Integrated proteogenomic and metabolomic characterization of papillary thyroid cancer with different recurrence risks

Ning Qu
Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China
Di Chen
Dalian Institute of Chemical Physics
Ben Ma
Fudan University Shanghai Cancer Center
Lijun Zhang
Ganmei Affiliated Hospital of Kunming Medical University
Yuting Wang
Fudan University Shanghai Cancer Center
Hongping Wang
Putuo Hospital, Shanghai University of Traditional Chinese Medicine
Zhaoxian Ni
Fudan University Shanghai Cancer Center
Wen Wang
Dalian Institute of Chemical Physics https://orcid.org/0000-0003-0986-8853
Tian Liao
Fudan University Shanghai Cancer Center
Jun Xiang
Fudan University Shanghai Cancer Center
Yu-Long Wang
Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, 200032, China
https://orcid.org/0000-0002-3830-7257
Shi Jin
The First Affiliated Hospital of Dalian Medical University
Dixin Xue
The Third Affiliated Hospital of Wenzhou Medical University
Weili Wu
The Third Affiliated Hospital of Wenzhou Medical University
Yu Wang

Department of Head and Neck Surgery, Fudan University Shanghai Cancer Center

Qing-Hai Ji

Fudan University Shanghai Cancer Center

Hui He

Fudan University Shanghai Cancer Center

Rong-Liang Shi

Fudan University Shanghai Cancer Center

Hai-long Piao

hpiao@dicp.ac.cn

Dalian Institute of Chemical Physics https://orcid.org/0000-0001-7451-0386

Article

Keywords:

Posted Date: June 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3036564/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on April 12th, 2024. See the published version at https://doi.org/10.1038/s41467-024-47581-1.

Abstract

Although papillary thyroid cancer (PTC) has a good prognosis, its recurrence rate is high and remains a core concern in the clinic. Molecular factors contributing to different recurrence risks (RRs) remain poorly defined. Here, we performed an integrative proteogenomic and metabolomic characterization of 102 Chinese PTC patients with different RRs. Genomic profiling revealed that mutations in *MUC16* and *TERT* promoter as well as multiple gene fusions like *NCOA4-RET* were enriched by the high RR. Integrative multi-omics analysis further described the multi-dimensional characteristics of PTC, especially in metabolism pathways, and delineated dominated molecular patterns of different RRs. Moreover, the PTC patients were clustered into four subtypes (CS1: low RR and BRAF-like; CS2: high RR and metabolism type, worst prognosis; CS3: high RR and immune type, better prognosis; CS4: high RR and BRAF-like) based on the omics data. Notably, the subtypes displayed significant differences considering BRAF and TERT promoter mutations, metabolism and immune pathway profiles, epithelial cell compositions, and various clinical factors (especially RRs and prognosis) as well as druggable targets. This study can provide insights into the complex molecular characteristics of PTC recurrences and help promote early diagnosis and precision treatment of recurrent PTC.

Introduction

Thyroid cancer (TC) is the most common malignant tumor of the endocrine system, and papillary thyroid cancer (PTC) is the most common type of thyroid malignancy. Although PTC is in general with low-grade malignancy and favourable long-term prognosis, the recurrence rate is relatively high with up to 20% of PTC patients having recurrences [1, 2]. The American Thyroid Association (ATA) risk stratification system categorizes the recurrence risks (RRs) into low, intermediate, and high levels based on several recurrence relevant clinical factors [1]. Uncovering molecular factors associated with PTC RR may promote early detection of PTC recurrences and better treatment. The elevated serum levels of Thyroglobulin (Tg) were found associated with PTC recurrence and have been applied for recurrence surveillance in clinical use [3]. Some recurrence relevant genes and microRNAs were identified based on the transcriptomics data [4, 5]. However, the molecular basis underlying different RRs are still not fully revealed.

High-throughput omics methods have been applied to explore the molecular atlas of PTC [4, 6–10]. Accordingly, the molecular landscape of PTC have been described. The high mutation frequencies in *BRAF*, *RAS*, *TERT* promoter and gene fusions involving *RET* have been widely observed [6, 11]. Proteomics and metabolomics studies described the remarkably altered protein and metabolite profiles in PTC [8, 10]. Existing omics based studies manifest that PTC is molecularly complex and the molecular characteristics underlying PTC recurrence require further more in-depth integrative investigations.

To obtain a more comprehensive perspective on the molecular landscape of PTC with different RRs, we performed an integrated proteogenomic and metabolomic investigation of PTC of 102 Chinese PTC patients. Our integrated analysis described the complicated and distinctive molecular features of the PTC patients and identified the RR relevant molecular landscape from the genomic, transcriptional, proteomic

and metabolism perspectives. We also redefined four molecular subtypes of PTC which not only possessed distinctive molecular characteristics but also showed significant differences in clinical and pathological scales, especially for RR patterns and recurrence-free prognosis. This multi-omics study holds immense potential in offering valuable data resources for unraveling the intricate molecular mechanisms of PTC recurrences, and the redefined molecular subtypes can significantly contribute to enhancing precision diagnosis and treatment of recurrent PTCs, thus leading to improved long-term survival rates.

Results

Overview of the multi-omics study of PTC

To comprehensively understand the molecular basis of PTC with different RRs, 102 PTC patients were collected, and five different types of omics including genomics, transcriptomics, metabolomics, proteomics and phosphorylated (phospho)-proteomics were performed (Fig. S1a). Whole exome sequencing (WES) based genomics data were from 97 tumor tissue samples and 33 paired normal tissues, the RNA-sequencing (RNA-seq) based transcriptomics data (16925 genes) were from 92 tumor tissue samples and 34 paired normal tissue samples, metabolomics profiling (503 metabolites) were conducted on 102 tumor tissue samples and 37 paired normal tissue samples, and proteomics (3147 proteins) and phospho-proteomics (652 phospho-proteins) profiling were performed on 37 paired tumor-normal tissues (Fig. S1a).

Genomic profiling of the PTC patients

An average of 74 nonsynonymous somatic point mutations and 2 indels were identified in the 97 Chinese PTC patients. Consistent with most genome studies about PTC [12, 13], the most frequent somatic mutation gene was *BRAF* (47%, all belong to V600E mutation, Fig 1a). In addition, frequently mutated cancer-associated genes also included *MUC16* (36%), *RNF213* (8%) and *MSH6* (7%) (Fig. 1a), showing higher mutation frequencies than the TCGA PTC dataset [13]. Here, the *MUC16* mutations were specifically enriched in the PTC patients with high RR (Fig. 1b), and also associated with multiple pathological factors including high RR (P=0.027), recurrence (P=0.010), metastatic lymph node size larger than 3cm (LNM.3cm) (P=0.018), T3 stage (P=0.032), N1b stage (P=0.0061) and M1 stage (P=0.021) (Fig. 2c, examined by hypergeometric distribution). Meanwhile, this Chinese PTC cohort did not contain mutations in RAS which were frequently mutated in previous PTC studies [12-14].

There were also frequent TERT promoter mutations (C228T, 14%) in the PTC patients. The mutations were also significantly enriched in the high RR patients (Fig. 2b and 2d), and frequently overlap with certain pathological or clinical factors including high RR (P=0.0026), recurrence (P=0.0030), LNM.3cm (P=0.011), extrathyroidal extension (ETE) (P=0.0013), lymph node metastasis (LNM) (P=0), or extranodal extension (ENE) (P=0.0077) (Fig. 2d, examined by hypergeometric distribution).

Gene rearrangements in *RET*, *NTRK* and *BRAF* have been frequently identified in PTC [15]. Here, *RET* fusions (*CCDC6-RET* 8%, *NCOA4-RET* 5%) were the most frequent fusions, and multiple *NTRK* fusions (*NTRK3-ETV6*, *TPR-NTRK1*, *ETV6-NTRK3*) were also identified (Fig. 1a). In addition, several other gene fusions (*FBXO25-SEPTIN14*, *TLK2-FAM157A*, *ZNF33B-NCOA4*) showing rare frequencies in previous PTC studies were also identified (Fig. 1a). Interestingly, several gene fusions (*NCOA4-RET*, *TLK2-FAM157A*, *ZNF33B-NCOA4*, *TPR-NTRK1*) also showed specific enrichment in the high RR (Fig. 1b).

Multi-omics based comparison of tumor and normal tissues of PTC patients

In addition to the genomic alterations, differentially expressed molecules (DEMs) were recognized by comparing between tumor and matched normal samples based on the multi-omics profiling data (Fig. S2). As a result, four types of DEMs, including 1674 genes (P<0.01, DeSeq2 [16]), 1864 proteins (P<0.01, Wilcox-test, paired), 391 phospho proteins (P<0.01, Wilcox-test, paired) and 334 metabolites (P<0.01, Wilcox-test, paired) were recognized (Fig. 2a). The most significant DEMs in mRNA level included *GGTLC3, C1QL1, PRG4*, etc. The most significant metabolites included free fatty acids (FFAs), serine, citric acid, triglycerides (TGs), sphingosines (SPHs), etc. Increased FFAs in PTC tumors have also been identified by other metabolomics studies [17]. Proteins like tenascin (TNC), and Fibronectin 1 (FN1) and dipeptidyl peptidase 4 (DPP4) and phospho-proteins of major vault protein (MVP) and Fibronectin 1 (FN1) showed remarkable up-regulation in the PTC tumor tissues, while proteins thyroid peroxidase (TPO), desmin (DES) and fatty acid binding protein 4 (FABP4), and phospho-proteins of thyroglobulin (Tg), DES, hemoglobin subunit delta (HBD) and hemoglobin subunit beta (HBB) were down-regulated in the PTC tumors (Fig. 2b). TNC was reported to show remarkably high expressions in medullary TC [18], here, we found it was also up-regulated in the PTC tumors. The up-regulation of FN1 was observed in all types of TC [19]. TPO, an essential enzyme for the production of thyroid hormones, is expressed mainly in normal thyroid cells [20]. The expression levels of TPO were decreased in the PTC tumors when compared to the normal ones.

Pathways enriched by the four types of DEMs were identified respectively (P<0.05, Hypergeometric Distribution). A large fraction of the enriched pathways were metabolism pathways even for the DEMs in terms of genes, proteins and phospho-proteins (Fig. 2a), and the pathways enriched by genes, proteins and phospho-proteins simultaneously all fell in metabolism pathways including valine, leucine and isoleucine degradation, pyruvate metabolism, glycolysis/gluconeogenesis, as well as arginine and proline metabolism (Fig. 2c), where glycolysis and pyruvate metabolism were also enriched by the differentially expressed metabolites (Fig. 2d). Meanwhile, multiple metabolism pathways like oxidative phosphorylation and citrate cycle were enriched by at least three types of DEMs (Fig. 2d). These together suggest the remarkable metabolic alterations of the PTC tumor tissues.

