Germline genetics encode the resistance, risk, and lymphatic metastasis of triple-negative breast cancer in the southern Chinese population

Mei Yang[†1], Yanhui Fan[†2], Qiangzu Zhang[2], Shunhua Han[2], Xiaoling Li[1], Teng Zhu[1], Minyi Cheng[1], Juntao Xu[2], Ciqiu Yang[1], Hongfei Gao[1], Chunming Zhang[2], Michael Q. Zhang[3], You-Qiang Song[4], Gang Niu[2]*, Kun Wang[1]*

[1]Department of Breast Cancer, Cancer Center, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong, China

[2]Phil Rivers Technology, Beijing, China

[3]MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; School of Medicine, Tsinghua University, Beijing, China

[4]School of Biomedical Sciences, The University of Hong Kong, Hong Kong, China

[†]These authors contributed equally to the work.

*Correspondence should be addressed to Dr. Gang Niu at Phil Rivers Technology, 807, 8/F, Yisibo Software Building, Haitian Er Road, Shenzhen, Guangdong, China, or at g.niu@philrivers.com; or to Dr. Kun Wang at the Department of Breast Cancer, Cancer Center, Guangdong General Hospital, Guangdong Academy of Medical Sciences, 106 Zhongshan 2nd Road, Guangzhou, Guangdong, China, or at gzwangkun@126.com.

Abstract

BACKGROUND Triple-negative breast cancer (TNBC) patients generally have poor prognosis but could be led to a better outcome if the risk can be identified in the stage without symptoms. Previously, determining the hereditary risk relied solely on assessing the potential damage due to single mutations or individually damaged genes, which has led to fragmentation in our understanding of genetic risk, particularly when multiple factors are involved. It is currently difficult to obtain a complete picture of genetic risk that unites individual discoveries under a single theoretical frame, not to mention a category method for systematically assessing all individuals within a population.

METHODS We postulated that there is a persistent stress from certain deterministic pathogenic factors on the whole population, so that we can simply solve this problem by assessing the innate stress resistance in each individual. This stress resistance is encoded in germline DNA and can be disrupted by rare mutations in certain genes. We developed a method to statistically infer whether a set of cancer related pathways in individual subjects would be differentially active or repressed driven by all currently mutated genes.

RESULTS We divided all 141 subjects including 29 TNBC patients and 112 normal females from Southern China into five categories of TNBC risk: very high risk, high risk, average, low risk, and zero risk. Approximately 4.5% and 31% were evaluated to be at very high risk of TNBC in normal population and patients, respectively. Whereas around 25% would have zero risk of TNBC in normal population.

Surprisingly, lymphatic metastasis correlated with the risk of disease (r^2 = 0.99, p = 0.0035) in patient population.

CONCLUSIONS Our findings suggested that a health human genome could encode an ability fully protecting the individual from a persistent neoplastic threat.

Introduction

According to data from the China National Central Cancer Registry, the incidence (268,600 new cases in 2015) and mortality (69,500 death cases in 2015) of breast cancer has continued to increase in China over the past two decades, and is fast becoming a major health concern in women.[1,2] Triple-negative breast cancer (TNBC) is defined by ≤1% estrogen receptor positive tumor cells, progesterone receptor negativity, and normal HER2-receptor expression detected by immunohistochemistry or in situ hybridization analysis (separately or combined).[3] Among all types of breast cancer, TNBC makes up 15%–20% of cases, usually affecting younger women.[4] Compared with other types of breast cancer, patients with TNBC generally have poor prognosis due to the combination of its inherent aggressive clinical behavior and lack of molecular targets for therapy.[5,6] On the other hand, studies have shown that screening and appropriate treatment can reduce breast cancer mortality. The contribution of screening and treatment in reducing cancer mortality differs according to the breast cancer subtype, of which the TNBC subtype benefits the most.[7] This has spurred major efforts in identifying potential molecular features that can be used to assess the risk for TNBC for surveillance, early detection, and timely treatment. The goal is a better outcome to prolong survival in patients with TNBC and to reduce TNBC-specific mortality.[7,8]

Identification of patients at risk for TNBC by genetic testing has been proven to help reduce breast cancer mortality.[9] Correspondingly, using a small number of high-penetrance mutations with a much larger number of low-penetrance variants to build a risk model for breast cancer is widely accepted. Much effort has focused on trying to identify and characterize high- and low-penetrance susceptibility genes. The well-

known inherited BRCA1/2 (Breast Cancer genes 1 and 2) mutations are considered to be the most powerful predictors of the risk of developing breast cancer.[10,11] However, those studies have proven to be very complex and they have not provided conclusive data.[12] Furthermore, BRCA1/2, the known high-penetrance genes account for no more than 25% of breast cancer cases based on prior studies and mathematical modeling.[13,14] The remaining more than 75% of the familial risk for breast cancer is still unexplained and the extent of the contribution of an individual's genotype to the risk for breast cancer[10] is still unknown. Moreover, healthy women with breast cancer risk genotypes, but without symptoms, are not included in the current risk models.

In this paper, we postulated a persistent stress from certain deterministic pathogenic factors on the whole population, so that we can simply solve the problem by assessing the innate stress resistance in each individual. This resistance is encoded in germline DNA and can be disrupted by rarely occurring mutations in certain genes. We developed a multiple-step method to statistically infer the resistance for each individual by measuring function of a set of pathways disrupted by any rare mutations. Consequently, a population without symptoms could be categorized into 5 resistance levels, in which 4.45% were classified as very high risk/very low resistance and, amazingly, up to 25% were classified as zero risk/fully protected. Interestingly, lymph node status is strongly associated with the five levels. Our method provides a new way to truly understand the nature of health and the possibly final weapon inside us against cancers.

