Genes & Diseases (2023) 10, 228–238

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

FULL LENGTH ARTICLE

Etiological roles of core promoter variation in triple-negative breast cancer

Teng Huang, Jiaheng Li, San Ming Wang*

Cancer Centre and Institute of Translational Medicine, Faculty of Health Sciences, Ministry of Education Frontiers Science Center for Precision Oncology, University of Macau, Taipa, Macao SAR 999078, China

Received 14 July 2021; received in revised form 26 December 2021; accepted 12 January 2022
Available online 24 February 2022

KEYWORDS
Core promoter;
RNA-seq;
Triple-negative breast cancer;
Variation;
Whole exome sequencing

Abstract Abnormal gene expression plays key role in cancer development. A core promoter is located around the transcriptional start site. Through interaction between core promoter sequences and transcriptional factors, core promoter controls transcriptional initiation. We hypothesized that in cancer, core promoter sequences could be mutated to interfere the interaction with transcriptional factors, resulting in altered transcriptional initiation and abnormal gene expression and cancer development. We used triple-negative breast cancer (TNBC) as a model to test our hypothesis. We collected genome-wide core promoter variants from 279 TNBC genomes. After extensive filtering of normal genomic polymorphism, we identified 19,427 recurrent somatic variants in 1,238 core promoters of 1,274 genes and 1,694 recurrent germline variants in 272 core promoters of 294 genes. Many of the affected genes were oncogenes and tumor suppressors. Analysis of RNA-seq data from the same patient cohort identified increased or decreased gene expression in 439 somatic and 85 germline variants-affected genes, and the results were validated by luciferase reporter assay. By comparing with the core promoter variation data from 610 unclassified breast cancer, we observed that core promoter variants in TNBC were highly TNBC-specific. We further identified the drugs targeting the genes with core promoter variation. Our study demonstrates that core promoter is highly mutable in cancer, and can play etiological roles in TNBC and other types of cancer through influencing transcriptional initiation.

Abbreviations: CDX, Chinese Dai in Xishuangbanna; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; EAS, East Asian; ER, estrogen receptor; EVDC, Exome-based variant detection in core-promoters; GO, Gene Ontology; HER2, human epidermal growth factor receptor 2; JPT, Japanese in Tokyo; KHV, Kinhin Ho Chi Minh City; SRA, Sequence read archive; TSS, transcriptional start site; TNBC, triple-negative breast cancer; PR, progesterone receptor; WES, Whole exome sequencing.

* Corresponding author. Faculty of Health Sciences, University of Macau, Taipa, Macao SAR 999078, China.
E-mail address: sanmingwang@um.edu.mo (S.M. Wang).
Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2022.01.003

Introduction

Gene expression is under tight regulation through precise interaction between cis-sequences and trans-transcriptional factors in the regulatory region. Core promoter locates around transcriptional start site (TSS) with multiple conserved cis-motifs.[1,2] The interaction between the cis-sequences and the trans-factors of RNA polymerase II, TFIIB and TFIID etc. in core promoter forms the transcriptional initiation complex to regulate transcriptional initiation.[3–5] Variation in core promoter sequences can interfere with the precise cis–trans interaction therefore the proper organization of the transcriptional initiation complex, causing altered transcription initiation and pathological consequences.[6–8] The core promoter variation in telomerase reverse transcriptase (*TERT*) in melanoma is a typical example. In *TERT* core promoter, each of the two somatic mutations at −72 (C > T), −50 (C > T) (TSS as +1) creates a new Ets binding site, causing enhanced Ets binding and increased *TERT* expression, contributing to telomere maintenance and melanoma.[9,10] After decades of study, however, core promoter variation in *TERT* remains as a few exceptional cases in connection of core promoter variation to cancer etiology. Despite rich biological knowledge of the important roles of core promoters in controlling transcription initiation, it remains largely elusive whether core promoter variation plays etiological roles in cancer; and despite the rapid progress of cancer genome studies, core promoter region in cancer has not been systematically characterized so far. There is no comprehensive data to show whether core promoter is mutable as in other parts of the cancer genomes. Abnormal gene expression in cancer has been traditionally studied through measuring the abundance of the processed mRNA, which is the processed product far downstream from transcriptional initiation, rather than measuring the nascent transcripts coming immediately after transcription initiation.

Breast cancer is the most common cancer in women. Approximately 10–20% of breast cancer are triple-negative breast cancer (TNBC), as characterized by the absence of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2).[11–13] TNBC is also enriched with *BRCA1* and *BRCA2* mutation-associated breast cancers.[14] Compared to other types of breast cancer, TNBC has the features of early onset, highly aggressive, metastatic and recurrent, with poor prognosis, and lack of specific drug targets. TNBC has its unique gene expression signature differentiating it from other types of breast cancer, highlighting that abnormal gene expression plays critical roles in TNBC etiology.[12,15,16] However, it remains largely unclear what causes the abnormal gene expression in TNBC.

We hypothesized that core promoter in TNBC could be highly mutable in contributing to its abnormal gene expression. In this study, we used the Exome-based Variant Detection in Core-promoters (EVDC) method to identify somatic and germline variants in the core promoter region in 279 TNBC genomes.[17] We systematically characterized the variants, the affected genes, their impact on gene expression, and their TNBC-specificity. Data from our study revealed the high prevalence of core promoter variation in TNBC and highlights that core promoter variation can play more important roles in cancer etiology than currently known.

Materials and methods

Data sources

Whole exome sequencing (WES) data from TNBC patients (n = 279) and RNA-seq data (n = 360) and paired normal tissues (n = 88) from the same TNBC study[13] were from Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra, SRP157974). WES data from the unclassified breast cancer (n = 610) were from SRA data-base (https://www.ncbi.nlm.nih.gov/sra, SRP218807).[18]

Identification of core promoter variants

The EVDC method was used to collect core promoter sequences from exome sequences.[17] Briefly, exome data were converted to FASTQ format using SRA Toolkit. BWA was used to map exome sequences to the human genome reference sequences (hg19). The resulting SAM files were converted into BAM files and sorted using SAMtools. Duplicates were removed and read group information was added using Picard. The BAM files were further processed using Genome Analysis Toolkit for variant calling. The called variants were annotated using ANNOVAR Toolkit for gene-based and filter-based annotations. The filter-based annotation was used to distinguish known or novel variants and to search allele frequencies. Variants matched in databases of dbSNP, 1000 Genome, ESP, ExAC, gnomAD, and ClinVar were considered as germline variants. Polymorphic germline variants were further filtered using multiple Chinese and non-Chinese-derived genome data sets including the PGG.Han project,[19] the ChinaMap project,[20] the 1000 Genome Project East Asian (EAS) data[21] and other resources[18,22–27] (Table S1). After filtering, the variants present in >140 cases (50%, somatic) and 14 cases (5%, germline) of the 279 TNBC cases were classified as normal polymorphic variants and eliminated, the variants present only in single cases were considered as private variants and also eliminated. The remaining variants were used for further analysis. Plots showing core promoter variants were generated using R trackViewer package.[28] Variation data in breast cancer were downloaded from Genomic Data Commons Data Portal of TCGA.[29] LiftOver tool[30] was used to convert variant position from hg38 to hg19.

