Cancer Genetics 239 (2019) 36–45

Contents lists available at ScienceDirect

Cancer Genetics

journal homepage: www.elsevier.com/locate/cancergen

Original Article

Dysregulated expression of repetitive DNA in ER+/HER2- breast cancer

Cihangir Yandım[a,b], Gökhan Karakülah[b,c,*]

[a] İzmir University of Economics, Faculty of Engineering, Department of Genetics and Bioengineering, 35330, Balçova, İzmir, Turkey
[b] İzmir Biomedicine and Genome Center (IBG), Dokuz Eylül University Health Campus, 35340, İnciraltı, İzmir, Turkey
[c] İzmir International Biomedicine and Genome Institute (iBG-İzmir), Dokuz Eylül University, 35340, İnciraltı, İzmir, Turkey

ARTICLE INFO

Article history:
Received 12 April 2019
Revised 3 July 2019
Accepted 5 September 2019

Keywords:
Breast cancer
Repetitive DNA
Survival
Transcriptome analysis
Bioinformatics

ABSTRACT

Limited studies on breast cancer indicated pathogenic changes in the expressions of some repeat elements. A global analysis was much needed within this context to distinguish the most significant repeats from more than a thousand repeat motifs. Utilising a previously presented RNA-seq dataset, we studied expression changes of all repeats in ER+/HER2- human breast tumour samples obtained from 22 patients in comparison to matched normal tissues. Fifty six (56) repeat subtypes including satellites and transposons were found to be differentially expressed and most of them were novel for breast cancer. HERVKC4-int and HERV1_LTRc, whose expressions correlated well with that of the estrogen receptor gene ESR1, were upregulated at the highest level. REP522 and D20S16 satellites were also significantly upregulated along with insignificant increases in the expressions of other satellites including HSATI and BSR/beta. Interestingly, expressions of REP522 and D20S16 correlated with many key breast cancer pathway (e.g. BRCA1, BRCA2, AKT1, MTOR, KRAS) and survival genes; possibly highlighting their importance in the carcinogenesis of breast. Additional differentially expressed elements such as L1P and various MER transposons also exhibited a similar pattern. Finally, our repeat enrichment analysis on the promoters of differentially expressed genes revealed further links between additional repeats and nearby genes.

Introduction

Repetitive DNA elements, which are traditionally thought to be 'evolutionary junk', make up more than half of the human genome [1,2]. In contrast to this underestimating saying, current evidence points out the importance of repeats in human development, physiology and disease [3–5]. Though the functional contribution of repeats to genome is still yet to be fully characterised, we already know that their dysregulation is associated with neurodegenerative and autoimmune disorders, as well as cancer; and that the types of repeats involved in certain diseases often vary from one type of pathology to the other [6–8].

There are numerous types of repeat elements in the human genome. Tandem repeats, such as the pericentromeric/centromeric satellites, are particularly pivotal in forming constitutive heterochromatin and maintaining it, and thereby providing the solid platform for healthy kinetochores and successful cell divisions [9–11]. On the other hand, their subtelomeric or telomeric counterparts play a major role in establishing a resilient structure at the end of chromosomes; protecting them from the erosive forces during cell division [12]. Moreover, micro- and minisatellites are thought to have gene regulatory functions that have been evolved over long periods of time [13]. Importantly, abnormal expression levels of otherwise-heterochromatic satellites were linked to cancer. Perhaps, the most well-known example is the overexpression of HSAT-II in pancreatic adenocarcinoma [14,15]. In this case, sequestering of DNA damage proteins [16] and the formation of RNA:DNA hybrids along with disruptions in the heterochromatic architecture are thought to be the key players in the pathogenesis [17,18]. In addition to satellites, interspersed elements such as Long Interspersed (LINE), Short Interspersed (SINE) and Long Terminal Repeat (LTR) containing transposable elements are known to cause insertional mutagenesis on genes, which also comes across as a pathogenic force associated with various types of cancers [19–21].

The progressing events in the axis of genomic distortion caused by the expression of repetitive DNA was often linked to genomic instability [20,22], which also plays a key role in the development of breast cancer [23,24]. It was confirmed that HSATI and HSATII were overexpressed in a percentage of clinical specimens and cell lines of breast cancer [25,26]. Furthermore, abnormal expressions of LINE-1 and Alu elements [27,28], and LTR retrotransposons (mainly the HERVK family) were also shown to be dysregulated in breast cancer [29-32]. HERVK expression was linked to breast cancer's metastatic progression [33] and it may serve as a

* Corresponding author.
E-mail address: gokhan.karakulah@ibg.edu.tr (G. Karakülah).

https://doi.org/10.1016/j.cancergen.2019.09.002
2210-7762/

biomarker for patients [34,35]. All of these reports on repetitive DNA expression in breast cancer so far focused on an isolated repeat type and a high-resolution analysis of the human repeatome was much needed; particularly knowing the fact that there are more than a thousand subtypes of repeats in the human genome [36]. A global analysis would indeed help for determining the subtypes of repeats, their strand specific expression as well as their correlations with cancer related genes.

To address these issues and perform a detailed analysis on the RNA expressions of repetitive elements, we searched for a publicly available human breast cancer dataset with matched normal tissues. To our surprise, we could not find a suitable dataset in TCGA (The Cancer Genome Atlas) [37–39] and ICGC (International Cancer Genome Consortium) [40] databases for this type of analysis as all of these datasets in these consortiums were subject to library preparation with poly(A) pre-selection, which reportedly interferes with the detection of many types of repeats and thereby does not allow a uniform analysis [41]. We could only employ a GEO (Gene Expression Omnibus, #GSE103001) RNA-Seq dataset, which was prepared without the poly(A) bias with the purpose of detecting non-coding sense and anti-sense transcripts in ER+/HER2-tumour and adjacent healthy tissue specimens obtained from 22 breast cancer patients [42]. Our analysis on the expressions of all repeat elements confirmed previously published studies and revealed many new elements that were not mentioned in the breast cancer literature before. Indeed, some of these novel elements were not previously reported within the context of any cancer.

Results

A global expression analysis of the repeatome reveals key repetitive elements in breast cancer

Given the reported changes in the chromatin environment in breast cancer and the limited number of repeat elements reported within this context, we wondered if there is a global change in the expression of the whole repeatome in breast cancer. We used HOMER NGS analysis suite to analyse the expression levels of 1395 RepeatMasker repeats (http://repeatmasker.org) and compared the expression levels of these repeats with the expression levels of UCSC annotated genes. HOMER is one of the commonly used tools for estimating the transcription abundances of both coding and non-coding regions including the repeats in a strand specific manner [43-46]. The relative percentage of the number of reads obtained from all repeats with respect to the total number of any reads (repeat + non-repeat) did not exhibit a statistically significant change in tumour tissues when compared to normal tissues (Fig. 1A, upper panel). However, the distribution of these percentages was slightly changed (Fig. 1A upper panel) and some patients exhibited a global change in their repeat expression profiles (Supp. Fig. S1**).** A similar trend was observed for global gene expression levels (Fig. 1A, lower panel). Moreover, we realised that transcripts arisen from satellite repeats only constituted a small fraction of total repeat transcripts but this was clearly increased in cancer tissues (0.5% vs 0.8%) (Fig. 1B). Small increases were seen for the percentage reads of LTRs, LINE and SINE elements. When a principal component analysis (PCA) was performed, a clear distinction was observed in gene expression profiles of normal versus tumour tissues (Supp. Fig. S2A). This was in agreement with the previously published study, which produced and analysed the same dataset [42]. However, there was a varying degree of difference in repeat expression levels for each individual patient; with some patients exhibiting an obvious dysregulated repeat expression profile whereas others only displayed slightly different changes (Supp. Figs. S1 and S2B).

