



Complex Age- and Cancer-Related Changes in Human Blood Transcriptome—Implications for Pan-Cancer Diagnostics

Fei Qi¹, Fan Gao¹, Ye Cai¹, Xueer Han¹, Yao Qi¹, Jiawen Ni¹, Jianfeng Sun², Shengquan Huang³, Shaohua Chen³, Chunlin Wu^{3*} and Philipp Kapranov^{1*}

¹School of Medicine, Institute of Genomics, Huaqiao University, Xiamen, China, ²Department of Bioinformatics, Technische Universität München, Freising, Germany, ³Department of Pathology, Second Affiliated Hospital of Fujian Medical University, Quanzhou, China

OPEN ACCESS

Edited by:

Yun Zheng,
Kunming University of Science and
Technology, China

Reviewed by:

Changning Liu,
Xishuangbanna Tropical Botanical
Garden (CAS), China
Jin Chen,
University of Kentucky, United States

*Correspondence:

Chunlin Wu
wuchunlin120@sina.com
Philipp Kapranov
philippk08@hotmail.com

Specialty section:

This article was submitted
to RNA,
a section of the journal
Frontiers in Genetics

Received: 25 July 2021

Accepted: 16 September 2021

Published: 15 October 2021

Citation:

Qi F, Gao F, Cai Y, Han X, Qi Y, Ni J,
Sun J, Huang S, Chen S, Wu C and
Kapranov P (2021) Complex Age- and
Cancer-Related Changes in Human
Blood Transcriptome—Implications for
Pan-Cancer Diagnostics.
Front. Genet. 12:746879.
doi: 10.3389/fgene.2021.746879

Early cancer detection is the key to a positive clinical outcome. While a number of early diagnostics methods exist in clinics today, they tend to be invasive and limited to a few cancer types. Thus, a clear need exists for non-invasive diagnostics methods that can be used to detect the presence of cancer of any type. Liquid biopsy based on analysis of molecular components of peripheral blood has shown significant promise in such pan-cancer diagnostics; however, existing methods based on this approach require improvements, especially in sensitivity of early-stage cancer detection. The improvement would likely require diagnostics assays based on multiple different types of biomarkers and, thus, calls for identification of novel types of cancer-related biomarkers that can be used in liquid biopsy. Whole-blood transcriptome, especially its non-coding component, represents an obvious yet under-explored biomarker for pan-cancer detection. In this study, we show that whole transcriptome analysis using RNA-seq could indeed serve as a viable biomarker for pan-cancer detection. Furthermore, a class of long non-coding (lnc) RNAs, very long intergenic non-coding (vlinc) RNAs, demonstrated superior performance compared with protein-coding mRNAs. Finally, we show that age and presence of non-blood cancers change transcriptome in similar, yet not identical, directions and explore implications of this observation for pan-cancer diagnostics.

Keywords: pan-cancer, biomarker, transcriptome, lncRNA, vlincRNA, aging, peripheral blood, liquid biopsy

INTRODUCTION

Cancer is a leading cause of death worldwide; however, it is widely known that early detection of primary tumors can significantly reduce mortality and improve outcome in cancer patients (Siegel et al., 2020). Therefore, a number of tumor screening solutions have been developed for several types of cancer, for example, colonoscopy for colon cancer, mammography for breast cancer, and others. However, such tests are limited to few specific types of cancer and are often fairly invasive. Thus, a strong need exists for non-invasive tests that can simultaneously identify multiple cancers and can be used for universal screening for the disease (Ahlquist, 2018). Indeed, several recent studies based on analysis of various components of peripheral blood have shown that such pan-cancer non-invasive

detection methods are feasible (Srivastava and Hanash, 2020). Cohen et al. developed CancerSEEK method based on a combination of detection of mutations in cell-free (cf) DNA and specific proteins in peripheral blood (Cohen et al., 2018). The method had a median sensitivity of 70% for the detection of eight common cancer types (Cohen et al., 2018). Lennon et al. later applied this approach to a cohort of ~10 K patients to identify 26 cancers undetected by the typical standard-of-care methods (Lennon et al., 2020). The study by Liu et al. based on analysis of methylation patterns of cfDNA in peripheral blood could identify 12 common cancers with a sensitivity of 67.3% and all 50 cancers tested in that study with a sensitivity of 43.9% (Liu et al., 2020). Best et al. have shown that cancer patients could be differentiated from normal patients with 96% accuracy using RNA-seq analysis of transcriptome from tumor-educated platelets (Best et al., 2015).

However, despite the obvious promise, these methods also have limitations. Most of all, the sensitivity of cancer detection is still relatively low, especially for the early-stage cancers where detection is most desirable. For example, sensitivity for stage I cancers was 18% compared with >43% for stages II and above for all 50 cancers investigated in the study by Liu et al. (2020). A similar trend was also found in the study by Cohen et al. where median sensitivity of detecting stage I cancers dropped to 43% compared with over 70% for stages II and above (Cohen et al., 2018). One likely avenue of improvement is the introduction of additional types of biomarkers into detection methods. In fact, the CancerSEEK method is based on two different types of biomarkers—cfDNA and circulating proteins (Cohen et al., 2018). Therefore, here we explored the potential of whole-blood transcriptome to detect seven types of non-blood cancers. We specifically tested the performance of protein coding and non-coding transcripts, with the latter being represented by the class of vlincRNAs. Finally, since cancer incidence correlates with age, we explored transcriptome changes caused by cancer and normal aging and their implications for transcriptome-based cancer detection.

MATERIALS AND METHODS

RNA-Seq

Peripheral blood samples were collected from 75 Chinese females between 24 and 82 years old (**Supplementary Table S1**) into Tempus Blood RNA Tubes (Thermo), and total RNA was isolated using Tempus Spin RNA Isolation Kit (Thermo) following the procedure of the manufacturer. The 75 individuals consist of 30 apparently healthy persons and 45 patients of cancer at various non-blood tissues, specifically breast, esophagus, stomach, thyroid, rectum, colon, and uterus (**Supplementary Table S1**). Construction of RNA-seq libraries was conducted by first removing globin mRNA and rRNA by Globin-Zero Gold rRNA Removal Kit followed by strand specific, lncRNA-seq protocol that included both polyA+ and polyA- RNA species and was performed by the Novogene corporation (Beijing). Sequencing was performed using the Illumina HiSeq X Ten

platform and paired-end 150-bp (PE150) strategy on a 10-gigabase (GB) scale by the Novogene corporation (Beijing).

RNA-Seq Data Analysis

Expression levels of genes were estimated based on the RNA-seq data using Salmon software (Patro et al., 2017) for the reference human transcriptome (GRCh38) from the Ensembl database (Zerbino et al., 2018) and 2,721 vlincRNA transcripts taken from previous publications (St Laurent et al., 2013; Caron et al., 2018). PCA was performed for all the 75 samples using the DESeq2 package (Love et al., 2014, 2) in R environment (R Core Team, 2020) based on the variance stabilizing transformation (Anders and Huber, 2010) of the raw read counts of genes. Five hundred genes were used for the PCA analysis, selected by the highest variance of the gene expression levels across all samples. The differential expression analyses between cancer and normal samples were performed using the DESeq2 package (Love et al., 2014, 2) in R environment (R Core Team, 2020). In the analyses, age of individuals was added as a term in the design formular (the design formular for the DESeq2 package became “~ age + phenotype”) to include it as a covarying factor and, thus, eliminate its influence. The threshold for identifying differentially expressed genes was FDR-adjusted p -value <0.1 and absolute value of \log_2 (expression fold change) > \log_2 (1.2). In differential expression analysis of the training dataset, only the 53 samples in the training dataset (see below for details) were included and 900 differentially expressed genes (DEGs) were derived (**Supplementary Table S2**). In the differential expression analysis for all samples, all the 75 samples were included and 2,124 DEGs were derived (**Supplementary Table S5**).

Genes with age-covarying expression were identified using only the 30 normal samples. A gene would be identified as age-covarying if its normalized expression levels correlated with the ages of samples (FDR-adjusted p -value <0.1, two-sided Spearman's rank test). The normalization was performed using the DESeq2 package with the “median of ratios” method (Love et al., 2014). A total of 609 genes were found as covarying with age (**Supplementary Table S4**).