The multi-omics based pathway analysis enable a comprehensive description of the pathway alteration. Taken glycolysis as one example (Fig. 2e), we observed that although most enzymes were down-

regulated considering the mRNA expressions (e.g., *HK1*, *PGM1*), many of them were up-regulated in the protein or phospho-protein levels (e.g., ENO1, PCK1, PDHA1), suggesting the complicated post-transcriptional modifications in the glycolysis process of PTC patients. Meanwhile, the metabolite changes were mainly reflected in the reduced levels of glucose, fructose, glycerate-3P and pyruvate and increased levels of lactate (i.e., L-Lactate) (Fig. 2e). Increased levels of lactate in TC and many other caner types have been widely reported [9]. The multi-omics based pathway alteration patterns help further explain potential mechanisms including alteration of the direct enzyme LDHA (in both mRNA and protein levels) and associated up/down-stream changes (e.g., ENO1, PKM and pyruvate).

In addition to the metabolism pathways, two cell death relevant pathways, necroptosis and apoptosis, were also significantly enriched by the DEMs in terms of metabolites, proteins and phospho-proteins (Fig. 2d). Most of the DEMs in the necroptosis and apoptosis pathways were up-regulated in the PTC tumor tissues than the normal tissues (Fig. S3a-S3b).

Multi-omics based molecular features of PTC with different RRs

The molecular expression features underlying different RRs of PTC were also characterized (Fig. 3a). The high RR PTC patients showed higher expression levels in multiple lipids like TGs, FFAs and the other metabolites like histamine and kynurenine. The high RR also displayed higher expressions in genes like *MMP13, CST1, COL11A1,* proteins like Tg, PTRRG, VWA1 and phospho-proteins like EPPK1, ALDH1A1 and LAMC1. The intermediate RR PTC patients showed higher expressions in metabolites like several FFAs and kynurenine, genes like *IGFN1*, *LOC391322* and *ZNRD1*, proteins like FTL, FABP5 and APOB, and phospho-proteins like C1QB, HBB and HBD. The low RR patients showed higher expressions in metabolites like PG (18:2_18:2) and OAHFA (18:2_18:1), genes like *JSRP1*, *TCAP* and *TNNI2*, proteins like ACADL, ABHD11 and FN1 and phospho-proteins like TNC, FN1 and POSTN. Comparing to the alterations between tumor and normal samples, the high RR PTC tumor samples showed reversed alterations compared to intermediate or low RR ones considering various molecules (Fig. 3a and Fig. S4a-S4d). For instance, although the PTC tumor samples showed significant reduced protein levels in Tg compared to the normal samples, the high RR PTC samples were with higher protein expressions of Tg comparing to other tumor samples (Fig. 3a and Fig. S4a-4d).

The expression profiles of the RR relevant molecules especially for the FFAs (FFA 26:2, FFA 24:2, FFA 26:4) and several proteins or phospho-proteins were highly associated (Fig. 3b, spearman correlation > 0.65 or spearman correlation < -0.65). The FFA 26:2, Tg, FN1 and 5-Lipoxygenase (ALOX5), phospho-FN1 and phospho-TNC harbored a relative hub position in the correlation network, suggesting their crucial roles in interactive regulations or signaling communications. ALOX5, as a non-heme iron-containing enzyme, can catalyze the peroxidation of polyunsaturated fatty acids [21]. Aberrant expression of ALOX5 has been observed in various types of cancers including PTC [22]. Here, we also found ALOX5 showed

specific low expressions in high RR PTC patients, and its alterations were associated with changes in many FFAs (Fig. 3b-d).

The different RRs also displayed remarkable differences in the pathway profiles. For high RR, the metabolites in various metabolism pathways, e.g., biosynthesis of amino acids and glycolysis, were up-regulated, while the protein levels of metabolic enzymes were mainly down-regulated (Fig. 3c). Except of the direct metabolism enzymes, there were other proteins showing remarkable associations with metabolite changes in PTC (Fig. 3f), e.g., protein Tg, phospho-protein MSN (Fig. 3g-3h). For the other pathways, the high RR showed up-regulations in PI3K-AKT and TGF-beta signaling pathways (based on the mRNA expressions) and thyroid hormone synthesis (based on the protein expressions) (Fig. S4d).

Integrative correlation analysis of the multi-omics data

Based on a supervised multi-omics integrative analysis method called DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents) [23] to simultaneously maximize the correlations among different types of omics and identify key molecules which can discriminate different sample groups (i.e., high, intermediate and low RR groups and the normal sample group). As a result, the general correlations between metabolism, proteomics and phospho-proteomics were high (no less than 0.88) suggesting common information among metabolism, proteomics and phospho-proteomics. By contrast, the correlations between transcriptomics and the other types of omics were relatively low (less than 0.6, Fig. 4a), implying the complicated post-transcription modifications.

The key molecules were further clustered into four network modules based on their expression profiles and inter-correlations, and different modules showed distinctive expression profiles (Fig. 4b) and interaction patterns (Fig. 4c). The molecules in the first module (M1) were mainly composed of extracellular matrix (ECM) relevant proteins including FBLN5, NID1, NID2, COL4A2, TINAGL1, VWA1 and different chains of the laminin proteins (LAMA4, LAMB1, LAMC1) [24], they showed high inter-correlations and possessed higher expression levels in the high RR groups than the other tumor samples (Fig. 4b-4c), highlighting the key role of ECM interactions in PTC recurrence. The second module (M2) was composed of multiple metabolism relevant phospho-proteins like PGK1, PSMF1, PRDX1 and TOM1, they showed lower expressions in the tumor tissues, especially the high RR tumor tissues (Fig. 4b). Their expressions were also associated with the proteins IGKV3-15, RPS27A, genes *MYH11* and *HTR2A* (Fig. 4c). The third module (M3) was the largest module, and molecules in M3 mainly showed higher expressions in the tumor tissues (regardless of the RRs) than the matched normal tissues (Fig. 4b). There were three sub-modules in M3 which were aggregated by metabolites, genes and proteins/phospho-proteins (Fig. 4c, M3). The inter-correlated metabolites in M3 were mainly lipids including phosphatidylcholines (PCs), phosphatidylethanolamines (PEs) and sphingomyelins (SMs). A large fraction of the proteins/phosph-proteins were involved in autophagy (STAT3/ATP6V1B2/ATP6V1C1/HGS/LAMTOR2/VPS13C/HSP90AA1) [25]. The genes were involved in glutathione metabolism (*GGTLC3*/*GGT2*) [25], immune response (*IFNE*/*PRSS2*) [25] and exocytosis

secretion of thyroid-stimulating hormone (*TRHR*). Meanwhile, *TRHR, C1QL4* and *AMY1B* possessed inter-connection positions in the network of M3. Molecules in the fourth module (M4) mainly showed higher expressions in the intermediate and low RR groups (Fig. 4b). Metabolites including FFAs, Diacylglycerols (DGs) and Phosphatidylglycerols (PGs) and the fatty acid binding protein FABP5 formed an intermediated layer linking the genes and proteins or phospho-proteins in the network of M4, and multiple proteins and phospho-proteins (ERAP2/CYBB/CD74/DYNC1H1/RAB7A) in M4 were involved in antigen processing and presenting [25] (Fig. 4c), indicating the potential interactions between fatty acid metabolism and adaptive immune functions in PTC.

Integrative stratification of PTC patients into four subtypes based on transcriptomics and metabolomics

Notably, although most high RR samples showed low expressions of molecules in M4, part of them also show similar expression profiles with the intermediate and low RR samples (Fig. 4b), implying alternative molecule subtypes different from the ATA risk classification may exist. We re-stratified the PTC patients into four subtypes based on a consensus integrative clustering analysis (see Methods, Fig. S5a) of the transcriptomics and metabolomics data (proteomics and phospho-proteomics were not considered here since the two types of omics were highly associated with metabolomics).

The four redefined subtypes showed significant differences in the transcriptional and metabolism profiles (Fig. 5a) as well as multiple clinical and mutation features including RR ($P = 1.61\times10^{-5}$, chi-square test), T stage ($P = 4.17\times10^{-2}$, chi-square test), N stage ($P = 2.45\times10^{-3}$, chi-square test), BRAF mutation ($P = 5.67\times10^{-3}$, chi-square test), TERT promoter mutation ($P = 3.49\times10^{-3}$ for overlap between CS4 and TERT promoter mutation, examined by hypergeometric distribution) (Fig. 5a-5b) and ENE ($P = 6.95\times10^{-3}$, chi-square test, Fig. S5b). The subtype CS1 contained more low RR patients, while the other three subtypes CS2 to CS4 had more high RR patients (Fig. 5b). The subtype CS2 and CS3 had more T3 stage, N1b stage patients but had less BRAF and TERT promoter mutations (Fig. 5b). The subtype CS4 had the highest BRAF and TERT promoter mutation frequencies (Fig. 5b). Moreover, the four subtypes possessed different prognosis outcomes in terms of recurrence free survival (Fig. 5c), where the subtype CS2 and CS3 respectively showed the worst and best prognosis among the three high RR enriched subtypes (Fig. 5d-5e). By contrast, the prognosis differences based only on the RRs were not significant (Fig. S5c). Taken together, the transcriptomics and metabolomics data help redefine four meaningful PTC subtypes.

Multi-dimensional characterization of the four PTC subtypes

The four subtypes were with remarkably distinctive molecular profiles, and each subtype possessed various specifically up or down regulated genes (Fig. 6a) and metabolites (Fig. 6b). Plasminogen (PLG) was reported to show significantly lower expressions in serum samples of PTC patients than the nodular

goiter patients [26]. Here, the mRNA expression levels of *PLG* were higher in the low RR dominated subtype CS1 than the other subtypes (Fig. 6a). *HSP6A* was found to be a potential biomarker to predict the prognosis of TC [27]. *ECM1* is associated with tumor invasiveness and poor prognosis in various cancer types [28]. Both *HSP6A* and *ECM1* showed CS2 specific higher expressions (Fig. 6a). Considering metabolites, the subtype CS1 had higher levels in FFAs, PGs, Fructose 1,6–diphosphate, etc.; the subtype CS2 had higher levels in stachydrine; the subtype CS3 had higher levels in TGs, citric acid, etc.; while the subtype CS4 showed higher levels in acylcarnitines, adenosine, histamine, etc. (Fig. 6b).

In addition to the molecular features, the four subtypes also showed differences in the other key aspects including tumor sizes, number of metastatic lymph nodes, tumor differentiation scores (TDSs) [6], BRAF-scores and RAS-scores. The subtype CS2 and CS3 showed larger tumor sizes than the subtype CS1 (Fig. 6c). The subtype CS1 had less number of metastatic lymph nodes than the other subtypes (Fig. 6d). The subtype CS2 showed higher TDSs than the subtype CS1 and CS4, and the subtype CS3 had higher TDSs than the subtype CS1 (Fig. 6e). Moreover, both subtype CS2 and CS3 showed higher RAS scores and lower BRAF scores comparing to the subtype CS1 and CS4 (Fig. 6f-6g). Correspondingly, the subtypes CS2 and CS3 had lower BRAF mutation frequencies than CS1 and CS4 (Fig. 5b).

