previously described.[27,40] The dataset contains whole transcriptome data from blood together with genetic and clinical data of 1266 individuals whose race was reported as white. All subjects, including 678 PD cases and 588 controls, were combined as a single study population for the current SVA and eQTL analyses (Figure 1(A)). The structural variant caller Delly2 (https://github.com/dellytools/delly)[42] was used with default settings to genotype structural variants from the PPMI cohort. The structural variants located within the coordinates of the human-specific SVAs and available on the UCSC genome browser (https://genome.ucsc.edu/) were extracted with ±100bp. Structural variants classified as deletions in these regions were inspected manually to determine if the break points were consistent with the reported coordinates of the reference human-specific SVAs. Gene association analysis of SVA RIPs was performed using logistic

regression in PLINK (v1.07)[43] with data from both the control and PD samples, and *P*-values adjusted for multiple testing (Bonferroni correction). In addition to reference RIPs, non-reference RIPs that presented in an individual's genome, but not the genome reference, were included in this study as non-reference SVA (NR_SVA). The read length of the whole-genome-sequencing data was 150bp, and, on average, there were 837 million reads per genome, with the coverage more than 30×. RNA-seq reads were also 150bp in length, and average number of reads per sample was 31 million.

Fastq files of whole-blood RNAseq data were downloaded from the PPMI database and mapped to hg19 (GRCh37) by STAR using GENCODE v19. The Salmon software[44] was used to call transcript counts from the reads in conjunction with the index and decoy based on the reference genome hg 38. Salmon-generated quant files were imported into R using *tximport* function from the tximport package of R. Raw counts were extracted with the *DESeqDataSetFromTximport* function and normalized using the median-of-ratios method implemented in the DESeq2 package. The counts were divided by sample-specific size factors determined by the median ratio of gene counts relative to the geometric mean per gene. PD cases and controls were combined, and transcript expression signals were tabulated after importing Salmon files in the R format to prepare them for the eQTL analysis. Altogether, 236,186 transcripts previously described were used for the initial analysis in combination with all the identified TEs, including the eight SVA within the MHC genomic region.[27]

**Table 1.** Ten SVA RIPs within or flanking the MHC genomic region (see Figure 2).

| SVA ID | SVA | SVA start and end position | Subfamily | TE width | DNA | MHC block | TE insert | Genes flanking TE indels |
|--------|-----|----------------------------|-----------|----------|-----|-----------|-----------|--------------------------|
| | Alias | GRCh38/hg38, chr6 | | bp | Strand | Region | Site | |
| NR_SVA_377 | | 29731783 | | | | Alpha cI | L2b | HLA-F-AS1 intron 1 |
| R_SVA_24 | SVA-HA | 29932088–29933753 | SVA_A | 1666 | pos | Alpha cI | ERV3-16A3 | HCG4B and HLA-A |
| R_SVA_25 | SVA-HC | 31243861–31245322 | SVA_F | 1462 | pos | Beta cI | MLT2C2/ ERVL | Psors1 C3/HCG27 and HLA-C |
| R_SVA_26 | SVA-MIC | 31453746–31456553 | SVA_D | 2808 | neg | Beta cI | L1M4 | MICA and HCP5 |
| NR_SVA_380 | | 32546835 | | | | Delta cII | L1M5 | DRB5 and DRB6 |
| R_SVA_27 | SVA-DRB1 | 32594194–32596780 | SVA_F | 2587 | neg | Delta cII | | DRB1 and DQA1 |
| R_SVA_85 | SVA-DPA1 | 33058946–33060797 | SVA-D | 1852 | neg | Epsilon cII | | DOA and DPA1 |
| NR_SVA_381 | | 33062533 | | | | Epsilon cII | L1MEc | DOA and DPA1 |
| R_SVA_28 | | 33525379–33526168 | SVA_F | 790 | pos | Outside | MER21B | ZBTB9 and BAK1 |
| NR_SVA_382 | | 33535745 | | | | Outside | L1ME5 | ZBTB9 and BAK1 |

SVA: SINE-VNTR-Alu; TE: transposable elements; MHC: major histocompatibility complex.
NB. Some SVA are inserted within ancient TE fragments such as L2, L1MEc, L1M4, L1M5 and ERV3_16A3 internal fragment. SVA TE aliases reported by Kulski *et al.*[20,21,39] R_SVA is the reference genome GRCh38/hg38, whereas NR_SVA is absent in GRCh38/hg38. The top eight SVA are within the MHC genomic region as indicated by the column showing the MHC block region with their presence in the alpha and beta class I (cI) blocks, and in the delta and epsilon class II (cII) blocks. R_SVA_28 and NR_SVA_382 are inserted "outside" the MHC genomic region on the centromeric side of the extended class II (cII) region (Figure 2).

## eQTL analysis

Matrix eQTL was used to calculate the genetic loci of SVA regulating the expression transcript variants.[45] The additive linear model was used with covariates, age, and sex, with the threshold of the FDR set at 0.05. During eQTL analysis, local (*cis*) and distant (*trans*) quantitative loci were called, but ultimately this depended on whether the loci were within the boundaries of the MHC genomic region ranging between the telomeric *HLA-F* and the centromeric *COL11A2* loci. Matrix eQTL reported effect-size estimates as *beta* values or slope coefficients. The correction for multiple testing of eQTL was performed using FDR, and only the results that remained significant after FDR correction are reported here. For pairwise comparisons between the genotype, a Wilcoxon test was used, and *P*-values were challenged with the Bonferroni multiple comparison test.

## SVA_24 structural polymorphisms and differential expression of *HLA-A*

A sample group of 550 individuals from the PPMI cohort was used for this study to determine which SVA structural characteristics affect the expression of HLA-A variants. Whole blood RNA-seq data obtained from 371 PD patients and 179 healthy controls were normalized and filtered for batch effects/quality control using open-source statistical programming platform, R, and the DESeq2 analysis package.[27] Count data were compared between long and short SVAs, as well as CT alleles A and B. Differential expression of *HLA-A* transcripts was analyzed via likelihood ratio tests and nonparametric multiple comparisons (Kruskal–Wallis), stratified by SVA length (long L, short S or absent A) and CT allele genotype data as outlined in Figure 1(B). Kruskal–Wallis (one-way analysis of variance) tests were carried out using the R package "ggpubr."

## Results

### Features of eight SVA RIPs within the MHC genomic region

The genomic locations and main characteristics of 10 SVA RIPs within or flanking the MHC genomic region that were identified within the MHC genomic region of 1266 individuals from the PPMI cohort database are listed in Table 1. These include SVA IDs, aliases, start and end positions within the hg38 genomic reference sequence, the subfamilies, sizes, and the flanking genes. The SVA insertions labeled as R_SVA are located within the hg38 genomic reference sequence at UCSC browser (https://genome.ucsc.edu), whereas the NR_SVA are present in PPMI individuals but are absent from the hg38 genomic reference sequence. Five of the R_SVA locations listed in Table 1 as TE aliases, *SVA-HA*, *SVA-HC*, *SVA-MIC*, *SVA-DRB1*, and *SVA-DPA1* were identified in previous studies.[20,21,39]