Training and Evaluation of Machine Learning Models

The 75 samples were randomly split into two datasets: 1) a training dataset containing 53 samples, and 2) an independent test dataset containing 22 samples (**Supplementary Table S1**). The ratio between the numbers of normal and cancer samples was kept in the split.

The DEG-classifier, vlinc-classifier, and non-vlinc-classifier were all trained using the MLSeq package (Goksuluk et al., 2019) in R environment (R Core Team, 2020) with the SVM model of a linear kernel, “deseq-vst” preprocessing parameter and repeated k -fold cross validation ($k = 5$, repeated five times), by the raw read counts of all the 900 genes, 120 vlincRNAs, and 780 non-vlincRNA genes of the DEGs from the training dataset, respectively. A fake read count of 1 was added to all genes to avoid the problem of division by 0 in the training process.

The bench classifier and age classifier were trained using decision tree model embedded in the scikit-learn python package (Pedregosa et al., 2011). The bench classifier was trained based on the in-cancerous-status probabilities of the 53 samples from the training dataset outputted by the DEG classifier. The age classifier was trained with the above probabilities as well as the ages of the 53 individuals from the training dataset, which were classified into six groups (21–30, 31–40, 41–50, 51–60, 61–70, and 71–90; see **Supplementary Table S1**) and then one-hot encoded. The classifiers were used to predict the cancerous status of the 22 samples in the test dataset.

The performance of all classifiers was evaluated by the resulting confusion matrix, accuracy, precision, recall, and F1 score.

Co-Expression Analysis

The gene co-expression network analysis was performed using the WGCNA package (Langfelder and Horvath, 2008) in R environment (R Core Team, 2020). A total of 7,581 genes which were the union of the genes with the top 10% median absolute deviation (MAD) across all the 75 samples, the DEGs of cancer from all the 75 samples, and the age-covarying genes were included in this analysis. First, the 75 samples were clustered based on the distances calculated by the expression levels of those genes. From the clusters, two outlier samples were identified and excluded from the analysis (**Supplementary Figure S1**). Then, the gene co-expression networks were built based on the remaining 73 samples.

GO Analysis

The enrichment analyses of GO terms were performed using the clusterProfiler package (Yu et al., 2012) in R environment (R Core Team, 2020). Significantly enriched items were identified by the threshold of FDR-adjusted p -value < 0.05 .

Overlap Analysis

Overlaps between sets of genes or GO terms were analyzed using the GeneOverlap package (Shen, 2020) in R environment (R Core Team, 2020). Overlaps with the results from studies by Chatsirisupachai et al. (2019) and Peters et al. (2015) were limited to non-vlincRNA genes since vlincRNAs were not included in those studies.

RESULTS

Blood Transcriptome Profile can Potentially Serve as Pan-Cancer Biomarker

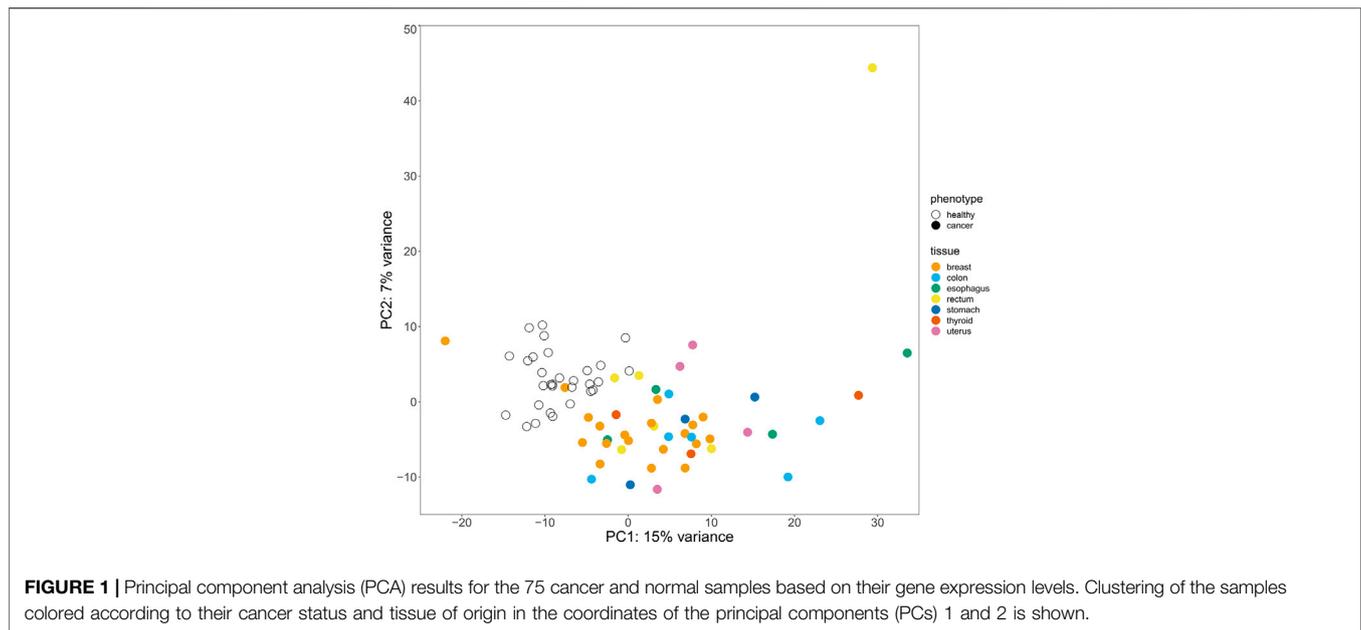
This work was based on the analysis of peripheral blood transcriptome from 75 females including 30 healthy persons aged 25–78 years old with five samples per decade of age and 45 patients with various non-blood cancers (breast, esophagus, stomach, thyroid, rectum, colon, and uterus; **Supplementary Table S1**). The relative fraction of the different cancers among the 45 samples was kept approximately similar to the occurrence of these cancers in females in China (Cao et al., 2020). Of the 43 cancers with staging information, 13 (30%), 17 (40%), 12 (28%),

and 1 (2%) were represented correspondingly by stages I, II, III, and IV (**Supplementary Table S1**). RNA from peripheral blood samples was subjected to RNA-seq analysis to estimate relative level of expression of all annotated human genes and the newly discovered class of long non-coding (lnc) RNA—vlincRNAs. These transcripts represent very long RNA molecules (minimum length of 50 kb) that are preferentially polyA– and retained in the nucleus (Kapranov et al., 2010; St Laurent et al., 2013; St Laurent et al., 2016). The functions of most of vlincRNAs are unknown; however, some of these transcripts have been directly implicated in cellular senescence (Lazorthes et al., 2015) and control of DNA replication (Heskett et al., 2020). Also, using expression analysis, these transcripts have been implicated in the control of cell cycle, carcinogenesis, pluripotency, and early development (St Laurent et al., 2013; St Laurent et al., 2016). The rationale for the inclusion of this type of transcripts in this study was high cell type specificity of their expression (Kapranov et al., 2010; St Laurent et al., 2013; St Laurent et al., 2016) and ability to discriminate various types of cancers (Caron et al., 2018).

Principal component analysis (PCA) (Bro and Smilde, 2014) performed based on expression levels of both protein coding mRNAs and vlincRNAs revealed that the cancer and normal samples could be discriminated quite well even though clustering by the tissue of origin of cancer was not apparent (**Figure 1**). This result indicated that the blood transcriptome could, in principle, serve as a pan-cancer biomarker even if the tissue of origin information may not be easily attainable from this approach. Therefore, we then built a computational classifier to predict cancerous status based on the blood transcriptome profiles using an SVM machine learning model with no feature selection procedures (Goksuluk et al., 2019). We first randomly split the 75 samples into two datasets: the training dataset containing 53 samples and the independent test dataset containing 22 samples (**Supplementary Table S1**). The ratio between the numbers of normal and cancer samples was kept in the split. Then, we performed differential expression analysis for the training dataset and identified a total of 900 differentially expressed genes (DEGs) between the cancer and normal samples (**Supplementary Table S2**). The read counts for the 900 DEGs from the training dataset samples were used as input into an SVM (support vector machine) learning model with a linear kernel, and the model was trained using repeated k -fold cross validation ($k = 5$, repeated five times). The trained classifier was evaluated on the test dataset, and resulted in 0.77 accuracy, 0.72 precision, 1.0 recall, and 0.84 F1 score (“DEG-classifier” in **Table 1**; confusion matrix in **Supplementary Table S3**). This result proved the assumption that the peripheral blood transcriptome profile could be used as a feature to predict the cancerous status of an individual.