Furthermore, the pathway profiles for the four subtypes were also identified. Although no significant differences in terms of the tumor sizes, number of metastatic lymph nodes, TDSs, BRAF and RAS scores were observed between the subtype CS2 and CS3 (Fig. 6c-6g, Fig. S6a), they displayed noteworthy opposite trends in the pathway profiles (Fig. 6h-6i). For the metabolism pathways, the mRNA expressions of enzymes in various metabolism pathways were up-regulated for CS2 and down-regulated for CS3 (Fig. 6h). Reversely, most immune relevant pathways were up-regulated for CS3 but down-regulated for CS2 (Fig. 6i). The up-regulation in various metabolism enzymes in CS2 imply a high metabolite consumption and can partly explain why few metabolites show higher levels in the CS2 subtype (Fig. 6b).

Single cell RNA-seq analysis has been performed to illustrate the tumor microenvironment of PTC [29]. We used a deconvolution method to predict the tumor microenvironment compositions of the PTC tumor samples based on the bulk sample RNA-seq data in our study and a previously reported single cell RNA-seq dataset of PTC [29] (see also Methods). As result, the four subtypes also displayed distinctive cell compositions, especially for the epithelium sub-populations (Fig. 6j, Fig. S6b), implying the different subtypes are oriented from different types of malignant thyrocytes (Fig. 6j).

Potential targets or biomarkers of different subtypes

According to the clinical, molecular and pathway features of the four subtypes, we summarized the four subtypes as low RR and BRAF-like (CS1), high RR and metabolism type (CS2), high RR and immune type (CS3), and high RR and BRAF-like (CS4).

Candidate druggable targets for each subtype were identified (Fig. 7a). These distinctive expression profiles of the four subtypes in the druggable targets, suggesting that different therapeutic strategies

should be applied to different subtypes in PTC. Carbonic anhydrase 12 (CA12/CAXII) is one metabolic enzyme and has emerged as a one promising cancer therapeutic target [30]. Germinal center kinase (GCK) has been identified as a therapeutic target in multiple myeloma with RAS mutation [31]. Both CA12 and GCK can be candidate targets for the high RR and metabolism subtype (CS2) as both CA12 and GCK showed higher expressions for CS2 (Fig. 7a). PRMT8 can be a potential therapeutic target for colon tumor [32]. CARTPT may affect thyroid stimulating hormone levels due to its association with the neuroendocrine function of hypothalamus [33]. PRMT8 and CARTPT can be potential therapeutic targets of the high RR and immune subtype (CS3, Fig. 7a). CEACAM6 has been regarded as a potential target for cancer immunotherapies [34], and it showed a CS4 specific up-regulation among the four subtypes (Fig. 7a).

In addition, metabolites which showed significantly associations with these potential targets were also identified (Fig. 7b). These metabolites can be potential biomarkers to indicate different targets. For instance, CEACAM6 showed high positive correlations with PE (16:0p_16:1) and PE (16:0p_18:1), then high levels of the two PEs in certain PTC patients may suggest the applicability of the CEACAM6 targets.

Furthermore, to predict the four meaningful PTC subtypes based on the omics data, we constructed a subtype predictor based on the expression levels of the top ranked genes and metabolites (see Methods). The predictor can classify the four subtypes accurately, with the areas under the receiver operator characteristic curves (AUCs) 1, 0.97, 0.97, 0.99 for CS1 to CS4 on the testing dataset (Fig. 7c). Meanwhile, the predicted probabilities of being the subtype CS2 were associated with recurrence-free survival where high CS2 probabilities were associated with unfavourable prognosis (Fig. 7d).

Discussion

In this study, we performed an integrative investigation of 102 Chinese PTC patients based on multi-omics profiling including WES, transcriptomics, metabolomics, proteomics and phospho-proteomics. We identified the molecular and pathway characteristics of these Chinese PTC patients from multi-omics perspectives. Consistent with previous reports, common mutations in PTC like BRAF, TERT promoter and gene fusions involved in RET were also revealed. However, no RAS mutation was identified, probably due to the geographical limitations of the sample collection. Additionally, frequent mutations in *MUC16* (36%) were also observed. Although few MUC16 mutation was reported in PTC studies, *MUC16* showed high mutation frequencies and was associated with prognosis in various types of cancers like gastric cancer [35], glioma [36] and melanoma [37], and *MUC16* mutation was found to be associated with better response to immune checkpoint inhibitors in solid tumors [38]. Molecular characters in terms of metabolites, genes, proteins and phospho-proteins were described, these molecular alterations especially focused on the metabolism processes, especially for the glycolysis and pyruvate metabolism.

Meanwhile, the molecular differences of PTC with different levels of recurrence risks were also portrayed. Mutations in MUC16, TERT promoter and various gene fusions were specifically enriched in the high RR

PTC patients. The multi-omics based molecular expressional patterns of different RRs were comprehensively described. It has been reported that elevated post lobectomy serum Tg can be used to predict high recurrence risk [3]. However, the other information about the molecular characteristics of different RRs in PTC was limited. Here, we found the high RR was associated with elevated levels in triglycerides, genes *MMP13* and *CST1*, proteins Tg, PTPRG and VWA1 and phosphorylated EPPK1, ALDH1A1 and LAMC1, etc. Expression of MMP-13 was reported to be associated with PTC invasion and metastasis [39]. CST1 up-regulation was found to facilitate cell proliferation, motility, epithelial-mesenchymal transition and stemness in PTC [40]. LAMC1 was reported to show a higher level in samples from PTC patients with metastasis [41]. The specific molecular features of intermediate and low recurrence risks were also described (Fig. 4). The complicated correlations between molecules were also investigated. Interestingly, the high RR showed specific high expressions in the ECM relevant protein dominant correlation network and low expressions in the FFA centered correlation network. Moreover, the metabolites showed high correlations with proteins and phospho-proteins in general. These findings therefore add to the atlas of biomarkers or targets that may be applied in the diagnosis and treatment of recurrent PTC.

PTC patients were usually stratified into high, intermediate and low RR according to the ATA recommendations. In this study, we re-stratified the PTC patients into four molecular subtypes based on an integrative clustering analysis using both transcriptomics and metabolomics. The four subtypes were different from the original RR groups, where the first subtype was featured by low RR patients while the other subtypes were mainly composed of the high and intermediate RRs. Biologically, these four subtypes showed remarkable differences in terms of hallmark mutations, PTC relevant gene and metabolite expressions, epithelial cell compositions, as well as metabolism and immune pathway profiles. Clinically, the subtypes were also enriched by different pathological factors including the disease stages, lymph node metastasis and tumor invasion states. Importantly, the four subtypes showed a significant difference in prognosis, where the subtype CS2 (high RR, less BRAF mutations, up-regulation in metabolism) showed the most unfavourable recurrence-free prognosis outcomes, and the subtype CS3 (high RR, less BRAF mutations, up-regulations in immune pathways) showed a relatively good prognosis among the three high RR subtypes, highlighting the important roles of metabolism and immune pathways in PTC recurrence. Moreover, the expressional patterns of druggable targets for the four subtypes were distinctive, suggesting subtype specific treatment may be needed. The redefined subtypes suggest ATA risk stratification should not be used as one single predictor, other molecular profiles should be taken into consideration as well to do better and precision management of PTC recurrences.

Overall, we revealed the molecular basis of PTC with different RRs and proposed an effective molecular stratification strategy. We also illustrated the PTC subtypes relevant molecular characteristics, identified potential drug targets, constructed subtype predictors and highlighted the important role of metabolism in PTC, thus can provide guidance for PTC stratification and promoting precision diagnosis and treatment.

Materials and methods

Clinical sample collection

The consecutive samples used for this study were selected from patients diagnosed with PTC from Oct 2014 to Jul 2021 at Fudan University Shanghai Cancer Center (FUSCC) in China. The clinical information of the enrolled PTC patients were also recorded, including age, gender, sex, tumor size, lymph node metastasis (LNM), extrathyroidal extension (ETE), extranodal extension (ENE), number of metastatic lymph nodes (LNM.No), TNM staging (AJCC cancer staging system 8th edition) and RR stratification (2015 ATA guideline [42]). Each patient provided a written informed consent for his/her specimens and information to be used for research and stored in the hospital database, and this study was approved by the Ethical Committee of the FUSCC. All procedures performed in our study were in accordance with the ethical standards of our institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

DNA library preparation and WES

The exome DNA sequences were enriched from 0.4 μg genomic DNA using Agilent SureSelect Human All Exon V6 kit according to manufacturer's protocol. DNA fragments were end repaired and phosphorylated, followed by A-tailing and ligation at the 3'ends with paired-end adaptors. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR reaction. Then, libraries hybridize with liquid phase with biotin labeled probe, and use magnetic beads with streptomycin to capture the exons of genes. Captured libraries were enriched in a PCR reaction to add index tags to prepare for sequencing. Products were purified using the AMPure XP system (Beckman Coulter, Beverly, USA), DNA concentration was measured by Qubit®3.0 Flurometer (Invitrogen, USA), libraries were analyzed for size distribution by NGS3K/Caliper and quantified by real-time PCR (3 nM). At last, DNA library were sequenced on Illumina for paired end 150 bp reads. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using Illumina PE Cluster Kit (Illumina, USA) according to the manufacturer's instructions. After that, the DNA libraries were sequenced on Illumina platform and 150 bp paired-end reads were generated.

RNA library preparation and RNA-seq

Briefly, mRNA was extracted and purified from total RNA of the fresh frozen tissues using poly-T oligo-attached magnetic beads. RNA integrity was measured using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Fragmentation was carried out using divalent cations under elevated temperature in First Strand Synthesis Reaction Buffer (5X). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H). Second strand cDNA was synthesized by DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, Adaptor with hairpin loop structure were ligated to prepare for hybridization. To select cDNA fragments of preferentially 370 ~ 420 bp n length, the library fragments were purified with the AMPure XP system (Beckman Coulter, Beverly, USA). Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. At last, PCR products were purified (AMPure XP system) and

library quality was assessed on the Agilent Bioanalyzer 2100 system. The library preparations were sequenced on an Illumina Novaseq platform and 150 bp paired-end reads were generated.

Metabolomics profiling

Sample preparation. Samples were prepared by extracting metabolites through a chloroform/methanol/water system. In brief, sheared tissues were weighed and then 500 μL methanol with internal standards (including 50 μM L-methionine sulfone and 50 μM D-camphor-10-sulfonic acid for capillary electrophoresis-mass spectrometry [CE-MS] analysis; carnitine C2:0_d3 at 0.8 μg/mL, carnitine C8:0_d3 at 0.8 μg/mL, carnitine C16:0_d3 at 0.5 μg/mL, palmitic acid-d3 at 0.8 μg/mL, ceramide d18:1-d7/18:0 at 0.8 μg/mL, lyso-phospatidylcholine (LPC) 17:0-d5 at 0.8 μg/mL, phosphatidylcholine (PC) 17:0/22:4-d5 at 0.8 μg/mL, and triacylglycerol (TAG) 15:0/18:1/15:0-d5 at 0.8 μg/mL for liquid chromatography-mass spectrometry [LC-MS] analysis) were added. Mixed grinding apparatus (Scientz-48) was used for homogenization (35 Hz, 2 minutes) followed by the addition of 500 μL chloroform and vortex for 30 seconds. After phase breaking using 200 μL water and centrifugation (13,000 g, 4°C, 15 min), the resulting extract was divided into three fractions: one for CE-MS, one for carnitine and acyl-carnitines analysis by LC-MS, and one sample was used for LC-MS-based lipidomics. 300 μL hydrophilic layer was transferred for ultrafiltration through a 5-kDa cutoff filter (Millipore, cat. UFC3LCCNB-HMT). Simultaneously, the quality control sample was prepared by combining the aqueous phase from each sample and then filtered. Then samples were vacuum dried and stored at -80°C until CE-MS analysis. For carnitine and acyl-carnitines analysis, 150 μL hydrophilic layer and 100 μL hydrophobic layer were freeze-dried. Quality control sample was also prepared by combining the aqueous phase and then vacuum dried to evaluate the analytical quality. 300 μL hydrophobic layer was collected and freeze-dried for lipidomics analysis. At the same time, the quality control sample was prepared by combining the hydrophobic layer from each sample and then vacuum dried.

