Twin Research and Human Genetics (2023), **26**, 269–276
doi:10.1017/thg.2023.42

CAMBRIDGE UNIVERSITY PRESS

Article

Cap Analysis of Gene Expression Clarifies Transcriptomic Divergence Within Monozygotic Twin Pairs

Hirokazu Katoh[1,2], Hiroaki Asai[3], Keiko Takemoto[4], Rie Tomizawa[2], Chika Honda[2], Mikio Watanabe[2,5], Osaka Twin Research Group[6], and Tomoyuki Honda[1,2,3]

[1]Department of Virology, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Science, Okayama, Japan, [2]Center for Twin Research, Osaka University Graduate School of Medicine, Osaka, Japan, [3]Division of Virology, Department of Microbiology and Immunology, Osaka University Graduate School of Medicine, Osaka, Japan, [4]Institute for Life and Medical Sciences, Kyoto University, Kyoto, Japan, [5]Department of Biomedical Informatics, Osaka University Graduate School of Medicine, Osaka, Japan and [6]Center for Twin Research, Osaka University Graduate School of Medicine, Osaka, Japan

Abstract

Phenotypic variation is the result of gene expression based on complex interaction between genetic and environmental factors. It is well known that genetic and environmental factors influence gene expression, but our understanding of their relative importance remains limited. To obtain a hint for the understanding of their contributions, we took advantage of monozygotic twins, as they share genetic and shared environmental factors but differ in nonshared factors, such as environmental differences and stochastic factors. In this study, we performed cap analysis of gene expression on three pairs of twins and clustered each individual based on their expression profiles of annotated genes. The dendrogram of annotated gene transcripts showed a monophyletic clade for each twin pair. We also analyzed the expression of retrotransposons, such as human endogenous retroviruses (HERVs) and long interspersed nuclear elements (LINEs), given their abundance in the genome. Clustering analyses demonstrated that HERV and LINE expression diverged even within monozygotic twin pairs. Thus, HERVs and LINEs are more susceptible to nonshared factors than annotated genes. Motif analysis of differentially expressed annotated genes suggests that specificity protein/Krüppel-like factor family transcription factors are involved in the expression divergence of annotated gene influenced by nonshared factors. Collectively, our findings suggest that expressions of annotated genes and retrotransposons are differently regulated, and that the expression of retrotransposons is more susceptible to nonshared factors than annotated genes.

Keywords: Monozygotic twins; Cap analysis of gene expression; Genes; Retrotransposons; Environmental factors

(Received 26 April 2023; revise received 20 August 2023; accepted 22 August 2023; First Published online 17 October 2023)

Evaluating the relative contributions of genetic and environmental factors in phenotypic outcome remains a major question in biology. Researchers have long taken advantage of the identical genomes of monozygotic (MZ) twins to evaluate genetic and environmental influences on traits and on gene expression (Boomsma et al., 2002). In the classical twin modeling, phenotypic variance is decomposed of genetic, shared environmental, and nonshared factors (McAdams et al., 2021). Nonshared factors are composed of twin-specific environmental differences and stochastic factors. Early microarray studies have revealed that the proportion of differentially expressed genes (DEGs) within MZ twin pairs was low compared to those between unrelated individuals (Sharma et al., 2005), and that the variance of the gene expression within MZ twin pairs is smaller than those of siblings or unrelated individuals (Cheung et al., 2003). Powell et al. (2012) calculated correlations of the expression levels of 9555 genes within MZ twin pairs (50 pairs in total) and estimated heritability of each gene. Mean heritability was 0.38 and 0.32 for lymphoblastoid cell lines and whole blood, respectively (Powell et al., 2012). In the MuTHER (Multiple Tissue Human Expression Resource) project, 856 twins (one-third MZ and two-thirds dizygotic [DZ], aged from 38.7 to 84.6 years) were recruited and genomewide expression profiling was performed across multiple tissues. The mean heritability estimates of expressed transcripts was 0.16–0.26 (Grundberg et al., 2012). Wright et al. (2014) profiled gene expression of peripheral blood in 2752 twins (690 MZ twin pairs and 618 DZ twin pairs, median of age of 32 years) by using microarray and estimated heritability of genes by using the classic twin modeling. They showed that the mean heritability was 0.10–0.14. Ouwens et al. (2020) also estimated the heritability of gene expression in whole blood as 0.20 using the twin RNA-seq data (1497 adult individuals, including 459 MZ twin pairs, and 150 DZ twin pairs, aged from 17.6 to 79.6). These previous studies imply that, while genetic factors regulate gene expression, nongenetic factors appear to have a larger influence on gene expression on a genomewide scale.

In addition to genes coding for proteins and functional RNAs, the human genome harbors many repetitive sequences, including transposable elements (TEs; Lander et al., 2001). Classified based on their mechanism of propagation (Bhat et al., 2022; Hoyt et al., 2022; Piégu et al., 2015), retrotransposons are the largest class of TEs. Retrotransposons move within the genome via a copy-and-paste

