**Supplementary materials**

**Supplementary methods**

**Sample recruitment and ethics statement**

In this study, we recruited 500 patients with BCS that was associated with unknown causes, including tumor invasion, aspergillosis, and Bechet’s syndrome, at the Fifth Affiliated Hospital, Sun Yat-sen University. The results of the clinical diagnoses were verified by manual reviews. We also enrolled 696 individuals as healthy controls. All healthy participants underwent medical check-ups in our hospital and were confirmed to have no history of BCS. These data were used as internal data to construct the model and evaluate the prediction accuracy of the model. In addition, we recruited an additional 30 BCS patients and 316 healthy individuals. These samples were used as independent external data to validate the performance of the model. Written informed consent was obtained from all participants. This study was approved by the Institutional Review Board of the Fifth Affiliated Hospital, Sun Yat-sen University (K05‑1) and the Institutional Review Board of BGI (BGI-IRB 21097).

**Whole-exome sequencing and quality control**

Five milliliters of venous blood was taken from each of the participants, and genomic DNA was extracted according to the manufacturer's standard procedure for the MagPure Buffy Coat DNA Midi KF Kit (Magen, Guangzhou, China). Genomic DNA was fragmented using Segmentase (BGI, Shenzhen, China) to generate small DNA fragments (e.g., 100-500 bp), which were passed over magnetic beads to enrich the fragments that were 280-320 bp long. The ends were filled in, and then base "A" was added to the 3' end to allow the DNA fragment to be ligated to an adapter bearing base "T" at the 3' end. The resulting DNA fragments were amplified by ligation-mediated polymerase chain reaction, purified, and then used to create a library. The library was enriched by array hybridization using the MGIEasy Exome Capture V4 Probe Set (MGI, Shenzhen, China), which was followed by elution and postcapture amplification. The products were then subjected to an Agilent 2100 Bioanalyzer to estimate the magnitude of enrichment. All amplified libraries were subsequently sent to BGI for circularization and sequencing on the MGIseq-2000 platform (BGI, Shenzhen, China) with an average depth of ~120X (PE100). For the original sequencing data, we removed some low-quality reads based on previously reported methods to generate "clean reads"1. The “clean reads” were then aligned to the human genome reference (hg19) using Burrows Wheeler Aligner (BWA) software, which was followed by removal of PCR duplicates with the Picard tool, local indel realignment, base quality score recalibration, and joint variant calling with GATK HaplotypeCaller. We applied the Variant Quality Score Recalibration (VQSR) method to filter out potential low-quality variants with the default datasets and used the parameters recommended by the GATK toolkit.

For the VCF file obtained above, we conducted stringent quality control based on the published PRS guidelines2. VCFTools software was utilized to further remove unqualified variants and samples, and specific parameters were defined as follows3: --remove-filtered-all, --remove-filtered-geno-all, --remove-indels, --minGQ 20, --minDP 30, --mac 3, --max-meanDP 100, --max-missing-count 1, --hwe 1e-6, --min-alleles 2, and --max-alleles 2. We then used PLINK2 software to retain single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAF)>0.05 and that were located on the autosomes and eliminated those individuals with high affinity (--rel-cutoff 0.125) and high heterozygosity (e.g., F coefficients that were greater than 3 standard deviations from the mean)4.

**Development of the PRS model to predict BCS**

**Training and test datasets**

To develop and test the PRS model, we followed best practices to ensure unbiased model performance estimates by developing the models using datasets that were distinct from the datasets that were used to test the model performance2. We randomly split 90% of the samples from the data after quality control and used these as the training data and the remaining 10% were used as the test data. The proportions of case and control samples in the test set and training set were consistent with the dataset before splitting.

**Genome-wide association study (GWAS)**

GWAS are usually used to assess whether a large number of genetic variations, mainly SNPs, are significantly related to traits or diseases5. In this study, a GWAS was conducted using PLINK2 software. For the binary BCS variables, we used logistic regression for model fitting, and the SNPs were considered to be independent of each other. In addition, we performed principal component analysis (PCA) on the SNP data of the training set and extracted the top 10 principal components (PCs) of all samples, as well as their sex and age. These data were used as covariates to eliminate the potential impacts of population stratification and other cryptic relatedness on the association analysis.