Mass spectrometry. CE-MS analysis was conducted on CE (G7100A, Agilent) couple to time of flight (TOF) mass spectrometry (G6224A, Agilent). The fused silica capillary (50 μm i.d. × 80 cm, Human Metabolome Technologies (HMT), Japan) was used for sample separation and the temperature of the capillary was at 20°C. Two analysis modes were performed. Detailed CE-MS methods were as previously described [43]. LC-MS analysis was performed by an ACQUITY UPLC system (Waters) coupled with a tripleTOF™ 5600 plus mass spectrometer (AB SCIEX). For acyl-carnitines analysis, the mobile phases consisted of phase A = water + 0.1% formic acid and phase B = acetonitrile + 0.1% formic acid. Development and validation of this rapid method was described before [44]. Lipidomics analysis was conducted through the C8 AQUITY column (2.1 mm × 100 mm × 1.7 μm, Waters, Milford, MA) and liquid chromatography was performed with phase A = 40:60 water: acetonitrile + 10 mM ammonium acetate and phase B = 90:10 2-propanol: acetonitrile + 10 mM ammonium acetate. Detailed LC-MS methods were performed as previously described [45]. In both ESI (+) and ESI (−) modes, TOF MS full scan and information-dependent acquisition (IDA) were performed in parallel to acquire high resolution MS and tandem-MS data simultaneously. In the positive mode, ion source gas 1 and gas 2 were set to 50 psi, curtain gas to 35 psi, temperature to 500°C, ion spray voltage floating (ISVF) to 5500 V, and collision

energy (CE) to 30 V with a collision energy spread (CES) of ± 15 V. In the negative mode, ion source gas 1 and gas 2 were set to 55 psi, curtain gas to 35 psi, temperature to 550°C, ISVF to − 4500 V, and CE to − 30 V with CES of ± 10 V. In the IDA setting, candidate ions with top 5 intensity were selected and subjected to high resolution tandem-MS analysis. All samples were randomized with respect to run order to avoid batch effects. Additionally, the quality control samples were identically inserted into the analytical sequence to monitor the reproducibility of the analytical method.

Metabolite Identification, Quantification, and Data Normalization. For CE-MS-based metabolites, the qualitative analysis of metabolites was performed using the pre-analyzed metabolite standard library (HMT), and internal standards were used to adjust the migration time and standardize the metabolite intensity. Peak extraction and identification were carried out with Quantitative Analysis Software (Agilent). Acyl-carnitines identification was based on the mass-to-charge ratio (m/z), retention time and MS/MS pattern. Lipid identification was based on exact mass and MS/MS pattern. The applied database search engines were HMDB (http://www.hmdb.ca/), Metlin (https://metlin.scripps.edu) and LIPID MAPS (http://www.lipidmaps.org/). Peakview workstation (AB SCIEX) was used to check MS/MS information of metabolites and Multiquant (AB SCIEX) was used to obtain the peak areas of identified metabolites. The raw data from CE-MS and LC-MS were normalized by corresponding internal standards and tissue weight to minimize errors arising from the sample pretreatment and analysis procedures as much as possible.

Proteomics and phospho-proteomics profiling

Sample preparation. The samples were homogenized in lysis buffer consisting of 2.5% SDS/100 mM Tris-HCl (pH 8.0). Then the samples were subjected to treatment with ultra sonication. After centrifugation, proteins in the supernatant were precipitated by adding 4 times of pre-cooled acetone. The protein pellet was dissolved in 8 M Urea/100 mM Tris-Cl. After centrifugation, the supernatant was used for the reduction reaction (10 mM DTT, 37°C for 1 h), followed by an alkylation reaction (40 mM iodoacetamide, room temperature/dark place for 30 min). Protein concentration was measured by the Bradford method. Urea was diluted below 2 M using 100 mM Tris-HCl (pH 8.0). Trypsin was added at a ratio of 1:50 (enzyme: protein, w/w) for overnight digestion at 37°C. The next day, TFA was used to bring the pH down to 6.0 to end the digestion. After centrifugation (12000×g, 15 min), the supernatant was subjected to peptide purification using Sep-Pak C18 desalting column. The peptide eluate was vacuum dried and stored at -20°C for later use. Phosphopeptide enrichment was performed totally according to a previous study [46].

LC-MS/MS analysis. LC-MS/MS data acquisition was carried out on an Orbitrap Exploris 480 mass spectrometer coupled with an Easy-nLC 1200 system. Peptides were loaded through auto-sampler and separated in a C18 analytical column (75μm × 25cm, C18, 1.9μm, 100Å). Mobile phase A (0.1% formic acid) and mobile phase B (80% ACN, 0.1% formic acid) were used to establish the separation gradient. A constant flow rate was set at 300 nL/min. For DDA mode analysis, each scan cycle consists of one full-scan mass spectrum (R = 60 K, AGC = 300%, max IT = 20 ms, scan range = 350–1500 m/z) followed by 20 MS/MS events (R = 15 K, AGC = 100%, max IT = auto, cycle time = 2 s). HCD collision energy was set to 30. Isolation window for precursor selection was set to 1.6 Da. Former target ion exclusion was set for 35 s.

Database search. MS raw data were analyzed with MaxQuant v1.6.6 using the Andromeda database search algorithm. Spectra files were searched against the UniProt Human proteome database using the following parameters: LFQ mode was checked for quantification; Variable modifications, Oxidation (M), Acetyl (Protein N-term) & Deamidation (NQ); Fixed modifications, Carbamidomethyl (C); Digestion, Trypsin/P; The MS1 match tolerance was set as 20 ppm for the first search and 4.5 ppm for the main search; the MS2 tolerance was set as 20 ppm; Match between runs was used for identification transfer. Search results were filtered with 1% FDR at both protein and peptide levels. Proteins denoted as decoy hits, contaminants, or only identified by sites were removed, the remaining identifications were used for further quantification analysis.

WES data analysis

Adaptors and low-quality reads of the WES sequencing data were removed by Trimmatic (v 0.39) [47], and the data quality was examined by fastqc (v 0.11.9) [48]. Then, the sequencing data were aligned to the human genome reference (GRCh38/hg38) using BWA (v 0.7.17) [49] and samtools (v 1.8) [50]. The somatic gene mutations of tumor samples with matched normal samples sequenced were called by the function of VarScan2 (v 2.4.4) [51], and the variants of tumor-only samples were called based on the function mpileup2cns of VarScan2 and stringent downstream filters. The called variants were annotated with Annovar [52] (version updated in 2020-06-08) according to multiple databases including refGene, knownGene, Exome Aggregation Consortium (ExAC03), Catalogue Of Somatic Mutations In Cancer (cosmic70), avsnp147, 1000 Genomes Project (2015_08), exome sequencing project (esp6500siv2_all) and clinvar_20220320.

To obtain high quality somatic variants for the tumor-only samples, stringent downstream filters were used. These filters included a base coverage of a minimum of 200 read depth and variant allele fraction (VAF) of 20% in tumor, and the variants should be at a frequency higher than 1% in the 1000 Genomes Project, ESP6500 or ExAC database, or present in the COSMIC with two or more occurrences.

TERT promoter mutation

The telomerase reverse transcriptase (TERT) promoter mutation (C228T/C250T) is determined using amplification-refractory mutation system quantitative polymerase chain reaction (ARMS-qPCR) as reported in the previous study [53].

RNA-seq data quantification

RNA-seq reads were aligned to the human genome reference (GRCh38/hg38) using HISAT2 (v2.0.5) [54]. HTseq (v 2.0.2) [55] was utilized to count the read numbers of each gene. Normalized gene expression matrix was obtained based on the counts function of DESeq2 [16] (v 1.26.0) with parameter normalized = TRUE.

Gene fusion

Gene fusions were detected from the RNA-seq data using arriba (v 2.3.0) [56] and STAR-fusion (v 1.4.0, https://github.com/STAR-Fusion/STAR-Fusion/wiki). The results from the two methods were further annotated and filtered base on annoFuseData (https://github.com/d3b-center/annoFuseData), and only fusions with JunctionReadCount > 3 and evaluated as high or median confidence were retained.

Differential expression analysis between tumor and normal samples

For the RNA-seq data, DESeq2 [16] (v 1.26.0) was applied to find the differentially expressed genes between paired tumor and normal samples. For the metabolism/proteomics/phospho-proteomics data, the differential expressions of each molecule was examined by Wilcox-Test (paired, two sided). Log2FCs between the paired tumor and normal samples were also calculated. P-values were adjusted by Benjamini & Hochberg method.

Multi-omics characterization of PTCs with different RRs.

For each RR type, we recognized the RR type associated molecules by comparing the expressions of molecules (metabolites, mRNAs, proteins, phosph-proteins) between this RR type and the other two RR types using Wilcox-Test (un-paired, two sides) and corresponding Log2FCs were calculated.

Integrative correlation analysis

The DIABLO [23] method (mixOmics R package, v 6.10.9) was applied to the four types of omics data (transcriptomics, metabolomics, proteomics, phosph-proteomics, and only molecules showed significant differences between tumor and normal tissues or specific type of RR were taken into account), with the samples covered by all four types of omics and labeled as high RR, intermediate RR, low RR and normal tissue. The DIABLO method aims to obtain the common information across multi-omics data by selecting a subset of molecules which not only maximize the inter-correlations among omics but also discriminate between different phenotypic labels. The expressional matrixes and the labels were taken as the input of DIABLO, the latent component number was set as 3, and the number of representative molecules to select for each latent component considering each type of input omics data was set as 20.

Integrative clustering of PTC patients

Firstly, we tried to cluster the PTC patients based on both transcriptomics and metabolomics profiles. Here, ten different multi-omics clustering methods including SNF, PINSPlus, NEMO, COCA, LRAcluster, ConsensusClustering, CIMLR, MoCluster, iClusterBayes, IntNMF were performed on our data using the MOVICS [57] R package (v 0.99.17). These methods generated ten clustering records, and a similarity matrix describing to what extant different samples were grouped into the same clusters in terms of the ten clustering records was obtained. Then, hierarchical clustering algorithm was applied on the consensus similarity matrix, and the number of clusters was set as 4 (to ensure each cluster with was with more than 10 samples).