The locations of the eight SVA retrotransposon polymorphic integrations within the MHC genomic region from the *HLA-F* gene in class I to the *HLA-DPA3* gene in class II are shown in Figure 2. Four of the SVA RIPs are within the MHC class I region and four are within the class II region relative to 297 known coding and noncoding gene loci within these two regions and the MHC class III region. We also included two SVA insertions, *R_SVA_28* and *NR_SVA_382*, located between the *ZBTB9* and *BAK1* genes that are ~48 kb outside the centromeric end of the MHC class II region for comparison of the effects of MHC and non-MHC SVA insertions on MHC gene expression. The MHC class I genomic region is subdivided into three HLA duplicated and regulatory regions or polymorphic frozen blocks, known as the alpha block with *HLA-A*, the kappa block with *HLA-E*, and the beta block with *HLA-C*, *HLA-B*, and the two MIC genes, *MICA* and *MICB*.
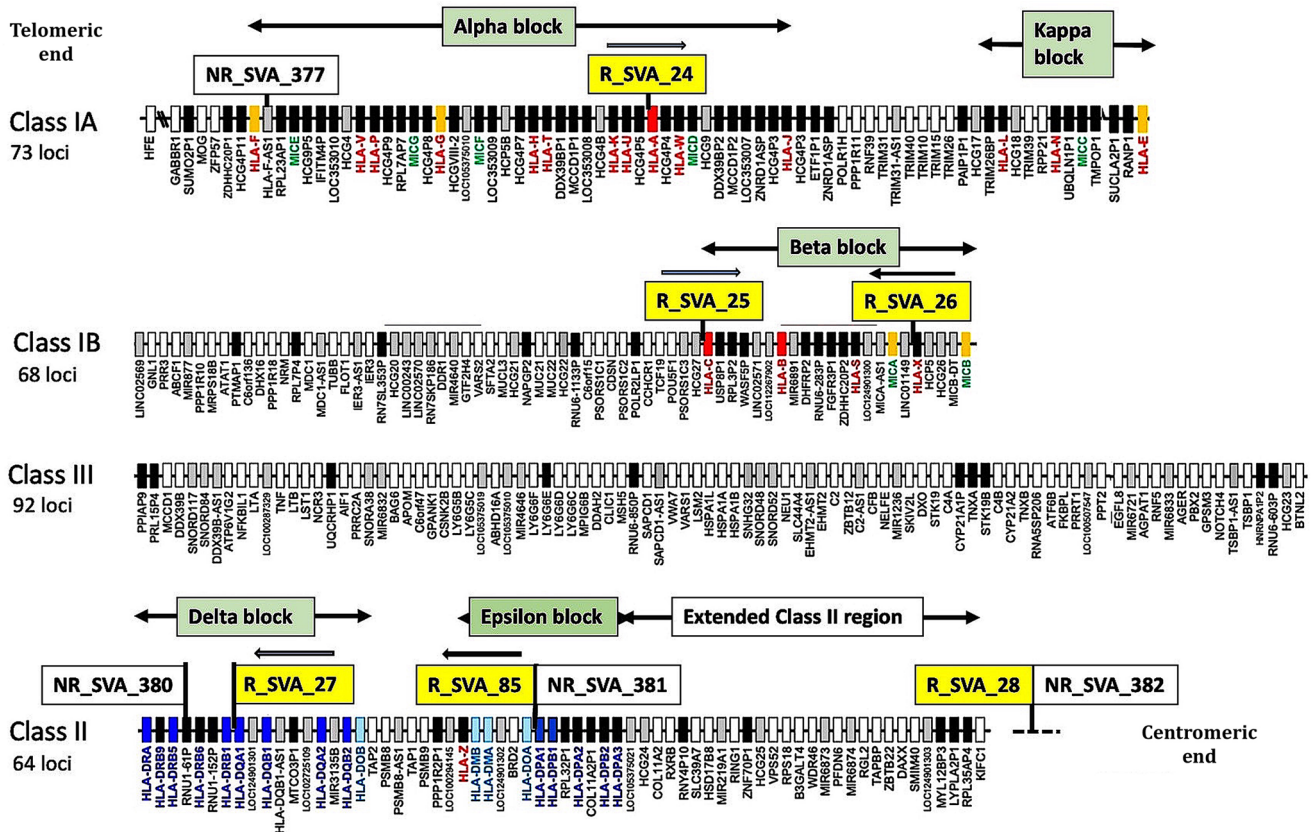
**Figure 2.** Gene map of MHC genomic region showing the clusters of coding and noncoding gene loci in the class I, class III, and class II regions. The alpha, kappa, and beta blocks within the class I genomic region and the delta and beta blocks of the class II region are labeled in green boxes with horizontal double arrows showing the spread of genes within each block. The extended class II region is indicated by the white box. Positions of the eight SVA integrations within the MHC genomic region are indicated by labeled boxes over black vertical lines. Two SVA insertions that are located between the ZBTB9 and BAK1 genes ~48 kb from KIFC1 at the centromeric end are shown slightly outside the extended MHC class II region. The gene map corresponds to the genomic coordinates of 29,602,228 (GABBR1) to 33,410,226 (KIFC1) on chromosome 6 in the human genome GRCh38.p13 primary assembly of the NCBI map viewer. White or colored (orange, red, and blue) vertical boxes, gray, and black vertical boxes show protein-coding genes, noncoding RNAs (ncRNAs), and pseudogenes, respectively. Red, green, and blue letters indicate HLA class I, MIC, and class II genes, respectively.
Source: Adapted from Shiina *et al*.[3,4]

**Table 2.** SVA genotype frequencies and number of *cis* and *trans* genes and transcripts affected in eQTL analysis of 1266 PD cases and controls.

| SVA ID | n | SVA genotype frequencies | | | | Number of cis and trans genes affected | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | AA | AP | PP | AP, PP | cis Genes | cis RNA variants | trans Genes | trans RNA variants |
| NR_SVA_377 | 1266 | 0.88 | 0.11 | 0.01 | 0.12 | 14 | 28 | 30 | 30 |
| R_SVA_24 | 1231 | 0.61 | 0.30 | 0.20 | 0.40 | 24 | 47 | 3 | 3 |
| R_SVA_25 | 1258 | 0.83 | 0.16 | 0.01 | 0.17 | 34 | 72 | 39 | 39 |
| R_SVA_26 | 1262 | 0.33 | 0.47 | 0.21 | 0.6 | 17 | 31 | 0 | 0 |
| NR_SVA_380 | 1265 | 0.74 | 0.2 | 0.02 | 0.26 | 22 | 40 | 6 | 6 |
| R_SVA_27 | 1261 | 0.76 | 0.23 | 0.01 | 0.24 | 23 | 48 | 4 | 4 |
| R_SVA_85 | 1221 | 0.02 | 0.30 | 0.68 | 0.98 | 5 | 8 | 2 | 2 |
| NR_SVA_381 | 1266 | 0.64 | 0.32 | 0.05 | 0.36 | 6 | 10 | 0 | 0 |
| R_SVA_28 | 1264 | 0.00 | 0.03 | 0.96 | 0.99 | 1 | 1 | 200 | 214 |
| NR_SVA_382 | 1266 | 0.97 | 0.03 | 0.03 | 0.07 | 1 | 1 | 451 | 488 |

SVA: SINE-VNTR-Alu; eQTL: expression quantitative trait loci; PD: Parkinson's disease; AA: totally absent; AP: absent and present; PP: totally present.
The SVA genotypes are AA, AP, and PP.

## *Cis* eQTL effects and genotype frequencies of SVAs

The SVA genotype frequencies and the number of *cis* and *trans* genes and transcripts significantly upregulated or downregulated in the eQTL analysis of 1266 individuals are shown in Table 2. The frequency of SVA insertions within the MHC of individuals varied between 0.116 for *NR_SVA_377* and 0.980 for *R_SVA_85*. The homozygous SVA insertion (genotype AA) frequency is substantially lower than for the

heterozygous SVA insertion (genotype AP) frequency for all the SVA integrations within the MHC except for *R_SVA_85*, which at 99% integration is close to complete fixation with the human population, while the other MHC SVA are still in the process of balancing selection or evolution toward fixity or homozygosity by purifying selection. Both *R_SVA_85* and *NR_SVA_381* that are inserted between the *HLA-DOA* and *-DPA1* genes within the epsilon block of the MHC class II region influenced fewer genes (five and six, respectively) than the other SVA insertions within either class I or class II regions, which affected between 14 and 34 *cis* genes. The two SVA insertions, *NR_SVA_380* and *R_SVA_27* near the *HLA-DRB1* gene, modified the transcription of 22 and 23 *cis* genes, respectively. *R_SVA_25*, which is located 23 kb telomeric of the *HLA-C* gene within the beta block of the class I region, coordinately modified the expression of 38 *cis* genes and 72 RNA variants within and between the different MHC polymorphic blocks including genes in the alpha block (*HLA-A, -H, -K, -F*), kappa block (*HLA-L*), beta block (*HLA-B, HLA-C, MICA, MICB*), and delta block (*HLA-DRB5, -DRB1, -DQB1, -DQB2*), and non-HLA genes in the intergenic regions located between the blocks and within the class III region (Supplemental Tables 1 and 2).

Supplemental Table 1 provides all the *cis* and *trans* genes and transcripts regulated by the MHC SVA. The type of SVA subfamily is correlated weakly with the frequency of the SVA insertions within individuals. However, the youngest SVA_F subfamily members are at a lower frequency than the older SVA-A and SVA_D subfamily members (Table 1).

The statistically significant effects (*P*-values, FDRs and beta values) of the MHC class I and class II SVAs on the *cis* gene transcripts within the class I, class II, and class III regions are shown in Supplementary Table 2. The beta value, generated from linear regression model in Matrix eQTL package, represents the change in mRNA expressions with positive (upregulation) or negative (downregulation) associations.

The location of the class I and class II SVA insertions, respectively, and the relative positions of the genes that are regulated by these SVA are shown in Figures 3 and 4. For example, Figure 3(A) shows that *NR_SVA_377* regulated four genes centromeric and nine genes telomeric of its insertion site in the region of the *HLA-F AS1* gene within the alpha block. Moreover, the regulation by *SVA_377* was limited mostly to the alpha block and its telomeric extended region (11 of 14 genes). In contrast, *R_SVA_24* modified the transcription of 23 genes with a broader scope both within and between the blocks ranging from 12 genes in the alpha block to two genes in the delta block (Figure 3(B)). Whereas the class II SVA, *NR_SVA_380* and *R_SVA_27*, had a broader reach on genes from the alpha to the epsilon block including the class III region, the influence of *R_SVA_85* and *NR_SVA_381* on gene transcription was limited mostly to the delta and epsilon blocks of the class II region (Figure 4(A) to (D)).

