OXFORD

Briefings in Bioinformatics, 21(5), 2020, 1742–1755

doi: 10.1093/bib/bbz118
Advance Access Publication Date: 28 October 2019
Problem solving protocol

Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer

Siqi Bao, Hengqiang Zhao, Jian Yuan, Dandan Fan, Zicheng Zhang, Jianzhong Su and Meng Zhou

Corresponding authors: Meng Zhou, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, P. R. China. E-mail: zhoumeng@wmu.edu.cn; Jianzhong Su, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, P. R. China. E-mail: sujz@wibe.ac.cn

Abstract

Emerging evidence revealed the critical roles of long non-coding RNAs (lncRNAs) in maintaining genomic instability. However, identification of genome instability-associated lncRNAs and their clinical significance in cancers remain largely unexplored. Here, we developed a mutator hypothesis-derived computational frame combining lncRNA expression profiles and somatic mutation profiles in a tumor genome and identified 128 novel genomic instability-associated lncRNAs in breast cancer as a case study. We then identified a genome instability-derived two lncRNA-based gene signature (GILncSig) that stratified patients into high- and low-risk groups with significantly different outcome and was further validated in multiple independent patient cohorts. Furthermore, the GILncSig correlated with genomic mutation rate in both ovarian cancer and breast cancer, indicating its potential as a measurement of the degree of genome instability. The GILncSig was able to divide TP53 wide-type patients into two risk groups, with the low-risk group showing significantly improved outcome and the high-risk group showing no significant difference compared with those with TP53 mutation. In summary, this study provided a critical approach and resource for further studies examining the role of lncRNAs in genome instability and introduced a potential new avenue for identifying genomic instability-associated cancer biomarkers.

Key words: genome instability; mutator phenotype; long non-coding RNAs

Siqi Bao is a graduate student at the School of Biomedical Engineering, Wenzhou Medical University. Her research interests include bioinformatics and translational medicine.

Hengqiang Zhao is a student at the School of Biomedical Engineering, Wenzhou Medical University. His research interests include bioinformatics and disease systems biology.

JianYuan is a research assistant at the School of Biomedical Engineering, Wenzhou Medical University. His research interests include bioinformatics and cancer epigenetics.

Dandan Fan is a student at the School of Biomedical Engineering, Wenzhou Medical University. Her research interests include bioinformatics and cancer epigenetics.

Zicheng Zhang is a graduate student at the School of Biomedical Engineering, Wenzhou Medical University. His research interests include bioinformatics and translational medicine.

Jianzhong Su is a professor at the School of Biomedical Engineering, Wenzhou Medical University. His research interests in bioinformatics and cancer epigenetics.

Meng Zhou is an associate professor at the School of Biomedical Engineering, Wenzhou Medical University. His research interests include bioinformatics and computational RNomics.

Submitted: 24 June 2019; **Received (in revised form):** 29 July 2019

Introduction

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death in women, which nearly account for a quarter of female cancer cases [1]. The death rate for breast cancer dropped by 40% from 1989 to 2016 because of the increased implementation of mammographic screening and adjuvant systemic therapies for newly diagnosed cases. However, the long-term survival and prognosis of terminal breast cancer patients are still poor, particularly in developing countries [2, 3]. Currently, patient features (such as age, tumor grade, nodal involvement, ER, PR, HER2, etc.) are widely used to predict the progression or recurrence of patients with breast cancer [4]. It is well known that breast cancer is a quite complex disease characterized by molecular and clinical heterogeneity, as is the case for its development, progress and response to treatment [3]. Therefore, there is an urgent need to identify novel biomarkers to more accurately assess the clinical outcomes of breast cancer patients.

Genomic instability has been reported as one of the hallmarks of cancer [5]. More importantly, it is suggested that genomic instability is an important prognostic factor and the accumulation of genomic instability is associated with tumor progression and survival [6, 7]. Although the molecular basis of genomic instability is not fully understood, aberrant transcriptional and post-transcriptional regulation have been implicated in genome instability [8], demonstrating the potential of molecular signature as quantitative measurement of genomic instability. For example, Habermann *et al.* [9] analyzed gene expression profiling of 48 breast cancer specimens and identified a 12-gene genomic instability signature. A subsequent study by Wang *et al.* constructed DNA damage response-related miRNA-regulatory network and identified 10-miRNA signature that is associated with genome instability and outcome of ovarian cancer (OV) [10–12]. Long non-coding RNAs (lncRNAs) are broadly defined as transcripts that are larger than 200 nt and do not appear to have the potential of protein coding [13]. During the past several years, increasing evidence both *in vitro* and *in vivo* suggests that lncRNAs play important roles across diverse biological process [14, 15], especially that aberrant expression of lncRNAs may have an impact on cell proliferation, tumor progression or metastasis [16]. Several lncRNAs such as *H19* [17], *MALAT1* [18] and *PCA3* [19] are recognized to be highly expressed in tumor tissues prior to the application of next-generation sequencing (NGS) technologies. With the development of the NGS technology, large-scale evaluation of lncRNAs expression profiles in cancer and normal tissues can be performed, and substantial lncRNAs have been found to be abnormally expressed in various cancers, but the function of these lncRNAs are still largely unknown [20–23]. Emerging evidence revealed the critical roles of lncRNAs in maintaining genomic instability [24, 25]. For example, a recent study from Mendell *et al.* [26] showed that a specific lncRNA, non-coding RNA activated by DNA damage (*NORAD*), interacts with proteins involved in DNA replication and repair and contributes to genomic stability. Betts *et al.* [27] identified two lncRNAs, *CUPID1* and *CUPID2*, which control the expression of genes involved in DNA repair. LncRNA *DDSR1* contributes to genomic stability by binding DNA damage proteins or regulating the expression of DNA damage-related genes [28, 29]. Although several lncRNAs have been shown to be involved in genomic stability, genome instability-associated lncRNAs and their clinical significance in cancers still remain largely unexplored.

In this study, we tried to develop a mutator hypothesis-derived computational frame combining lncRNA expression profiles and somatic mutation profiles in a tumor genome to explore the possibility of lncRNA signature as an indicator of genomic stability, thereby improving its prognostic utility.

Materials and methods

Data collection

Clinical features, RNA-seq expression data and somatic mutation information of female patients with breast tumors were collected from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). LncRNA expression data were downloaded from TANRIC database (http://bioinformatics.mdanderson.org/main/TANRIC:Overview, version 1.0.6) [30]. A total of 795 female samples with paired lncRNA and mRNA expression profiles, survival information, somatic mutation information and common clinicopathological characteristics were retained for further study. All of the breast cancer patients used in this study was divided into two patient sets according to the batch, named training set and testing set separately. The training set consisted of TCGA batches 47, 56, 61, 72, 74, 81 and 85 with a total of 398 patients, which was used to identify prognostic lncRNA signature and build prognostic risk model. The testing set contained batches 93, 96, 103, 109, 117, 120, 124, 136, 142, 147, 155, 167, 177, 202 and 216 with a total of 397 patients, which was used to independently validate the performance of the prognostic risk model. Another two independent breast cancer validation sets GSE3494 [31] and GSE31448 [32] with large sample size and common clinicopathological characteristics were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31448).

In addition, somatic mutation information and the corresponding lncRNAs expression data of 294 OV patients were also downloaded from TCGA and TANRIC database. A brief summary of clinical and pathological characteristics was shown in Table 1.

Identification of genome instability-associated lncRNAs

To identify genome instability-associated lncRNAs, a mutator hypothesis-derived computational frame combining lncRNA expression profiles and somatic mutation profiles in a tumor genome was developed as shown in Figure 1: (i) the cumulative number of somatic mutations for each patient was computed; (ii) patients were ranked in decreasing order of the cumulative number of somatic mutations; (iii) the top 25% of patients were defined as genomic unstable (GU)-like group, and the last 25% were defined genomically stable (GS)-like group; (iv) expression profiles of lncRNAs between the GU group and GS group were compared using significance analysis of microarrays (SAM) method; (v) differentially expressed lncRNAs (fold change > 1.5 or <0.67 and false discovery rate (FDR) adjusted $P < 0.05$) were defined as genome instability-associated lncRNAs.