Differential gene expression analysis

RNA-seq data from cancer and non-cancer samples were used for the analysis. Differential gene expression between cancer and non-cancer samples was determined by using HISAT2 (version 2.2.0), SAMtools (version 1.9), StringTie (version 2.1.1) and DESeq2 (version 1.26.0) following the instructions in each program. Volcano plots showing differential expression were generated using R ggplot2 package.[31]

Luciferase reporter assay

Human embryonic kidney 293 cells (293 cells) were grown in Dulbecco's modified Eagle's media/Nutrient Mixture culture medium with 10% fetal bovine serum, 100 IU/ml penicillin and 100 IU/ml streptomycin sulfate. The core promotor sequences containing the wild-type and mutated bases were synthesized and cloned into KpnI/MluI-digested pGL3 luciferase reporter vector (Table S5). Each clone was validated by Sanger sequencing. Fifty μg of pGL3 containing corresponding core promoter sequences and 5 μg of control pRL *Renilla* luciferase reporter vector were mixed, and co-transfected into the 293 cells using Lipofectamine 3000 (Thermo Fisher SCIENTIFIC, MS, USA). Forth-eight hours after the transfection, the cells were harvested to measure luciferase activity using the Dual-Luciferase Reporter Assay System following the instruction (Promega, WI, USA). Normalized luciferase activity was calculated as the ratio of *firefly* luciferase activity to *Renilla* activity.

$$E_l = E_f / E_r$$

where E_f is the activity of the *firefly* luciferases, E_r is the activity of the *Renilla* luciferases, and E_l is the luciferase activity.

Functional analysis

For the genes with core promoter variation, their functional categories and pathways were analyzed using Gene Ontology (GO) knowledgebase[32] and GeneCards.[33] The drugs targeting the genes with core promoter variants and altered expression were identified from DrugBank,[34] CMap,[35] LINCS[36] and GEO.[37] GO terms and DrugBank drugs were identified using Metascape.[38] Drugs from CMap, LINCS and GEO were identified using Drug Gene Budger.[39]

Statistics analysis

Student's *t*-test was used to compare the variation types between cancer and non-cancer from the 1000 Genome Project, and *P*-value < 0.05 was considered as significant. Adjusted *P*-value < 0.05 by using Benjamini-Hochberg procedure and fold changes ≥ 1.5 by using DESeq2 were considered as significantly difference for the differentially expressed genes. *P*-value < 0.05 by using the hypergeometric test and overlap ≥ 3 by using the Metascape were considered as significant in enrichment analysis. In dual-luciferase reporter assay, Student's *t*-test was used to test the statistically significant between mutant and wild-type, *P*-value < 0.05 and fold changes ≥ 1.5 were considered as significantly difference.

Results

Core promoter variation in TNBC genome

Figure 1 outlines the analytic process of our study. We collected the core promoter sequences from 279 TNBC patients.[13] We called variants in the collected sequences, and removed polymorphic variants by filtering through genome sequences derived from Chinese and non-Chinese populations (see methods for details).

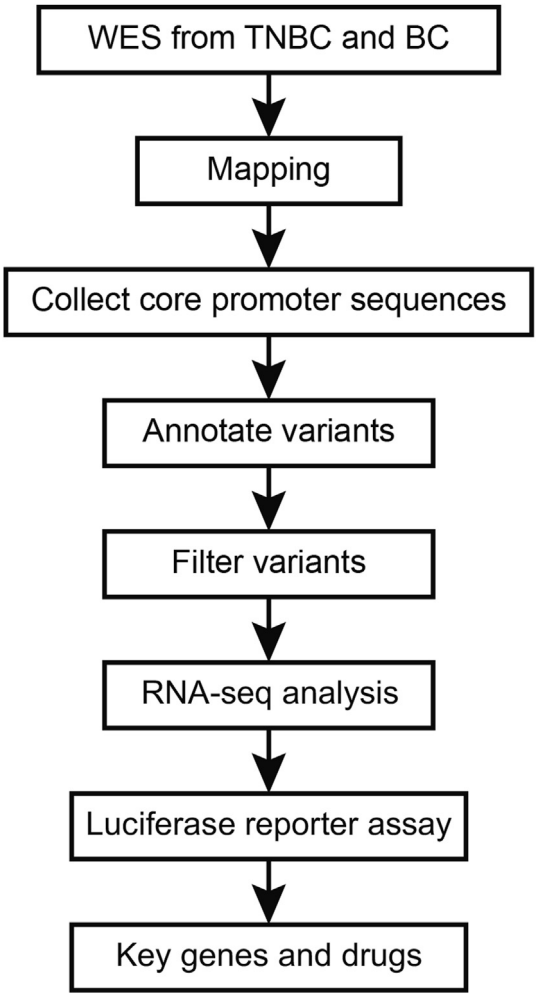

Figure 1 Scheme of the analytic process. WES, whole exome sequences; TNBC, triple-negative breast cancer; BC, unclassified breast cancer.

We identified a total of 19,427 recurrent somatic variants (present in ≥ 2 carriers, 70 variants per TNBC case on average), composed of 1,940 distinct somatic variants in 1,238 core promoters of 1,274 genes (Fig. 2A and Table 1A; Table S2A, S3A, B). Of the 1,940 somatic variants, the most frequent type was substitution (55.3%), and the rests were insertion (20.9%) and deletion (23.8%); 99.2% were absent in the COSMIC database, 34.2% were located at simple repetitive sequences or microsatellites in the core promoters of 202 genes (Table S3A, B and Fig. 2B). We also identified 1,694 recurrent germline variants (present in ≥ 2 carriers, 6 variants per TNBC case on average), composed of 496 distinct variants in 272 core promoters of 294 genes (Fig. 2C and Table 1A; TableS2B, S3C, D). Of the variants identified, the most frequent type was insertion (46.2%), and the rests were substitution (23.0%) and deletion (30.8%). Of the 496 variants, 37.1% were located at simple repetitive sequences or microsatellites in the core promoters of 75 genes (Table S3C, D and Fig. 2D). There was no sharing of the same position between somatic and germline variants, but there

Table 1 Summary of variants identified in TNBC core promoters.