Aging and Cancer Have Similar Effects on Blood Transcriptome

Another crucial property of the individuals in our dataset, which has been found to correlate with occurrence of various cancers, is age (Henry et al., 2010; Aramillo Irizar et al., 2018; Parikh et al., 2018; Chatsirisupachai et al., 2019; Rozhok and DeGregori, 2019).



Indeed, the average age of the 45 cancer patients used in this study was ~54 years. To investigate whether the age parameter improves the prediction of cancerous status, we built another classifier (termed “age-classifier”) by training a decision tree model with the ages of the 53 individuals in the training dataset and their probabilities of the cancerous status derived from the DEG-classifier. The decision tree model was used here since ensemble models like tree-based models perform better in capturing distinguishing patterns from complex data (Garg and Mishra, 2018). To reveal the true contribution of the age parameter to the prediction (i.e., ruling out the improvement purely due to the additional training process), a “bench-classifier” was also built by the same procedure but only trained with the probabilities of the cancerous status of a sample, and was used as the benchmark in the comparison. As shown in **Table 1**, the additional decision tree model improved the prediction, however, the addition of the age parameter did not (confusion matrices in **Supplementary Table S3**). A likely explanation of these results is that aging-associated blood transcriptomic changes were present in the 900 DEGs used to train the models, and thus have already been learnt by the DEG-classifier.

This prompted us to investigate whether age- and cancer-related transcriptome changes indeed shared some similarities. As the first step, we identified 609 genes covarying with age

(termed aging-related genes; **Supplementary Table S4**) based on RNA-seq data from the 30 normal samples (**Supplementary Table S1**). We first validated our findings by comparing the age-related genes with aging-associated transcriptome signatures from two studies. The first study by Peters et al. identified 1,497 genes differentially expressed with chronological age in peripheral blood transcriptome from 14,983 individuals (Peters et al., 2015). The second study by Chatsirisupachai et al. identified 1) 1,260 cellular senescence signature genes by a meta-analysis of 20 replicative senescence microarray datasets mostly based on cultured human cells, and 2) age-related transcriptome changes in 26 different human tissues (Chatsirisupachai et al., 2019).

As seen in **Figure 2A**, genes up- or downregulated with age identified by us showed highly significant overlap with the corresponding genes from the study of Peters et al. Furthermore, the vast majority of genes in common showed the same direction of change (**Figure 2A**). Likewise, similar results were obtained when age-related DEGs identified by us were compared with the cellular senescence signature genes of Chatsirisupachai et al. (**Figure 2B**). We then compared age-related transcriptome changes found in this study in blood with those found by Chatsirisupachai et al. in 17 of the 26 non-blood tissues having >50 aging-associated DEGs. As seen in **Figures 2C,D**, the blood age-related transcriptomic changes showed similarity with tissue-specific changes in 5/17 tissues (blood vessel, brain, breast, heart, and prostate). The only tissue that showed opposite changes was the uterus. This is in line with the previous finding that the aging-associated transcriptomic changes in uterus behaved differently with the other tissues and were opposite to the expression of the cellular senescence signature genes (Chatsirisupachai et al., 2019). Taken together, these results strongly suggest that the transcriptome profiles derived in this study are consistent with the previous findings. However, the significance of overlap of the age-related

TABLE 1 | Performance of different classifiers in prediction of cancerous status.

Classifier	Accuracy	Precision	Recall	F1 score
DEG-classifier	0.77	0.72	1.0	0.84
Vlinc-classifier	0.86	0.86	0.92	0.89
Non-vlinc-classifier	0.73	0.68	1.0	0.81
Age-classifier	0.91	0.87	1.0	0.93
Bench-classifier	0.91	0.87	1.0	0.93

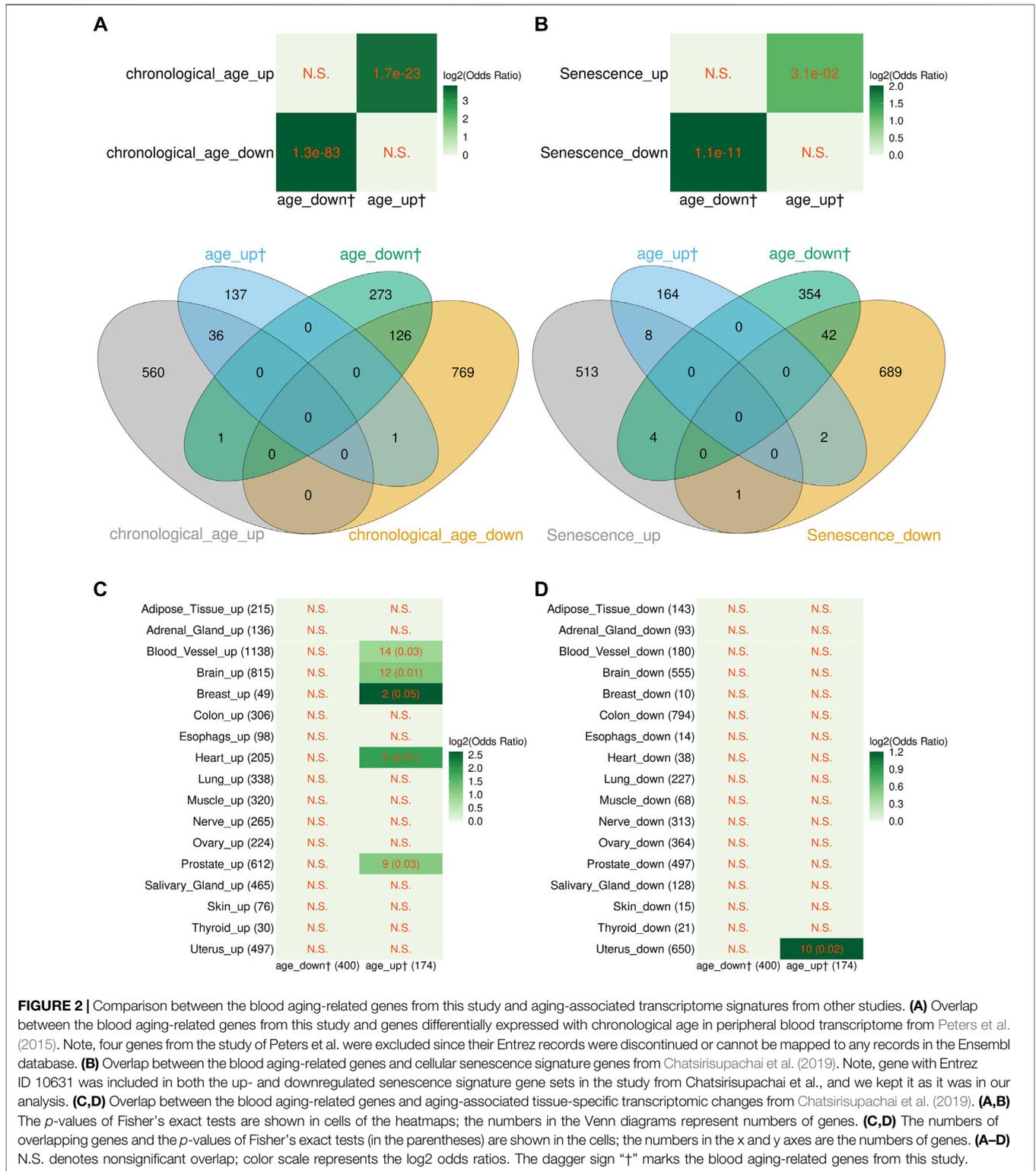
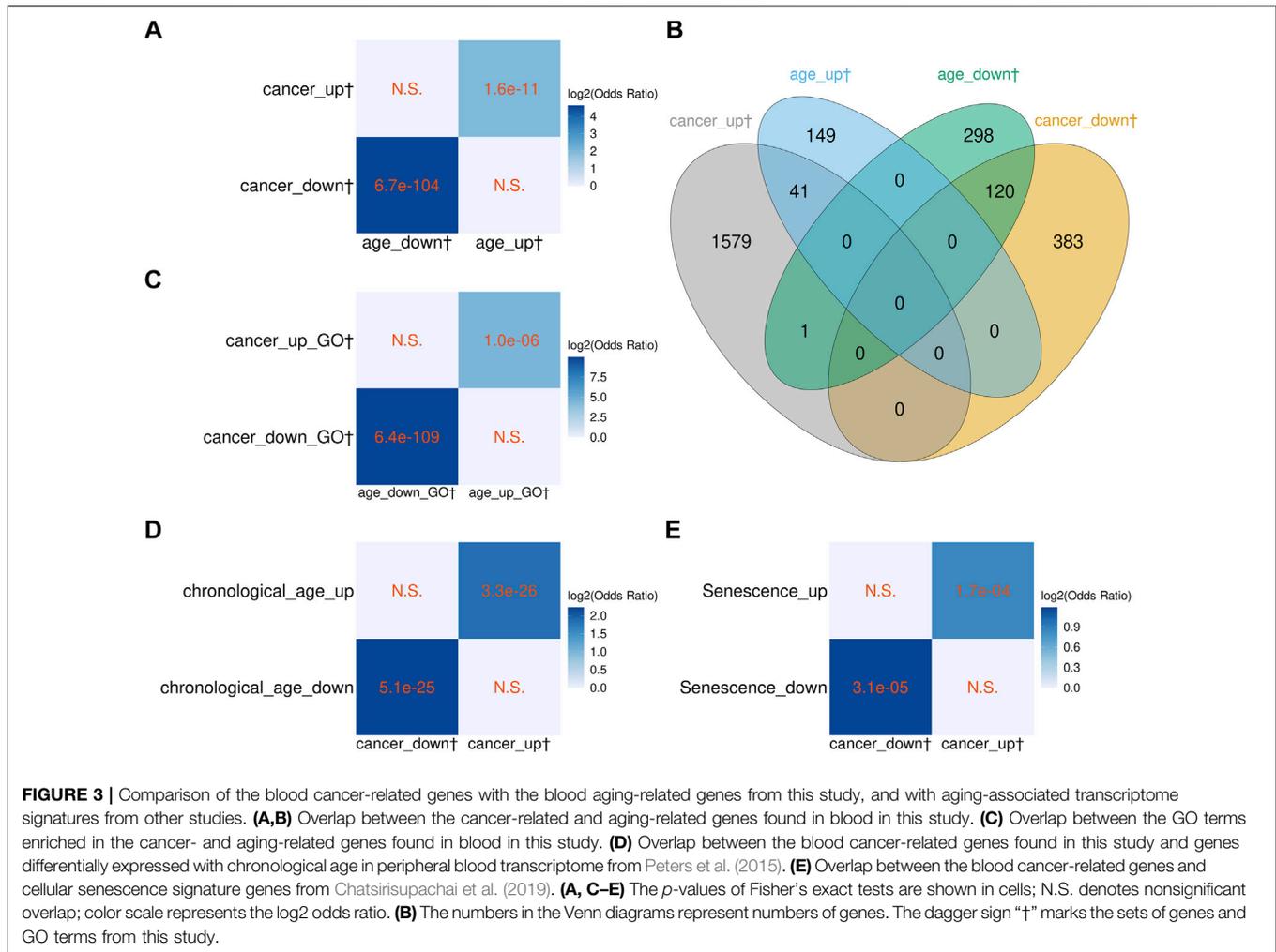