There are approximately 62 coding and noncoding non-HLA genes located between the MHC alpha and kappa blocks, and the kappa and beta blocks. *SVA_24, SVA_25* and
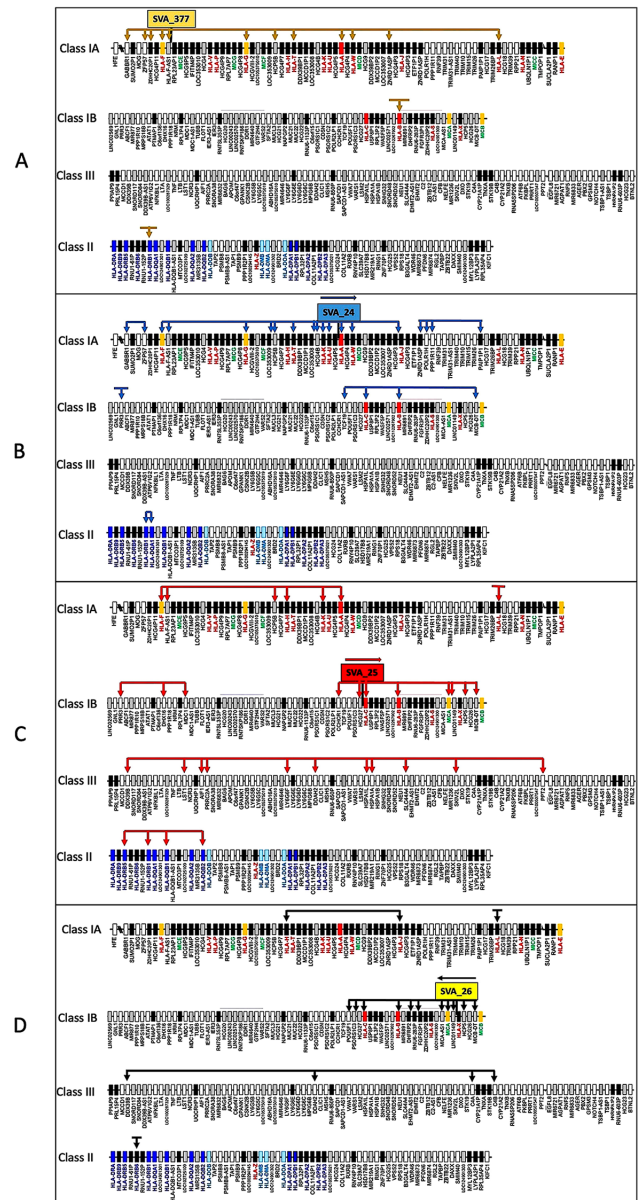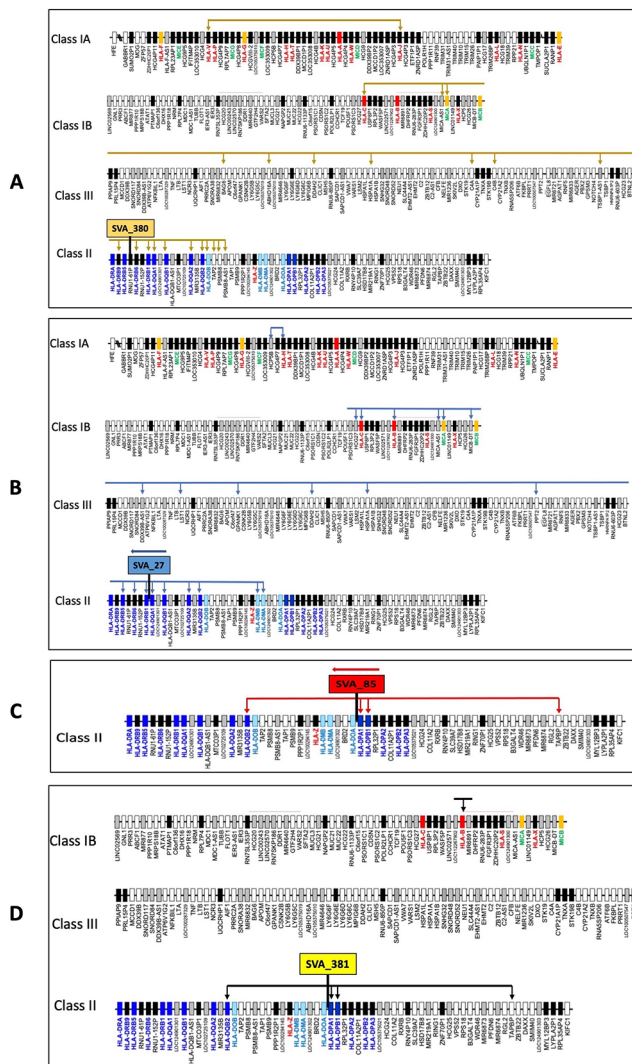


**Figure 3.** The *cis* eQTL modulated by the four polymorphic SVA inserted within the MHC class I genomic region. Positions of SVA integration are indicated by labeled, colored boxes over black vertical lines. Horizontal arrows above the SVA boxes indicate the 5′ to 3′ direction of the retrotransposons. Genes modulated by SVA_377, SVA_24, SVA_25, and SVA_26 are indicated by (A) orange/yellow, (B) blue, (C) red, and (D) black vertical arrows, respectively.

*SVA_26* affected 13 of the non-HLA genes (Figure 3), further highlighting the SVA regulatory role to interconnect the multilocus expression of different HLA and non-HLA genes within the MHC. Despite a separation of over one megabase of nucleotide sequence, two of the four SVA in the alpha block (*SVA_25* and *SVA_26*) and two of the four SVA in the delta block (*NR_SVA_380* and *R_SVA_27*) coregulated some of the non-HLA genes in the class III region including the *complement factor* genes *C4A* and *C4B*, the *spliceosome RNA helicase* gene *DDX39B*, the *leukocyte-specific transcript 1 protein* gene *LST1*, a *dimethylarginine dimethylaminohydrolase* gene

**Figure 4.** The *cis* eQTL modulated by the four polymorphic SVA inserted within the MHC class II genomic region. Positions of SVA integration are indicated by labeled, colored boxes over black vertical lines. Horizontal arrows above the SVA boxes indicate the 5′ to 3′ direction of the retrotransposons. Genes modulated by SVA_380, SVA_27, SVA_85, and SVA_381 are indicated by (A) orange/yellow, (B) blue, (C) red, and (D) black vertical arrows, respectively.

*DDAH2* involved in nitrous oxide regulation, the *heat shock protein 70* gene *HSPA1A*, and the *lysosomal palmitoyl-protein thioesterase 2* gene *PPT2*. All these gene products have roles in the inflammatory response and may act to mitigate autoimmune disease. For example, *DDX39B* is an RNA–DNA helicase with known functions in mRNA splicing and nuclear export that regulates inflammatory responses,[46,47] and interacts with influenza A viral proteins[48] and the pattern recognition receptor pathway that affects sensitivity to DNA damaging chemicals.[49] *LST1* regulates leukocyte abundance in lymphoid organs and inflammatory response in the gut, and the gene exons undergo extensive alternative splicing, giving rise to both membrane-bound (encoded by exon 3) and soluble isoforms.[50]

The non-HLA genes in the MHC class I region between the kappa and beta blocks impacted by the MHC class I SVA

were *POLR1H, PPP1R11, PAIP1P1, PRR3, DHX16, MDC1, CCHCR1, POU5F1,* and *TCF19* (Figure 2). Most of the non-HLA gene transcription levels modified by the MHC class I SVAs are involved with various mRNA and protein activities and processes such as transcription factors, RNA polymerase, ubiquitin-protein ligase, helicases, and RNA and DNA binding (Supplemental Table 3). The MHC class II SVAs had no significant influence on these genes.

## SVA effects on *trans* eQTL

The different gene transcripts and chromosomes that were affected outside (*trans*) the MHC genomic region by the eight MHC polymorphic SVA are listed in Supplemental Table 4. Six of the eight SVA had eQTL *trans* effects. Only the regulatory effects of the *R_SVA_26* insertion within class I and the *NR_SVA_381* insertion within class II seemed to be limited totally to the MHC genomic region. The other SVA insertions within the MHC had *trans* effects involving all chromosomes except chromosomes 9 and 18. *NR_SVA_377* and *R_SVA_25* had the greatest number of *trans* results. For comparison, we included *R_SVA_28* and *NR_SVA_38* that are located outside the MHC between the *ZBTB9* and *BAK1* genes 48 kb from *KIFC1* at the telomeric end of the extended MHC class II region (Table 2, Figure 2). *ZBTB9* enables sequence-specific DNA binding and transcription protein binding activity and is involved in the regulation of transcription by RNA polymerase II.[51] *BAK1* (*BCL2 Antagonist/Killer 1*) is a member of the BCL2 protein family and functions as a pro-apoptotic regular involved in a wide variety of cellular activities.[52] *R_SVA_28* and *NR_SVA_382* each effected one MHC gene (*TAP2* and *HLA-DPB2*) and 200 and 451 genes, respectively, outside the MHC (Supplemental Table 1). Most SVA *trans* effects were split between coding and noncoding genes, but a biological significant number of coding changes appear to involve neurological functions and development (Supplemental Table 1).

