รายงานวิจัยฉบับสมบูรณ์

โครงการ

“การวิเคราะห์โปรตีโอมิกส์ของมะเร็งร่วมบริเวณคอหอยส่วนล่างและหลอดอาหารในผู้ป่วยภาคใต้ของไทย เพื่อหาตัวบ่งชี้ทางชีวภาพ”

“Identification of candidate biomarkers for synchronous hypopharyngeal and esophageal cancer in Southern Thais using oncoproteomics analysis”

โดย ผศ.ดร.พญ. ณฐินี จินาวัฒน์
สังกัด คณะแพทยศาสตร์ รพ.รามาธิบดี มหาวิทยาลัยมหิดล

29 มีนาคม 2562

สัญญาเลขที่ RSA5780065

รายงานวิจัยฉบับสมบูรณ์

โครงการ
“การวิเคราะห์โปรตีโอมิกส์ของมะเร็งร่วมบริเวณคอหอยส่วนล่างและหลอดอาหารในผู้ป่วยภาคใต้ของไทย เพื่อหาตัวบ่งชี้ทางชีวภาพ”
“Identification of candidate biomarkers for synchronous hypopharyngeal and esophageal cancer in Southern Thais using oncoproteomics analysis”

ผู้วิจัย ผศ.ดร.พญ.ณฐินี จินาวัฒน์
สังกัด คณะแพทยศาสตร์ รพ.รามาธิบดี มหาวิทยาลัยมหิดล

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย
และ มหาวิทยาลัยมหิดล

กิตติกรรมประกาศ

โครงการวิจัยนี้สำเร็จได้ด้วยความช่วยเหลือและการสนับสนุนอย่างดียิ่งจากบุคคลและหน่วยงานหลายฝ่ายดังต่อไปนี้

ขอขอบคุณ คณะอาจารย์, นักวิจัย, นักวิทยาศาสตร์ จากหน่วยงานและห้องปฏิบัติการต่างๆ ของคณะแพทยศาสตร์โรงพยาบาลรามาธิบดี และคณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์ คณะแพทยศาสตร์จุฬาลงกรณ์มหาวิทยาลัย ที่เข้าร่วมดำเนินการวิจัยเป็นอย่างดี เอื้อเฟื้อสถานที่และทรัพยากรในการศึกษาวิจัย สนับสนุนข้อมูลทางคลินิก และตัวอย่างตรวจสำหรับการวิจัย ขอขอบคุณ ผู้ป่วยมะเร็งทุกท่านที่เสียสละเข้าร่วมโครงการวิจัย

สุดท้ายนี้ขอขอบคุณ สำนักงานกองทุนสนับสนุนการวิจัย และมหาวิทยาลัยมหิดล ที่ได้สนับสนุนทุนวิจัย และการดำเนินงานของโครงการฯเป็นอย่างดีตลอดมา

บทคัดย่อ

รหัสโครงการ: RSA5780065
ชื่อโครงการ:“การวิเคราะห์โปรตีโอมิกส์ของมะเร็งร่วมบริเวณคอหอยส่วนล่าง และหลอดอาหารในผู้ป่วยภาคใต้ของไทยเพื่อหาตัวบ่งชี้ทางชีวภาพ”
ชื่อนักวิจัยและสถาบัน: ผศ.ดร.พญ.ณฐินี จินาวัฒน์ คณะแพทยศาสตร์ รพ.รามาธิบดี มหาวิทยาลัยมหิดล
อีเมล์: jnatini@hotmail.com
ระยะเวลาโครงการ: 16 มิถุนายน 2557 - 29 มีนาคม 2562

บทคัดย่อ: ผู้ป่วยมะเร็งศีรษะและลำคอ มีโอกาสเสี่ยงสูงที่จะเกิดมะเร็งตำแหน่งที่สองตามมา ผู้ป่วยที่มีมะเร็งตำแหน่งที่สองนี้จะมีการพยากรณ์โรคที่แย่กว่า และอัตรารอดชีวิตที่ต่ำกว่า เมื่อเปรียบเทียบกับผู้ป่วยที่ไม่มีมะเร็งตำแหน่งที่สอง การศึกษานี้มีวัตถุประสงค์เพื่อศึกษาตัวชี้วัดทางชีวภาพสำหรับพยากรณ์โอกาสเสี่ยงในการเกิดมะเร็งตำแหน่งที่สองในผู้ป่วยมะเร็งศีรษะและลำคอ โดยใช้ชิ้นเนื้อพาราฟินบล็อกของก้อนเนื้องอกตำแหน่งแรก เริ่มแรกได้ทำการศึกษาโปรตีนทั้งหมดในเซลล์ ด้วยวิธีแมสสเปกโตรเมทรีจากชิ้นเนื้อมะเร็งและชิ้นเนื้อปกติจำนวน 16 คู่ เพื่อค้นหาโปรตีนที่มีการแสดงออกต่างกันในผู้ป่วยที่มีหรือไม่มีมะเร็งตำแหน่งที่สอง จากนั้นได้คัดเลือกโปรตีนที่มีความสำคัญและมีแนวโน้มที่จะพัฒนาต่อเป็นตัวบ่งชี้วัดทางชีวภาพมาตรวจสอบประสิทธิภาพในกลุ่มผู้ป่วยใหม่จำนวน 49 คน ที่ไม่เกี่ยวข้องกับกลุ่มเดิมที่ศึกษาไปแล้ว โดยศึกษาการแสดงออกของยีนที่คัดเลือกมาด้วยวิธีนาโนสตริง และใช้กรรมวิธีทางชีวสถิติ เพื่อคัดเลือกกลุ่มของยีนที่มีความสำคัญ ผลการศึกษาพบว่าโปรตีนจำนวน 32 โปรตีนมีความแตกต่างในกลุ่มผู้ป่วยที่มีหรือไม่มีมะเร็งตำแหน่งที่สอง จากการตรวจสอบประสิทธิภาพโดยการแสดงออกของยีน พบว่าแบบแผนการแสดงออกของยีน *ITPR3, KMT2D, EMILIN1* และอายุของผู้ป่วยมีความไวร้อยละ 88 และความจำเพาะร้อยละ 75 ในการพยากรณ์การเกิดมะเร็งตำแหน่งที่สองในกลุ่มผู้ป่วยมะเร็งศีรษะและลำคอที่ไม่ได้เกิดจากการติดเชื้อ HPV นอกจากนี้ยังพบว่าระดับ *ITPR3* และ*DSG3* ที่สูงจะสัมพันธ์กับอัตราเร็วในการเกิดมะเร็ง

ตำแหน่งที่สอง กล่าวโดยสรุปการศึกษานี้ได้เสนอแบบแผนการแสดงออกของยีนที่มีโอกาสพัฒนาเป็นตัวชี้วัดทางชีวภาพสำหรับพยากรณ์การเกิดมะเร็งตำแหน่งที่สองในผู้ป่วยมะเร็งศีรษะและลำคอ เพื่อนำไปประยุกต์ในการรักษาและติดตามผู้ป่วยกลุ่มนี้ เพื่อผลการรักษาที่ดีขึ้นในอนาคต

คำหลัก: ตัวชี้วัดทางชีวภาพ / มะเร็งตำแหน่งที่สอง / มะเร็งศีรษะและลำคอ / แมสสเปกโตรเมทรี / นาโนสตริง

Abstract

Project Code: RSA5780065

Project Title: "Identification of candidate biomarkers for synchronous hypopharyngeal and esophageal cancer in Southern Thais using oncoproteomics analysis"

Investigator: Natini Jinawath, MD, PhD, DABMGG Faculty of Medicine Ramathibodi Hospital, Mahidol University

E-mail Address: jnatini@hotmail.com

Project Period: June 16, 2014 - March 29, 2019

Abstract: Patients with head and neck squamous cell carcinoma (HNSCC) are at increased risk of developing a second primary malignancy (SPM), which is associated with poor prognosis and early death. To help improve clinical outcome, we aimed to identify biomarkers for SPM risk prediction using the routinely obtained formalin-fixed paraffin-embedded (FFPE) tissues of the index HNSCC. LC-MS/MS was initially performed for candidate biomarker discovery in 16 pairs of primary HNSCC FFPE tissues and their matched normal mucosal epithelia from HNSCC patients with or without SPM. The 32 candidate proteins differentially expressed between HNSCCs with and without SPM were identified. Among these, 30 selected candidates and seven more from literature review were further studied using NanoString nCounter gene expression assay in an independent cohort of 49 HNSCC patients. Focusing on the p16-negative cases, we showed that a multivariate logistic regression model comprising the expression levels of *ITPR3*, *KMT2D*, *EMILIN1*, and the patient's age can accurately predict SPM occurrence with 88% sensitivity and 75% specificity. Furthermore, using Cox proportional hazards regression analysis and survival analysis, high expression levels of *ITPR3* and *DSG3* were found to be significantly associated with shorter time to SPM development

(log-rank test P = 0.017). In summary, we identified a set of genes whose expressions may serve as the prognostic biomarkers for SPM occurrence in HNSCCs. In combination with the histopathologic examination of index tumor, these biomarkers can be used to guide the optimum frequency of SPM surveillance, which may lead to early diagnosis and better survival outcome.

Keywords: BIOMARKERS / SECOND PRIMARY MALIGNANCY / HEAD AND NECK SQUAMOUS CELL CARCINOMA / LC-MS/MS / NANOSTRING

Research summary

1. Objectives

1) To discover molecular carcinogenesis mechanisms underlying primary head and neck squamous cell carcinomas (HNSCC) and their synchronous or metachronous second primary malignancies (SPM)
2) To identify candidate biomarkers specific to synchronous/ metachronous SPM, which can be used as SPM early detection biomarkers
3) To develop clinical test for SPM early detection and optimize SPM surveillance protocol for primary HNSCC patients based on their primary tumor tissues' protein/gene expression profiles (precision cancer medicine)

2. Introduction

Head and neck cancer is the seventh most common cancer worldwide, with over 600,000 new cases in 2012 (1). About 90% of all head and neck cancers are squamous cell carcinomas (HNSCC). HNSCC is associated with an elevated likelihood of developing second primary malignancy (SPM), which is defined as a second malignancy that presents either simultaneously or after the diagnosis of an index tumor. The overall incidence of SPMs in HNSCC patients has been reported to range from 5.6% to 35.9% (2, 3), with an annual incidence ranging from 3.2% to 4% (4). The common sites of SPM are the head and neck, lung, and esophagus (5). Several risk factors for SPM in HNSCC patients have been reported including the location of index tumor, patient age, cigarette smoking, alcohol consumption, and betel nut chewing (4).

Despite the progress in cancer molecular biology, the exact underlying molecular mechanism of SPM is still poorly understood. Field cancerization is one of the widely accepted concepts that explains SPM tumorigenesis (6). According to this concept, an area of the upper aerodigestive tract is considered as a field that is continuously exposed to a diversity of carcinogens resulting in multiple precancerous genetic changes, even though the mucosal epithelia still retain their normal histologic appearances. These precancerous fields may eventually become malignant leading to multiple SPM occurrences (7). Recently, Curtius et al. has reviewed the evolutionary process that results in field creation (8). In this context, a cancerized field is both enabled by and causes alterations in the tissue microenvironment. Measurements of the cancerized field evolution hold considerable promise as a new class of biomarker for cancer risk.

SPMs are known to have a major negative impact on HNSCC patients. The prognoses of HNSCC patients who develop SPM are worse than those with index malignancies alone (9), and it is also a leading long-term cause of mortality in HNSCC patients (10). Early diagnosis of SPMs often allows less invasive or curative treatments. However, current clinical screening and surveillance methods including panendoscopy and PET/CT scan still have clinical limitations (11, 12). Panendoscopy is a relatively invasive procedure and can result in serious complications such as esophageal perforation (11). While PET/CT scan demonstrates high sensitivity for detecting synchronous SPMs, its findings can be false-positive, and may miss small and/or superficial synchronous SPMs (13). Furthermore, the attempts to use chemoprevention for SPM risk reduction have so far been unsuccessful (14). For these reasons,

novel screening approaches such as molecular biomarkers for SPM risk prediction are very much needed.

Advances in omics technology have resulted in a better understanding of HNSCC carcinogenesis, which in turn lead to the development of novel molecular biomarkers for HNSCC that can be successfully translated into routine clinical practice. Human papillomavirus (HPV) has emerged as an important etiologic factor of HNSCCs, particularly in oropharyngeal cancers. HPV status is currently used as a prognostic biomarker for HNSCC; HPV-positive HNSCCs have a favorable prognosis and may benefit from less aggressive treatment regimens (15). The recently released 8th edition of the American Joint Committee on Cancer (AJCC) staging manual, Head and Neck section, has acknowledged the importance of using p16 immunostaining as a surrogate marker of HPV status, and issued a distinct staging classification for HPV-associated oropharyngeal cancer, separating it from cancer of the oropharynx related to other causes (16). Despite these advances, to date, very few studies have focused on SPM, particularly on a biomarker discovery aspect (17, 18), and so far none has been successfully translated into clinics.

In this study, we hypothesized that the cancerized fields in HNSCC patients with or without SPM are diverse, leading to distinct protein or gene expression patterns that can be exploited as biomarkers. We then aimed to identify molecular biomarkers for SPM risk prediction using the routinely collected formalin-fixed paraffin-embedded (FFPE) tissue biopsies of the index HNSCCs. Using liquid chromatography-tandem mass spectrometry (LC-MS/MS), we identified a set of candidate biomarkers differentially expressed in the primary tumors of HNSCC patients with or without SPM after a minimum follow-up period of three years. Next, we utilized a NanoString nCounter gene expression assay

to study the ability of selected biomarkers to predict SPM development and the time to SPM occurrence in another independent cohort of HNSCC patients. Our study permitted the development of the first tissue-based gene-expression biomarker panel for SPM that can easily be implemented in the routine pathology practice.

3. Materials and Methods

Patient selection and sample description

All patients were clinically diagnosed with head and neck cancer by the otolaryngologists or oncologists and had a histopathological diagnosis of squamous cell carcinoma. The patients with only a single primary tumor after a minimum follow-up period of approximately three years were defined as "P_{index} group". The HNSCC patients who developed SPM during the same follow-up period were defined as "P_{SPM} group". The clinical diagnosis of SPM was made based on the applied recommendation from Warren and Gates' criteria (19) as follows: i) both index and secondary tumors had histologic confirmation of squamous cell carcinoma; ii) the two malignancies were anatomically separated by more than 2 cm of normal mucosa; iii) the possibility of the SPM being a metastasis from the index tumor was excluded. SPM was further classified as synchronous SPM when the diagnosis of both index tumor and SPM was made simultaneously or within 6 months after the index tumor, or as metachronous SPM when the diagnosis of SPM was made longer than 6 months after the index tumor (20).

Cohort 1 consisted of 16 HNSCC patients from Songklanagarind Hospital (Prince of Songkla University, Songkhla, Thailand). A pair of FFPE tissue blocks containing the

index tumor and normal mucosal epithelia was collected for each patient, resulting in a total of 32 samples for proteomic analysis. In addition, all SPM samples were previously confirmed to be SPMs, not metastatic tumors, by comparing the loss of heterozygosity (LOH) patterns identified by single nucleotide polymorphism array between each index HNSCC and its matched SPM (21). Cohort 2 was an independent multi-center cohort comprising 49 HNSCC patients from Ramathibodi Hospital (Mahidol University, Bangkok, Thailand), Siriraj Hospital (Mahidol University, Bangkok, Thailand), and Songklanagarind Hospital. Only FFPE tissue blocks containing the index tumor were collected for each patient. This study was approved by the Institutional Review Board (IRB) of all participating hospitals.

Sample preparation

FFPE tissue sections were prepared. Previously archived hematoxylin and eosin stained tissue slides were evaluated for tumor content and density by pathologists. The areas with at least 70% tumor cells and normal mucosal epithelial cells were marked for manual macrodissection using a needle tip or scalpel. For LC-MS/MS, dissected FFPE tissue was prepared for protein digestion using the filter-assisted sample preparation (FASP) method (22) and physically disrupted by sonication. For NanoString gene expression assay, total RNA was isolated using High Pure FFPET RNA Isolation Kit (Roche) according to the manufacturer's instructions.

Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS)

LC–MS/MS analysis was performed as a service by Bioproximity, LLC (Chantilly, VA, USA). In brief, proteins were digested with trypsin, and digested peptides were desalted using C18 stop-and-go extraction (STAGE) tips (23). Peptides were then fractionated by strong anion exchange STAGE tip chromatography (24). LC was performed on an Easy nanoLC II HPLC system (Thermo Fisher Scientific). The LC was interfaced to a dual pressure linear ion trap mass spectrometer (LTQ Velos, Thermo Fisher Scientific) via nano-electrospray ionization. MS data were processed and the Mascot Generic Format (MGF) files were searched using X!!Tandem, k-score scoring algorithms and OMSSA (open mass spectrometry search algorithm). The common Repository of Adventitious Proteins (cRAP) and the Ensembl release 69 were used as protein databases for the searches. Proteins were required to have 2 or more unique peptides across the analyzed samples with E-value scores of 0.01 or less and protein E-value scores of 0.0001 or less (25). The spectral count of each identified protein was initially scaled by the total and/or maximum peptide counts of the same patient. All scaled spectral counts were added with 0.01 to avoid division by zero, and each scaled spectral count of tumor tissue was then divided by the scaled spectral count of the corresponding normal tissue from the same patient to normalize interindividual variations.