FIGURE 2 | Comparison between the blood aging-related genes from this study and aging-associated transcriptome signatures from other studies. **(A)** Overlap between the blood aging-related genes from this study and genes differentially expressed with chronological age in peripheral blood transcriptome from Peters et al. (2015). Note, four genes from the study of Peters et al. were excluded since their Entrez records were discontinued or cannot be mapped to any records in the Ensembl database. **(B)** Overlap between the blood aging-related genes and cellular senescence signature genes from Chatsirisupachai et al. (2019). Note, gene with Entrez ID 10631 was included in both the up- and downregulated senescence signature gene sets in the study from Chatsirisupachai et al., and we kept it as it was in our analysis. **(C,D)** Overlap between the blood aging-related genes and aging-associated tissue-specific transcriptomic changes from Chatsirisupachai et al. (2019). **(A,B)** The *p*-values of Fisher’s exact tests are shown in cells of the heatmaps; the numbers in the Venn diagrams represent numbers of genes. **(C,D)** The numbers of overlapping genes and the *p*-values of Fisher’s exact tests (in the parentheses) are shown in the cells; the numbers in the x and y axes are the numbers of genes. **(A–D)** N.S. denotes nonsignificant overlap; color scale represents the log2 odds ratios. The dagger sign “†” marks the blood aging-related genes from this study.

genes from our study was much higher with the genes found by Peters et al. compared with the senescence- and age-related genes found in cultured cells and non-blood tissues (Figure 2), suggesting that age-related transcriptome changes depend on tissue type.

The significant overlaps between the age-related genes found by us and those found in previous studies indicated that biologically meaningful transcriptomic changes could be faithfully detected in our relatively small sample dataset. Therefore, as the next step, we explored the relationship

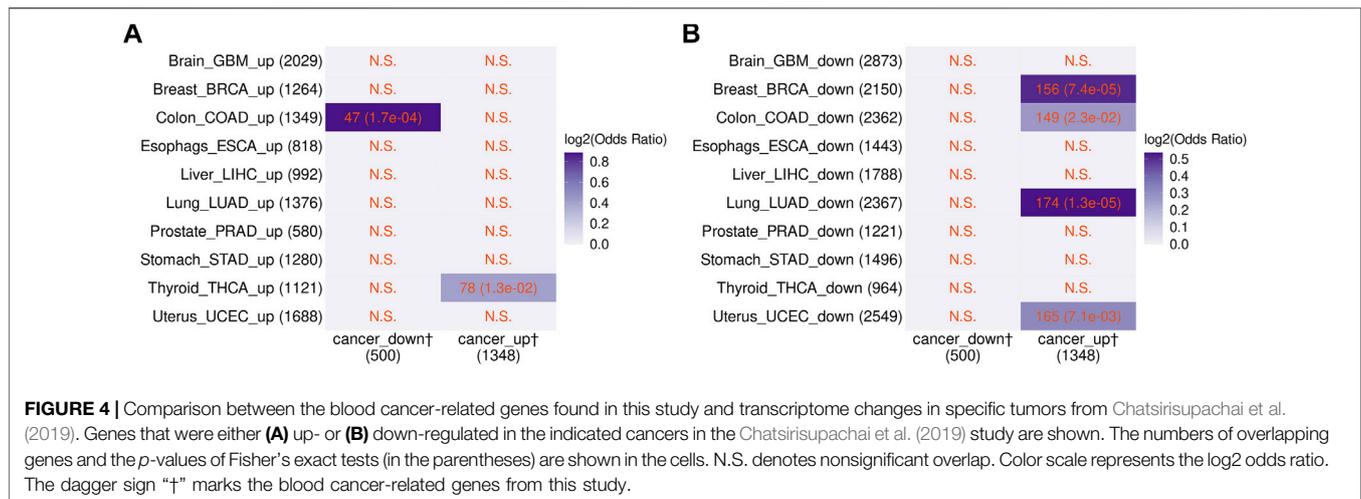


between cancer- and age-related changes in blood transcriptomes. We identified 2,124 genes differentially expressed between the cancer and normal samples (termed cancer-related genes; **Supplementary Table S5**) and compared them with the 609 aging-related genes (both groups included protein-coding genes and vlincRNAs). As can be seen in **Figures 3A,B**, aging and cancer-associated transcriptomic changes followed similar trajectory. Specifically, of the 503 genes downregulated in cancer, 120 were also negatively correlated with age and none was positively correlated with age. Reciprocally, of the 1,621 gene upregulated in cancer, 41 were also positively correlated with age, while only one was negatively correlated. Furthermore, the overlaps of genes changing in the same direction were statistically significant (**Figure 3A**).