Items	Core promoter variants	
	Somatic	Germline
A. General features		
Total	19,427	1694
Distinct	1940	496
Co-promoter with variants	1238	272
Absent in COSMIC database	1925	496
Non-repetitive	1276	312
Repetitive	664	184
Type		
Substitution	1073	114
Insertion	406	229
Deletion	461	153
Gene affected	1274	294
Average number of mutation/case	70	6
B. Variation frequency in motifs		
Total	2339	693
MTE_box2	731	306
DCE_box3	317	47
BREd	271	50
Inr	228	56
DPE	198	57
DTIE	140	37
TCT	96	30
BREu	82	14
MTE_box1	72	7
Ets	58	38
DCE_box2	40	12
E-Box	40	13
DCE_box1	35	14
SP1	12	0
XCPE1	10	1
TCF	5	0
XCPE2	3	0
TATA box	1	11
C. Transition, transversion and Ts/Tv ratio		
Transition		
G > A	144	16
C > T	120	21
A > G	35	6
T > C	29	8
Total	328	51
Transversion		
C > A	235	11
G > C	49	10
C > G	47	6
A > C	12	2
G > T	189	15
T > G	19	5
A > T	15	8
T > A	29	6
Total	595	63
Ts/Tv ratio[a]	1.10	1.62

[a] Ts/Tv ratio was calculated by 2xTs/Tv.

were 24 non-repetitive and 14 repetitive variant-containing genes shared by both somatic and germline groups. The total number of genes with somatic and germline core promoter variants reached to 1,530, over 7% of which were in the human genome.

The core promoter variants were highly enriched at core promoter motifs (Table 1B). For example, there were 731 somatic variants and 306 germline variants located at the MTE box2 motif. Consistent with its highly stable nature,[17] only 1 somatic and 11 germline variants were located at the TATA box. Lack of variants in TATA box served as a valuable internal control for the reliability of the core promoter variants identified in the study.

We calculated Ts/Tv ratio for the identified germline core promoter variants. Of the 114 single-base germline substitutions, the Ts/Tv ratio was 1.62 (Table 1C). This rate was significantly lower than the 3.25—3.81 [CHB (Han Chinese in Bejing): 3.65; CHS (Southern Han Chinese): 3.28; CDX (Chinese Dai in Xishuangbanna): 3.25; JPT (Japanese in Tokyo): 3.81; KHV (Kinh in Ho Chi Minh City): 3.37] in the core promoter region (Student's *t*-test: *P*-value = 7.32e-05) in EAS (East Asian) population from the 1000 Genome data.[40]

Effects of core promoter variation on gene expression

To test if core promoter variation can alter gene expression, we compared the expression of the genes with core-promoter variation in TNBC to the non-cancer control using the RNA-seq data from the same groups. Of the 1,274 genes with somatic core promoter variants in TNBC, 439 (34.5%) had altered expression including 261 increased and 178 decreased expression, with *PRDM13* as the highest of 67.0-fold increased expression and *LINC00445* as the highest of 10.8-fold decreased expression (Fig. 3A, B and Table S4A); of the 294 genes with germline core promoter variation in TNBC, 85 (28.9%) had altered expression including 37 with increased and 48 with decreased expression, with *LINC00221* as the highest of 10.0-fold increased expression and *ANKRD20A12P* as the highest of 21.9-fold decreased expression (Fig. 3C, D and Table S4B).

We used the dual-luciferase reporter assay to validate the altered gene expression caused by core promoter variation. Based on the functional importance of the genes carrying the variation, significance of the altered expression level by expression data analysis, and core promoter sequence features for designing the mutant and constructs, we selected core promoters in 10 genes (*BRCA2*, *FANCB*, *PRDM13*, *SLIT2*, *MAGEC2*, *HOXB13*, *MMP10*, *LINC00445*, *CYP4F22* and *GPM6A*) for the test. BRCA2 and FANCB are known in maintaining genome stability; SLIT2 suppresses tumor growth and metastasis through participating in "negative regulation of cell growth" and "negative regulation of cell migration" pathways; HOXB13 is a transcription factor associated with cancer; GPM6A involves in cellular differentiation and migration; *PRDM13* had the highest altered expression of 67-fold and *MAGEC2* was the second; *LINC00445* had highly decreased expression level;

