Comprehensive assessments of germline deletion structural variants reveal the association between prognostic MUC4 and CEP72 deletions and immune response gene expression in colorectal cancer patients

Lin Peng-Chan
National Cheng Kung University https://orcid.org/0000-0002-9424-1985
Hui-O Chen
National Cheng Kung University
Chih-Jung Lee
National Cheng Kung University
Yu-Min Yeh
National Cheng Kung University
Meng-Ru Shen
National Cheng Kung University
Jung-Hsien Chiang (jchiang@mail.ncku.edu.tw)
National Cheng Kung University

Primary research

Keywords: Whole genome sequencing, Cancer risk, Deletion structural variants, MUC4, CEP72

Posted Date: September 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-79420/v1

Version of Record: A version of this preprint was published on January 11th, 2021. See the published version at https://doi.org/10.1186/s40246-020-00302-3.

Comprehensive assessments of germline deletion structural variants reveal the association between prognostic MUC4 and CEP72 deletions and immune response gene expression in colorectal cancer patients

Peng-Chan Lin[1,2,3,4], Hui-O Chen[1] , Chih-Jung Lee[1] , Yu-Min Yeh[3,4] , Meng-Ru Shen[5,6,7] , Jung-Hsien Chiang[1,2]

[1]Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan, Taiwan.

[2]Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan.

[3]Department of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

[4]Department of Internal medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

[5]Graduate Institute of Clinical Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

[6]Department of Obstetrics and Gynecology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Taiwan.

[7]Department of Pharmacology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Taiwan.

Correspondence:

Jung-Hsien Chiang, Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Cheng Kung University, Tainan, Taiwan

Tel: +886-6-2757575, Ext. 62534

E-mail: jchiang@mail.ncku.edu.tw

Abstract

Background

Functional disruptions by large germline genomic structural variants in susceptible genes are known risks for cancer. Few studies have used deletion structural variants (DSVs) to predict cancer risk with neural networks or studied the relationship between DSVs and immune gene expression to stratify prognosis.

Methods

Whole-genome sequencing (WGS) data was analyzed with the blood samples of 192 cancer and 499 noncancer subjects with or without family cancer history (FCH). Ninety-nine colorectal cancer (CRC) patients had immune response gene expression data. To build the cancer risk predictive model and identify DSVs in familial cancer, we used joint calling tools and attention-weighted model. The survival support vector machine (survival-SVM) was used to select prognostic DSVs.

Results

We identified 671 DSVs that could predict cancer risk. The area under the curve (AUC) of receiver operating characteristic curve (ROC) of attention-weighted model was 0.71. The 3 most frequent DSV genes observed in cancer patients were identified as ADCY9, AURKAPS1, and RAB3GAP2 ($p < 0.05$). We identified 65 immune-

associated DSV markers for assessing cancer prognosis ($P < 0.05$). The functional protein of MUC4 DSV gene interacted with MAGE1expresssion, according to the STRING database. The causal inference model showed that deleting the CEP72 DSV gene could affect the recurrence-free survival (RFS) of IFIT1 expression.

Conclusions

We established an explainable attention-weighted model for cancer risk prediction and used the survival-SVM for prognostic stratification by using DSV and immune gene expression datasets. It can provide the genetic landscape of cancer patients and help predict the clinical outcome.

Keywords: Whole genome sequencing, Cancer risk, Deletion structural variants, MUC4, CEP72

Introduction

Large-scale germline structural variants, especially deletion structural variants (DSVs), can affect gene expression with a partial or complete loss of gene function and increased risk of cancer in patients. Traditional genome-wide association studies (GWAS) have revealed the association of numerous single nucleotide polymorphism (SNP) traits and germline risk[1, 2]. However, several studies have highlighted the biases and inaccuracies of the polygenic risk score (PRS), a sum of cancer-associated alleles, when predicting cancer risk in individuals from different populations[3]. Linkage disequilibrium and variant frequencies in different ethnicities play an important role[4]. Most SNPs, except for high or moderated penetrance inherited cancer susceptibility genes such as *BRCA* or *MLH* in Mendelian hereditary cancer syndromes[5], have no effect on health or cancer development.

For complex cancer traits, one gene with SNPs is not sufficient to cause cancer, but rather multiple genes with non-Mendelian inheritance are needed[6]. The WGS are increasingly being used to establish the mutation landscape of cancer risk as well as detect DSVs and protein-coding mutations. Instead of SNPs, multiple cancer-associated DSVs have become more widely used for cancer risk assessment[7, 8]. However, the role of DSVs in germline risk and cancer prognosis has not been sufficiently understood in the general Asian population.

Prediction models are important when classifying individuals for predicting risk and survival stratification in order to minimize the impact of cancer and optimize treatment[9]. The application of machine learning techniques, such as deep learning (DL) and inherited risk genomic variation analysis, is rapidly developing[6, 10]. As DL has improved the ability to predict inherited cancer genomic susceptibility, we focused on DL as an attention-weighted model with multilayer perceptrons (MLPs)[11], which can reveal the importance of each DSV for predicting cancer risk. Additionally, we used the survival support vector machine (survival-SVM) for selecting the features of prognostic DSVs.

In this study, we will describe the prediction of germline cancer risk with DSV detection and the impact of DSVs in cancer patients with and without family cancer history (FCH). For assessing prognosis, we used a machine leaning model for survival stratification and demonstrated the biological relevance of germline DSVs and the expression of tumor microenvironment immune response genes.

Results

Study design and workflow

To develop the risk and prognostic stratification model, we collected genomic and clinical information, including FCH, such as survival and FCH, from 192 cancer patients at National Cheng Kung University Hospital (NCKUH) and 499 normal subjects without cancer in the Taiwan Biobank[12] with four aims. First, we aimed to build the cancer risk prediction model with germline DSVs. Second, we studied the spectrum and frequency of DSV genes in cancer patients with or without FCH. Third, we aimed to observe whether genes with DSVs would impact the immune response gene expression within tumor microenvironment. Fourth, we stratified the cancer patients' clinical outcomes by immune-related DSVs and investigated the relationship and biological relevance of the DSVs. Fig. 1 shows the overall workflow of this study. We applied feature extraction and selection methods to analyze genomics data for the detection of cancer-associated, immune-associated, and prognosis-associated DSVs. We utilized the PopDel[13] tool to detect germline DSVs. The WGS data of cancer patients and noncancer subjects were input simultaneously for joint calling. A total of 14,772 autosomal DSVs with sizes ranging from 500 to 10,000 base pairs were called simultaneously across all samples. We focused our analysis on DSVs occurring in at least 1% of the samples of both cancer and noncancer populations at minor allele

frequency (MAF) above 5%[14]. A total of 2,919 DSVs that passed the filtering criteria were further used to build a classification model.

Predicting cancer risk with whole genome DSVs and MLP

Germline genomic DSVs are known to be associated with increased risk for cancer[4], and several studies have reportedly applied machine learning tools for developing prediction models[15]. To learn the importance of each DSV for classifying cancer or noncancer samples, we consider the attention-weighted model to be the final approach. Furthermore, the attention-weighted model had the best performance to predict cancer risk. Herein, several machine learning strategies for classification were applied and evaluated. We used an SVM with linear kernel and logistic regression (LR), both of which were well-known linear models. We also used random forests (RF) to test nonlinear results. Moreover, neural network strategies, such as multilayer perceptron and attention-weighted models, were also adopted (Fig. 2A).

The area under the curve (AUC) from the receiver operating characteristic curve (ROC) and the performance of models (i.e. sensitivity) are crucial for clinical use. Among these methods, the attention-weighted model (AUC = 0.71, sensitivity = 0.58) performed the best with 2,919 DSVs (Fig. 2B and Supplementary Fig. 1). All of the model's performance were improved with 671 cancer-associated DSVs. In total, 671 of 2,919 significant DSVs were selected for the prediction of cancer risk with positive weights

from the attention-weighted model (Supplementary Table 1A). There were no demographic biases in the population data (Supplementary Fig. 2). The size and distribution of deletions on each chromosome were no different between cancer patients and noncancer subjects (Supplementary Fig. 3). The cancer and non-cancer samples could be distinguished with 671 DSVs in principal components analysis (PCA), however, there were no differences between each type of cancer (Supplementary Fig. 4A). Further analyses were conducted to determine the genes with DSVs and relevant pathways related to different cancer types. The result indicated that the DSVs in SNTG2, PCMT1, DACT2, CBX3, ATP11A, and SHC2 were associated with breast cancer. DSVs in SGSM2 and LHFPL3 were relevant to colorectal cancer. Whereas ADAP1, DLGAP2, ERC1 and PPP6R2 were related to gynecologic cancer (Supplementary Fig. 4B and C).