Immunohistochemistry (IHC)

A Leica BOND-MAX™ automated staining system (Leica Biosystems) was utilized according to the manufacturer's protocol for anti-CALML3 (1:1000, PA5-30232, Thermo Fisher Scientific), and anti-CKMT2 (1:200, PA5-28591, Thermo Fisher Scientific). A

Ventana BenchMark XT automated staining system (Ventana Medical Systems) was used according to the manufacturer's instructions for anti-DSG3 (1:50, MAB1720, R&D Systems), anti-PLOD1 (1:50, NBP2-31885, Novus Biologicals), and anti-p16 (CINtec® p16 Histology, 705-4713, Ventana Medical Systems). Blinded scoring was independently performed by two pathologists using the same scoring criteria (Supplementary Table S1).

Digital gene expression quantification using NanoString nCounter™

A NanoString nCounter™ Gene Expression Assay (Nanostring Technologies) with the custom-designed CodeSets containing 40 genes was performed following the standard protocol. Data processing was conducted using nSolver Analysis Software v3.0 (NanoString Technologies). Background correction was made by subtracting the "mean+2SD" value of the negative controls from the raw counts. Adjusted raw counts were then normalized with the geometric mean of the positive controls in each sample. The resulting data was normalized again using the geometric mean of three housekeeping genes. Normalized data was standardized using mean centering then divided by the standard deviation to obtain the Z scores for subsequent statistical analyses.

Bioinformatic and statistical analysis

Hierarchical cluster analysis (HCA) using average linkage method was performed using Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm), and visualized by Java TreeView version 1.1.6r4 (http://jtreeview.sourceforge.net). Functional annotation of candidate proteins was performed using DAVID 6.8

(https://david.ncifcrf.gov/). Statistical analyses were performed using PASW Statistics version 18.0 for Windows (SPSS Inc.), GraphPad Prism version 6.0 for Windows (GraphPad Software), R-statistical software version 3.3.1 (R Foundation, http://www.r-project.org), and ROCR package (26).

4. Results

Subject characteristics

Two independent cohorts of HNSCC patients were included in this study. Among the 16 patients in cohort 1, 5 (31.25%) and 11 (68.75%) patients belonged to the P_{index} and P_{SPM} groups, respectively. The median duration of follow-up to ensure no occurrence of SPM was 30.8 (29.7-86.4) months in the P_{index} group. All SPMs were esophageal squamous cell carcinomas. No statistically significant difference was detected between the clinical characteristics of patients in the P_{index} and P_{SPM} groups.

In cohort 2, of the 49 HNSCC patients examined, 21 (42.86%) and 28 (57.14%) patients belonged to the P_{index} and P_{SPM} groups, respectively. The median duration of follow-up was 54.1 (35.3-94.5) months in the P_{index} group. All SPMs were squamous cell carcinomas located in the esophagus, head and neck, or lung. The patients' age was the only clinical characteristic with statistically significant difference between the P_{index} and P_{SPM} groups (P = 0.027). The detailed clinicopathological and demographical characteristics of patients in both cohorts are summarized in Table 1.

The HPV status of HNSCC tumors was also evaluated using p16 IHC. The difference in HPV status based on p16-positivity between the P_{index} and P_{SPM} groups from the two cohorts was not statistically significant (Fisher's exact test P = 1.000). Of note,

only 3 cases out of the 49 HNSCCs in cohort 2 were p16-positive. The p16 IHC results are shown in Supplementary Table S2.

LC-MS/MS analysis and validation by IHC

The detailed experimental outline of this study is summarized in Figure 1. We first performed LC-MS/MS in 16 pairs of tumor-normal FFPE tissues from P_{index} and P_{SPM} groups (cohort 1) as a means to discover candidate biomarkers for SPM. A total of 2,816 proteins were initially identified. Proteins detected in less than 10% of all samples were then filtered out, leaving a total of 2,101 proteins for unsupervised hierarchical cluster analysis (HCA). The resulted dendrogram showed that these protein profiles could correctly classify the tumor and normal tissue samples (Figure 2A). We further investigated the protein profiles within the tumor or normal tissue subgroups separately whether they could correctly define each sample as P_{index} or P_{SPM} by performing HCA of differentially expressed proteins (t-test $P < 0.05$). The results interestingly demonstrated that P_{index} and P_{SPM} samples could be accurately clustered in both tumor and normal tissue subgroups (Supplementary Figure S1).

In addition, IHC analysis of representative proteins was performed to validate the accuracy of LC-MS/MS results. Four proteins that were at least 2-fold up- (PLOD1) or down-regulated (CKMT2, DSG3, CALML3) in tumors as compared to normal tissues, and expressed in more than half of the samples, were selected. The results demonstrated that IHC analysis showed the same up- or down-regulated trend between normal and tumor tissues as in the LC-MS/MS data in 3 of 4 proteins (DSG3, CALML3, PLOD1) (Figure 2B-C).

Identification of candidate proteins biomarkers for SPM

To screen for the differentially expressed proteins between HNSCC patients with and without SPM, the protein expression ratio of tumor to matched normal tissue of each patient was analyzed. The inclusion criteria for candidate proteins included: i) proteins expressed in more than 50% of the samples, ii) proteins showing at least 3-fold difference in expression level between the P_{index} and P_{SPM} groups, and iii) proteins exhibiting statistically significant differential expression between the two groups (t-test $P < 0.05$). Based on these criteria, 32 candidate proteins were identified (Supplementary Table S3). HCA using expression ratio of these proteins demonstrated that all cohort 1 samples could be correctly divided into P_{index} and P_{SPM} groups (Figure 3).

We further carried out gene-annotation enrichment analysis using DAVID 6.8 to identify enriched biological functions of the candidate proteins. Among the up-regulated proteins in P_{SPM} samples, the statistically significant gene ontology (GO) term ($P < 0.001$) with highest enrichment score (3.88) was “intermediate filament” (KRT37, KRT12, KRT86, KRT24). The other significant GO terms ($P < 0.05$) were “cytoplasm” (UBE2N, PA2G4, S100A7, ARPC2, SERPINB5, TRIM29, CRABP2, EIF5A, RPS20, AHNAK, KRT24), and “poly(A) RNA binding” (UBE2N, PA2G4, EIF5A, RPS20, AHNAK). In contrast, the significant GO terms with highest enrichment score (1.31) among the down-regulated proteins were “extracellular exosome” and “extracellular matrix” (PTGES3, PLOD1, FBLN2, UGDH, CSRP1, ECM1, EMILIN1) ($P < 0.01$) (Figure3 and Supplementary Table S4A-B).

Development of a predictive model for SPM occurrence by NanoString nCounter gene expression analysis

Considering the difficulties in translating tissue-based quantitative protein biomarkers into clinical practice, we set out to alternatively utilize a NanoString nCounter gene expression

assay to optimize the biomarker selection in an independent cohort (cohort 2). The 40-gene custom panel consisted of the 30 candidate proteins (of the 32 candidates, two did not have specific CodeSets available), 3 housekeeping genes, and 7 frequently mutated genes in HNSCCs with synchronous nodal metastasis or metachronous recurrence from literature review (27, 28) (Supplementary Table S5). HPV-positive and HPV-negative HNSCCs are known to exhibit different clinical and molecular characteristics (16). Since the number of p16-positive cases in our cohort was limited (3 out of 49 cases), we focused on the 46 p16-negative HNSCC patients for further analyses.

To identify biomarkers that could accurately determine the development of SPM, univariate logistic regression analysis of the standardized NanoString gene expression levels and selected clinical variables of the 46 p16-negative HNSCC patients was carried out. The expression levels of three genes (*ITPR3*, *FAT1, KMT2D)* and the patient's age at diagnosis were statistically significantly associated with the development of SPM ($P < 0.05$) (Figure 4A). The strongest risk factor for the development of SPM was high *ITPR3* level. An increase in one standard deviation (SD) of *ITPR3* level was associated with 3.27 times higher risk of SPM development (OR = 3.27; 95%CI [1.36-11.04]; $P = 0.025$). Other significant risk factors included high *FAT1* (OR = 2.25; 95%CI [1.13-5.49]; $P = 0.038$) and *KMT2D* levels (OR = 2.20; 95%CI [1.10-5.18]; $P = 0.043$). To minimize the effect of small sample size, we performed 5,000 permutation tests to correct for non-asymptotic properties of *P*-values by reshuffling the observed data. The resulting *P*-values were reported as empirical *P*-values. The analysis showed that the expression levels of two more genes (*EMILIN1* and *ECM1*) were also statistically significantly associated with SPM development (empirical $P < 0.05$). The results of univariate logistic regression of all genes are shown in

Supplementary Table S6. Comparing between P_{index} and P_{SPM} groups, the expression levels of *ITPR3*, *FAT1, KMT2D*, and *ECM1* were higher in P_{SPM} samples, whereas the level of *EMILIN1* was lower (Figure 4B).

We next aimed to identify the best combination of biomarkers for SPM development prediction by analyzing the six statistically significant variables ($P < 0.05$ or empirical $P < 0.05$) from univariate logistic regression (*ITPR3*, *FAT1, KMT2D*, *EMILIN1*, *ECM1,* and patient's age*)* using multivariate logistic regression analysis. The most parsimonious final model based on Akaike's Information Criterion (AIC), Bayesian information criterion (BIC), and deviance test was the combination of *ITPR3*, *KMT2D*, *EMILIN1*, and the patient's age. The AIC, and BIC values of this model were 50.82 and 59.96, respectively. The detailed results of all multivariate logistic regression models and the final selection are shown in Supplementary Table S7 and S8.

To evaluate the performance of the selected multivariate model in predicting SPM occurrence, a receiver operating characteristic (ROC) analysis was performed and an area under the ROC curve (AUC) value was calculated as 0.86 (95% CI [0.75-0.97]). Moreover, this model has a sensitivity of 88.46%, a specificity of 75.00% and an accuracy of 82.61% for predicting the occurrence of SPM. We further performed a leave-one-out cross-validation over the same 46 samples (cohort 2) to assess the robustness of our model. The resulting AUC, sensitivity, specificity and accuracy values were 0.80 (95% CI [0.66-0.94]), 76.92%, 70.00%, 73.91%, respectively (Figure 4C). These new values were not significantly different from the original results, confirming the validity of our predictive model.

Identification of biomarkers for predicting time to SPM development

Next, we focused on HNSCC patients who eventually developed SPM (P_{SPM} group) in order to identify the biomarkers associated with time to subsequent SPM occurrence. Univariate Cox regression analysis using gene expression levels of the 26 p16-negative HNSCCs from cohort 2 was performed. The levels of *ITPR3, DBI, AHNAK, IGHV3-49, CALML3, ARPC2, DSG3,* and *KRT37* were significantly associated with a shorter time to SPM development ($P < 0.05$) (Figure 5A). The strongest association was with *ITPR3* level. An increase in one SD of *ITPR3* level was associated with 2.68 times higher risk of SPM development at any given time after the index tumor diagnosis (HR = 2.68; 95%CI [1.53-4.72]; $P = 0.001$). The complete results of univariate Cox regression analysis including the empirical *P*-values generated by 5,000 permutation tests are shown in Supplementary Table S9. Using multivariate Cox proportional hazards regression analysis, the best model was the combination of *ITPR3* and *DSG3* (Supplementary Table S10). Moreover, by dividing the standardized *ITPR3* and *DSG3* level by the sample mean of the 26 P_{SPM} patients, survival analysis showed that the patients with p16-negative HNSCC whose index tumors exhibited high *ITPR3* and *DSG3* expression levels had the shortest time interval between the diagnosis of an index HNSCC to subsequent SPM development (log-rank test $P = 0.017$) (median time difference between the high and low risk groups based on four combinations of *ITPR3* & *DSG3* levels = 394 days) (Figure 5B and Supplementary Table S11).

Comparison of the expression pattern of candidate biomarkers between protein and mRNA levels

It is known that protein and mRNA expression levels often do not directly correlate (29). To investigate whether our candidate genes originally discovered by proteomics study

shared the same differential expression patterns as their protein equivalents, we compared the standardized gene expressions of the top 10 statistically significant genes identified by logistic regression and Cox regression analyses to their standardized protein levels obtained from LC-MS/MS. Seven of the 10 biomarkers (*AHNAK, ARPC2, CALML3, DBI, DSG3, EMILIN1, KRT12*) demonstrated a similar up- or down-regulation trend between mRNA and protein levels in the tumor samples across the two patient cohorts (Figure 6).

5. Discussion

In this study, we identified a set of tissue-based biomarkers for predicting SPM occurrence and time to SPM development in HNSCC patients using a combination of high-throughput shotgun proteomics and targeted gene expression analysis. These SPM risk prediction biomarkers can help guide clinical management of HNSCC patients, particularly in the frequency of SPM surveillance after the diagnosis of an index tumor, and the choice of treatments. The expected long-term benefit is an improvement in overall survival of HNSCC patients, especially those who are eligible for curative or less invasive therapy. Moreover, one of the clinical strengths of this study is the use of index HNSCC FFPE tissue as the preferred material for biomarker discovery. This specimen is routinely acquired for histopathological diagnosis, which makes it easier to incorporate our biomarker panel into the current clinical practice. We believe that our study is the first to identify prognostic biomarkers that can accurately predict SPM occurrence, thus opening the door to the possibility of clinical application of tissue-based biomarkers for SPM in HNSCC patients.

Mass spectrometry (MS) has been extensively used to discover novel protein biomarkers (33). Proteins are the functional molecules in the cell, and thus are the key players that represent actual cellular physiology. For tissue-based protein biomarkers, several laboratory methods such as IHC and targeted MS can be used to develop the validated markers into quantitative clinical assays but not without limitations. To date, IHC is routinely performed in clinics to determine both the qualitative and semi-quantitative aspects of a protein biomarker, however it relies heavily on the quality of antibody and still lacks interpretation standardization, resulting in poor reproducibility (34). While targeted mass spectrometry can do marker multiplexing and does not rely on antibody, it is technically complex and carry a high developing cost (35). The development of tissue-based high-throughput gene expression assays has greatly improved the accuracy and reproducibility of quantitative measurement in tissue biomarker studies. Currently, several such assays have been integrated into clinical practice. For example, the Prosigna breast cancer prognostic gene signature assay based on NanoString nCounter analysis system was approved by U.S. Food and Drug Administration in 2013 as a prognostic assay for distant recurrence risk in breast cancer patients (36). In this study, we sequentially utilized LC-MS/MS and NanoString nCounter system in two independent cohorts of HNSCC patients in order to identify the best set of SPM prediction biomarkers. In light of its clinical translatability, technical reproducibility, and compatibility with small biopsied FFPE tissues (37), the NanoString digital gene expression platform was chosen for biomarker panel development.

In cohort 1, the protein profiles were markedly different between HNSCCs and their matched normal squamous mucosa, which is in line with previous studies (17, 38). Interestingly, we

also observed distinct protein expression profiles between the P_{index} and P_{SPM} subgroups when analyzing the tumor and normal tissue samples separately. These results suggest that tissue microenvironments of both tumor and surrounding normal mucosal cells of the P_{index} and P_{SPM} patients may exhibit varying levels of genetic diversity, resulting in distinct protein and/or gene expression signatures. These findings strongly support our hypothesis that the aberrant field effect in the biopsied index tumors can be used to predict whether the patients are at risk of subsequent SPM development. It is proposed that genetic diversity can be utilized as a marker for field evolvability (8). In a prospective pilot study by Roesch-Ely et al., by analyzing the protein profiles of mucosal biopsies from the oropharynx, hypopharynx, and three different regions of esophagus in HNSCC patients and controls, tumor relapse was correctly predicted (17). Taken together, these findings firmly established an altered field as a promising cancer risk prediction marker.

Of note, among the candidate proteins differentially expressed between P_{index} and P_{SPM} samples, up-regulated proteins in P_{SPM} group were significantly enriched with proteins associated with intermediate filament, whereas those involved with extracellular matrix (ECM) were down-regulated. Overexpression of keratins, the intermediate filament-forming proteins of epithelial cells, is associated with enhanced tumor cell migration and invasion through interactions with extracellular environment (39). ECM proteins are associated with both tumor suppression and tumor promotion. Interestingly, the majority of down-regulated ECM proteins in P_{SPM} samples including ECM1, EMILIN1, and FBLN2 have demonstrated tumor suppressive roles in various cancers (40-42). Based on our findings, aberrant cytoskeletal activity and altered cell motility as a result of reduced

expression of ECM-associated tumor suppressors are among the major molecular mechanisms underlying SPM development.

Recently, a large HNSCC genomics study from the Cancer Genome Atlas (TCGA) consisting of mostly Caucasian patients revealed that 64% and 6% of oropharyngeal and non-oropharyngeal tumors, respectively, are HPV-positive (28). In contrast, studies from Thailand reported that HPV status was positive in only 26.09% of oropharyngeal cancers and in none of the 80 non-oropharyngeal cancers tested (30). This fact is also reflected in our largely p16-negative cohorts. In addition, multiple studies have reported that patients with HPV-negative oropharyngeal SCC have a higher risk of SPM development than HPV-positive oropharyngeal SCC patients (31, 32). Since most of our patients were p16-negative and HPV-negative HNSCC patients are more likely to have SPM, we decided to focus on the p16-negative cases in cohort 2 for our gene expression analyses.