Statistical analysis

Hierarchical cluster analyses were performed using Euclidean distances and Ward's linkage method. Univariate and multivariate Cox proportional hazard regression analysis was used to evaluate the association between the expression level of genome instability-associated lncRNA and overall survival. Based on the

Table 1. Clinical information for three BRCA patients sets in this study

Covariates		Training set ($n = 398$)	Testing set ($n = 397$)	TCGA set ($n = 795$)	P-value
Age, no (%)	Young (<58)	178(44.7)	198(49.9)	376(47.3)	0.167[a]
	Old (≥58)	220(55.3)	199(50.1)	419(52.7)	
Progesterone receptor status, no (%)	Positive	253(63.6)	260(65.5)	513(64.5)	0.108[a]
	Negative	139(34.9)	110(27.7)	249(31.3)	
	Unknown	6(1.5)	27(6.8)	33(4.2)	
Estrogen receptor status, no (%)	Positive	299(75.1)	288(72.5)	587(73.8)	0.725[a]
	Negative	94(23.6)	84(21.2)	178(22.4)	
	Unknown	5(1.3)	25(6.3)	30(3.8)	
Her2 receptor status, no (%)	Positive	73(18.4)	54(13.6)	127(16.0)	0.154[a]
	Negative	213(53.5)	215(54.2)	428(53.8)	
	Unknown	112(28.1)	128(32.2)	240(30.2)	
Pathologic stage, no (%)	I	62(15.6)	74(18.6)	136(17.1)	0.248[b]
	II	225(56.5)	232(58.4)	457(57.5)	
	III	81(20.4)	86(21.7)	167(21.0)	
	IV	12(3.0)	2(0.5)	14(1.8)	
	Unknown	18(4.5)	3(0.8)	21(2.6)	
TP53 mutation status, no (%)	With	148(37.2)	133(33.5)	281(35.3)	0.311[a]
	Without	250(62.8)	264(66.5)	514(64.7)	
Vital status, no (%)	Alive	346(86.9)	332(83.6)	678(85.3)	0.224[a]
	Dead	52(13.1)	65(16.4)	117(14.7)	

[a] Chi square test
[b] Wilcoxon rank sum test

coefficients from the multivariate regression analysis and the expression levels of prognostic genome instability-associated lncRNAs, we constructed a genome instability-derived lncRNA signature (GILncSig) for outcome prediction as follows:

$$\text{GILncSig (patient)} = \sum_{i=1}^{n} \text{coef}\left(\text{lncRNA}_i\right) * \text{expr}\left(\text{lncRNA}_i\right),$$

where GILncSig (patient) is a prognostic risk score for the breast cancer patient. lncRNAi represents the ith prognostic lncRNA, and expr (lncRNAi) is the expression level of lncRNAi for the patient. coef (lncRNAi) represents the contribution of lncRNAi to prognostic risk scores that were obtained from the regression coefficient of multivariate Cox analysis. The median score of the patients in the training set was used as a risk cutoff to classify patients into the high-risk group with high GILncSig or low-risk group with low GILncSig.

The Kaplan–Meier method was used to calculate the survival rate and median survival for each prognostic risk group, and the log-rank test was used to assess the difference in survival between the high-risk group and the low-risk group with a significant level of 5%. Multivariate Cox regression and stratified analysis were used to assess the independence of GILncSig from other key clinical factors. Hazard ratio (HR) and 95% confidence interval (CI) were calculated by Cox analysis. The performance of the GILncSig also was evaluated by the time-dependent receiver operating characteristic (ROC) curve. All statistical analyses were performed using R-version 3.5.2.

Functional enrichment analysis

We computed the Pearson correlation coefficients to measure the correlation between the paired expression of lncRNAs and mRNAs, and the top 10 mRNAs were considered as co-expressed lncRNA-associated partners. To predict the potential functions of lncRNAs, we performed functional enrichment analysis of co-expressed lncRNA-associated mRNA partners to determine significantly enriched Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. The functional enrichment analysis was performed using clusterProfiler software in R- version 3.5.2 [33].

Result

Identification of genomic instability-related lncRNAs in breast cancer patients

To identify lncRNAs associated with genomic instability, the cumulative number of somatic mutations per patient was calculated and sorted in the decreasing order, and the top 25% ($n = 199$) and the last 25% ($n = 189$) of the patients were assigned to GU-like group and GS-like group, based on the cumulative number of somatic mutations. Then lncRNA expression profiles of the 199 patients in the GU-like group and 189 patients in the GS-like group were compared to find lncRNAs with significant differences. Using the SAM method, a total of 128 lncRNAs were considered to be significantly differentially expressed with their fold change value greater than 1.5 or less than 0.67 and their FDR-adjusted P-value less than 0.05. Among them, 31 lncRNAs were found to be upregulated and 97 to be downregulated in GU-like group (Supplementary Table 1). Unsupervised hierarchical clustering analysis was conducted on 795 samples of the TCGA set using the set of 128 differentially expressed lncRNAs. As shown in Figure 2A, all 795 samples were clustered into two groups according to the expression levels of the 128 differentially expressed lncRNAs. The somatic mutation pattern is significantly different between the two groups. The group with higher cumulative somatic mutations was named as GU-like group, and the other group was named as GS-like group. The median value of somatic cumulative mutations in the GU-like group was significantly higher than that in the GS-like group (57 versus 28.5, $P < 0.001$, Mann–Whitney U test; Figure 2B). We next compared the expression level of *UBQLN4* gene (a newly identified driver of genomic instability) between the GS-like group and GU-like

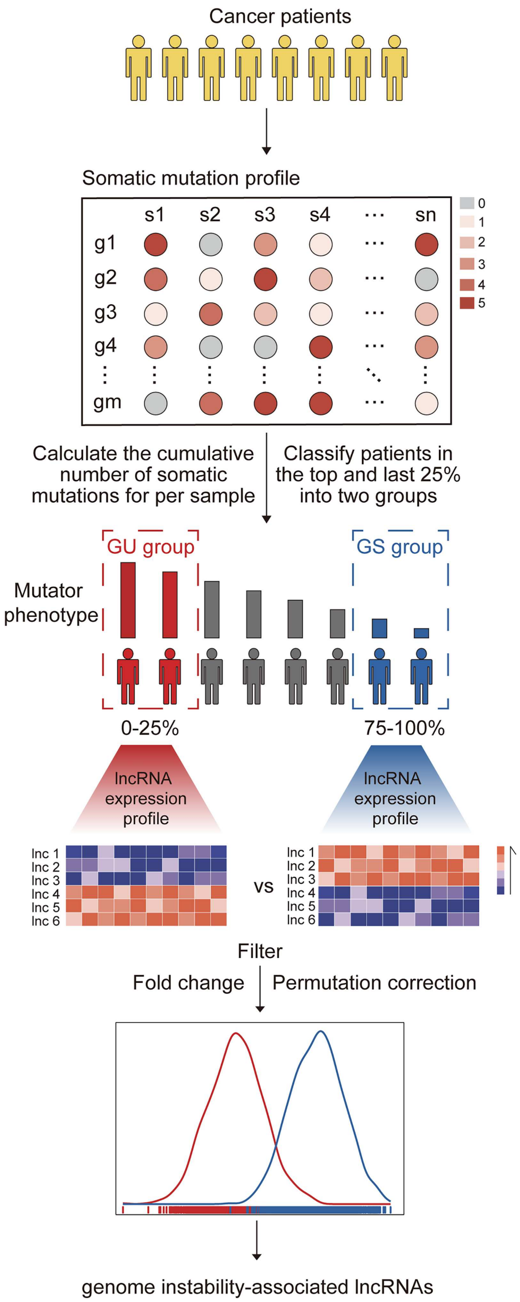
Cancer patients
Somatic mutation profile
s1
s2
s3
s4
···
sn
g1
g2
g3
g4
gm
0
1
2
3
4
5
Calculate the cumulative number of somatic mutations for per sample
Classify patients in the top and last 25% into two groups
GU group
GS group
Mutator phenotype
0-25%
75-100%
lncRNA expression profile
lncRNA expression profile
lnc 1
lnc 2
lnc 3
lnc 4
lnc 5
lnc 6
vs
Filter
Fold change
Permutation correction
genome instability-associated lncRNAs

Figure 1. Computational overview of genomic instability-related lncRNAs detection. Somatic mutation profile was built. The columns reflect breast cancer samples, and the rows reflect genes. The value reflects the number of altered sites for each gene on each sample. Samples were divided into two groups, GU group (patients' mutator phenotype ranked in the top 25%) and GS group (patients' mutator phenotype ranked in the last 25%), according to their mutator phenotype. Genomic instability-related lncRNAs were detected by comparing the lncRNA expression profile between GU group and GS group.

group. As shown in Figure 2C, the expression of *UBQLN4* in the GU-like group is significantly higher than that in the GS-like group ($P < 0.001$, Mann–Whitney U test).

To determine whether the potential functions and pathways involved in 128 lncRNAs are associated with genomic instability, we performed functional enrichment analysis to predict potential functions. We first measured the expression correlation between the 128 differentially expressed lncRNAs and protein-coding genes (PCGs) and obtained lncRNA-correlated PCGs, which were the top 10 PCGs most correlated with each lncRNAs. An lncRNAs–mRNA co-expression network was constructed in which the nodes are lncRNAs and mRNAs, and if they are related to each other, the lncRNA and mRNA are linked together (Figure 2D). GO analysis of lncRNA-correlated PCGs revealed that mRNAs in this network are significantly associated with the formation and development of genomic instability, including cell cycle checkpoint, nucleotide-excision repair and DNA integrity checkpoint (Figure 2E). In light of KEGG pathway analysis of lncRNA-correlated PCGs, 11 significantly enriched pathways were found, most of which are linked to genomic instability, including cell cycle, transcriptional misregulation in cancer, nucleotide excision repair and mismatch repair (Figure 2E). These results indicated that the 128 differentially expressed lncRNAs are involved in genome instability, and their altered expression may destruct the genomic stability of cells by disrupting the balance of the lncRNA-related PCGs regulatory network to affect the normal gene damage repair pathways leading to an increase in genomic instability. Based on the above results, 128 differentially expressed lncRNAs were considered as candidate genome instability-associated lncRNAs (GUlncRNAs).