The mutational landscape of DSVs and their significance in familial cancer patients

The germline mutational landscape of DSVs plays an important role in cancer patients with or without FCH. Patients who have one or more blood relatives within third-degree suffering from any types of cancer are considered having family cancer history. The odds ratios were estimated to identify which genes with DSVs were associated with FCH. We chose the top 10 DSV genes associated with an increased risk

of cancer (odds ratios (OD) >1) and the bottom 10 DSV genes associated with a decreased risk of cancer (OD <1) from 671 DSVs associated with cancer risk (Fig. 3A). The top 10 genes frequently observed in cancer patients were *ADCY9*, *RAB3GAP2*, *AURKAPS1*, *EYS*, *SHC2*, *DPP6*, *FREM2*, *ESR1*, *TBC1D22A*, and *ACTN2*. The ten genes frequently observed in noncancer subjects were *SNTG2*, *LHFPL3*, *DACT2*, *NKAIN2*, *KALRN*, *ABR*, *LMNTD1*, *PLEKHA7*, *DOC2B*, and *ADPRHL1* (Supplementary Table 2). We also studied the prevalence and spectrum of well-known germline cancer susceptibility genes in our subjects[16]. The frequencies of 26 cancer susceptibility genes are shown in Fig. 3B. Deletions in the *FANCA*, *POLD1*, and *STK11* genes were observed in cancer patients only. The frequency of gene deletions was almost the same between cancer and noncancer subjects. The mutational landscape of DSV genes is shown in Fig. 3C. There were 57 cancer-associated DSV genes with a P-value < 0.05 in the cancer and noncancer groups (Supplementary Table 3).

In this study, we found a higher incidence of FCH in cancer patients than in noncancer subjects (Fig. 4A). Moreover, certain DSV genes were associated with cancer or noncancer subjects with or without FCH (Fig. 4B and C). *MGAT4C*, *HSPA4L*, *ZSCAN5A*, *LOC100505841*, and *NALCN* gene deletions were associated with cancer patients without FCH ($p<0.05$), while *SMYD3* and *NKD2* DSV genes were associated with cancer patients with FCH ($p<0.05$). *HHIPL2*, *XPO1*, *SALRNA1*, *ZBTB45*,

ANP32AP1, *ACTR3BP5*, *LOC100129138*, *GPR45*, and *CAB39L* gene deletions were associated with noncancer objects without FCH ($p<0.05$), while *RAB9BP1*, *LOC101928523*, and *MALRD1* gene deletions were related to noncancer patients with FCH ($p<0.05$). Consequently, we inferred that subjects with FCH carrying *SMYD3* or *NKD2* gene deletions may have a higher cancer incidence. As illustrated in Fig. 4D, the volcano plot shows eight significant DSV genes based on the Cox's proportional hazards model for survival analysis (Supplementary Table 4).

The clinical impact of immune gene expression-related DSVs in colorectal cancer patients

The host immune system differentially participate in the tumor microenvironment. Cancer often develops because of the immune system disturbance caused and functional disorder. The germline DSVs influence aberrant gene expression in tumors.[17] Therefore, we studied the functions associated with 160 immune gene expression-associated DSVs with correlation coefficients of > 0.3, which were selected based on the point-biserial correlation to understand the clinical impact of their deletions (Supplementary Table 1B). There are six categories of immune gene functions: housekeeping, checkpoint pathways, cytokine signaling, lymphocyte markers, lymphocyte regulation, and tumor characterization. A total of 57 DSV genes were correlated with the six functional immune response categories; the *PTPRN2* gene

deletion had the highest frequency (Fig. 5A and Supplementary Table 5). *STNG2* and *LOC105376360* gene deletions (Fig. 5A) were related to lymphocyte regulation and housekeeping (p-adjusted value less than 0.05), while *CEP72* and *ZZEF1* gene deletions had high occurrences in the housekeeping and cytokine signaling categories, respectively (Fig. 5A).

We selected 65 prognosis-associated DSVs by using survival support vector machine (survival-SVM)[18], which had the highest predicted score for survival SVM. We used 65 prognosis-associated DSVs among 160 immune-associated DSV genes and constructed a heatmap (Fig. 5B and Supplementary Table 1C). These prognosis-associated DSVs were grouped into poor (33 recurrence-associated DSVs) and better (32 nonrecurrence-associated DSVs) prognostic groups using Cox's proportional hazards model. There were more poor prognostic deletions in the tumor characterization functional category (e.g. *MUC4* and *PTPRN2* gene deletions) and better prognostic deletions in the lymphocyte regulation functional category (Fig. 5B). We then stratified the patients into two groups by prognostic deletions that have different clinical outcomes. Group 1(G1) was the patient who has more recurrence-associated DSVs than nonrecurrence-associated DSVs. According to the Kaplan-Meier curve, these patients in G1 experienced a poor clinical outcome ($p<0.05$) (Fig. 5C). Patients in group 2 (G2) had better outcomes whose nonrecurrence-associated DSVs

are more than recurrence-associated DSVs.

The biological relevance of germline DSVs and tumor microenvironment immune genes

The tumor microenvironment can affect prognosis and shape therapeutic resistance[19]. Overexpression of the immune *MAGEA1* gene, a member of the *MAGEA* gene family, in tumor and stromal cells is associated with a poor prognosis and an ideal candidate for tumor immunotherapy[20, 21], *MAGE1* was highly expressed in a previous study on colorectal cancer[21]. In our data, we showed that colorectal cancer patients with germline *MUC4* gene deletion experienced a poor clinical outcome (Fig. 6A). Seven of 13 patients with a germline *MUC4* gene deletion experienced recurrence. Moreover, the *MUC4* gene deletion was positively correlated with *MAGE1* expression, which indicated that SV deletion resulted in increased *MAGE1* expression (Fig. 6A). With the use of the STRING database[22], we also demonstrated protein-protein interactions between the transmembrane mucin family, including *MUC4* and *MAGE1* (Fig. 6B). The functional protein association networks indicated that the MUC4 gene deletion might influence the expression of *MAGE1*. We hypothesized that germline DSVs could affect immune MAGEA1 expression and correlate with a poor prognosis.

Here, we also showed that eight prognostic DSVs can affect RFS by expressing tumor microenvironment immune genes. In our cohort, the eight prognostic deletions were

correlated with immune gene expression and survival in colorectal cancer stage III patients (Supplementary Fig. 5 and 6). To understand the cause-effect relationship of this result, we applied causal modeling and implemented the PC algorithm by R package CompareCausalNetworks[23]. The PC algorithm uses conditional independence tests for model selection in graphical modeling with directed acyclic graphs[24]. Our results showed that deletion of the oncogene *CEP72* could affect RFS by *IFIT1* immune expression. *IFIT1* is an abundant product of interferon-stimulating genes that correlates with a poor prognosis in cancer[25].

In this study, we demonstrated the possible biological relevance of the *MUC4* gene deletion and *MAGE1* expression and found the causal relationships among *CEP72* gene deletion, *IFIT1*, and RFS (Fig. 6D). These results indicate that germline DSVs might affect prognosis by expressing tumor microenvironment immune genes.

Discussion

Advances in machine learning technologies have led to the use of deep learning prediction models for cancer prevention. Here, we applied WGS of germline DSVs for predicting cancer risk and machine learning methods for assessing immune-related prognosis. Our results highlighted the following: (i) a cancer risk predictive model was established with 671 DSVs and an attention-weighted neural network; (ii) potential markers for inherited cancer risk were identified in cancer patients with or without FCH; (iii) 57 DSVs were correlated with six immune functional categories; (iv) 65 prognostic deletions were identified in order to construct a survival model for clinical outcome stratification; and (v) the possible mechanisms and biological relevance of 2 germline deletions in the expression of two immune genes were presented. Germline WGS and immune gene expression profiling are excellent tools for predicting cancer and stratifying prognosis in colorectal cancer patients.

Traditionally, a small subset of gene alteration features that could predict and classify types of cancer were selected by different machine learning models[26]. However, gene-gene interactions can significantly complicate the search for disease-associated genes. Genes play various essential roles in cancer biology, and each gene carries a different weight importance in the clinical outcome. Deep learning can employ an automatic weight learning feature that can allow complex predictions. In this study, we

built a deep learning classification model to identify unique biological features that can differentiate between cancer and noncancer subjects. Using population-based designs, we identified 671 DSVs associated with the risk of cancer. We found that PCA could distinguish between cancer and noncancer subjects using these 671 DSVs.