For cohort 2, in addition to the candidate genes from our proteomics study, seven genes from recent genomics literatures (27, 28) were included in the gene expression panel. *TP53* and *NOTCH1* are among the most commonly mutated genes in HNSCC (28). A recent whole exome sequencing study of HNSCCs has reported that *C17orf104* and *ITPR3* are specifically mutated in synchronous nodal metastases but not in the primary cancers, while *DDR2* is exclusively mutated in metachronous recurrent tumors (27). Additionally, *FAT1* and *KMT2D* are found to be mutated in both primary HNSCCs and their nodal metastases (27). Hence, it is of interest to investigate whether their expressions are associated with SPM risk. In this study, the combination of three genes *(ITPR3*, *KMT2D*, *EMILIN1)* and patient's age allowed for the most accurate discrimination between P_{index} and P_{SPM} groups with a sensitivity of 88.46% and a specificity of 75.00%. Of note, while these numbers will require

validation in the future studies using larger sample size, our panel showed a comparable level of performance to that of the currently used clinical gene expression tests in breast cancer including Oncotype Dx (sensitivity 71-85%, specificity 55-66% for high and intermediate risk groups) and MammaPrint (sensitivity 83-92%, specificity 41-59%) (43).

In addition to its implications in HNSCC (27), increased expression of ITPR3, a major intracellular Ca^{2+} release channel, is associated with enhanced tumor proliferation and invasion in breast and colorectal cancers (44, 45). *KMT2D*, which encodes a histone H3 lysine 4 methyltransferase, is one of the most commonly mutated genes in cancers (46). High KMT2D expression has recently been reported to promote tumor progression by inducing epithelial-mesenchymal transition (EMT), and is a predictor of poor prognosis in esophageal squamous cell carcinoma (Ref). In contrast, *EMILIN1*, which encodes an ECM glycoprotein associated with the development of elastic tissues, plays a suppressive role in tumor growth, tumor lymphatic vessel formation, and metastatic spread to lymph nodes *(41).* A recent report has suggested that α9β1 integrin, of which EMILIN1 is a ligand, is the major integrin involved in regulating HNSCC cell migration on ECM (47). In our study, increased *ITPR3* and *KMT2D* levels, as well as reduced *EMILIN1* are significantly associated with SPM in HNSCC. Their exact roles in SPM tumorigenesis remain to be explored.

SPM can be classified into either a synchronous or metachronous tumor. Since survival outcomes between synchronous and metachronous SPM are different (5, 9), it is clinically useful to be able to predict the time to SPM development. The combination of two genes, *ITPR3* and *DSG3*, was statistically significantly associated with the time to SPM occurrence in our study. *DSG3*, a component of cell-cell

junctions, is overexpressed in HNSCCs, and its inhibition significantly suppresses tumor growth (48). DSG3 has also been proposed as a predictive biomarker for cervical lymph node micrometastasis in oral cancer (Ref). In line with their roles in promoting tumor proliferation and progression, high expression of both *ITPR3* and *DSG3* contributed to the shortest time to SPM development in patients with p16-negative HNSCC.

It is generally known that transcript levels by themselves are by no means an accurate predicting factor for protein levels in many scenarios (29). In this study, the majority of top candidate biomarkers remarkably showed a similar pattern of differential protein and mRNA expressions in the tumor samples across the P_{index} and P_{SPM} subgroups of both cohorts. This finding partly explains a smooth transition from protein to transcript level of this particular set of markers in regard to SPM risk prediction, and may also render future biomarker assay development the flexibility and feasibility to use either gene expression or targeted proteomic platforms.

According to the current follow-up recommendation by the National Comprehensive Cancer Network (NCCN) guidelines for head and neck cancers (version 2.2018) (49), history and physical exams including a complete head and neck exam; and mirror and fiberoptic examination should be performed at least every three months during the first year, and at least every six months in the second year. Based on our results, all patients with high *ITPR3* and *DSG3* levels developed SPM within approximately 13 months (400 days) after the diagnosis of an index tumor. Therefore, it may be beneficial to provide closer monitoring of SPM for this group of patients during their first years. Moreover, even though some synchronous SPMs develop around the same time as the index tumor, they often cannot be clinically detected until their sizes are big enough. The use of our

biomarkers should help increase early diagnosis of both subclinical synchronous SPMs and subsequent metachronous SPMs through a more rigorous surveillance protocol.

Limitations of this study include the relatively small sample size and limited follow-up time to SPM occurrence. Others include the use of a cohort of convenience where adequate FFPE tissues was available, and the use of one tissue sample per tumor, which may not represent intratumor heterogeneity.

In summary, our biomarkers demonstrate great potential as a companion prognostic test for SPM risk prediction in routine clinical practice. The expected long-term benefit of SPM early detection is an improvement in overall survival of HNSCC patients, especially those who are eligible for curative or less invasive therapy. The results are compelling and warrant future validation studies in larger HNSCC cohorts. The protein signatures also hint at tumor-ECM interactions as a major player in SPM tumorigenesis. Further functional studies of these biomarkers may better clarify the clinical utility and their roles in SPM development.

6. References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-E86.

2. González-García R, Naval-Gías L, Román-Romero L, et al. Local recurrences and second primary tumors from squamous cell carcinoma of the oral cavity: a retrospective analytic study of 500 patients. Head Neck. 2009;31(9):1168-80.

3. Yamashita T, Araki K, Tomifuji M, et al. Clinical features and treatment outcomes of Japanese head and neck cancer patients with a second primary cancer. Asia Pac J Clin Oncol. 2017;13(3):172-8.
4. Lee DH, Roh J-L, Baek S, et al. Second cancer incidence, risk factor, and specific mortality in head and neck squamous cell carcinoma. Otolaryngol Head Neck Surg. 2013;149(4):579-86.
5. Vaamonde P, Martin C, Labella T. Second primary malignancies in patients with cancer of the head and neck. Otolaryngol Head Neck Surg. 2003;129(1):65-70.
6. Slaughter DP, Southwick HW, Smejkal W. “Field cancerization” in oral stratified squamous epithelium. Clinical implications of multicentric origin. Cancer. 1953;6(5):963-8.
7. Braakhuis BJ, Tabor MP, René Leemans C, et al. Second primary tumors and field cancerization in oral and oropharyngeal cancer: molecular techniques provide new insights and definitions. Head Neck. 2002;24(2):198-206.
8. Curtius K, Wright NA, Graham TA. An evolutionary perspective on field cancerization. Nat Rev Cancer. 2018;18(1):19-32.
9. Di Martino E, Sellhaus B, Hausmann R, et al. Survival in second primary malignancies of patients with head and neck cancer. J Laryngol Otol. 2002;116(10):831-8.
10. Baxi SS, Pinheiro LC, Patil SM, et al. Causes of death in long-term survivors of head and neck cancer. Cancer. 2014;120(10):1507-13.
11. Clayburgh DR, Brickman D. Is esophagoscopy necessary during panendoscopy? Laryngoscope. 2017;127(1):2-3.
12. Haerle SK, Strobel K, Hany TF, et al. 18F-FDG-PET/CT versus panendoscopy for the detection of synchronous second primary tumors in patients with head and neck squamous cell carcinoma. Head Neck. 2010;32(3):319-25.

13. Yabuki K, Kubota A, Horiuchi C, et al. Limitations of PET and PET/CT in detecting upper gastrointestinal synchronous cancer in patients with head and neck carcinoma. Eur Arch Otorhinolaryngol. 2013;270(2):727-33.
14. Khuri FR, Lee JJ, Lippman SM, et al. Randomized phase III trial of low-dose isotretinoin for prevention of second primary tumors in stage I and II head and neck cancer patients. J Natl Cancer Inst. 2006;98(7):441-50.
15. Kang H, Kiess A, Chung CH. Emerging biomarkers in head and neck cancer in the era of genomics. Nat Rev Clin Oncol. 2015;12(1):11-26.
16. Lydiatt WM, Patel SG, O'Sullivan B, et al. Head and neck cancers—major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(2):122-37.
17. Roesch-Ely M, Leipold A, Nees M, et al. Proteomic analysis of field cancerization in pharynx and oesophagus: a prospective pilot study. J Pathol. 2010;221(4):462-70.
18. Hildebrandt MA, Lippman SM, Etzel CJ, et al. Genetic variants in the PI3K/PTEN/AKT/MTOR pathway predict head and neck cancer patient second primary tumor/recurrence risk and response to retinoid chemoprevention. Clin Cancer Res. 2012;18(13):3705-13.
19. Warren S, Gates O. Multiple primary malignant tumors: A survey of the literature and a statistical study. Am J Cancer. 1932;16:1358-414.
20. Hong WK, Lippman SM, Itri LM, et al. Prevention of second primary tumors with isotretinoin in squamous-cell carcinoma of the head and neck. N Engl J Med. 1990;323(12):795-801.
21. Sunpaweravong S, Sunpaweravong P, Dechaphunkul T, et al. Clinico-molecular study of synchronous head and neck

squamous cell carcinoma (HNSCC) and esophageal squamous cell carcinoma (ESCC). J Clin Oncol. 2015;33(15_suppl):e17080
22. Wisniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods. 2009;6(5):359-62.
23. Rappsilber J, Ishihama Y, Mann M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem. 2003;75(3):663-70.
24. Wiśniewski JR, Zougman A, Mann M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res. 2009;8(12):5674-8.
25. Skeie JM, Mahajan VB. Proteomic interactions in the mouse vitreous-retina complex. PLoS One. 2013;8(11):e82140.
26. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940-1.
27. Hedberg ML, Goh G, Chiosea SI, et al. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. J Clin Invest. 2016;126(1):169-80.
28. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-82.
29. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165(3):535-50.
30. Pongsapich W, Jotikaprasardhna P, Lianbanchong C, et al. Human Papillomavirus Infection in Oral Cavity and Oropharyngeal Cancers: Are They the Same Story? J Med Assoc Thai. 2016;99(6):684-9.
31. Martel M, Alemany L, Taberna M, et al. The role of HPV on the risk of second primary neoplasia in patients with oropharyngeal carcinoma. Oral Oncol. 2017;64:37-43.

32. Gan SJ, Dahlstrom KR, Peck BW, et al. Incidence and pattern of second primary malignancies in patients with index oropharyngeal cancers versus index nonoropharyngeal head and neck cancers. Cancer. 2013;119(14):2593-601.
33. Crutchfield CA, Thomas SN, Sokoll LJ, et al. Advances in mass spectrometry-based clinical biomarker discovery. Clin Proteomics. 2016;13(1):1.
34. Lin F, Chen Z. Standardization of diagnostic immunohistochemistry: literature review and geisinger experience. Arch Pathol Lab Med. 2014;138(12):1564-77.
35. Meng Z, Veenstra TD. Targeted mass spectrometry approaches for protein biomarker verification. J Proteomics. 2011;74(12):2650-9.
36. Győrffy B, Hatzis C, Sanft T, et al. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res. 2015;17(1):11.
37. Veldman-Jones MH, Brant R, Rooney C, et al. Evaluating robustness and sensitivity of the NanoString technologies nCounter platform to enable multiplexed gene expression analysis of clinical samples. Cancer Res. 2015;75(13):2587-93.
38. Roesch-Ely M, Nees M, Karsai S, et al. Proteomic analysis reveals successive aberrations in protein expression from healthy mucosa to invasive head and neck cancer. Oncogene. 2007;26(1):54-64.
39. Karantza V. Keratins in health and cancer: more than mere epithelial cell markers. Oncogene. 2011;30(2):127-38.
40. Gao F, Xia Y, Wang J, et al. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. Genome Biol. 2014;15(12):533.

41. Danussi C, Petrucco A, Wassermann B, et al. An EMILIN1-negative microenvironment promotes tumor cell proliferation and lymph node invasion. Cancer Prev Res. 2012;5(9):1131-43.
42. Law E, Cheung A, Kashuba V, et al. Anti-angiogenic and tumor-suppressive roles of candidate tumor-suppressor gene, Fibulin-2, in nasopharyngeal carcinoma. Oncogene. 2012;31(6):728-38.
43. Berg AO, Armstrong K, Botkin J, et al. Recommendations from the EGAPP Working Group: can tumor gene expression profiling improve outcomes in patients with breast cancer? Genet Med. 2009;11(1):66-73.
44. Mound A, Rodat-Despoix L, Bougarn S, et al. Molecular interaction and functional coupling between type 3 inositol 1, 4, 5-trisphosphate receptor and BK Ca channel stimulate breast cancer cell proliferation. Eur J Cancer. 2013;49(17):3738-51.
45. Shibao K, Fiedler MJ, Nagata J, et al. The type III inositol 1, 4, 5-trisphosphate receptor is associated with aggressiveness of colorectal carcinoma. Cell Calcium. 2010;48(6):315-23.
46. Rao RC, Dou Y. Hijacked in cancer: the MLL/KMT2 family of methyltransferases. Nat Rev Cancer. 2015;15(6):334-46.
47. Gopal S, Veracini L, Grall D, et al. Fibronectin-guided migration of carcinoma collectives. Nat Commun. 2017;8:14105.
48. Chen Y, Chang J, Lee L, et al. DSG3 is overexpressed in head neck cancer and is a potential molecular target for inhibition of oncogenesis. Oncogene. 2007;26(3):467-76.
49. National Comprehensive Cancer Network. Head and Neck Cancers (Version 2.2018 - June 20, 2018) 2018 [Available from: https://www.nccn.org/professionals/physician_gls/PDF/head-and-neck.pdf.

7. Figure and Table Legends

Figure 1. Overview and experimental design of the study. The biomarker identification steps were sequentially conducted using liquid chromatography-tandem mass spectrometry (LC-MS/MS) in cohort 1 (Left), and NanoString gene expression assay in cohort 2 (Right). Multiple statistical analyses were performed to identify prognostic biomarkers for second primary malignancy (SPM) occurrence and time to SPM development. P_{index}: index HNSCC without SPM, P_{SPM}: index HNSCC with SPM, FFPE: formalin-fixed paraffin-embedded, IHC: immunohistochemistry, ROC: receiver operating characteristic.

Figure 2. Protein expression profiling by LC-MS/MS and validation by IHC. (A) Dendrogram of unsupervised hierarchical cluster analysis (HCA) of the 2,101 protein profiles across the tumor samples and their matched normal mucosal tissues in cohort 1. **(B)** Representative IHC images (x10 magnification) of the four selected proteins (CKMT2, DSG3, CALML3, and PLOD1). N: normal mucosal epithelial cells, T: tumor cells. **(C)** Box plot showing the comparison between normalized expression ratios (T/N) of the four selected proteins analyzed by IHC and LC-MS/MS.

Figure 3. Hierarchical cluster analysis (HCA) of the candidate proteins. A heatmap showing the 32 candidate proteins differentially expressed between P_{index} and P_{SPM} subgroups with statistical significance (t-test $P < 0.05$) across cohort 1 (Left), and the summary of gene-annotation enrichment analysis of up- and down-regulated proteins of each subgroup using DAVID bioinformatics resources 6.8 (Right). The color scale is shown at the upper left corner. GO: Gene Ontology, P_{index}:

index HNSCC without SPM, P_{SPM}: index HNSCC with SPM, *Statistically significant ($P < 0.05$), **Statistically significant ($P < 0.01$).

Figure 4. Identification of biomarkers for SPM risk prediction by NanoString gene expression analysis. (A) The statistically significant variables identified by univariate logistic regression across P_{index} and P_{SPM} subgroups of the p16-negative tumors in cohort 2 (n = 46). The empirical *P*-values were generated by 5,000 permutation tests. **(B)** Box plot depicting expression levels of the five candidate genes identified by univariate logistic regression across P_{index} and P_{SPM} subgroups of the p16-negative tumors in cohort 2. All pairwise comparisons show statistically significant differential expressions. **(C)** The predictive performance of the selected multivariate logistic regression model (*ITPR3-KMT2D-EMILIN1*-patient's age) in classifying SPM risk. The robustness of the model was assessed by performing a leave-one-out cross-validation for each sample in the same cohort. The unit of patient's age is year. *Statistically significant ($P < 0.05$), **Statistically significant ($P < 0.01$), OR: odds ratio, CI: confidence interval, CV: cross-validation, AUC: area under the ROC curve, SN: sensitivity, SP: specificity, AC: accuracy.

Figure 5. Identification of biomarkers for predicting time to SPM development. (A) The statistically significant genes identified by univariate Cox regression across P_{SPM} patients with p16-negative tumor in cohort 2 (n = 26). The empirical *P*-values were generated by 5,000 permutation tests. **(B)** Kaplan-Meier curves showing the proportions of P_{SPM} patients with p16-negative tumor stratified by standardized *ITPR3* and *DSG3* expression. High and low levels of *ITPR3* and *DSG3*

expression were defined as standardized gene levels above or below mean, respectively. Patients with high *ITPR3* and high *DSG3* expression levels had the shortest time to SPM development (log-rank test $P = 0.017$) HR: Hazard ratio, CI: confidence interval, *Statistically significant ($P < 0.05$), **Statistically significant ($P < 0.01$).

Figure 6. Comparison of the expression patterns between protein and mRNA levels of the candidate biomarkers across the P_{index} and P_{SPM} subgroups. The standardized NanoString gene expressions of the top 10 statistically significant genes from cohort 2 were compared to their standardized protein levels in the tumor samples of cohort 1 as quantified by LC-MS/MS. Color codes represent the standardized levels of protein and mRNA expression from low (blue) to high (red).

Table

Table 1. Clinical characteristics of patients.