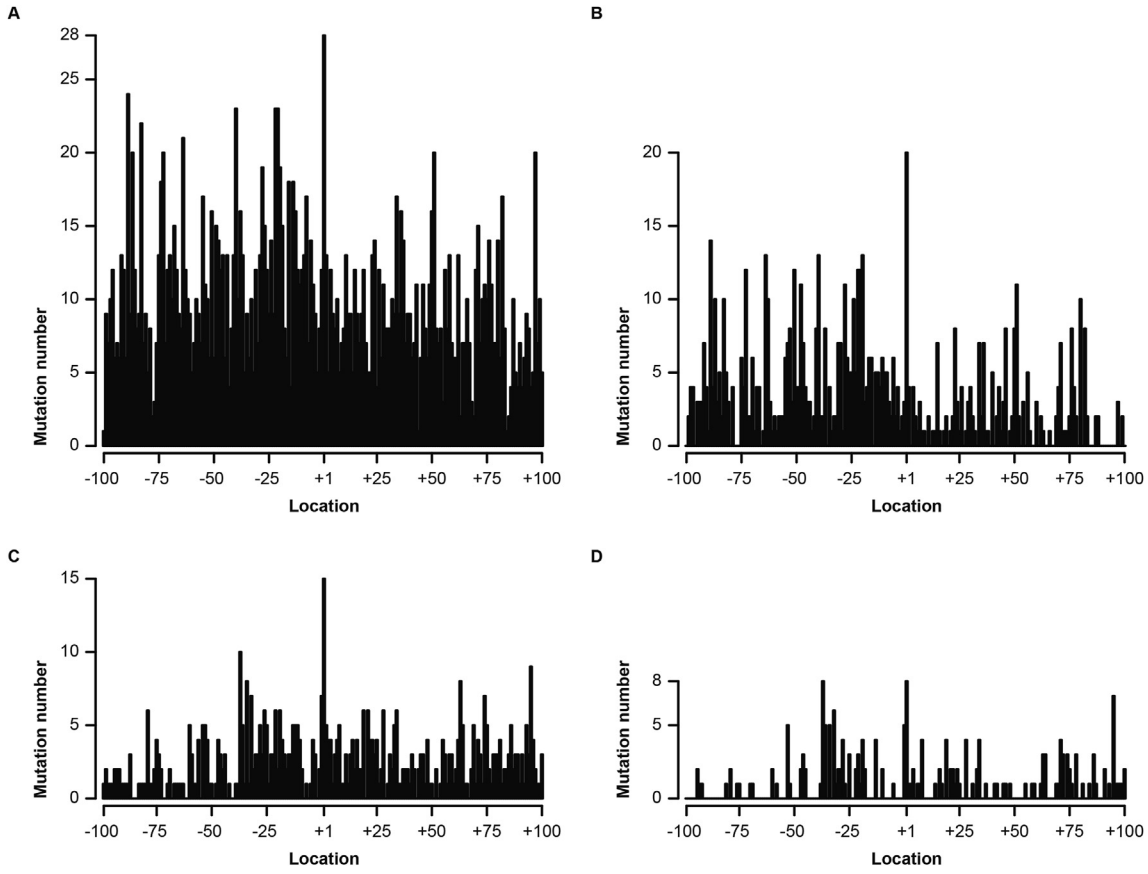

Figure 2 Variant distribution in TNBC core promoter region. **(A)** Distribution of somatic variants in core promoter region. **(B)** Distribution of somatic variants in simple repetitive sequence of core promoter region. **(C)** Distribution of germline variants in core promoter region. **(D)** Distribution of germline variants in simple repetitive sequence of core promoter region.

CYP4F22 had highly decreased expression participating in drug metabolism; *MMP10* was drug target. We generated the mutated core promoters containing the variants identified in TNBC, and tested their expression by referring to the corresponding wild-type core promoters. Of the 10 mutated core promoters, 7 (70%) had significantly affected luciferase activities (*BRCA2*, *FANCB*, *SLIT2*, *MMP10*, *MAGEC2*, *GMP6A* and *LINC00445*, $P < 0.05$), of which *BRCA2*, *FANCB*, *SLIT2* and *MAGEC2* had increased luciferase activities, *MMP10*, *GMP6A* and *LINC00445* had decreased luciferase activities (Fig. 3E and Table S5). Of the 7 genes tested, 5 were consistent with the results from RNA-seq data analysis: RNA-seq data showed that *BRCA2*, *FANCB* and *MAGEC2* had 3.4-fold, 3.1-fold, and 48.7-fold increased expression respectively, whereas luciferase reporter assay showed 3.7-fold, 7.7-fold, and 13.3-fold increased expression respectively; RNA-seq data showed that *GPM6A* and *LINC00445* had 5.8-fold and 10.8-fold decreased expression, whereas luciferase reporter assay showed 8.9-fold and 7.0-fold decreased expression, respectively.

Core promoter variation in cancer-related genes

Of the genes with core promoter variation, many are classical oncogenes or tumor suppressors (Table 2A; Table S2, S6). For example, PRDM13 is a histone methyltrans-ferase and negative regulator of RNA polymerase II. A C > A somatic variation at +58 in 2 TNBC cases generated a new Inr motif, causing a 67-fold increased expression in TNBC over the control, which was the highest among the 261 increasingly expressed genes with core promoter variation (Fig. 4A); MAGEC2 enhances ubiquitination and involves in liver cancer development. An A > G somatic variation at −47 in 2 TNBC cases generated a new Inr motif, causing a 48.7-fold increased expression; BRCA2 plays key roles in double-strand break repair through homologous recombination. A G > T somatic variation at +34 in 2 TNBC cases (BioSample accession number: SAMN09838023 and SAMN09838068) disrupted the MTE_box2 motif and created a new putative DPE motif (Fig. 4B), leading to a 3.4-fold increased expression in TNBC (Table S4A). For example, a 2.9-fold change of expression (primary breast cancer: paired normal breast tissue = 1517.8: 526.4) was observed in SAMN0988023; FANCB involves in DNA damage repair in Fanconi anemia pathway. A GG > TT somatic substitution at −67 to −68 was present in 51 TNBC cases. This variant created a new Ets motif, causing a 3.1-fold increased expression in TNBC; SLIT2 suppresses tumor growth and metastasis through participating in "negative regulation of cell growth" and "negative regulation of cell migration" pathways. A GA > CC somatic variation at −89 to −88 in *SLIT2* core promoter was present in 55 TNBC cases. This variant deleted an existing MTE_box2 and a DCE_box3 but