Methods

Study design and patients

The general design of the study is presented in Fig. 1. This study was approved by the research ethics committee at Guangdong General Hospital, Guangdong Academy of Medicals Sciences. Written informed consent was obtained from all participants to allow the use of banked tissues (including white blood cells and oral epithelial cells), and for collection of pathological data and clinical follow-up data. The characteristics of the TNBC patients are described in supplemental Table 1.

Whole-exome sequencing

Oral epithelial cells or peripheral blood leukocytes were collected from individuals. Each sequenced sample was prepared according to the Illumina protocols. Paired-end multiplex sequencing of samples was performed on the Illumina HiSeq X Ten sequencing platform.

Variant calling

Paired-end raw sequence reads were mapped to the human reference genome (UCSC hg19) using Burrows-Wheeler Aligner[15] using default settings. Variants calling was carried out using the HaplotypeCaller module in Genome Analysis Toolkit[16] following GATK Best Practices. The variants then were annotated by ANNOVAR[17] based on RefGene.

Variant coding

Genes were coded as 1 or 0 based on whether or not there was any mutation in the gene. The Z-score was calculated using both mutation and expression data from the

COSMIC Cell Line Project.[18] Briefly, for each gene, all cell lines were separated into two groups based on whether or not they carried this gene mutation, and the differences between the average expression of each gene could then be calculated. We randomly assigned group labels to each cell line while maintaining group size, and then calculated the difference between the two simulated groups. This process was repeated 10000 times and the Z-score was calculated as the standardized value of the difference in the average expression.

$$Z = \frac{x - \mu}{\sigma}$$

Where μ and σ are the mean and standard deviation, respectively.

Statistical analysis

Method implemented in NMRCLUST[19] was used for clustering. This method uses the average linkage to define how clusters are built up, followed by penalty function that simultaneously optimizes the number of clusters and the average spread of the clusters. Naïve bayes[20] that was implemented in R package caret[21] was used to build the binary classifiers. We first randomly split the data into training group (75%) and testing group (25%) while preserve the overall class distribution of the data. Then the 10-fold cross validation was used to search the optimal parameters. All statistical analyses were conducted using an in-house developed script in Perl, Matlab, and R.

Results

The study recruited 29 TNBC patients and 112 healthy aged women (78.49 ± 1.24 years). We found 10^7 SNPs from their genomic DNA of each subject, of which 10^4 were rare (MAF < 0.01) germline mutations, which were used for coding. Those rare germline variants were coded into around 300 genes with mutations (Fig. 1, left panel). Using the reference database (Fig. 1, right panel), we expanded the number of deterministically regulated expression genes to about 1000 which will be compared to the signatures of cancer related pathways. Finally, we projected the rare germline variants onto a spectrum of active/resting pathways less than one hundred.

Rare germline mutation profiles are exceptionally sparse

A germline variant with a frequency of less than 0.01 was defined as a rare germline variant. The average numbers of genes with rare germline variants in healthy people ranged from 220 to 300 as shown in Fig. 2B, which appeared to have a normal distribution. When comparing the number of mutated genes between any two individuals, we found between 4 and 34 shared genes, which also appeared to have a normal distribution (Fig. 2C). This suggests that most people have about 220 to 300 genes with rare germline variants, but there were only about 4 to 34 common mutated genes between any two individuals.

These rare variants were then coded into genes. In the cluster diagram (Fig. 2A), the variants in the genes were high-dimensional, sparse, discrete, and nonstandard. Almost no similar variation patterns were observed between any two individuals, which indicated each person was unique. This is well illustrated in the upper panel of Fig. 2F. The Pearson correlation coefficient histogram was narrow and evenly

distributed toward zero with a median value was < 0.05. Therefore, few rules could be inferred based on their gene mutation patterns, even being given sharp and meaningful phenotypes in current subjects.

We next coded the variants by their presence in cancer dependencies[22] and clustered them based on their effect on gene expression. We could infer, to some extent, a little more regularity from the cluster diagram (Fig. 2D). The correlation coefficient histogram shifted to the right of the X-axis and the range extended to from -0.3 to 0.5 (Fig. 2F, middle panel), indicating there might be some complicated rules in the gene expression.

When the rare variants effect on gene expression were projected onto the pathway activity, we observed better clustering (Fig. 2E). The correlation coefficient histogram was no longer a normal distribution and shifted more to the right of the X-axis and the range extended to from -0.5 to 0.9, with a median value was 0.5 (Fig. 2F, lower panel). When the germline variant information was gradually mapped onto the low-dimensional functional feature space, the internal regularity of the population gradually increased, suggesting that individuals with similar phenotypes might follow the same rules in this low-dimensional feature space.

Most pathways are downregulated in TNBC patients

When we projected the rare germline variants of the 29 TNBC patients onto those pathways, the cluster diagram showed a regular pattern (Fig. 3A). However, the patterns in the cluster diagram appeared to be complex and could not be generalized by specific pathways. Next, we compared the pathway activities between the 112

healthy people and the 29 TNBC patients to try to characterize the features in TNBC patients. The upregulated pathways activity did not contribute to TNBC patient clustering (Fig. 3B, asterisk line), whereas the downregulated pathways activity greatly contributed to the TNBC clustering (Fig. 3B, dot line). When each pathway's activity was further examined (Fig. 3C), we found 83% of the pathway activities were downregulated, indicating a repressed trend in most pathways in TNBC. In addition, the Z-score cutoff value tended to the minimum (Fig. 3B), the more the pathway activity was downregulated, suggesting the risk/resistance for TNBC could be predicted by a small number of pathways. With the Z-score cutoff value set to -4 (Fig. 3D), the risk/resistance of TNBC could be predicted by just three pathways.