We next investigated the expression levels of all subtypes of main repeat classes individually by calculating the fold change by dividing the read number obtained for the specific repeat (without the strand bias) in the tumour sample with that of the matched normal tissue for each patient. We came up with 56 repeat elements that were significantly and differentially expressed in tumours ($\log_2$(Fold change) > 0.5, FDR < 0.01) (Fig. 1C and D, Supp. Table S1). Among these, HERVKC4-int (int: internal region); a member of the previously breast cancer-linked HERVK family [29–35,47–50] was the most significantly upregulated repeat. LTR14, which is the LTR element associated with HERVKC4 [51], was also significantly upregulated along with HERV1_LTRc, HERV9-int, LTR14C and HERVK14C-int. Moreover, some members of the simple repeat family, DNA transposons including MERs and UCONs as well as previously reported LINE1 (L1P) elements [27,28] and two classical satellites; predominantly subtelomeric REP522 [52] and predominantly pericentromeric D20S16 [53] were upregulated. When we checked the distribution of fold changes from all patients, we could see that some repeats were differentially expressed in almost all tumour samples (e.g. HERVKC4-int, HERV1_LTRc, D20S16) whereas others did not strictly follow this trend (Fig. 1D). In order to see if there is a strand bias in the expressions of these elements, we performed a similar analysis one more time by using the aligned reads from positive and negative strands separately (Supp. Fig. S3). Those which showed the highest upregulation in Fig. 1C still remained in the list however others such as HERV1_LTRc, D20S16 and MER72B only appeared in the differentially expressed repeats list when only the reads from positive strand were considered in the analysis.

A comprehensive analysis of the expression of classical satellites

It is well known that classical satellites, whose tandemly repeating motif could range from 80 bp to 1500 bp, are predominantly found in centromeric/pericentromeric or telomeric regions [54]. Expression of pericentromeric satellites were tightly linked to *de novo* heterochromatin formation in mammalian development [10,55–58]. There are 35 satellites defined for human in the Repbase [36]. HSATI was the primary classical satellite reported to be abnormally expressed *per se* in breast cancer so far and this is known to contribute to the carcinogenesis of the breast [25,26,59,60]. Given the importance of heterochromatic classical satellites in maintaining genomic stability [16,60], we checked the expression levels of all satellites and calculated an FDR value for each of them after filtering out satellites that have less than 0.5 TPM reads. We ended up with 24 satellites, three of which had significant FDR values (*<0.01) in terms of fold change (Fig. 2A). When FDR values were not considered, REP522 was again at the top of the list with the highest upregulation in tumour samples. This was followed by the previously reported HSATI. Even though the FDR value for HSATI (FDR = 0.157) was more than the 0.01 threshold**,** this result could still be an indication of its importance in breast cancer. D20S16 (FDR = 0.0004) and LSAU (beta satellite, FDR = 0.1755) elements were upregulated more than the $\log_2$(fold change) > 0.5 threshold whereas the pericentromeric GSAT (gamma satellite, FDR=0.3804) repeat was downregulated with the caveat that the latter two were not statistically significant. Interestingly, pancreatic cancer-associated HSATII [14,15] did not show an upregulation in ER+/HER2- breast cancer samples.

Strand specific expression of pericentromeric satellites were shown to be influential during development [10], hence we checked if the satellite dysregulation was due to strand specific expression by calculating strand specific fold change ratios (tumour/normal tissue) in satellite RNA expression for each patient (Fig. 2B). For REP522, the expression of the positive strand was slightly more elevated whereas for D20S16, expression of the

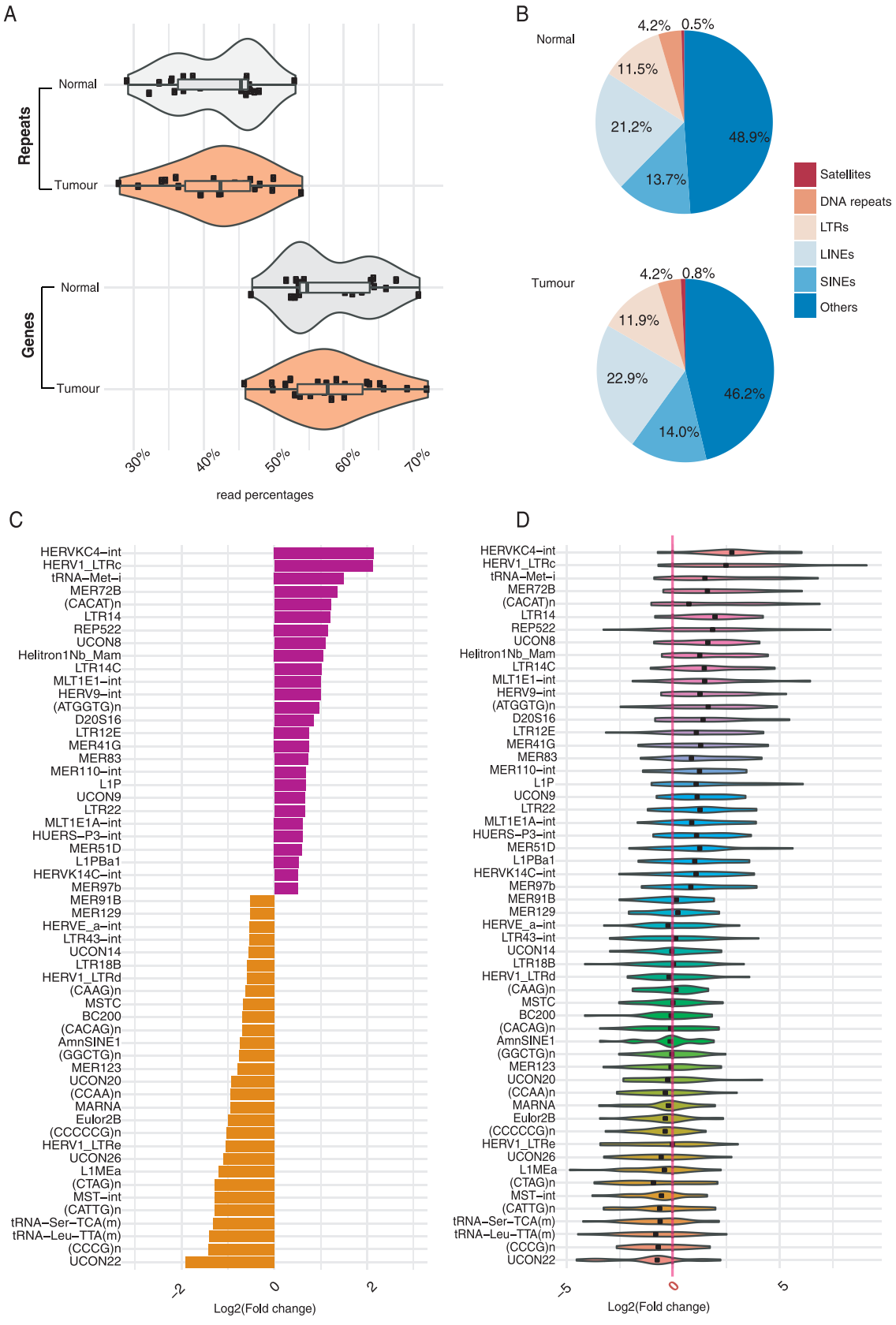

Fig. 1. Analysis of global changes in the expression of repeatome and differentially expressed repeat elements in breast tumours. A Read percentages of repeats and genes in tumour versus matched normal tissues. Read percentage was calculated by dividing the number of reads obtained from all repeats by the number of reads obtained from UCSC genes. B Read percentages of individual repeat classes. Read percentages were calculated by the number of reads obtained from all members of the given repeat class by the read number obtained from the whole repeatome. C Fold changes in differentially expressed repeats in breast tumours. D Violin plots representing the distributions of fold changes in differentially expressed repeats in tumours. For C and D, fold changes of repeats were calculated by dividing the read number obtained from the given repeat in tumour versus matched normal tissue for each patient. A filter of $\log_2$(Fold change) > 0.5 and FDR < 0.01 was applied.