We also investigated the overlap between aging- and cancer-related genes in the space of GO terms enriched in genes up/downregulated with aging or cancer (**Supplementary Tables S6, S7**, respectively). As seen in **Figure 3C**, the GO terms (biological process) enriched in aging- and cancer-related genes showed significant overlaps, indicating that the blood transcriptome

changes associated with aging and cancer are also similar in terms of functional relatedness. We then compared the 1,848 cancer-related non-vlincRNA genes with the age-related genes found by Chatsirisupachai et al. and Peters et al. and, as expected, found statistically significant overlap with both studies for genes changing the same direction, i.e., up- or downregulated with age (**Figures 3D,E**). Again, the overlap was higher with Peters et al. (**Figure 3D**) likely because the same tissue type was used.

Chatsirisupachai et al. reported that age-related transcriptome changes had a tendency to be opposite to those found in tumors (Chatsirisupachai et al., 2019). Therefore, based on the results above, it would be expected that transcriptome changes in the blood of cancer patients would also be opposite to those occurring in tumors. To test this, we compared the 1,848 cancer-related non-vlincRNA genes found in the blood with genes up- and downregulated in tumors originating from 10 non-blood tissues found by Chatsirisupachai et al. Indeed, we found that the cancer-related transcriptome changes in the blood were opposite to the changes in tumors from 4/10 tissues (breast, colon, lung, and



uterus; **Figure 4**). Only thyroid cancers showed changes in the same direction, in line with transcriptomic changes in cancers of thyroid origin being different from those in tumors originating from other tissues (Chatsirisupachai et al., 2019).

All in all, these results show that aging- and cancer-associated transcriptomic changes found in this study in the blood were similar to each other, and also to age-related signatures found in other studies, particularly in the blood, but also in other tissues. However, cancer-associated transcriptome changes in the blood were opposite to those found in actual tumors.

In-Depth Analysis of Similarities in Blood Transcriptome Profiles Associated With Cancer Status and Normal Aging

To gain further insight into blood transcriptome changes associated with normal aging and cancer status, we conducted a co-expression network analysis for the aging- and cancer-associated whole-body level transcriptomic changes. Co-expression network analysis does not depend on fixed fold change thresholds and can identify genes with consistent, although low-magnitude, changes in expression. Genes were sorted by MAD across all the 75 samples and the top 10% of them were included in this analysis. Furthermore, the above identified aging- and cancer-related genes were also included resulting in a total of 7,581 genes. The analysis revealed 12 gene expression modules (**Figure 5; Supplementary Table S8**; see *Materials and Methods* for details). Then, for each module, its eigengene—a vector representing the overall expression of genes within the module—was used to calculate the correlation (Pearson's *r*) between the module and the age and cancerous status of the samples. As seen in **Figure 5**, with the threshold of *p*-value <0.05 and absolute value of correlation coefficient >0.4, 4 of the 12 modules showed significant correlation with age and/or cancerous status: 1) modules blue and turquoise correlated with both the age of normal samples and the cancerous status; and 2) modules yellow and black correlated only with the cancerous

status of samples. Of the 3,834 genes contained in these four modules, most (~89%) were represented by the genes whose expression changed in both age- and cancer-related fashion: the up- (turquoise module) and downregulated genes (blue module) accounted for correspondingly 2,320 (60.5%) and 1,092 (28.5%) genes (**Figure 5**). This result strongly supports the findings above that most genes whose expression in blood changes with age also have concomitant change in response to cancer in non-blood tissues and vice versa. None of the 12 modules showed correlation with the tissue of origin of the cancer (**Supplementary Figure S2**) suggesting that this finding is not limited to cancer derived from a specific tissue.

To further understand the functional properties of these modules, GO enrichment analysis was performed on genes found within the modules. Enriched GO terms were identified with adjusted *p*-value threshold of <0.05 (**Supplementary Table S9**). The enriched GO terms of each module were then summarized and shown in **Figure 6**. Interestingly, the yellow module containing genes downregulated only in cancer showed enrichment in 15 various DNA repair-related GO terms, including DNA damage checkpoint, cellular response to DNA damage stimulus, double-strand break repair, recombinational repair, etc., and involving 37 genes (**Supplementary Table S10**). Furthermore, the module was enriched in functions related to DNA replication, cell cycle, and p53 signal transduction. The black module containing genes upregulated only in cancer showed enrichment in functions related to chromosome organization, immune system development, response to cytokine, cell cycle, and gene silencing. On the other hand, the turquoise and blue modules had very characteristic GO profiles. Genes upregulated in both aging and cancer genes (module turquoise) were enriched in various immune and stress related functions such as immune response, cytokine production, stress-activated signal transduction, inflammatory response, response to oxidative stress, and others (**Figure 6**). Genes downregulated in both aging and

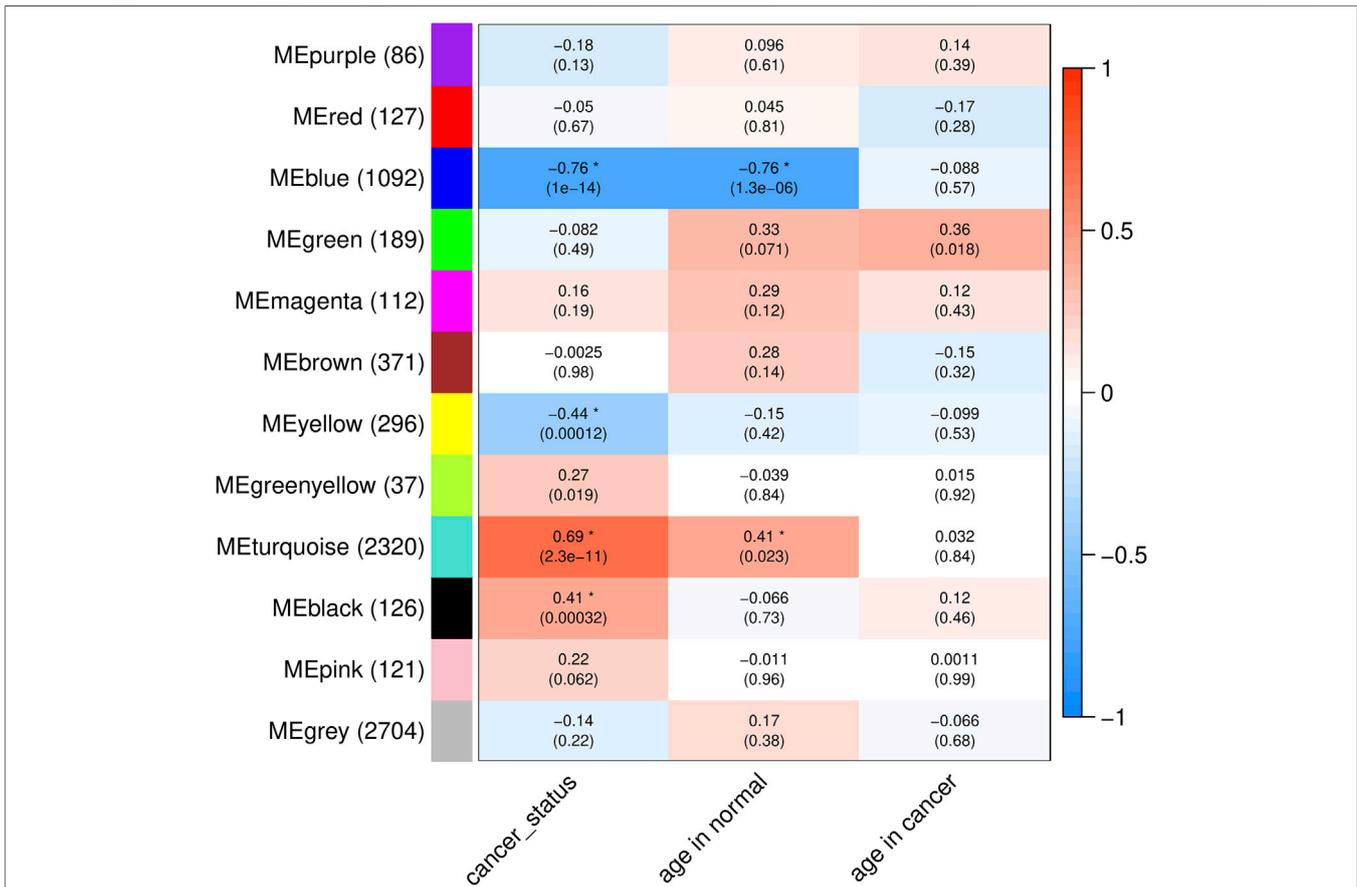


FIGURE 5 | Summary of gene co-expression modules. The colors and numbers in the cells represent the correlation coefficients (Pearson’s *r*) between the eigengenes of the modules and the traits of the samples (cancer status, ages of normal samples, and ages of cancer samples, shown on the bottom). The numbers in the parentheses are the *p*-values of two-sided Pearson’s tests. Asterisks mark the significant correlations under the threshold of *p*-value <0.05 and absolute value of correlation coefficient >0.4. Names of modules and numbers of genes in each module are shown on the left.