**PRS estimations with different methods**

We used the PLINK, LASSOSUM6, BLUP7, and BayesA8 methods to calculate the PRS risk scores, and their descriptions are as follows:

PLINK is a simple additive model that does not require training. This method calculates scores by simply adding the effect values of the SNPs in the GWAS summary data. The effect values are typically reported as log odds ratios (e.g., log(OR)).

LASSOSUM is a method for calculating LASSO/Elastic Net estimates of a linear regression problem given summary statistics from GWAS and genome-wide meta-analyses, which account for linkage disequilibrium (LD) via a reference panel.

BLUP is the Best Linear Unbiased Prediction. The linear mixed model was y = μ + Xb + e, where y was the response variable, μ was the intercept, X were the input features, b was the regression coefficient, and e was the residual coefficient.

BayesA was constructed using the R package bWGR. It assumes that each SNP has an effect and follows a normal distribution and that the effect variance follows a scaled inverted chi-square distribution, χ–2(ν, S), where S is a scale parameter and ν is the number of degrees of freedom. Gibbs sampling is introduced into Markov chain Monte Carlo theory (MCMC) to calculate the effect sizes of the SNPs.

All methods were evaluated using different SNP feature sets. The SNPs were filtered using the GWAS association p values at thresholds of 10−1, 10−2, 10−3, 10−4 and 10−5. These were also tested using the full SNP feature set without any filtering.