Development of a genomic instability-derived lncRNA signature for outcome prediction in the training set

To further investigate the prognostic roles of these candidate genome instability-associated lncRNAs, 795 breast cancer patients from the TCGA project were divided into the training set ($n = 398$, batch 47, 56, 61, 72, 74, 81 and 85) and the testing set ($n = 397$, batch 93, 96, 103, 109, 117, 120, 124, 136, 142, 147, 155, 167, 177, 202 and 216) according to their batches. To screen for prognostic-related lncRNAs, univariate Cox proportional hazard regression analysis was used to analyze the relationship between expression levels of 128 genome instability-associated lncRNAs and OS in the training set, and 10 genome instability-associated lncRNAs were found to be significantly associated with the prognosis of breast cancer patients ($P < 0.05$; Table 2). In addition, we hoped to single out lncRNAs with the independent prognostic value from these 10 candidate lncRNAs, and then multivariate Cox proportional hazards regression analysis was carried out among the 10 candidate lncRNAs and common clinical features such as age, pathological stage, estrogen receptor status and progesterone receptor status. Finally, 2 of 10 candidate lncRNAs (*RP11-358L4.1* and *LINC02207*) were identified as independent prognostic lncRNAs as they retained their prognostic significance in multivariate Cox ($P < 0.05$). Then a genomic instability-derived lncRNA signature (GILncSig) was constructed to assess the prognosis risk of breast cancer patients based on the coefficients of multivariate Cox analysis and the expression level of two independent prognostic genomic instability-associated lncRNAs as follow: GILncSig score = (0.2408 × expression level of *RP11-358L4.1*) + (−0.5927 × expression level of *LINC02207*). Of the GILncSig, the coefficient of lncRNA *RP11-358L4.1* was positive, indicating that it may be a risky factor because its high expression was associated with a poor prognosis, while the other lncRNA *LINC02207* tended to be a protective factor, and its high expression was associated with longer survival. The risk score of each patient in the training set was obtained by using the GILncSig, and then these patients were classified into different prognostic groups using the median risk score (0.141) as a threshold. The group, in which patient's scores are equal

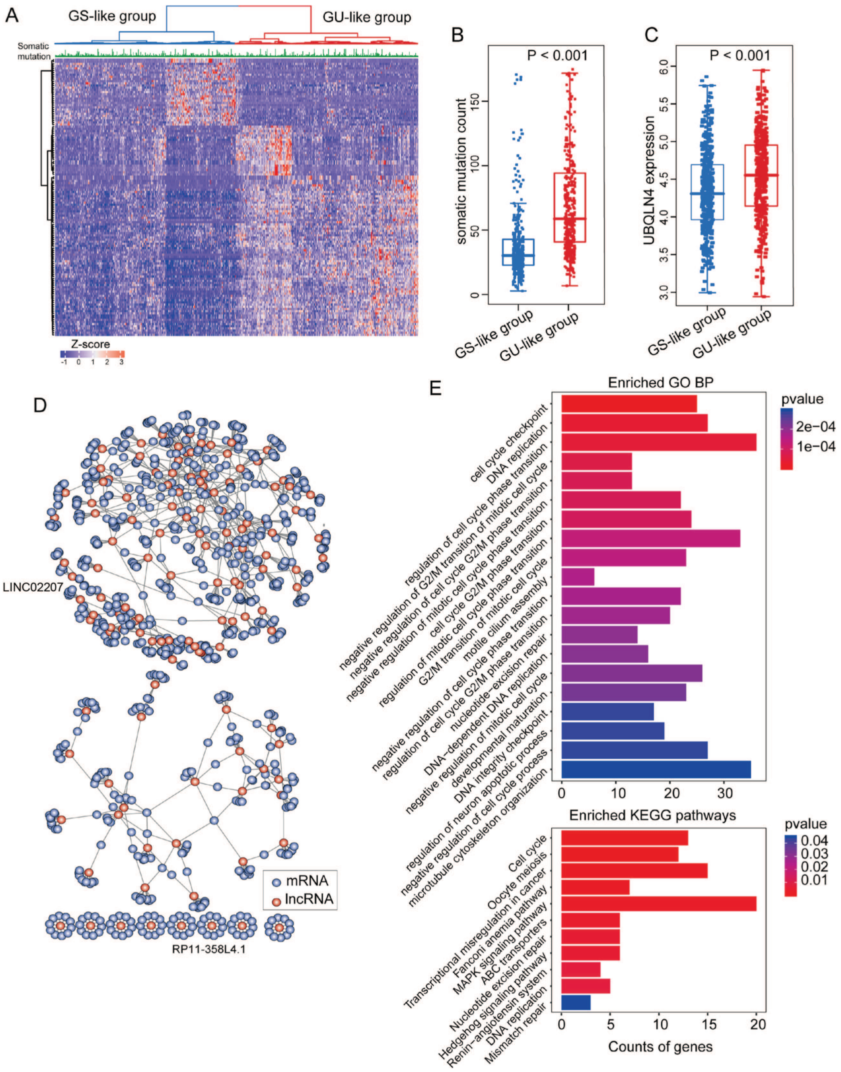

Figure 2. Identification and functional annotations of genomic instability-related lncRNAs in patients with breast cancer. (**A**) Unsupervised clustering of 795 breast cancer patients based on the expression pattern of 128 candidate genomic instability-related lncRNAs. The left blue cluster is GS-like group, and the right red cluster is GU-like group. (**B**) Boxplots of somatic mutations in the GU-like group and GS-like group. Somatic cumulative mutations in the GU-like group are significantly higher than those in the GS-like group. (**C**) Boxplots of UBQLN4 expression level in the GU-like group and GS-like group. The expression level of *UBQLN4* in the GU-like group is significantly higher than that in the GS-like group. Horizontal lines: median values. Statistical analysis was performed using the Mann–Whitney U test. (**D**) Co-expression network of genomic instability-related lncRNAs and mRNAs based on the Pearson correlation coefficient. The red circles represent lncRNAs, and the blue circles represent mRNAs. (**E**) Functional enrichment analysis of GO and KEGG for mRNAs co-expressed lncRNAs.

to or higher than the threshold, is named high-risk group, and the other group is named low-risk group. Kaplan–Meier analysis showed that the survival outcomes of patients in the low-risk group are significantly better than patients in the high-risk group (median OS 20.4 years versus 10.8 years, $P = 0.002$, log-rank test; Figure 3A). The survival rate of the high-risk group was 23% at 5 years and that of the low-risk group was 29%. In the univariate analysis, the HR of high-risk group versus low-risk group for overall survival was 2.382 (95% CI 1.338–4.240, $P = 0.003$; Figure 3A). The time-dependent ROC curves analysis of the GILncSig yielded an area under curve (AUC) of 0.747 (95% CI: 0.636–0.858; Figure 3B). We sorted the patients in the training set according to their score and observed how the expression levels of the GILncSig, the count of somatic mutations in patients and the expression level of *UBQLN4* change with the increasing score (Figure 3C). For patients with high scores, the expression level of the risk lncRNA *RP11-358L4.1* was upregulated, while the protective lncRNA *LINC02207* was downregulated. In contrast, the GILncSig in patients with low scores showed opposite expression patterns. Comparison analysis showed significant differences in somatic mutation pattern and *UBQLN4* expression pattern between the patients in the high-risk and low-risk group. As shown in Figure 3D, the number of somatic mutation of patients in the high-risk group is significantly higher compared to that of patients in the low-risk group (median somatic mutation counts 47 versus 36, $P < 0.001$, Mann–Whitney U test; Figure 3D). Moreover, *UBQLN4* also revealed higher expression levels in high-risk patients than in the low-risk patients ($P = 0.016$, Mann–Whitney U test; Figure 3E). It is well known that genomic instability is an important feature of OV. Therefore, we test the GILncSig in TCGA OV patients. The cumulative number of somatic mutations and the risk score for each OV patient was calculated as did in breast cancer. Comparison analysis showed that the somatic mutation

Table 2. Univariate Cox regression analyses of the 10 of 128 genome instability-related lncRNAs associated with overall survival in BRCA

Ensembl ID	Gene Symbol	Genomic location	Coefficient	HR	95% CI	*P*-value
ENSG00000234996	LOC148709	chr1:202,830,882-202,844,369	0.237	1.267	1.069–1.502	0.006
ENSG00000163597	SNHG16	chr17:74,553,848-74,561,430	0.208	1.232	1.050–1.445	0.010
ENSG00000261712	RP11-358L4.1	chr15:79,705,700-79,706,228	0.259	1.295	1.051–1.596	0.015
ENSG00000258476	LINC02207	chr15:94,406,338-94,421,585	−0.596	0.551	0.335–0.908	0.019
ENSG00000238273	AC108058.1	chr2:105,979,495-105,995,296	−0.801	0.449	0.224–0.901	0.024
ENSG00000262136	AC092115.3	chr16:69,760,916-69,776,466	0.259	1.295	1.033–1.624	0.025
ENSG00000233694	LINC02579	chr2:64,834,109-64,843,616	0.308	1.361	1.037–1.785	0.026
ENSG00000254414	RP11-182 J1.1	chr15:85,175,129-85,177,218	0.323	1.381	1.004–1.900	0.047
ENSG00000266904	LINC00663	chr19:19,868,175-19,887,232	0.264	1.303	1.003–1.691	0.047
ENSG00000231881	AL109615.2	chr6:44,041,650-44,045,689	0.282	1.326	1.001–1.757	0.049

counts of OV patients with high score were significantly higher than that with low score (median 74.5 versus 59.5, $P = 0.051$, Mann–Whitney U test; Figure 3F). All OV patients were classified into a high-mutation group and a low-mutation group, and the GILncSig scores of patients in the high-mutation group were marginally significantly higher than those in the low-score patients ($P = 0.064$, Mann–Whitney U test; Figure 3G).