Corresponding author: Tomoyuki Honda; Email: thonda@okayama-u.ac.jp

Cite this article: Katoh H, Asai H, Takemoto K, Tomizawa R, Honda C, Watanabe M, Osaka Twin Research Group, and Honda T. (2023) Cap Analysis of Gene Expression Clarifies Transcriptomic Divergence Within Monozygotic Twin Pairs. *Twin Research and Human Genetics* **26**: 269–276, https://doi.org/10.1017/thg.2023.42

mechanism, first forming RNA intermediates, being reverse-transcribed to DNA, and then integrating at a new genomic site. Retrotransposons are further separated into two subclasses: the first encodes their own catalytic enzymes and comprises human endogenous retroviruses (HERVs) and long interspersed nuclear elements (LINEs); the other consists of short interspersed nuclear elements (SINEs) and composite retroelement SINE-VNTR-*Alus* (SVAs), which are nonautonomous and rely on LINE-encoded proteins for retrotransposition (Konkel & Batzer, 2010). While gene-coding sequences comprise about 1.5% of the human genome, HERVs and LINEs account for approximately one-third (Lander et al., 2001). In general, epigenetic mechanisms (e.g., DNA methylation and histone modification) silence HERV and LINE expression, preventing their potential threat to host genome stability (Sammarco et al., 2022). Although the regulatory mechanism of retrotransposon expression is not fully understood, HERV and LINE expression is considered to be regulated by genetic and environmental factors, similarly to the protein-coding genes. For example, external stimuli, such as UV radiation, infections and chemicals, as well as internal stimuli, such as hormones and cytokines, stimulate transcription of HERVs (Durnaoglu et al., 2021). Addition of hypomethylation reagents increases LINE-1 mRNA expression and activates retrotransposition in cancer cells (Chénais, 2022). We previously found that therapeutic vaccination for human visceral leishmaniasis and post kala azar dermal leishmaniasis upregulated transcripts containing MLT-int of ERVs, which belong to the mammalian apparent long terminal repeat (LTR)-retrotransposon (MaLR) family (T. Honda et al., 2019; Osman et al., 2017). To further investigate the contribution of genetic and environmental factors on HERV and LINE expression, a twin study is useful. As MZ twins share genetic factors and shared environmental factors, difference in retrotransposon/gene expression within MZ twin pairs is explained by the influence of nonshared factors. To our best knowledge, at present there is no study investigating expression profiles of retrotransposons in twins or estimating the heritability of retrotransposon expression using the classical twin modeling. Thus, the influence of factors shared or nonshared by MZ twins on the regulation of retrotransposon expression remains unclear.

Here, we quantified the expression of annotated genes, as well as of HERVs and LINEs, from blood samples of three MZ twin pairs by using cap analysis of gene expression (CAGE; Arner et al., 2015; Forrest et al., 2014). CAGE is more quantitative than RNA-seq because CAGE enables counting the number of transcripts without PCR amplification. The primary objective of our pilot study is to evaluate divergence of gene and retrotransposon expression within MZ twin pairs, between unrelated individuals, and within individuals. Clustering analysis of annotated gene expression showed a monophyletic clade for each twin pair. On the other hand, retrotransposon expression was more divergent than annotated gene expression even within MZ twin pairs. Motif analysis of differentially expressed annotated genes (DEGs) in MZ twins revealed enrichment of sequences potentially bound by Specificity protein 1 (Sp1) and Sp2 transcriptional factors (TFs). In contrast, TFs involved in the regulation of retrotransposons differentially expressed in each pair of MZ twins did not converge to specific TFs. These results imply that there are distinct regulatory mechanisms for expression between annotated genes and retrotransposons. Our findings provide basic information for further understanding of how gene and retrotransposon expression is influenced by factors shared or nonshared by MZ twins.

Materials and Methods

Participants

Participants were recruited from the registry of the Center for Twin Research at Osaka University (C. Honda et al., 2019). Three pairs of MZ twins (T1, T2, and T3) participated in this study (Table 1). Individuals in the pairs were designated by sample ID with lower case of 'a' or 'b'. To evaluate divergence of gene expression in an individual along with time, we recruited two participants, T1-a and T2-a, who agreed to provide blood samples at multiple time points. T1-a was 21 years old, having blood samples taken at four time points (Wave1–Wave4). T2-a was 51 years old, having samples taken at two time points (Wave1 and Wave2). Sample IDs of temporal samples from an individual can be designated with a suffix such as '2y2w', which means about 2 years and 2 weeks after Wave1. Difference of expression profile between temporal samples represents divergence within individuals. As T1 and T2 were youths (15−24 years) and adults (25−64 years) respectively, based on the definition of Statistics Canada (Statistics Canada, 2017), we then recruited the third pair as T3-a and T3-b, being 66 years old, from seniors (65 years and over) to determine whether divergence within MZ twin pairs becomes larger with aging. Written informed consent was obtained from all participants, and the Ethics Committee of Osaka University approved the protocol (No. 696). Blood samples were taken at 9:00 after over 12 h of fasting. Genomic DNA was isolated from peripheral blood mononuclear cells using a commercial kit (QIAamp DNA Mini Kit; QIAGEN, Hilden, Germany). Zygosity was confirmed via perfect matching of 15 short tandem repeat (STR) loci using the PowerPlex® 16 System (Promega, Madison, WI, USA).

Cap Analysis of Gene Expression (CAGE)

Total RNA was extracted from peripheral blood. To assess quality, Bioanalyzer (Agilent) was used to ensure that RIN (RNA integrity number) was over 7.0, and A260/280 and A260/230 ratios were over 1.7. First-strand cDNA was transcribed to the 5′ end of capped RNAs and attached to CAGE barcode tags. Sequenced CAGE tags were mapped to the human GRCh38 genome after discarding ribosomal or non-A/C/G/T base containing RNAs. Reads were quality-filtered using fastp (Chen et al., 2018) and mapped using HISAT2 (Kim et al., 2019). Reads with mapping quality (MAPQ) ≥ 20 were collected. Strand-specific coverage of CAGE tags was counted using bedtools at a single base-resolution (Quinlan & Hall, 2010). Then, CAGE tags were classified based on their genomic locations as transcripts. The 5'-capping site of annotated gene transcripts lies on the genomic region of transcription start sites (TSSs) defined by FANTOM5 (Abugessaisa et al., 2017). Similarly, HERV and LINE transcripts were defined based on the annotated regions in the RepeatMasker (https://www.repeatmasker.org/). The number of 5'-capping sites of transcripts within a given region were counted in total by using bigWigAverageOverBed (Kent Utility Tools: https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/). Counts were normalized as tags per million (TPM) after scaling with normalization factors calculated using Relative Log Expression (RLE) by using the edgeR package of R 4.2.2 (Robinson et al., 2009).