Figure 1:

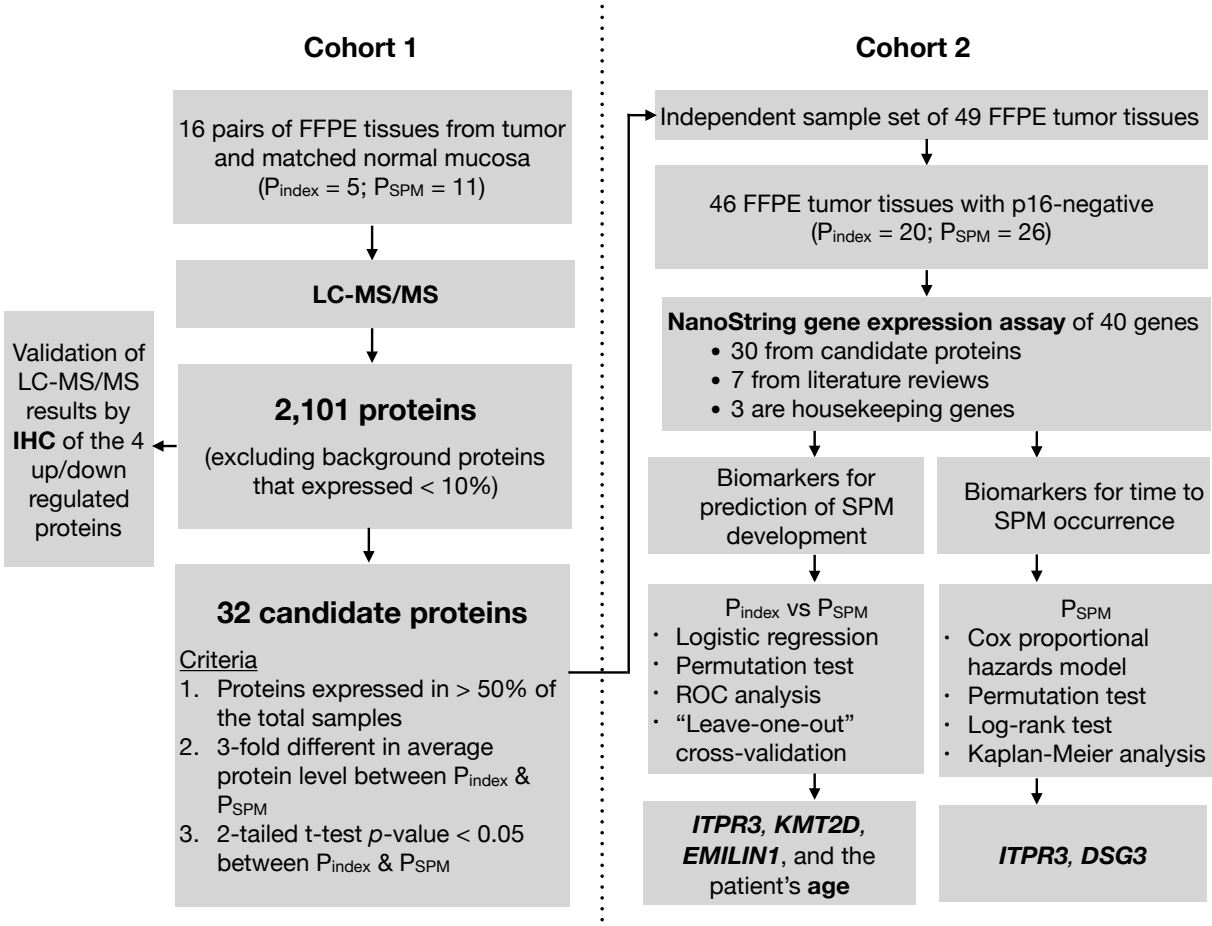

Cohort 1
Cohort 2
16 pairs of FFPE tissues from tumor and matched normal mucosa
(P_{index} = 5; P_{SPM} = 11)
LC-MS/MS
2,101 proteins
(excluding background proteins that expressed < 10%)
Validation of LC-MS/MS results by IHC of the 4 up/down regulated proteins
32 candidate proteins
Criteria
1. Proteins expressed in > 50% of the total samples
2. 3-fold different in average protein level between P_{index} & P_{SPM}
3. 2-tailed t-test p-value < 0.05 between P_{index} & P_{SPM}
Independent sample set of 49 FFPE tumor tissues
46 FFPE tumor tissues with p16-negative
(P_{index} = 20; P_{SPM} = 26)
NanoString gene expression assay of 40 genes
• 30 from candidate proteins
• 7 from literature reviews
• 3 are housekeeping genes
Biomarkers for prediction of SPM development
Biomarkers for time to SPM occurrence
P_{index} vs P_{SPM}
· Logistic regression
· Permutation test
· ROC analysis
· "Leave-one-out" cross-validation
P_{SPM}
· Cox proportional hazards model
· Permutation test
· Log-rank test
· Kaplan-Meier analysis
ITPR3, KMT2D, EMILIN1, and the patient's age
ITPR3, DSG3

Figure 2:

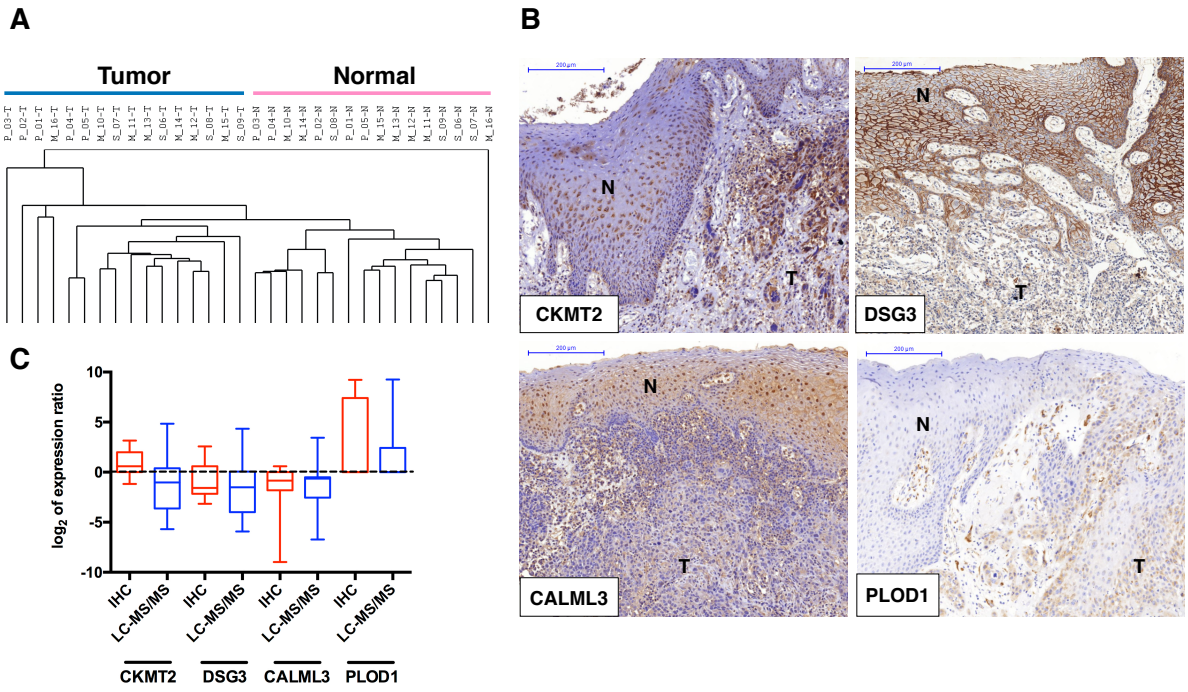
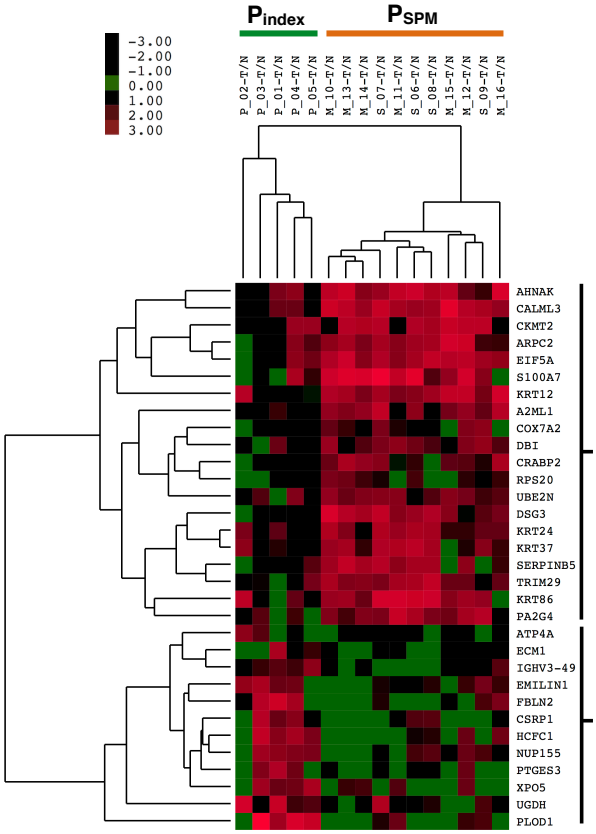

A
Tumor
Normal
B
N
T
CKMT2
DSG3
CALML3
PLOD1
C
log2 of expression ratio
10
5
0
-5
-10
IHC
LC-MS/MS
CKMT2
DSG3
CALML3
PLOD1

Figure 3:



GO term	Enrichment Score	*P*	Protein
Intermediate filament	3.88	<0.001**	KRT37, KRT12, KRT86, KRT24
Cytoplasm	0.88	0.024*	UBE2N, PA2G4, S100A7, ARPC2, SERPINB5, TRIM29, CRABP2, EIF5A, RPS20, AHNAK, KRT24
Poly(A) RNA binding	0.69	0.034*	UBE2N, PA2G4, EIF5A, RPS20, AHNAK

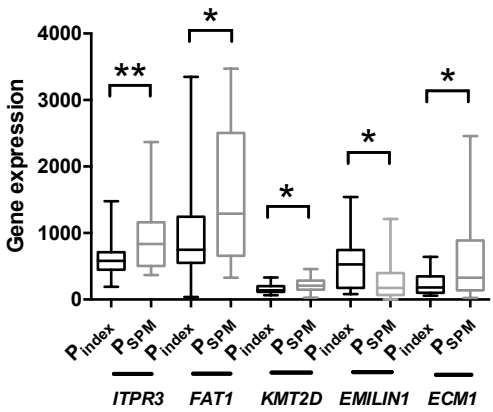
GO term	Enrichment Score	*P*	Protein
Extracellular exosome / Extracellular matrix	1.31	<0.01**	PTGES3, PLOD1, FBLN2, UGDH, CSRP1, ECM1, EMILIN1

Figure 4:

A

Univariate logistic regression			
	OR (95% CI)	***P***	***Empirical P***
Clinical variable			
Age	0.93 (0.87-0.99)	**0.025***	
Markers			
ITPR3	3.27 (1.36-11.04)	**0.025***	**0.007****
FAT1	2.25 (1.13-5.49)	**0.038***	**0.023***
KMT2D	2.20 (1.10-5.18)	**0.043***	**0.031***
EMILIN1	0.52 (0.25-0.97)	0.055	**0.046***
ECM1	2.71 (1.13- 9.73)	0.069	**0.033***

B



C

Predictive model: *ITPR3-KMT2D-EMILIN1*-patient's age

	AUC (95% CI)	**%SN (95% CI)**	**%SP (95% CI)**	**%AC (95% CI)**
Samples cohort 2 (n = 46)	0.86 (0.75-0.97)	88.46% (69.85-97.55)	75.00% (50.90-91.34)	82.61% (68.58-92.18)
CV cohort 2 (n = 46)	0.80 (0.66-0.94)	76.92% (56.35-91.03)	70.00% (45.72-88.11)	73.91% (58.87-85.73)

Figure 5:

A

Univariate Cox regression

	HR (95% CI)	*P*	*Empirical P*
Markers			
ITPR3	2.68 (1.53-4.72)	**0.001****	**0.001****
DBI	1.73 (1.14-2.62)	**0.010***	**0.011***
AHNAK	1.71 (1.13-2.57)	**0.010***	**0.023***
IGHV3-49	1.57 (1.05-2.33)	**0.026***	**0.045***
CALML3	1.57 (1.03-2.39)	**0.035***	**0.044***
ARPC2	1.53 (1.03-2.27)	**0.037***	0.058
DSG3	1.70 (1.03-2.82)	**0.039***	**0.046***
KRT37	1.51 (1.01-2.24)	**0.045***	0.069

B

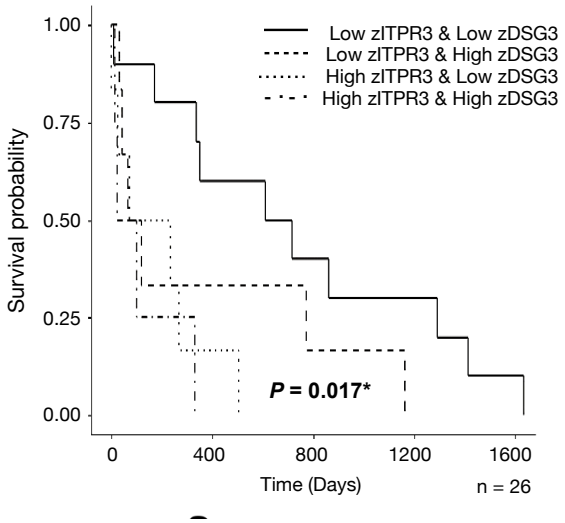


Figure 6:

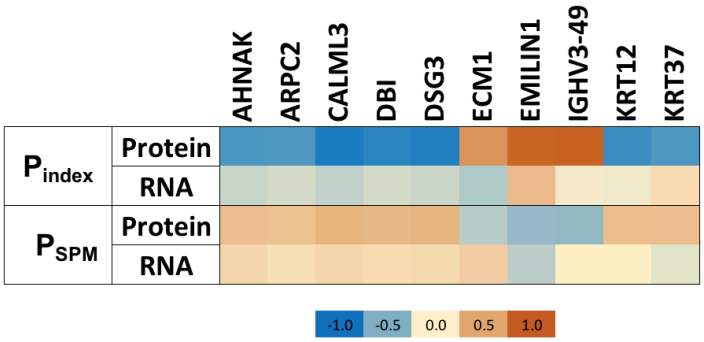

AHNAK
ARPC2
CALML3
DBI
DSG3
ECM1
EMILIN1
IGHV3-49
KRT12
KRT37
P index
Protein
RNA
P SPM
Protein
RNA
-1.0
-0.5
0.0
0.5
1.0

Table 1.

	Cohort 1				Cohort 2			
	P_{index}	P_{SPM}	Total	*P*-value	P_{index}	P_{SPM}	Total	*P*-value
No. of patients	5	11	16		21	28	49	
Age								
mean (years)	61	57	58	0.387[a]	64	57	60	***0.027***[a*]
<65	4	9	13	1.000[c]	11	22	33	0.053[b]
≥65	1	2	3		10	6	16	
Gender								
Male	5	10	15	1.000[c]	18	28	46	0.072[c]
Female	0	1	1		3	0	3	
Smoking status								
Never	0	1	1	1.000[c]	2	2	4	1.000[c]
Former[d]	1	2	3		7	9	16	
Active	2	5	7		11	16	27	
Unknown	2	3	5		1	1	2	
Alcohol consumption								
Never	2	1	3	0.500[c]	5	1	6	0.128[c]
Former[d]	0	2	2		4	5	9	
Active	1	3	4		10	18	28	
Unknown	2	5	7		2	4	6	
Site of primary (Index) tumor								
Oral cavity	0	1	1	1.000[c]	6	4	10	0.172[c]
Oropharynx	0	2	2		3	8	11	
Hypopharynx	4	7	11		5	12	17	
Larynx	1	1	2		6	4	10	
Nose & Paranasal sinus	0	0	0		1	0	1	
AJCC stage (7th ed) of primary (Index) tumor								
I	0	3	3	0.646[c]	2	3	5	0.105[c]
II	0	0	0		2	7	9	
III	3	4	7		7	2	9	
IV	2	4	6		10	16	26	
HPV status								
p16 (+)	1	2	3	1.000[c]	1	2	3	1.000[c]
p16 (-)	4	7	11		20	26	46	
Unknown	0	2	2		0	0	0	
Median time from inde tumor diagnosis to last follow up or death (months)	30.8 (29.7-8(				54.1 (35.3-94			
Classification of SPM								
Synchronous SPM		4				12		
Metachronous SPM		7				16		
Site of SPM								
Head and Neck		0				3		
Esophagus		11				21		
Lung		0				4		
Median time from inde tumor diagnosis to SPN occurrence (months)		14.1 (0-118.6)				10.0 (0-65.3)		

[a]t-test, [b]Chi-Square test, [c]Fisher's exact test, [d]Quit smoking at least 1 year before this study began, [*]Statistically significant ($P < 0.05$)

Synchronous SPM = SPM diagnosed simultaneously or within 6 months after the index tumor.
Metachronous SPM = SPM diagnosed more than 6 months after the index tumor.