Three pathways can predict the risk of TNBC

To test our model, we chose three repressed pathways to predict the risk/resistance of TNBC. From Fig. 4A, we can see the average AUC appears to be a normal distribution, with a median value of 0.75. Fig. 4B shows the ROC curve with 95% confidence interval (CI) of false positive rate. With Z-score cutoff values set as from -1 to 0, the prediction accuracy was more than 70% (Fig. 4C) and false discovery rate was less than 30% (Fig. 4D), indicating high efficacy for predicting TNBC risk/resistance by three downregulated pathways.

Clustering people into five categorizes base on three pathways

We used the germline features of the three identified pathways to categorize the pooled study sample (112 healthy people and 29 TNBC patients) into groups corresponding to the level of TNBC risk/resistance. The result suggested that the appropriate number of cluster groups was 16 with a penalty < 22 (Fig. 5A). For each

group, we compared the proportion, Z score, and ratio of proportions between healthy people and TNBC patients (Fig. 5B), and found some groups had values that were very close. The groups with similar values were then combined to give the final five groups (Fig. 5C), which indicated five different levels of TNBC risk/resistance in this population. We found 31.03% TNBC patients and 4.45% healthy people fell into class 1 (very high risk/very low resistance) with a Z-score of 8, which indicated those with this pathway pattern had a TNBC risk/resistance about eight standard deviations above/below the mean risk/resistance. The proportion of TNBC patients was more than twice compared to the proportion of healthy people in class 2 (high risk/low resistance). Class 3 (average risk/resistance) included similar proportions of healthy people and TNBC patients, which suggested the risk of TNBC in this group was not related to germline variants. Notably, healthy people in class 4 (low risk/high resistance) might have protective factors against TNBC, as they had a lower risk of TNBC. Likewise, class 5 (zero risk) included 25% healthy people, but no TNBC patients, which indicated those with this pathway pattern had no risk of TNBC. In terms of predictive value, 4.45% healthy people (class 1) had very high risk/very low resistance of TNBC, which indicated the elimination trend; 16% healthy people (class 1 and 2) were at high risk of TNBC; and nearly 50% healthy people (class 4 and 5) were protected from TNBC. When the five groups were sorted by Z-scores, the distribution of the pathway activities appeared regular (Fig. 5D, left panel), according to the risk/resistance level. For individuals with certain germline backgrounds of rare genetic variants, the risk of TNBC could be significantly different. We offer a web-based service for assessment of the resistance/risk level using our methods to help other users of interest to evaluate the resistance/risk level of their patients or healthy subjects if the users provide the germline mutation information to us. The service is

available at http://philrivers.vicp.io:9900/.

Risk/Resistance classes are associated with lymph node metastasis

Next, we analyzed the clinical features of the TNBC patients. We observed a strong association (r^2 = 0.99, p = 0.0035) between Lymph node status and the five classes (Fig. 5D, right panel). We found 67% patients in class 1 had lymph node metastasis, indicating aggressive pathway activity patterns. Likewise, 44% patients in class 2 and 25% in class 3 also had lymph node metastasis. No patients in class 4 had lymph node metastasis. Correspondingly, the median onset age of patients in class 1 was younger than that in class 2 and 3 but is not statistical significant (Fig. 5E). To a certain extent, we were able to confirm the activation of these three pathways was related to the risk of TNBC and its progression.

Discussion

The aim of this study was to gain a deeper understanding of the genetic features of TNBC. With regards to gene mutations, gene expressions, and cancer pathways, we found the distance between genetic codes of any two individuals was generally becoming closer. When we projected the rare germline variants onto cancer pathways, significant genetic features were identified in the 29 TNBC patients. Furthermore, when pooling all 112 healthy people and 29 TNBC patients, we were able classify them into five risk/resistance groups for TNBC based on three repressed pathways. In these five groups, each individual with a certain pathway activity pattern had a different degree of risk/resistance for TNBC (in a spectrum from very high to zero risk). People with low or no risk (class 4 and 5) might have protective factors against TNBC, whereas those with very high risk/low resistance (class 1) will develop TNBC.

Surprisingly, the risk/resistance patterns were closely associated with lymph node status, and very high risk/low resistance patients (class 1) tended to have lymph node metastasis, whereas low risk/high resistance patients (class 4) did not have lymph node metastasis, even in those with the same tumor staging. Because lymph node status has been identified as a significant independent prognostic factor for TNBC progress, mortality, and overall survival,[23-25] our rare germline variants-based risk/resistance patterns can be correlated to TNBC prognosis. The results may have important implications for our understanding of TNBC pathogenesis, allowing more aggressive treatments and monitoring of certain subgroups at risk of TNBC.

Cancer is a complex genetic disease that is a result of the accumulation of genomic alterations.[26] Germline variations predispose individuals to cancer and somatic alterations initiate and trigger the progression of cancer.[26] Although genetic data, especially rare variants, could provide useful risk prediction,[27] more efficient and accurate genetic-association models need to be established. We know that DNA and its structure holds massive amounts of information, which can be represented by binary values (on or off[28]). Changes in multiple DNA sequences could contribute to one gene expression ("many to one"). Therefore, gene expression is continuous and informative. In this study, we collected all rare genetic variants in 112 healthy people and coded those rare variants into gene expressions. We obtained more regular patterns compared with gene mutation patterns. However, the data from rare genetic variants were high-dimensional and nonstandard. When we projected the genetic rare events onto cancer pathways, the cluster diagram appeared to be regular. We also showed that several individuals with similar phenotypes might share similar pathway

activity patterns. Using this method, we found many TNBC patients shared similar characteristics and cancer pathways. When we compared the pathway patterns between healthy people and TNBC patients, we found three repressed pathways in TNBC. Surprisingly, based on only these three pathway activity patterns, we were able to grade the TNBC risk/resistance into different classes of risk/resistance. Most importantly, using this model, we were able to definitively identify subpopulations that have high risk for TNBC and those that have no risk.