Motif Discovery and Identification of Transcription Factors

To perform motif enrichment analysis, we selected the top 200 transcripts showing the highest magnitude of absolute fold change (|FC|) as differentially expressed transcripts and

https://doi.org/10.1017/thg.2023.42 Published online by Cambridge University Press

Table 1. Characteristics of MZ twin samples

			Wave1[a]	Wave2[b]		Wave3[c]		Wave4[d]	
ID[e]	Gender	Zygosity	Age	ID[e]	Days from Wave1	ID[e]	Days from Wave2	ID[e]	Days from Wave3
T1-a	Male	MZ	21	T1-a-2y	734	T1-a-2y2w	14	T1-a-2y4w	14
T1-b	Male	MZ	21						
T2-a	Female	MZ	51	T2-a-3y	1235				
T2-b	Female	MZ	51						
T3-a	Female	MZ	66						
T3-b	Female	MZ	66						

[a]Time point of the first blood sample taken from individuals.
[b]Time points of the second blood sample taken from individuals.
[c]Time points of the third blood sample taken from individuals.
[d]Time points of the fourth blood sample taken from individuals.
[e]Sample IDs used in this study.

3000 background transcripts with the lowest |FC| from a comparison within MZ twin pairs. Genomic sequences of a defined region with both sides of flanking 500 bases were retrieved from the GRCh38 genome using bedtools (Quinlan & Hall, 2010). Six to 20 bases in length of consensus sequences enriched in the retrieved sequences of differentially expressed transcripts were discovered using HOMER (Heinz et al., 2010); the p value cut-off was <.01. These sequences were matched against the human motif database (HOCOMOCO Human v11CORE) to identify associated TFs using the Tomtom program of MEME Suite 5.5.1 (https://meme-suite.org/meme/tools/tomtom); the E-value cut-off was <0.05.

Cluster Analysis and Statistics

We used Spearman's rank correlation of expression levels of annotated genes and retrotransposons as an indicator of the resemblance between the samples. Spearman's rank correlations were calculated between a pair of samples using the R 4.2.2, with significance threshold of $p = .001$ (Bonferroni corrected). The distance between samples was defined as: 1 – correlation. A dendrogram of hierarchical clustering was generated using Ward's method in R 4.2.2. Violin plots were generated using the NumPy, Pandas, Matplotlib, and Seaborn packages of Python 3.7.3. One-way analysis of variance (ANOVA) and Student's t test were performed, with significance thresholds of $p = .05$ and $p = .016$ (Bonferroni corrected) respectively.

Results

To evaluate divergence of gene and retrotransposon expression within MZ twin pairs, between unrelated individuals, and within individuals at different time points, we recruited participants as described in Materials and Methods. Table 1 summarizes the biological properties of participants and time points of blood sampling for CAGE. T1-a and T1-b were male MZ twins at 21 years old at Wave1. T2-a and T2-b were female MZ twins at 51 years old at Wave1. To investigate expression divergence within individuals, additional sampling at different time points was performed for T1-a (Wave2, Wave3, and Wave4) and for T2-a (Wave2). The third pair, T3, was recruited to check whether expression divergence becomes larger with aging. T3-a and T3-b were female MZ twins aged 66 years at Wave1. Mapping statistics of raw FASTQ reads against the GRCh38 reference genome are presented in Supplementary Table S1. Over 95% of reads were mapped to the genome, and over 83% of reads were uniquely mapped in the samples.

Divergent Expression of Annotated Gene Transcripts

The number of transcripts from regions corresponding to annotated TSSs was counted at a single-base resolution by CAGE. All correlations of annotated gene expression were significant (Supplementary Table S2, $p < .001$, Bonferroni corrected). We then performed hierarchical clustering based on Spearman's correlations for annotated gene transcripts (Figure 1A). As expected, the dendrogram of annotated gene transcripts showed a monophyletic clade for each twin pair (Figure 1A), indicating that annotated gene transcripts are influenced by factors shared with each twin pair. Although a previous study has found that the number of genes differentially expressed in MZ twin pairs increases with age (Viñuela et al., 2018), we did not detect increased divergence of annotated gene expression with age (correlations within T1-a and T1-b, T2-a and T2-b, and T3-a and T3-b were 0.817, 0.820, 0.818 respectively).

We categorized correlations into four groups based on the blood relationship and time points between two samples: (1) inter-individual (referred to as Unrelated), which represents comparison between unrelated individuals at the Wave1; (2) intra-MZ twin pairs (referred to as MZ), which represents comparison within MZ twins at the Wave1; (3) intra-individual (referred to as IND), which represents comparison within individuals; (4) remained, which represents comparison between a temporal sample (Wave2, Wave3, or Wave4) with unrelated individual. In the following sections, a group, to which the correlation belongs, is represented as a superscript, except for the remained (4) group (Supplementary Table S2).