Many hereditary cancer syndromes have now been defined and attributed to specific germline inherited mutations. Cancer development is related to accumulating genetic alterations. In this study, we studied the evolution pattern of DSVs in cancer patients with or without FCH. We found that subjects with FCH had a higher incidence of developing cancer and may have initially inherited three DSV genes, namely, *MALRD1*, *LOC101928523*, and *RAB9BP1*. They developed cancer after acquiring two DSV genes: *NKD2* and *SMYD3*. However, patients without FCH may have a different evolution pattern of DSVs. Initially, they inherited nine DSV genes— *CAB39L*, *GPR45*, *LOC1001291138*, *ACTR3BP5*, *ANP32AP1*, *ZBTB45*, *SALRNA1*, *XPO1*, and *HHIPL2*—and developed cancer after acquiring five DSV genes— *MGAT4A*, *HSPA4L*, *ZSCAN5A*, *LOC100505841*, and *NALCN*. We focused on eight signaling pathways associated with the aforementioned DSV genes[27]. The most significant pathway enriched with DSV genes for subjects with FCH was metabolic regulation while for subjects without FCH was transport regulation. These results imply that subjects with

or without FCH may develop cancer through different signaling pathways. These DSVs may become useful screening markers.

The result from each classification was the average after five-fold cross validation. The 192 cancer patients and 499 noncancer samples data were divided into a training set and testing set. We randomly chose 80% samples as the training data and 20% samples as the testing set in each fold. Due to the DSVs analysis was started with BAM file and lack of samples, there was no other public data can be used as validation data.

Genetic alterations from nature vs nurture: What determines cancer risk and prognosis? We hypothesized that germline DSVs mold the tumor microenvironment and immune gene expression, impacting the clinical outcome. In this study, we wanted to examine the correlation of germline deletions and immune response genes to understand the potential mechanisms by which the tumor microenvironment can affect clinical outcomes[28]. We classified germline structural deletions by the expression of tumor microenvironment-based immune response-associated genes. There were significantly poorer prognostic deletions in the tumor characterization category and better prognostic deletions in the lymphocyte regulation category. Eight prognostic deletions associated with immune gene expression were identified, including *HGF*, *CDKN2A*, and *ITGB1*. They were also reported as poor prognostic factors in a previous study[29].

Beyond the traditional signaling factor statistical survival model, we used the survival-SVM and Cox's proportional hazards model to select 65 prognostic deletions. We proposed a method to classify risk and nonrisk groups by prognostic deletions and identified 57 prognostic DSVs as possible markers for survival stratification and prognosis assessment. From the bioinformatics database and casual inference model, we also demonstrated that immune-associated gene expression may influence the clinical outcome of some germline deletions. The possible mechanisms which affects tumor microenvironment survival was shown, but further molecular validation is needed.

Conclusions

In conclusion, we used genomic data, including WGS and immune gene expression data, and two explainable machine learning models to establish cancer risk predictive models and a prognosis assessment tool that could be useful for cancer prevention and potential therapeutic strategies.

Materials and methods

Enrollment of cancer patients and noncancer healthy subjects

A total of 192 cancer patients, including eight with breast cancer, 120 with colorectal cancer, 29 with endometrial cancer, and 35 with ovarian cancer, were recruited for the study at the NCKUH between January 2015 and January 2017. Follow-up continued through October 2018. Clinical information (detailed family cancer history (FCH)), tissue, and blood samples for DNA extraction and WGS were collected at the time of enrollment. The NCKUH institutional review board approved this study (A-ER-103-395 and A-ER-104-153), and all participants provided written informed consent. WGS, health, and lifestyle data of 499 noncancer Taiwanese people were obtained from the Taiwan Biobank as reference (Fig. 1). Of all 99 CRC patients, the distribution of gender was almost the same. The median age of these patients was 58 years. The prevalent primary tumor site was left colon (80.8%). Family cancer history and recurrence were not significant different. There was no significant difference between recurrence and tumor characteristics, such as tumor site, tumor invasion stage (T) or nodal stage (N) (Supplementary table 6).

Germline WGS

Genomic DNA from collected blood samples was quantified with a Qubit fluorescence assay (Thermo Fisher Scientific) and sheared with an S2 instrument (Covaris). Library

preparation was carried out using the TruSeq DNA PCR-Free HT Kit (Illumina). Individual DNA libraries were measured with 2100 Bioanalyzer (Agilent) qPCR and Qubit (Thermo Fisher Scientific). All flow cells were sequenced on a HiSeq 2500 sequencer (Illumina) using SBS kit V4 chemistry (Illumina). FastQC was used to check read quality, and the resulting reads were aligned to the hg19 reference genome with the BWA-MEM algorithm[30]. The identification of SNPs and indels and genotyping were performed across all samples simultaneously using standard hard filtering parameters or variant quality score recalibration according to GATK Best Practices recommendations[31]. WGS was performed with a minimum, median coverage of 30X.

Immune response gene expression data

Cancer tissues with immune response gene expression profile data were obtained from 99 colorectal cancer patients. RNA was prepared from formalin-fixed paraffin-embedded (FFPE) tissue that was extracted with the RecoverAll Total Nucleic Acid Isolation Kit (Thermo Fisher Scientific). RNA concentration was determined on an Invitrogen™ Qubit™ Fluorometer with the Qubit™ RNA High Sensitivity Assay (Thermo Fisher Scientific). Twenty nanograms of RNA was used for each reverse transcription reaction, and cDNA was prepared with the SuperScript™ IV VILO™ Master Mix Kit. Immune response libraries were prepared using the Ion AmpliSeq™ Kit for Chef DL8 with the Ion Chef™ System and according to instructions in the

Oncomine™ Immune Response Research Assay user guide (Pub. No. MAN0015867). The raw gene expression data were preprocessed using Torrent Suite (Thermo Fisher Scientific) and normalized with the min-max feature scaling approach.

Statistical analysis

The chi-square test and Fisher's exact test were used to assess the differences between groups. Kaplan–Meier curves were used to evaluate RFS, which was defined as the time between surgery and cancer recurrence. A P-value < 0.05 was considered statistically significant.

Machine learning model and analysis

Detecting DSVs and data preprocessing

We detected germline DSVs in cancer and noncancer subjects simultaneously with PopDel from whole genome DNA sequencing data[13]. DSVs were then filtered by the minor allele frequency (MAF). A MAF greater than or equal to 0.05 and occurring in at least 1% of the sample in each population were subjected to further analysis.

Selecting DSVs for the cancer risk and immune expression correlation model

We designed an attention-weighted model[32] to select important DSVs (Fig. 2A). This model is a MLP model based on the attention mechanism. During the learning process, the model automatically adjusts the weight of every DSV. The main aim of this model is to predict subjects with or without cancer. We used the deletion vector for each

sample as the input of the attention-weighted model and then adopted binary cross-entropy as a loss function. After training the model, we obtained the weight of each DSV. We then selected cancer risk associated DSVs with positive weights, which are important when classifying cancer and noncancer samples. We correlated cancer risk associated DSVs and immune gene expression data from 99 colorectal cancer patients. The gene expression data were normalized. An immune expression correlation table was established with the point-biserial correlation[33], which was used to correlate continuous variables with dichotomous variables, to determine the relationship between DSVs and immune gene expression.

Prognostic candidate genes and survival stratification

There are many survival analysis that using the machine learning approach to achieve predicted results, especially survival-SVM[18] can have better results. We can also know the importance of each DSVs to the model and it can also be more interpretable. We selected prognosis-associated candidate DSVs by using the survival-SVM[18], which is the approach that can be used to predict the event time duration based on a given set of features. Therefore, we do feature selection base on the survival-SVM, which can select the most predictive prognosis associated DSVs in model. The candidate DSVs were clustered into two groups: the recurrence-associated DSV group and the nonrecurrence-associated DSV group. We measured the hazard ratio (HR) of

each candidate deletion using Cox's proportional hazards model, which represents the probability of recurrence by giving the survival time of patients. We determined that DSVs with a positive log (hazard ratio) were recurrence-associated deletions, while DSVs with a negative log were nonrecurrence-associated deletions. The prognostic DSVs were selected with statistical significance in the hazard model. We used the Kaplan–Meier method for survival analysis to compare the differences between the two survival curves using the log-rank test[34].

Acknowledgments

The authors gratefully acknowledge the significant contribution of LY Hung for their helpful critiques and suggestions. We also acknowledge the essential work of the Taiwan Biobank in collecting the population-based genome data.

Authors' contributors

Conception and study design: PC Lin, HO Chen, JH Chiang; Development of methodology: PC Lin, CJ Lee, JH Chiang; Acquisition of data: PC Lin, YM Yeh, MR Shen; Statistical and computational analysis: PC Lin, HO Chen, CJ Lee, JH Chiang; Writing, review, and/or revision of the manuscript: PC Lin, HO Chen, YM Yeh, MR Shen, JH Chiang; Study supervision: JH Chiang; All authors have read and approved the manuscript. All authors agree for publication

Availability of data and materials

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Ethics approval and consent to participate

This study was approved by the institutional review board of National Cheng Kung University Hospital (NCKUH) (A-ER-103-395 and A-ER-104-153) and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent.