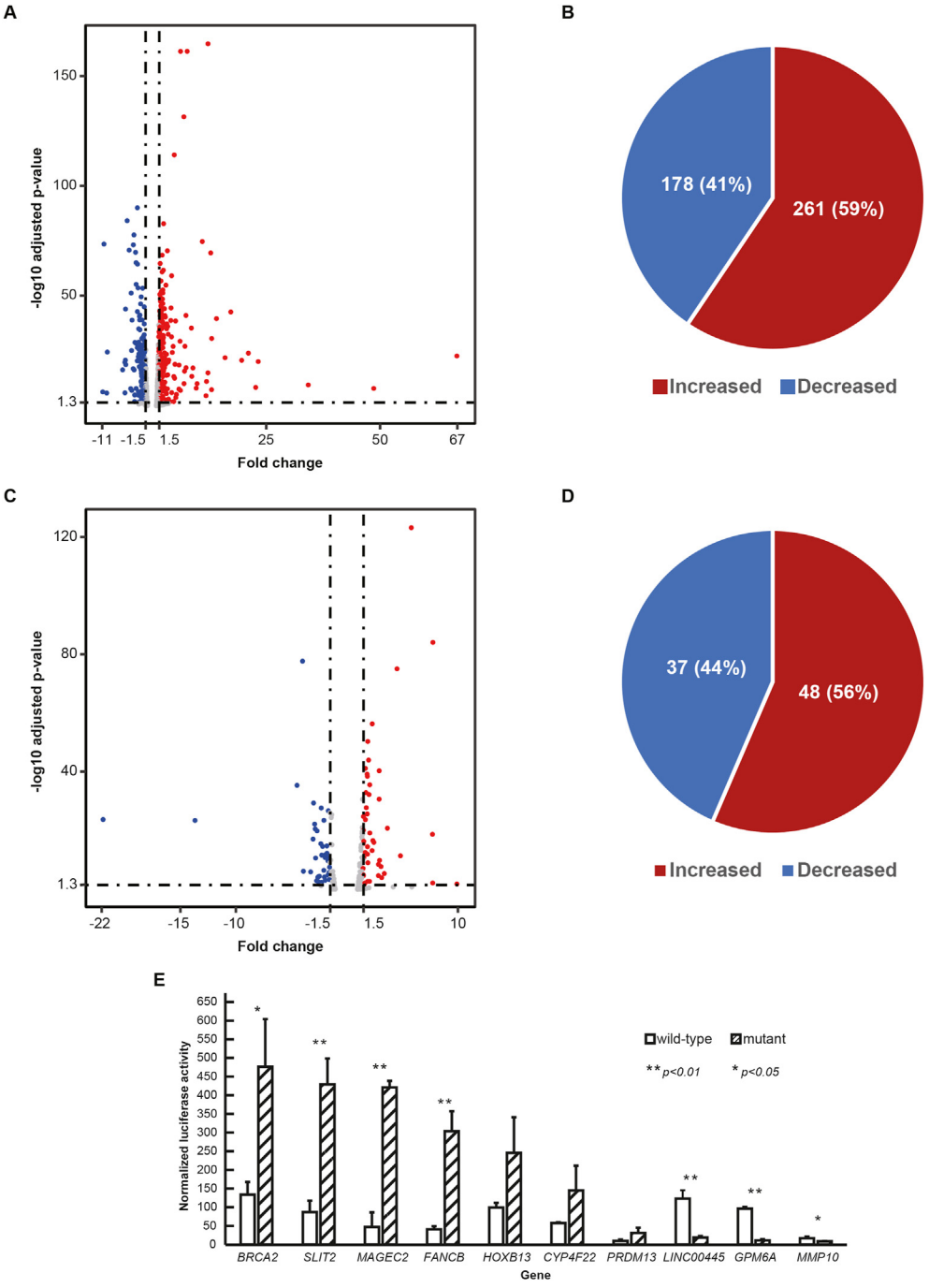

Figure 3 Core promoter variant-containing genes with altered gene expression in TNBC. The volcano plot showed the differential expression of genes with core promoter variation in TNBC and non-TNBC based on RNA-seq data. X-axis represents fold changes of increased or decreased expression, and Y-axis represents distribution of the genes with altered expression at -log10 scale (adjusted *P*-value). Red dots: increasingly expressed genes with statistically significant. Blue dots: decreasingly expressed genes with statistically significant. The pie charts display the number of differential expression genes. **(A)** Differential expression of the genes with somatic core promoter variation. **(B)** Proportion of genes of somatic core promoter variation with altered gene expression. **(C)** Differential expression of genes with germline core promoter variation. **(D)** Proportion of genes of germline core promoter variation with altered gene expression. **(E)** Luciferase activity by reporter assay. Dual-Luciferase Reporter Assay was performed to verify the impact of core promoter variation on gene expression in the 10 selected genes. Luciferase activities were compared with these from their corresponding wild-type controls. Three independent tests were performed for each gene.

Table 2 Examples of functional important genes with core promoter variation.

Gene	Co-promoter	#Carrier	Fold change	Breast cancer gene panel	Type
A. Examples of cancer related genes					
PRDM13	+58	2	+67.0		somatic
MAGEC2	−47	2	+48.7		somatic
HOXB13	+66	2	+19.7		somatic
MMP10	+2	2	+17.3		somatic
MAFA	−99	7	+7.2		somatic
BRCA2	+34	2	+3.4		somatic
FANCB	−67	51	+3.1		somatic
PLAG1	−44	2	+2.0		somatic
NFKB2	−55	2	+1.5		somatic
CYP4F22	−40	2	−9.8		somatic
GPM6A	+32	2	−5.8		somatic
NRG1	+1	3	−4.5		somatic
SPRY2	+86	6	−3.9		somatic
VIT	+36	2	−3.9		somatic
SLIT2	−88	55	−3.5		somatic
DLC1	−52	2	−3.1		somatic
CX3CR1	−60	2	−2.3		somatic
JUN	−40	6	−1.9		somatic
HOXA11	−61	7	+7.8		germline
STYXL1	−90	2	+1.8		germline
B. Signature genes included in breast cancer gene panels					
CHI3L2	−21	50	+3.6	Hu306|Sorlie500	somatic
SAT1	−69	18	+2.0	Sorlie500	somatic
FAM110A	+70	4	+2.0	Sorlie500	somatic
LMO7	−9	2	+1.9	Sorlie500	somatic
CDH3	+74	2	+1.9	PAM50|Sorlie500	somatic
GNPNAT1	+6	2	+1.6	Sorlie500	somatic
BTG3	+99	2	+1.6	Sorlie500	somatic
AK2	−53	2	+1.6	Hu306	somatic
ACSL3	+5	8	+1.6	Hu306	somatic
NME4	−25	2	+1.5	Sorlie500	somatic
FZD6	−45	110	+1.5	Sorlie500	somatic
NRG1	+1	3	−4.5	Sorlie500	somatic
PLAT	+52	2	−2.5	Sorlie500	somatic
SALL2	+27	2	−2.3	Hu306	somatic
CX3CL1	+28	5	−2.0	Sorlie500	somatic
CITED2	−88	2	−1.7	Sorlie500	somatic
ITGB1BP1	−77	2	−1.5	Sorlie500	somatic
FGF18	+83	2	−1.5	MammaPrint	somatic
ACTR3B	+63	2	+1.6	PAM50	germline
SERINC1	+21	2	−1.6	Sorlie500	germline
C. Categories of core promoter variant-containing genes in tumorigenesis					
Response to growth factor					somatic|germline
Female sex differentiation					somatic|germline
Chemotaxis					somatic|germline
Blood vessel development					somatic|germline
Proximal promoter sequence-specific DNA binding					somatic
Transmembrane receptor protein tyrosine kinase signaling pathway					somatic
Signaling pathways regulating pluripotency of stem cells					somatic
Positive regulation of cell death					germline
Mesoderm development					germline