Of the 85 gene transcripts impacted outside the MHC, 57 were expressed by known coding genes, six novel transcripts, nine long intergenic non-protein coding RNA, four antisense RNA, and nine pseudogenes. Among the coding genes were the five immunity genes, *immunoglobulin heavy variable 2-26 (IGHV2-26), immunoglobulin lambda variable 9-49 (IGLV9-49), T cell receptor beta variable 7-3 (TRBV7-3), T cell receptor alpha variable 20 (TRAV20),* and *T cell receptor alpha variable 22 (TRAV22),* and *interleukin 1 receptor type 2 (IL1R2).* Cytidine/uridine monophosphate kinase 2 is a rheostat for macrophage homeostasis, ubiquitin-specific peptidase 13 regulates HMGB1 DNA binding stability and is implicated in multiple diseases, whereas calnexin is a chaperone protecting protein structure, and the heat shock protein 90 beta family member 1 protects against oxidative stress and has a regulatory role in neurodevelopment. On the contrary, Fas cell surface death receptor leads to programmed death (apoptosis), mitogen-activated protein kinase 1 acts as an essential component of the MAP kinase signal transduction pathway, and the progesterone receptor regulates the expression of target genes in response to the progesterone steroid
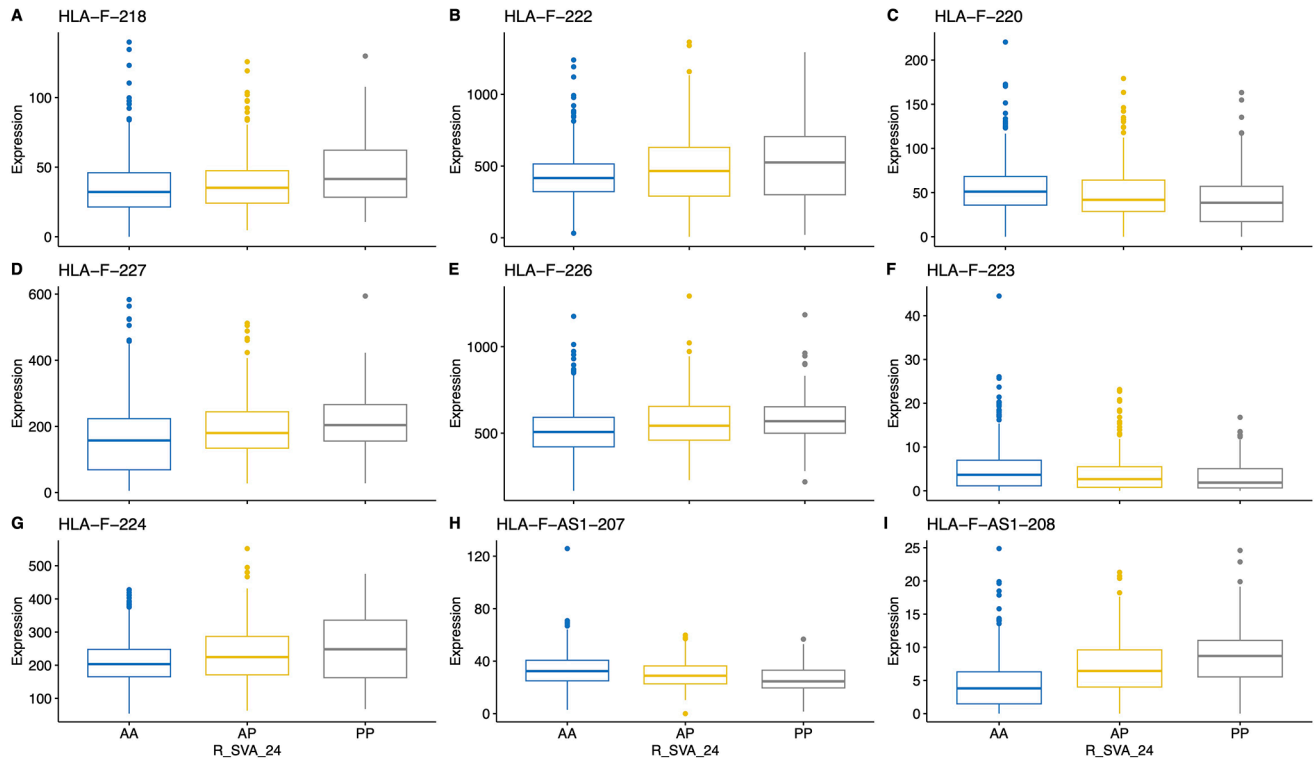
**Figure 5.** Box plots of regulation of the expression of seven different HLA-F transcript sizes (A to G) and two HLA-F-AS1 transcript sizes (H and I) by R_SVA_24 RIPs as totally absent (AA, *n* = 745), absent and present (AP, *n* = 364), or total present (PP, *n* = 122) on the chromosomes of 1231 individuals. All transcripts were significantly affected as shown by the *P*-values and FDR and beta values in Supplemental Tables 1 and 2. The panels (A), (B), (D), (E), (G), and (I) show upregulation of expression and panels (C), (F), and (H) show downregulation of gene expression.

hormone. Some of the translated transcripts regulated by SVA have neurological functions such as PEST proteolytic signal containing nuclear protein (ubiquitination activity), gamma-aminobutyric acid type A receptor subunit beta2 (encephalopathy), acetylcholinesterase (neurotransmitter), and contactin 1 (glycosylphosphatidylinositol (GPI))-anchored neuronal membrane protein that functions as a cell adhesion molecule and that might be associated with neuropathic pain, demyelinating neuropathies, and with the formation of axon connections in the developing nervous system.

## SVA effects on locus-specific differential transcript expression of HLA and non-HLA genes

The SVA differential effects on transcript variants of MHC genes including HLA and non-HLA genes are presented in Supplemental Table 2. SVA affected all the classical class I (*HLA-A, -B, -C*) and class II (*HLA-DR, -DQ, -DP*) genes, and some of the coregulated differentially expressed nonclassical HLA genes (*HLA-F, -G, -H*), HLA class I pseudogenes (*HLA-K, -U, -V, -W, -L, -J, -T*), and HLA class II pseudogenes (*HLA-DRB9, -DRB6*). None of the class I or class II SVA significantly impacted on the expression of the nonclassical class I *HLA-E* gene, although two of them downregulated the *HLA-L* pseudogene in the kappa block. SVA affected the *HERV-16* related lncRNA genes, *HCP5* in the beta block[29] and *HCP5B* in the alpha block,[53] as well as the *HCG4B* lncRNA in the alpha block and *HCG27* lncRNA in the beta block. It is

noteworthy that the differentially expressed transcript isoforms were mostly upregulated, but that some were downregulated simultaneously.

Box plots of transcriptional activities based on the SVA genotypes (1, absent–absent; 2, absent–present; 3, present–present) for the PPMI cohort are shown in Figures 5 and 6 and Supplemental Figure 1. Figure 5 shows box plots of regulation of the expression of seven different *HLA-F* transcript sizes (A to G) and two *HLA-F-AS1* transcript sizes (H and I) by *R_SVA_24* RIPs; whereas Figure 6 shows box plots of regulation of the expression of nine different HLA class II transcript sizes (A to I) and three *TAP2* transcript sizes by *NR_SVA_380* RIPs as two SVA regulatory examples. In general, most box plots showed that the number of transcripts for various expressed HLA and non-HLA genes are upregulated due to the presence of the SVA insertion. Most SVA insertion homozygotes are at a significantly lower frequency than the other genotypes (Table 2). Thus, the genes that are upregulated by low frequency SVA homozygotes or heterozygotes might occur in fewer individuals than the larger number of heterozygotes and homozygotes without the SVA insertions, ranging from less than 0.5% and 21% depending on the SVA insertion in the MHC class I region, and less than 1.3% and 97% depending on the SVA insertion in the MHC class II region (Table 2). In contrast, the *R_SVA_85* and *R_SVA_28* insertion frequencies of >90% suggest that most individuals within the cohort would have SVA upregulated genes, whereas a few homozygous individuals without the SVA insertions might have the same genes downregulated.
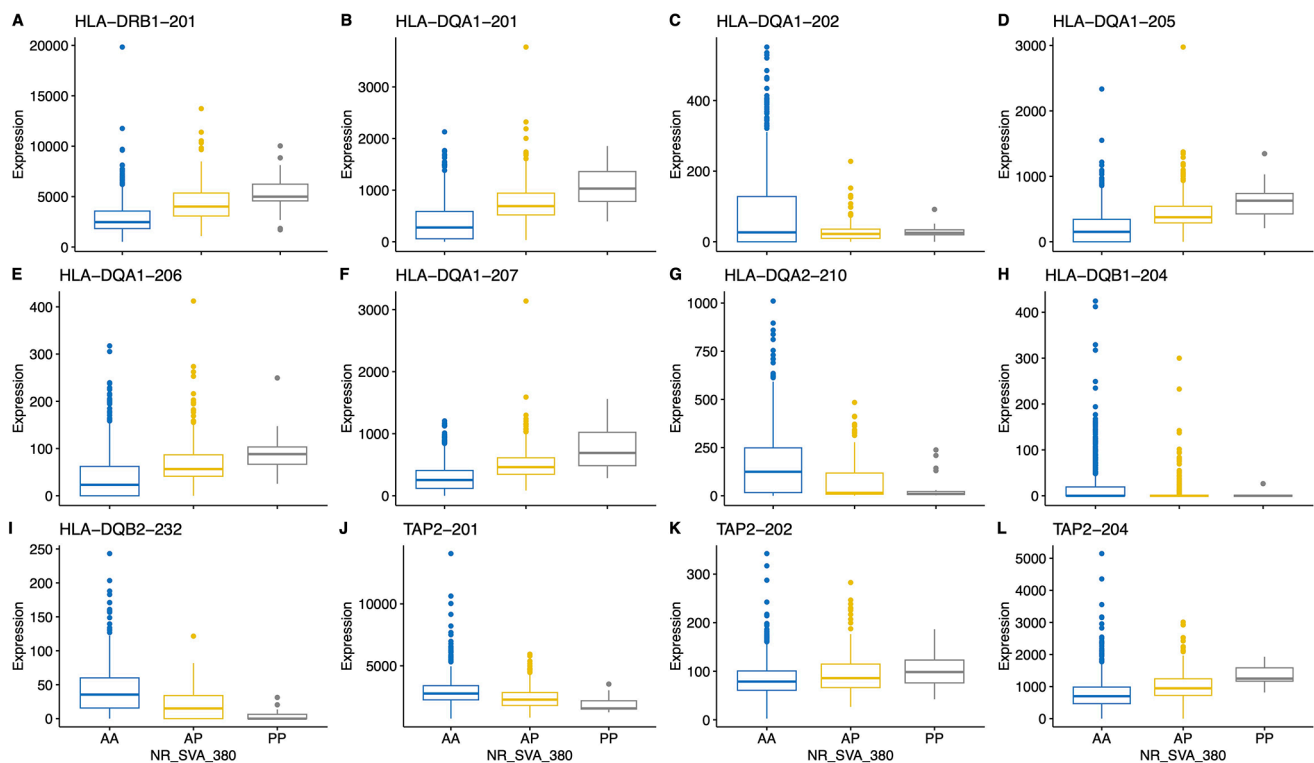
**Figure 6.** Box plots of regulation of the expression of nine different HLA class II transcript sizes (A to I) and three TAP2 transcript sizes by NR_SVA_380 RIPs as totally absent (AA, $n=935$), absent and present (AP, $n=311$), or total present (PP, $n=19$) on the chromosomes of 1266 individuals. All transcripts were significantly affected as shown by the $P$-values and FDR and beta values in Supplemental Tables 1 and 2. The panels (A), (B), (D), (E), (F), (K), and (L) show upregulation of expression and panels (C), (G), (H), (I), and (J) show downregulation of gene expression.