Funding

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan under Research Grant of MOST-108-2634-F-006-006 and MOST 108-2634-F-006-011 and Ministry of Health and Welfare (MOHW108-TDU-B-211-124018 and MOHW108-TDU-B-211-133003). All authors have read and approved the manuscript

Competing interests

The authors declare no competing interests.

Consent for publication

Not applicable

Provenance and peer review

Not commissioned; externally peer reviewed.

Reference

1. ParkS, SupekF, LehnerB. Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. Nat Commun. 2018;9.

2. BunielloA, MacarthurJAL, CerezoM, HarrisLW, HayhurstJ, MalangoneC, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.

3. DeLa VegaFM, BustamanteCD. Polygenic risk scores: A biased prediction? Genome Med. 2018;10.

4. DuncanL, ShenH, GelayeB, MeijsenJ, ResslerK, FeldmanM, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10.

5. TsaousisGN, PapadopoulouE, ApessosA, AgiannitopoulosK, PepeG, KampouriS, et al. Analysis of hereditary cancer syndromes by using a panel of genes: Novel and multiple pathogenic mutations. BMC Cancer. 2019;19.

6. KimBJ, KimSH. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. Proc Natl Acad Sci U S A. 2018;115:1322–7.

7. Escala-GarciaM, GuoQ, DörkT, CanisiusS, KeemanR, DennisJ, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. Br J Cancer.

2019;120:647–57.

8. TenesaA, DunlopMG. New insights into the aetiology of colorectal cancer from genome-wide association studies. Nature Reviews Genetics. 2009;10:353–8.

9. WangX, OldaniMJ, ZhaoX, HuangX, QianD. A review of cancer risk prediction models with genetic variants. Cancer Informatics. 2014;13:19–28.

10. EraslanG, AvsecŽ, GagneurJ, TheisFJ. Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics. 2019;20:389–403.

11. CruzJA, WishartDS. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics. 2006;2:59–77.

12. ChenCH, YangJH, ChiangCWK, HsiungCN, WuPE, ChangLC, et al. Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. Hum Mol Genet. 2016;25.

13. KehrB, HelgadottirA, MelstedP, JonssonH, HelgasonH, JonasdottirA, et al. Diversity in non-repetitive human sequences not found in the reference genome. Nat Genet. 2017;49:588–93.

14. AudanoPA, SulovariA, Graves-LindsayTA, CantsilierisS, SorensenM, WelchAME, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176:663-675.e19.

15. GuoX, ShiJ, CaiQ, ShuXO, HeJ, WenW, et al. Use of deep whole-genome

sequencing data to identify structure risk variants in breast cancer susceptibility genes. Hum Mol Genet. 2018;27:853–9.

16. ZhangJ, WalshMF, WuG, EdmonsonMN, GruberTA, EastonJ, et al. Germline mutations in predisposition genes in pediatric cancer. N Engl J Med. 2015;373:2336–46.

17. HanahanD, WeinbergRA. Hallmarks of cancer: The next generation. Cell. 2011;144:646–74.

18. PölsterlS, NavabN, KatouzianA. Fast training of support vector machines for survival analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. p. 243–59.

19. BinnewiesM, RobertsEW, KerstenK, ChanV, FearonDF, MeradM, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018;24:541–50.

20. FanipakdelA, Seilanian ToussiM, RezazadehF, Mohamadian RoshanN, JavadiniaSA. Overexpression of cancer-testis antigen melanoma-associated antigen A1 in lung cancer: A novel biomarker for prognosis, and a possible target for immunotherapy. J Cell Physiol. 2019;234:12080–6.

21. MaoY, TangQ, FanW, TangX, XuL, ZhuJ, et al. A novel MAGE-A1-IgG antibody for lung adenocarcinoma. J Clin Oncol. 2017;35 15_suppl:e20085–e20085.

22. FranceschiniA, SzklarczykD, FrankildS, KuhnM, SimonovicM, RothA, et al. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41.

23. Heinze-DemlC, MaathuisMH, MeinshausenN. Causal Structure Learning. Annu Rev Stat Its Appl. 2018;5:371–91.

24. SpirtesP, GlymourC, ScheinesR. Causation, prediction, and search, 2nd edition. 2000.

25. PiduguVK, WuMM, YenAH, PiduguHB, ChangKW, LiuCJ, et al. IFIT1 and IFIT3 promote oral squamous cell carcinoma metastasis and contribute to the anti-tumor effect of gefitinib via enhancing p-EGFR recycling. Oncogene. 2019;38:3232–47.

26. SohKP, SzczurekE, SakoparnigT, BeerenwinkelN. Predicting cancer type from tumour DNA signatures. Genome Med. 2017;9:1–11.

27. AshburnerM, BallCA, BlakeJA, BotsteinD, ButlerH, CherryJM, et al. Gene ontology: Tool for the unification of biology. Nature Genetics. 2000;25:25–9.

28. YuH, KortylewskiM, PardollD. Crosstalk between cancer and immune cells: Role of STAT3 in the tumour microenvironment. Nature Reviews Immunology. 2007;7:41–51.

29. KondouR, IizukaA, NonomuraC, MiyataH, AshizawaT, NagashimaT, et al. Classification of tumor microenvironment immune types based on immune response-

associated gene expression. Int J Oncol. 2019;54:219–28.

30. LiH. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;00:1–3. http://arxiv.org/abs/1303.3997.

31. DepristoMA, BanksE, PoplinR, GarimellaKV., MaguireJR, HartlC, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–501.

32. VaswaniA, ShazeerN, ParmarN, UszkoreitJ, JonesL, GomezAN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Neural information processing systems foundation; 2017. p. 5999–6009.

33. KornbrotD. Point Biserial Correlation. In: Wiley StatsRef: Statistics Reference Online. 2014.

34. BewickV, CheekL, BallJ. Statistics review 12: Survival analysis. Critical Care. 2004.

Figure Legends

Fig. 1. Study design and workflow. Study design and overall workflow of WGS analysis of germline DSVs and immune gene expression for cancer risk prediction and survival stratification. In total, 192 cancer patients **(i)**—comprised of 120 with colorectal cancer, 29 with endometrial cancer, 35 with ovarian cancer, and eight with breast cancer—were enrolled in the study group, and 499 noncancer subjects **(i)** were included in the reference group. Genomic data, including WGS, gene expression, clinical outcome, and FCH, were collected. First, we used the PopDel method **(ii)** to detect DSVs and perform data preprocessing **(ii)** from the WGS analysis of all subjects. The cancer risk predictive model **(iii)** was built with an attention-weighted model. We also studied DSVs in familial cancer **(iv)**. Second, we examined the relationship between DSVs and the tumor microenvironment **(v)**. Immune gene expression data were normalized. We constructed an immune gene expression-associated DSV correlation matrix with the point-biserial correlation. Third, a machine learning method with a survival support vector machine (survival-SVM) and Kaplan-Meier survival analysis was applied to examine prognosis and survival **(vi)**.

Fig. 2. Feature selection of DSVs to distinguish cancer and noncancer subjects. a

The architecture of the attention-weighted model for selecting the cancer risk DSV features. The primary purpose was to classify cancer or noncancer subjects by the neural network. This was a MLP model based on the attention mechanism. We used *n* samples (x_n) as input in the attention-weighted model: every sample had *m* (Del_m) filtered deletions. A value of 1 in the deletion vector indicates that the sample has the specific deletion, while 0 implies no deletion. A weighted vector ($\underset{W}{\rightarrow}$) is associated to the input layer to identify the importance of each deletion (red color gamut). Additionally, an embedding layer (*E* represents the embedding table, *e* denotes the embedding size) is applied to reduce the feature size and each deletion. We took the sum of each column and obtained a vector that can represent the information of the input deletion features ($r_{n,e}$); this is the input of multilayer perceptron. The output of MLP utilizes the SoftMax layer. The output labels are cancer or noncancer subjects.

b Performance of five machine learning strategies (attention-weighted model, MLP SVM, RF and LR) for cancer risk prediction with different number of features (2,919 and 671 cancer-associated DSVs). The attention-weighted model was more sensitive (AUC=0.71, sensitivity=0.57) than the other methods. All of the models performance are improved with 671 cancer-associated DSVs.

c PCA plot by cancer-associated DSVs. Red dots represent cancer subjects, and blue dots represent noncancer subjects. A total of 671 cancer-associated DSVs with positive

weights were used for PCA. DSVs can distinguish cancer and noncancer subjects.