Our model is a multivariate composite assessment model, in which all rare genetic variants of an individual are used to estimate the possibility of downstream events, which can be used to grade the risk of TNBC. At the genetic level, many tiny driving forces are counted together to determine the occurrence of certain downstream events, such as cancer. Our models give a very well-defined understanding of hereditary predispositions for TNBC. However, larger population-based prospective studies are required to validate these findings. With appropriate study populations, we can establish germline risk models for other pathological types, which could lead to general mechanisms. On the other hand, there is no patient in class 5 but 25% healthy people fall into this category, which indicates about a quarter of people have no risk of TNBC. Further research is needed to study whether these people have innate resistance that is coded by genetics to most type of cancers and to investigate the underlying mechanism of resistance to cancer. These results may revolutionize our current understanding of cancer and provide new strategy for realizing precision healthy.

Acknowledgments

We would like to thank Jin Gu for the fruitful discussions and advice. The present study was supported in part by the National Natural Science Foundation of China (grant no. 81202076), the Guangzhou Science and Technology Program (grant no. 2014J2200007), and the Natural Science Foundation of Guangdong (grant no. 2017A030313882).

Author Contributions

GN and YF conceived the experiments. KW and MY designed the experiments. MY, XL, TZ, MC, JX, CY and HG performed the experiments. YF, QZ, SH, CZ, YQS, MZ and GN analyzed the data. MY, YF, QZ, and GN wrote the paper. All authors discussed the results and contributed to the final manuscript.

References

1. Jiang X, Tang H, Chen T. Epidemiology of gynecologic cancers in China. J Gynecol Oncol 2018;29:e7.
2. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66:115-32.
3. Stovgaard ES, Nielsen D, Hogdall E, Balslev E. Triple negative breast cancer - prognostic role of immune-related factors: a systematic review. Acta Oncol 2018;57:74-82.
4. Reis-Filho JS, Tutt AN. Triple negative tumours: a critical review. Histopathology 2008;52:108-18.
5. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat Rev Clin Oncol 2016;13:674-90.
6. Malorni L, Shetty PB, De Angelis C, et al. Clinical and biologic features of triple-negative breast cancers in a large cohort of patients with long-term follow-up. Breast Cancer Res Treat 2012;136:795-804.
7. Plevritis SK, Munoz D, Kurian AW, et al. Association of Screening and Treatment With Breast Cancer Mortality by Molecular Subtype in US Women, 2000-2012. JAMA 2018;319:154-64.
8. Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med 2005;353:1784-92.
9. Lobo M, Lopez-Tarruella S, Luque S, et al. Evaluation of Breast Cancer Patients with Genetic Risk in a University Hospital: Before and After the Implementation of a Heredofamilial Cancer Unit. J Genet Couns 2017;DOI:10.1007/s10897-017-0187-3.
10. Fackenthal JD, Olopade OI. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. Nat Rev Cancer 2007;7:937-48.
11. Jara L, Morales S, de Mayo T, Gonzalez-Hormazabal P, Carrasco V, Godoy R. Mutations in BRCA1, BRCA2 and other breast and ovarian cancer susceptibility genes in Central and South American populations. Biol Res 2017;50:35.
12. Nathanson KL, Wooster R, Weber BL. Breast cancer genetics: what we know and what we need. Nat Med 2001;7:552-6.
13. Walsh T, Casadei S, Coats KH, et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA 2006;295:1379-88.
14. Antoniou AC, Easton DF. Models of genetic susceptibility to breast cancer. Oncogene 2006;25:5898-905.
15. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. q-bioGN 2013;arXiv:1303.3997v2.
16. McKenna AH, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164-e.
18. Iorio F, Knijnenburg Theo A, Vis Daniel J, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016;166:740-54.
19. Kelley LA, Gardner SP, Sutcliffe MJ. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein Engineering, Design and Selection 1996;9:1063-5.
20. Zhang Z. Naïve Bayes classification in R. Annals of Translational Medicine 2016;4:241.

21. Kuhn M. Building Predictive Models in R Using the caret Package. 2008 2008;28:26.
22. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a Cancer Dependency Map. Cell 2017;170:564-76.e16.
23. Urru SAM, Gallus S, Bosetti C, et al. Clinical and pathological factors influencing survival in a large cohort of triple-negative breast cancer patients. BMC Cancer 2018;18:56.
24. Hatoum HA, Jamali FR, El-Saghir NS, et al. Ratio between positive lymph nodes and total excised axillary lymph nodes as an independent prognostic factor for overall survival in patients with nonmetastatic lymph node-positive breast cancer. Ann Surg Oncol 2009;16:3388-95.
25. Rosa Mendoza ES, Moreno E, Caguioa PB. Predictors of early distant metastasis in women with breast cancer. J Cancer Res Clin Oncol 2013;139:645-52.
26. Low SK, Zembutsu H, Nakamura Y. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Sci 2017;DOI:10.1111/cas.13463.
27. Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. Genet Epidemiol 2017;DOI:10.1002/gepi.22092.
28. The management of ductal carcinoma in situ (DCIS). The Steering Committee on Clinical Practice Guidelines for the Care and Treatment of Breast Cancer. Canadian Association of Radiation Oncologists. CMAJ 1998;158 Suppl 3:S27-34.