Downregulated genes shown in Supplementary Figure 1 for the *R_SVA_24* insertion homozygotes included *HLA-F-220* and *HLA-F-223*, *HLA-G-207*, *HLA-A-203*, *HLA-A-204*, *HLA-W-207*, and *POLR1H-203*. The downregulated genes for the *R_SVA_25* insertion homozygotes were *HLA-H-201*, *HLA-L-205*, *CCHCR1-201*, *CCHCR1-203*, *CCHCR1-209*, *MICA-214*, *HCP5-202*, *HCP5-206*, *DDX39B-208*, *DDX39B-214*, and *HLA-DQB2-232*. Thus, these SVAs and others have differential effects at the level of transcription induction and reduction and splicing of introns and exons. Many of the downregulated gene transcripts might not be translated into proteins (Supplemental Table 2).

## Effect of R_SVA_24 (alias *SVA-HA*) structural polymorphisms on differential expression of *HLA-A*

The *R_SVA_24* insertion has a long and short variant and a separate internal variable number of tandem repeats (VNTR) – CT length genotypes that affect the expression of the different *HLA-A* isoforms (Figure 7). The violin plots shown in Figure 8C to J displayed the absence of a normal distribution, with many containing multiple peaks. The $P$-values generated by these tests were all highly significant ($P=4.4 \times 10^{-2}$ up to $2.22 \times 10^{-16}$), bar 2. The impact of absent versus short SVA on *HLA-A-204*, and CT alleles A and B on *HLA-A-201* expression, had $P$-values of 0.24 and 0.64, respectively (Figure 8(C) and (J)). Long and short SVAs, and different CT alleles were associated with differential expression and the impacts of these variations differed in a transcript-dependent manner.

For example, the presence of either SVA length (short more so than long) significantly increased *HLA-A-202* versus the absence of the SVA (Figure 8(D)). *HLA-A-203* expression was greater in the presence of the long SVA than the short (Figure 8(E)), whereas expression of *HLA-A-201* was significantly suppressed by either SVA (Figure 8(F)). Following this trend, expression of *HLA-A-202* was significantly reduced by the presence of CT allele A versus allele B (Figure 8(H)), while the reverse was observed in *HLA-A-203* (Figure 8(I)). Those $P$-values generated by likelihood ratio tests and multiple comparisons were substantially weaker for CT variants than SVA length (Figure 8(G) to (J) versus (C) to (F)). This might be caused by the large disparity in sample numbers for each data type. From all 550 samples genotyped, 486 had length metadata, whereas only 106 had viable CT allele calls.

Figure 9 shows the violin plots for the differential expression of *HLA-A* transcription variants stratified according to *R_SVA_24* genotypes, homozygous present (PP), homozygous absent (AA), or heterozygous present–absent (PA). *HLA-A-204*, *HLA-A-202*, *HLA-A-203* were upregulated significantly, whereas *HLA-A-201* was downregulated.

## Discussion

Recent RNAseq studies of a PPMI cohort have shown that hundreds of SVA RIPs regulate or modify coding and non-coding genes, pseudogenes, and TEs at the transcription level within various genomic regions and across different chromosomes.[27,40,54–56] Some SVA RIPs were associated with
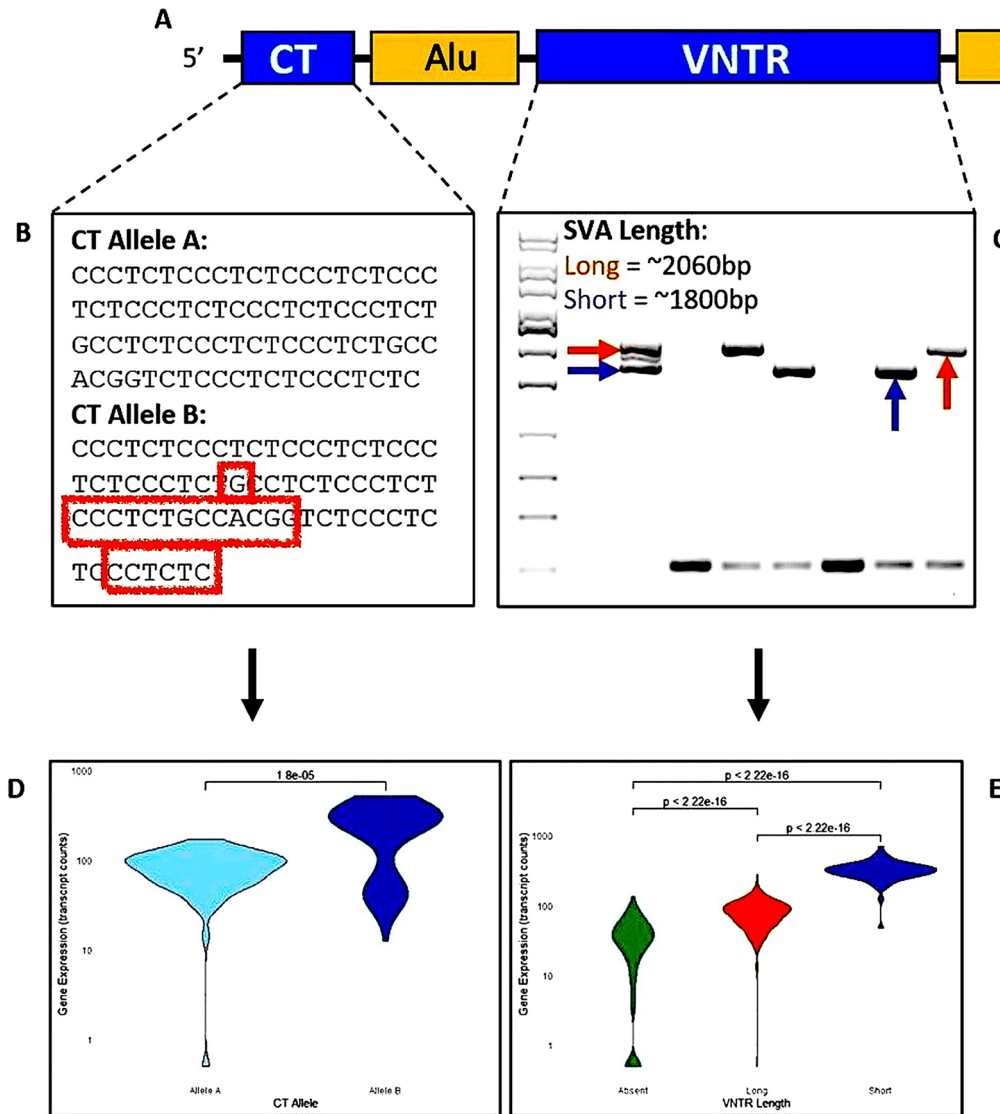
**Figure 7.** Variable SVA structures and effects on HLA-A gene transcription. (A) SVA structure diagram, highlighting highly polymorphic CT and VNTR regions. CT and VNTR copy numbers were thought to underlie SVA length variation. (B) Nucleotide sequences of CT alleles A and B. Nucleotide differences are boxed in red. (C) Exemplar PCR/gel electrophoresis of SVA showing two bands representing each length variant, long and short, indicated by red and purple arrows. (D and E) Example plots (HLA-A-202) generated for HLA-A expression against SVA CT allele A (light blue) and B (dark blue), and VNTR (length) variation, absent (green), long (red), short (purple). Y-axis transcript numbers (counts). X-axis SVA_24 CT allele A ($n=51$) or B ($n=55$) (D), SVA_24 genotype, absent ($n=268$), long ($n=157$), or short ($n=61$) (E).

progression markers of PD, such as R_SVA_24 that was significantly associated with changes in the Hoehn and Yar stage (measure of symptom progression) in PD subjects.[40] Here, we expanded the previous analyses of *SVA_24, SVA_26,* and *SVA_27*[27,40] to include an additional five SVAs and other targeted gene loci, differentially expressed transcripts and spliced variants expressed by the targeted genes within the MHC genomic region. The differentially expressed transcript isoforms were mostly upregulated in accordance with previous observations of SVA effects on gene transcription in the human genome.[27] However, some of the transcripts of the genes that were either upregulated or downregulated might not be translated into functional proteins. For example, according to the Ensembl database, *HLA-C-201, -C-203, -C-204* transcripts translate a full-length protein of