Correlations of annotated gene expression differed between the three groups ($p < .05$, ANOVA; Supplementary Table S3). $\text{Correlation}^{\text{Unrelated}}$ was lower than $\text{correlation}^{\text{MZ}}$ and $\text{correlation}^{\text{IND}}$ ($p < .016$, Student's t test), but $\text{correlation}^{\text{MZ}}$ and $\text{correlation}^{\text{IND}}$ were comparable (Supplementary Table S3 and Figure 1B). These results clearly demonstrate that MZ twins have similar expression profiles of annotated genes despite nonshared factors.

https://doi.org/10.1017/thg.2023.42 Published online by Cambridge University Press

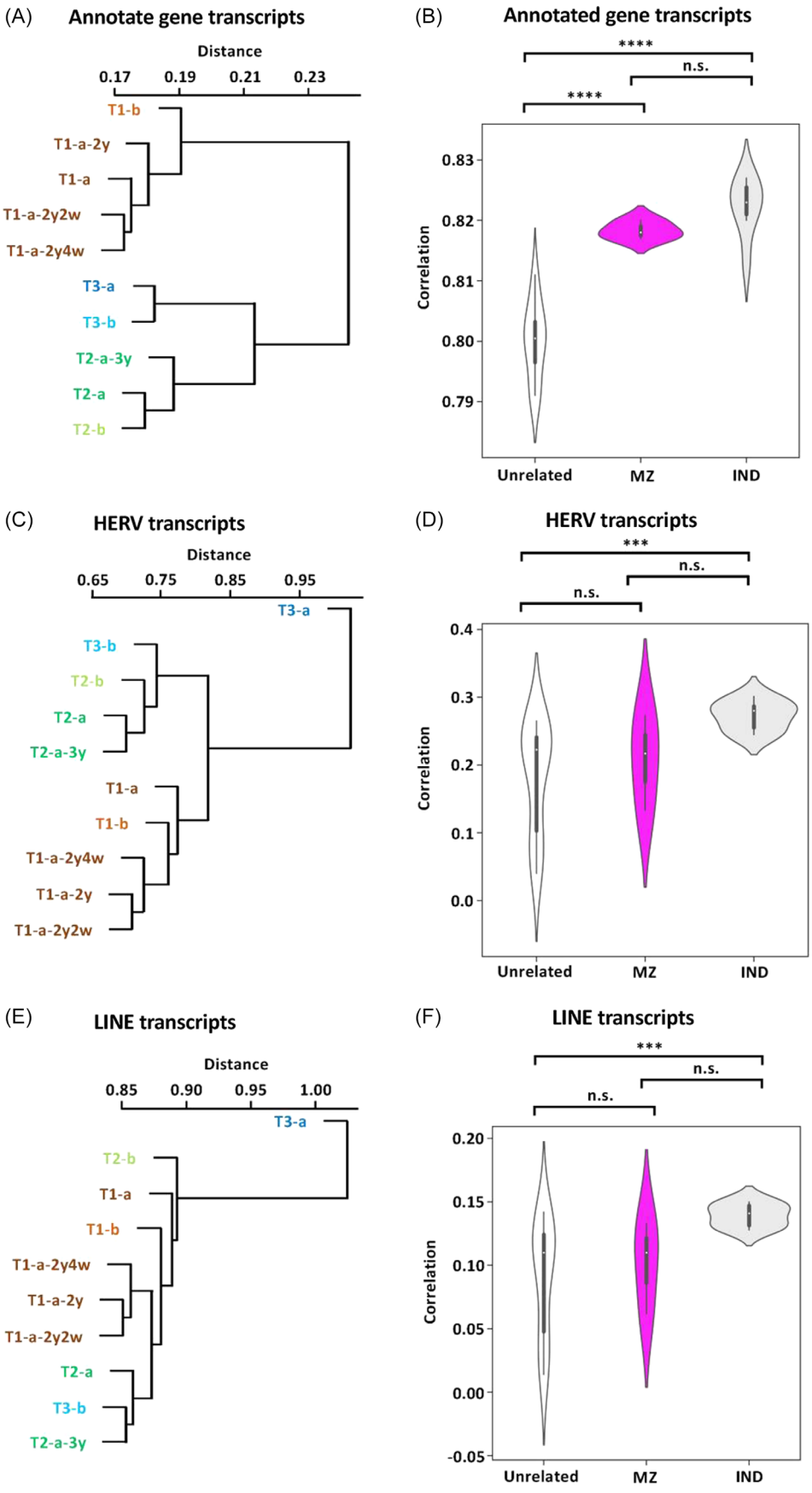

Figure 1. Hierarchical clustering of gene expression profiles and violin plots of correlations between two samples. Dendrogram of annotate gene (A), HERV (C), and LINE (E) transcripts are depicted. Sample IDs from T1, T2, and T3 are depicted in brown, green, and blue respectively. 'a' or 'b' in the ID represents the individual in a MZ twin pair with dark or light in similar color. Sample IDs without suffix represent samples at the time point Wave1. Temporal samples from an individual with a suffix are as follows. T1-a-2y, two years after the time point of Wave1 of the T1-a. T1-a-2y2w, two weeks after the time point of Wave2 of T1-a. T1-a-2y4w, four weeks after the time point of Wave2 of T1-a. T2-a-3y, three years after the time point of Wave1 of T2-a. Correlations of annotated gene (B), HERV (D), and LINE (F) transcripts were categorized by blood relationships and time points between two samples (left, Unrelated: comparison between unrelated individuals; middle, MZ: comparison within MZ twin pairs; right, IND: comparison within individuals) and their distributions are plotted. ***$p < .005$; ****$p < .001$; n.s., nonsignificant. *P* value thresholds were corrected by Bonferroni method.

https://doi.org/10.1017/thg.2023.42 Published online by Cambridge University Press