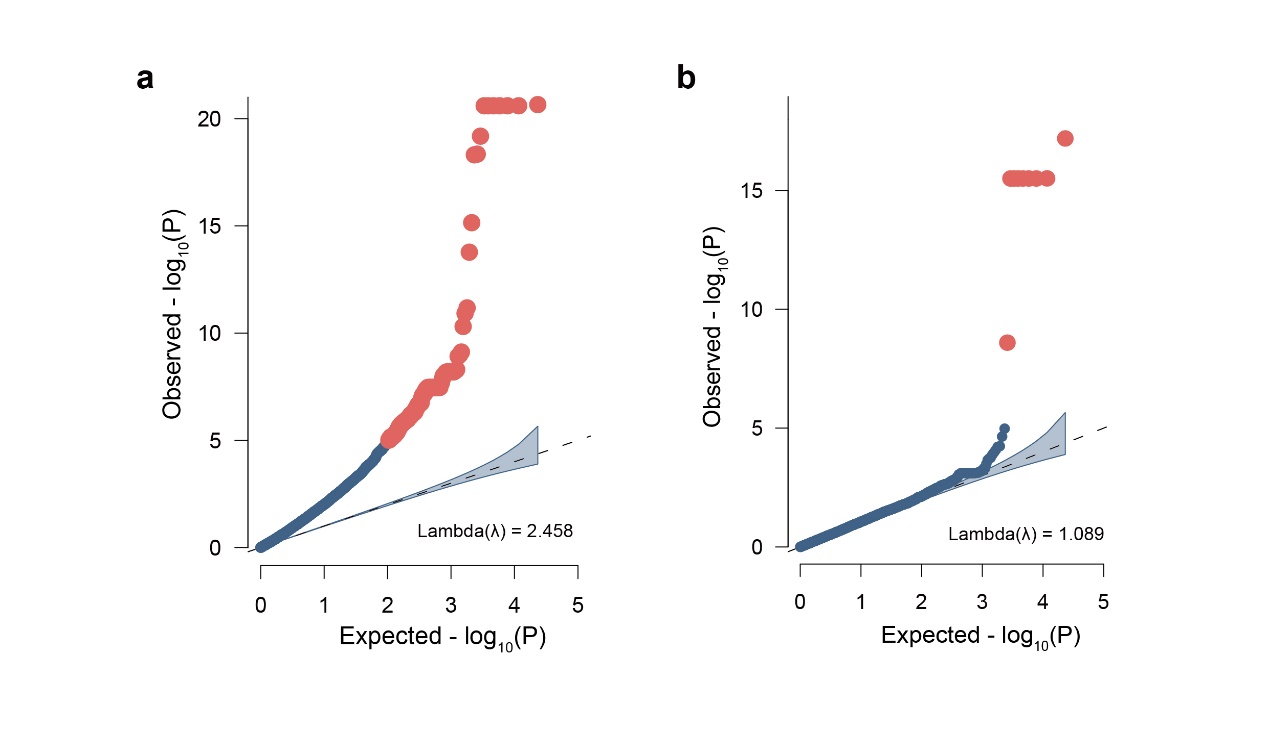
**Model evaluation and interpretation**

Using the risk scores that were provided by each method, the receiver operating characteristic (ROC) curve was drawn, and the area under the ROC curve (AUC) value was calculated. The AUC is currently considered to be the standard method to assess the accuracy of predictive models, with AUC=1 representing perfect performance and 0.50 indicating a random guess. In addition, we calculated the accuracy, sensitivity and other indicators and used these indicators to determine the best prediction method for BCS. Furthermore, we extracted the weight information of the SNPs corresponding to the model and the genes where the SNPs were located. We then used Enrichr9 to perform enrichment analysis on these genes to evaluate their enrichment levels in different gene sets.

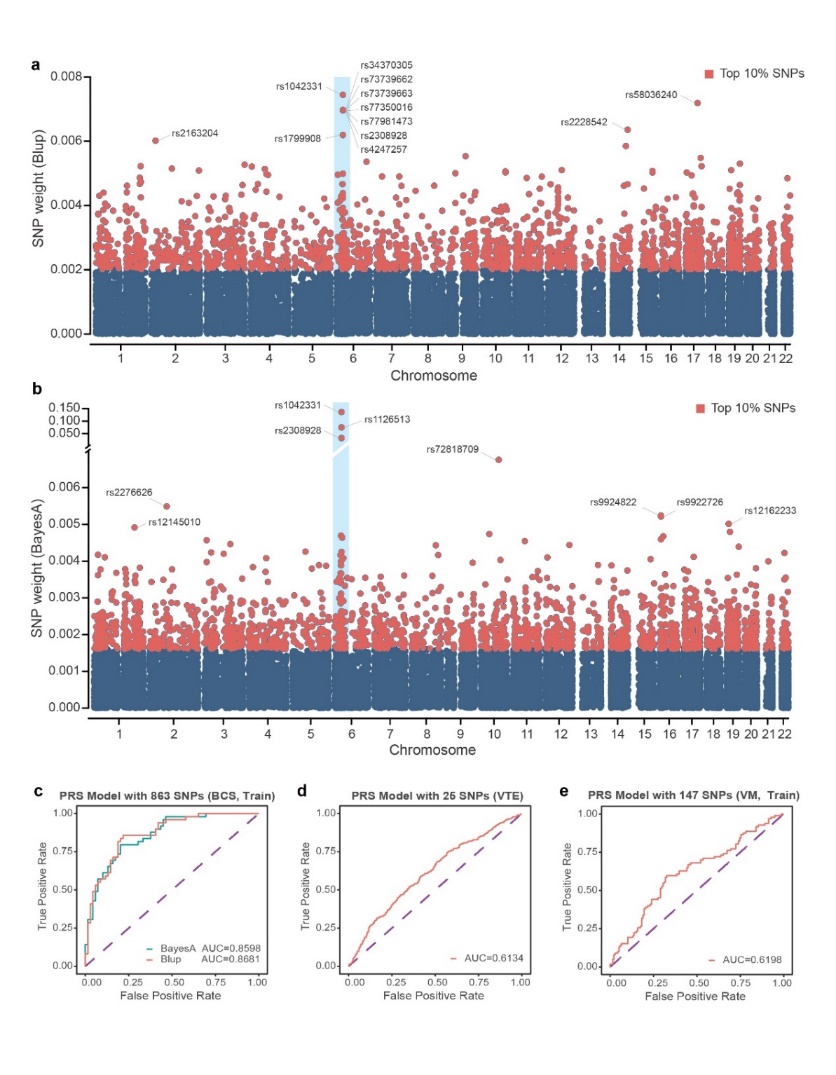
**Relevant disease PRS model construction to predict BCS risk**