261aa-366aa, *HLA-C-202* translates a truncated protein of 131aa, whereas the other *HLA-C* transcripts translate no protein because of retained introns or other unknown modifications. It is unclear what the outcome or purpose of the differential overexpression of the HLA classical genes might be in relation to immunity with the generation of different-sized HLA antigens that might bind and present peptides at the cellular membranes or be secreted as soluble antigens, and how often such molecular and cellular processes occur normally or pathologically. None of the SVA RIPs had any significant effects on the regulation of the *HLA-E* gene, a gene with two alleles that seems to act independently of the other classical and nonclassical class I HLA genes.[57] Similar trends of shared and independent gene regulation were observed for the SVA RIPs in the class II region. Six of the eight SVA
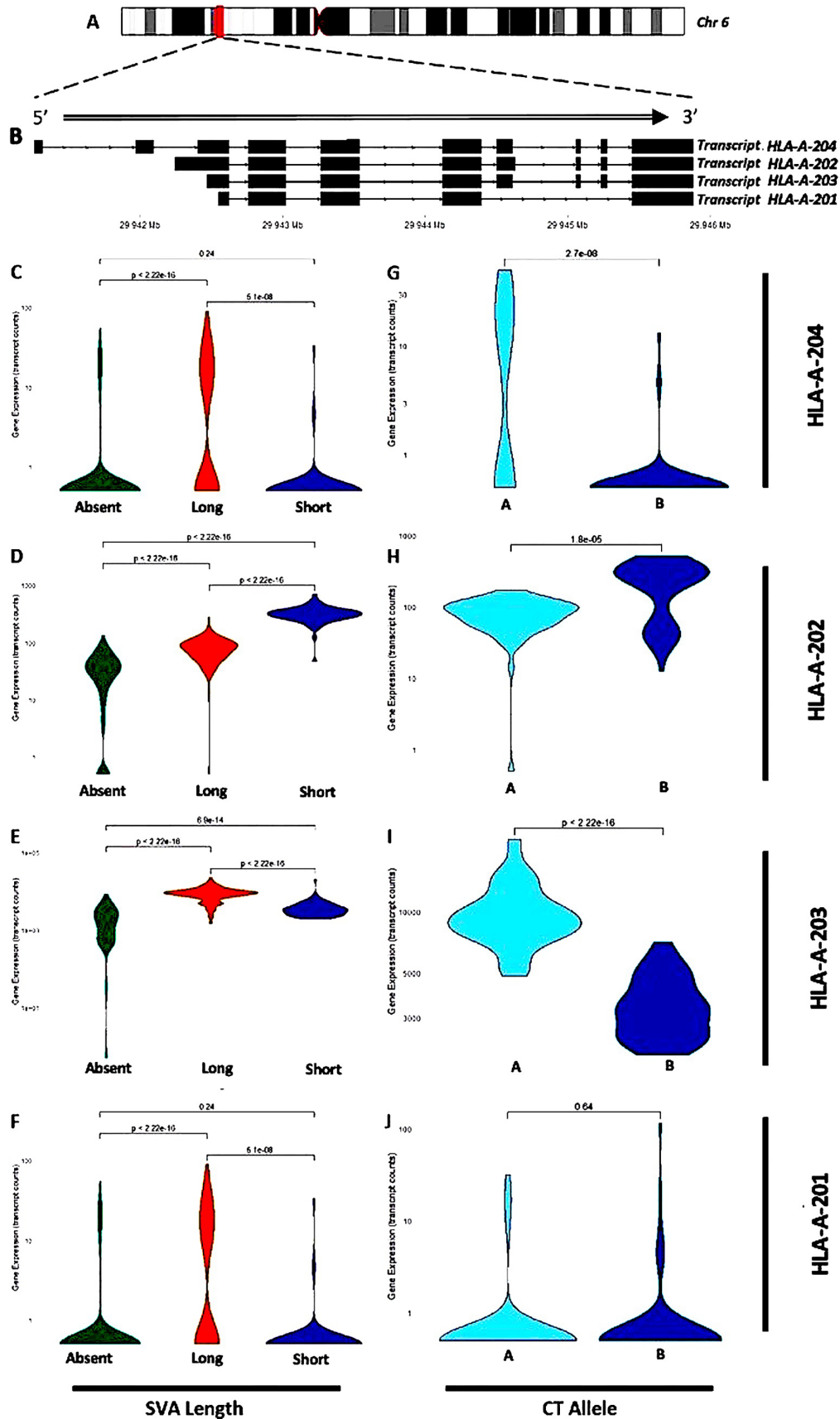
**Figure 8.** Kruskal–Wallis multiple comparisons of HLA-A transcript expression relative to SVA length and CT allele. (A) Chromosome 6 ideogram, with location of the HLA-A locus shown in red. (B) Gene models of each HLA-A transcript (data derived from ensembl.org), exons indicated by solid blocks, introns by arrowed lines. Transcription start sites (TSS; left most exon) are unique for each transcript. All shown exons are protein coding. (C to F). Violin plots of HLA-A expression by SVA length genotype, absent ($n = 268$), long ($n = 157$), or short ($n = 61$). (G to J). Violin plots of HLA-A expression by CT allele genotype A ($n = 51$) or B ($n = 55$). Y-axis transcript numbers (counts). X-axis SVA_24 genotypes as length or CT allele.
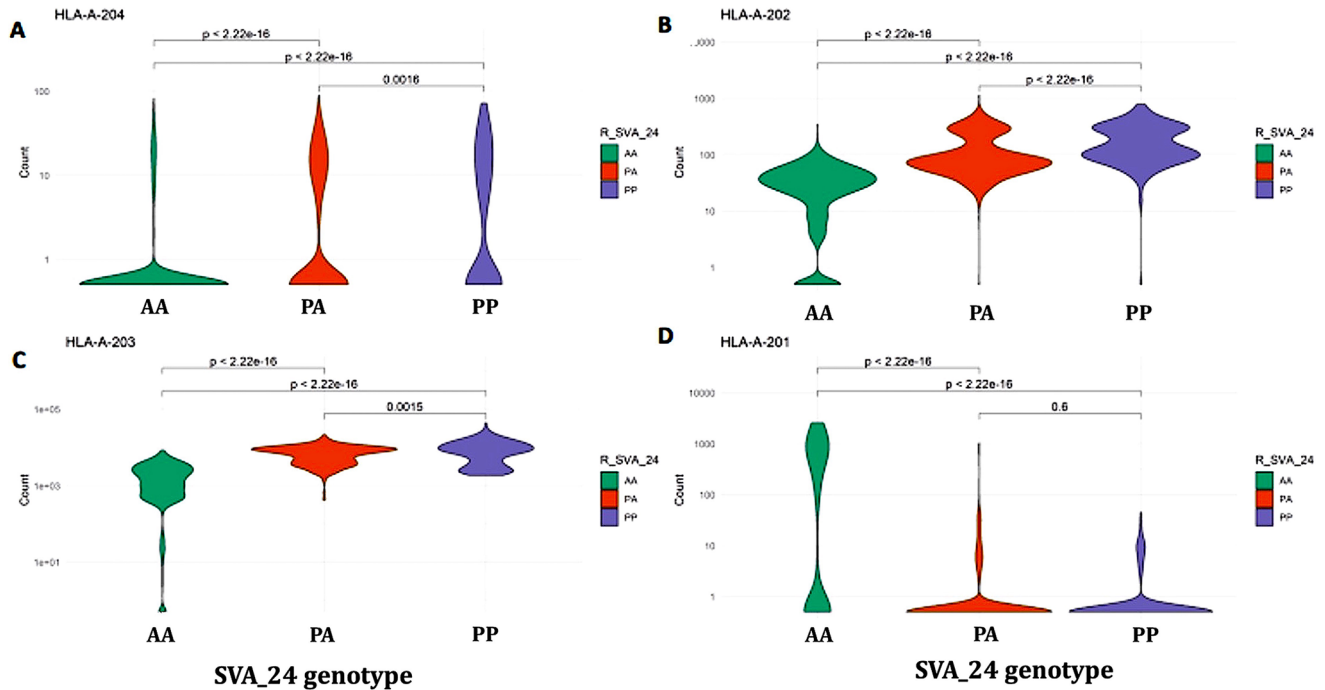
**Figure 9.** Expression of HLA-A isoforms. (A) HLA-A-204, (B) HLA-A-202, (C) HLA-A-203, and (D) HLA-A-201, stratified according to SVA_24 genotypes, homozygous present (PP), homozygous absent (AA), or heterozygous present–absent (PA). *Y*-axis transcript numbers (counts). *X*-axis SVA_24 genotype, AA, PA, or PP.

RIPs also had eQTL *trans* effects outside the MHC region. Only the regulatory effects of *R_SVA_26* insertion within class I and NR_SVA_381 insertion within class II seemed to be limited to the MHC genomic region. The other SVA insertions within the MHC had *trans* effects involving all chromosomes except chromosomes 9 and 18. Thus, possible epistatic effects of these SVA on multiple gene interactions both inside and outside the MHC might either debilitate or protect regarding the pathogenesis of PD.

Koks *et al.*[27] showed that *R_SVA_27* significantly upregulated *HLA-B, HLA-DRB1*, and *HLA-DRB5* and that *R_SVA_26* also significantly upregulated *HLA-B* in an additive manner. In the present study, we observed that these two SVA RIPs and the six others regulated clusters of genes both within and outside the MHC in different ways, but occasionally with overlapping or possible additive effects on some of the same genes. Because of the different frequencies for the SVA RIPs within the MHC region of the PPMI cohort, it is likely that overlapping and additive effects of the expressed genes overall were low. Although we did not perform calculations or statistics for these potential additive dosage effects, we did note their coexistence in some of the same individuals. The overlapping and expression of genes regulated by the different MHC SVAs probably depend on the MHC haplotypes within the PPMI cohort for which we have no HLA allelic information at present. The eight SVA RIPs in this study were limited to the alpha and beta blocks of the HLA class I region and the beta and epsilon blocks of the MHC class II region. Three of these SVA RIPs annotated within the reference genome, GRCh38.p14/hg38 (https://genome.ucsc.edu) and were described previously in MHC haplotyped reference cell-lines and in the populations of European Australians, Japanese, and African Americans.[54]

The *R_SVA_24* (alias *SVA-HA*), *R_SVA_25* (alias *SVA-HC*), and *R_SVA_26* (alias *SVA-MIC*) retrotransposons are inserted at a relatively low frequency in Australian European populations and strongly associated with the ancestral MHC haplotype 7.1AH, but not the 8.1AH.[20,21,39] *SVA-HA*, which is referred to as *R_SVA_24* in this study, was first detected in Australian Europeans, Japanese, and African Americans at a homozygous frequency of about 10% and a strong LD with the *HLA-A3/A11/A30*, *HLA-C7*, and *HLA-B7* alleles.[39] This is half the homozygous insertion frequency of 19.8% for *R_SVA_24* in the PPMI cohort (Table 2). Coincidently, 7.1AH along with other HLA class II alleles were associated with PD.[15] Although NR_SVA_381 (genotype PP) was the only MHC SVA RIP associated significantly with the risk of PD (data not shown), many other SVA RIPs outside of the MHC genomic region of the PPMI cohort were found to modify the progression of PD.[27,38,40,54–56] Also, there are many other conserved (fixed) and structurally polymorphic SVA within the MHC than just the eight that we have examined in the PPMI cohort, and they might regulate gene expression in other population groups and disease cohorts.[58]