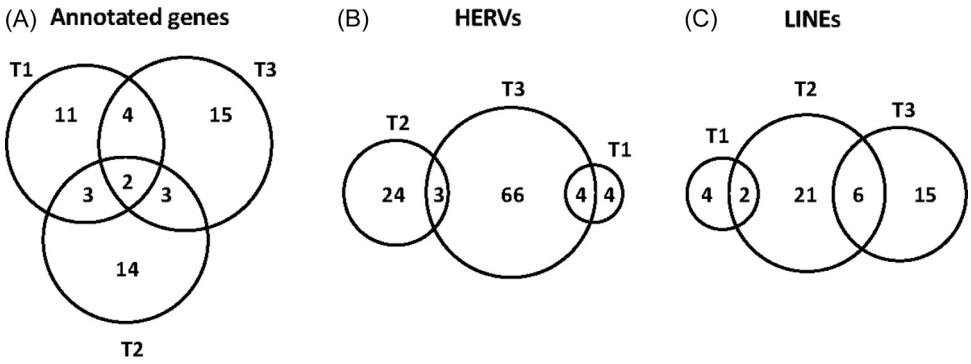

Figure 2. Venn diagrams of the number of TFs involved in the regulation of annotated genes and retrotransposons differentially expressed in T1, T2, and T3. (A) annotated genes. (B) HERVs. (C) LINEs.

Divergent Expression of Retrotransposon Transcripts

Next, we focused on the expression profiles of HERV and LINE transcripts. All correlation of HERV and LINE expression were significant (Supplementary Tables S4 and S5, $p < .001$, Bonferroni corrected). We performed an ANOVA following Student's t test for correlations of annotated gene, HERV, and LINE transcript expression (Supplementary Table S6). We found that the mean correlation of annotated gene expression was higher than those of HERV and LINE expression ($p < .016$). Because the mean correlation was higher than those of HERV and LINE expression even within MZ twin pairs, these results suggest that retrotransposon expression is more susceptible to nonshared factors, such as different environment, than annotated gene expression.

The dendrogram of HERV expression showed high divergence for T3, although T1 and T2 still constituted monophyletic clades (Figure 1C). $\text{Correlation}^{\text{Unrelated}}$ and $\text{correlation}^{\text{MZ}}$ of HERV expression seemed more divergent than $\text{correaltion}^{\text{IND}}$ (Figure 1D), although only the difference between mean $\text{correlation}^{\text{Unrelated}}$ and mean $\text{correlation}^{\text{IND}}$ was significant ($p < .016$). The dendrogram of LINE expression revealed that no MZ twin pairs constitute a monophyletic clade (Figure 1E). $\text{Correlation}^{\text{Unrelated}}$ and $\text{correlation}^{\text{MZ}}$ of LINE expression seemed more divergent than $\text{correaltion}^{\text{IND}}$, although only the difference between mean $\text{correlation}^{\text{Unrelated}}$ and mean $\text{correlation}^{\text{IND}}$ was significant ($p < .016$), similarly to those of HERV expression (Figure 1F). Thus, HERV and LINE expression profiles appear to diverge even within genetically identical MZ twin pairs, confirming the impact of nonshared factors on their expression.

Identification of TFs Binding to Enriched Consensus Motifs

Finally, we searched for consensus sequences to identify TFs that could be potentially influenced by nonshared factors. For annotated gene transcripts, we discovered 89, 79, and 82 consensus sequences enriched around the TSSs of DEGs in T1, T2, and T3 respectively. After comparing against the human motif database, we identified 20, 22, and 24 TFs for T1, T2, and T3 respectively. Two TFs were common among the three MZ twin pairs (Figure 2A): Sp1 and Sp2. Our findings suggest that these two TFs might be involved in divergence of annotated gene expression influenced by nonshared factors.

We performed the same analysis for HERVs and LINEs. For HERVs, 56, 72, and 264 consensus sequences enriched in the retrieved genomic regions of differentially expressed HERVs were discovered in T1, T2, and T3 respectively. By comparing against the human motif database, 8, 27, and 73 TFs were identified for MZ T1, T2, and T3 respectively (Figure 2B). For LINEs, 66, 67, and 135 enriched consensus sequences and 6, 29, and 21 TFs were discovered for T1, T2, and T3 respectively. However, consensus motifs or TFs that influence the divergence of HERV or LINE expression in each MZ twin did not converge to any specific motifs or TFs (Figure 2C). These results imply that a wide variety of pathways are involved in divergence of retrotransposon expression within MZ twin pairs.

Discussion

This is a pilot study using a small sample size to obtain a hint for understanding of the impact of factors shared or nonshared by MZ twin pairs on gene and retrotransposon expression. Consistent with earlier reports (Cheung et al., 2003; Sharma et al., 2005), we confirmed that annotated gene expression was more highly correlated within MZ twin pairs than between unrelated individuals (Figure 1A and 1B). These results indicate that annotated gene expression is substantially influenced by shared factors within MZ twin pairs. Shared factors within MZ twin pairs are composed of genetic and shared environment factors. As previous twin studies have suggested that nongenetic factors appear to have a larger influence on gene expression on a genomewide scale (Grundberg et al., 2012; Ouwens et al., 2020; Powell et al., 2012; Wright et al., 2014), shared environment factors and/or unknown shared factors may have an impact on annotated gene expression. We also found that divergence of annotated gene expression within individuals was comparable to that within MZ twin pairs (Figure 1A and 1B). These results suggest that MZ twin pairs are identical regarding divergence of annotated gene expression. A previous twin study has suggested that the number of genes differentially expressed within MZ twin pairs increased with age (Viñuela et al., 2018). However, we did not detect increased divergence of annotated gene expression with age. As this discrepancy may be due to the small sample size in our study, further investigation regarding age-related divergence in gene expression using a larger sample size is required.