Fig. 3. The frequency spectrum of DSVs in cancer and noncancer subjects and germline cancer susceptibility gene analysis. a Bar plot of the top 10 DSV genes with significant odds ratios and p-adjusted values <0.05 by the false discovery rate (FDR) in the cancer group and noncancer group separately. The x-axis indicates the percentage of subjects who carry DSV genes, while the y-axis represents the DSV genes. **b** Heatmap of 57 DSV genes and clinical information. Genes with an odds ratio and P-adjusted value < 0.05 by the FDR were selected. The clinical information includes sex, age, and FCH. **c** Bar plot of 26 DSV genes intersected in 565 germline cancer susceptibility genes in cancer and noncancer subjects. The x-axis indicates the well-known cancer susceptibility genes, and the y-axis indicates the frequency of the genes in cancer and noncancer subjects. FANCA, POLD1, and STK11 gene deletions occurred only in the cancer group.

Fig. 4. DSV genes and survival analysis in cancer patients with or without family cancer history. a Table of the association between cancer and FCH. The subjects who had FCH had a higher risk of developing cancer 1.89 [1.33-2.68] than the subjects without FCH (Fisher's exact test p= 0.0003). **b** Fisher's exact test and odds ratio were applied to measure the relationship between each DSV gene and FCH. Forest plot of

cancer patients with and without FCH. The DSV genes are *SMYD3* and *NKD2* in cancer patients with FCH. The DSV genes are *MGAT4C*, *HSPA4L*, *ZSCAN5A*, *LOC100505841*, and *NALCN* in cancer patients without FCH. **c** Forest plot of noncancer subjects with and without FCH. The DSV genes are *MALRD1*, *LOC101928523*, and *RAB9BP1* in noncancer subjects with FCH. There are nine DSV genes in noncancer subjects without FCH. **d** Point plot of the log2 hazard ratio DSV genes and log10 (p-value). The size of the point indicates the frequency of the DSV gene in cancer objects, and the red marks indicate the eight DSV genes with a p-value < 0.05. Blue points (*MUC4* and *CEP72* gene deletions) show the validated results.

Fig. 5. The correlation of DSVs and immune gene expression and prognostic stratification. a Heatmap of 57 DSV genes related to six functional immune expression categories. The ratio indicates the frequency of DSV genes related to immune expression. Only two DSV genes, *STNG2* and *LOC105376360*, have P-adjusted values <greatergreater.05. **b** Heatmap of 65 immune-related DSV genes and clinical outcomes. There are six functional immune categories: tumor characterization (green), lymphocyte regulation (purple), lymphocyte marker (orange), cytokine signaling (blue), checkpoint pathways (brown), and housekeeping (light blue). The value shown is the point biserial correlation coefficient in the heatmap. There are poorer prognostic DSV

genes correlated with the immune functional tumor characterization category and better prognostic DSV genes related to the immune functional lymphocyte regulation category. **c** RFS by Kaplan-Meier survival plots. The patients were stratified into G1 (orange) and G2 groups (blue) by prognostic deletions that have different clinical outcomes. The G2 group had better clinical outcome than the G1 group.

Fig. 6. Protein-protein interactions and the causal inference model. a Kaplan-Meier survival plot of the *MUC4* gene DSVs. *MUC4* (d-) indicates that cancer patients have no *MUC4* gene deletion. *MUC4* (d+) indicates that cancer patients have the *MUC4* gene deletion. RFS indicates recurrence-free survival. The survival analysis showed that patients with *MUC4* (d-) had a better clinical outcome ($p = 0.027$), and colorectal cancer patients with the *MUC4* gene deletion were associated with increased immune gene (*MAGE1*) expression. The right column shows the *MUC4* gene deletion and *MAGEA1* expression correlation plot ($r = 0.35$). **b** The STRING database was used to show protein-protein interactions of the transmembrane mucin family, including *MUC1*, *MUC4*, and *MAGEA1*. **c** Kaplan-Meier survival plot of *CEP72* DSV. *CEP72* (d-) indicates that cancer patients have no *CEP72* gene deletion. *CEP72* (d+) indicates that cancer patients have a *CEP72* gene deletion. RFS indicates recurrence-free survival. The survival analysis showed that patients with *CEP72* (d+) had a better clinical outcome ($p = 0.012$), and patients with the *CEP72* gene deletion were associated with

decreased immune gene (*IFIT1*) expression. The left figure shows the *CEP72* gene deletion and *IFIT1* expression correlation plot (r = 0.36). **d** Causal inference model of DSVs, immune gene expression, and RFS. Gray circles represent DSV genes, white circles represent immune expression genes, and the black circle represents RFS. The arrow indicates the causal effect pair. The red arrow pair indicates the RFS causal inference-associated pairs. The causal inference model showed that *CEP72* could affect RFS by *IFIT1*.

Figures

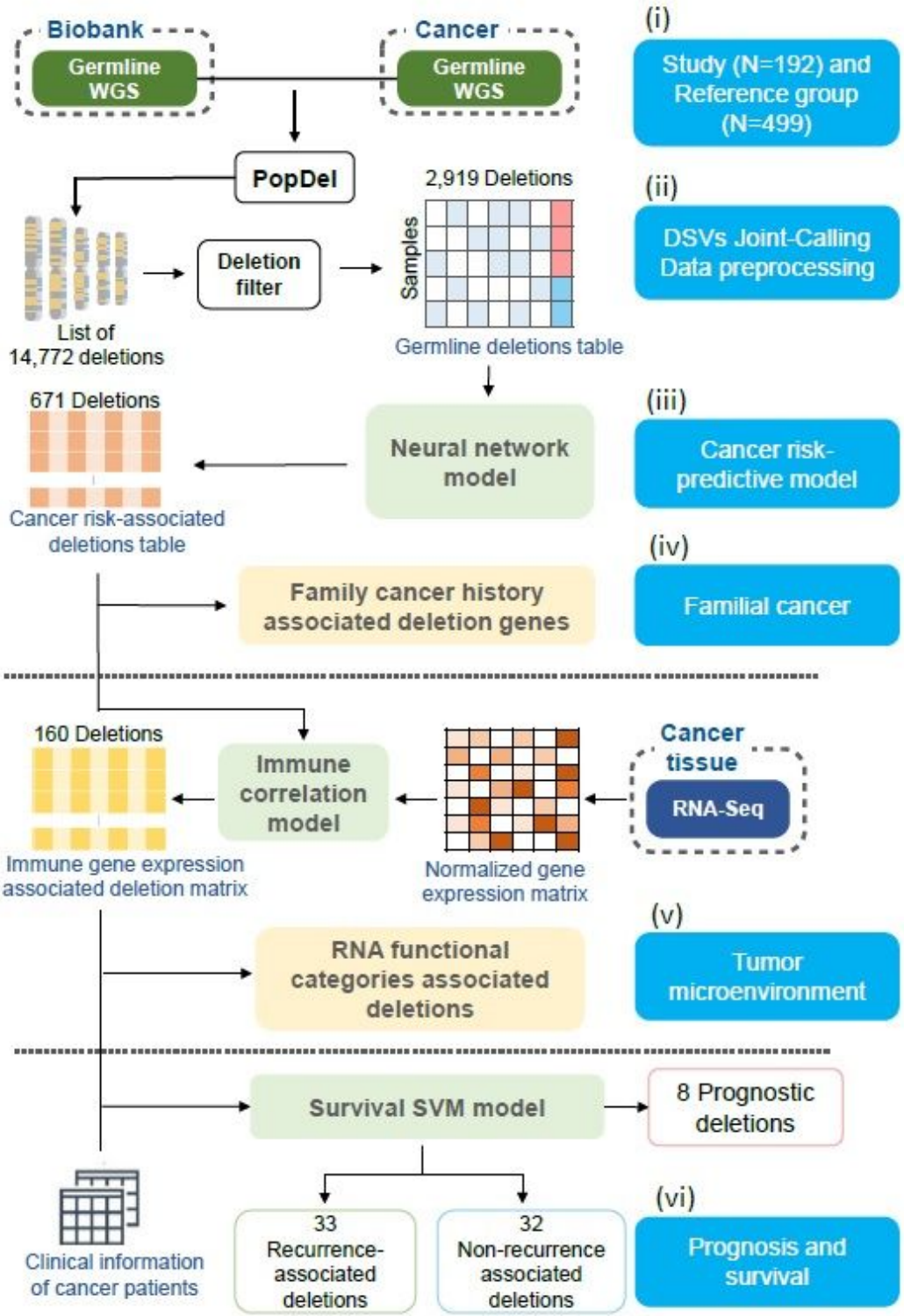

Figure 1

Study design and workflow. Study design and overall workflow of WGS analysis of germline DSVs and immune gene expression for cancer risk prediction and survival stratification. In total, 192 cancer patients (i)—comprised of 120 with colorectal cancer, 29 with endometrial cancer, 35 with ovarian cancer, and

eight with breast cancer—were enrolled in the study group, and 499 noncancer subjects (i) were included in the reference group. Genomic data, including WGS, gene expression, clinical outcome, and FCH, were collected. First, we used the PopDel method (ii) to detect DSVs and perform data preprocessing (ii) from the WGS analysis of all subjects. The cancer risk predictive model (iii) was built with an attention-weighted model. We also studied DSVs in familial cancer (iv). Second, we examined the relationship between DSVs and the tumor microenvironment (v). Immune gene expression data were normalized. We constructed an immune gene expression-associated DSV correlation matrix with the point-biserial correlation. Third, a machine learning method with a survival support vector machine (survival-SVM) and Kaplan-Meier survival analysis was applied to examine prognosis and survival (vi).