According to previous report, we found that some BCS patients had vascular abnormalities10, which are reminiscent of clinically different vascular malformations (VMs) with thrombotic events11. Further research found that they may be caused by defects in the development of blood vessels that are caused by congenital mutations12. Venous thromboembolism (VTE) has also been found to share the same genetic risk factors as BCS, such as Leiden mutation of Factor V, protein C, and protein S13. To explore their genetic commonalities, we attempted to predict the BCS risk based on reported PRS models or genetic risk information for these diseases. By predicting the outcomes, we can assess whether the risk factors for these diseases have risk trends that are consistent for BCS.

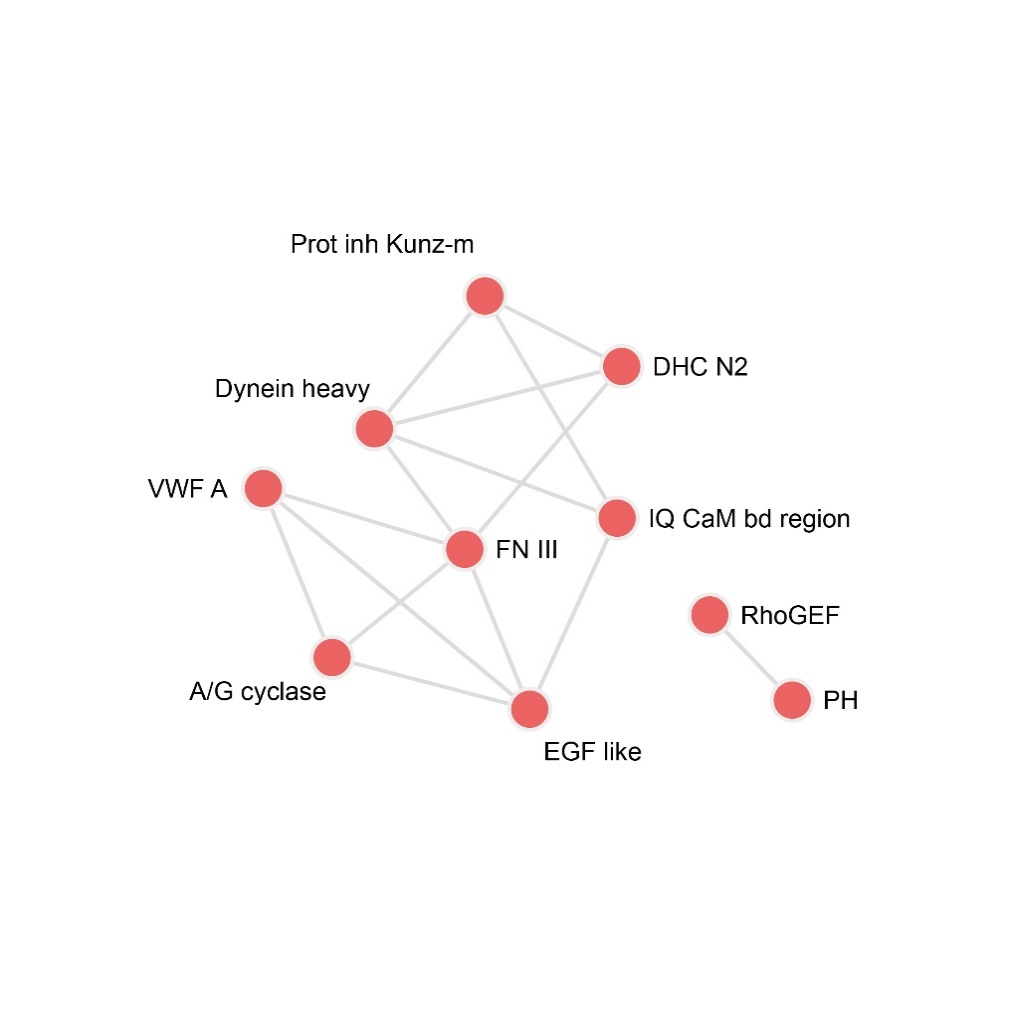
We created 2 PRS models. 1) VTE 25 SNP model: We downloaded the PRS model of VTE (pgs000043) from the PGS catalog project (https://www.pgscatalog.org/). It was mainly constructed based on European individuals and contained 297 variants. These variants intersected with our project data, and a total of 25 variants were retained. We extracted the weights of these variants, constructed a new model and used it to score our 1171 samples. 2) VM 147 SNP model: since no reported PRS model associated with VM was found, we performed an extensive literature search and collected a large number of genes reported to be involved in VM (Supplementary Table 5). A total of 147 SNPs located in these genes were filtered out in the dataset of this project. Training was performed on our training set (n=936) by using the BayesA method, and scoring was performed on the test set (n=235).