Independent validation of GILncSig in the breast cancer data set with RNA-seq platform

To examine the robustness of the GILncSig, the GILncSig was then tested for its prognostic performance in the independent TCGA testing set of 397 patients. When the same GILncSig and risk cutoff as those derived from the training set was applied to the testing set, 397 patients of the testing set was classified into the high-risk group ($n = 177$) and low-risk group ($n = 220$) with significantly different overall survival. As shown in Figure 4A, the overall survival of 220 patients in the high-risk group was much poorer than 177 patients in the low-risk group (median OS 7.85 versus 10.85 years, $P = 0.044$, log-rank test). Similar to the training set, the survival rate was 24% in the high-risk group at 5 years lower than 28% in the low-risk group. In the univariate analysis, the HR of high-risk group versus low-risk group for overall survival was 1.658 (95% CI: 1.010–2.722, $P = 0.046$; Figure 4A). The time-dependent ROC curves analysis of the GILncSig in the testing set yielded an AUC of 0.610 (95% CI: 0.482–0.738; Supplementary Figure 1A). The expression of the GILncSig and the distribution of somatic mutation count and *UBQLN4* expression in the testing samples were illustrated in Figure 4B. The somatic mutation pattern was significantly different between the patients in the high-risk and low-risk group (median 46 versus 29.5, $P < 0.001$, Mann–Whitney U test; Figure 4C). The expression level of *UBQLN4* in the high-risk group was observed to be marginally significantly higher than that in the low-risk group ($P = 0.084$, Mann–Whitney U test; Figure 4C).

The prognostic performance of the GILncSig in the TCGA set was similar to the above results. Patients of the TCGA set were assigned to the high-risk group ($n = 376$) and low-risk group ($n = 419$), which the median survival of patients in the high-risk group was shorter than the patients in the low-risk group (10.6 versus 18.1 years, $P < 0.001$, log-rank test; Figure 4D). The survival rate was 23% in the high-risk group at 5 years lower than 28% in the low-risk group. The HR of high-risk group versus low-risk group for overall survival was 1.861 (95% CI: 1.289–2.688, $P = 0.001$; Figure 4D). The expression of the GILncSig and the distribution of somatic mutation count and *UBQLN4* expression in the TCGA samples were illustrated in Figure 4E. The time-dependent ROC curves analysis was applied to the TCGA set, and consistent results were observed as above (AUC = 0.673, 95% CI: 0.585–0.761; Supplementary Figure 1B). The distribution of the somatic mutation counts was significantly different between the patients in the high-risk and low-risk group (median somatic mutation counts 47 versus 33, $P < 0.001$, Mann–Whitney U test; Figure 4F). The expression level of *UBQLN4* in the high-risk group was significantly higher than that in the low-risk group ($P = 0.003$, Mann–Whitney U test; Figure 4F).

Further validation of the GILncSig in two external independent breast cancer data sets with microarray platform

To pursue a cross-platform validation of the GILncSig in other independent data sets with different platforms, we reannotated the common microarray array and found that only one lncRNA (*RP11-358L4.1*) of two lncRNAs in the GILncSig was covered by GSE3494 and GSE31448 with large sample size and common clinicopathological characteristics. Therefore, we examined the association of *RP11-358L4.1* with breast cancer and genome genomic instability in these two independent microarray data sets. As shown in GSE3494 set, there is a good correlation between *RP11-358L4.1* expression and tumor grade. The expression level of *RP11-358L4.1* among different patients with different grade is significantly different ($P < 0.001$, Kruskal–Wallis Test; Figure 5A). Furthermore, patients with lymph node metastasis tended to express high *RP11-358L4.1* compared to those without lymph node metastasis ($P = 0.004$, Mann–Whitney U test; Figure 5A). Similar results were observed in GSE31448 data set. As shown in Figure 5B, high *RP11-358L4.1* expression is significantly associated with tumor progression ($P = 0.002$, Kruskal–Wallis Test; Figure 5B) in the GSE31448 (unavailable lymph node metastasis status). Finally, we further explored the association of *RP11-358L4.1* expression with *UBQLN4* expression and found that *UBQLN4* expression in patients with high *RP11-358L4.1* expression is significantly higher than that in patients with low *RP11-358L4.1* expression in both GEO data sets (GSE3494 $P = 0.046$; GSE31448 $P = 0.047$; Mann–Whitney U test; Figure 5C). These results are consistent with those observed in the training set and TCGA testing set.

Performance comparison of the GILncSig with existing lncRNA-related signatures in survival prediction

We further compared the prediction performance of the GILncSig with two recently published lncRNA signatures: 5-lncRNA

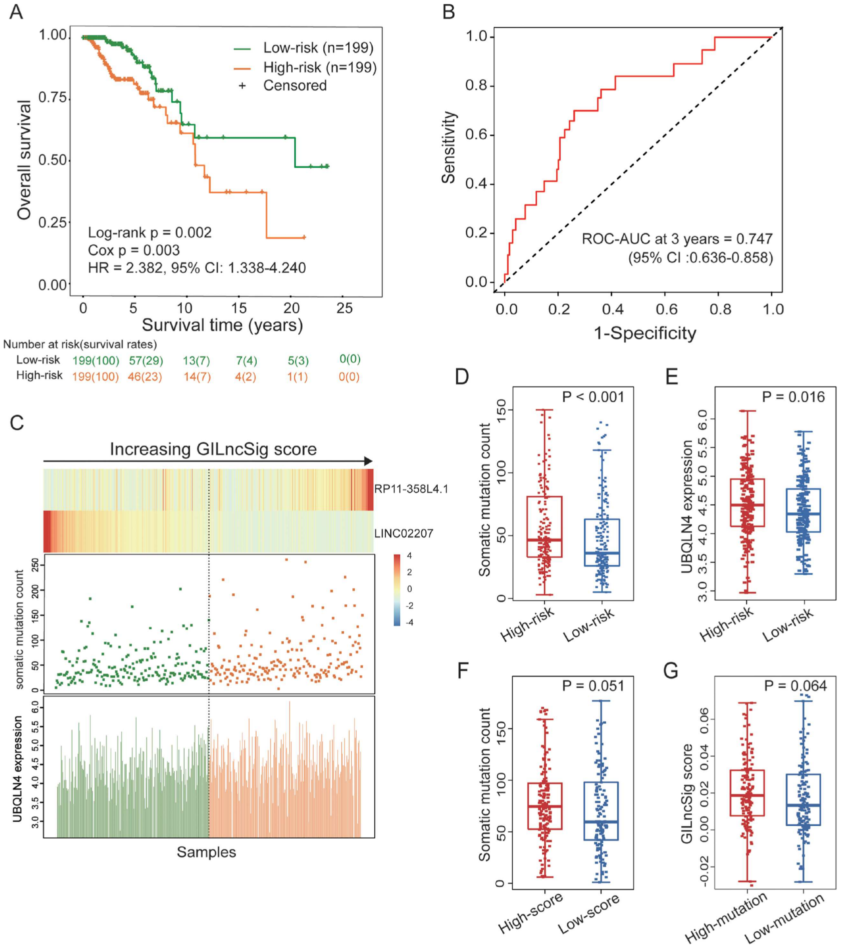

Figure 3. Identification of the genomic instability-derived lncRNA signature (GILncSig) for outcome prediction in the training set. (**A**) Kaplan–Meier estimates of overall survival of patients with low or high risk predicted by the GILncSig in the training set. Statistical analysis was performed using the log-rank test and univariate Cox analysis. (**B**) Time-dependent ROC curves analysis of the GILncSig at 3 years. (**C**) LncRNA expression patterns and the distribution of somatic mutation and *UBQLN4* expression with increasing GILncSig score. The distribution of somatic cumulative mutations (**D**) and *UBQLN4* expression in the (**E**) in the high- and low-risk groups for breast cancer patients. The red represents the high-risk group, and the blue represents the low-risk group. (**F**) Distribution of somatic cumulative mutations in the high- and low-risk groups for TCGA OV patients. The red represents the high-risk group, and the blue represents the low-risk group for TCGA OV patients. (**G**) Boxplots of GILncSig score in the High-mutation group and Low-mutation group for TCGA OV patients. Horizontal lines: median values. Statistical analysis was performed using the Mann–Whitney U test.

signature derived from Li's study (hereinafter referred to as LilncSig) [34] and 12-lncRNA signature derived from Sun's study (hereinafter referred to as SunlncSig) [35] using the same TCGA patient cohort. As shown in Figure 6, the AUC at 3 years of OS for the GILncSig is 0.673, which is significantly higher than that of LilncSig (AUC = 0.549) and SunlncSig (AUC = 0.509) (Figure 6). Furthermore, the number of lncRNAs included in the GILncSig is smaller than that included in the LilncSig (5 lncRNAs) and SunlncSig (12 lncRNAs). These results demonstrated the better prognostic performance of the GILncSig in predicting survival than two recently published lncRNA signatures.