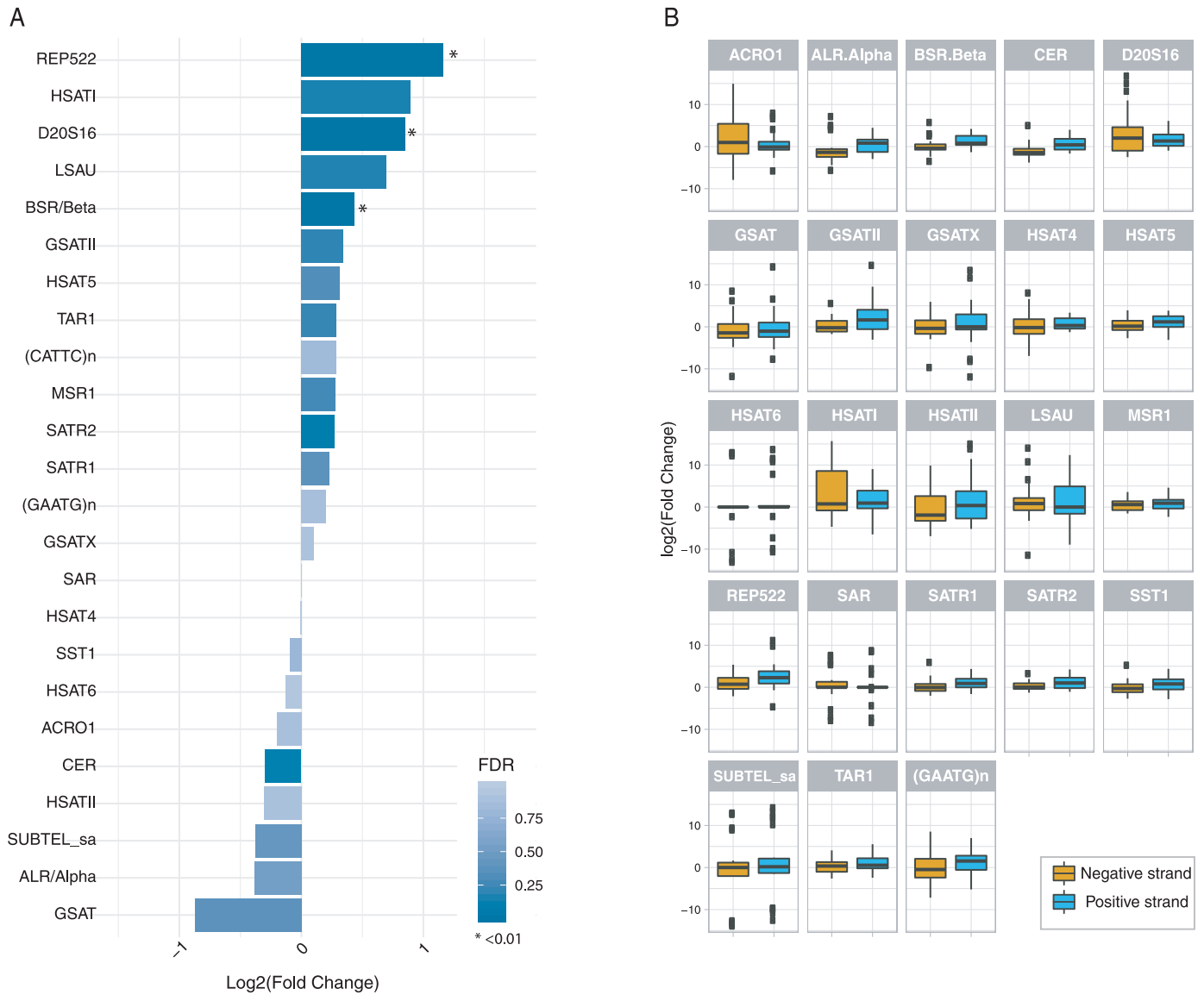

Fig. 2. Expression fold changes of satellite repeats in breast tumours. A Fold changes in the expressions of all satellites. All satellites (TPM > 0.5) were considered for the fold change analysis (tumour vs normal). (*FDR < 0.01) B Strand specific analysis of fold changes in satellite expression (tumour vs normal).

negative strand was emphasised for at least in a proportion of patients. Overall, none of the satellites had a dramatic difference in their expression levels when positive and negative strands were compared side by side.

KEGG breast cancer genes and their expression correlations with repeats

To see if there is a correlation with the expressions of repeats and genes which are influential in breast cancer pathogenesis, we extracted the list of genes from the KEGG pathway [61,62] of breast cancer (pathway #hsa05224) and calculated the Pearson's correlation (r) coefficients of these genes in terms of expression with that of the 56 dysregulated repeats (as given in Fig. 1C) and all 24 satellite repeats (as given in Fig. 2A). Supp. Fig. S4 summarises these correlations in two separate heatmaps.

The highly upregulated HERVKC4-int, which is a member of the previously reported HERVK family in the context of breast cancer [29–35,47–50], and its flanking long terminal repeat LTR14, as well as HERV1_LTRc did not show a significant correlation with the vast majority of KEGG breast cancer genes (Supp. Fig. S4A). Still, HERVKC4-int and HERV1_LTRc interestingly correlated well with the expression of the estrogen receptor gene ESR1 (Fig. 3). Moreover, expressions of various DNA transposons (MERs) and L1 element L1P correlated with the expressions of many KEGG breast cancer genes. More intriguingly, the expressions of two phenomenal breast cancer genes BRCA1 and BRCA2 [63] were in good correlation with the expressions of REP522 and D20S16 (Fig. 3), and many other types of satellites (Supp. Fig. S4B). Expressions of these particular satellites, which were significantly upregulated in tumours, also correlated well with the expressions of well-known oncogenes such as AKT1, KRAS and MTOR (Fig. 3); which were crucially involved in breast cancer's pathogenesis [64–67].

Breast cancer survival genes and their expression correlations with repeats

We performed a similar correlation analysis on the expressions of dysregulated repeats and top survival genes in breast cancer patients. The list of top 100 breast cancer survival linked genes were obtained from the GEPIA webtool [68], which performs an overall survival analysis based on gene expression using a Log-rank (a.k.a.