Retrotransposon expression profiles were highly divergent (Figure 1C, 1E, and Supplementary Table S6), implying that retrotransposon expression is substantially influenced by nonshared factors and is, therefore, considered to be regulated by mechanisms distinct from annotated gene expression. Among retrotransposons, HERVs comprise at least 8% of the human genome (Kazazian & Moran, 2017) and are thought to be remnants of ancestral viral infections (Durnaoglu et al., 2021; Suntsova et al., 2015). While the majority are defective, several phylogenetically distinct HERVs can produce retroviral proteins (Bannert & Kurth, 2004; Chan et al., 2019). HERVs have been shown to contribute to various biological events, such as cell-cell fusion during placentation (Mi et al., 2000), establishment and/or maintenance of the

https://doi.org/10.1017/thg.2023.42 Published online by Cambridge University Press

pluripotency of embryonic stem cells (Lu et al., 2014; Macfarlan et al., 2012), and activation of immune-related genes (Chuong et al., 2016). Dysregulation of HERVs may be involved in various human diseases, including autoimmune disorders, neurological disorders, infectious diseases, and cancer (Gonzalez-Cao et al., 2016; Suntsova et al., 2015). Among LINEs accounting for approximately 17% of the human genome, LINE-1 (L1) is the only active class of autonomous retrotransposons in humans (Richardson et al., 2015). L1-encoded ORF2p contains a reverse transcriptase domain and an endonuclease domain, which are responsible for L1-mediated retrotransposition. L1 can lead to retrotransposition of L1 itself and other mobile elements, such as *Alu* and SVA (Dewannieux et al., 2003; Raiz et al., 2012). Although its biological significance remains unclear, L1 is also hypothesized to contribute to biological processes; for example, the L1 antisense promoter plays a role in cell proliferation (T. Honda et al., 2020). On the other hand, L1 dysregulation has detrimental effects on health and host genome stability, with multiple studies linking L1 to various cancers (T. Honda, 2016; Kemp & Longworth, 2015; Sciamanna et al., 2014; Sciamanna et al., 2018). For example, the role of de novo L1 insertions is reported in hepatocellular carcinoma (Shukla et al., 2013); Kaposi's sarcoma-associated herpesvirus causes cellular transformation via L1 activation (Nakayama et al., 2019); and endonuclease activity of ORF2p can cause DNA damage through the formation of double-strand breaks (Gasior et al., 2006; Kines et al., 2014). Our results showed that the expression profiles of HERVs and LINEs diverged even within MZ twin pairs. Considering the physiological significance of retrotransposons, differences in their expression likely contribute to intra-MZ phenotypic divergence.

We also identified consensus TFs that were potentially influenced by differences within MZ twin pairs. Thus, versatile GC-rich element-binding TFs, Sp1 and Sp2, were identified in all intra-MZ twin comparisons (Figure 2A). The Sp1 and Sp2, belonging to the Sp subfamily of the Sp/KLFs superfamily, were identified as potential TFs involved in gene expression influenced by environmental factors. They are characterized by a highly conserved DNA-binding domain near the C-terminus, which recognizes the GC (consensus sequence: GGGGCGGGG) and GT/CACC (GGTGTGGGG) boxes (Philipsen & Suske, 1999). Sp/KLFs are expressed ubiquitously, activating or repressing the transcription of many genes in response to physiological and pathological stimuli. For example, Sp family members are key mediators of gene expression induced by hormones (Solomon et al., 2008), which are notably influenced by environmental stimuli. Sp1 is involved in the development of atherosclerosis, including inflammation, lipid metabolism, plaque stability, vascular smooth muscle cell proliferation, and endothelial dysfunction (Jiang et al., 2022). Together with previous studies, our findings suggest that the Sp subfamily plays a major role in differential gene expression within MZ twin pairs. In contrast, we did not identify any common TFs for retrotransposons (Figure 2B and 2C). These results suggest that difference in retrotransposon expression within MZ twin pairs is influenced by a combination of wide variety of TFs. Our findings will benefit future efforts to further clarify the molecular mechanisms of how gene and retrotransposon expression is influenced by environmental differences.

In conclusion, this pilot study using a small sample size provides a hint for understanding of how the expression of genes and retrotransposons is influenced by factors shared or nonshared by MZ twins. We demonstrated that retrotransposon expression is more variable than gene expression even within MZ twin pairs, suggesting that factors nonshared by MZ twin pairs, such as environmental or stochastic influence, play substantial roles in retrotransposon expression. As retrotransposon activation potentially causes human diseases such as cancers, we hypothesized that environmental factors may contribute to disease development via retrotransposon expression modulation. If this hypothesis is true, targeting relevant environmental factors may prevent disease development. Undoubtedly, retrotransposons exhibit divergent expression even within genetically identical MZ twin pairs, which may contribute to any phenotypic discordance within MZ twin pairs. Further studies on the molecular basis of the divergence of gene and retrotransposon expression influenced by nonshared factors within MZ twin pairs will improve our understanding of phenotypic diversity and benefit the development of effective therapeutic interventions for human diseases.

Supplementary material. For supplementary material accompanying this paper visit https://doi.org/10.1017/thg.2023.42

Data availability. Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments. CAGE was conducted in collaboration with DNAFORM (Japan).

Financial support. This study was supported in part by JSPS KAKENHI (grant numbers 15K08496, 18H02664, 18K19449) and grants from the Takeda Science Foundation (T.H.).

Ethical standards. This study was approved by the Ethics Committee of Osaka University (No. 696). Written informed consent was obtained from participants. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Competing interests. The authors have no conflicts of interest to declare.