Independence of the GILncSig from other clinical factors

To evaluate whether the prognostic value of the GILncSig was independent of common clinical variables, multivariate Cox regression analyses were performed on age, progesterone receptor status, estrogen receptor status, pathologic stage and our GILncSig-based prognostic risk score model. The results of the multivariate analyses suggested that the GILncSig was significantly correlated to overall survival in each set when adjusted for age, progesterone receptor status, estrogen receptor status and pathologic stage (Table 3). In addition to the GILncSig, there were other two clinical factors, age and pathologic stage, that were observed to be significant in the multivariate analysis. Therefore, stratification analysis was performed to determine whether the GILncSig possessed a prognostic value that was independent of the age and pathologic stage. Patients in the TCGA set were stratified into a young-patient ($n = 376$) and an old-patient group ($n = 419$) according to the median age (age = 58) of the total patients. Using the GILncSig, patients in each age group could be further divided into high-risk or low-risk group. There was a significant difference in overall survival between the high-risk and low-risk groups in the young-patient group (log-rank test $P = 0.005$; Figure 7A) as was in the old-patient group (log-rank test $P = 0.056$; Figure 7B). Next, all breast cancer patients were also stratified by pathologic stage, which patients with pathologic stage I or II were combined into an early-stage group ($n = 593$) and that with pathologic stage III or IV were combined into a late-stage group ($n = 181$). The GILncSig could classify the patients with pathologic stage I or II into high-risk group

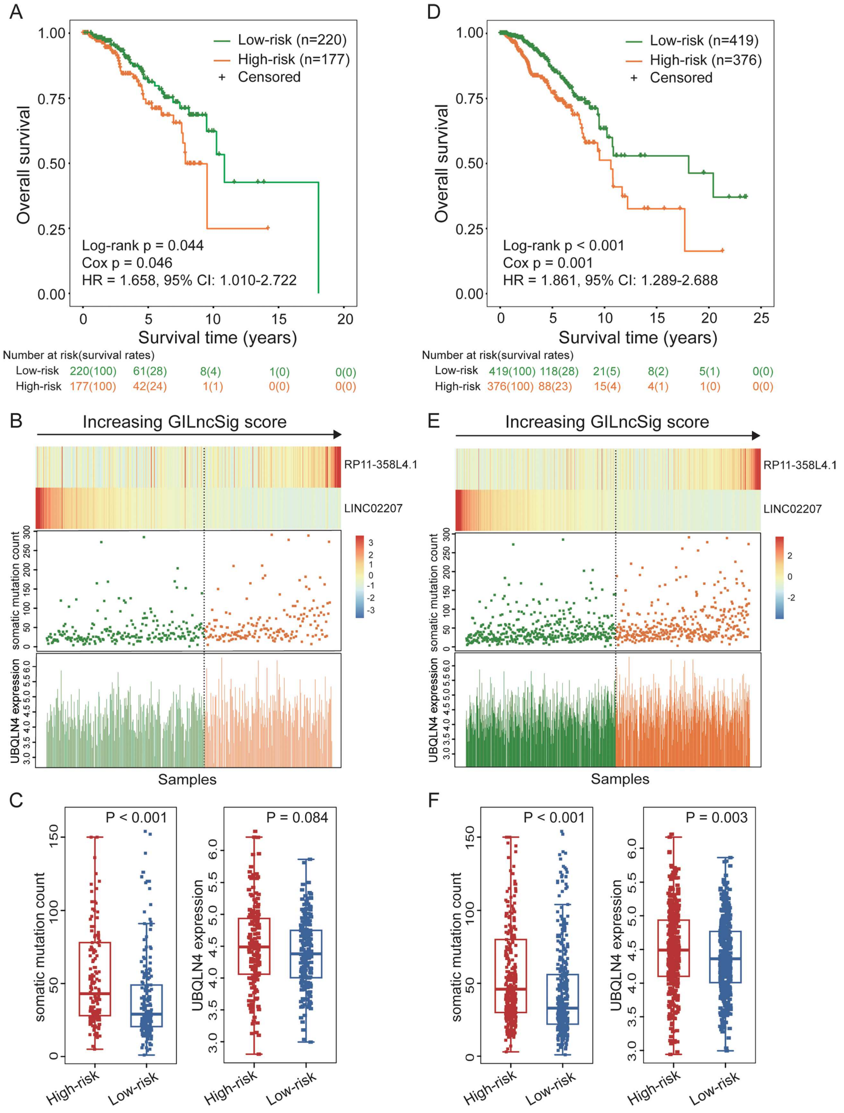

Figure 4. Performance evaluation of the GILncSig in the TCGA set. Kaplan–Meier estimates of overall survival of patients with low or high risk predicted by the GILncSig in the testing set (**A**) and TCGA set (**B**). Statistical analysis was performed using the log-rank test and univariate Cox analysis. LncRNA expression patterns and the distribution of somatic mutation count distribution and *UBQLN4* expression for patients in high- and low-risk groups in the testing set (**C**) and TCGA set (**D**). The distribution of somatic mutation and *UBQLN4* expression in patients of high- and low-risk groups in the testing set (**E**) and TCGA set (**F**). Horizontal lines: median values. Statistical analysis was performed using the Mann–Whitney U test.

($n = 279$) and low-risk group ($n = 314$), and the overall survival was significantly different between the two groups (log-rank test $P = 0.008$; Figure 7C). Similarly, the GILncSig could also be used to separate patients with pathologic stage III or IV into high-risk group ($n = 85$) and low-risk group ($n = 96$), and there was a statistically significant difference in overall survival between the two groups (log-rank test $P = 0.021$; Figure 7D). These results indicated that the GILncSig was an independent prognostic factor associated with overall survival in breast cancer patients.

The GILncSig predicts outcome better than TP53 mutation status

Further analysis showed that the proportion of patients with TP53 mutations in the high-risk group was significantly higher than that in the low-risk group among the training set, testing set and TCGA set (Figure 8A). In the training set, 91 patients (54.3%) in the high-risk group possessed TP53 mutations, significantly higher than 57 patients (28.6%) in the low-risk group (chi-square test $P < 0.001$). Similar results were found in the testing set and the TCGA set. In the testing set, 84 patients (52.5%) in the high-risk group possessed TP53 mutations, significantly higher than 49 patients (22.3%) in the low-risk group (chi-square test $P < 0.001$). In the TCGA set, 175 patients (53.5%) in the high-risk group possessed TP53 mutations, significantly higher than 106 patients (25.3%) in the low-risk group (chi-square test $P < 0.001$). Furthermore, the expression level of *RP11-358L4.1* in patients with TP53 mutations is significantly higher than those without TP53 mutations ($P = 0.003$, Mann–Whitney U test; Figure 8B) in the GSE3494 data set. In GSE31448, patients with TP53 mutations tended to express higher *RP11-358L4.1* compared to those without TP53 mutations although this did not reach but approached significance ($P = 0.132$, Mann-Whitney U test; Figure 8B). These results indicated that the GILncSig are also correlated with TP53

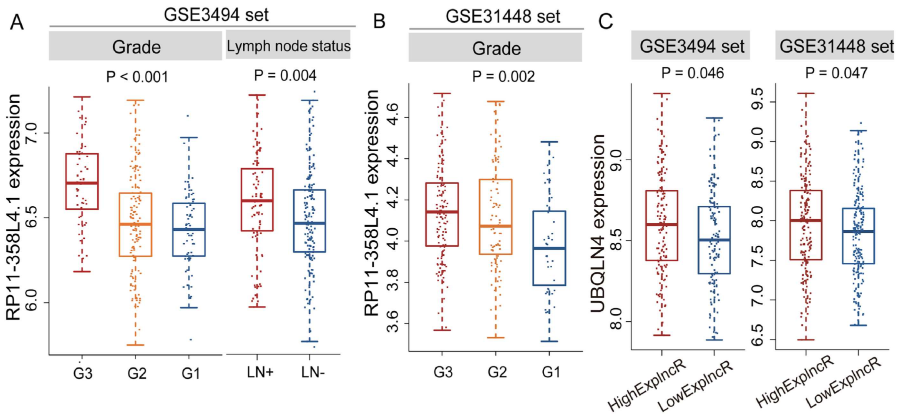

Figure 5. Performance evaluation of the GILncSig in two external independent GEO data sets. (**A**) Boxplots for *RP11-358L4.1* expression among patients with different grade and lymph node status in GSE3494 set. (**B**) Boxplots for *RP11-358L4.1* expression among patients with different grade in GSE31448 set. (**C**) Boxplots for *UBQLN4* expression levels among patients with high and low *RP11-358L4.1* expression. Comparison among three groups was performed by Kruskal–Wallis Test, and comparison between two groups was performed by Mann–Whitney U test.