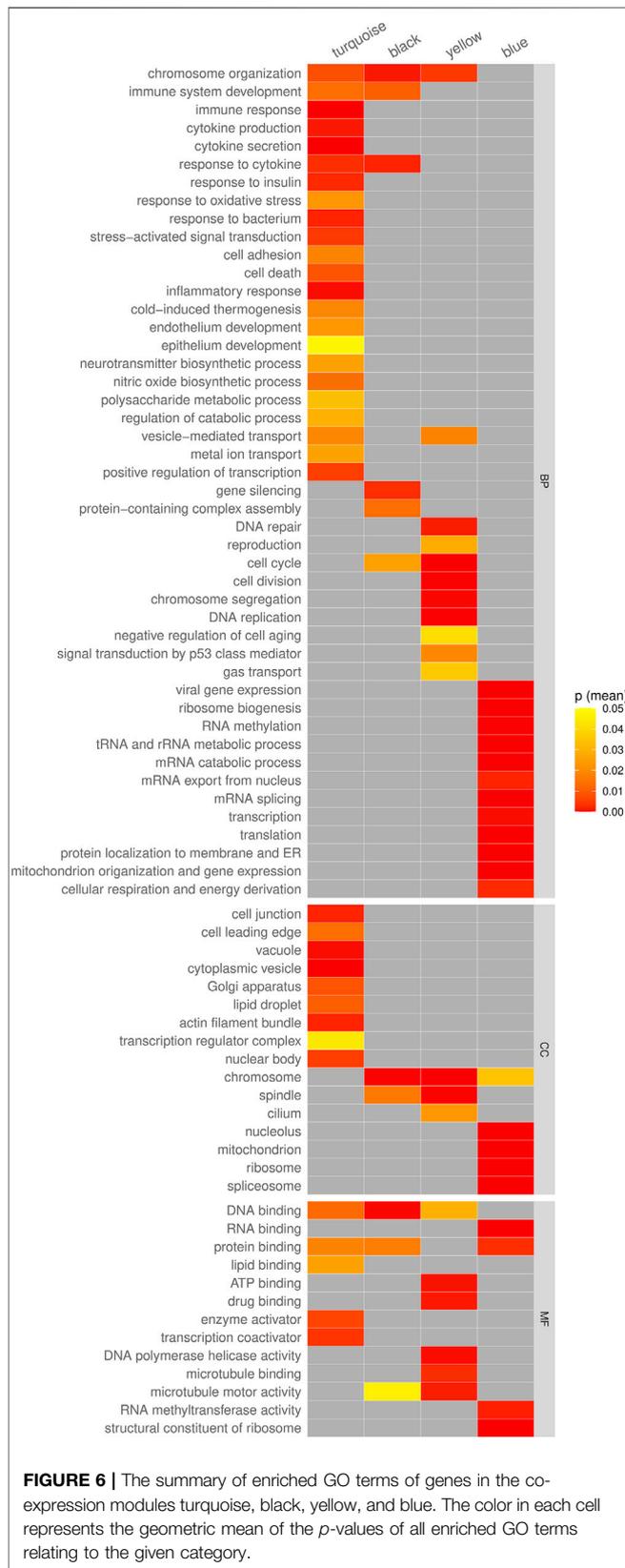
cancer (module blue) were enriched in various functions associated with normal functioning of cell such as translation, transcription, RNA metabolism and respiration, and energy production (Figure 6).

Strikingly, the magnitude of age- and cancer-related changes were different between the genes up- (module turquoise) and downregulated (module blue) in both processes (Figure 7; Supplementary Figure S3). The downregulated genes exhibited a tendency to have higher association with aging than the upregulated ones (Figure 7; Supplementary Figure S3). For this analysis, the Spearman correlations shown in Figure 7 between gene expression levels and age for the downregulated genes were first inverted (multiplied by -1), and then compared with the unmodified Spearman correlations for the upregulated genes (Supplementary Figure S3). The former had a tendency to be higher than the latter with *p*-value of 8.22e⁻⁹⁰ (Wilcoxon Rank Sum Test). On the other hand, the upregulated genes had a tendency to have a higher fold change difference in cancer than the downregulated ones (Supplementary Figure S3; *p*-value = 1.16e⁻⁶⁷, Wilcoxon rank

sum test). In other words, among the genes associated with aging and cancer, downregulation of expression was more pronounced in normal aging than upregulation; however, the opposite was the case for the cancer-related genes. Consistent with this observation, overlap with the senescence- and chronological age-related genes from the studies of Chatsirisupachai et al. and Peters et al. was much more significant for the down- than upregulated genes (Figures 2A,B). All in all, these results suggest that while cancer and normal aging share transcriptomic profiles, these profiles are not identical and can be used to separate cancerous status from normal aging.

VlincRNAs Potentially Represent Superior Biomarkers for Liquid Biopsies

Previously, vlincRNAs were shown to represent transcripts with a high cell type-specific pattern of expression (Kapranov et al., 2010; St Laurent et al., 2013; St Laurent et al., 2016) and a promising class of biomarkers for classification of human cancers (Caron et al., 2018). To test the utility of these



transcripts as cancer biomarkers for liquid biopsies, we compared predictions made based either on the vlincRNAs or non-vlincRNA genes. The 900 genes that were used to train the SVM model consisted of 120 vlincRNAs (**Supplementary Figure S4**) and the 780 non-vlincRNA genes. These two groups of genes were then used to train the classifier separately, while the training and evaluation processes were unchanged; and thus, two classifiers were generated: the “vlinc-classifier” and the “non-vlinc-classifier.” These two classifiers were evaluated on the test dataset, and the results are shown in **Table 1**. The vlinc-classifier resulted in a better outcome than the non-vlinc-classifier in terms of accuracy, precision, and F1 score (confusion matrices in **Supplementary Table S3**). Moreover, vlinc-classifier was also more accurate than classifier based on all DEGs (the DEG-classifier; **Table 1**). Notably, only one cancer sample was wrongly classified as normal among the five classifiers listed in **Table 1** (sample X551 by the vlinc-classifier; **Supplementary Table S3**), likely due to this sample being early stage I of cancer (**Supplementary Table S1**). This result suggests that the sensitivity of blood-based transcriptome to detect early cancers is fairly high (3/4, 4/4, 4/4, and 1/1 for stage I, II, III and IV, respectively).

Furthermore, vlincRNAs showed a clear tendency to be upregulated in cancer. The 2,124 cancer-related genes contained 276 vlincRNAs of which 273 were upregulated and only three downregulated. Overall, the 273 vlincRNAs represented ~17% of 1,621 genes upregulated in cancer compared with only 0.6% (3/503) of the downregulated genes. The enrichment of vlincRNAs among the upregulated genes was significant at *p*-value $< 2.2e^{-16}$ (chi-squared test). VlincRNAs also had a similar albeit weaker tendency to be upregulated in normal aging. Of the 609 age-related genes, vlincRNAs represented 19/419 or 4.5% of downregulated genes and 16/190 or 8.4% of the upregulated ones (*p*-value = 0.085, chi-squared test). Finally, the tendency toward upregulation in both cancer and normal aging was also evident based on the co-expression analysis. VlincRNAs represented 401/2,320 or ~17% of genes in the module turquoise and only 18/1,092 or ~1.6% of genes in the module blue (*p*-value $< 2.2e^{-16}$, chi-squared test).