****

**Supplementary Figure 1. Quantile–Quantile plot.** (a) QQ plot of BCS before covariates were added. (b) QQ plot of BCS after covariates were added.



**Supplementary Figure 2. PRS model interpretation and genetic commonality analysis.** **a-b)** The weight distribution diagram of the BLUP and BayesA optimal model, the weight is the absolute value of the SNPs effect value in the model. **c)** The AUC result of the BLUP and BayesA model constructed based on the intersection of the top 10% of SNPs in a and b. **d)** The AUC result of VTE 25 SNPs PRS model. **e)** The AUC result of VM 147 SNPs PRS model.



**Supplementary Figure 3.** **The functional association network constructed by Enrichr**

**References**

1. Wei X, Ju X, Yi X, et al. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. *PloS one* 2011;6(12):e29500.

2. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. *Nature protocols* 2010;5(9):1564-73.

3. AlbersCornelis A, DePristoMark A, HandsakerRobert E, et al. The variant call format and VCFtools. *Bioinformatics* 2011

4. Purcell S, Neale B, Todd-Brown K, et al. Ferreira 1500 MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, 1501 Sham PC (2007) PLINK: A tool set for whole-genome 1502 association and population-based linkage analyses. *Am J 1503 Hum Genet*;81:559-75.

5. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *Jama* 2008;299(11):1335-44.

6. Mak S, Porsch R, Choi S, et al. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* 2017;41(6)

7. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 2011;88(1):76-82.

8. Xavier A, Muir WM, Rainey KM. bWGR: Bayesian whole-genome regression: Oxford University Press, 2020.

9. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 2016(W1):W90-W97.

10. Sakamoto N, Koizumi K, Asahina Y, et al. Primary Budd–Chiari syndrome due to complex venous anomalies. *Abdominal imaging* 1997;22(5):499-501.

11. Liu H, Xing X, Yue H, et al. Angiographically occult cerebral vascular malformation mimicking cortical venous sinus thrombosis: case report. *Acta Neurologica Belgica* 2020;120(5):1209-11.

12. Martinez-Corral I, Zhang Y, Petkova M, et al. Blockade of VEGF-C signaling inhibits lymphatic malformations driven by oncogenic PIK3CA mutation. *Nature communications* 2020;11(1):1-14.

13. Zöller B, Svensson PJ, Dahlbäck B, et al. Genetic risk factors for venous thromboembolism. *Expert Review of Hematology* 2020;13(9):971-81.