8. Supplementary figures and tables

Please see the link below for supplementary information;
https://www.nature.com/articles/s41379-019-0211-2

9. Appendix

1) List of published manuscripts supported by RSA5780065

1: Bunbanjerdsuk S, Vorasan N, Saethang T, Pongrujikorn T, Pangpunyakulchai D, Mongkonsiri N, Arsa L, Thokanit N, Pongsapich W, Anekpuritanang T, Ngamphaiboon N, Jinawath A, Sunpaweravong S, Pisitkun T, Suktitipat B, Jinawath N. Oncoproteomic and gene expression analyses identify prognostic biomarkers for second primary malignancy in patients with head and neck squamous cell carcinoma. Mod Pathol. 2019 Feb 8. doi: 10.1038/s41379-019-0211-2. [Epub ahead of print] PubMed PMID: 30737471. JIF 6.655

2: Shiao MS, Chiablaem K, Charoensawan V, Ngamphaiboon N, Jinawath N. Emergence of Intrahepatic Cholangiocarcinoma: How High-Throughput Technologies Expedite the Solutions for a Rare Cancer Type. Front Genet. 2018 Aug 15;9:309. doi: 10.3389/fgene.2018.00309. eCollection 2018. Review. PubMed

PMID: 30158952; PubMed Central PMCID: PMC6104394. **JIF 4.151**

3: **Jinawath N**, Shiao MS, Norris A, Murphy K, Klein AP, Yonescu R, Iacobuzio-Donahue C, Meeker A, Jinawath A, Yeo CJ, Eshleman JR, Hruban RH, Brody JR, Griffin CA, Harada S. Alterations of type II classical cadherin, cadherin-10 (CDH10), is associated with pancreatic ductal adenocarcinomas. Genes Chromosomes Cancer. 2017 May;56(5):427-435. doi: 10.1002/gcc.22447. Epub 2017 Mar 7. PubMed PMID: 28124395. **JIF 3.96**

4: **Jinawath N**, Bunbanjerdsuk S, Chayanupatkul M, Ngamphaiboon N, Asavapanumas N, Svasti J, Charoensawan V. Bridging the gap between clinicians and systems biologists: from network biology to translational biomedical research. J Transl Med. 2016 Nov 22;14(1):324. Review. PubMed PMID: 27876057; PubMed Central PMCID: PMC5120462. **JIF 3.786**

2) Reprints of published manuscript

Modern Pathology
https://doi.org/10.1038/s41379-019-0211-2

ARTICLE

Oncoproteomic and gene expression analyses identify prognostic biomarkers for second primary malignancy in patients with head and neck squamous cell carcinoma

Sacarin Bunbanjerdsuk[1,2] · Nutchavadee Vorasan[3] · Thammakorn Saethang[4] · Tanjitti Pongrujikorn[1] · Duangjai Pangpunyakulchai[5] · Narongsak Mongkonsiri[5] · Lalida Arsa[6] · Nintita Thokanit[7] · Warut Pongsapich[8] · Tauangtham Anekpuritanang[9] · Nuttapong Ngamphaiboon[10] · Artit Jinawath[6] · Somkiat Sunpaweravong[11] · Trairak Pisitkun[4] · Bhoom Suktitipat[3,12,13] · Natini Jinawath[1,7,13]

Received: 16 September 2018 / Revised: 21 December 2018 / Accepted: 6 January 2019

Abstract

Patients with head and neck squamous cell carcinoma are at increased risk of developing a second primary malignancy, which is associated with poor prognosis and early death. To help improve clinical outcome, we aimed to identify biomarkers for second primary malignancy risk prediction using the routinely obtained formalin-fixed paraffin-embedded tissues of the index head and neck cancer. Liquid chromatography-tandem mass spectrometry was initially performed for candidate biomarker discovery in 16 pairs of primary cancer tissues and their matched normal mucosal epithelia from head and neck squamous cell carcinoma patients with or without second primary malignancy. The 32 candidate proteins differentially expressed between head and neck cancers with and without second primary malignancy were identified. Among these, 30 selected candidates and seven more from literature review were further studied using NanoString nCounter gene expression assay in an independent cohort of 49 head and neck cancer patients. Focusing on the p16-negative cases, we showed that a multivariate logistic regression model comprising the expression levels of *ITPR3*, *KMT2D*, *EMILIN1*, and the patient's age can accurately predict second primary malignancy occurrence with 88% sensitivity and 75% specificity. Furthermore, using Cox proportional hazards regression analysis and survival analysis, high expression levels of *ITPR3* and *DSG3* were found to be significantly associated with shorter time to second primary malignancy development (log-rank test $P = 0.017$). In summary, we identified a set of genes whose expressions may serve as the prognostic biomarkers for second primary malignancy occurrence in head and neck squamous cell carcinomas. In combination with the histopathologic examination of index tumor, these biomarkers can be used to guide the optimum frequency of second primary malignancy surveillance, which may lead to early diagnosis and better survival outcome.

Introduction

Head and neck cancer is the seventh most common cancer worldwide, with over 600,000 new cases in 2012 [1]. About 90% of all head and neck cancers are squamous cell carcinomas. Head and neck squamous cell carcinoma is associated with an elevated likelihood of developing second primary malignancy, which is defined as a second malignancy that presents either simultaneously or after the diagnosis of an index tumor. The overall incidence of second primary malignancies in head and neck squamous cell carcinoma patients has been reported to range from 5.6 to 35.9% [2, 3], with an annual incidence ranging from 3.2 to 4% [4]. The common sites of second primary malignancy are the head and neck, lung, and esophagus [5]. Several risk factors for second primary malignancy in head and neck squamous cell carcinoma patients have been reported including the location of index tumor, patient age, cigarette smoking, alcohol consumption, and betel nut chewing [4].

Despite the progress in cancer molecular biology, the exact underlying molecular mechanism of second primary

Supplementary information The online version of this article (https://doi.org/10.1038/s41379-019-0211-2) contains supplementary material, which is available to authorized users.

✉ Natini Jinawath
natini.jin@mahidol.ac.th
jnatini@hotmail.com

Extended author information available on the last page of the article

Published online: 08 February 2019

malignancy is still poorly understood. Field cancerization is one of the widely accepted concepts that explains second primary malignancy tumorigenesis [6]. According to this concept, an area of the upper aerodigestive tract is considered as a field that is continuously exposed to a diversity of carcinogens resulting in multiple precancerous genetic changes, even though the mucosal epithelia still retain their normal histologic appearances. These precancerous fields may eventually become malignant leading to multiple second primary malignancy occurrences [7]. Recently, Curtius et al. has reviewed the evolutionary process that results in field creation [8]. In this context, a cancerized field is both enabled by and causes alterations in the tissue microenvironment. Measurements of the cancerized field evolution hold considerable promise as a new class of biomarker for cancer risk.

Second primary malignancies are known to have a major negative impact on head and neck cancer patients. The prognoses of head and neck cancer patients who develop second primary malignancy are worse than those with index malignancies alone [9], and it is also a leading long-term cause of mortality in head and neck cancer patients [10]. Early diagnosis of second primary malignancies often allows less invasive or curative treatments. However, current clinical screening and surveillance methods including panendoscopy and positron emission tomography/computed tomography (PET/CT) scan still have clinical limitations [11, 12]. Panendoscopy is a relatively invasive procedure and can result in serious complications such as esophageal perforation [11]. While PET/CT scan demonstrates high sensitivity for detecting synchronous second primary malignancies, its findings can be false-positive, and may miss small and/or superficial synchronous second primary malignancies [13]. Furthermore, the attempts to use chemoprevention for second primary malignancy risk reduction have so far been unsuccessful [14]. For these reasons, novel screening approaches such as molecular biomarkers for second primary malignancy risk prediction are very much needed.

Advances in omics technology have resulted in a better understanding of head and neck cancer carcinogenesis, which in turn lead to the development of novel molecular biomarkers for this type of cancer that can be successfully translated into routine clinical practice. Human papillomavirus (HPV) has emerged as an important etiologic factor of head and neck cancers, particularly in oropharyngeal cancers. HPV status is currently used as a prognostic biomarker for head and neck cancer; HPV-positive head and neck squamous cell carcinomas have a favorable prognosis and may benefit from less aggressive treatment regimens [15]. The recently released 8th edition of the American Joint Committee on Cancer (AJCC) Staging Manual, Head and Neck Section, has acknowledged the importance of using p16 immunostaining as a surrogate marker of HPV status, and issued a distinct staging classification for HPV-associated oropharyngeal cancer, separating it from cancer of the oropharynx related to other causes [16]. Despite these advances, to date, very few studies have focused on second primary malignancy, particularly on a biomarker discovery aspect [17, 18], and so far none has been successfully translated into clinics.

In this study, we hypothesized that the cancerized fields in head and neck squamous cell carcinoma patients with or without second primary malignancy are diverse, leading to distinct protein or gene expression patterns that can be exploited as biomarkers. We then aimed to identify molecular biomarkers for second primary malignancy risk prediction using the routinely collected formalin-fixed paraffin-embedded tissue biopsies of the index head and neck squamous cell carcinomas. Using liquid chromatography-tandem mass spectrometry, we identified a set of candidate biomarkers differentially expressed in the primary tumors of head and neck squamous cell carcinoma patients with or without second primary malignancy after a minimum follow-up period of 3 years. Next, we utilized a NanoString nCounter gene expression assay to study the ability of selected biomarkers to predict second primary malignancy development and the time to second primary malignancy occurrence in another independent cohort of head and neck squamous cell carcinoma patients. Our study permitted the development of the first tissue-based gene expression biomarker panel for second primary malignancy that can easily be implemented in the routine pathology practice.

Materials and methods

Patient selection and sample description

All patients were clinically diagnosed with head and neck cancer by the otolaryngologists or oncologists and had a histopathological diagnosis of squamous cell carcinoma. The patients with only a single primary tumor after a minimum follow-up period of approximately 3 years were defined as "P_{index} group". The head and neck squamous cell carcinoma patients who developed second primary malignancy during the same follow-up period were defined as "P_{SPM} group". The clinical diagnosis of second primary malignancy was made based on the applied recommendation from Warren and Gates' criteria [19] as follows: (i) both index and secondary tumors had histologic confirmation of squamous cell carcinoma; (ii) the two malignancies were anatomically separated by more than 2 cm of normal mucosa; (iii) the possibility of the second primary malignancy being a metastasis from the index tumor was

excluded. Second primary malignancy was further classified as synchronous second primary malignancy when the diagnosis of both index tumor and second tumor was made simultaneously or within 6 months after the index tumor, or as metachronous second primary malignancy when the diagnosis of second tumor was made longer than 6 months after the index tumor [20].

Cohort 1 consisted of 16 head and neck squamous cell carcinoma patients from Songklanagarind Hospital (Prince of Songkla University, Songkhla, Thailand). A pair of formalin-fixed paraffin-embedded tissue blocks containing the index tumor and normal mucosal epithelia was collected for each patient, resulting in a total of 32 samples for proteomic analysis. In addition, all second primary malignancy samples were previously confirmed to be second primary malignancies, not metastatic tumors, by comparing the loss of heterozygosity patterns identified by single nucleotide polymorphism array between each index head and neck squamous cell carcinoma and its matched second tumor [21]. Cohort 2 was an independent multi-center cohort comprising 49 head and neck squamous cell carcinoma patients from Ramathibodi Hospital (Mahidol University, Bangkok, Thailand), Siriraj Hospital (Mahidol University, Bangkok, Thailand), and Songklanagarind Hospital. Only formalin-fixed paraffin-embedded tissue blocks containing the index tumor were collected for each patient. This study was approved by the Institutional Review Board of all participating hospitals.

Sample preparation

Formalin-fixed paraffin-embedded tissue sections were prepared. Previously archived hematoxylin and eosin stained tissue slides were evaluated for tumor content and density by pathologists. The areas with at least 70% tumor cells and normal mucosal epithelial cells were marked for manual macrodissection using a needle tip or scalpel. For liquid chromatography-tandem mass spectrometry, dissected tissue was prepared for protein digestion using the filter-assisted sample preparation method [22] and physically disrupted by sonication. For NanoString gene expression assay, total RNA was isolated using High Pure FFPET RNA Isolation Kit (Roche) according to the manufacturer's instructions.

Liquid chromatography-tandem mass spectrometry

Liquid chromatography-tandem mass spectrometry analysis was performed as a service by Bioproximity, LLC (Chantilly, VA, USA). In brief, proteins were digested with trypsin, and digested peptides were desalted using C18 stop-and-go extraction (STAGE) tips [23]. Peptides were then fractionated by strong anion exchange STAGE tip chromatography [24]. Liquid chromatography was performed on an Easy nanoLC II HPLC system (Thermo Fisher Scientific). The liquid chromatography was interfaced to a dual pressure linear ion trap mass spectrometer (LTQ Velos, Thermo Fisher Scientific) via nano-electrospray ionization. Data from mass spectrometry were processed and the Mascot generic format (MGF) files were searched using X!!Tandem, k-score scoring algorithms and OMSSA (open mass spectrometry search algorithm). The common Repository of Adventitious Proteins and the Ensembl release 69 were used as protein databases for the searches. Proteins were required to have two or more unique peptides across the analyzed samples with *E*-value scores of 0.01 or less and protein *E*-value scores of 0.0001 or less [25]. The spectral count of each identified protein was initially scaled by the total and/or maximum peptide counts of the same patient. All scaled spectral counts were added with 0.01 to avoid division by zero, and each scaled spectral count of tumor tissue was then divided by the scaled spectral count of the corresponding normal tissue from the same patient to normalize interindividual variations.

Immunohistochemistry

A Leica BOND-MAX™ automated staining system (Leica Biosystems) was utilized according to the manufacturer's protocol for anti-CALML3 (1:1000, PA5-30232, Thermo Fisher Scientific), and anti-CKMT2 (1:200, PA5-28591, Thermo Fisher Scientific). A Ventana BenchMark XT automated staining system (Ventana Medical Systems) was used according to the manufacturer's instructions for anti-DSG3 (1:50, MAB1720, R&D Systems), anti-PLOD1 (1:50, NBP2-31885, Novus Biologicals), and anti-p16 (CINtec® p16 Histology, 705-4713, Ventana Medical Systems). Blinded scoring was independently performed by two pathologists using the same scoring criteria (Supplementary Table S1).

Digital gene expression quantification using NanoString nCounter®

A NanoString nCounter® Gene Expression Assay (NanoString Technologies) with the custom-designed CodeSets containing 40 genes was performed following the standard protocol. Data processing was conducted using nSolver Analysis Software v3.0 (NanoString Technologies). Background correction was made by subtracting the "mean + 2 standard deviation" value of the negative controls from the raw counts. Adjusted raw counts were then normalized with the geometric mean of the positive controls in each sample. The resulting data was normalized again using the geometric mean of three housekeeping genes. Normalized data was standardized using mean centering then divided by

Table 1 Clinical characteristics of patients

	Cohort 1				Cohort 2			
	P_{index}	P_{SPM}	Total	*P*-value	P_{index}	P_{SPM}	Total	*P*-value
No. of patients	5	11	16		21	28	49	
Age								
Mean (years)	61	57	58	0.387[a]	64	57	60	**0.027**[a]*
< 65	4	9	13	1.000[c]	11	22	33	0.053[b]
≥ 65	1	2	3		10	6	16	
Gender								
Male	5	10	15	1.000[c]	18	28	46	0.072[c]
Female	0	1	1		3	0	3	
Smoking status								
Never	0	1	1	1.000[c]	2	2	4	1.000[c]
Former[d]	1	2	3		7	9	16	
Active	2	5	7		11	16	27	
Unknown	2	3	5		1	1	2	
Alcohol consumption								
Never	2	1	3	0.500[c]	5	1	6	0.128[c]
Former[d]	0	2	2		4	5	9	
Active	1	3	4		10	18	28	
Unknown	2	5	7		2	4	6	
Site of primary (Index) tumor								
Oral cavity	0	1	1	1.000[c]	6	4	10	0.172[c]
Oropharynx	0	2	2		3	8	11	
Hypopharynx	4	7	11		5	12	17	
Larynx	1	1	2		6	4	10	
Nose & Paranasal sinuses	0	0	0		1	0	1	
Stage[e] of primary (Index) tumor								
I	0	3	3	0.646[c]	2	3	5	0.105[c]
II	0	0	0		2	7	9	
III	3	4	7		7	2	9	
IV	2	4	6		10	16	26	
HPV status								
p16 (+)	1	2	3	1.000[c]	1	2	3	1.000[c]
p16 (−)	4	7	11		20	26	46	
Unknown	0	2	2		0	0	0	
Median time from index tumor diagnosis to last follow-up or death (months)	31 (30–86)				54 (35–95)			
Classification of SPM								
Synchronous SPM		4				12		
Metachronous SPM		7				16		
Site of SPM								
Head and Neck		0				3		
Esophagus		11				21		
Lung		0				4		
Median time from index tumor diagnosis to SPM occurrence (months)		14 (0–119)				10 (0–65)		

SPM second primary malignancy, *Synchronous SPM* second primary malignancy diagnosed simultaneously or within 6 months after the index tumor, *Metachronous SPM* second primary malignancy diagnosed more than 6 months after the index tumor

[a]*t* test

[b]Chi-Square test

[c]Fisher's exact test

[d]Quit smoking at least 1 year before this study began

[e]The American Joint Committee on Cancer staging (7th edition)

*Statistically significant ($P < 0.05$)

the standard deviation to obtain the *Z* scores for subsequent statistical analyses.

Bioinformatics and statistical analysis

Hierarchical cluster analysis using average linkage method was performed using Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm), and visualized by Java TreeView version 1.1.6r4 (http://jtreeview.sourceforge.net). Functional annotation of candidate proteins was performed using DAVID 6.8 (https://david.ncifcrf.gov/). Statistical analyses were performed using PASW Statistics version 18.0 for Windows (SPSS Inc.), GraphPad Prism version 6.0 for Windows (GraphPad Software), R-statistical software version 3.3.1 (R Foundation, http://www.r-project.org), and ROCR package [26].

Results

Subject characteristics

Two independent cohorts of head and neck squamous cell carcinoma patients were included in this study. Among the 16 patients in cohort 1, 5 (31%) and 11 (69%) patients belonged to the P_{index} and P_{SPM} groups, respectively. The median duration of follow-up to ensure no occurrence of second primary malignancy was 31 (30–86) months in the P_{index} group. All second primary malignancies were esophageal squamous cell carcinomas. No statistically significant difference was detected between the clinical characteristics of patients in the P_{index} and P_{SPM} groups.