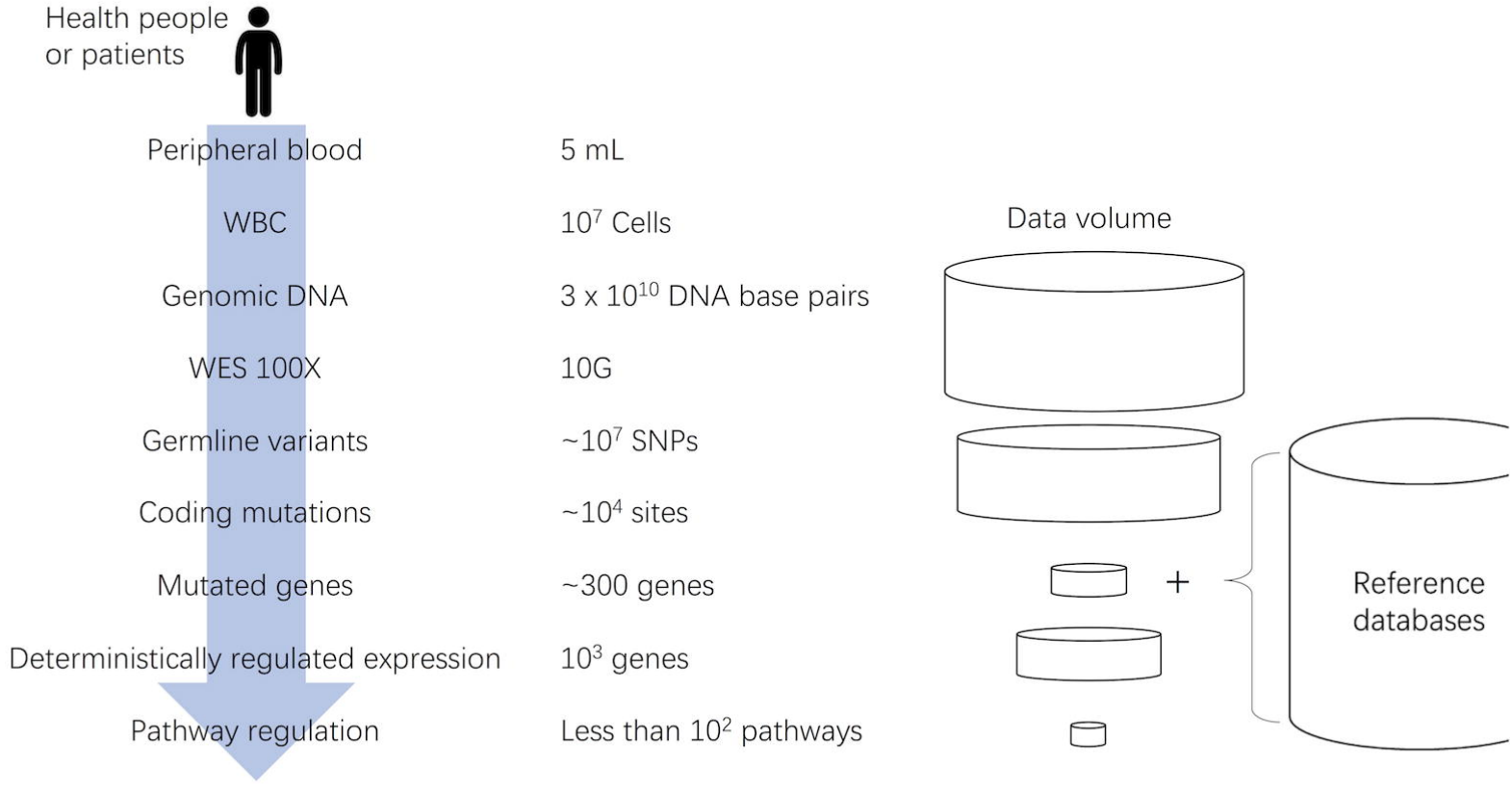
Figure 1. Flow diagram illustrating the workflow developed for this study. GATK best practices workflow was used for variants calling on whole exome sequencing data. Annotation was performed using ANNOVAR. Variants were filtered based on frequency and function. Then variants were mapped to genes and pathways for analysis.