TDS, BRAF scores and RAS scores

We calculated the mean log2-transformed expression levels of 16 thyroid function relevant genes defined by the TCGA-PTC study [6] as the TDS scores. Similarly, the BRAF and RAS scores were calculated based on the mean expression of the upregulated signature genes in the BRAFV600E-mutated and RAS-mutated samples from the TCGA-PTC mRNA expressions.

Single cell RNA-seq based prediction of cell compositions in the PTC bulk sample RNA-seq data

The single cell RNA-seq data of PTC samples (GSE184362, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184362) as well as the annotated cell types were utilized as the input of dampened weighted least squares (DWLS R package, v 0.1.0, https://CRAN.R-project.org/package=DWLS) to train a cell decovolution model. Then, the trained DWLS model was applied on the transcriptomics data of the PTC bulk tissue samples to estimate the potential cell compositions of each tissue sample and a cell type composition matrix was obtained.

Identification of druggable targets of different subtypes

For each subtype, the subtype associated genes and metabolites were recognized by examining the differential expressions between one subtype and the other three subtypes using DESeq2 for transcriptomics and limma for metabolomics. Then, the top-10 significant and specifically highly expressed genes for each subtype was recognized using the runMarker function of the MOVICS package. Then, druggable targets among the top-ranked subtype associated genes were selected based on the DGIdb database [58].

Subtype prediction model

The transcriptomics and metabolomics data of the 97 PTCs were partitioned into training (60%) and testing datasets (40%). Importances of the metabolites and genes in predicting the subtypes were estimated based on random forest (RF) [59] method using the randomForest R package (v 4.6–14). Then, a subtype predictor was trained based on the expressional profiles of the top-30 metabolites and top-30 genes using the RF method. The model was applied on the testing dataset, and the prediction performance was evaluated by ROC curves.

Statistical analysis

Detailed computational and statistical methods are reported in the Methods or figure legends. All statistical analyses were performed by R (v 3.6.3 and v 4.0.4). The statistical tests were two-sided by default, and one-sided tests were specifically stated.

Declarations

Acknowledgement

This study is supported by the National Natural Science Foundation of China Grants (No.81972625 No.82072951, No.82174133, No.82203052), Liaoning Revitalization Talents Program (XLYC2002035), Liaoning Science and Technology Innovation Funding (20230101-JH2/1013), Science and Technology Innovation Fund (Youth Science and Technology Star) of Dalian (No. 2021RQ009), Innovation program of science and research from the DICP, CAS (DICP I202129, DICP I202209).

Author contribution

Yu W., Q. J., H. H., H-l.P. and R. S. conceived the project, H-l.P. and R.S. supervised the project. N. Q. and B. M. designed the project and performed most of experiments, D. C. performed the computational data analysis. N. Q., B. M., Yuting W., Z. N, T. L., J. X. and Yulong W. collected clinical samples and the relevant clinical information. L. Z., H. W., S. J., D. X. and W. Wu helped to analyze the results from a clinical perspective. W. Wang conducted the metabolism profiling experiments. N. Q., D. C, B. M. and H-l.P. wrote the manuscript with input from all other authors.

Competing Interests

The authors have declared that no competing interest exist.

References

1. Duh, Q.Y., et al., *Back so soon? Is early recurrence of papillary thyroid cancer really just persistent disease? Discussion*. Surgery, 2018. 163(1): p. 123–123.
2. Hay, I.D., R.H. Grogan, and Q.Y. Duh, *A study of recurrence and death from papillary thyroid cancer with 27 years of median follow-up DISCUSSION*. Surgery, 2013. 154(6): p. 1446–1447.
3. Xu, S., et al., *Predictive Value of Serum Thyroglobulin for Structural Recurrence Following Lobectomy for Papillary Thyroid Carcinoma*. Thyroid, 2021. 31(9): p. 1391–1399.
4. Chien, M.N., et al., *Recurrence-associated genes in papillary thyroid cancer: An analysis of data from The Cancer Genome Atlas*. Surgery, 2017. 161(6): p. 1642–1650.
5. Nieto, H.R., et al., *Recurrence of Papillary Thyroid Cancer: A Systematic Appraisal of Risk Factors*. J Clin Endocrinol Metab, 2022. 107(5): p. 1392–1406.
6. Cancer Genome Atlas Research, N., *Integrated genomic characterization of papillary thyroid carcinoma*. Cell, 2014. 159(3): p. 676–90.
7. Jiang, N., et al., *Plasma Lipidomics Profiling Reveals Biomarkers for Papillary Thyroid Cancer Diagnosis*. Front Cell Dev Biol, 2021. 9: p. 682269.
8. Abdullah, M.I., et al., *Tissue and serum samples of patients with papillary thyroid cancer with and without benign background demonstrate different altered expression of proteins*. PeerJ, 2016. 4: p. e2450.
9. Ciavardelli, D., et al., *Metabolic Alterations of Thyroid Cancer as Potential Therapeutic Targets.* Biomed Research International, 2017. 2017.

10. Du, Y., et al., *Serum Metabolomics Study of Papillary Thyroid Carcinoma Based on HPLC-Q-TOF-MS/MS*. Front Cell Dev Biol, 2021. 9: p. 593510.
11. Abdullah, M.I., et al., *Papillary Thyroid Cancer: Genetic Alterations and Molecular Biomarker Investigations*. Int J Med Sci, 2019. 16(3): p. 450–460.
12. Li, M., et al., *Genomic characterization of high-recurrence risk papillary thyroid carcinoma in a southern Chinese population*. Diagnostic Pathology, 2020. 15(1).
13. Agrawal, N., et al., *Integrated Genomic Characterization of Papillary Thyroid Carcinoma*. Cell, 2014. 159(3): p. 676–690.
14. Nikiforov, Y.E. and M.N. Nikiforova, *Molecular genetics and diagnosis of thyroid cancer*. Nature Reviews Endocrinology, 2011. 7(10): p. 569–580.
15. Yakushina, V.D., L.V. Lerner, and A.V. Lavrov, Gene Fusions in Thyroid Cancer. Thyroid, 2018. 28(2): p. 158–167.
16. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome Biology, 2014. 15(12).
17. Leng, J., et al., *Application of isotope-based carboxy group derivatization in LC-MS/MS analysis of tissue free-fatty acids for thyroid carcinoma*. J Pharm Biomed Anal, 2013. 84: p. 256–62.
18. Shi, X., et al., *Integrated proteogenomic characterization of medullary thyroid carcinoma*. Cell Discov, 2022. 8(1): p. 120.
19. Geng, Q.S., et al., *Over-Expression and Prognostic Significance of FN1, Correlating With Immune Infiltrates in Thyroid Cancer*. Front Med (Lausanne), 2021. 8: p. 812278.
20. Ruf, J. and P. Carayon, *Structural and functional aspects of thyroid peroxidase*. Arch Biochem Biophys, 2006. 445(2): p. 269–77.
21. Sun, Q.Y., H.H. Zhou, and X.Y. Mao, *Emerging Roles of 5-Lipoxygenase Phosphorylation in Inflammation and Cell Death.* Oxidative Medicine and Cellular Longevity, 2019. 2019.
22. Kummer, N.T., et al., *Arachidonate 5 lipoxygenase expression in papillary thyroid carcinoma promotes invasion via MMP-9 induction*. Journal of Cellular Biochemistry, 2012. 113(6): p. 1998–2008.
23. Singh, A., et al., *DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays*. Bioinformatics, 2019. 35(17): p. 3055–3062.
24. Shao, X.H., et al., *MatrisomeDB 2.0: 2023 updates to the ECM-protein knowledge database*. Nucleic Acids Research, 2022.
25. Carbon, S., et al., *The Gene Ontology resource: enriching a GOld mine*. Nucleic Acids Research, 2021. 49(D1): p. D325-D334.
26. Wang, Y.C., et al., *Complement C4-A and Plasminogen as Potential Biomarkers for Prediction of Papillary Thyroid Carcinoma*. Frontiers in Endocrinology, 2021. 12.
27. Zhang, W., et al., *Identification of novel immune-related molecular subtypes and a prognosis model to predict thyroid cancer prognosis and drug resistance*. Frontiers in Pharmacology, 2023. 14.

28. Lee, K.M., et al., *ECM1 regulates tumor metastasis and CSC-like property through stabilization of beta-catenin*. Oncogene, 2015. 34(50): p. 6055–65.
29. Pu, W.L., et al., *Single-cell transcriptomic analysis of the tumor ecosystems underlying initiation and progression of papillary thyroid carcinoma*. Nature Communications, 2021. 12(1).
30. Li, G.Z., et al., *Carbonic Anhydrase XII is a Clinically Significant, Molecular Tumor-Subtype Specific Therapeutic Target in Glioma with the Potential to Combat Invasion of Brain Tumor Cells*. Oncotargets and Therapy, 2021. 14: p. 1707–1718.
31. Li, S.R., et al., *Targeting the GCK pathway: a novel and selective therapeutic strategy against RAS-mutated multiple myeloma*. Blood, 2021. 137(13): p. 1754–1764.
32. Lin, H.S., et al., *Protein arginine methyltransferase 8 gene enhances the colon cancer stem cell (CSC) function by upregulating the pluripotency transcription factor*. Journal of Cancer, 2018. 9(8): p. 1394–1402.
33. Burgos, J.R., B.M. Iresjo, and U. Smedh, *MCG101-induced cancer anorexia-cachexia features altered expression of hypothalamic Nucb2 and Cartpt and increased plasma levels of cocaine- and amphetamine-regulated transcript peptides.* Oncology Reports, 2016. 35(4): p. 2425–2430.
34. Burgos, M., et al., *Prognostic value of the immune target CEACAM6 in cancer: a meta-analysis*. Ther Adv Med Oncol, 2022. 14: p. 17588359211072621.
35. Li, X., et al., *Association of MUC16 Mutation With Tumor Mutation Load and Outcomes in Patients With Gastric Cancer*. JAMA Oncol, 2018. 4(12): p. 1691–1698.
36. Ferrer, V.P., *MUC16 mutation is associated with tumor grade, clinical features, and prognosis in glioma patients*. Cancer Genetics, 2023. 270: p. 22–30.
37. Wang, Z., et al., *Effect of MUC16 mutations on tumor mutation burden and its potential prognostic significance for cutaneous melanoma*. Am J Transl Res, 2022. 14(2): p. 849–862.
38. Zhang, L., X. Han, and Y. Shi, *Association of MUC16 Mutation With Response to Immune Checkpoint Inhibitors in Solid Tumors*. JAMA Netw Open, 2020. 3(8): p. e2013201.
39. Wang, J.R., et al., *Expression of MMP-13 is associated with invasion and metastasis of papillary thyroid carcinoma*. Eur Rev Med Pharmacol Sci, 2013. 17(4): p. 427–35.
40. Ding, J., et al., *Silencing of cystatin SN abrogates cancer progression and stem cell properties in papillary thyroid carcinoma*. FEBS Open Bio, 2021. 11(8): p. 2186–97.
41. Tong, Y., et al., *Radiogenomic Analysis of Papillary Thyroid Carcinoma for Prediction of Cervical Lymph Node Metastasis: A Preliminary Study*. Front Oncol, 2021. 11: p. 682998.
42. Haugen, B.R., et al., 2015 *American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer.* Thyroid, 2016. 26(1): p. 1-133.
43. Zeng, J., et al., *Metabolomics study of hepatocellular carcinoma: discovery and validation of serum potential biomarkers by using capillary electrophoresis-mass spectrometry*. J Proteome Res, 2014.