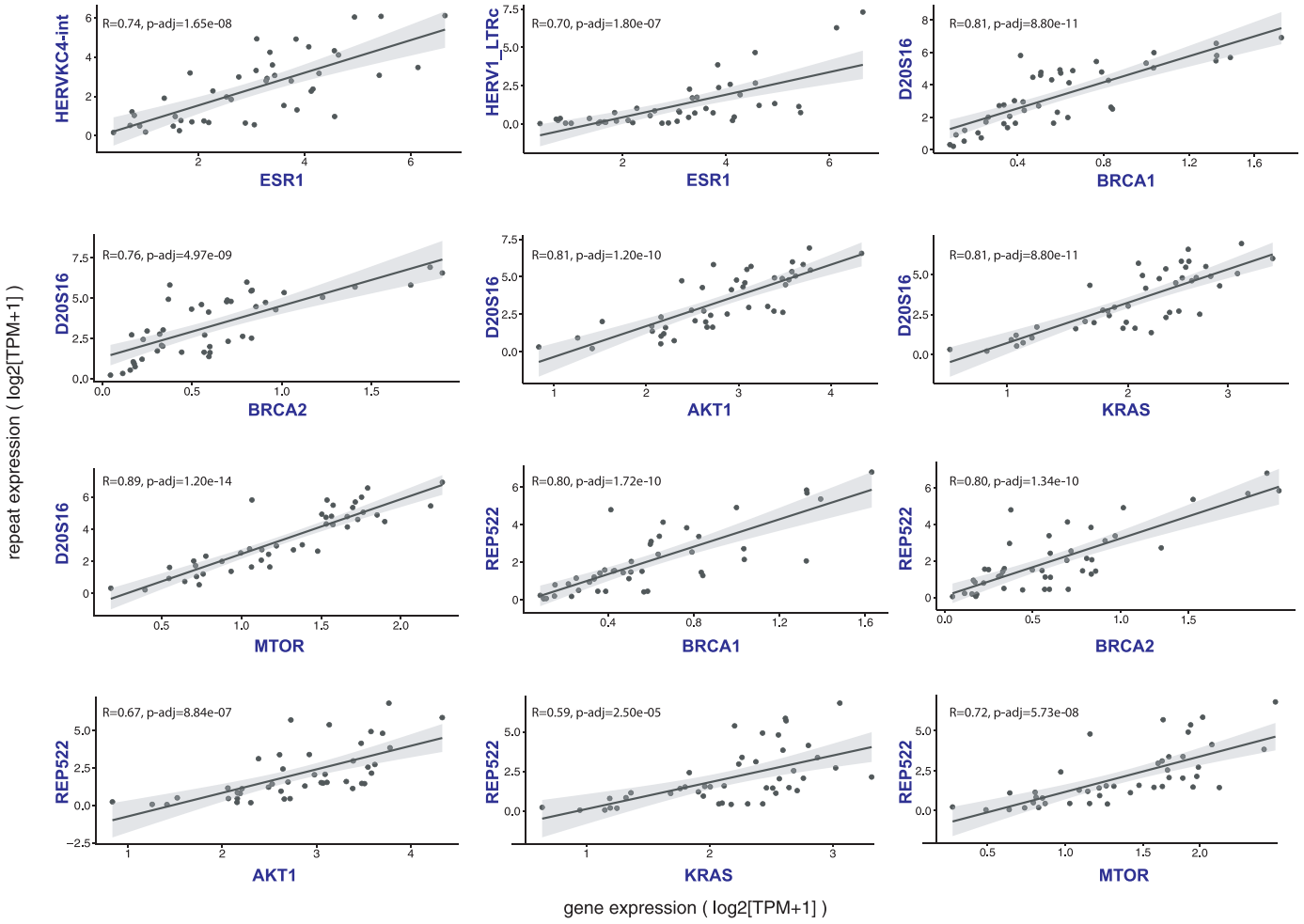

Fig. 3. Expression correlations of HERVKC4-int, HERV1_LTRc, D20S16 and REP522 with vital genes involved in breast cancer pathogenesis. R correlation coefficients and the corresponding Benjamini–Hochberg adjusted *p*-values were indicated on the scatter plots.

Mantel-Cox) as an hypothesis test using 1085 primary breast tumour data extracted from the TCGA database [38,39]. The list of top 100 genes, whose differential expression is related to survival in breast cancer according to this analysis, is given in Supp. Table S2. Representative Kaplan–Meier plots on the expressions of these genes and their impact on overall survival were given in Supp. Fig. S5. We calculated r correlation coefficients for the expressions of these survival genes with the expressions of repeat elements. After filtering the genes that have a TPM value less than 0.5, we ended up with 50 survival genes. We drew heatmaps to visualise the correlations for 56 differentially expressed repeat elements and all satellite repeats. A group of repeat elements (Fig. 4A, shown in blue) showed medium to high correlation levels with many survival genes. Highly upregulated HERVKC4_int and HERV1_LTRc only correlated with a few survival genes. On the other hand; satellites, REP522 and D20S16 highly correlated ($r > 0.7$) with 10 and 18 of survival genes respectively, and they showed moderate levels ($0.5 < r < 0.7$) of correlations with many others. Some of the survival genes correlated with almost all members of the classical satellite family (Fig. 4B).

Enriched repeat motifs within the promoters of differentially expressed genes in breast cancer

Repeat elements have the potential to interfere with the regulation of nearby genes when located within their promoters [69,70]. To see if the differentially expressed genes in breast tumours are enriched by certain repeat motifs within their promoters, we first identified dysregulated genes ($|\log_2(\text{fold change})| > 1$ and FDR < 0.05) in tumour samples versus matched normal tissues (Supp. Table S3). Then, we ran a gene ontology analysis on these genes using DAVID [71] (Supp. Table S4). For all differentially expressed genes falling into each GO term category, we checked how many times any repeat motif appears in their promoters using the reference genome hg19. We set the limit for the promoter length to 1 kb as the vast majority of human genes were reported to have their promoters within the 1 kb flanking region of their 5′ ends [72]. Next, we calculated an enrichment score for each repeat by dividing its occurrence rate in 1 kb promoters of all differentially expressed genes listed in each GO term with its occurrence rate in the whole reference genome followed by a Fisher's exact test. We summarised the result on a dot-plot for enriched repeats and their relevant GO terms, which contain at least five differentially expressed genes that are associated with at least one type of repeat motif (Fig. 5A).

Two of the significantly and differentially expressed repeats, AmnSINE1 and MER91B appeared in the enriched repeats list. Even though not listed in the statistically empowered differentially expressed repeats given in Fig. 1C; L3b, LTR16C, MER135 and Plat_L3 elements were associated with the most number of GO terms. The violin plots representing the expression fold changes of these previously unlisted repeats show that there is always a considerable portion of patients, which differentially express these enriched repeats in their tumour samples versus their matched normal tissues