In cohort 2, of the 49 head and neck squamous cell carcinoma patients examined, 21 (43%) and 28 (57%) patients belonged to the P_{index} and P_{SPM} groups, respectively. The median duration of follow-up was 54 (35–95) months in the P_{index} group. All second primary malignancies were squamous cell carcinomas located in the esophagus, head and neck, or lung. The patients' age was the only clinical characteristic with statistically significant difference between the P_{index} and P_{SPM} groups ($P = 0.027$). The detailed clinicopathological and demographical characteristics of patients in both cohorts are summarized in Table 1.

The HPV status of head and neck tumors was also evaluated using p16 immunohistochemistry. The difference in HPV status based on p16-positivity between the P_{index} and P_{SPM} groups from the two cohorts was not statistically significant (Fisher's exact test $P = 1.000$). Of note, only three cases out of the 49 head and neck squamous cell carcinomas in cohort 2 were p16-positive. The p16 immunostaining results are shown in Supplementary Table S2.

Liquid chromatography-tandem mass spectrometry analysis and validation by immunohistochemistry

The detailed experimental outline of this study is summarized in Fig. 1. We first performed liquid chromatography-tandem mass spectrometry in 16 pairs of tumor-normal formalin-fixed paraffin-embedded tissues from P_{index} and P_{SPM} groups (cohort 1) as a means to discover candidate biomarkers for second primary malignancy. A total of 2816 proteins were initially identified. Proteins detected in <10% of all samples were then filtered out, leaving a total of 2101 proteins for unsupervised hierarchical cluster analysis. The resulted dendrogram showed that these protein profiles could correctly classify the tumor and normal tissue samples (Fig. 2a). We further investigated the protein profiles within the tumor or normal tissue subgroups separately whether they could correctly define each sample as P_{index} or P_{SPM} by performing hierarchical cluster analysis of differentially expressed proteins (*t* test $P < 0.05$). The results interestingly demonstrated that P_{index} and P_{SPM} samples could be accurately clustered in both tumor and normal tissue subgroups (Supplementary Figure S1).

In addition, immunohistochemistry analysis of representative proteins was performed to validate the accuracy of liquid chromatography-tandem mass spectrometry results. Four proteins that were at least two-fold up- (PLOD1) or downregulated (CKMT2, DSG3, CALML3) in tumors as compared to normal tissues, and expressed in more than half of the samples, were selected. The results demonstrated that immunohistochemistry analysis showed the same up- or downregulated trend between normal and tumor tissues as in the proteomics data in three of four proteins (DSG3, CALML3, PLOD1) (Fig. 2b–c).

Identification of candidate proteins biomarkers for second primary malignancy

To screen for the differentially expressed proteins between head and neck squamous cell carcinoma patients with and without second primary malignancy, the protein expression ratio of tumor to matched normal tissue of each patient was analyzed. The inclusion criteria for candidate proteins included: (i) proteins expressed in more than 50% of the samples, (ii) proteins showing at least three-fold difference in expression level between the P_{index} and P_{SPM} groups, and (iii) proteins exhibiting statistically significant differential expression between the two groups (*t* test $P < 0.05$). Based on these criteria, 32 candidate proteins were identified (Supplementary Table S3). Hierarchical cluster analysis using expression ratio of these proteins demonstrated that all cohort 1 samples could be correctly divided into P_{index} and P_{SPM} groups (Fig. 3).

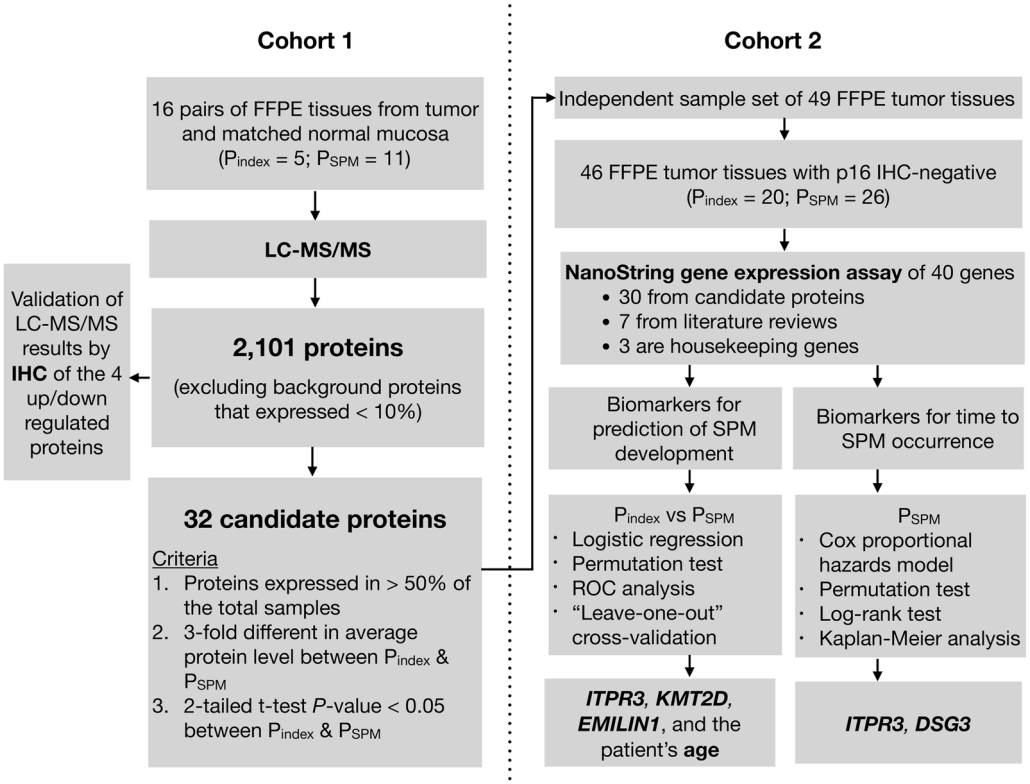


Fig. 1 Overview and experimental design of the study. The biomarker identification steps were sequentially conducted using liquid chromatography-tandem mass spectrometry in cohort 1 (left), and NanoString gene expression assay in cohort 2 (right). Multiple statistical analyses were performed to identify prognostic biomarkers for second primary malignancy occurrence and time to second primary malignancy development. SPM second primary malignancy, P_{index} index head and neck squamous cell carcinoma without second primary malignancy, P_{SPM} index head and neck squamous cell carcinoma with second primary malignancy, FFPE formalin-fixed paraffin-embedded, LC-MS/MS liquid chromatography-tandem mass spectrometry, IHC immunohistochemistry, ROC receiver operating characteristic

We further carried out gene-annotation enrichment analysis using DAVID 6.8 to identify enriched biological functions of the candidate proteins. Among the upregulated proteins in P_{SPM} samples, the statistically significant gene ontology term ($P < 0.001$) with highest enrichment score (3.88) was "intermediate filament" (KRT37, KRT12, KRT86, KRT24). The other significant gene ontology terms ($P < 0.05$) were "cytoplasm" (UBE2N, PA2G4, S100A7, ARPC2, SERPINB5, TRIM29, CRABP2, EIF5A, RPS20, AHNAK, KRT24), and "poly(A) RNA binding" (UBE2N, PA2G4, EIF5A, RPS20, AHNAK). In contrast, the significant terms with highest enrichment score (1.31) among the downregulated proteins were "extracellular exosome" and "extracellular matrix" (PTGES3, PLOD1, FBLN2, UGDH, CSRP1, ECM1, EMILIN1) ($P < 0.01$) (Fig. 3 and Supplementary Table S4A-B).

Development of a predictive model for second primary malignancy occurrence by NanoString nCounter gene expression analysis

Considering the difficulties in translating tissue-based quantitative protein biomarkers into clinical practice, we set out to alternatively utilize a NanoString nCounter gene expression assay to optimize the biomarker selection in an independent cohort (cohort 2). The 40-gene custom panel consisted of the 30 candidate proteins (of the 32 candidates, two did not have specific CodeSets available), three housekeeping genes, and seven frequently mutated genes in head and neck squamous cell carcinomas with synchronous nodal metastasis or metachronous recurrence from literature review [27, 28] (Supplementary Table S5). HPV-positive and HPV-negative head and neck cancers are known to exhibit different clinical and molecular characteristics [16]. Since the number of p16-positive cases in our cohort was limited (3 out of 49 cases), we focused on the 46 p16-negative head and neck squamous cell carcinoma patients for further analyses.

To identify biomarkers that could accurately determine the development of second primary malignancy, univariate logistic regression analysis of the standardized NanoString gene expression levels and selected clinical variables of the 46 p16-negative head and neck squamous cell carcinoma patients was carried out. The expression levels of three genes (*ITPR3*, *FAT1, KMT2D)* and the patient's age at diagnosis were statistically

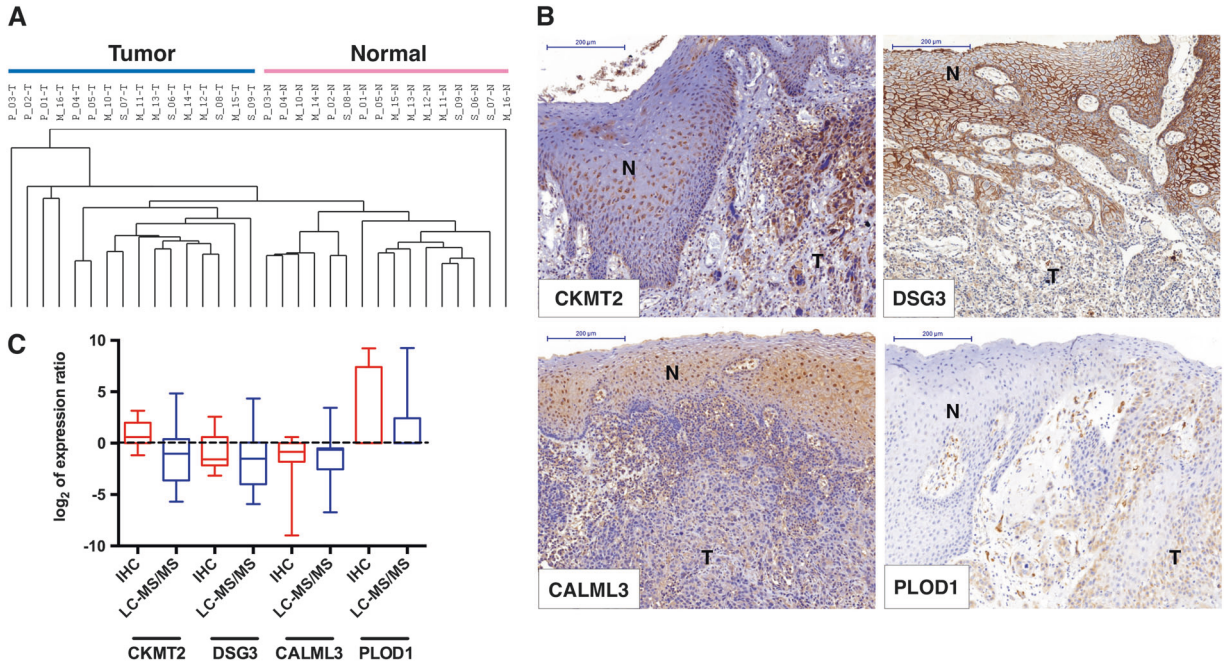
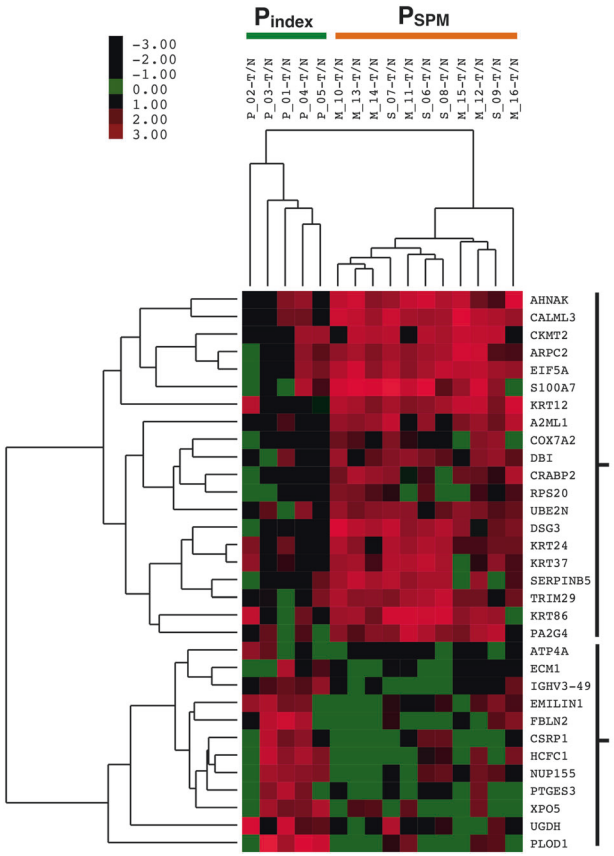


Fig. 2 Protein expression profiling by liquid chromatography-tandem mass spectrometry and validation by immunohistochemistry. **a** Dendrogram of unsupervised hierarchical cluster analysis of the 2101 protein profiles across the tumor samples and their matched normal mucosal tissues in cohort 1. **b** Representative immunohistochemistry images (×10 magnification) of the four selected proteins (CKMT2, DSG3, CALML3, and PLOD1). N: normal mucosal epithelial cells, T: tumor cells. **c** Box plot showing the comparison between normalized expression ratios (T/N) of the four selected proteins analyzed by immunohistochemistry and liquid chromatography-tandem mass spectrometry. IHC immunohistochemistry, LC-MS/MS liquid chromatography-tandem mass spectrometry

significantly associated with the development of second primary malignancy ($P < 0.05$) (Fig. 4a). The strongest risk factor for the development of second primary malignancy was high *ITPR3* level. An increase in one standard deviation of *ITPR3* level was associated with 3.27 times higher risk of second primary malignancy development (odds ratio = 3.27; 95% confidence interval (CI) [1.36–11.04]; $P = 0.025$). Other significant risk factors included high *FAT1* (odds ratio = 2.25; 95% CI [1.13–5.49]; $P = 0.038$) and *KMT2D* levels (odds ratio = 2.20; 95% CI [1.10–5.18]; $P = 0.043$). To minimize the effect of small sample size, we performed 5000 permutation tests to correct for non-asymptotic properties of *P*-values by reshuffling the observed data. The resulting *P*-values were reported as empirical *P*-values. The analysis showed that the expression levels of two more genes (*EMILIN1* and *ECM1*) were also statistically significantly associated with second primary malignancy development (empirical $P < 0.05$). The results of univariate logistic regression of all genes are shown in Supplementary Table S6. Comparing between P_{index} and P_{SPM} groups, the expression levels of *ITPR3*, *FAT1, KMT2D*, and *ECM1* were higher in P_{SPM} samples, whereas the level of *EMILIN1* was lower (Fig. 4b).

We next aimed to identify the best combination of biomarkers for second primary malignancy development prediction by analyzing the six statistically significant variables ($P < 0.05$ or empirical $P < 0.05$) from univariate logistic regression (*ITPR3*, *FAT1, KMT2D*, *EMILIN1*, *ECM1*, and patient's age) using multivariate logistic regression analysis. The most parsimonious final model based on Akaike's information criterion, Bayesian information criterion, and deviance test was the combination of *ITPR3*, *KMT2D*, *EMILIN1*, and the patient's age. The Akaike's information criterion, and Bayesian information criterion values of this model were 50.82 and 59.96, respectively. The detailed results of all multivariate logistic regression models and the final selection are shown in Supplementary Table S7 and S8.

To evaluate the performance of the selected multivariate model in predicting second primary malignancy occurrence, a receiver operating characteristic analysis was performed and an area under the curve value was calculated as 0.86 (95% CI [0.75–0.97]). Moreover, this model has a sensitivity of 88.46%, a specificity of 75.00% and an accuracy of 82.61% for predicting the occurrence of second primary malignancy. We further performed a leave-one-out cross-validation over the same 46 samples (cohort 2) to assess the robustness of our model. The resulting area under the curve, sensitivity, specificity and accuracy values were 0.80 (95% CI [0.66–0.94]), 76.92%, 70.00%, and 73.91%, respectively (Fig. 4c). These new values were not significantly different



Gene Ontology term	Enrichment Score	*P*	Protein
Intermediate filament	3.88	<0.001**	KRT37, KRT12, KRT86, KRT24
Cytoplasm	0.88	0.024*	UBE2N, PA2G4, S100A7, ARPC2, SERPINB5, TRIM29, CRABP2, EIF5A, RPS20, AHNAK, KRT24
Poly(A) RNA binding	0.69	0.034*	UBE2N, PA2G4, EIF5A, RPS20, AHNAK

Gene Ontology term	Enrichment Score	*P*	Protein
Extracellular exosome / Extracellular matrix	1.31	<0.01**	PTGES3, PLOD1, FBLN2, UGDH, CSRP1, ECM1, EMILIN1

Fig. 3 Hierarchical cluster analysis of the candidate proteins. A heatmap showing the 32 candidate proteins differentially expressed between P_{index} and P_{SPM} subgroups with statistical significance (*t* test $P < 0.05$) across cohort 1 (left), and the summary of gene-annotation enrichment analysis of up- and downregulated proteins of each subgroup using DAVID bioinformatics resources 6.8 (right). The color scale is shown at the upper left corner. P_{index}: index head and neck squamous cell carcinoma without second primary malignancy, P_{SPM}: index head and neck squamous cell carcinoma with second primary malignancy. *Statistically significant ($P < 0.05$), **statistically significant ($P < 0.01$)

from the original results, confirming the validity of our predictive model.