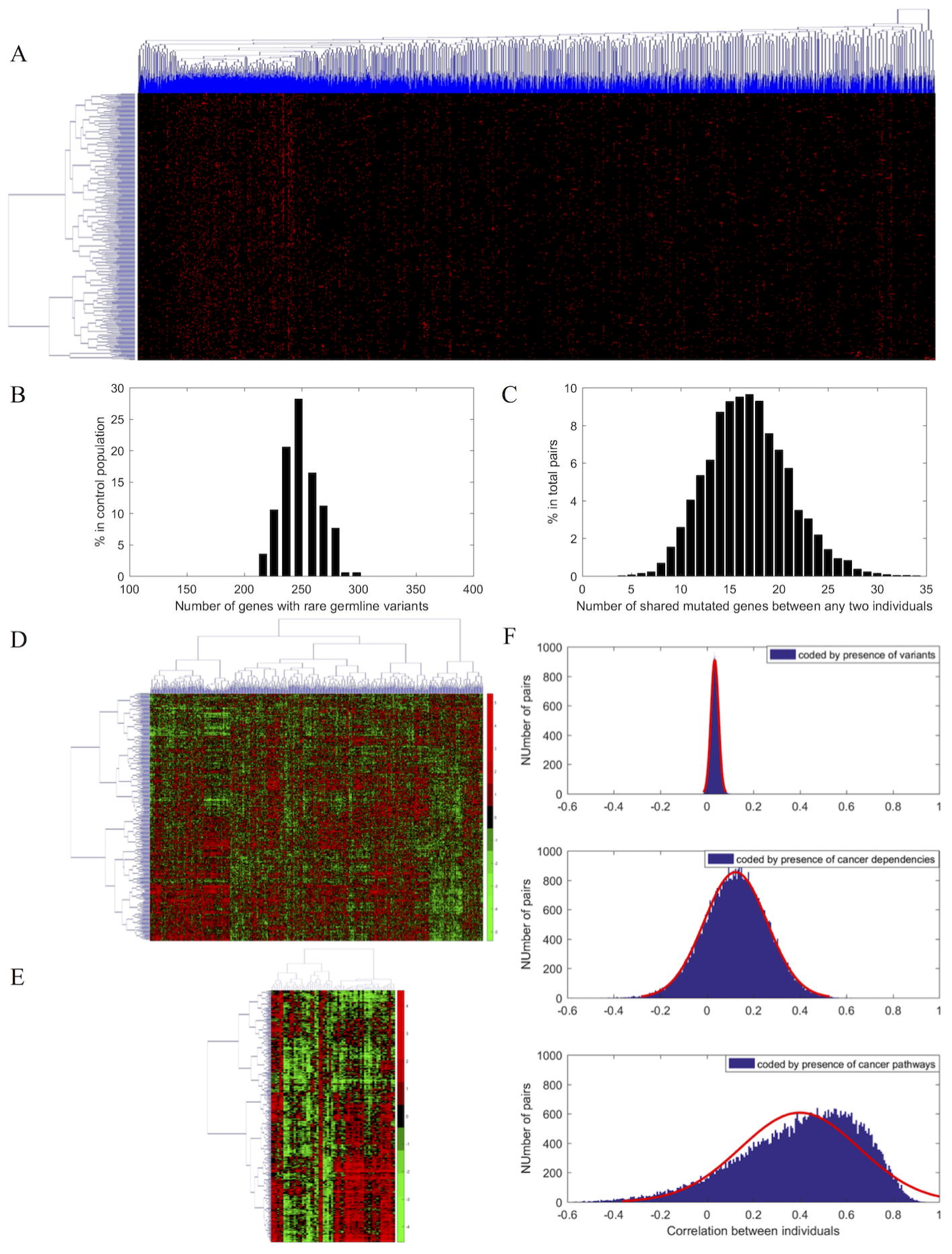
Figure 2. Relationship between any two individuals. (A) Heatmap shows the present and absent of rare variants on genes. Columns represent genes and rows represent individuals. (B) The histogram of the number of genes with rare germline variants. (C) The histogram of the number of shared mutated genes between any two individuals. (D) Heatmap shows the present and absent of rare variants on cancer dependencies. Columns represent cancer dependencies and rows represent individuals. (E) Heatmap shows the present and absent of rare variants on pathways. Columns represent pathways and rows represent individuals. (F) The histogram of correlation coefficient between any two individuals based on variants present or absent on genes (upper panel), cancer dependencies (middle panel) and pathways (lower panel).

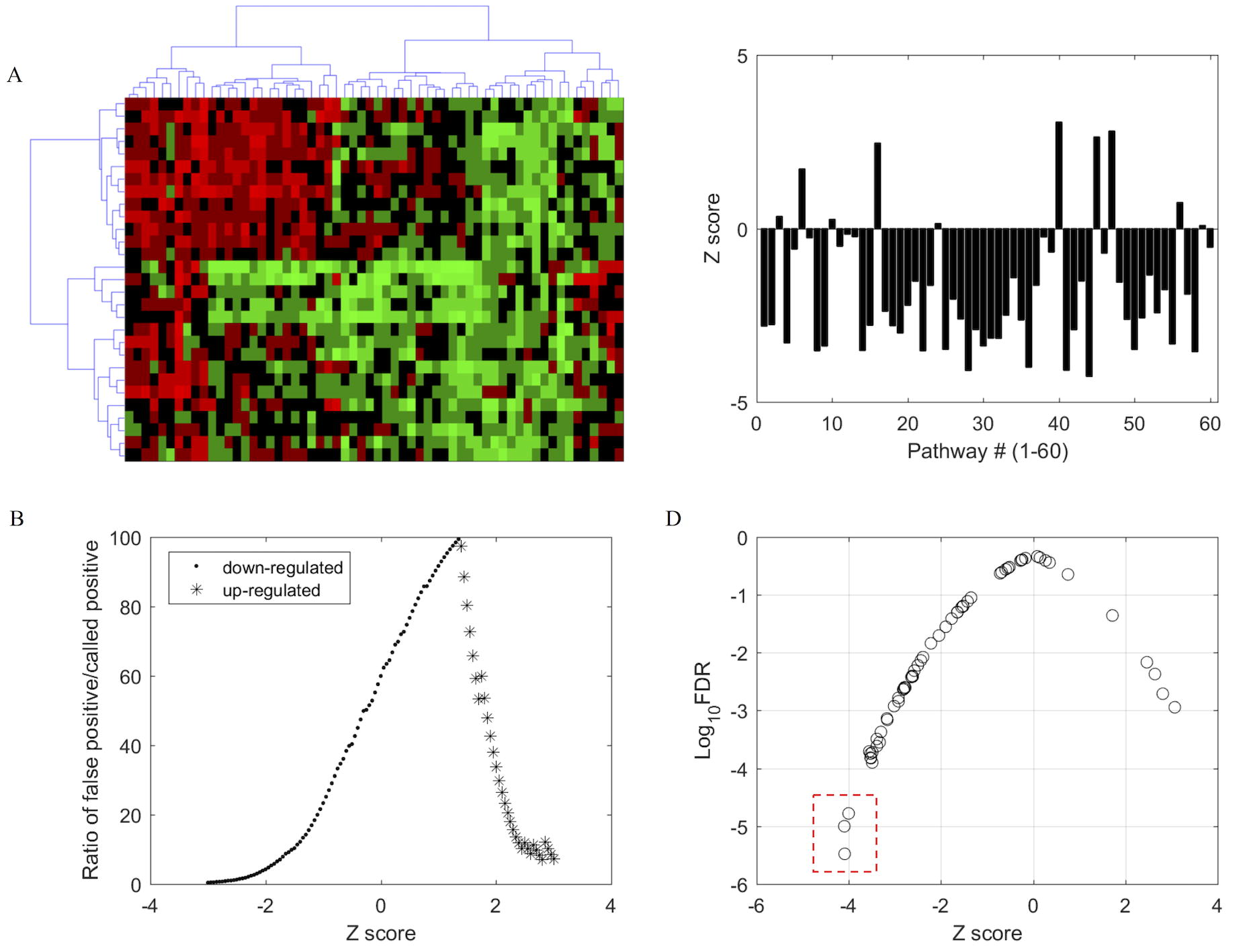
Figure 3. Contribution of pathways in TNBC clustering. (A)Heatmap shows the pattern of pathways in TNBC patients. Columns represent pathways and rows represent patients. (B) Scatterplots showing the relationship between Z score and the ratio of false positive to called positive in both down-regulated pathways and up-regulated pathways. (C) States of pathways. Z score larger than 0 represents up-regulated and less than 0 represents down-regulated. (D) Scatterplots showing the relationship between Z score and the logarithm of false discovery rate with base 10.