References

Abugessaisa, I., Noguchi, S., Hasegawa, A., Harshbarger, J., Kondo, A., Lizio, M., Severin, J., Carninci, P., Kawaji, H., & Kasukawa, T. (2017). FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Scientific Data*, *4*, 170107. https://doi.org/10.1038/sdata.2017.107

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T. C., Alhendi, A. M. N., Arner, P., Axton, R., Baillie, J. K., Beckhouse, A., Bodega, B., Briggs, J., Brombacher, F., … **Hayashizaki, Y.** (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, *347*, 1010–1014. https://doi.org/10.1126/science.1259418

Bannert, N., & Kurth, R. (2004). Retroelements and the human genome: New perspectives on an old relation. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 14572–14579. https://doi.org/10.1073/pnas.0404838101

Bhat, A., Ghatage, T., Bhan, S., Lahane, G. P., Dhar, A., Kumar, R., Pandita, R. K., Bhat, K. M., Ramos, K. S., & Pandita, T. K. (2022). Role of transposable elements in genome stability: Implications for health and disease. *International Journal of Molecular Sciences*, *23*, 7802. MDPI. https://doi.org/10.3390/ijms23147802

Boomsma, D., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics*, *3*, 872–882. https://doi.org/10.1038/nrg932

Chan, S. M., Sapir, T., Park, S.-S., Rual, J.-F., Contreras-Galindo, R., Reiner, O., & Markovitz, D. M. (2019). The HERV-K accessory protein Np9

controls viability and migration of teratocarcinoma cells. *PLOS ONE*, *14*, e0212970. https://doi.org/10.1371/journal.pone.0212970

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chénais, B. (2022). Transposable elements and human diseases: Mechanisms and implication in the response to environmental pollutants. *International Journal of Molecular Sciences*, *23*, 2551. https://doi.org/10.3390/ijms23052551

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M., & Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, *33*, 422–425. https://doi.org/10.1038/ng1094

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, *351*, 1083–1087. https://doi.org/10.1126/science.aad5497

Dewannieux, M., Esnault, C., & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, *35*, 41–48. https://doi.org/10.1038/ng1223

Durnaoglu, S., Lee, S. K., & Ahnn, J. (2021). Human endogenous retroviruses as gene expression regulators: Insights from animal models into human diseases. *Molecules and Cells*, *44*, 861–878. https://doi.org/10.14348/MOLCELLS.2021.5016

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., De Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*, 462–470. https://doi.org/10.1038/nature13182

Gasior, S. L., Wakeman, T. P., Xu, B., & Deininger, P. L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology*, *357*, 1383–1393. https://doi.org/10.1016/j.jmb.2006.01.089

Gonzalez-Cao, M., Iduma, P., Karachaliou, N., Santarpia, M., Blanco, J., & Rosell, R. (2016). Human endogenous retroviruses and cancer. *Cancer Biology and Medicine*, *13*, 483–488. https://doi.org/10.20892/j.issn.2095-3941.2016.0080

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T. P., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, S. Y., Glass, D., Travers, M., Min, J. L., Ring, S., Ho, K., … Spector, T. D. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*, 1084–1089. https://doi.org/10.1038/ng.2394

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Honda, C., Watanabe, M., Tomizawa, R., & Sakai, N. (2019). Update on Osaka University Twin Registry: An overview of multidisciplinary research resources and biobank at Osaka University Center for Twin Research. *Twin Research and Human Genetics*, *22*, 597–601. https://doi.org/10.1017/thg.2019.70

Honda, T. (2016). Links between human LINE-1 retrotransposons and hepatitis virus-related hepatocellular carcinoma. *Frontiers in Chemistry*, *4*, 21. https://doi.org/10.3389/fchem.2016.00021

Honda, T., Nishikawa, Y., Nishimura, K., Teng, D., Takemoto, K., & Ueda, K. (2020). Effects of activation of the LINE-1 antisense promoter on the growth of cultured cells. *Scientific Reports*, *10*, 22136. https://doi.org/10.1038/s41598-020-79197-y

Honda, T., Takemoto, K., & Ueda, K. (2019). Identification of a retroelement-containing human transcript induced in the nucleus by vaccination. *International Journal of Molecular Sciences*, *20*, 2875. https://doi.org/10.3390/ijms20122875

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*, *376*, eabk3112. https://doi.org/10.1126/science.abk3112

Jiang, J. F., Zhou, Z. Y., Liu, Y. Z., Wu, L., Nie, B. Bin, Huang, L., & Zhang, C. (2022). Role of Sp1 in atherosclerosis. *Molecular Biology Reports*, *49*, 9893–9902. https://doi.org/10.1007/s11033-022-07516-9

Kazazian, H. H., & Moran, J. V. (2017). Mobile DNA in health and disease. *New England Journal of Medicine*, *377*, 361–370. https://doi.org/10.1056/nejmra1510092

Kemp, J. R., & Longworth, M. S. (2015). Crossing the LINE toward genomic instability: LINE-1 retrotransposition in cancer. *Frontiers in Chemistry*, *3*, 68. https://doi.org/10.3389/fchem.2015.00068

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kines, K. J., Sokolowski, M., De Haro, D. L., Christian, C. M., & Belancio, V. P. (2014). Potential for genomic instability associated with retrotranspositionally-incompetent L1 loci. *Nucleic Acids Research*, *42*, 10488–10502. https://doi.org/10.1093/nar/gku687

Konkel, M. K., & Batzer, M. A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in Cancer Biology*, *20*, 211–221. https://doi.org/10.1016/j.semcancer.2010.03.001

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921. https://doi.org/10.1038/35057062

Lu, X., Sachs, F., Ramsay, L. A., Jacques, P. É., Göke, J., Bourque, G., & Ng, H. H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural and Molecular Biology*, *21*, 423–425. https://doi.org/10.1038/nsmb.2799

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D., & Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*, 57–63. https://doi.org/10.1038/nature11244