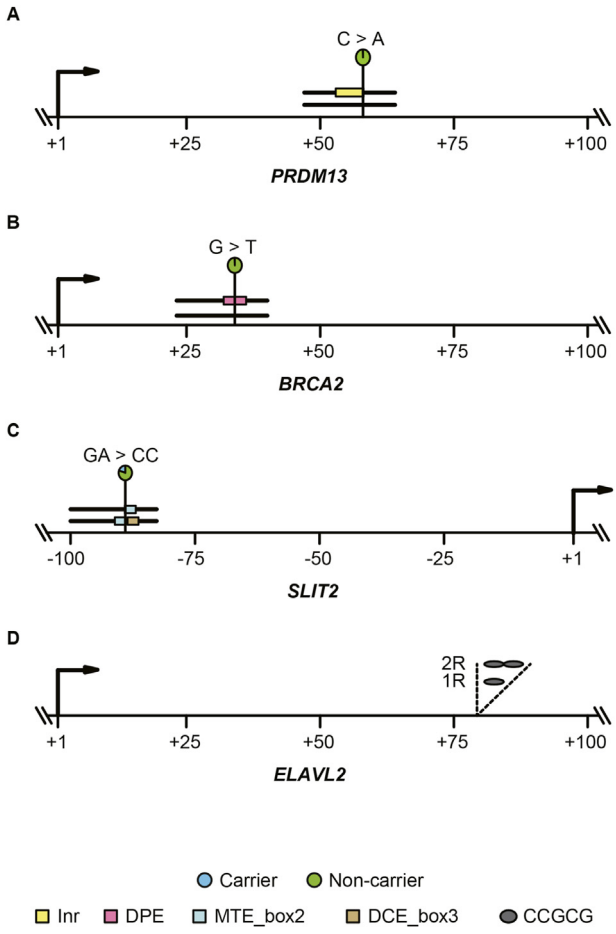

Figure 4 Examples of genes with core promoter variation. The lollipops represent the variations and the pies in lollipops show the proportion of variant carriers and non-carriers. The upper line represents the variant-altered sequences and the bottom line the reference sequences. The boxes in the lines represent the variant-affected core promoter motifs. **(A)** Core promoter variation in *PRDM13* in TNBC. A C > A variant at +58 created a new Inr motif in the core promoter of *PRDM13*. **(B)** Core promoter variation in *BRCA2* in TNBC. A G > T variant created a DPE motif at +34 in the core promoter of *BRCA2*. **(C)** Core promoter variant in *SLIT2* in TNBC. A GA > CC variant at −89 to −88 disrupted an MTE_box2 and a DCE_box3, and generated a new MTE_box2 in *SLIT2* core promoter. **(D)** Core promoter variant in *ELAVL2* in TNBC. A simple repetitive sequence (CCGCG) was inserted at +80 of the core promoter of *ELAVL2* in 6 TNBC cases, generated 1R (CCGCG) and 2R (CCGCGCCGCG) genotypes.

generated a new MTE_box2 motif, causing 3.5-fold decreased expression in TNBC (Fig. 4C).

It is well known that simple repetitive sequences such as Variable Number of Tandem Repeats are enriched in regulatory region contributing to gene expression regulation.[41–44] Many identified core promoter variants were located at the simple repetitive sequences, indicating that simple repetitive sequences were vulnerable for core promoter variation in TNBC (Table S3). For example, ELAVL2 is a 3′UTR binding protein involving in post-transcriptional regulation of gene expression. A "CCGCGCCGCG"-like simple repetitive sequence was inserted at its +80 in 6 TNBC cases, causing an 8.7-fold increased expression (Fig. 4D); BMF is a BCL2 family member involving in apoptotic regulation. A set of 7 "ACACACACAC"-like simple repeat were inserted at −83 of *BMF* core promoter in 14 TNBC cases, causing a 2.1-fold increased expression; TNFSF8 is a member of tumor necrosis factor ligand family. A "TGTGTGTGTG"-like simple repetitive sequence was deleted at −20, −22, and −26 in 20 TNBC carriers, causing a 2.3-fold increased expression.

Gene expression in TNBC has been extensively studied using gene panels specifically designed for breast cancer applications, such as the Mamma Print,[45] PAM50, Hu306, and Sorlie500.[46] We searched the core promoter variation in the genes included in these panels, and identified 20 genes with core promoter variation (Table 2B). The results indicate that core promoter variation can also contribute to the gene expression signature in TNBC and other types of breast cancer.

We performed functional annotation for the genes affected with core promoter variation and observed that the genes were enriched with the functional pathways highly relevant to cancer development, such as "Response to growth factors", "Chemotaxis", "Blood vessel development", "Proximal promoter sequence-specific DNA binding", "Transmembrane receptor protein tyrosine kinase signaling pathway", "Signaling pathways regulating pluripotency of stem cells", "Positive regulation of cell death" and "Mesoderm development" (Table 2C; Table S6).

TNBC-specificity of core promoter variation

To investigate the specificity of the core promoter variation in TNBC, we collected core promoter variation from unclassified breast cancer (n = 610) and compared the data with the core promoter variants from TNBC. Both TNBC and the unclassified breast cancer patients were from Shanghai region, therefore, shared the same ethnic genetic background and environment. In the 610 unclassified breast cancer cases, we identified 258 distinct recurrent germline variants in 151 core promoters of 172 genes. Similar to TNBC core promotors, the unclassified breast cancer had significantly lower Ts/Tv ratio of 1.71 than the non-cancer population (Student's t-test: P = 8.92e-05). Comparison of the core promoter variation and variation-affected genes between TNBC and the unclassified breast cancer groups showed that core-promoter variation in TNBC was highly TNBC-specific, and the core-promoter variation in the unclassified breast cancer was also highly unclassified breast cancer-specific (Fig. 5). Comparison of the core promoter variants in the simple repetitive sequences also showed the same feature between the two cancer groups (Fig. 5C).

Drugs targeting the genes with core promoter variation and altered expression

The genes with core promoter variation and altered expression provide potential drug targets for TNBC treatment. From the DrugBank, we respectively identified 271 and 111 existing drugs and compounds targeting 39

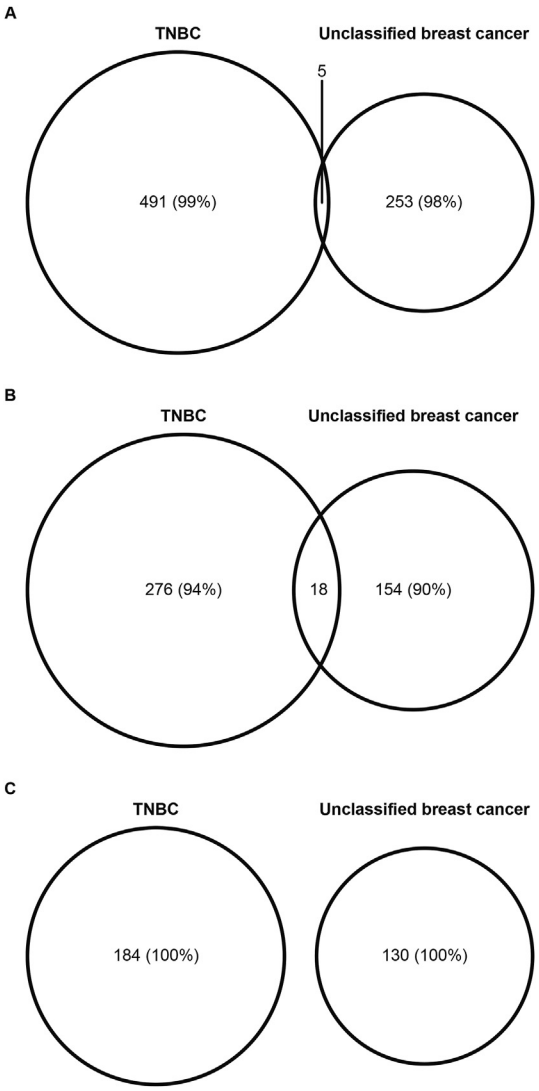

Figure 5 Comparison of the core promoter variation between TNBC and unclassified breast cancer. **(A)** Comparison of all core promoter variations. It showed that 99% of core promoter variation in TNBC were TNBC-specific. **(B)** Comparison of variant-affected genes. It showed that 94% of core promoter variant-affected genes were TNBC-specific. **(C)** Comparison of core promoter variant-affected simple repetitive sequences. It shows that all of the affected simple repetitive sequences in TNBC was TNBC-specific. The same features also applied for the unclassified breast cancer.