The eight MHC SVA insertions described in the present analysis are transcribed in whole blood samples and homozygous lymphoblastoid cell lines (unpublished findings). However, there are at least another 10 SVA within different loci of the human MHC genomic region including at least four that appear to have been fixed in human and chimpanzee genomes.[20,21,39] Two SVA RIPs integrated within the MHC class III region, one within an intron of the *STK19* gene centromeric of *C4A*[59] and the other within an intron of the *C2* gene,[60] were among the first SVA to be discovered and mapped precisely within the human genome.[61] Also, there are SVA-RIP near *HLA-C* (SVA-HC) and *HLA-B* (SVA-B)[20,21,39]

that were not identified in the present analysis. We found no evidence here that these other SVA are transcribed or are associated with eQTL of MHC genes. The *P* values (corrected for multiple testing by the Bonferroni test), FDRs, and beta values that we obtained using an eQTL software computing package in this study is an extension of our previous published findings.[27,40] These statistical measures are reasonably reliable with FDR providing additional corrections for multiple testing.[62] However, it is possible that our high level of statistical stringency has lost some of the lower level biologically significant SVA effects on HLA and non-HLA genes. In this regard, our analysis may have underestimated rather than overestimated some of the SVA effects. Also, our study was limited to the blood cells of the PPMI cohort, whereas the SVA effects on the MHC genes are likely to vary in different tissues or cells, population groups, cohorts, disease groups, and individuals with different MHC haplotypes. Since we have not genotyped or haplotyped the individuals in our PPMI cohort, we could not make direct associations of the SVA effects with MHC haplotypes or HLA alleles. In this regard, our study is preliminary, and many other studies are still needed to follow-up and progress our findings.

We also examined in more detail the structural and regulatory effects of *R_SVA_24* on *HLA-A* gene expression than previously reported.[27,40] Size variations in the SVA that were associated with the differential expression of *HLA-A* transcripts (Figure 9) may have been in part mediated by transcription factor binding to the VNTR region.[40,56] For example, transcription factors YB-1 and CTCF were shown to bind to a polymorphic VNTR in the *SLC6A4* gene in a copy number-specific manner.[63] Such *in vitro* studies have yet to be performed on the MHC SVA sequences. The violin or box plots (Figures 5 to 9) nevertheless further highlight the differential effects of the SVA RIPs such as *R_SVA_24* on the splicing mechanisms involved in generating the different HLA transcript sizes that in turn probably affect the HLA antigen structures, membrane attachments, solubility, and binding to beta-2-microglobulins, T cell and killer cell receptors. Three of the four SVA insertions (*NR_SVA_377, R_SVA_24, R_SVA_25*) within the MHC class I region modified the *HLA-A* gene expression and transcription variants, whereas none of the four SVA insertion within the class II region had any effect on *HLA-A* gene expression. This provides further evidence for the hypothesis that SVAs have a differential regulatory role in HLA transcription, and that this role is dependent on the SVA integration site, the SVA variable sequence type, and possibly also the gene allele.

The present analysis of eight SVA RIPs within the MHC genomic region further highlights the importance of these REs in the multigenic regulation of HLA and non-HLA gene expression (eQTL) both within (*cis*) and outside (*trans*) the MHC at the level of (1) homozygosity and heterozygosity of the SVA RIPs, and (2) the expressed transcripts as spliced variants of different sizes and exonic representations. Moreover, these SVA RIPs are modulators or regulators of multigenetic expression that integrate the MHC with other genomic regions as part of a complicated and dynamic network or "interactome" in regulating human health and disease.[64,65] The mechanism by which these SVA regulate multiple genes over long distances

both within and outside the MHC on various chromosomes is not known. Various biochemical and genetical mechanisms of long-range chromosomal interactions and gene regulation probably have a role including epigenetic processes, chromatin loop formation, nuclear matrix proteins, and RNA,[66,67] and hundreds of genomic anchors that organize chromatin loop domains across the MHC.[68] Since the eight MHC SVA insertions that we described in the present analysis are transcribed, it is likely that the SVA RNAs regulate both the multigene networks and alternative splicing outcomes, involving mechanisms such as SVA exon capture[69,70] and/or interaction with distant promoters and enhancers by way of 3D chromatin structures.[65,67] The mechanisms involved in the long-range regulation of multigene expression by SVA insertions is a challenge that still needs considerable research and modeling with coherent and testable hypotheses.

### AUTHORS' CONTRIBUTIONS

Conceptualization, SK, JKK, and JPQ; methodology, SK; software, ALP; validation, ALP, LDM, VJB, and AF; formal analysis, SK, AF, and LDM; investigation, JKK, JPQ; resources, SK, JPQ; data curation, ALP, VJB, LDM, and AF; writing – original draft preparation, JKK; writing – review and editing, JKK, JPQ, and SK; visualization, JKK and SK; supervision, SK; project administration, SK; funding acquisition, SK. All authors have read and agreed to the published version of the manuscript.

### DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### ETHICAL APPROVAL

### FUNDING

### DATA AVAILABILITY

Raw data are available from the PPMI website (www.ppmi-info.org/data).

### ORCID IDS

Alexander Fröhlich https://orcid.org/0000-0001-9083-7485
Vivien J Bubb https://orcid.org/0000-0003-2763-7004
Sulev Koks https://orcid.org/0000-0001-6087-6643

### SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

## REFERENCES

1. Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev* 1999;**167**:275–304

2. Vandiedonck C, Knight JC. The human major histocompatibility complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic* 2009;**8**:379–94

3. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* 2004;**64**:631–49

4. Shiina T, Hosomichi K, Inoko H, Kulski JK. The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* 2009;**54**:15–39

5. Kulski JK, Suzuki S, Shiina T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var* 2022;**9**:49

6. Tait BD. The importance of establishing genetic phase in clinical medicine. *Int J Immunogenet* 2022;**49**:1–7

7. Barker DJ, Maccari G, Georgiou X, Cooper MA, Flicek P, Robinson J, Marsh SGE. The IPD-IMGT/HLA database. *Nucleic Acids Res* 2023;**51**:D1053–60

8. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, Hart E, Howe K, Jackson DK, Palmer S, Roberts AN, Sims S, Stewart CA, Traherne JA, Trevanion S, Wilming L, Rogers J, de Jong PJ, Elliott JF, Sawcer S, Todd JA, Trowsdale J, Beck S. Variation analysis and gene annotation of eight MHC haplotypes: the MHC haplotype project. *Immunogenetics* 2008;**60**:1–18

9. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet* 2013;**14**:301–23

10. Yunis EJ, Larsen CE, Fernandez-Viña M, Awdeh ZL, Romero T, Hansen JA, Alper CA. Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. *Tissue Antigens* 2003;**62**:1–20

11. Goodin DS, Khankhanian P, Gourraud PA, Vince N. Highly conserved extended haplotypes of the major histocompatibility complex and their relationship to multiple sclerosis susceptibility. *PLoS ONE* 2018;**13**:e0190043

12. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. *Hum Immunol* 2007;**68**:779–88

13. Neville MJ, Lee W, Humburg P, Wong D, Barnardo M, Karpe F, Knight JC. High resolution HLA haplotyping by imputation for a British population bioresource. *Hum Immunol* 2017;**78**:242–51

14. Smith WP, Vu Q, Li SS, Hansen JA, Zhao LP, Geraghty DE. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. *Genomics* 2006;**87**:561–71

15. Wissemann WT, Hill-Burns EM, Zabetian CP, Factor SA, Patsopoulos N, Hoglund B, Holcomb C, Donahue RJ, Thomson G, Erlich H, Payami H. Association of Parkinson disease with structural and regulatory variants in the HLA region. *Am J Hum Genet* 2013;**93**:984–93

16. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, Tan M, Kia DA, Noyce AJ, Xue A, Bras J, Young E, von Coelln R, Simón-Sánchez J, Schulte C, Sharma M, Krohn L, Pihlstrøm L, Siitonen A, Iwaki H, Leonard H, Faghri F, Gibbs JR, Hernandez DG, Scholz SW, Botia JA, Martinez M, Corvol JC, Lesage S, Jankovic J, Shulman LM, Sutherland M, Tienari P, Majamaa K, Toft M, Andreassen OA, Bangale T, Brice A, Yang J, Gan-Or Z, Gasser T, Heutink P, Shulman JM, Wood NW, Hinds DA, Hardy JA, Morris HR, Gratten J, Visscher PM, Graham RR, Singleton AB, 23andMe Research Team, System Genomics of Parkinson's Disease Consortium, International Parkinson's Disease Genomics Consortium. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* 2019;**18**:1091–102

17. Yu E, Ambati A, Andersen MS, Krohn L, Estiar MA, Saini P, Senkevich K, Sosero YL, Sreelatha AAK, Ruskey JA, Asayesh F, Spiegelman D, Toft M, Viken MK, Sharma M, Blauwendraat C, Pihlstrøm L, Mignot E, Gan-Or Z. Fine mapping of the HLA locus in Parkinson's disease in Europeans. *NPJ Park Dis* 2021;**7**:84

18. Lokki M, Paakkanen R. The complexity and diversity of major histocompatibility complex challenge diseasee association studies. *HLA* 2018;**93**:3–15

19. Traherne JA. Human MHC architecture and evolution: implications for disease association studies. *Int J Immunogenet* 2008;**35**:179–92

20. Kulski JK, Suzuki S, Shiina T. SNP-density crossover maps of polymorphic transposable elements and HLA genes within MHC class I haplotype blocks and junction. *Front Genet* 2021;**11**:594318

21. Kulski JK, Suzuki S, Shiina T. Haplotype shuffling and dimorphic transposable elements in the human extended major histocompatibility complex class II region. *Front Genet* 2021;**12**:665899

22. Cun Y, Shi L, Kulski JK, Liu S, Yang J, Tao Y, Zhang X, Shi L, Yao Y. Haplotypic associations and differentiation of MHC class II polymorphic alu insertions at five loci with HLA-DRB1 alleles in 12 minority ethnic populations in China. *Front Genet* 2021;**12**:636236