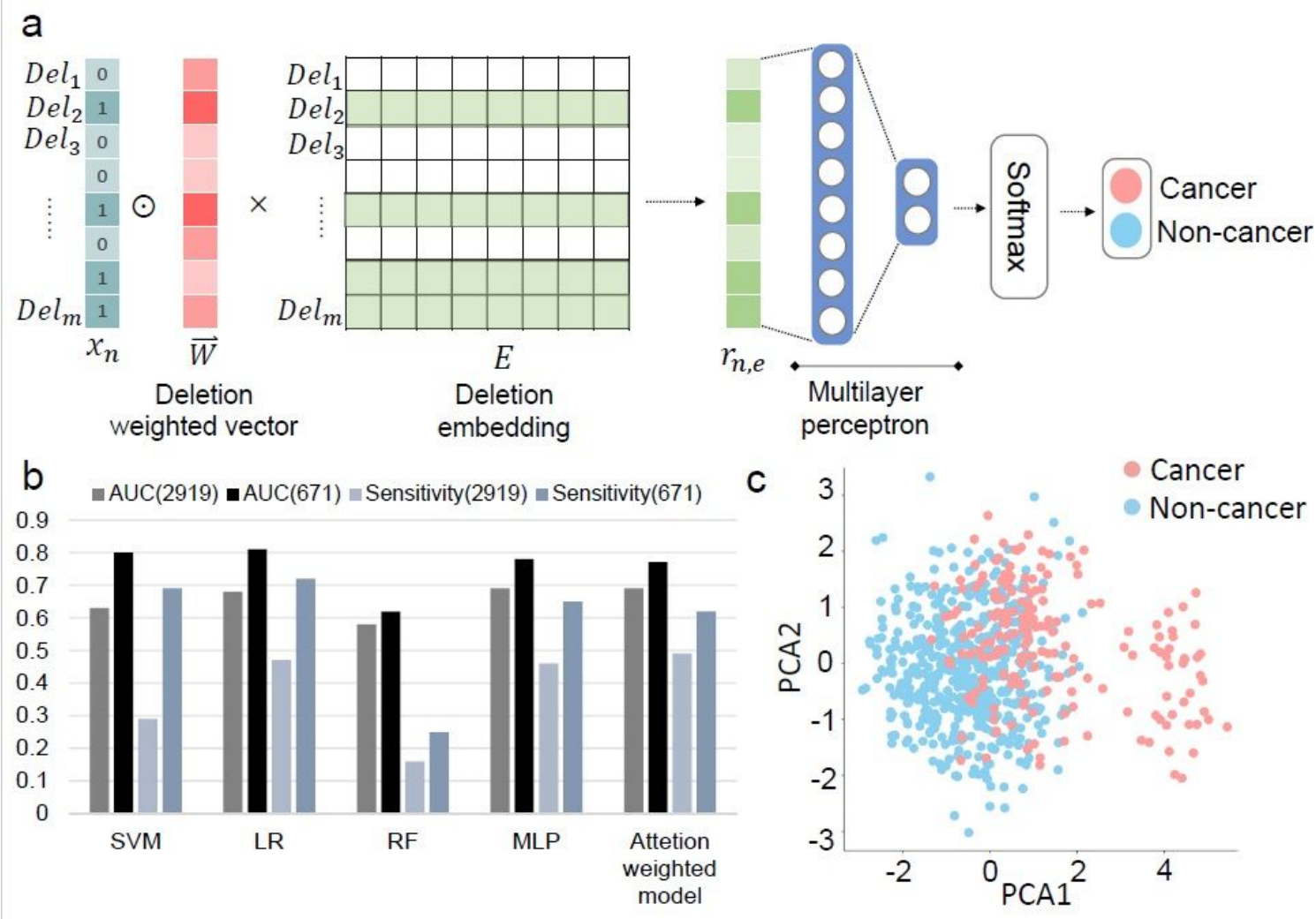

Figure 2

Feature selection of DSVs to distinguish cancer and noncancer subjects. a The architecture of the attention-weighted model for selecting the cancer risk DSV features. The primary purpose was to classify cancer or noncancer subjects by the neural network. This was a MLP model based on the attention mechanism. We used n samples (□□□□) as input in the attention-weighted model: every sample had m (□□□□□□□□) filtered deletions. A value of 1 in the deletion vector indicates that the sample has the specific deletion, while 0 implies no deletion. A weighted vector (□□□) is associated to the input layer to identify the

importance of each deletion (red color gamut). Additionally, an embedding layer (E represents the embedding table, e denotes the embedding size) is applied to reduce the feature size and each deletion. We took the sum of each column and obtained a vector that can represent the information of the input deletion features (□□□□,□□); this is the input of multilayer perceptron. The output of MLP utilizes the SoftMax layer. The output labels are cancer or noncancer subjects. b Performance of five machine learning strategies (attention-weighted model, MLP SVM, RF and LR) for cancer risk prediction with different number of features (2,919 and 671 cancer-associated DSVs). The attention-weighted model was more sensitive (AUC=0.71, sensitivity=0.57) than the other methods. All of the models performance are improved with 671 cancer-associated DSVs. c PCA plot by cancer-associated DSVs. Red dots represent cancer subjects, and blue dots represent noncancer subjects.A total of 671 cancer-associated DSVs with positive weights were used for PCA. DSVs can distinguish cancer and noncancer subjects.

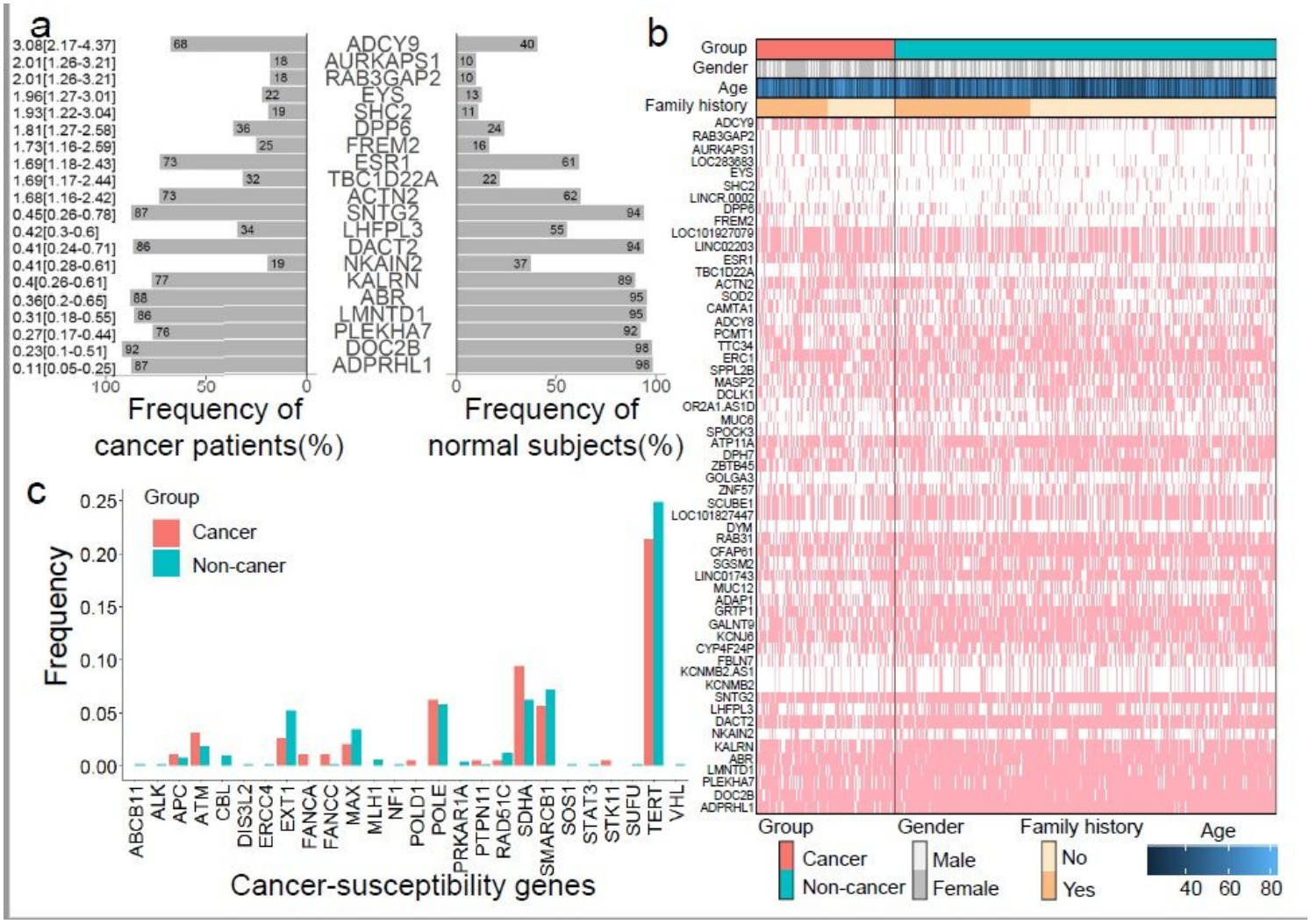

Figure 3

The frequency spectrum of DSVs in cancer and noncancer subjects and germline cancer susceptibility gene analysis. a Bar plot of the top 10 DSV genes with significant odds ratios and p-adjusted values <0.05 by the false discovery rate (FDR) in the cancer group and noncancer group separately. The x-axis indicates the percentage of subjects who carry DSV genes, while the y-axis represents the DSV genes. b