DISCUSSION

This work represents a proof-of-principle study showing that analysis of whole transcriptome of peripheral blood can serve as a basis or at least as a component of a pan-cancer diagnostic test. While the ability to distinguish tissue of origin of cancer is not obvious from these results, the transcriptome approach appears to be able to detect early-stage cancers quite well. We further show that lncRNAs, represented by the subclass of vlincRNAs used in this study, embody a superior class of biomarkers compared with the protein-coding mRNAs. Possible reasons for it could be known high cell-type specificity of expression

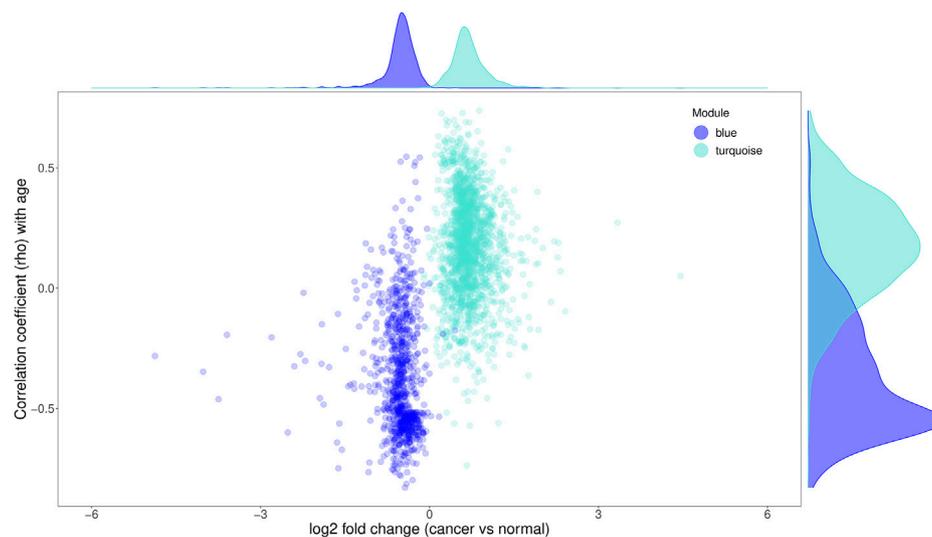


FIGURE 7 | The difference in the magnitude of age- and cancer-related changes between the up- (module turquoise) and downregulated (module blue) genes. The x-axis is the log₂ value of the fold change between cancer and normal samples of each gene; the y-axis is the Spearman correlation coefficients (ρ) between gene expression levels and age.

of lncRNAs in general (Derrien et al., 2012) and vlincRNAs in particular (Kapranov et al., 2010; St Laurent et al., 2013; St Laurent et al., 2016). These results suggest that transcriptome-based pan-cancer blood-based diagnostics should also include these transcripts. Since vlincRNAs like other lncRNAs tend to be polyA⁻, it means that RNA-seq assays should be based on whole, rRNA-depleted RNA rather than just on the polyA⁺ fraction.

In the process of this work, we discovered that normal aging and cancer induce somewhat similar changes in transcriptome. For example, genes up- or downregulated in cancer also had a tendency to be, respectively, up- or downregulated in aging and vice versa while the reciprocal situation (i.e., upregulation in one condition while being downregulated in another) almost never happened. On one hand, this similarity is not surprising since cancer incidence correlates with age (Campisi, 2013). On the other hand, however, the transcriptomic profiles associated with these two conditions also had remarkable differences. Blood transcriptome changes associated with normal aging were dominated by downregulation of various functions associated with normal cell functioning and to a lesser extent, upregulation of immune and stress-related function. Reciprocally, the cancer-related changes were most dominated by upregulation of the immune and stress-related functions and to a lesser extent, downregulation of functions related to normal functioning of cells. Interestingly, immune-related functions were also enriched in genes upregulated only in cancers (the black module). Considering statistically significant overlap of aging-related genes found in the blood in this study with those found in non-blood tissues in other studies, it is reasonable to suggest that our observations relate not only to changes happening in the peripheral blood, but also those taking place throughout the body. In other words, our results reveal complex balance of the two

types of opposing phenomena—slowdown of normal cell functioning and increase in immune and stress related functions—with the final outcome of this complex interplay potentially signifying whether aging is “normal” or “cancer-prone.”

Currently, it is not clear whether these transcriptome-derived phenomena underlie mechanistic reasons for differences in “normal” vs. “cancer-prone” aging, and if they do, which phenomenon is primary. Still, the higher enrichment of immune-related functions is consistent with numerous previous studies linking inflammation and cancer [reviewed in Greten and Grivennikov (2019)]. In fact, secondary messengers produced during inflammation (e.g., cytokines and growth factors) can promote a number of processes associated with tumorigenesis such as cell growth, de-differentiation, and others [reviewed in Greten and Grivennikov (2019)]. In this respect, the lesser slowdown of normal cellular functions observed in cancers compared with the normal aging is consistent with the growth-promoting effects of the higher levels of inflammation associated with cancers that counteracts the general cellular slowdown effects of normal aging.

Interestingly, vlincRNAs had a tendency to be upregulated in both processes, especially in cancers. While implicated in control of certain biological processes (see above), mostly these transcripts represent yet a not-well-understood group of transcripts just like most lncRNAs (Gao et al., 2020). This study further underlines potential roles of these transcripts in aging and cancers. Even though, the mechanisms of their involvement in these processes are not known, this work strongly argues that not only vlincRNAs should be included in the biomarker discovery screens, but they could also

represent yet unknown components involved in normal aging and cancer-related processes. Notably, we also found downregulation of DNA repair-related genes in cancer patients. DNA damage is widely assumed to play a central role in aging, cancer, and other age-related diseases (Hoeijmakers, 2009; Maynard et al., 2015; Ou and Schumacher, 2018). At present, it is not clear whether individuals prone to cancer have intrinsically low levels of expression of the DNA repair-related transcripts, or their downregulation is a part of transcriptome changes caused by cancer in blood cells.

Overall, our work suggests that transcriptome changes happening during normal aging and cancer are both similar and quite complex. Additional studies with much larger cohorts are needed to fully address the utility of peripheral blood transcriptome as pan-cancer diagnostic marker and to ensure that the resulting model is applicable to the general population that includes both genders, multiple races, and ethnicities and does not suffer from overfitting. The final test of the applicability of transcriptome-based or transcriptome-including early cancer diagnostics methods would then need to be carried out on a large cohort of people with no prior knowledge of cancer to ascertain whether these methods can detect relatively small number of early cancers similar to the study by Lennon et al. based on CancerSEEK (Lennon et al., 2020). However, the current proof-of-principle study provides the foundation framework for these future endeavors by showing that precise transcriptome-based pan-cancer diagnostics is feasible and it requires comprehensive profiling of all cellular RNAs, protein-coding and non-coding, polyA+ and polyA-, rather than focusing on a select few biomarker genes.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the GSA-Human repository, accession number HRA001249.