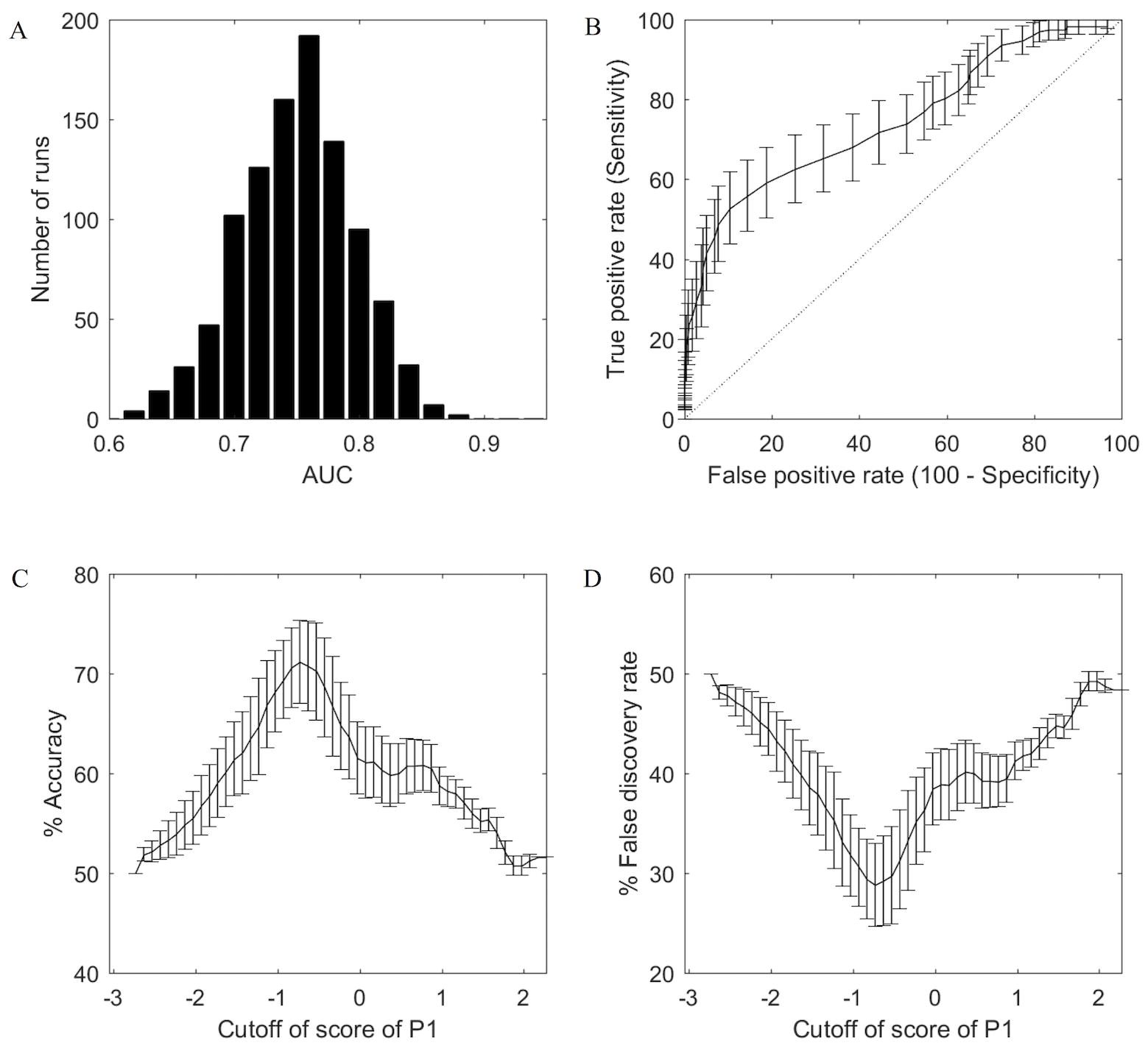
Figure 4. Performance of TNBC risk prediction. (A) Histogram of AUC. (B) ROC curve with 95% CI of the false positive rate. (C) Distribution of accuracy with 95% CI. (D) Distribution of false discovery rate with 95% CI.