McAdams, T. A., Rijsdijk, F. V., Zavos, H. M. S., & Pingault, J. B. (2021). Twins and causal inference: Leveraging nature's experiment. *Cold Spring Harbor Perspectives in Medicine*, *11*, a039552. https://doi.org/10.1101/cshperspect.a039552

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., Lavallie, E., Tang, X., Edouard, P., Howes, S., Keith, J. C. Jr., McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved. *Nature*, *403*, 785–789. https://doi.org/10.1038/35001608

Nakayama, R., Ueno, Y., Ueda, K., & Honda, T. (2019). Latent infection with Kaposi's sarcoma-associated herpesvirus enhances retrotransposition of long interspersed element-1. *Oncogene*, *38*, 4340–4351. https://doi.org/10.1038/s41388-019-0726-5

Osman, M., Mistry, A., Keding, A., Gabe, R., Cook, E., Forrester, S., Wiggins, R., Di Marco, S., Colloca, S., Siani, L., Smith, D. F., Aebischer, T., Kaye, P. M., & Lacey, C. J. (2017). A third generation vaccine for human visceral leishmaniasis and post kala azar dermal leishmaniasis: First-in-human trial of ChAd63-KH. *PLoS Neglected Tropical Diseases*, *11*, e0005527. https://doi.org/10.1371/journal.pntd.0005527

Ouwens, K. G., Jansen, R., Nivard, M. G., van Dongen, J., Frieser, M. J., Hottenga, J. J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., Franke, L., Heijmans, B. T., 't Hoen, P. A. C., van Meurs, J., Brooks, A. I.; BIOS Consortium; Penninx, B. W. J. H., & Boomsma, D. I. (2020). A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. *European Journal of Human Genetics*, *28*, 253–263. https://doi.org/10.1038/s41431-019-0511-5

Philipsen, S., & Suske, G. (1999). A tale of three fingers: The family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Research*, *27*, 2991–3000. https://doi.org/10.1093/nar/27.15.2991

Piégu, B., Bire, S., Arensburger, P., & Bigot, Y. (2015). A survey of transposable element classification systems — A call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular*

Phylogenetics and Evolution, *86*, 90–109. https://doi.org/10.1016/j.ympev.2015.03.009

Powell, J. E., Henders, A. K., McRae, A. F., Wright, M. J., Martin, N. G., Dermitzakis, E. T., Montgomery, G. W., & Visscher, P. M. (2012). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Research*, *22*, 456–466. https://doi.org/10.1101/gr.126540.111

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W. H., Löwer, R., & Schumann, G. G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Research*, *40*, 1666–1683. https://doi.org/10.1093/nar/gkr863

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The influence of LINE-1 and SINE Retrotransposons on mammalian genomes. *Microbiology Spectrum*, *3*. https://doi.org/10.1128/microbiolspec.mdna3-0061-2014

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Sammarco, I., Pieters, J., Salony, S., Toman, I., Zolotarov, G., & Lafon Placette, C. (2022). Epigenetic targeting of transposon relics: Beating the dead horses of the genome? *Epigenetics*, *17*, 1331–1344. https://doi.org/10.1080/15592294.2021.2022066

Sciamanna, I., Gualtieri, A., Piazza, P. V., & Spadafora, C. (2014). Regulatory roles of LINE-1-encoded reverse transcriptase in cancer onset and progression. *Oncotarget*, *5*, 8039–8051. https://doi.org/10.18632/oncotarget.2504

Sciamanna, I., Serafino, A., & Spadafora, C. (2018). LINE-1-encoded reverse transcriptase in the genesis and therapy of cancer. *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, *22*(2). https://doi.org/10.4267/2042/68766

Sharma, A., Sharma, V. K., Horn-Saban, S., Lancet, D., Ramachandran, S., & Brahmachari, S. K. (2005). Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays. *Physiological Genomics*, *21,* 117–123. https://doi.org/10.1152/physiolgenomics.00228.2003

Shukla, R., Upton, K. R., Muñoz-Lopez, M., Gerhardt, D. J., Fisher, M. E., Nguyen, T., Brennan, P. M., Baillie, J. K., Collino, A., Ghisletti, S., Sinha, S., Iannelli, F., Radaelli, E., Dos Santos, A., Rapoud, D., Guettier, C., Samuel, D., Natoli, G., Carninci, P., … **Faulkner, G. J.** (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*, *153*, 101–111. https://doi.org/10.1016/j.cell.2013.02.032

Solomon, S. S., Majumdar, G., Martinez-Hernandez, A., & Raghow, R. (2008). A critical role of Sp1 transcription factor in regulating gene expression in response to insulin and other hormones. *Life Sciences*, *83*, 305–312. https://doi.org/10.1016/j.lfs.2008.06.024

Statistics Canada. (2017). Age categories, life cycle groupings. https://www.statcan.gc.ca/en/concepts/definitions/age2

Suntsova, M., Garazha, A., Ivanova, A., Kaminsky, D., Zhavoronkov, A., & Buzdin, A. (2015). Molecular functions of human endogenous retroviruses in health and disease. *Cellular and Molecular Life Sciences*, *72*, 3653–3675. https://doi.org/10.1007/s00018-015-1947-6

Viñuela, A., Brown, A. A., Buil, A., Tsai, P. C., Davies, M. N., Bell, J. T., Dermitzakis, E. T., Spector, T. D., & Small, K. S. (2018). Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Human Molecular Genetics*, *27*, 732–741. https://doi.org/10.1093/hmg/ddx424

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T. H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., … **Boomsma, D. I.** (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, *46*, 430–437. https://doi.org/10.1038/ng.2951

https://doi.org/10.1017/thg.2023.42 Published online by Cambridge University Press