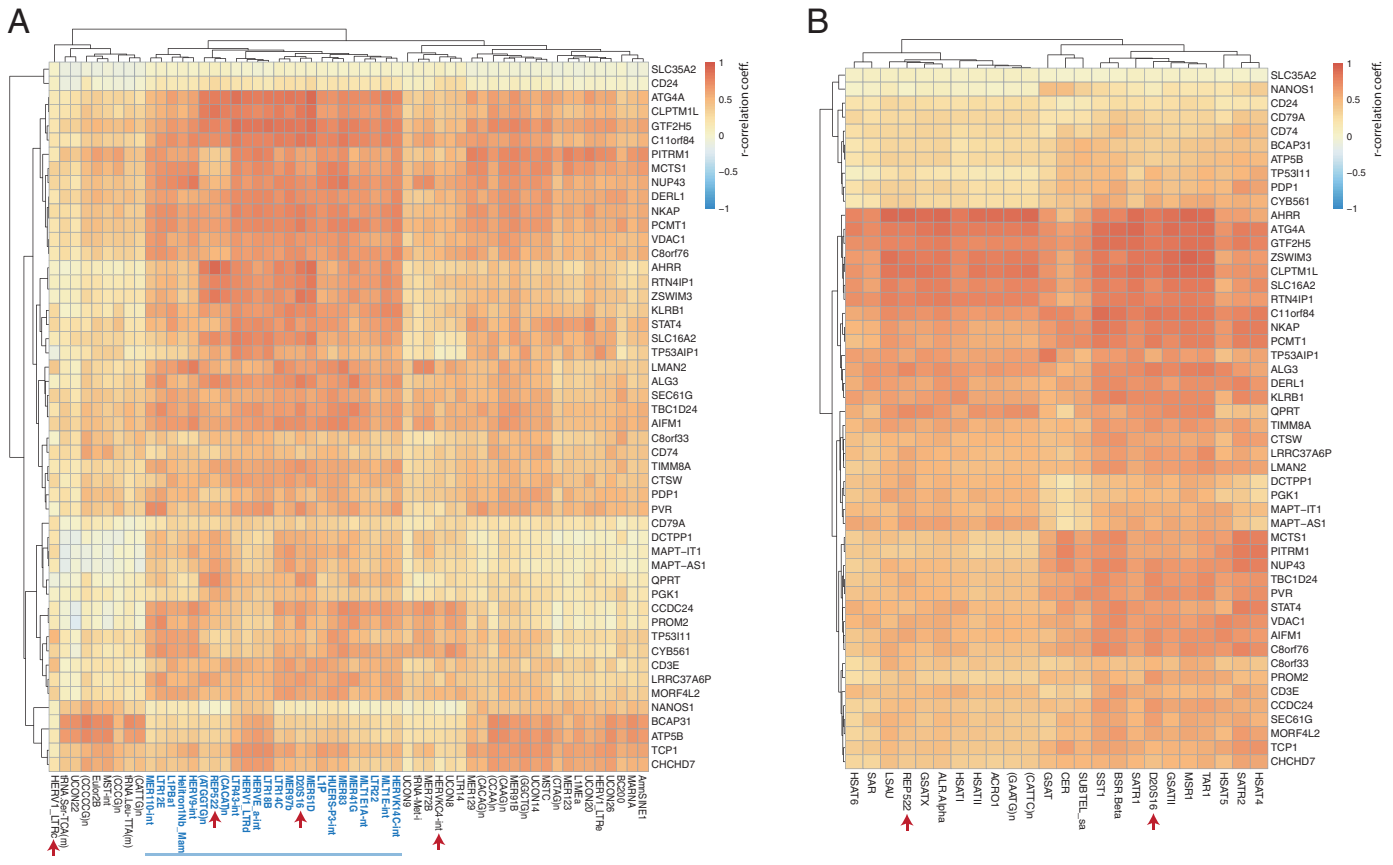

Fig. 4. Expression correlations of differentially expressed repeats and satellites with breast cancer survival genes. A Heatmap representing the expression correlations of differentially expressed repeats (as shown in Fig. 1C) with GEPIA survival genes. A cluster of repeat elements that exhibit higher correlation levels were indicated in blue. B Heatmap representing the expression correlations of all satellites (as shown in Fig. 2A) with GEPIA survival genes [68]. A,B Survival genes were obtained by using the "most significant survival genes" tool of the GEPIA webtool [68], which lists survival genes based on TCGA breast cancer datasets. Genes that have TPM < 0.5 were not included in the analysis. Annotation arrows indicate repeat elements, which showed significant correlations with breast cancer genes shown in Fig. 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Fig. 5B). Expressions of repeats also showed varying levels of correlation with the expression of genes in the relevant GO term list (Supp. Fig. S6). Particularly, LTR16C correlated at moderate to high levels ($r > 0.5$) with the genes in most of the GO term modules.

Discussion

Understanding genomic instability and the pathogenic genomic rearrangements taking place in breast cancer will help in evaluating patient prognosis and designing efficient therapies. Even though repetitive DNA makes up at least half of the human genome and its expression is linked to genomic instability, genome-wide studies often disregard the information obtained from repeats or they use sample preparation methods that does not allow one to quantify repetitive DNA transcripts. Our analysis on a published dataset with 22 patients points out abnormalities in the expression statuses of many repeat elements in breast cancer. Here, we analysed ER+/HER2- breast cancer in terms of molecular pathology. ER+/HER2- breast cancer makes up at least half of the breast cancer cases [73–75]. Still, even further molecular heterogenetiy exists among this specific subtype of breast cancer patients [38,73]. To further stratify the samples in terms of molecular heterogeneity and analyse changes in expression accordingly, one could make use of personalised perturbation profiling. A diligent pipeline was developed by Menche et al. [76] with the caveat that it was not tested for RNA sequencing data. However, such advanced perturbation methods could also be applied to RNA sequencing in future.

Our results suggest that there is not a dramatic global change in the expression of the whole repeatome in breast tumours that are ER+/HER2-. The PCA analysis showed that expression patterns of genes behave differently than that of repeats. Knowing the fact that repeats have characteristic and distinct chromatin modifications when compared to genes [75,77,78], this result suggests that repeat regulation in ER+/HER2- breast cancer is likely to be biologically different in comparison to genes. Our study highlights robust changes in the expressions of particular elements from different repeat families and the effects of their expression could be explored further with experimental studies. In addition to the previous literature suggesting the dysregulation of HERVK [29–35,47–50] and HSATI elements [59,60] in breast cancer, our global approach helped to characterise the behaviours of all repeat classes in this context. We realised that dysregulation of repeats were not limited to those stated before. The satellite class of repeats showed the highest increase in expression in the global scale when all satellite transcripts were compared in tumour versus normal samples. Though satellite transcripts accounted only for 0.5% of normal tissues, this number was increased to 0.8% (approximately 60% increase) in tumour tissues, making the satellites the most influenced repeat class in ER+/HER2-breast cancer at a global scale (Fig. 1B). Given the contribution of satellite transcripts in modulating the genome architecture [10,16,18,57,79], their dysregulation may potentially be a driving force for genomic instability in breast cancer. The knockout of the breast cancer suppressor gene BRCA1 caused the derepression of the pericentromeric HSATI satellites, which in turn triggered genomic instability and mitotic