Table 3. Univariate and Multivariate Cox regression analysis of the GILncSig and overall survival in different patient sets

Variables		Univariable model			Multivariable model		
		HR	95% CI	*P*-value	HR	95% CI	*P*-value
Training set (*n* = 398)							
GILncSig	High/Low	2.382	1.338–4.240	0.003	3.045	1.562–5.938	0.001
Age		1.032	1.009–1.056	0.005	1.050	1.022–1.077	<0.001
Progesterone receptor status	Positive/Negative	0.731	0.417–1.280	0.273	0.289	0.125–0.667	0.004
Estrogen receptor status	Positive/Negative	1.106	0.574–2.133	0.763	1.633	0.648–4.112	0.298
Her2 receptor status	Positive/Negative	0.908	0.301–2.742	0.864			
Pathologic stage	(III + IV)/(I + II)	3.672	2.029–6.646	<0.001	4.146	2.229–7.710	<0.001
Pathologic N	Metastasis/N0	2.319	1.250–4.304	0.008			
Testing set (*n* = 397)							
GILncSig	High/Low	1.658	1.010–2.722	0.046	1.823	1.038–3.203	0.037
Age		1.034	1.015–1.053	<0.001	1.040	1.020–1.061	<0.001
Progesterone receptor status	Positive/Negative	0.937	0.543–1.616	0.814	1.092	0.478–2.499	0.834
Estrogen receptor status	Positive/Negative	0.773	0.433–1.379	0.383	1.039	0.414–2.613	0.934
Her2 receptor status	Positive/Negative	2.311	1.173–4.554	0.015			
Pathologic stage	(III + IV)/(I + II)	2.253	1.348–3.766	0.002	3.150	1.802–5.507	<0.001
Pathologic N	Metastasis/N0	2.099	1.209–3.643	0.008			
TCGA set (*n* = 795)							
GILncSig	High/Low	1.861	1.289–2.688	0.001	2.176	1.441–3.286	<0.001
Age		1.033	1.019–1.048	<0.001	1.041	1.025–1.057	<0.001
Progesterone receptor status	Positive/Negative	0.841	0.570–1.241	0.384	0.601	0.335–1.078	0.088
Estrogen receptor status	Positive/Negative	0.929	0.602–1.433	0.739	1.325	0.689–2.549	0.399
Her2 receptor status	Positive/Negative	1.606	0.914–2.819	0.099			
Pathologic stage	(III + IV)/(I + II)	2.793	1.901–4.102	<0.001	3.44	2.293–5.160	<0.001
Pathologic N	Metastasis/N0	2.288	1.519–3.447	<0.001			

mutation status and may be the mutation marker for the TP53 gene. It is well known that TP53 maintains genomic stability and TP53 mutations are correlated with worse survival and can be used as an independent prognostic marker in breast cancer [36–40]. Therefore, we further test whether the GILncSig could predict outcome better than TP53 mutation status. When the GILncSig was applied to patients with TP53-sequence wild type (TP53-wt), the GILncSig separated TP53-wt patients into low- (*n* = 313) and high-risk groups (*n* = 201) with significantly different survival (median OS 10.85 years versus 9.34 years, *P* = 0.005, log-rank test) (Figure 8C). More interestingly, TP53-wt patients in the low-risk group (defined as TP53-wt/wt-like) have better outcome than those with TP53-sequence mutation type (TP53-mt) (*P* = 0.064, log-rank test), whereas TP53-wt patients in the high-risk group (defined as TP53-wt/mt-like) reveals similar outcome with those TP53-mt patients (*P* = 0.251, log-rank test). Figure 8D showed the survival curves of four risk groups classified by the GILncSig, which were TP53 wt/wt-like group, TP53 mt/wt-like group, TP53 wt/mt-like group and TP53 mt/mt-like group (median OS 10.85 years versus 20.42 years versus 9.34 years versus 17.69 years, *P* = 0.01, log-rank test). As shown in Figure 8D, the survival curve of TP53 mt/mt-like group was more similar to TP53 wt/mt-like group but not that similar to the TP53 mt/wt-like group. Therefore, these findings indicated that the GILncSig may have greater prognostic significance than TP53 mutation status alone.

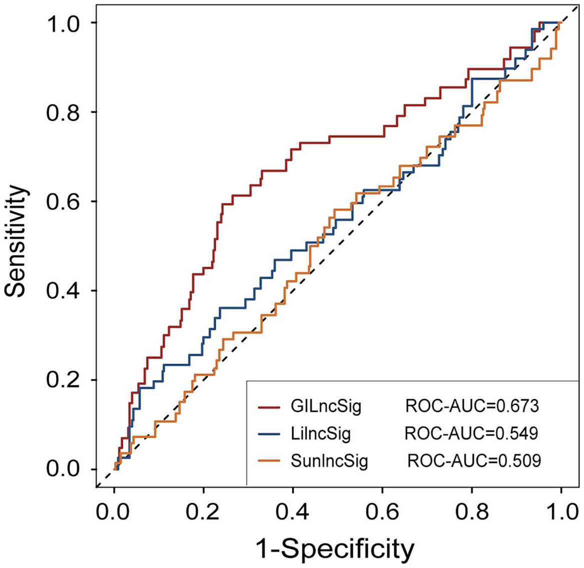

Figure 6. The ROC analysis at 3 years of overall survival for the GILncSig, LilncSig and SunlncSig.

Discussion

During the past years, many efforts were made to study the initiation, development and treatment of breast cancer [41–44]. Traditional histopathological features of tumor size, stage and grade and the Nottingham Prognostic Index (NPI) are still used as the most important prognostic factors in breast cancer, and patients are stratified to different therapeutic groups depending on their pathological features [45–48]. However, the clinical outcome of breast cancer patients remains highly heterogeneous due to the limitations of the traditional clinicopathological features [49]. Genomic instability has been reported to be not only the ubiquitous feature of most cancers [5, 50, 51] but also is one of the influence factors of breast cancer prognosis [52]. The genomic instability plays essential and dominating roles in cancer progression and recurrence, indicating that the pattern and degree of genomic instability have important diagnostic and prognostic implications [52, 53]. However, the quantitative measure for the degree of genomic instability has been a major challenge. Increasing evidence has suggested that aberrant transcriptional or epigenetic changes contributed to genomic instability [54]. Continuous efforts are made to identify genomic instability-related PCGs and microRNAs and develop gene or miRNA signature for predicting genomic instability [9, 53, 55].

More recently, lncRNAs, a novel class of ncRNAs, have been shown to be an important component of tumor biology, and dysregulated expression ncRNAs in cancer was related to disease progression and may have potential as prognostic markers for patients [56–58]. Recent advances in the understanding of functional mechanisms of lncRNAs have led to the realization that lncRNAs also are essential for genomic stability, such as *NORAD* [59] and *GUARDIN* [60]. Although some efforts have been made, genome-wide identification of genome instability-associated lncRNAs and systematic exploration on their clinical significance in cancers are still in its infancy. Therefore, we developed a computational frame in identifying genome instability-associated lncRNAs combining lncRNA expression and tumor mutator phenotype. Then we combined lncRNA expression profiles with somatic mutation profiles of breast cancers as a case study and identified 128 novel genome instability-associated lncRNAs. Based on functional analysis of genes that co-expressed with the 128 genomic instability-related lncRNAs, our observations suggested that the genes that co-expressed with the 128 lncRNAs were enriched in cell cycle checkpoint and nucleotide-excision repair. Cell cycle checkpoint is a set of check mechanisms in the cell cycle that ensure the order, integrity and fidelity of the main events, and uncorrected errors in the cell cycle will lead to genomic instability, a distinct characteristic of cancer [61–63]. Nucleotide excision repair can specifically prevent mutations caused by environmental carcinogens and thus maintain genomic stability to some extent [64]. We also observed enrichment for genes involved in pathways such as cell cycles, transcriptional misregulation in cancer, Fanconi anemia pathways, nucleotide excision repair and mismatch repair, which are associated with genomic instability [64–71]. We further examined whether genomic instability-related lncRNAs could allow the prediction of clinical outcome and resulted in a lncRNA signature (GILncSig) comprising two genomic instability-related lncRNAs (*RP11-358L4.1* and *LINC02207*). The GILncSig separated patients into two risk groups with significantly different survival in the training set, which was validated on the independent testing set. Furthermore, the GILncSig are significantly correlated with tumor mutator phenotype and *UBQLN4* expression in breast cancer and OV, both of which are important indicators of genomic instability. In addition, similar results were also found in two external GEO data sets. We consistently observed the association between *RP11-358L4.1* expression and breast cancer phenotype or *UBQLN4* expression (Figure 5). After a careful literature search, we found that the biological function of these two lncRNAs in the GILncSig has not been reported until now. But, we found that the lncRNA; *LINC02207* is located in chromosome 15q26 known for breast cancer susceptibility loci in previous genome-wide association analysis [72]. Another lncRNA *RP11-358L4.1* is located in chromosome 15q25 region that was newly reported to be involved in disease-associated genomic instability [73]. These validation results in multiple data sets across different technology platforms, together with literature mining results, indicated that the GILncSig not only predicts the prognosis but also is an indicator of genomic instability of cancer patients.