increasingly and 17 decreasingly expressed genes with somatic core promoter variation. We also respectively identified 34 and 14 existing drugs and compounds targeting 8 increasingly and 6 decreasingly expressed genes with germline core promoter variation (Table S7). We also respectively identified candidate agonists for decreasingly expressed genes and antagonists for increasingly expressed genes from CMap, LINCS and GEO. For example, FABP6 is involved in fatty acid uptake, transport, metabolism, and development of colorectal cancer. A C > A variant at −34 in 8 TNBC carriers caused a 23.3-fold increased expression. There are 3 FABP6-targeting drugs, Cholic Acid ($C_{24}H_{40}O_5$), N-Cholylglycine and Taurocholic Acid in DrugBank; MMP10 is involved in angiogenic and apoptotic pathways, and promotes cervical tumor progression. Two MMP10-targeting drugs, Marimastat ($C_{15}H_{29}N_3O_5$) and N-Isobutyl-N-[4-Methoxyphenylsulfonyl] Glycyl Hydroxamic Acid, were identified; KIF11 is involved in chromosome positioning, centrosome separation and bipolar spindle formation during mitosis. 11 drugs targeting KIF11 were present in DrugBank. For example, Monastrol ($C_{14}H_{16}N_2O_3S$) prevents centrosome migration and mitosis by blocking KIF11 activity. These existing drugs are readily testable in clinic treatment of TNBC cases.

Discussion

The high prevalence of somatic and germline variation in core promoter region of TNBC genomes revealed by our study indicates that core promoter in cancer genome is highly mutable. As such, core promoter provides a new paradigm to study the mechanisms of abnormal gene expression in TNBC and in other types of cancer.

From 155,281 variants in breast cancer generated by TCGA, we identified 2,618 core promoter variants in 2,032 genes, and generated Table S8. The results support our hypothesis that core promoter is mutable in breast cancer.

Core promoter variation can substantially influence gene expression as demonstrated in our expression analysis that about a third of the genes with either somatic or germline variants had expression changes. The changes can be either increased or decreased expression (Fig. 3), and the increased ones were more than the decreased ones in both somatic and germline variant groups (Table S4A, B). The increased expression was observed in core promoter variation in *TERT*, in which the two C > T mutations at −50 and −72 in the core promoter of *TERT* led to increased expression of *TERT* and tumorigenesis.[9,10] The increased expression was also seen in *BRCA2* in our study; a G > T variant at +34 created a DPE motif in the core promoter of *BRCA2* (Fig. 4B) and caused an increased *BRCA2* expression in RNA-seq data (3.4-fold increase) and luciferase reporter assay (3.7-fold increase). In the absence of coding variants in *BRCA2* in the patients,[13] the TNBC may adapt homologous recombination by increasing *BRCA2* expression to maintain TNBC genome stability. It is noted that no core promoter variation was present in ER, PR, and HER2 in TNBC. Therefore, lack of ER, PR, and HER2 expression in TNBC is not due to core promoter variation but by other possible mechanisms, such as epigenetic modification or variation in other cis-elements in distal regulatory region. It is necessary to indicate that gene expression is a dynamic process, most of the genes except housekeeper genes are developmentally regulated and tissue-specific. The RNA-seq data used in our analysis were generated at a given time point. Therefore, 34.5% genes affected were likely the low threshold for the effects of core promoter mutation on gene expression.

The number of somatic variant-affected genes was much more than these with germline variants. This can be related to the randomness nature of somatic variation, within the sequence region of a given length, the chance of somatic variation should be much higher than germline variation as

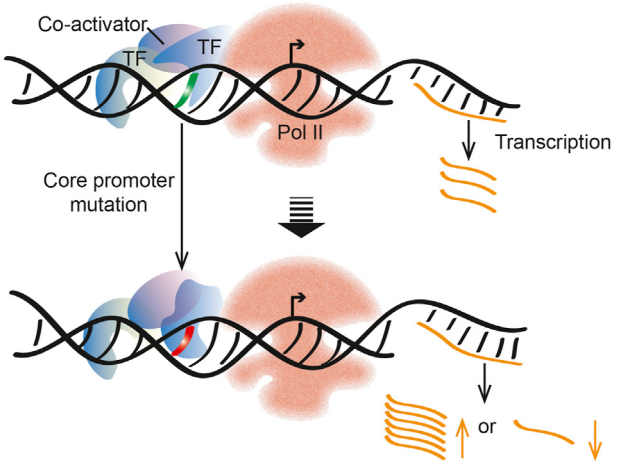

Figure 6 Model of core promoter variation and altered gene expression. It shows that variants in core promoter interfere with the cis—trans interaction and the organization of transcriptional initiation complex, leading to increased or decreased transcription initiation. Green line represents wild-type and red line represents mutant.

it follows mendelian genetics. This feature is also reflected by the differences of somatic variants collected in COSMIC database and germline variants collected in ClinVar database: there are 38,343,693 coding somatic variants in COSMIC database but only 982,763 coding germline variants in ClinVar database. The lower matching rate of the core promoter somatic variants in COSMIC database is likely due to the lack of non-coding somatic variants in COSMIC database. The fact that there were 38,343,693 coding somatic variants but only 15,916,617 non-coding variants in COSMIC supports this explanation, considering that coding region account for only 1—2% of the genome. To ensure high reliability of the somatic variants, we also filtered the variants called from TNBC extensively with the variation data from multiple types of normal human populations and germline variants from the same patient cohort to ensure the reliability of the identified somatic variants.

Based on the data from our study, we propose a model to explain the relationship between core promoter variation and altered gene expression in TNBC (Fig. 6): the motifs in core promoter are short with 3—5 bases only. A slight base change in core promoter sequences within the motif can not only easily interfere with the motif but also easily create a new motif; the variants located out of motif could alter the spatial relationship of core promoter sequences with trans-factors. As such, core promoter variation can influence transcription initiation, leading to either increased or decreased expression, contributing to TNBC etiology.

In summary, our study provides a genome-wide view for core promoter variation in TNBC, and highlights that core promoter variation can be a new paradigm to study the mechanism of abnormal gene expression in cancer.