13(7): p. 3420–31.

44. Ouyang, Y., et al., *A high throughput metabolomics method and its application in female serum samples in a normal menstrual cycle based on liquid chromatography-mass spectrometry*. Talanta, 2018. 185: p. 483–490.
45. Xuan, Q., et al., *Development of a High Coverage Pseudotargeted Lipidomics Method Based on Ultra-High Performance Liquid Chromatography-Mass Spectrometry*. Anal Chem, 2018. 90(12): p. 7608–7616.
46. Yao, Y.T., et al., *An immobilized titanium (IV) ion affinity chromatography adsorbent for solid phase extraction of phosphopeptides for phosphoproteome analysis*. Journal of Chromatography A, 2017. 1498: p. 22–28.
47. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. 30(15): p. 2114–20.
48. Andrews, S., *FASTQC. A quality control tool for high throughput sequence data*. 2010.
49. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. 25(14): p. 1754–60.
50. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. 25(16): p. 2078–9.
51. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing*. Genome Res, 2012. 22(3): p. 568–76.
52. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data*. Nucleic Acids Res, 2010. 38(16): p. e164.
53. Yu, P.C., et al., *Arms-qPCR Improves Detection Sensitivity of Earlier Diagnosis of Papillary Thyroid Cancers With Worse Prognosis Determined by Coexisting BRAF V600E and Tert Promoter Mutations*. Endocr Pract, 2021. 27(7): p. 698–705.
54. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements*. Nat Methods, 2015. 12(4): p. 357–60.
55. Anders, S., P.T. Pyl, and W. Huber, *HTSeq–a Python framework to work with high-throughput sequencing data*. Bioinformatics, 2015. 31(2): p. 166–9.
56. Uhrig, S., et al., *Accurate and efficient detection of gene fusions from RNA sequencing data*. Genome Res, 2021. 31(3): p. 448–460.
57. Lu, X., et al., *MOVICS: an R package for multi-omics integration and visualization in cancer subtyping*. Bioinformatics, 2021. 36(22–23): p. 5539–5541.
58. Griffith, M., et al., *DGIdb: mining the druggable genome*. Nat Methods, 2013. 10(12): p. 1209–10.
59. Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5–32.

Figures

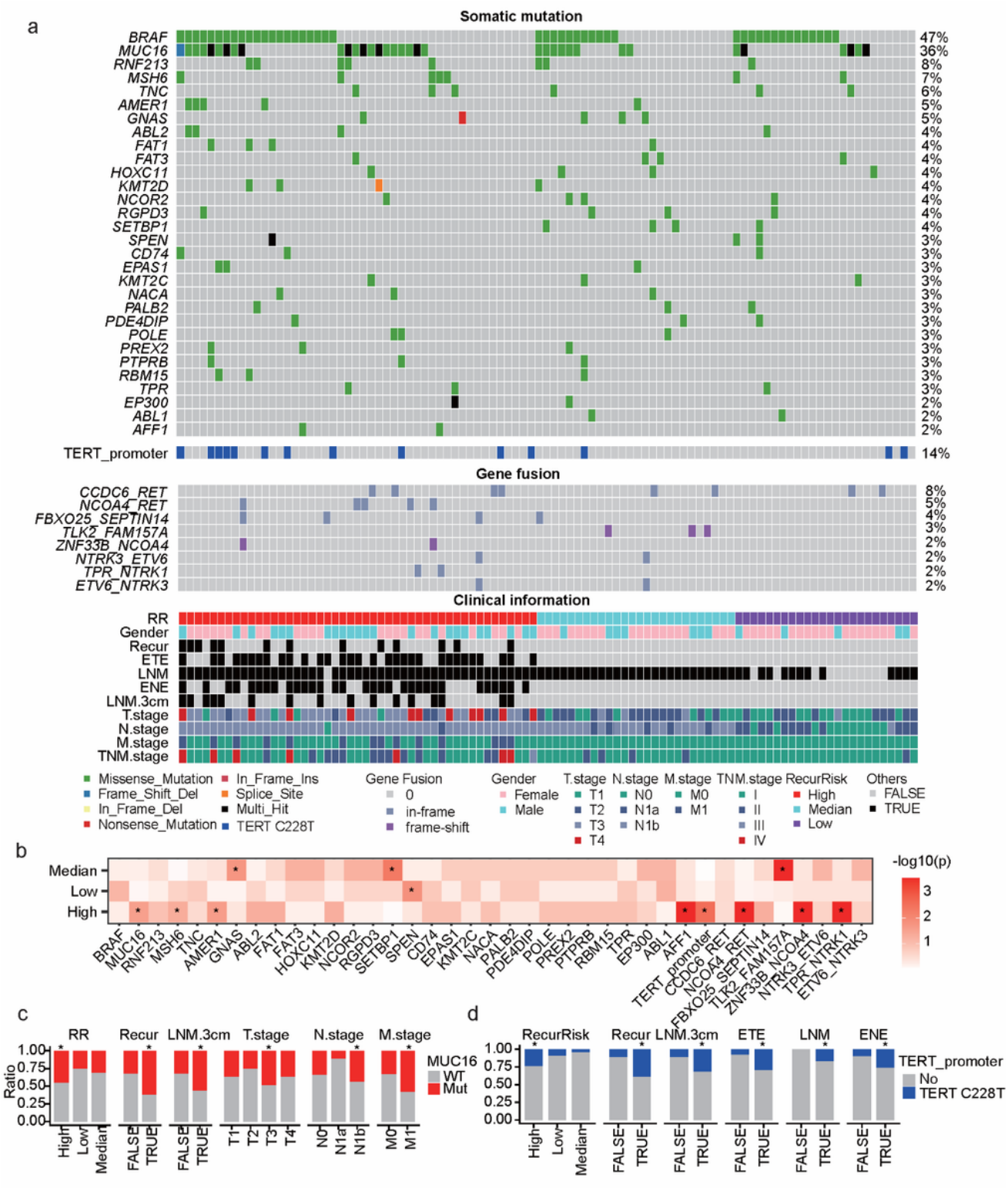

Figure 1

Genetic profile of the PTC patients with different recurrent risk

a. Genetic profile and associated clinical information of 97 PTC patients.

b. Mutations with significant enrichment in one type of RR.

c-d. Clinical features showed significant enrichment in the samples with mutations in *MUC16*(c) or *TERT* promoter(c). Overlap significance was examined by hypergeometric distribution. *: P<0.05.

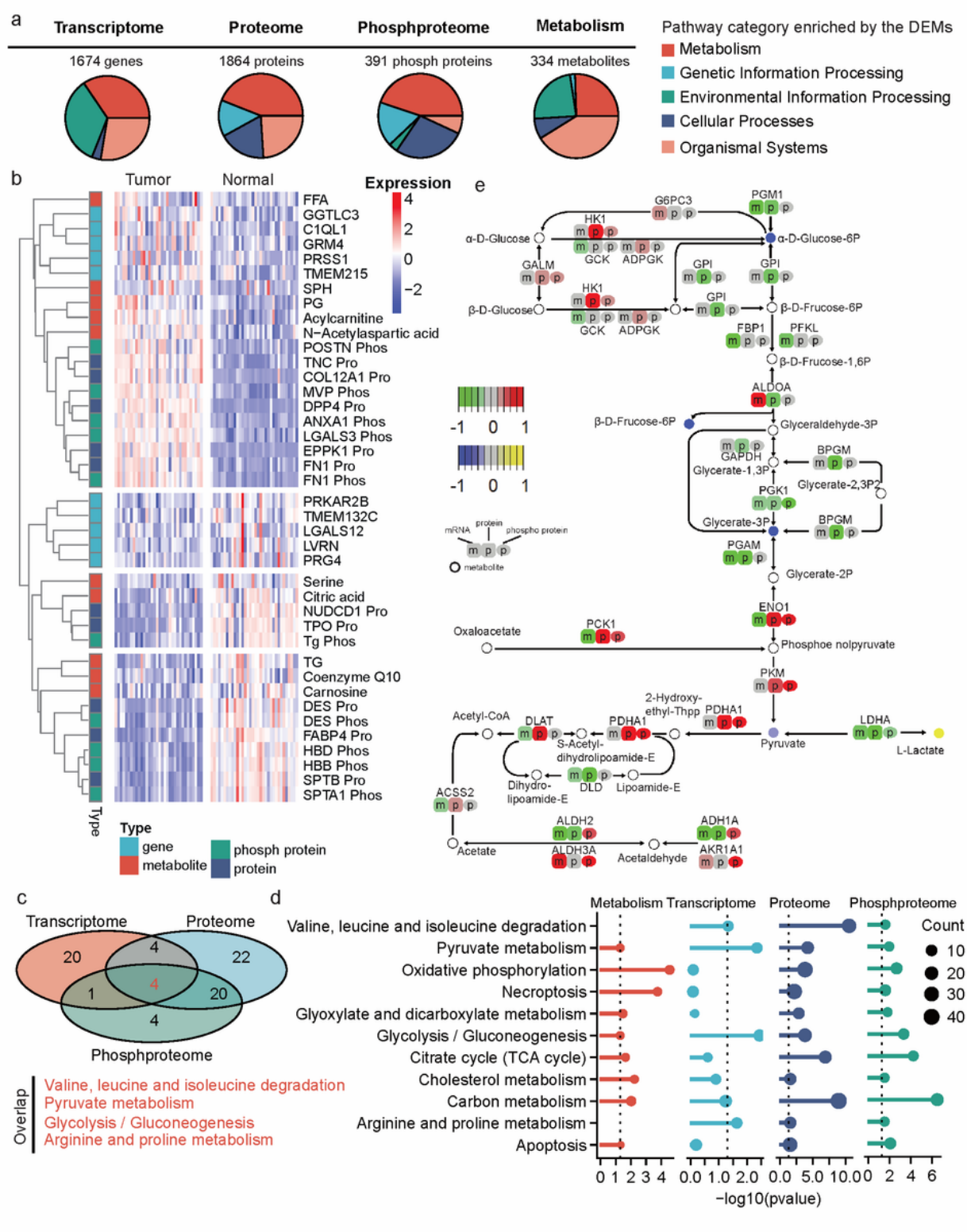

Figure 2

Multi-omics based profiling of the tumor and normal thyroid samples

a. Pathway categories enriched by the DEMs.

b. Heatmap showing the top-rank DEMs. Only molecules with the top-10 significant p-values for each omics type are listed. The metabolite levels for TG, FFA, PG and SPH were the summarized abundances of different kinds of TG, FFA, PG and SPH. The suffix Pro and Phos respectively represent proteins and phospho-proteins.