In line with previous research, TP53 mutation ratio of patients in the high-risk group was significantly higher than patients in the low-risk group according to the GILncSig, suggesting that the GILncSig could capture the TP53 mutation status. Furthermore, we also observed that the GILncSig could significantly distinguish different clinical outcome in TP53 wild-type patients. TP53 wide-type patients in the low-risk group had significantly longer survival than those with TP53-sequence mutation type. In contrast, TP53 wide-type patients in the high-risk group were not significantly different from those with TP53-sequence mutation type. The significant difference in survival between high-risk and low-risk of TP53 wide-type patients suggested that the GILncSig may have greater prognostic significance than TP53 mutation status alone and are capable of identifying intermediate subtype group existing with partial TP53 functionality in TP53 wild-type patients.

Though our study provides important insights to better evaluate genome instability and the prognosis of breast cancer patients, it still has some limitations that require further study. Although the GILncSig was validated in the TCGA data set and two GEO data sets profiled by different technology platforms, more independent data sets were needed to validate the GILncSig to ensure its robustness and reproducibility. In addition, the GILncSig was identified using our computational frame based

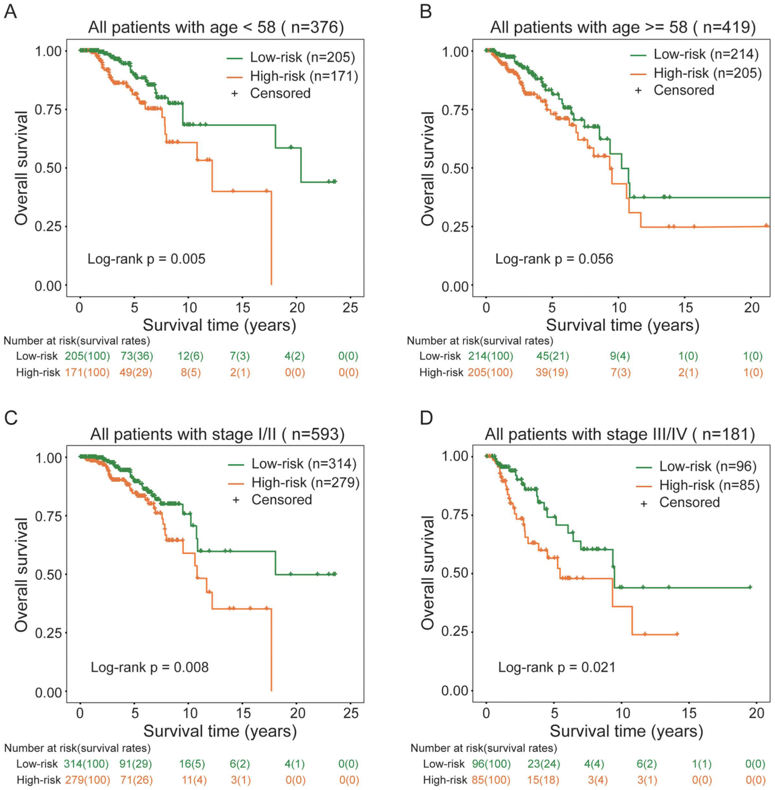

Figure 7. Stratification analyses by age and stage. Kaplan–Meier curve analysis of overall survival in high- and low-risk groups for young patients (**A**) and old patients (**B**). Kaplan–Meier curve analysis of overall survival in high- and low-risk groups for early-stage patients (**C**) and late-stage patients (**D**). Statistical analysis was performed using the log-rank test and univariate Cox analysis

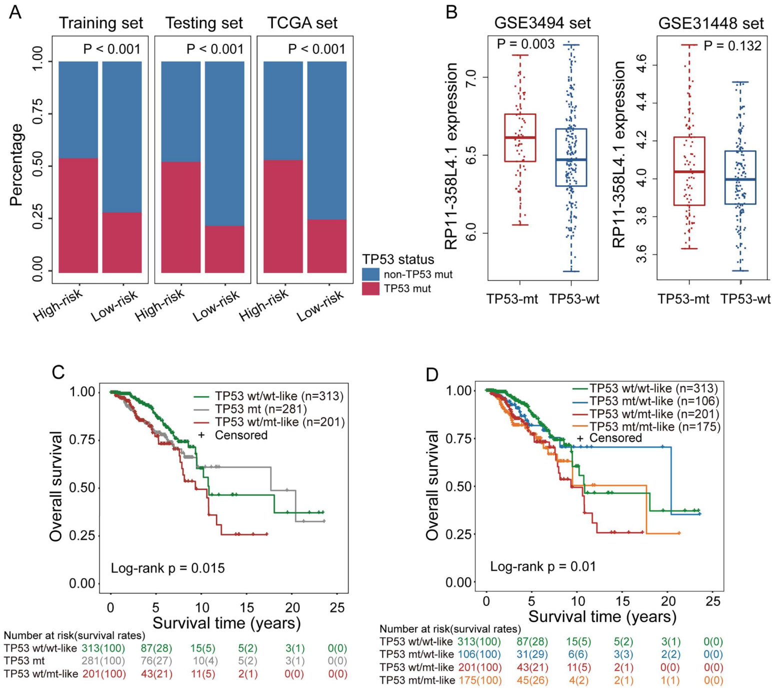

Figure 8. Relationship between the GILncSig and TP53 somatic mutation. (**A**) The proportion of TP53 mutation in high- and low-risk groups in the training set, testing set and the TCGA set. (**B**) Boxplots for *RP11-358L4.1* expression between patients with and without TP53 mutation in the GSE3494 and GSE31448 data sets. (**C**) Kaplan–Meier curve analysis of overall survival in TP53 wt/wt-like and TP53 wt/mt-like group according to the GILncSig and TP53-mt group, only TP53 wild-type patients were examined. (**D**) Kaplan–Meier curve analysis of overall survival is shown for patients classified according to TP53 mutation status and the GILncSig. Statistical analysis was performed using the log-rank test. TP53-wt: TP53-sequence wild type; TP53-mt: TP53-sequence mutation type.

on mutator hypothesis; therefore, further functional studies are required by experimental biologists to understand the regulatory mechanisms of the GILncSig in maintaining genome instability.

Conclusion

This work proposed a mutator hypothesis-derived computational frame to identify genome instability-associated lncRNAs, which provides a critical approach and resource for further studies examining the role of lncRNAs in genome instability. By combining lncRNA expression profiles with somatic mutation profiles and clinical information of breast cancers as a case study, we identified a genome instability-derived lncRNA signature as an independent prognostic marker to stratify risk subgroups for patients with breast cancer, which was successfully validated on the independent patient cohort. Through further prospective validation, the GILncSig may have important implications for genome instability and customized decision-making in breast cancer patients.

Authors' contributions

M.Z. and J.S. conceived and designed the experiments. S.B., H.Z., J.Y., D.F. and Z.Z. analyzed data. M.Z. and S.B. wrote this manuscript. All authors read and approved the final manuscript.

Key Points

- A mutator hypothesis-derived computational frame combining lncRNA expression profiles and somatic mutation profiles in a tumor genome was developed to identify genomic instability-associated lncRNAs.
- A genome instability-derived two lncRNA-based gene signature (GILncSig) was identified to stratify patients into high- and low-risk groups with significantly different outcome and was validated in multiple independent patient cohorts.
- The GILncSig correlated with genomic mutation rate in both ovarian cancer and breast cancer and can act as a measurement of the degree of genome instability.
- The GILncSig has greater prognostic significance than TP53 mutation status alone and is capable of identifying intermediate subtype group existing with partial TP53 functionality in TP53 wild-type patients.