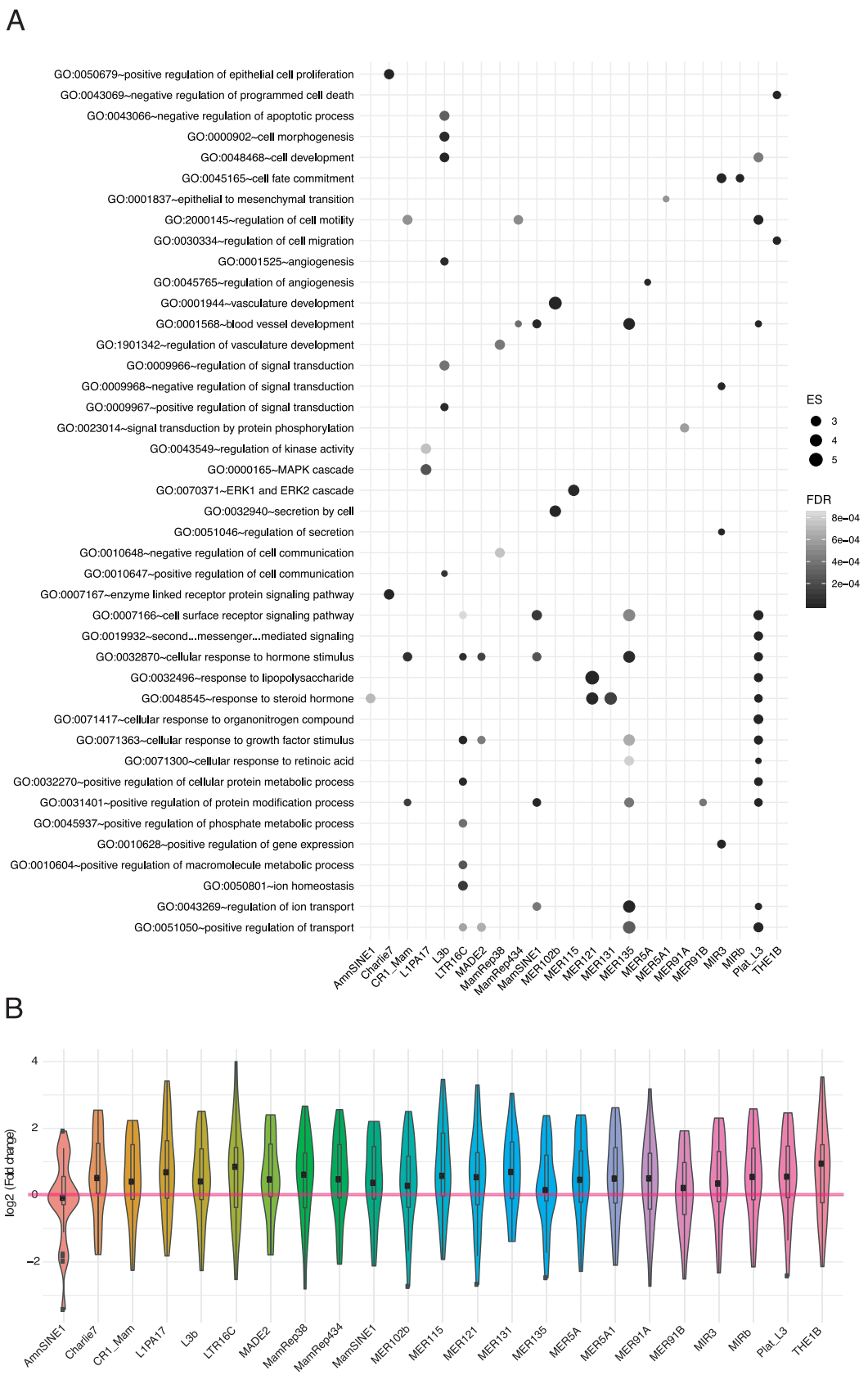

Fig. 5. Enrichment of repeat elements in the promoters of differentially expressed genes and the expression fold changes of these repeats in tumours. A Enrichment levels of repeats in the promoters of genes falling into different GO term categories. Differentially expressed genes were listed using the edgeR tool ($|\log_2(\text{fold change})| > 1$ and FDR < 0.05). GO terms for these genes were obtained using the DAVID webtool. An occurrence enrichment score was calculated for repeats that occur at the promoters of at least five genes present in the corresponding GO term. Calculated enrichment score and the FDR for this analysis were indicated on the dot-plot. Only repeats with an enrichment score of > 2 and FDR < 0.001 were given. Supplementary Fig. 6 points out that the expressions of repeats that are enriched in gene promoters also correlate well with the expressions of those particular genes. B Violin plots representing the distributions of expression fold changes in enriched repeats in differentially expressed gene promoters.

catastrophe in breast cancer cells due to the disruption of heterochromatic mechanisms [59,60]. Here, we are reporting that additional elements such as REP522 and D20S16 are also dysregulated in ER+/HER2- breast cancer (Fig. 1C) and whether they have similar effects to HSATI, should be explored with further studies. A study reported that promoters of genes and lncRNAs associated in various cancer types, including the breast cancer, are overlapping with REP522 elements [80], which are mainly subtelomeric interspersed repeats of 1.8 kb in length [36]. Our study directly demonstrates the dysregulated overexpression of REP522 satellites *per se* in ER+/HER2- breast cancer. In addition, D20S16 seems as another crucial satellite that is novel for breast cancer. D20S16 is an interspersed satellite with a repeating unit of 98 bp whose contribution to genome architecture is unknown [36]. Furthermore, even though not statistically significant, it is worth mentioning that a number of patients upregulated beta satellites (BSR/Beta) and its associated repeat LSAU. The functions of these repeats, if any, are also generally uncharacterised. Among the 10 members of the HERVK family, HERVKC4-int (a.k.a. HML10) seemed to be the most upregulated one (Fig. 1C). A previous study also reported this upregulation in primary patient tumours with a hybridisation array [48]. In that previous report, however, HERVKC4 was not the most outstanding member of the HERVK family in terms of expression upregulation, having in mind that the molecular pathologies of the patients (ER and HER2 statuses) could be different. The robust upregulation of LTR14, which is the flanking LTR to HERVKC4-int [81], also supports our finding within the patient group covered in this study. In addition to those mentioned here, LINE element L1P and various other elements including members of MER and UCON also warrant further experimental research according to our results.

Our correlation analysis with breast cancer KEGG pathway genes and GEPIA survival genes may help scientists to dissect the most important repeats within the pathogenesis of breast cancer (Figs. 3, 4 and Supp. Fig. S5). We realised that the most upregulated elements, HERVKC4-int and HERV1_LTRc, were not among the repeats that showed the highest correlation with breast cancer pathway or survival genes in ER+/HER2- breast cancer. Still, their expression interestingly correlated well with the estrogen receptor gene ESR1; perhaps implying the estrogen responsiveness in the mechanistic insight of this upregulation, which is a potential possibility to be addressed with experimental studies in future. On the other hand, satellite elements REP522 and D20S16 correlated with more breast cancer KEGG pathway genes compared to the upregulated HERVs. Finally, the correlations with crucial cancer related genes such as BRCA1, BRCA2, AKT1, KRAS and MTOR were also noteworthy, as were the correlations with GEPIA survival genes. Together, these results could indeed indicate a functional importance of repeat elements in carcinogenesis and should be explored with future experimental studies. The relatively high number of correlating survival and breast cancer genes with satellites specifically brings these ancient members of our genome under spotlight.

Our distribution analysis of repeats on differentially expressed gene promoters (Fig. 5) demonstrated that certain repeats could play a role in the regulation of nearby genes, at least in a cohort of patients. However, this latter analysis did not outstandingly reveal those repeats that were significantly upregulated. This analysis was done with a reference genome and combined analyses of RNA and DNA from the same patient in future could put this dimension into a further clarity. Moreover, preparing RNA samples for sequencing without poly(A) pre-selection would help the community to analyse the repeatome's contribution to cancer more in depth, specifically in large cohort studies. According to the results presented in this study, there are many mechanistic and functional aspects worth studying on repeat dysregulation in breast cancer. Developing therapies that target repeat expression may reduce the genomic instability and hence interfere with the speed of tumour evolution in breast cancer, which could have implications on metastasis and resistance.

Methods

RNA-seq data collection and processing

The whole transcriptome sequencing data of 22 matched normal-tumour pairs (44 RNA-seq libraries in total) introduced in a recent study [42] were extracted from Sequence Read Archive database [82] (SRA Accession: SRP116023) using SRA Tool Kit v.2.9.0 with "*fastq-dump –gzip –skip-technical –readids –dumpbase –clip –split-3*" command. Sequencing reads of each sample were then aligned to the human reference genome hg19 with HISAT2, a sensitive splice-aware aligner, using "*hisat2 -x {ht2-idx} –1 {infile} –2 {infile} -S {outfile}*" parameters [83]. For the measurement of RNA expression in gene and repeat regions across 44 samples, we utilised HOMER NGS analysis suite v4.10 (http://homer.ucsd.edu/homer/ngs/) [45]. We particularly included the HOMER tool in our analysis pipeline as it allows one to quantify strand specific repeat expression. The following commands were used for gene and repeat expression analysis, respectively: "*analyzeRepeats.pl rna hg19 -raw / -tpm -condenseL1 -d {infile} > {outfile}*" and "*analyzeRepeats.pl repeats hg19 -raw / -tpm -condenseL1 -d {infile} > {outfile}*".

To detect both genes and repeats differentially expressed between normal and tumour tissues, we utilized edgeR package v3.24.3 [84] of R statistical computation environment v3.4.4 (http://www.R-project.org). As recommended by the edgeR's reference guideline, we firstly applied Trimmed Mean of M-values (TMM) normalisation to the count values from HOMER and we employed a generalised linear model with tissue type and patient as factors. Then, dispersion estimation was done with estimateDisp function, and paired differential expression analysis between tumour and normal tissues within patients were performed using glmFit and glmLRT functions of edgeR.

Analysis of the expression correlations

KEGG breast cancer pathway (#hsa05224) genes were extracted from https://www.genome.jp/kegg/ and GEPIA top 100 overall survival genes were obtained from http://gepia.cancer-pku.cn. For the latter, median was used as the cut-off point for the classification of low and high expression groups. Expression correlations of these KEGG or GEPIA genes were calculated using the cor.test function with method="pearson" in the R environment.

Enrichment analysis of repeats

Using the RepeatMasker annotation, which was downloaded from the UCSC database, we determined overlapping genomic regions of 1 kb upstream regions of differentially expressed genes which are significantly enriched in GO terms. To achieve this, we utilised the intersectBed command of bedtools [85]. To figure out if certain repeat elements were enriched in these regions, we calculated an enrichment score (ES) for repeat elements as the following:

$ES_X = (r/R)/(g/G)$

- *where R is the total number of all repeats located within 1 kb upstream regions of the genes associated with the GO term,*
- *G is the total number of all repeats located within 1 kb upstream regions of all genes in the genome,*
- *r is the total number of repeat of interest located within 1 kb upstream regions of the genes associated with the GO term,*
- *g is the total number of repeat of interest located within* 1 kb *upstream regions of all genes in the genome.*

Statistical significances of ESs for repeats were calculated with the Fisher's Exact Test as reported previously [86]. Repeats, which were linked to at least five genes and exhibited an ES of ≥ 2 with a corrected p-value < 0.01, were considered for downstream analysis.