Figure 5. Clustering analysis results and its correlation with clinical features. (A) Determining the number of clusters with different penalty values. (B) Distribution of TNBC patients and controls in the 16 clusters (C) Distribution of TNBC patients and controls in the combined 5 clusters. (D) Left panel: heatmap shows the pattern of pathways in different clusters. Columns represent pathways and rows represent clusters. Right panel: Statistics of lymph node and onset age of patients in each cluster. (E) Boxplots showing the onset age of different clusters.

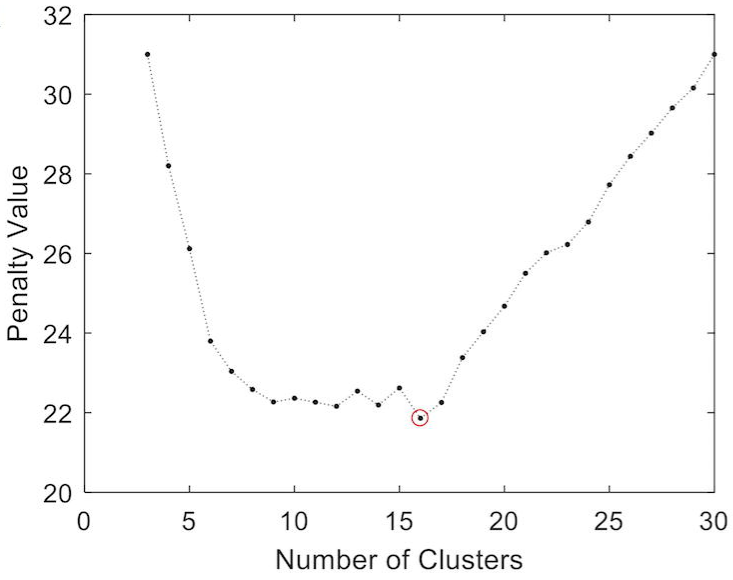

Health people
or patients
Peripheral blood
5 mL
WBC
10^7 Cells
Genomic DNA
3 x 10^{10} DNA base pairs
WES 100X
10G
Germline variants
~10^7 SNPs
Coding mutations
~10^4 sites
Mutated genes
~300 genes
Deterministically regulated expression
10^3 genes
Pathway regulation
Less than 10^2 pathways
Classification of subjects & Rule extraction
Data volume
+
Reference
databases


A
B
% in control population
Number of genes with rare germline variants
C
% in total pairs
Number of shared mutated genes between any two individuals
D
E
F
coded by presence of variants
coded by presence of cancer dependencies
coded by presence of cancer pathways
NUmber of pairs
Correlation between individuals


A
5
0
-5
Z score
0
10
20
30
40
50
60
Pathway # (1-60)
B
100
80
60
40
20
0
Ratio of false positive/called positive
down-regulated
up-regulated
-4
-2
0
2
4
Z score
D
0
-1
-2
-3
-4
-5
-6
Log10FDR
-6
-4
-2
0
2
4
Z score


A
Number of runs
AUC
B
True positive rate (Sensitivity)
False positive rate (100 - Specificity)
C
% Accuracy
Cutoff of score of P1
D
% False discovery rate
Cutoff of score of P1

A



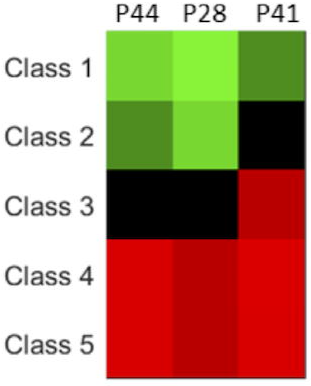
B

class	% in TN	% in Norm	std in Norm	Z score	% in Norm/% in TN	
8	17.24%	2.69%	2.58%	5.6361	0.1563	Class 1
5	13.79%	1.80%	2.11%	5.6917	0.1306	
6	6.90%	4.44%	3.35%	0.7337	0.6440	Class 2
7	3.45%	0.90%	1.51%	1.6843	0.2606	
12	20.69%	6.25%	3.91%	3.6943	0.3023	
14	6.90%	11.49%	5.20%	-0.8826	1.6655	Class 3
11	10.34%	11.68%	5.05%	-0.2646	1.1291	
13	6.90%	7.17%	4.13%	-0.0660	1.0395	
4	6.90%	6.24%	3.84%	0.1721	0.9042	
16	3.45%	11.64%	5.09%	-1.6078	3.3746	Class 4
2	3.45%	10.73%	5.01%	-1.4531	3.1120	
10	0.00%	7.98%	4.40%	-1.8155	Inf	Class 5
9	0.00%	7.16%	4.14%	-1.7285	Inf	
15	0.00%	6.26%	3.86%	-1.6237	Inf	
1	0.00%	1.79%	2.11%	-0.8448	Inf	
3	0.00%	1.78%	2.13%	-0.8368	Inf	

C

class	% in TN	% in Norm	std in Norm	Z score	% in Norm/% in TN
1	31.03%	4.45%	3.32%	8.00	0.14
2	31.03%	11.65%	5.15%	3.77	0.38
3	31.03%	36.58%	7.75%	-0.72	1.18
4	6.90%	22.26%	6.70%	-2.29	3.23
5	0.00%	25.05%	7.04%	-3.56	Inf

D



Class	Number of Patient	% of N in class	Onset age (median±error)
1	9	0.67	55.0±6.65
2	9	0.44	58.5±9.99
3	9	0.25	62.0±6.02
4	2	0	-
5	0	-	-

E