Identification of biomarkers for predicting time to second primary malignancy development

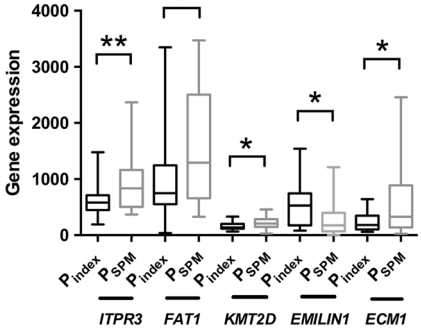
Next, we focused on head and neck squamous cell carcinoma patients who eventually developed second primary malignancy (P_{SPM} group) in order to identify the biomarkers associated with time to subsequent second primary malignancy occurrence. Univariate Cox regression analysis using gene expression levels of the 26 p16-negative head and neck squamous cell carcinomas from cohort 2 was performed. The levels of *ITPR3, DBI, AHNAK, IGHV3-49, CALML3, ARPC2, DSG3*, and *KRT37* were significantly associated with a shorter time to second primary malignancy development ($P < 0.05$) (Fig. 5a). The strongest association was with *ITPR3* level. An increase in one standard deviation of *ITPR3* level was associated with 2.68 times higher risk of second primary malignancy development at any given time after the index tumor diagnosis (hazard ratio = 2.68; 95% CI [1.53–4.72]; $P = 0.001$). The complete results of univariate Cox regression analysis including the empirical *P*-values generated by 5000 permutation tests are shown in Supplementary Table S9. Using multivariate Cox proportional hazards regression analysis, the best model was the combination of *ITPR3* and *DSG3* (Supplementary Table S10). Moreover, by dividing the standardized *ITPR3* and *DSG3* level by the sample mean of the 26 P_{SPM} patients, survival analysis showed that the patients with p16-negative head and neck squamous cell carcinoma whose index tumors exhibited high *ITPR3* and *DSG3* expression levels had the shortest time interval between the diagnosis of an index head and neck squamous cell carcinoma to subsequent second primary malignancy development (log-rank test $P = 0.017$) (median time difference between the high and low risk groups based on four combinations of *ITPR3* and *DSG3* levels = 394 days) (Fig. 5b and Supplementary Table S11).

A

Univariate logistic regression

	Odds ratio (95% CI)	*P*	*Empirical P*
Clinical variable			
Age	0.93 (0.87-0.99)	**0.025***	
Markers			
ITPR3	3.27 (1.36-11.04)	**0.025***	**0.007****
FAT1	2.25 (1.13-5.49)	**0.038***	**0.023***
KMT2D	2.20 (1.10-5.18)	**0.043***	**0.031***
EMILIN1	0.52 (0.25-0.97)	0.055	**0.046***
ECM1	2.71 (1.13- 9.73)	0.069	**0.033***

B



C

Predictive model: *ITPR3-KMT2D-EMILIN1*-patient's age

	Area under the curve (95% CI)	%Sensitivity (95% CI)	%Specificity (95% CI)	%Accuracy (95% CI)
Samples cohort 2 (n = 46)	0.86 (0.75-0.97)	88.46% (69.85-97.55)	75.00% (50.90-91.34)	82.61% (68.58-92.18)
Cross-validation cohort 2 (n = 46)	0.80 (0.66-0.94)	76.92% (56.35-91.03)	70.00% (45.72-88.11)	73.91% (58.87-85.73)

Fig. 4 Identification of biomarkers for second primary malignancy risk prediction by NanoString gene expression analysis. **a** The statistically significant variables identified by univariate logistic regression across P_{index} and P_{SPM} subgroups of the p16-negative tumors in cohort 2 ($n = 46$). The empirical *P*-values were generated by 5000 permutation tests. **b** Box plot depicting expression levels of the five candidate genes identified by univariate logistic regression across P_{index} and P_{SPM} subgroups of the p16-negative tumors in cohort 2. All pairwise comparisons show statistically significant differential expressions. **c** The predictive performance of the selected multivariate logistic regression model (*ITPR3-KMT2D-EMILIN1*-patient's age) in classifying second primary malignancy risk. The robustness of the model was assessed by performing a leave-one-out cross-validation for each sample in the same cohort. The unit of patient's age is year. P_{index}: index head and neck squamous cell carcinoma without second primary malignancy, P_{SPM}: index head and neck squamous cell carcinoma with second primary malignancy. *Statistically significant ($P < 0.05$), **statistically significant ($P < 0.01$), CI confidence interval

A

Univariate Cox regression

	Hazard ratio (95% CI)	*P*	*Empirical P*
Markers			
ITPR3	2.68 (1.53-4.72)	**0.001****	**0.001****
DBI	1.73 (1.14-2.62)	**0.010***	**0.011***
AHNAK	1.71 (1.13-2.57)	**0.010***	**0.023***
IGHV3-49	1.57 (1.05-2.33)	**0.026***	**0.045***
CALML3	1.57 (1.03-2.39)	**0.035***	**0.044***
ARPC2	1.53 (1.03-2.27)	**0.037***	0.058
DSG3	1.70 (1.03-2.82)	**0.039***	**0.046***
KRT37	1.51 (1.01-2.24)	**0.045***	0.069

B

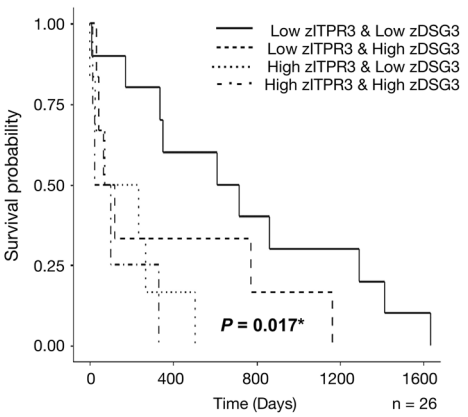


Fig. 5 Identification of biomarkers for predicting time to second primary malignancy development. **a** The statistically significant genes identified by univariate Cox regression across P_{SPM} patients with p16-negative tumor in cohort 2 ($n = 26$). The empirical *P*-values were generated by 5000 permutation tests. **b** Kaplan–Meier curves showing the proportions of P_{SPM} patients with p16-negative tumor stratified by standardized *ITPR3* and *DSG3* expression. High and low levels of *ITPR3* and *DSG3* expression were defined as standardized gene levels above or below mean, respectively. Patients with high *ITPR3* and high *DSG3* expression levels had the shortest time to second primary malignancy development (log-rank test $P = 0.017$). CI confidence interval, *statistically significant ($P < 0.05$), **statistically significant ($P < 0.01$)

Comparison of the expression pattern of candidate biomarkers between protein and mRNA levels

It is known that protein and mRNA expression levels often do not directly correlate [29]. To investigate whether our candidate genes originally discovered by proteomics study shared the same differential expression patterns as their protein equivalents, we compared the standardized gene expressions of the top 10 statistically significant genes identified by logistic regression and Cox regression analyses to their standardized protein levels obtained from liquid chromatography-tandem mass spectrometry. Seven of the 10 biomarkers (*AHNAK, ARPC2, CALML3, DBI, DSG3, EMILIN1, KRT12*) demonstrated a similar up- or downregulation trend between mRNA and protein levels in the tumor samples across the two patient cohorts (Fig. 6).

Discussion

In this study, we identified a set of tissue-based biomarkers for predicting second primary malignancy occurrence and time to second primary malignancy development in head and neck squamous cell carcinoma patients using a combination of high-throughput shotgun proteomics and targeted gene expression analysis. These second primary malignancy risk prediction biomarkers can help guide clinical management of head and neck squamous cell carcinoma patients, particularly in the frequency of second primary malignancy surveillance after the diagnosis of an index tumor, and the choice of treatments. The expected long-term benefit is an improvement in overall survival of head and neck cancer patients, especially those who are eligible for curative or less invasive therapy. Moreover, one of the clinical strengths of this study is the use of index head and neck squamous cell carcinoma formalin-fixed paraffin-embedded tissue as the preferred material for biomarker discovery. This specimen is routinely acquired for histopathological diagnosis, which makes it easier to incorporate our biomarker panel into the current clinical practice. We believe that our study is the first to identify prognostic biomarkers that can accurately predict second primary malignancy occurrence, thus opening the door to the possibility of clinical application of tissue-based biomarkers for second primary malignancy in head and neck squamous cell carcinoma patients.

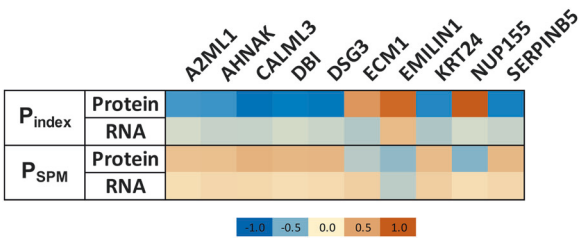


Fig. 6 Comparison of the expression patterns between protein and mRNA levels of the candidate biomarkers across the P_{index} and P_{SPM} subgroups. The standardized NanoString gene expressions of the top 10 statistically significant genes from cohort 2 were compared to their standardized protein levels in the tumor samples of cohort 1 as quantified by liquid chromatography-tandem mass spectrometry. Color codes represent the standardized levels of protein and mRNA expression from low (blue) to high (red)

Mass spectrometry has been extensively used to discover novel protein biomarkers [30]. Proteins are the functional molecules in the cell, and thus are the key players that represent actual cellular physiology. For tissue-based protein biomarkers, several laboratory methods such as immunohistochemistry and targeted mass spectrometry can be used to develop the validated markers into quantitative clinical assays but not without limitations. To date, immunohistochemistry is routinely performed in clinics to determine both the qualitative and semi-quantitative aspects of a protein biomarker, however, it relies heavily on the quality of antibody and still lacks interpretation standardization, resulting in poor reproducibility [31]. While targeted mass spectrometry can do marker multiplexing and does not rely on antibody, it is technically complex and carry a high developing cost [32]. The development of tissue-based high-throughput gene expression assays has greatly improved the accuracy and reproducibility of quantitative measurement in tissue biomarker studies. Currently, several such assays have been integrated into clinical practice. For example, the Prosigna breast cancer prognostic gene signature assay based on NanoString nCounter analysis system was approved by U.S. Food and Drug Administration in 2013 as a prognostic assay for distant recurrence risk in breast cancer patients [33]. In this study, we sequentially utilized liquid chromatography-tandem mass spectrometry and NanoString nCounter system in two independent cohorts of head and neck squamous cell carcinoma patients in order to identify the best set of second primary malignancy prediction biomarkers. In light of its clinical translatability, technical reproducibility, and compatibility with small biopsied formalin-fixed paraffin-embedded tissues [34], the NanoString digital gene expression platform was chosen for biomarker panel development.

In cohort 1, the protein profiles were markedly different between head and neck squamous cell carcinomas and their matched normal squamous mucosa, which is in line with previous studies [17, 35]. Interestingly, we also observed distinct protein expression profiles between the P_{index} and P_{SPM} subgroups when analyzing the tumor and normal tissue samples separately. These results suggest that tissue microenvironments of both tumor and surrounding normal mucosal cells of the P_{index} and P_{SPM} patients may exhibit varying levels of genetic diversity, resulting in distinct

protein and/or gene expression signatures. These findings strongly support our hypothesis that the aberrant field effect in the biopsied index tumors can be used to predict whether the patients are at risk of subsequent second primary malignancy development. It is proposed that genetic diversity can be utilized as a marker for field evolvability [8]. In a prospective pilot study by Roesch-Ely et al., by analyzing the protein profiles of mucosal biopsies from the oropharynx, hypopharynx, and three different regions of esophagus in head and neck squamous cell carcinoma patients and controls, tumor relapse was correctly predicted [17]. Taken together, these findings firmly established an altered field as a promising cancer risk prediction marker.

Of note, among the candidate proteins differentially expressed between P_{index} and P_{SPM} samples, upregulated proteins in P_{SPM} group were significantly enriched with proteins associated with intermediate filament, whereas those involved with extracellular matrix were downregulated. Overexpression of keratins, the intermediate filament-forming proteins of epithelial cells, is associated with enhanced tumor cell migration and invasion through interactions with extracellular environment [36]. Extracellular matrix proteins are associated with both tumor suppression and tumor promotion. Interestingly, the majority of downregulated extracellular matrix proteins in P_{SPM} samples including ECM1, EMILIN1, and FBLN2 have demonstrated tumor suppressive roles in various cancers [37–39]. Based on our findings, aberrant cytoskeletal activity and altered cell motility as a result of reduced expression of extracellular matrix-associated tumor suppressors are among the major molecular mechanisms underlying second primary malignancy development.

Recently, a large head and neck squamous cell carcinoma genomics study from The Cancer Genome Atlas (TCGA) consisting of mostly Caucasian patients revealed that 64% and 6% of oropharyngeal and non-oropharyngeal tumors, respectively, are HPV-positive [28]. In contrast, studies from Thailand reported that HPV status was positive in only 26.09% of oropharyngeal cancers and in none of the 80 non-oropharyngeal cancers tested [40, 41]. This fact is also reflected in our largely p16-negative cohorts. In addition, multiple studies have reported that patients with HPV-negative oropharyngeal squamous cell carcinoma have a higher risk of second primary malignancy development than HPV-positive oropharyngeal squamous cell carcinoma patients [42, 43]. Since most of our patients were p16-negative and HPV-negative head and neck squamous cell carcinoma patients are more likely to have second primary malignancy, we decided to focus on the p16-negative cases in cohort 2 for our gene expression analyses.

For cohort 2, in addition to the candidate genes from our proteomics study, seven genes from recent genomics literatures [27, 28] were included in the gene expression panel. *TP53* and *NOTCH1* are among the most commonly mutated genes in head and neck squamous cell carcinoma [28]. A recent whole exome sequencing study of head and neck squamous cell carcinomas has reported that *C17orf104* and *ITPR3* are specifically mutated in synchronous nodal metastases but not in the primary cancers, while *DDR2* is exclusively mutated in metachronous recurrent tumors [27]. Additionally, *FAT1* and *KMT2D* are found to be mutated in both primary head and neck squamous cell carcinomas and their nodal metastases [27]. Hence, it is of interest to investigate whether their expressions are associated with second primary malignancy risk. In this study, the combination of three genes *(ITPR3*, *KMT2D*, *EMILIN1)* and patient's age allowed for the most accurate discrimination between P_{index} and P_{SPM} groups with a sensitivity of 88.46% and a specificity of 75.00%. Of note, while these numbers will require validation in the future studies using larger sample size, our panel showed a comparable level of performance to that of the currently used clinical gene expression tests in breast cancer including Oncotype Dx (sensitivity 71–85%, specificity 55–66% for high and intermediate risk groups) and MammaPrint (sensitivity 83–92%, specificity 41–59%) [44].

In addition to its implications in head and neck squamous cell carcinoma [27], increased expression of ITPR3, a major intracellular Ca^{2+} release channel, is associated with enhanced tumor proliferation and invasion in breast and colorectal cancers [45, 46]. *KMT2D*, which encodes a histone H3 lysine 4 methyltransferase, is one of the most commonly mutated genes in cancers [47]. High KMT2D expression has recently been reported to promote tumor progression by inducing epithelial-mesenchymal transition (EMT), and is a predictor of poor prognosis in esophageal squamous cell carcinoma [48]. In contrast, *EMILIN1*, which encodes an extracellular matrix glycoprotein associated with the development of elastic tissues, plays a suppressive role in tumor growth, tumor lymphatic vessel formation, and metastatic spread to lymph nodes [38]. A recent report has suggested that α9β1 integrin, of which EMILIN1 is a ligand, is the major integrin involved in regulating head and neck squamous cell carcinoma cell migration on extracellular matrix [49]. In our study, increased *ITPR3* and *KMT2D* levels, as well as reduced *EMILIN1* are significantly associated with second primary malignancy in head and neck squamous cell carcinoma. Their exact roles in second primary malignancy tumorigenesis remain to be explored.

Second primary malignancy can be classified into either a synchronous or metachronous tumor. Since survival outcomes between synchronous and metachronous second primary malignancies are different [5, 9], it is clinically useful to be able to predict the time to second tumor development. The combination of two genes, *ITPR3* and

DSG3, was statistically significantly associated with the time to second primary malignancy occurrence in our study. *DSG3*, a component of cell–cell junctions, is overexpressed in head and neck squamous cell carcinomas, and its inhibition significantly suppresses tumor growth [50]. DSG3 has also been proposed as a predictive biomarker for cervical lymph node micrometastasis in oral cancer [51]. In line with their roles in promoting tumor proliferation and progression, high expression of both *ITPR3* and *DSG3* contributed to the shortest time to second primary malignancy development in patients with p16-negative head and neck squamous cell carcinoma.

It is generally known that transcript levels by themselves are by no means an accurate predicting factor for protein levels in many scenarios [29]. In this study, the majority of top candidate biomarkers remarkably showed a similar pattern of differential protein and mRNA expressions in the tumor samples across the P_{index} and P_{SPM} subgroups of both cohorts. This finding partly explains a smooth transition from protein to transcript level of this particular set of markers in regard to second primary malignancy risk prediction, and may also render future biomarker assay development the flexibility and feasibility to use either gene expression or targeted proteomic platforms.