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 61873193, 61871294). The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;**68**:394–424.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;**69**:7–34.
3. Wang Y, Zhang Y, Pan C, *et al*. Prediction of poor prognosis in breast cancer patients based on microRNA-21 expression: a meta-analysis. *PLoS One* 2015;**10**:e0118647.
4. Oakman C, Santarpia L, Di Leo A. Breast cancer assessment tools and optimizing adjuvant therapy. *Nat Rev Clin Oncol* 2010;**7**:725–32.
5. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability—an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 2010;**11**:220–8.
6. Ottini L, Falchetti M, Lupi R, *et al*. Patterns of genomic instability in gastric cancer: clinical implications and perspectives. *Ann Oncol* 2006;**17**(Suppl 7):vii97–102.
7. Suzuki K, Ohnami S, Tanabe C, *et al*. The genomic damage estimated by arbitrarily primed PCR DNA fingerprinting is useful for the prognosis of gastric cancer. *Gastroenterology* 2003;**125**:1330–40.
8. Tam AS, Sihota TS, Milbury KL, *et al*. Selective defects in gene expression control genome instability in yeast splicing mutants. *Mol Biol Cell* 2019;**30**:191–200.
9. Habermann JK, Doering J, Hautaniemi S, *et al*. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. *Int J Cancer* 2009;**124**:1552–64.
10. Wang T, Wang G, Zhang X, *et al*. The expression of miRNAs is associated with tumour genome instability and predicts the outcome of ovarian cancer patients treated with platinum agents. *Sci Rep* 2017;**7**: 14736.
11. Zeng X, Liu L, Lu L, *et al*. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 2018;**34**:2425–32.
12. Zhang X, Zou Q, Rodriguez-Paton A, *et al*. Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:283–91.
13. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* 2015;**22**:5–7.
14. Koziol MJ, Rinn JL. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* 2010;**20**:142–8.
15. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* 2013;**20**:300–7.
16. Sanchez Calle A, Kawamura Y, Yamamoto Y, *et al*. Emerging roles of long non-coding RNA in cancer. *Cancer Sci* 2018;**109**:2093–100.
17. Zhang Y, Shields T, Crenshaw T, *et al*. Imprinting of human H19: allele-specific CpG methylation, loss of the active allele in Wilms tumor, and potential for somatic allele switching. *Am J Hum Genet* 1993;**53**:113–24.
18. Luo JH, Ren B, Keryanov S, *et al*. Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology* 2006;**44**:1012–24.
19. de Kok JB, Verhaegh GW, Roelofs RW, *et al*. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res* 2002;**62**:2695–8.
20. Bartonicek N, Maag JL, Dinger ME. Long noncoding RNAs in cancer: mechanisms of action and technological advancements. *Mol Cancer* 2016;**15**:43.
21. Huarte M. The emerging role of lncRNAs in cancer. *Nat Med* 2015;**21**:1253–61.
22. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. *Cancer Discov* 2011;**1**:391–407.
23. Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. *Cancer Cell* 2016;**29**:452–63.
24. Liu H. Linking lncRNA to genomic stability. *Sci China Life Sci* 2016;**59**:328–9.

25. D'Alessandro G. *di Fagagna FdAJN-cRI*. NCRI: Long non-coding RNA in the control of genome stability and cancer phenotypes, 2018, 2.
26. Munschauer M, Nguyen CT, Sirokman K, *et al*. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* 2018;**561**:132–6.
27. Betts JA, Moradi Marjaneh M, Al-Ejeh F, *et al*. Long noncoding RNAs CUPID1 and CUPID2 mediate breast cancer risk at 11q13 by modulating the response to DNA damage. *Am J Hum Genet* 2017;**101**:255–66.
28. Polo SE, Blackford AN, Chapman JR, *et al*. Regulation of DNA-end resection by hnRNPU-like proteins promotes DNA double-strand break signaling and repair. *Mol Cell* 2012;**45**:505–16.
29. Sharma V, Khurana S, Kubben N, *et al*. A BRCA1-interacting lncRNA regulates homologous recombination. *EMBO Rep* 2015;**16**:1520–34.
30. Li J, Han L, Roebuck P, *et al*. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res* 2015;**75**:3728–37.
31. Miller LD, Smeds J, George J, *et al*. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* 2005;**102**:13550–5.
32. Sabatier R, Finetti P, Adelaide J, *et al*. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PLoS One* 2011;**6**:e27656.
33. Yu G, Wang LG, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7.
34. Li J, Wang W, Xia P, *et al*. Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in patients with breast cancer. *Int J Cancer* 2018;**143**: 2150–60.
35. Sun J, Chen X, Wang Z, *et al*. A potential prognostic long noncoding RNA signature to predict metastasis-free survival of breast cancer patients. *Sci Rep* 2015;**5**: 16553.
36. Huszno J, Grzybowska E. TP53 mutations and SNPs as prognostic and predictive factors in patients with breast cancer. *Oncol Lett* 2018;**16**:34–40.
37. Borresen-Dale AL. TP53 and breast cancer. *Hum Mutat* 2003;**21**:292–300.
38. Walerych D, Napoli M, Collavin L, *et al*. The rebel angel: mutant p53 as the driving oncogene in breast cancer. *Carcinogenesis* 2012;**33**:2007–17.
39. Gasco M, Shami S, Crook T. The p53 pathway in breast cancer. *Breast Cancer Res* 2002;**4**:70–6.
40. Pharoah PD, Day NE, Caldas C. Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *Br J Cancer* 1999;**80**:1968–73.
41. Jin Y, Desta Z, Stearns V, *et al*. CYP2D6 genotype, antidepressant use, and tamoxifen metabolism during adjuvant breast cancer treatment. *J Natl Cancer Inst* 2005;**97**:30–9.
42. Russo J, Russo IH. The role of estrogen in the initiation of breast cancer. *J Steroid Biochem Mol Biol* 2006;**102**: 89–96.
43. Spiegel D, Bloom JR, Kraemer HC, *et al*. Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *Lancet* 1989;**2**:888–91.
44. Wagner KU, Rui H. Jak2/Stat5 signaling in mammogenesis, breast cancer initiation and progression. *J Mammary Gland Biol Neoplasia* 2008;**13**:93–103.
45. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991;**19**:403–10.
46. Balslev I, Axelsson CK, Zedeler K, *et al*. The Nottingham prognostic index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). *Breast Cancer Res Treat* 1994;**32**:281–90.
47. Sundquist M, Thorstenson S, Brudin L, *et al*. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. South East Swedish Breast Cancer Study Group. *Breast Cancer Res Treat* 1999;**53**:1–8.
48. D'Eredita G, Giardina C, Martellotta M, *et al*. Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution. *Eur J Cancer* 2001;**37**:591–6.
49. Polyak K. Heterogeneity in breast cancer. *J Clin Invest* 2011;**121**:3786–8.
50. Bartkova J, Horejsi Z, Koed K, *et al*. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* 2005;**434**:864–70.
51. Gorgoulis VG, Vassiliou LV, Karakaidos P, *et al*. Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* 2005;**434**:907–13.
52. Kronenwett U, Ploner A, Zetterberg A, *et al*. Genomic instability and prognosis in breast carcinomas. *Cancer Epidemiol Biomarkers Prev* 2006;**15**:1630–5.
53. Mettu RK, Wan YW, Habermann JK, *et al*. A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types. *Int J Biol Markers* 2010;**25**:219–28.
54. Ferguson LR, Chen H, Collins AR, *et al*. Genomic instability in human cancer: molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. *Semin Cancer Biol* 2015;**35**:S5–S24.
55. Zhang S, Yuan Y, Hao D. A genomic instability score in discriminating nonequivalent outcomes of BRCA1/2 mutations and in predicting outcomes of ovarian cancer treated with platinum-based chemotherapy. *PLoS One* 2014;**9**:e113169.
56. Gupta RA, Shah N, Wang KC, *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;**464**:1071–6.
57. Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer? *Hum Mol Genet* 2010;**19**:R152–61.
58. Prensner JR, Iyer MK, Balbin OA, *et al*. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 2011;**29**:742–9.
59. Munschauer M, Nguyen CT, Sirokman K, *et al*. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* 2018;**561**:132–6.
60. Hu WL, Jin L, Xu A, *et al*. GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability. *Nat Cell Biol* 2018;**20**:492–502.
61. Barnum KJ, O'Connell MJ. Cell cycle regulation by checkpoints. *Methods Mol Biol* 2014;**1170**:29–40.
62. Weinert T, Lydall D. Cell cycle checkpoints, genetic instability and cancer. *Semin Cancer Biol* 1993;**4**:129–40.
63. Wenzel ES, Singh ATK. Cell-cycle checkpoints and aneuploidy on the path to cancer. *In Vivo* 2018;**32**:1–5.
64. Friedberg EC. How nucleotide excision repair protects against cancer. *Nat Rev Cancer* 2001;**1**:22–33.
65. Chakraborty U, Dinh TA, Alani E. Genomic instability promoted by overexpression of mismatch repair factors in yeast: a model for understanding cancer progression. *Genetics* 2018;**209**:439–56.

66. Garfinkel DJ, Bailis AM. Nucleotide excision repair, genome stability, and human disease: new insight from model systems. *J Biomed Biotechnol* 2002;**2**:55–60.
67. Harfe BD, Jinks-Robertson S. DNA mismatch repair and genetic instability. *Annu Rev Genet* 2000;**34**:359–99.
68. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell* 2013;**152**:1237–51.
69. Moldovan GL, D'Andrea AD. How the fanconi anemia pathway guards the genome. *Annu Rev Genet* 2009;**43**:223–49.
70. Palovcak A, Liu W, Yuan F, *et al*. Maintenance of genome stability by Fanconi anemia proteins. *Cell Biosci* 2017;**7**:8.
71. Tang W, Wan S, Yang Z, *et al*. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
72. Cai Q, Zhang B, Sung H, *et al*. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet* 2014;**46**: 886–90.
73. Maggiolini FAM, Cantsilieris S, D'Addabbo P, *et al*. Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLoS Genet* 2019;**15**: e1008075.