Heatmap of 57 DSV genes and clinical information. Genes with an odds ratio and P-adjusted value < 0.05 by the FDR were selected. The clinical information includes sex, age, and FCH. c Bar plot of 26 DSV genes intersected in 565 germline cancer susceptibility genes in cancer and noncancer subjects. The x-axis indicates the well-known cancer susceptibility genes, and the y-axis indicates the frequency of the genes in cancer and noncancer subjects. FANCA, POLD1, and STK11 gene deletions occurred only in the cancer group.

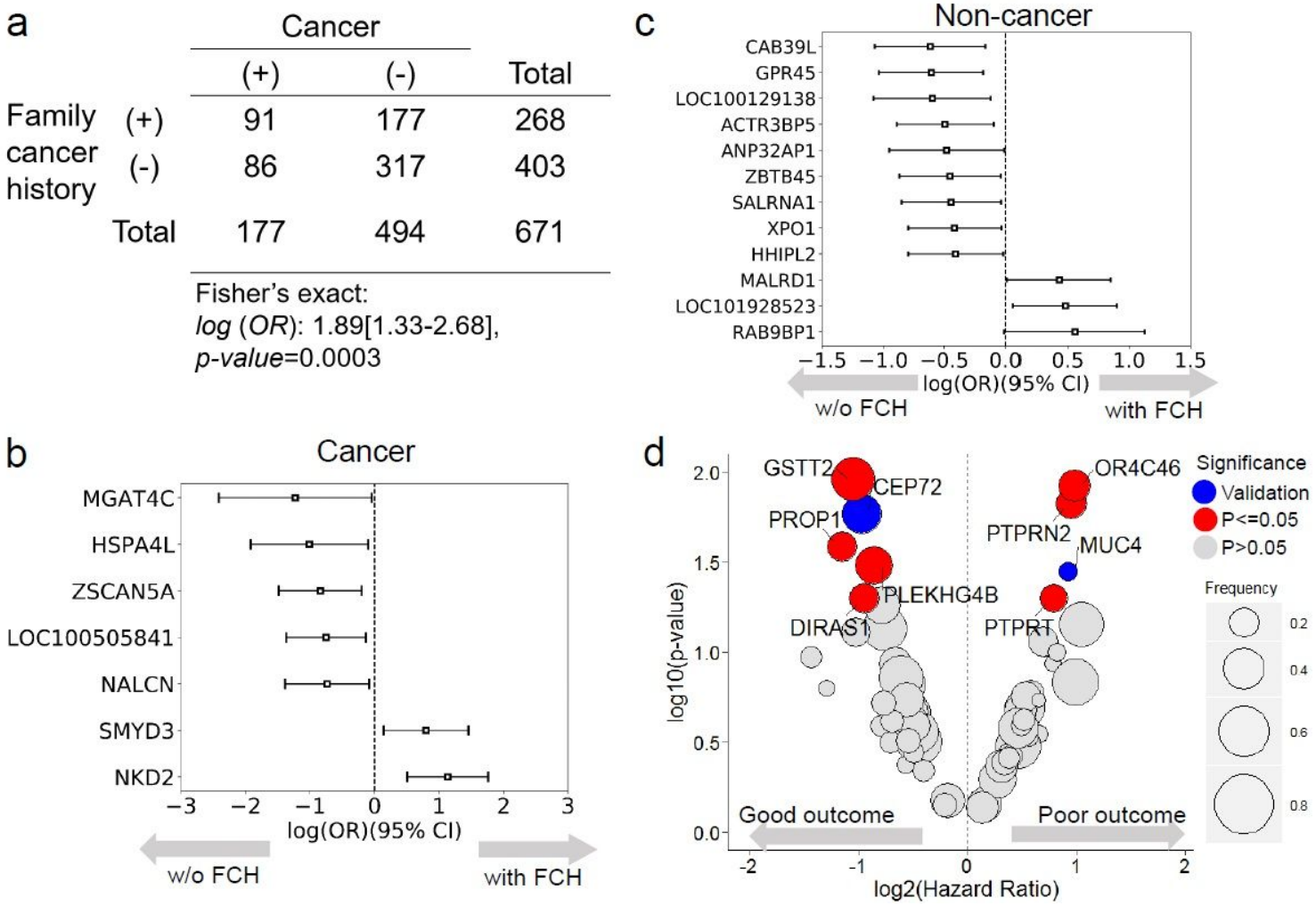

Figure 4

DSV genes and survival analysis in cancer patients with or without family cancer history. a Table of the association between cancer and FCH. The subjects who had FCH had a higher risk of developing cancer 1.89 [1.33-2.68] than the subjects without FCH (Fisher's exact test p= 0.0003). b Fisher's exact test and odds ratio were applied to measure the relationship between each DSV gene and FCH. Forest plot of cancer patients with and without FCH. The DSV genes are SMYD3 and NKD2 in cancer patients with FCH. The DSV genes are MGAT4C, HSPA4L, ZSCAN5A, LOC100505841, and NALCN in cancer patients without FCH. c Forest plot of noncancer subjects with and without FCH. The DSV genes are MALRD1, LOC101928523, and RAB9BP1 in noncancer subjects with FCH. There are nine DSV genes in noncancer subjects without FCH. d Point plot of the log2 hazard ratio DSV genes and log10 (p-value). The size of the

point indicates the frequency of the DSV gene in cancer objects, and the red marks indicate the eight DSV genes with a p-value < 0.05. Blue points (MUC4 and CEP72 gene deletions) show the validated results.