Conclusion

Our study demonstrates that core-promoter is highly mutable in cancer, and can play etiological roles in TNBC and other types of cancer through influencing transcriptional initiation.

Author contributions

Teng Huang: Data Curation, Methodology, Software, Formal analysis, Writing - Original Draft, Visualization; Jiaheng Li: Validation, Investigation, Writing - Original Draft; San Ming Wang: Conceptualization, Methodology, Writing - Review & Editing, Funding acquisition.

Conflict of interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the Macau Science and Technology Development Fund (No. 085/2017/A2, 0077/2019/AMJ), the University of Macau (No. SRG2017-00097-FHS, MYRG2019-00018-FHS, 2020-00094-FHS), the Faculty of Health Sciences, University of Macau (No. FHSIG/SW/0007/2020P and a startup fund) to SMW.

Acknowledgements

We are thankful for the Information and Communication Technology Office, University of Macau for providing the High-Performance Computing Cluster resource and facilities for the study.

Code availability

Public websites and code packages used in this study were described in the methods.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2022.01.003.

References

1. Vo Ngoc L, Wang YL, Kassavetis GA, et al. The punctilious RNA polymerase II core promoter. *Genes Dev*. 2017;31(13):1289—1301.
2. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol*. 2012;1(1):40—51.
3. Lubliner S, Regev I, Lotan-Pompan M, et al. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res*. 2015;25(7):1008—1017.
4. Sato MP, Makino T, Kawata M. Natural selection in a population of Drosophila melanogaster explained by changes in gene expression caused by sequence variation in core promoter regions. *BMC Evol Biol*. 2016;16:35.
5. Wray GA. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*. 2007;8(3):206—216.
6. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*. 2015;16(4):197—212.

7. Poulos RC, Thoms JA, Shah A, et al. Systematic screening of promoter regions pinpoints functional *cis*-regulatory mutations in a cutaneous melanoma genome. *Mol Cancer Res*. 2015;13(8):1218—1226.
8. Lappalainen T, Montgomery SB, Nica AC, et al. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet*. 2011;89(3):459—463.
9. Huang FW, Hodis E, Xu MJ, et al. Highly recurrent TERT promoter mutations in human melanoma. *Science*. 2013;339(6122):957—959.
10. Horn S, Figl A, Rachakonda PS, et al. TERT promoter mutations in familial and sporadic melanoma. *Science*. 2013;339(6122):959—961.
11. Kumar P, Aggarwal R. An overview of triple-negative breast cancer. *Arch Gynecol Obstet*. 2016;293(2):247—269.
12. Garrido-Castro AC, Lin NU, Polyak K. Insights into molecular classifications of triple-negative breast cancer: improving patient selection for treatment. *Cancer Discov*. 2019;9(2):176—198.
13. Jiang YZ, Ma D, Suo C, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell*. 2019;35(3):428—440.
14. Lips EH, Mulder L, Oonk A, et al. Triple-negative breast cancer: BRCAness and concordance of clinical features with BRCA1-mutation carriers. *Br J Cancer*. 2013;108(10):2172—2177.
15. Koboldt DC, Fulton RS, McLellan MD, et al, Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61—70.
16. Pereira B, Chin SF, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*. 2016;7:11479.
17. Kim YC, Cui J, Luo J, et al. Exome-based variant detection in core promoters. *Sci Rep*. 2016;6:30716.
18. Zeng C, Guo X, Wen W, et al. Evaluation of pathogenetic mutations in breast cancer predisposition genes in population-based studies conducted among Chinese women. *Breast Cancer Res Treat*. 2020;181(2):465—473.
19. Gao Y, Zhang C, Yuan L, et al. PGG.Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res*. 2020;48(D1):D971—D976.
20. Cao Y, Li L, Xu M, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res*. 2020;30(9):717—731.
21. Auton A, Brooks LD, Durbin RM, et al, 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68—74.
22. Lan T, Lin H, Zhu W, et al. Deep whole-genome sequencing of 90 Han Chinese genomes. *GigaScience*. 2017;6(9):1—7.
23. Song S, Tian D, Li C, et al. Genome variation map: a data repository of genome variations in BIG data center. *Nucleic Acids Res*. 2018;46(D1):D944—D949.
24. Du Z, Ma L, Qu H, et al. Whole genome analyses of Chinese population and de novo assembly of A northern Han genome. *Dev Reprod Biol*. 2019;17(3):229—247.
25. Wall JD, Stawiski EW, Ratan A, et al, GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019;576(7785):106—111.
26. Miao X, Li X, Wang L, et al. DSMNC: a database of somatic mutations in normal cells. *Nucleic Acids Res*. 2019;47(D1):D971—D975.
27. Wu D, Dou J, Chai X, et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell*. 2019;179(3):736—749.
28. Ou J, Zhu LJ. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. *Nat Methods*. 2019;16(6):453—454.
29. Weinstein JN, Collisson EA, Mills GB, et al, Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113—1120.
30. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996—1006.
31. Villanueva RAM, Chen ZJ. Meas-Interdiscip Res. *ggplot2: elegant graphics for data analysis*. 2nd ed. 2019;17:160—167(3).
32. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25—29.
33. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics*. 2016;54:1.30.1-1.30.33.
34. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074—D1082.
35. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313(5795):1929—1935.
36. Stathias V, Turner J, Koleti A, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res*. 2020;48(D1):D431—D439.
37. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(Database issue):D991—D995.
38. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523.
39. Wang Z, He E, Sani K, et al. Drug Gene Budger (DGB): an application for ranking drugs to modulate a specific gene based on transcriptomic signatures. *Bioinformatics*. 2019;35(7):1247—1248.
40. Gupta H, Chandratre K, Sinha S, et al. Highly diversified core promoters in the human genome and their effects on gene expression and disease predisposition. *BMC Genom*. 2020;21(1):842.
41. Weber JL, Wong C. Mutation of human short tandem repeats. *Hum Mol Genet*. 1993;2(8):1123—1128.
42. Sawaya S, Bagshaw A, Buschiazzo E, et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One*. 2013;8(2):e54710.
43. Gymrek M, Willems T, Guilmatre A, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*. 2016;48(1):22—29.
44. Cui J, Luo J, Kim YC, et al. Differences of variable number tandem repeats in *XRCC5* promoter are associated with increased or decreased risk of breast cancer in BRCA gene mutation carriers. *Front Oncol*. 2016;6:92.
45. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med*. 2016;375(8):717—729.
46. Huang CC, Tu SH, Lien HH, et al. Prediction consistency and clinical presentations of breast cancer molecular subtypes for Han Chinese population. *J Transl Med*. 2012;10(Suppl 1):S10.