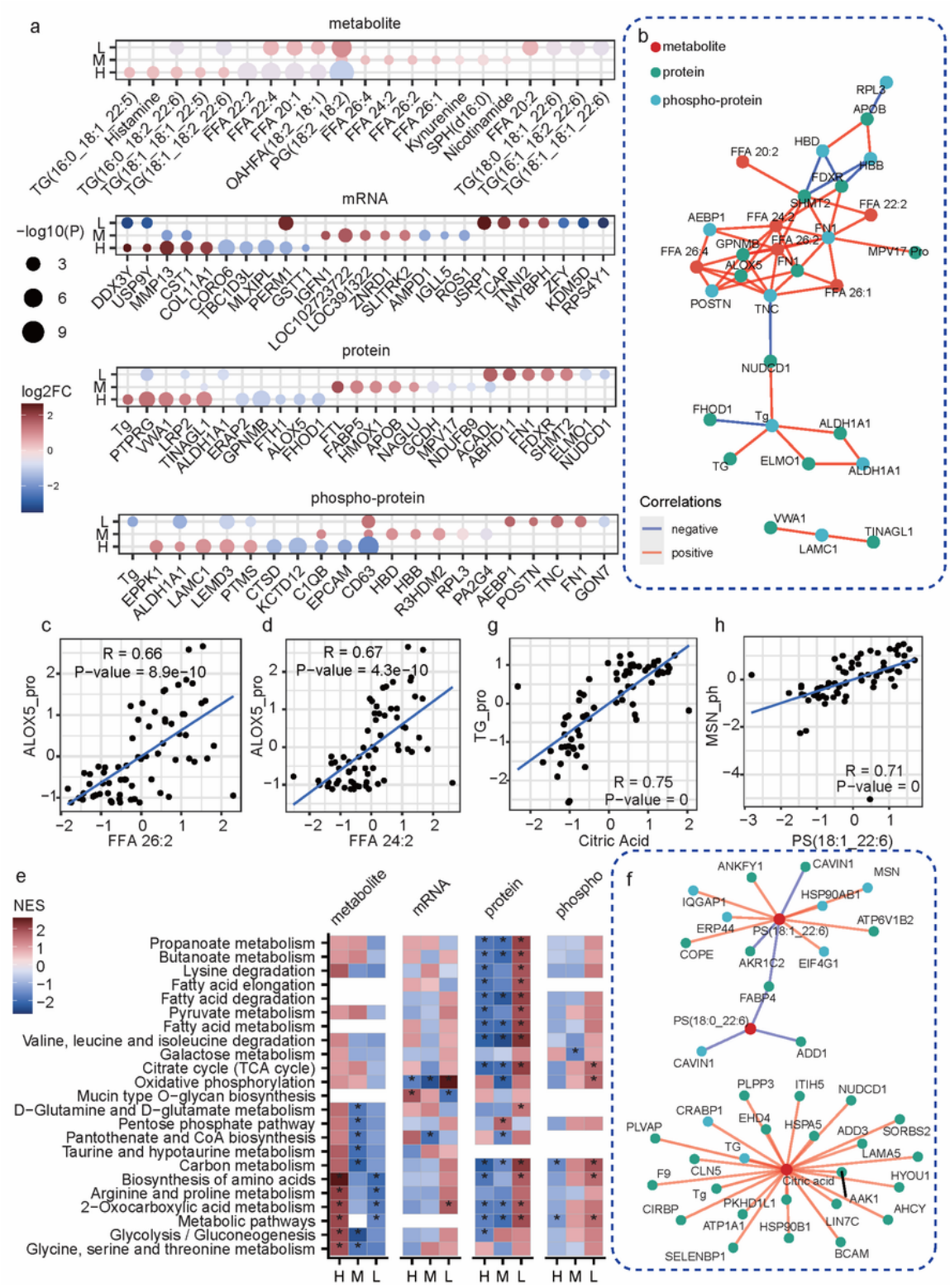

Figure 3

Multi-omics landscape of different recurrent risk

a. Molecules significant differential expressions in patients with different RR. H: high RR, M: intermediate RR; L: low RR.

b. Correlation network of the molecules showed significant differential expressions among different RRs.

c-d. Scatter plot showing the correlations between protein ALOX5 and metabolite FFA 26:2 (c) and FFA 24:2(d).

e. Metabolism pathway enrichment results of the RR relevant molecules. Fisher exact test, *: $P<0.05$;

f. Correlation network of the molecules. Only metabolites in the down-regulated metabolism pathways for the high-risk samples concerning proteomics were shown. The displayed edges are with correlation coefficients larger than 0.5.

g-h. Scatter plot showing the correlations between metabolite citric acid and protein TG (g), metabolite PS (18:1_22:6) and phospho-protein MSN (h).

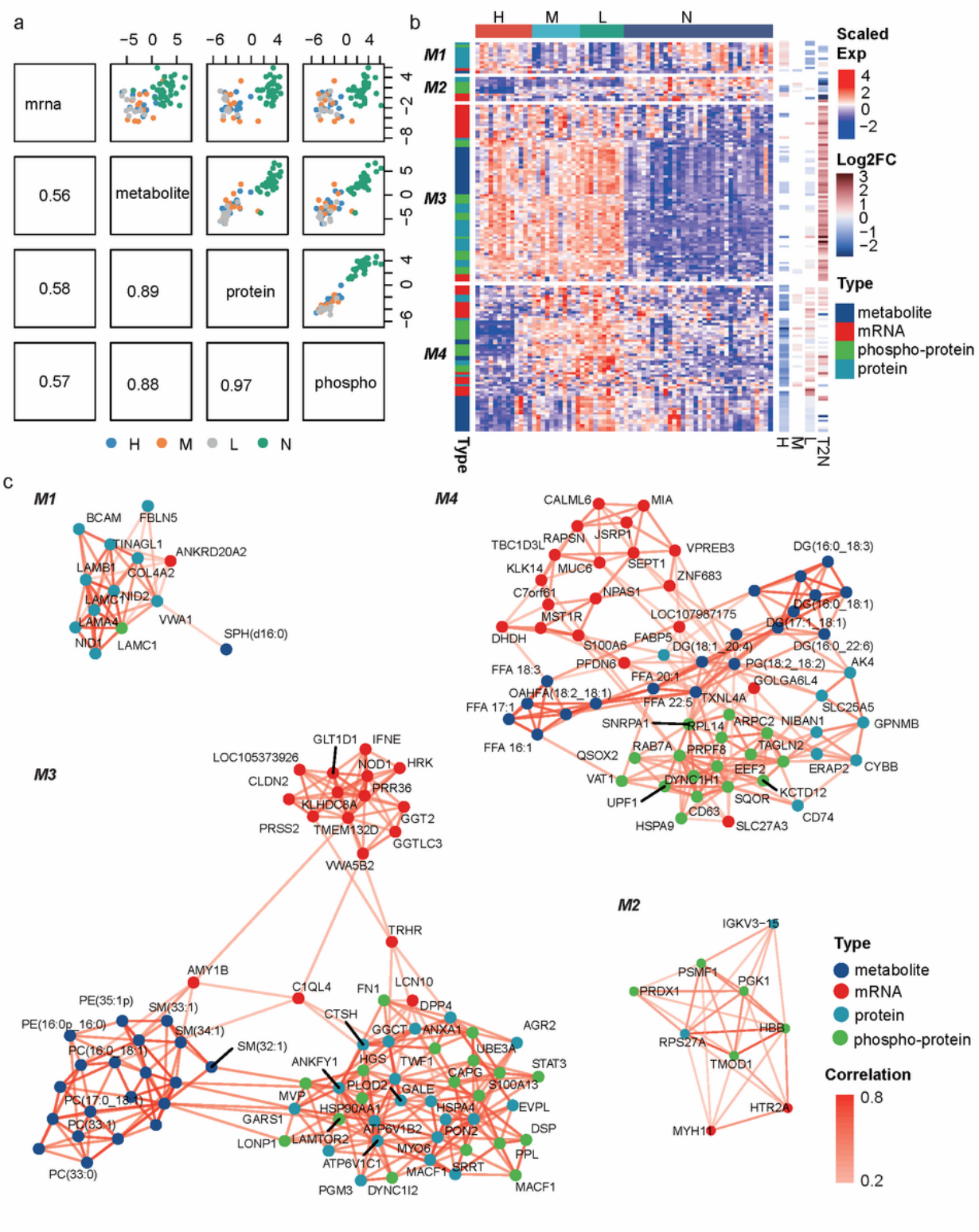

Figure 4

Correlations between different types of omics

a. Correlation between the first DIABLO component determined by different types of omics.

b. Heatmap of the most contributing molecules for the first two components. The four columns on the right showed the log2FC calculated for comparing different groups. H: high compared to intermediate and low risks; M: intermediate compared to high and low risks; L: low compared to high and intermediate risks; T2N: tumor compared to paired normal samples.

c. Correlation network for four molecule modules (M1 to M4) based on the multi-omics.