Statistical analysis and graphical representation

R computation environment was used for all statistical calculations. We employed fisher.test for ES calculations and pheatmap package of R (https://CRAN.R-project.org/package=pheatmap) was utilised to draw all the heatmaps, on which expression values were represented in the rows. Clustering was performed with the Euclidian method and prcomp function was used for the PCA analysis. Other graphics were obtained using the ggplot2 package [87].

Ethic approvals and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

Not applicable.

Disclosure

CY was a full-time faculty member at Izmir University of Economics and also an honorary research associate at Izmir Biomedicine and Genome Center while this study was being performed.

Data for reference

The datasets analysed during the current study are available in the GEO repository, with the accession number GSE103001.

Declaration of Competing Interest

Authors declare no competing interests.

CRediT authorship contribution statement

Cihangir Yandım: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Gökhan Karakülah:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Acknowledgements

We thank the scientists who submitted the publicly available GEO dataset (GSE103001).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cancergen.2019.09.002.

References

[1] de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011;7:e1002384.
[2] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[3] Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M. Transcription of tandemly repetitive DNA: functional roles. Chromosome Res 2015;23:463–477.
[4] Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 2017;18:71–86.
[5] Yandim C, Karakulah G. Expression dynamics of repetitive DNA in early human embryonic development. BMC Genomics 2019;20:439.
[6] Chenais B. Transposable elements in cancer and other human diseases. Curr Cancer Drug Targets 2015;15:227–42.
[7] Hancks DC, Kazazian HH Jr. Roles for retrotransposon insertions in human disease. Mob DNA 2016;7:9.
[8] Reilly MT, Faulkner GJ, Dubnau J, Ponomarev I, Gage FH. The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. 2013;33:17577–86.
[9] Maison C, Bailly D, Roche D, Montes de Oca R, Probst AV, Vassias I, Dingli F, Lombard B, Loew D, Quivy JP, et al. Sumoylation promotes de novo targeting of hp1alpha to pericentric heterochromatin. Nat. Genet. 2011;43:220–7.
[10] Probst AV, Okamoto I, Casanova M, El Marjou F, Le Baccon P, Almouzni G. A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev. Cell 2010;19:625–38.
[11] Younger ST, Rinn JL. Silent pericentromeric repeats speak out. In: Proceedings of the national academy of sciences of the United States of America, 112; 2015. p. 15008–9.
[12] Schoeftner S, Blasco MA. A 'higher order' of telomere regulation: telomere heterochromatin and telomeric rnas. EMBO J. 2009;28:2323–36.
[13] Bagshaw ATM. Functional mechanisms of microsatellite DNA in eukaryotic genomes. Genome Biol Evol 2017;9:2428–43.
[14] Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S, et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 2011;331:593–6.
[15] Kishikawa T, Otsuka M, Yoshikawa T, Ohno M, Yamamoto K, Yamamoto N, Kotani A, Koike K. Quantitation of circulating satellite RNAs in pancreatic cancer patients. JCI Insight 2016;1:e86646.
[16] Kishikawa T, Otsuka M, Yoshikawa T, Ohno M, Ijichi H, Koike K. Satellite rnas promote pancreatic oncogenic processes via the dysfunction of ybx1. Nat Commun 2016;7:13006.
[17] Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H, et al. Pericentromeric satellite repeat expansions through rna-derived DNA intermediates in cancer. In: Proceedings of the national academy of sciences of the United States of America, 112; 2015. p. 15148–53.
[18] Johnson WL, Yewdell WT, Bell JC, McNulty SM, Duda Z, O'Neill RJ, Sullivan BA, Straight AF. Rna-dependent stabilization of suv39h1 at constitutive heterochromatin. Elife 2017;6.
[19] Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, Lohr JG, Harris CC, Ding L, Wilson RK, et al. Landscape of somatic retrotransposition in human cancers. Science 2012;337:967–71.
[20] Anwar SL, Wulaningsih W, Lehmann U. Transposable elements in human cancer: causes and consequences of deregulation. Int J Mol Sci 2017;18.
[21] Burns KH. Transposable elements in cancer. Nat. Rev. Cancer 2017;17:415–24.
[22] Zeller P, Gasser SM. The importance of satellite sequence repression for genome stability. Cold Spring Harb. Symp. Quant. Biol. 2017;82:15–24.
[23] Kalimutho M, Nones K, Srihari S, Duijf PHG, Waddell N, Khanna KK. Patterns of genomic instability in breast cancer. Trends Pharmacol. Sci. 2019;40:198–211.
[24] Kwei KA, Kung Y, Salari K, Holcomb IN, Pollack JR. Genomic instability in breast cancer: pathogenesis and clinical implications. Mol Oncol 2010;4:255–66.
[25] Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB. Demethylated hsatii DNA and hsatii rna foci sequester prc1 and mecp2 into cancer-specific nuclear bodies. Cell Rep 2017;18:2943–56.
[26] Ichida K, Suzuki K, Fukui T, Takayama Y, Kakizawa N, Watanabe F, Ishikawa H, Muto Y, Kato T, Saito M, et al. Overexpression of satellite alpha transcripts leads to chromosomal instability via segregation errors at specific chromosomes. Int J Oncol 2018.
[27] Bratthauer GL, Cardiff RD, Fanning TG. Expression of line-1 retrotransposons in human breast cancer. Cancer 1994;73:2333–6.
[28] Park SY, Seo AN, Jung HY, Gwak JM, Jung N, Cho NY, Kang GH. Alu and line-1 hypomethylation is associated with her2 enriched subtype of breast cancer. PLoS One 2014;9:e100429.
[29] Burmeister T, Ebert AD, Pritze W, Loddenkemper C, Schwartz S, Thiel E. Insertional polymorphisms of endogenous herv-k113 and herv-k115 retroviruses in breast cancer patients and age-matched controls. AIDS Res. Hum. Retroviruses 2004;20:1223–9.
[30] Johanning GL, Malouf GG, Zheng X, Esteva FJ, Weinstein JN, Wang-Johanning F, Su X. Expression of human endogenous retrovirus-k is strongly associated with the basal-like breast cancer phenotype. Sci Rep 2017;7:41960.