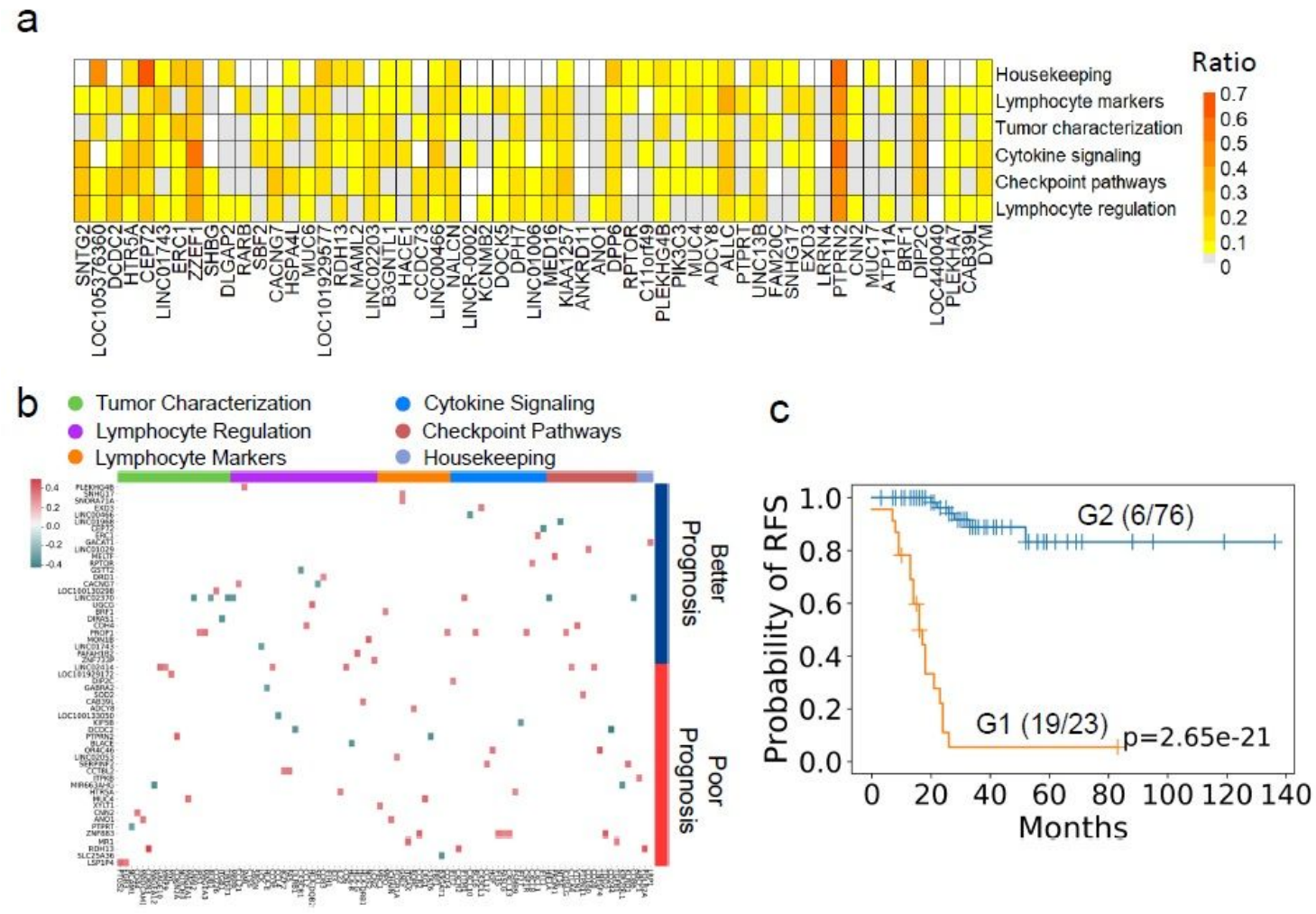

Figure 5

The correlation of DSVs and immune gene expression and prognostic stratification. a Heatmap of 57 DSV genes related to six functional immune expression categories. The ratio indicates the frequency of DSV genes related to immune expression. Only two DSV genes, STNG2 and LOC105376360, have P-adjusted values greatergreater.05. b Heatmap of 65 immune-related DSV genes and clinical outcomes. There are six functional immune categories: tumor characterization (green), lymphocyte regulation (purple), lymphocyte marker (orange), cytokine signaling (blue), checkpoint pathways (brown), and housekeeping (light blue). The value shown is the point biserial correlation coefficient in the heatmap. There are poorer prognostic DSV genes correlated with the immune functional tumor characterization category and better prognostic DSVgenes related to the immune functional lymphocyte regulation category. c RFS by Kaplan-Meier survival plots. The patients were stratified into G1 (orange) and G2 groups (blue) by prognostic deletions that have different clinical outcomes. The G2 group had better clinical outcome than the G1 group.

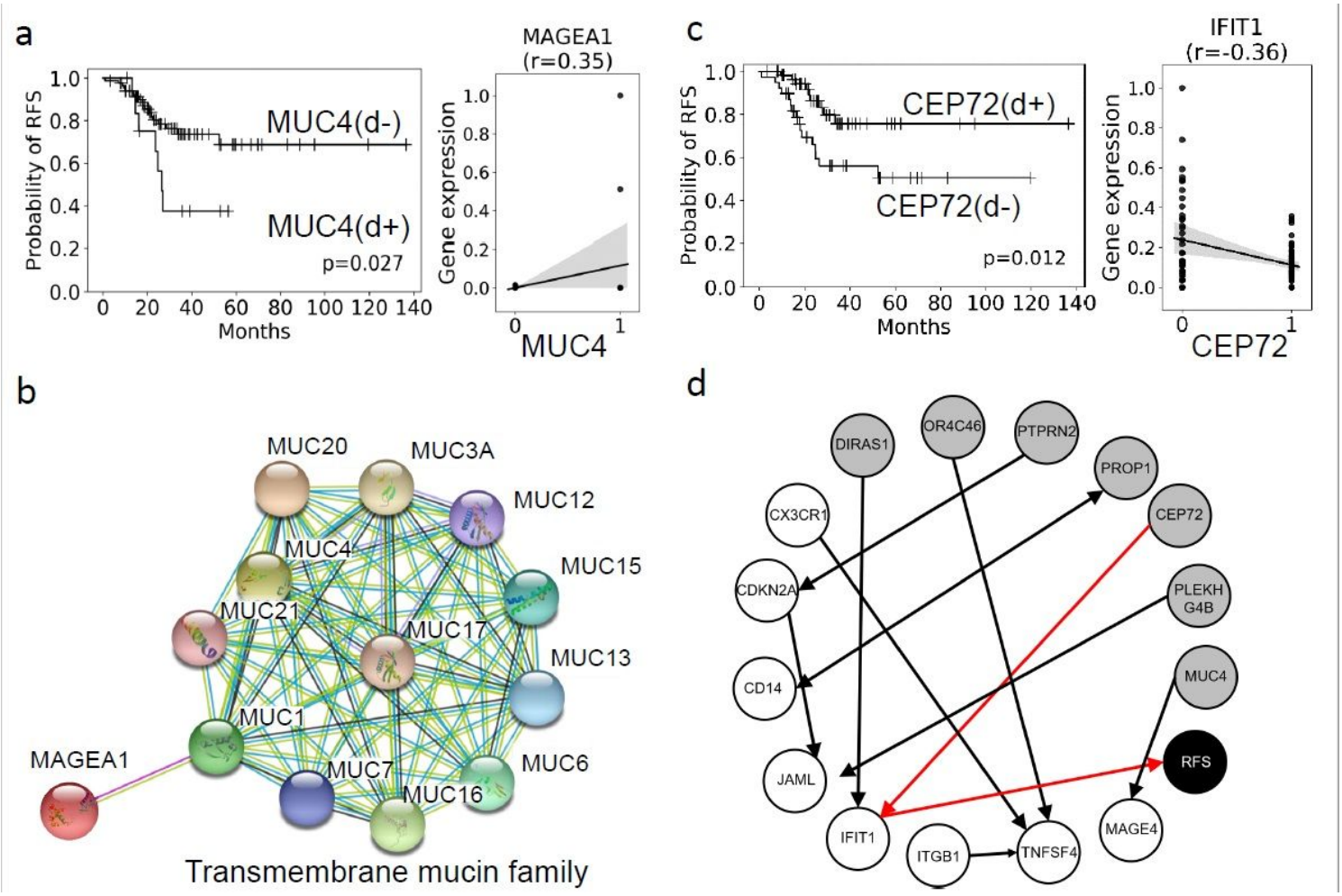

Figure 6

Protein-protein interactions and the causal inference model. a Kaplan-Meier survival plot of the MUC4 gene DSVs. MUC4 (d-) indicates that cancer patients have no MUC4 gene deletion. MUC4 (d+) indicates that cancer patients have the MUC4 gene deletion. RFS indicates recurrence-free survival. The survival analysis showed that patients with MUC4 (d-) had a better clinical outcome (p = 0.027), and colorectal cancer patients with the MUC4 gene deletion were associated with increased immune gene (MAGE1) expression. The right column shows the MUC4 gene deletion and MAGEA1 expression correlation plot (r = 0.35). b The STRING database was used to show protein-protein interactions of the transmembrane mucin family, including MUC1, MUC4, and MAGEA1. c Kaplan-Meier survival plot of CEP72 DSV. CEP72 (d-) indicates that cancer patients have no CEP72 gene deletion. CEP72 (d+) indicates that cancer patients have a CEP72 gene deletion. RFS indicates recurrence-free survival. The survival analysis showed that patients with CEP72 (d+) had a better clinical outcome (p = 0.012), and patients with the CEP72 gene deletion were associated with decreased immune gene (IFIT1) expression. The left figure shows the CEP72 gene deletion and IFIT1 expression correlation plot (r = 0.36). d Causal inference model of DSVs, immune gene expression, and RFS. Gray circles represent DSV genes, white circles represent immune expression genes, and the black circle represents RFS. The arrow indicates the causal effect pair. The red arrow pair indicates the RFS causal inference-associated pairs. The causal inference model showed that CEP72 could affect RFS by IFIT1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable6.CRCcharacteristics.docx
- Supplementarytable5.57Sixfunctionalcategories.xlsx
- Supplementarytable4.Coxproportional.xlsx
- Supplementarytable3.57Thespecificgenesinsubjects.xlsx
- Supplementarytable2.20ThefrequencyspectrumofDSVs.xlsx
- Supplementarytable1.DSVinformation.xlsx
- Supplementaryfigure6.SV20200129.pdf
- Supplementaryfigure5.SV20200129.pdf
- Supplementaryfigure420200818.pdf
- Supplementaryfigure3.SV20200129.pdf
- Supplementaryfigure2.SV20200129.pdf
- Supplementaryfigure1.SV20200820.pdf