23. Kulski JK, Mawart A, Marie K, Tay GK, AlSafar HS. MHC class I polymorphic Alu insertion (POALIN) allele and haplotype frequencies in the Arabs of the United Arab Emirates and other world populations. *Int J Immunogenet* 2019;**46**:247–62

24. Kulski JK, Shigenari A, Inoko H. Genetic variation and hitchhiking between structurally polymorphic Alu insertions and HLA-A, -B, and -C alleles and other retroelements within the MHC class I region. *Tissue Antigens* 2011;**78**:359–77

25. Lamontagne M, Joubert P, Timens W, Postma DS, Hao K, Nickle D, Sin DD, Pare PD, Laviolette M, Bossé Y. Susceptibility genes for lung diseases in the major histocompatibility complex revealed by lung expression quantitative trait loci analysis. *Eur Respir J* 2016;**48**:573–6

26. Lam TH, Shen M, Tay MZ, Ren EC. Unique Allelic eQTL Clusters in Human MHC Haplotypes. *G3 Genesgenomesgenetics* 2017;**7**:2595–604

27. Koks S, Pfaff AL, Bubb VJ, Quinn JP. Expression quantitative trait loci (eQTLs) associated with retrotransposons demonstrate their modulatory effect on the transcriptome. *Int J Mol Sci* 2021;**22**:6319

28. Smallegan MJ, Shehata S, Spradlin SF, Swearingen A, Wheeler G, Das A, Corbet G, Nebenfuehr B, Ahrens D, Tauber D, Lennon S, Choi K, Huynh T, Wieser T, Schneider K, Bradshaw M, Basken J, Lai M, Read T, Hynes-Grace M, Timmons D, Demasi J, Rinn JL. Genome-wide binding analysis of 195 DNA binding proteins reveals "reservoir" promoters and human specific SVA-repeat family regulation. *PLoS ONE* 2021;**16**:e0237055

29. Kulski JK. Long noncoding RNA HCP5, a Hybrid HLA class i endogenous retroviral gene: structure, expression, and disease associations. *Cells* 2019;**8**:480

30. Fort V, Khelifi G, Hussein SMI. Long non-coding RNAs and transposable elements: a functional relationship. *Biochim Biophys Acta Mol Cell Res* 2021;**1868**:118837

31. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res* 2016;**45**:gkw1286

32. Kwon YJ, Choi Y, Eo J, Noh YN, Gim JA, Jung YD, Lee JR, Kim HS. Structure and expression analyses of SVA elements in relation to functional genes. *Genomics Inform* 2013;**11**:142–8

33. Vandiedonck C, Taylor MS, Lockstone HE, Plant K, Taylor JM, Durrant C, Broxholme J, Fairfax BP, Knight JC. Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res* 2011;**21**:1042–54

34. Spirito G, Mangoni D, Sanges R, Gustincich S. Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *BMC Bioinformatics* 2019;**20**:495

35. Makałowski W, Gotea V, Pande A, Makałowska I. Transposable elements: classification, identification, and their use as a tool for comparative genomics. In: Anisimova M (ed.) Evolutionary genomics. New York: Springer, 2019, pp.177–207

36. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* 2006;**27**:323–9

37. Autio MI, Bin Amin T, Perrin A, Wong JY, Foo RS-Y, Prabhakar S. Transposable elements that have recently been mobile in the human genome. *BMC Genomics* 2021;**22**:789

38. Savage AL, Bubb VJ, Breen G, Quinn JP. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol Biol* 2013;**13**:101

39. Kulski JK, Shigenari A, Inoko H. Polymorphic SVA retrotransposons at four loci and their association with classical HLA class I alleles in Japanese, Caucasians and African Americans. *Immunogenetics* 2010;**62**:211–30

40. Pfaff AL, Bubb VJ, Quinn JP, Koks S. Reference SVA insertion polymorphisms are associated with Parkinson's disease progression and differential gene expression. *NPJ Park Dis* 2021;**7**:44

41. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, Simuni T, Jennings D, Tanner CM, Trojanowski JQ, Shaw LM, Seibyl J, Schuff N, Singleton A, Kieburtz K, Toga AW, Mollenhauer B, Galasko D, Chahine LM, Weintraub D, Foroud T, Tosun-Turgut D, Poston K, Arnedo V, Frasier M, Sherer T, Parkinson's Progression Markers Initiative. The Parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. *Ann Clin Transl Neurol* 2018;**5**:1460–77

42. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**:i333–9

43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75

44. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**:417–9

45. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 2012;**28**:1353–8

46. Pérez-Calero C, Bayona-Feliu A, Xue X, Barroso SI, Muñoz S, González-Basallote VM, Sung P, Aguilera A. UAP56/DDX39B is a major cotranscriptional RNA–DNA helicase that unwinds harmful R loops genome-wide. *Genes Dev* 2020;**34**:898–912

47. Schott G, Garcia-Blanco MA. MHC class III RNA binding proteins and immunity. *RNA Biol* 2021;**18**:640–6

48. Zhao L, Liu Q, Huang J, Lu Y, Zhao Y, Ping J. TREX (transcription/export)-NP complex exerts a dual effect on regulating polymerase activity and replication of influenza A virus. *PLoS Pathog* 2022;**18**:e1010835

49. Szymura SJ, Bernal GM, Wu L, Zhang Z, Crawley CD, Voce DJ, Campbell P-A, Ranoa DE, Weichselbaum RR, Yamini B. DDX39B interacts with the pattern recognition receptor pathway to inhibit NF-κB and sensitize to alkylating chemotherapy. *BMC Biol* 2020;**18**:32

50. Mulcahy H, O'Rourke KP, Adams C, Molloy MG, O'Gara F. LST1 and NCR3 expression in autoimmune inflammation and in response to IFN-γ, LPS and microbial infection. *Immunogenetics* 2006;**57**:893–903

51. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, Szigyarto CA, Skogs M, Takanen JO, Berling H, Tegel H, Mulder J, Nilsson P, Schwenk JM, Lindskog C, Danielsson F, Mardinoglu A, Sivertsson A, von Feilitzen K, Forsberg M, Zwahlen M, Olsson I, Navani S, Huss M, Nielsen J, Ponten F, Uhlén M. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;**13**:397–406

52. Chittenden T, Flemington C, Houghton AB, Ebb RG, Gallo GJ, Elangovan B, Chinnadurai G, Lutz RJ. A conserved domain in Bak, distinct from BH1 and BH2, mediates cell death and protein binding functions. *EMBO J* 1995;**14**:5589–96

53. Kulski JK, Gaudieri S, Martin A, Dawkins RL. Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. *J Mol Evol* 1999;**49**:84–97

54. Gianfrancesco O, Bubb VJ, Quinn JP. SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides* 2017;**64**:37

55. Gianfrancesco O, Geary B, Savage AL, Billingsley KJ, Bubb VJ, Quinn JP. The role of SINE-VNTR-Alu (SVA) retrotransposons in shaping the human genome. *Int J Mol Sci* 2019;**20**:5977

56. Quinn JP, Bubb VJ. SVA retrotransposons as modulators of gene expression. *Mob Genet Elem* 2014;**4**:e32102

57. Kanevskiy L, Erokhina S, Kobyzeva P, Streltsova M, Sapozhnikov A, Kovalenko E. Dimorphism of HLA-E and its disease association. *Int J Mol Sci* 2019;**20**:5496

58. Wang L, Norris ET, Jordan IK. Human retrotransposon insertion polymorphisms are associated with health and disease via gene regulatory phenotypes. *Front Microbiol* 2017;**8**:1418

59. Shen L, Wu L, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. *J Biol Chem* 1994;**269**:8466–76

60. Zhu ZB, Hsieh S-L, Bentley DR, Campbell RD, Volanakis JE. A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. *J Exp Med* 1992;**175**:1783–7

61. Ono M, Kawakami M, Takezawa T. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res* 1987;**15**:8725–37

62. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**57**:289–300

63. Klenova E, Scott AC, Roberts J, Shamsuddin S, Lovejoy EA, Bergmann S, Bubb VJ, Royer H-D, Quinn JP. YB-1 and CTCF differentially regulate the 5-HTT polymorphic intron 2 enhancer which predisposes to a variety of neurological disorders. *J Neurosci* 2004;**24**:5966–73

64. Paci P, Fiscon G, Conte F, Wang R-S, Farina L, Loscalzo J. Gene coexpression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Syst Biol Appl* 2021;**7**:3

65. Barnada SM, Isopi A, Tejada-Martinez D, Goubert C, Patoori S, Pagliaroli L, Tracewell M, Trizzino M. Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. *PLoS Genet* 2022;**18**:e1010225

66. Pappalardo XG, Barra V. Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics Chromatin* 2021;**14**:25

67. Trigiante G, Blanes Ruiz N, Cerase A. Emerging roles of repetitive and repeat-containing RNA in nuclear and chromatin organization and gene expression. *Front Cell Dev Biol* 2021;**9**:735527

68. Ottaviani D, Lever E, Mitter R, Jones T, Forshew T, Christova R, Tomazou EM, Rakyan VK, Krawetz SA, Platts AE, Segarane B, Beck S, Sheer D. Reconfiguration of genomic anchors upon transcriptional activation of the human major histocompatibility complex. *Genome Res* 2008;**18**:1778–86

69. Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH Jr. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* 2009;**19**:1983–91

70. Nadler MJS, Chang W, Ozkaynak E, Huo Y, Nong Y, Boillot M, Johnson M, Moreno A, Anderson MP Hominoid SVA-lncRNA AK057321 targets human-specific SVA retrotransposons in SCN8A and CDK5RAP2 to initiate neuronal maturation. *Commun Biol* 2023;**6**:347