REFERENCES

- Ahlquist, D. A. (2018). Universal Cancer Screening: Revolutionary, Rational, and Realizable. *npj Precision Onc* 2, 23. doi:10.1038/s41698-018-0066-x
- Anders, S., and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Nat. Prec* 1, 1. doi:10.1038/npre.2010.4282.2
- Aramillo Irizar, P., Schäuble, S., Esser, D., Groth, M., Frahm, C., Priebe, S., et al. (2018). Transcriptomic Alterations During Ageing Reflect the Shift from Cancer to Degenerative Diseases in the Elderly. *Nat. Commun.* 9, 327. doi:10.1038/s41467-017-02395-2
- Best, M. G., Sol, N., Kooi, I., Tannous, J., Westerman, B. A., Rustenburg, F., et al. (2015). RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell* 28, 666–676. doi:10.1016/j.ccell.2015.09.018
- Bro, R., and Smilde, A. K. (2014). Principal Component Analysis. *Anal. Methods* 6, 2812–2831. doi:10.1039/C3AY41907J
- Campisi, J. (2013). Aging, Cellular Senescence, and Cancer. *Annu. Rev. Physiol.* 75, 685–705. doi:10.1146/annurev-physiol-030212-183653

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics review boards of the Second Affiliated Hospital of Fujian Medical University and School of Medicine, Huaqiao University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PK conceived the project and supervised the analytical and the wet lab parts of project. FQ performed all the bioinformatics analyses. JS contributed to the methodology of the machine learning part. SH, SC, and CW performed all the clinical part of the project. FG, YC, XH, YQ, and JN performed the sample preparation. PK and FQ wrote the manuscript.

FUNDING

FQ is supported by the National Natural Science Foundation of China (Grant No. 32000462) and the Scientific Research Funds of Huaqiao University. PK is supported by the Natural Science Foundation of Fujian Province, China (Grant No. 2020J02006) and the Scientific Research Funds of Huaqiao University.

ACKNOWLEDGMENTS

We wish to thank Chaofan Huang (Huaqiao University) for excellent administrative support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.746879/full#supplementary-material>

- Cao, M., Li, H., Sun, D., and Chen, W. (2020). Cancer Burden of Major Cancers in China: A Need for Sustainable Actions. *Cancer Commun.* 40, 205–210. doi:10.1002/cac2.12025
- Caron, M., St-Onge, P., Drouin, S., Richer, C., Sontag, T., Busche, S., et al. (2018). Very Long Intergenic Non-Coding RNA Transcripts and Expression Profiles are Associated to Specific Childhood Acute Lymphoblastic Leukemia Subtypes. *PLoS One* 13, e0207250. doi:10.1371/journal.pone.0207250
- Chatsirisupachai, K., Palmer, D., Ferreira, S., and Magalhães, J. P. (2019). A human Tissue-Specific Transcriptomic Analysis Reveals a Complex Relationship Between Aging, Cancer, and Cellular Senescence. *Aging Cell* 18, e13041. doi:10.1111/acel.13041
- Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., et al. (2018). Detection and Localization of Surgically Resectable Cancers With a Multi-Analyte Blood Test. *Science* 359, 926–930. doi:10.1126/science.aar3247
- Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression. *Genome Res.* 22, 1775–1789. doi:10.1101/gr.132159.111

- Gao, F., Cai, Y., Kapranov, P., and Xu, D. (2020). Reverse-Genetics Studies Of Lncrnas-What We Have Learnt and Paths Forward. *Genome Biol.* 21, 93. doi:10.1186/s13059-020-01994-5
- Garg, D., and Mishra, A. (2018). Bayesian Regularized Neural Network Decision Tree Ensemble Model for Genomic Data Classification. *Appl. Artif. Intellig.* 32, 463–476. doi:10.1080/08839514.2018.1483115
- Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., et al. (2019). MLSeq: Machine Learning Interface for RNA-Sequencing Data. *Comp. Methods Programs Biomed.* 175, 223–231. doi:10.1016/j.cmpb.2019.04.007
- Greten, F. R., and Grivnenikov, S. I. (2019). Inflammation and Cancer: Triggers, Mechanisms, and Consequences. *Immunity* 51, 27–41. doi:10.1016/j.immuni.2019.06.025
- Henry, C. J., Marusyk, A., Zaberezhnyy, V., Adane, B., and DeGregori, J. (2010). Declining Lymphoid Progenitor Fitness Promotes Aging-Associated Leukemogenesis. *Proc. Natl. Acad. Sci.* 107, 21713–21718. doi:10.1073/pnas.1005486107
- Heskett, M. B., Smith, L. G., Spellman, P., and Thayer, M. J. (2020). Reciprocal Monoallelic Expression of ASAR lncRNA Genes Controls Replication Timing of Human Chromosome 6. *RNA* 26, 724–738. doi:10.1261/rna.073114.119
- Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* 361, 1475–1485. doi:10.1056/NEJMr0804615
- Kapranov, P., St Laurent, G., Raz, T., Oszolak, F., Reynolds, C. P., Sorensen, P. H., et al. (2010). The Majority of Total Nuclear-Encoded Non-Ribosomal RNA in a Human Cell is ‘Dark Matter’ Un-Annotated RNA. *BMC Biol.* 8, 149. doi:10.1186/1741-7007-8-149
- Langfelder, P., and Horvath, S. (2008). WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinform.* 9, 559. doi:10.1186/1471-2105-9-559
- Lazorthes, S., Vallot, C., Brioso, S., Aguirrebengoa, M., Thuret, J.-Y., St Laurent, G., et al. (2015). A vlincRNA Participates in Senescence Maintenance by Relieving H2AZ-Mediated Repression at the INK4 Locus. *Nat. Commun.* 6, 5971. doi:10.1038/ncomms6971
- Lennon, A. M., Buchanan, A. H., Kinde, I., Warren, A., Honushesky, A., Cohain, A. T., et al. (2020). Feasibility of Blood Testing Combined With PET-CT to Screen for Cancer and Guide Intervention. *Science* 369, eabb9601. doi:10.1126/science.abb9601
- Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., and Seiden, M. V.CCGA Consortium (2020). Sensitive and Specific Multi-Cancer Detection and Localization using Methylation Signatures in Cell-Free DNA. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* 31, 745–759. doi:10.1016/j.annonc.2020.02.01110.1016/j.annonc.2020.06.008
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of fold Change and Dispersion for RNA-seq Data With DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8
- Maynard, S., Fang, E. F., Scheibye-Knudsen, M., Croteau, D. L., and Bohr, V. A. (2015). DNA Damage, DNA Repair, Aging, and Neurodegeneration. *Cold Spring Harb. Perspect. Med.* 5, a025130. doi:10.1101/cshperspect.a025130
- Ou, H.-L., and Schumacher, B. (2018). DNA Damage Responses and p53 in the Aging Process. *Blood* 131, 488–495. doi:10.1182/blood-2017-07-746396
- Parikh, N., Shuck, R. L., Gagea, M., Shen, L., and Donehower, L. A. (2018). Enhanced Inflammation and Attenuated Tumor Suppressor Pathways are Associated With Oncogene-Induced Lung Tumors in Aged Mice. *Aging Cell* 17, e12691. doi:10.1111/ace1.12691
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* 14, 417–419. doi:10.1038/nmeth.4197
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., et al. (2015). The Transcriptional Landscape of Age in Human Peripheral Blood. *Nat. Commun.* 6, 8570. doi:10.1038/ncomms9570
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org>.
- Rozhok, A., and DeGregori, J. (2019). A Generalized Theory of Age-Dependent Carcinogenesis. *eLife* 8, e39950. doi:10.7554/eLife.39950
- Shen, L. (2020). GeneOverlap: Test and Visualize Gene Overlaps. Available at: <http://shenlab-sinai.github.io/shenlab-sinai/>.
- Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer Statistics, 2020. *CA A. Cancer J. Clin.* 70, 7–30. doi:10.3322/caac.21590
- Srivastava, S., and Hanash, S. (2020). Pan-Cancer Early Detection: Hype or Hope. *Cancer Cell* 38, 23–24. doi:10.1016/j.ccell.2020.05.021
- St Laurent, G., Shtokalo, D., Dong, B., Tackett, M. R., Fan, X., Lazorthes, S., et al. (2013). VlincRNAs Controlled by Retroviral Elements are a Hallmark of Pluripotency and Cancer. *Genome Biol.* 14, R73. doi:10.1186/gb-2013-14-7-r73
- St Laurent, G., Vyatkin, Y., Antonets, D., Ri, M., Qi, Y., Saik, O., et al. (2016). Functional Annotation of the Vlinc Class of Non-Coding RNAs Using Systems Biology Approach. *Nucleic Acids Res.* 44, 3233–3252. doi:10.1093/nar/gkw162
- Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (20182018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi:10.1093/nar/gkx1098

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Qi, Gao, Cai, Han, Qi, Ni, Sun, Huang, Chen, Wu and Kapranov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.