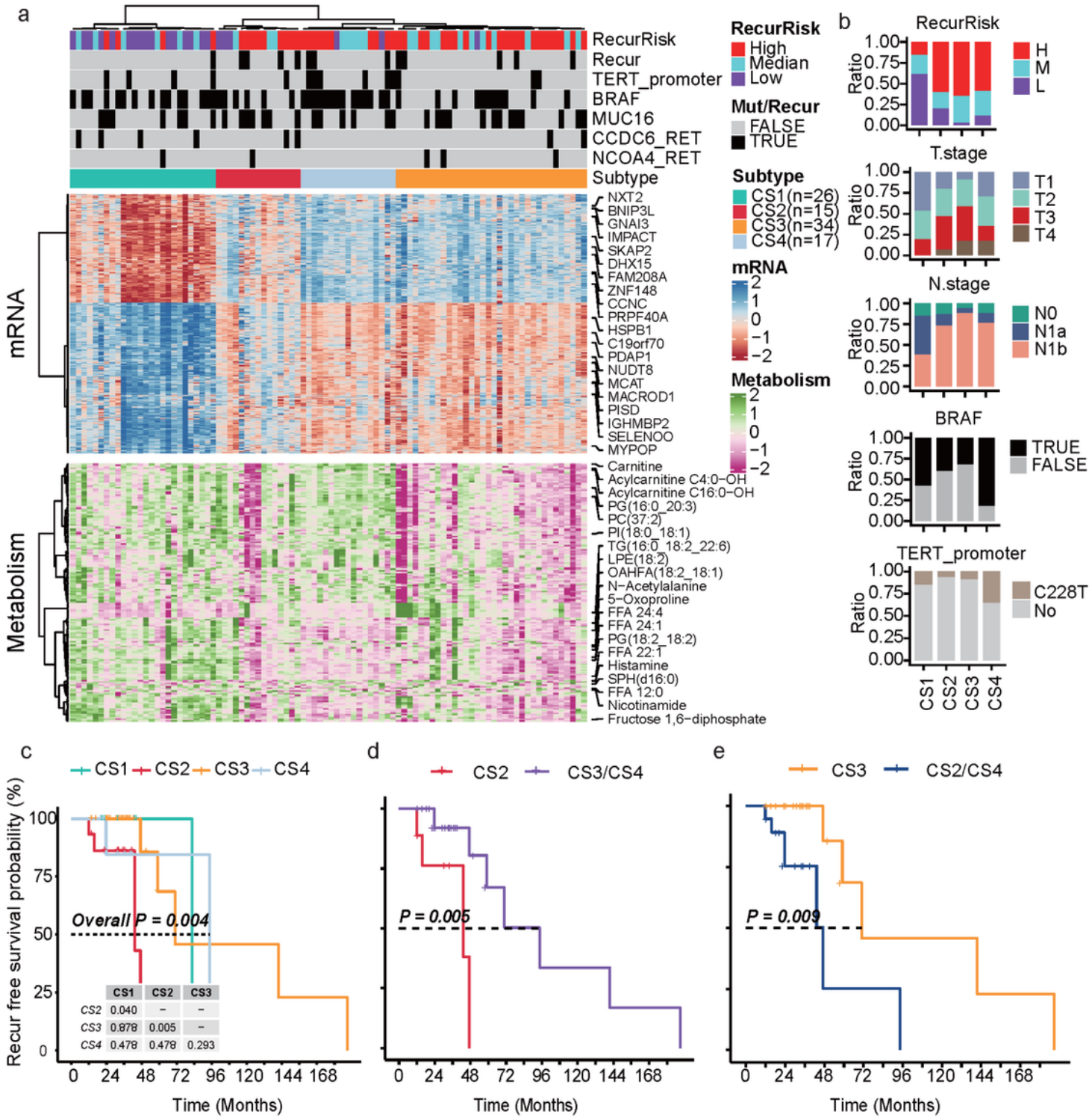

Figure 5

Integrative clustering of the PTC patients based on the transcriptome and metabolomics profile

a. Heatmap showing the clustering results of the PTC patients based on both transcriptomics and metabolomics.

b. Bar plot of the clinical features with significantly different distribution across the four clusters.

c. Kaplan Meier (KM) plot of the recurrence free survival probability of the patients in different clusters

d. KM-plot of the patients in the cluster CS2 and the other two high-risk enriched subtypes

e. KM-plot of the patients in the cluster CS3 and the other two high-risk enriched subtypes.

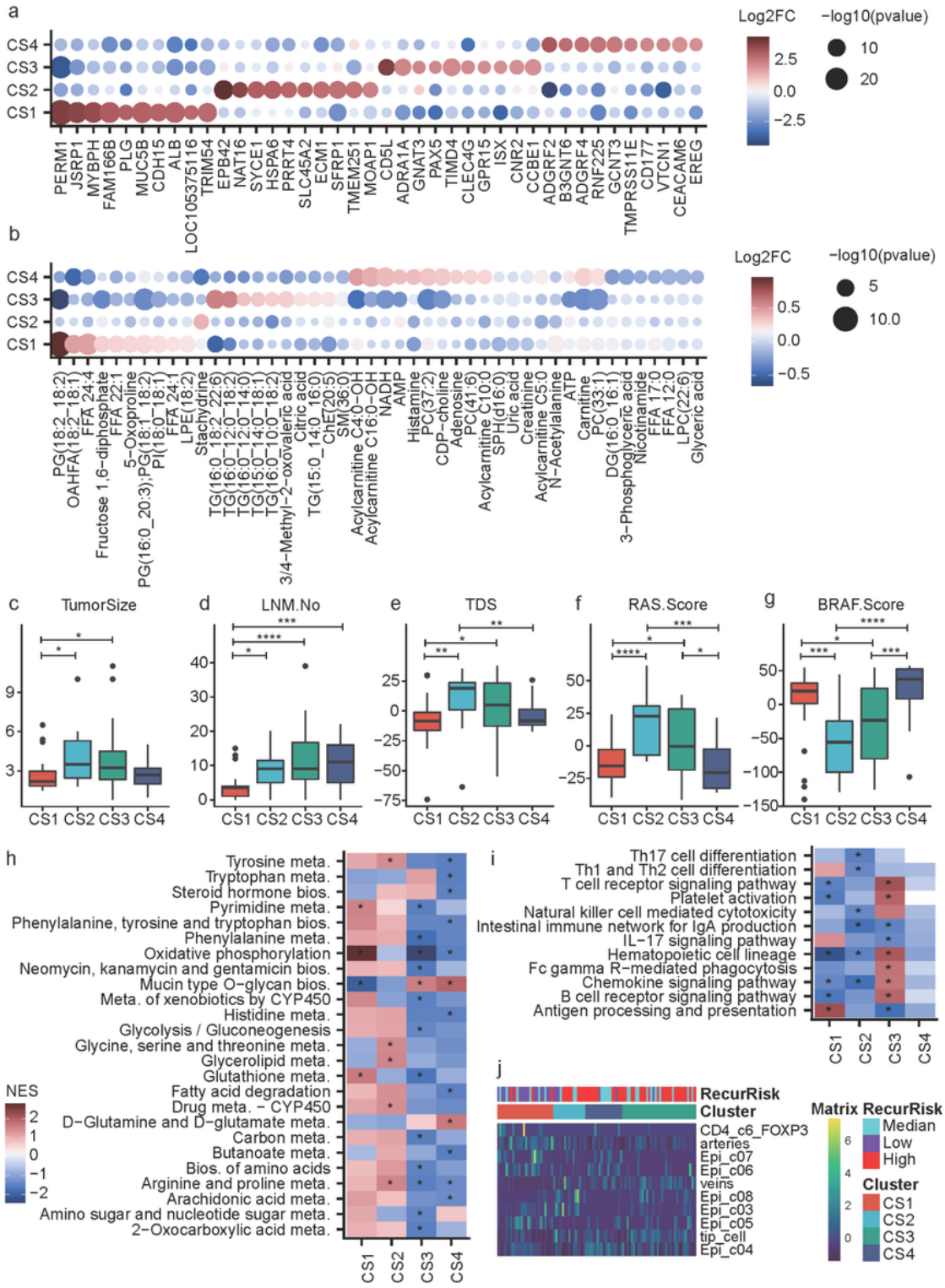

Figure 6

Characterization of the four PTC subtypes

a-b. Significantly up-regulated genes (a) and metabolites (b) in different clusters based on transcriptomics and metabolomics.

c-g. Boxplot of tumor size (c), number of metastatic lymph nodes (LNM.No) (d), tumor difference scores (TDS) (e), RAS score (f) and BRAF score (g) across the four clusters. T-test, *: P<0.05, **:P<0.01, ***:P<0.001, ****:P<0.0001

h-i. The metabolism (h) and immune pathway (i) enrichment results for the four clusters.

j. Heatmap showing different cell compositions of different clusters. Epi: epithelium. Only cell types showing significant differences (Kruskal test, P<0.05) for at least one subtype were displayed.

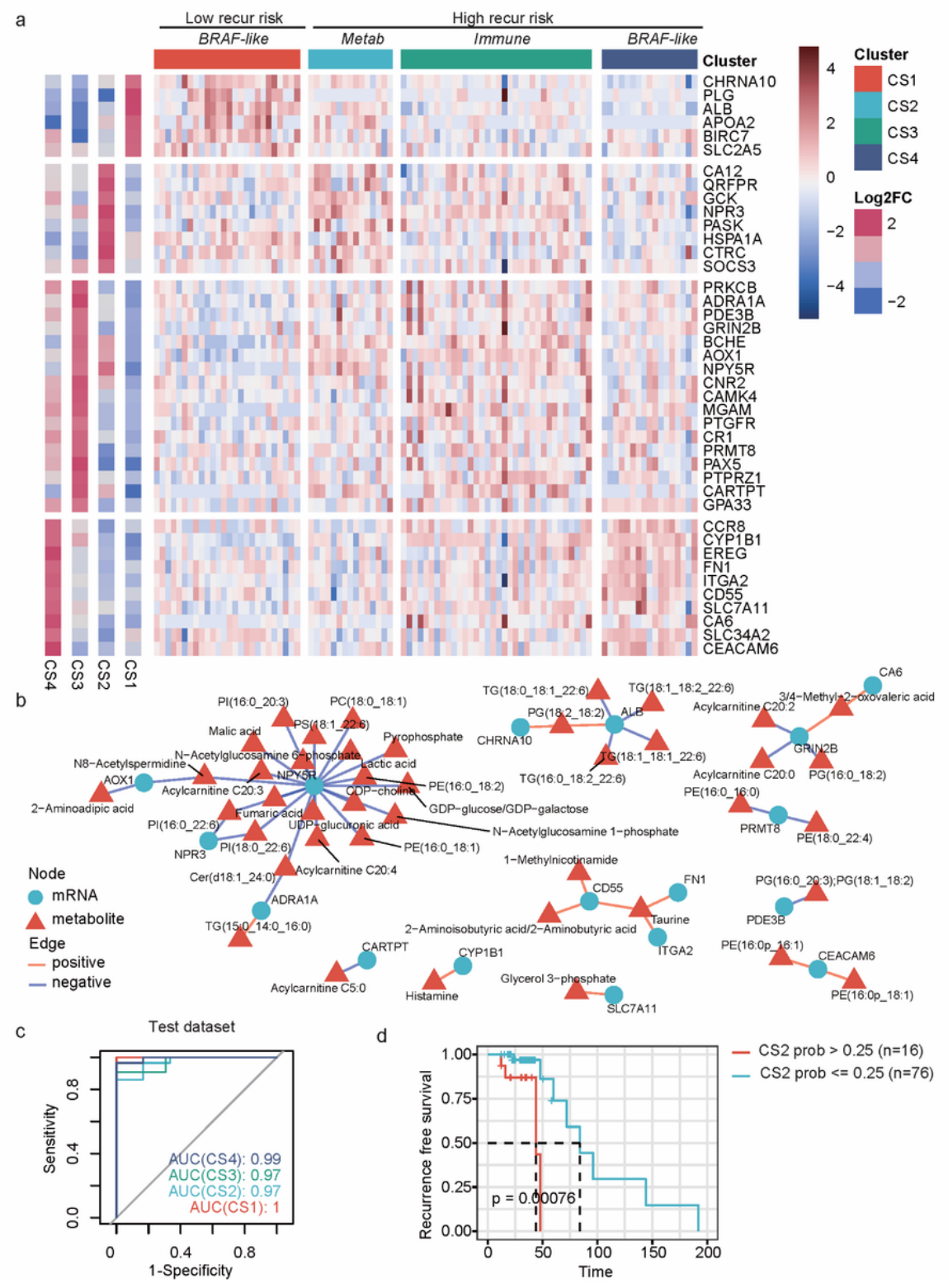

Figure 7

Potential biomarkers or targets of different subtypes

a. Heatmap of mRNA expression of the over-expressed druggable targets in the four clusters.

b. Correlations between metabolites and the druggable targets.

c. Receiver operator characteristic (ROC) curve for predictions of CS1 to CS4 subtypes in the test dataset.

d. KM-plot of the recurrence free survival curves of the PTC patients predicted with high and low CS2 probabilities (threshold was set as 0.25, since there were four subtypes).

Supplementary Files

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

- Supplementarymaterials20230522.docx