[31] Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL. Quantitation of herv-k env gene expression and splicing in human breast cancer. Oncogene 2003;22:1528–35.
[32] Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV. Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clin. Cancer Res. 2001;7:1553–60.
[33] Zhou F, Li M, Wei Y, Lin K, Lu Y, Shen J, Johanning GL, Wang-Johanning F. Activation of herv-k env protein is essential for tumorigenesis and metastasis of breast cancer cells. Oncotarget 2016;7:84093–117.
[34] Wang-Johanning F, Li M, Esteva FJ, Hess KR, Yin B, Rycaj K, Plummer JB, Garza JG, Ambs S, Johanning GL. Human endogenous retrovirus type k antibodies and mrna as serum biomarkers of early-stage breast cancer. Int J Cancer 2014;134:587–95.
[35] Zhao J, Rycaj K, Geng S, Li M, Plummer JB, Yin B, Liu H, Xu X, Zhang Y, Yan Y, et al. Expression of human endogenous retrovirus type k envelope protein is a novel candidate prognostic marker for human breast cancer. Genes Cancer 2011;2:914–22.
[36] Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 2015;6:11.
[37] Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell 2018;33:690–705 e699.
[38] Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
[39] Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015;163:506–19.
[40] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016;534:47–54.
[41] Solovyov A, Vabret N, Arora KS, Snyder A, Funt SA, Bajorin DF, Rosenberg JE, Bhardwaj N, Ting DT, Greenbaum BD. Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and t cell suppressive classes. Cell Rep 2018;23:512–21.
[42] Wenric S, ElGuendi S, Caberg JH, Bezzaou W, Fasquelle C, Charloteaux B, Karim L, Hennuy B, Freres P, Collignon J, et al. Transcriptome-wide analysis of natural antisense transcripts shows their potential role in breast cancer. Sci Rep 2017;7:17452.
[43] Bouvy-Liivrand M, Hernandez de Sande A, Polonen P, Mehtonen J, Vuorenmaa T, Niskanen H, Sinkkonen L, Kaikkonen MU, Heinaniemi M. Analysis of primary microrna loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture. Nucleic Acids Res 2017;45:12054.
[44] Gurzeler E, Aavik E, Laine A, Valkama T, Niskanen H, Huusko J, Kaikkonen MU, Yla-Herttuala S. Therapeutic effects of rosuvastatin in hypercholesterolemic prediabetic mice in the absence of low density lipoprotein receptor. Biochimica et biophysica acta. General subjects 2019;1863:481–90.
[45] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Mol. Cell 2010;38:576–89.
[46] Niskanen H, Tuszynska I, Zaborowski R, Heinaniemi M, Yla-Herttuala S, Wilczynski B, Kaikkonen MU. Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions. Nucleic Acids Res 2018;46:1724–40.
[47] Bannert N, Hofmann H, Block A, Hohn O. Hervs new role in cancer: from accused perpetrators to cheerful protectors. Front Microbiol 2018;9:178.
[48] Frank O, Verbeke C, Schwarz N, Mayer J, Fabarius A, Hehlmann R, Leib–Mosch C, Seifarth W. Variable transcriptional activity of endogenous retroviruses in human breast cancer. J Virol 2008;82:1808–18.
[49] Hohn O, Hanke K, Bannert N. Herv-k(hml-2), the best preserved family of hervs: endogenization, expression, and implications in health and disease. Front Oncol 2013;3:246.
[50] Wang-Johanning F, Radvanyi L, Rycaj K, Plummer JB, Yan P, Sastry KJ, Piyathilake CJ, Hunt KK, Johanning GL. Human endogenous retrovirus k triggers an antigen-specific immune response in breast cancer patients. Cancer Res. 2008;68:5869–77.
[51] Dangel AW, Mendoza AR, Baker BJ, Daniel CM, Carroll MC, Wu LC, Yu CY. The dichotomous size variation of human complement c4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among old world primates. Immunogenetics 1994;40:425–36.
[52] Roschenthaler F, Schable KF, Thiebe R, Zachau HG. Of orphons and uhos. Delimitation of the germline repertoire of human immunoglobulin kappa genes. Biol Chem Hoppe-Seyler 1992;373:177–86.
[53] Bowden DW, Krawchuk MD, Weaver EJ, Howard TD, Knowlton RG, Rao PN, Pettenati MJ, Hayworth R, Wagner BJ, Rothschild CB. D20s16 is a complex interspersed repeated sequence: genetic and physical analysis of the locus. Genomics 1995;25:394–403.
[54] Garrido-Ramos MA. Satellite DNA: an evolving topic. Genes (Basel) 2017;8.
[55] Jachowicz JW, Santenard A, Bender A, Muller J, Torres-Padilla ME. Heterochromatin establishment at pericentromeres depends on nuclear position. Genes Dev 2013;27:2427–32.
[56] Jagannathan M, Cummings R, Yamashita YM. A conserved function for pericentromeric satellite DNA. Elife 2018;7.
[57] Jagannathan M, Yamashita YM. Function of junk: pericentromeric satellite DNA in chromosome maintenance. Cold Spring Harb Symp Quant Biol 2018.
[58] Probst AV, Santos F, Reik W, Almouzni G, Dean W. Structural differences in centromeric heterochromatin are spatially reconciled on fertilisation in the mouse zygote. Chromosoma 2007;116:403–15.
[59] Zhu Q, Hoong N, Aslanian A, Hara T, Benner C, Heinz S, Miga KH, Ke E, Verma S, Soroczynski J, et al. Heterochromatin-encoded satellite RNAs induce breast cancer. Mol Cell 2018;70:842–53 e847.
[60] Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM. Brca1 tumour suppression occurs via heterochromatin-mediated silencing. Nature 2011;477:179–84.
[61] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61.
[62] Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
[63] Paul A, Paul S. The breast cancer susceptibility genes (brca) in breast and ovarian cancers. Front Biosci 2014;19:605–18.
[64] Calaf GM, Abarca-Quinones J. Ras protein expression as a marker for breast cancer. Oncol Lett 2016;11:3637–42.
[65] Hare SH, Harvey AJ. Mtor function and therapeutic targeting in breast cancer. Am J Cancer Res 2017;7:383–404.
[66] Kim RK, Suh Y, Yoo KC, Cui YH, Kim H, Kim MJ, Gyu Kim I, Lee SJ. Activation of kras promotes the mesenchymal features of basal-type breast cancer. Exp Mol Med 2015;47:e137.
[67] Paplomata E, O'Regan R. The pi3k/akt/mtor pathway in breast cancer: targets, trials and biomarkers. Ther Adv Med Oncol 2014;6:154–66.
[68] Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. Gepia: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45:W98–W102.
[69] Huda A, Marino-Ramirez L, Landsman D, Jordan IK. Repetitive DNA elements, nucleosome binding and human gene expression. Gene 2009;436:12–22.
[70] Shephard EA, Chandan P, Stevanovic-Walker M, Edwards M, Phillips IR. Alternative promoters and repetitive DNA elements define the species-dependent tissue-specific expression of the fmo1 genes of human and mouse. Biochem J 2007;406:491–9.
[71] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc 2009;4:44–57.
[72] Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. Nature 2005;436:876–80.
[73] Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, Watson M, Davies S, Bernard PS, Parker JS, et al. Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. J Natl Cancer Inst 2009;101:736–50.
[74] Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, Shi B. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res 2015;5:2929–43.
[75] Valla M, Vatten LJ, Engstrom MJ, Haugen OA, Akslen LA, Bjorngaard JH, Hagen AI, Ytterhus B, Bofin AM, Opdahl S. Molecular subtypes of breast cancer: long-term incidence trends and prognostic differences. In: Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 25; 2016. p. 1625–34.
[76] Menche J, Guney E, Sharma A, Branigan PJ, Loza MJ, Baribaud F, Dobrin R, Barabasi AL. Integrating personalized gene expression profiles into predictive disease-associated gene pools. NPJ Syst Biol Appl 2017;3:10.
[77] Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J 2005;24:800–12.
[78] Nishibuchi G, Dejardin J. The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. Chromosome Res 2017;25:77–87.
[79] Velazquez Camacho O, Galan C, Swist-Rosowska K, Ching R, Gamalinda M, Karabiber F, De La Rosa-Velazquez I, Engist B, Koschorz B, Shukeir N, et al. Major satellite repeat rna stabilize heterochromatin retention of suv39h enzymes by rna-nucleosome association and rna:DNA hybrid formation. Elife 2017;6.
[80] Kaczkowski B, Tanaka Y, Kawaji H, Sandelin A, Andersson R, Itoh M, Lassmann T, Hayashizaki Y, Carninci P, Forrest AR, et al. Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan–cancer biomarkers. Cancer Res 2016;76:216–26.
[81] Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. The dfam database of repetitive DNA families. Nucleic Acids Res 2016;44:D81–9.
[82] Leinonen R, Sugawara H, Shumway M. International nucleotide sequence database, C., the sequence read archive. Nucleic Acids Res 2011;39:D19–21.
[83] Kim D, Langmead B, Salzberg SL. Hisat: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60.
[84] Robinson MD, McCarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.
[85] Quinlan AR. Bedtools: the swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics 2014;47(11):11–34 12.
[86] Karakulah G, Suner A. Plantenrichment: a tool for enrichment analysis of transposable elements in plants. Genomics 2017;109:336–40.
[87] Wickham, H.; Sievert, C., *Ggplot2 : elegant graphics for data analysis.*second ed.; p xvi, 260 pages.