**Supplementary Table 1.** Summary of sequencing data and comparison of 1196 individual samples collected in the previous period.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | BCS samples | | BCS controls | |
|  | Mean | 25th-75th quantile | Mean | 25th-75th quantile |
| Raw Reads (All reads) | 185476216 | (160659952-206326932) | 171490686 | (153826163-183413069) |
| Mapped Reads | 185039725 | (159884660-205998269) | 171087306 | (153443428-183050974) |
| Fraction of Mapped Reads | 99.76 | (99.72-99.84) | 99.76 | (99.74-99.81) |
| Fraction of duplicated reads | 28.10 | (21.82-32.84) | 28.61 | (22.28-33.96) |
| Target Reads | 132820278 | (113562700-149155994) | 139119912 | (124430916-147809283) |
| Fraction of Target Reads | 71.87 | (70.71-75.63) | 81.07 | (79.65-82.49) |
| Average depth(rmdup) | 129.44 | (108.84-145.92) | 134.93 | (120.59-147.31) |
| Coverage in targeted region (>=30x) | 95.45 | (94.69-96.57) | 95.34 | (94.58-96.2) |
| Average depth in flanking region | 47.55 | (40.05-53.9) | 43.87 | (38.3-47.44) |
| Coverage in flanking region (>=30x) | 48.20 | (42.62-53.21) | 42.56 | (38.13-45.90) |

**Supplementary Table 2.** Comparison of age and gender between cases and controls in the BCS cohort.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Group | Number of samples | Average age | P value | Statistical method | Gender | | P value | Statistical method |
|  | Male | Female |
| Before sample  filtering\* | BCS samples | 500 | 45.43 | 1.66e-45 | Student t test | 212 | 288 | 0.02 | Fisher exact test |
| BCS controls | 696 | 35.35 | 247 | 446 |
| After sample  filtering | BCS samples | 483 | 45.59 | 1.23e-46 | Student t test | 207 | 276 | 0.01 | Fisher exact test |
| BCS controls | 688 | 35.33 | 244 | 444 |

\* In the data set before filtering, there are three samples in the BCS control group without clear gender information.

**Supplementary Table 3** - GWAS analysis result.txt

**Supplementary Table 4** - Enrichr analysis result.xlsx

**Supplementary Table 5.** the list of known genes associated with vascular malformation

|  |  |  |  |
| --- | --- | --- | --- |
| Gene symbol | | | |
| *MAP3K3* | *ENG* | *NAGA* | *VEGFB* |
| *ET* | *EPHB4* | *NEMO* | *VEGFR3* |
| *BMP9* | *EPO* | *NOTCH1* | *VEGFR2* |
| *MADH4* | *F5* | *PCNA* | *VEGFC* |
| *ALK1* | *FAM111B* | *PDCD10* | *VG5Q* |
| *ACE* | *FLT1* | *PIEZO1* | *RNF213* |
| *AKT1* | *FLT4* | *PIK3CA* | *NOTCH3* |
| *ACTA2* | *FLVCR2* | *PIK3R2* | *HTRA1* |
| *ACVRL1* | *FOXF1* | *PON1* | *VEGF* |
| *ANGPT1* | *FOXC2* | *PTEN* | *TGFBR1* |
| *ANTXR1* | *GDF2* | *RASA1* | *TGFBR2* |
| *ARHGAP31* | *GLMN* | *SLC7A14* | *ACVLR1* |
| *ATM* | *GNAQ* | *SLC2A10* | *IL1b* |
| *ATR* | *GLOMULIN* | *SMAD4* | *ITGB8* |
| *BMPR2* | *GUCY1A1* | *SNORD118* | *ANGPTL4* |
| *CCM2* | *IKBKG* | *SNRK* |  |
| *CCM1* | *IGFBP7* | *SOX18* |  |
| *CCM3* | *IL1RN* | *STAMBP* |  |
| *CCND2* | *IL6* | *STN1* |  |
| *CLEC14A* | *IDH1* | *TALDO1* |  |
| *COL4A1* | *IDH2* | *TEK* |  |
| *CTC1* | *JAK2* | *TIE1* |  |
| *CTD* | *KDR* | *TIE2* |  |
| *DLL4* | *KRAS* | *TMEM173* |  |
| *DOCK6* | *KRIT1* | *TREX1* |  |
| *DUSP5* | *LBR* | *VHL* |  |
| *EIF2AK4* | *MGC4607* | *VASN* |  |
| *ELMO2* | *MRE11* | *VEGFA* |  |

**Supplementary Table 6** - bayesA\_snp\_weight.txt

**Supplementary Table 7** - blup\_snp\_weight.txt