According to the current follow-up recommendation by the National Comprehensive Cancer Network (NCCN) Guidelines for Head and Neck Cancers (version 2.2018) [52], history and physical exams including a complete head and neck exam; and mirror and fiberoptic examination should be performed at least every 3 months during the first year, and at least every 6 months in the second year. Based on our results, all patients with high *ITPR3* and *DSG3* levels developed second primary malignancy within ~13 months (400 days) after the diagnosis of an index tumor. Therefore, it may be beneficial to provide closer monitoring of second primary malignancy for this group of patients during their first years. Moreover, even though some synchronous second primary malignancies develop around the same time as the index tumor, they often cannot be clinically detected until their sizes are big enough. The use of our biomarkers should help increase early diagnosis of both subclinical synchronous second primary malignancies and subsequent metachronous second primary malignancies through a more rigorous surveillance protocol.

Limitations of this study include the relatively small sample size and limited follow-up time to second primary malignancy occurrence. Others include the use of a cohort of convenience where adequate formalin-fixed paraffin-embedded tissues was available, and the use of one tissue sample per tumor, which may not represent intratumor heterogeneity.

In summary, our biomarkers demonstrate great potential as a companion prognostic test for second primary malignancy risk prediction in routine clinical practice. The expected long-term benefit of early detection of second primary malignancy is an improvement in overall survival of head and neck cancer patients, especially those who are eligible for curative or less invasive therapy. The results are compelling and warrant future validation studies in larger head and neck squamous cell carcinoma cohorts. The protein signatures also hint at tumor-extracellular matrix interactions as a major player in second primary malignancy tumorigenesis. Further functional studies of these biomarkers may better clarify the clinical utility and their roles in second primary malignancy development.

Acknowledgements We would like to thank Mr. Jeffrey Makin, Dr. Premyot Ngaotepprutaram, Dr. Tanasarun Watcharadilokkul, and Dr. Meng-Shin Shiao for their valuable comments. S. Bunbanjerdsuk is a recipient of research assistant scholarship from Faculty of Medicine Ramathibodi Hospital and Faculty of Graduate Studies, Mahidol University. N. Jinawath is a recipient of TRF Research Scholar Fund (RSA5780065), the research grants from Mahidol University-National Research Council of Thailand (NRCT), CEMB-PERDO, and the Ramathibodi Comprehensive Cancer Center. The NJ laboratory is partially supported by the Crown Property Bureau Foundation through Integrative Computational BioScience (ICBS) Center, Mahidol University.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359–E86.
2. González-García R, Naval-Gías L, Román-Romero L, Sastre-Pérez J, Rodríguez-Campo FJ. Local recurrences and second primary tumors from squamous cell carcinoma of the oral cavity: a retrospective analytic study of 500 patients. Head Neck. 2009;31:1168–80.
3. Yamashita T, Araki K, Tomifuji M, Tanaka Y, Harada E, Suzuki T, et al. Clinical features and treatment outcomes of Japanese head and neck cancer patients with a second primary cancer. Asia Pac J Clin Oncol. 2017;13:172–8.
4. Lee DH, Roh J-L, Baek S, Jung JH, Choi S-H, Nam SY, et al. Second cancer incidence, risk factor, and specific mortality in head and neck squamous cell carcinoma. Otolaryngol Head Neck Surg. 2013;149:579–86.
5. Vaamonde P, Martin C, Labella T. Second primary malignancies in patients with cancer of the head and neck. Otolaryngol Head Neck Surg. 2003;129:65–70.
6. Slaughter DP, Southwick HW, Smejkal W. "Field cancerization" in oral stratified squamous epithelium. Clin Implic multicentric Orig Cancer. 1953;6:963–8.
7. Braakhuis BJ, Tabor MP, René Leemans C, van der Waal I, Snow GB, Brakenhoff RH. Second primary tumors and field

cancerization in oral and oropharyngeal cancer: molecular techniques provide new insights and definitions. Head Neck. 2002;24:198–206.
8. Curtius K, Wright NA, Graham TA. An evolutionary perspective on field cancerization. Nat Rev Cancer. 2018;18:19–32.
9. Di Martino E, Sellhaus B, Hausmann R, Minkenberg R, Lohmann M, Esthofen MW. Survival in second primary malignancies of patients with head and neck cancer. J Laryngol Otol. 2002;116:831–8.
10. Baxi SS, Pinheiro LC, Patil SM, Pfister DG, Oeffinger KC, Elkin EB. Causes of death in long-term survivors of head and neck cancer. Cancer. 2014;120:1507–13.
11. Clayburgh DR, Brickman D. Is esophagoscopy necessary during panendoscopy? Laryngoscope. 2017;127:2–3.
12. Haerle SK, Strobel K, Hany TF, Sidler D, Stoeckli SJ. 18F-FDG-PET/CT versus panendoscopy for the detection of synchronous second primary tumors in patients with head and neck squamous cell carcinoma. Head Neck. 2010;32:319–25.
13. Yabuki K, Kubota A, Horiuchi C, Taguchi T, Nishimura G, Inamori M. Limitations of PET and PET/CT in detecting upper gastrointestinal synchronous cancer in patients with head and neck carcinoma. Eur Arch Otorhinolaryngol. 2013;270:727–33.
14. Khuri FR, Lee JJ, Lippman SM, Kim ES, Cooper JS, Benner SE, et al. Randomized phase III trial of low-dose isotretinoin for prevention of second primary tumors in stage I and II head and neck cancer patients. J Natl Cancer Inst. 2006;98:441–50.
15. Kang H, Kiess A, Chung CH. Emerging biomarkers in head and neck cancer in the era of genomics. Nat Rev Clin Oncol. 2015;12:11–26.
16. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and neck cancers—major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67:122–37.
17. Roesch-Ely M, Leipold A, Nees M, Holzinger D, Dietz A, Flechtenmacher C, et al. Proteomic analysis of field cancerization in pharynx and oesophagus: a prospective pilot study. J Pathol. 2010;221:462–70.
18. Hildebrandt MA, Lippman SM, Etzel CJ, Kim E, Lee JJ, Khuri FR, et al. Genetic variants in the PI3K/PTEN/AKT/MTOR pathway predict head and neck cancer patient second primary tumor/recurrence risk and response to retinoid chemoprevention. Clin Cancer Res. 2012;18:3705–13.
19. Warren S, Gates O. Multiple primary malignant tumors: a survey of the literature and a statistical study. Am J Cancer. 1932;16:1358–414.
20. Hong WK, Lippman SM, Itri LM, Karp DD, Lee JS, Byers RM, et al. Prevention of second primary tumors with isotretinoin in squamous-cell carcinoma of the head and neck. N Engl J Med. 1990;323:795–801.
21. Sunpaweravong S, Sunpaweravong P, Dechaphunkul T, Pongrujikorn T, Bunbanjerdsuk S, Jinawath N. Clinico-molecular study of synchronous head and neck squamous cell carcinoma (HNSCC) and esophageal squamous cell carcinoma (ESCC). J Clin Oncol. 2015;33:e17080.
22. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009;6:359–62.
23. Rappsilber J, Ishihama Y, Mann M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem. 2003;75:663–70.
24. Wiśniewski JR, Zougman A, Mann M. Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res. 2009;8:5674–8.
25. Skeie JM, Mahajan VB. Proteomic interactions in the mouse vitreous-retina complex. PLoS ONE. 2013;8:e82140.
26. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1.
27. Hedberg ML, Goh G, Chiosea SI, Bauman JE, Freilino ML, Zeng Y, et al. Genetic landscape of metastatic and recurrent head and neck squamous cell carcinoma. J Clin Invest. 2016;126:169–80.
28. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.
29. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell . 2016;165:535–50.
30. Crutchfield CA, Thomas SN, Sokoll LJ, Chan DW. Advances in mass spectrometry-based clinical biomarker discovery. Clin Proteomics. 2016;13:1.
31. Lin F, Chen Z. Standardization of diagnostic immunohistochemistry: literature review and geisinger experience. Arch Pathol Lab Med. 2014;138:1564–77.
32. Meng Z, Veenstra TD. Targeted mass spectrometry approaches for protein biomarker verification. J Proteom. 2011;74:2650–9.
33. Győrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res. 2015;17:11.
34. Veldman-Jones MH, Brant R, Rooney C, Geh C, Emery H, Harbron CG, et al. Evaluating robustness and sensitivity of the NanoString technologies nCounter platform to enable multiplexed gene expression analysis of clinical samples. Cancer Res. 2015;75:2587–93.
35. Roesch-Ely M, Nees M, Karsai S, Ruess A, Bogumil R, Warnken U, et al. Proteomic analysis reveals successive aberrations in protein expression from healthy mucosa to invasive head and neck cancer. Oncogene. 2007;26:54–64.
36. Karantza V. Keratins in health and cancer: more than mere epithelial cell markers. Oncogene. 2011;30:127–38.
37. Gao F, Xia Y, Wang J, Lin Z, Ou Y, Liu X, et al. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. Genome Biol. 2014;15:533.
38. Danussi C, Petrucco A, Wassermann B, Modica TME, Pivetta E, Belluz LDB, et al. An EMILIN1-negative microenvironment promotes tumor cell proliferation and lymph node invasion. Cancer Prev Res. 2012;5:1131–43.
39. Law E, Cheung A, Kashuba V, Pavlova T, Zabarovsky E, Lung H, et al. Anti-angiogenic and tumor-suppressive roles of candidate tumor-suppressor gene, Fibulin-2, in nasopharyngeal carcinoma. Oncogene. 2012;31:728–38.
40. Pongsapich W, Jotikaprasardhna P, Lianbanchong C, Siritantikorn S, Chongkolwatana C. Human Papillomavirus Infection in Oral Cavity and Oropharyngeal Cancers: Are They the Same Story? J Med Assoc Thai. 2016;99:684–9.
41. Pongsapich W, Eakkasem N, Siritantikorn S, Pithuksurachai P, Bongsabhikul K, Chongkolwatana C. Prevalence of HPV infection in hypopharyngeal and laryngeal squamous cell carcinoma at Thailand's largest tertiary referral center. Infect Agent Cancer. 2017;12:58.
42. Martel M, Alemany L, Taberna M, Mena M, Tous S, Bagué S, et al. The role of HPV on the risk of second primary neoplasia in patients with oropharyngeal carcinoma. Oral Oncol. 2017;64:37–43.
43. Gan SJ, Dahlstrom KR, Peck BW, Caywood W, Li G, Wei Q, et al. Incidence and pattern of second primary malignancies in patients with index oropharyngeal cancers versus index non-oropharyngeal head and neck cancers. Cancer. 2013;119:2593–601.
44. Berg AO, Armstrong K, Botkin J, Calonge N, Haddow J, Hayes M, et al. Recommendations from the EGAPP Working Group: can

tumor gene expression profiling improve outcomes in patients with breast cancer? Genet Med. 2009;11:66–73.
45. Mound A, Rodat-Despoix L, Bougarn S, Ouadid-Ahidouch H, Matifat F. Molecular interaction and functional coupling between type 3 inositol 1, 4, 5-trisphosphate receptor and BK Ca channel stimulate breast cancer cell proliferation. Eur J Cancer. 2013;49:3738–51.
46. Shibao K, Fiedler MJ, Nagata J, Minagawa N, Hirata K, Nakayama Y, et al. The type III inositol 1, 4, 5-trisphosphate receptor is associated with aggressiveness of colorectal carcinoma. Cell Calcium. 2010;48:315–23.
47. Rao RC, Dou Y. Hijacked in cancer: the MLL/KMT2 family of methyltransferases. Nat Rev Cancer. 2015;15:334–46.
48. Abudureheman A, Ainiwaer J, Hou Z, Niyaz M, Turghun A, Hasim A, et al. High MLL2 expression predicts poor prognosis and promotes tumor progression by inducing EMT in esophageal squamous cell carcinoma. J Cancer Res Clin Oncol. 2018;144:1025–35.
49. Gopal S, Veracini L, Grall D, Butori C, Schaub S, Audebert S, et al. Fibronectin-guided migration of carcinoma collectives. Nat Commun. 2017;8:14105.
50. Chen Y, Chang J, Lee L, Wang H, Liao C, Chiu C, et al. DSG3 is overexpressed in head neck cancer and is a potential molecular target for inhibition of oncogenesis. Oncogene. 2007;26:467–76.
51. Patel V, Martin D, Malhotra R, Marsh C, Doçi C, Veenstra T, et al. DSG3 as a biomarker for the ultrasensitive detection of occult lymph node metastasis in oral cancer using nanostructured immunoarrays. Oral Oncol. 2013;49:93–101.
52. National Comprehensive Cancer Network. Head and neck cancers (Version 2.2018 - June 20, 2018) 2018. https://www.nccn.org/professionals/physician_gls/PDF/head-and-neck.pdf.

Affiliations

Sacarin Bunbanjerdsuk[1,2] · Nutchavadee Vorasan[3] · Thammakorn Saethang[4] · Tanjitti Pongrujikorn[1] · Duangjai Pangpunyakulchai[5] · Narongsak Mongkonsiri[5] · Lalida Arsa[6] · Nintita Thokanit[7] · Warut Pongsapich[8] · Tauangtham Anekpuritanang[9] · Nuttapong Ngamphaiboon[10] · Artit Jinawath[6] · Somkiat Sunpaweravong[11] · Trairak Pisitkun[4] · Bhoom Suktitipat[3,12,13] · Natini Jinawath[1,7,13]

1 Program in Translational Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

2 Medical Genetics Center, Medical Life Sciences Institute, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand

3 Siriraj Center of Research Excellence in Bioinformatics and Clinical Data Management, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

4 Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

5 Immunohistopathology and Special Laboratory, Department of Pathology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

6 Molecular Histopathology Laboratory, Department of Pathology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

7 Ramathibodi Comprehensive Cancer Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

8 Department of Otorhinolaryngology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

9 Department of Pathology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

10 Medical Oncology Unit, Department of Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

11 Department of Surgery, Faculty of Medicine, Prince of Songkla University, Songkhla, Thailand

12 Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

13 Integrative Computational BioScience Center (ICBS), Mahidol University, Nakhon Pathom, Thailand

frontiers in Genetics

REVIEW
published: 15 August 2018
doi: 10.3389/fgene.2018.00309

Emergence of Intrahepatic Cholangiocarcinoma: How High-Throughput Technologies Expedite the Solutions for a Rare Cancer Type

Meng-Shin Shiao[1†], *Khajeelak Chiablaem*[2†], *Varodom Charoensawan*[3,4,5], *Nuttapong Ngamphaiboon*[6] *and Natini Jinawath*[2,4*]

[1] Research Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand, [2] Program in Translational Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand, [3] Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok, Thailand, [4] Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom, Thailand, [5] Systems Biology of Diseases Research Unit, Faculty of Science, Mahidol University, Bangkok, Thailand, [6] Medical Oncology Unit, Department of Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

OPEN ACCESS

Edited by:
Arvin Gouw,
Rare Genomics Institute,
United States

Reviewed by:
Theodora Katsila,
University of Patras, Greece
Ramu Elango,
Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Saudi Arabia

***Correspondence:**
Natini Jinawath
jnatini@hotmail.com;
natini.jin@mahidol.ac.th

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Genetic Disorders, a section of the journal Frontiers in Genetics

Received: *09 October 2017*
Accepted: *23 July 2018*
Published: *15 August 2018*

Citation:
Shiao M-S, Chiablaem K, Charoensawan V, Ngamphaiboon N and Jinawath N (2018) Emergence of Intrahepatic Cholangiocarcinoma: How High-Throughput Technologies Expedite the Solutions for a Rare Cancer Type. Front. Genet. 9:309. doi: 10.3389/fgene.2018.00309

Intrahepatic cholangiocarcinoma (ICC) is the cancer of the intrahepatic bile ducts, and together with hepatocellular carcinoma (HCC), constitute the majority of primary liver cancers. ICC is a rare disorder as its overall incidence is <1/100,000 in the United States and Europe. However, it shows much higher incidence in particular geographical regions, such as northeastern Thailand, where liver fluke infection is the most common risk factor of ICC. Since the early stages of ICC are often asymptomatic, the patients are usually diagnosed at advanced stages with no effective treatments available, leading to the high mortality rate. In addition, unclear genetic mechanisms, heterogeneous nature, and various etiologies complicate the development of new efficient treatments. Recently, a number of studies have employed high-throughput approaches, including next-generation sequencing and mass spectrometry, in order to understand ICC in different biological aspects. In general, the majority of recurrent genetic alterations identified in ICC are enriched in known tumor suppressor genes and oncogenes, such as mutations in *TP53*, *KRAS*, *BAP1, ARID1A*, *IDH1*, *IDH2*, and novel *FGFR2* fusion genes. Yet, there are no major driver genes with immediate clinical solutions characterized. Interestingly, recent studies utilized multi-omics data to classify ICC into two main subgroups, one with immune response genes as the main driving factor, while another is enriched with driver mutations in the genes associated with epigenetic regulations, such as *IDH1* and *IDH2*. The two subgroups also show different hypermethylation patterns in the promoter regions. Additionally, the immune response induced by host-pathogen interactions, i.e., liver fluke infection, may further stimulate tumor growth through alterations of the tumor microenvironment. For in-depth functional studies, although many ICC cell lines have been globally established, these homogeneous cell lines may not fully explain the highly heterogeneous genetic contents of this disorder. Therefore, the advent of patient-derived

xenograft and 3D patient-derived organoids as new disease models together with the understanding of evolution and genetic alterations of tumor cells at the single-cell resolution will likely become the main focus to fill the current translational research gaps of ICC in the future.

Keywords: intrahepatic cholangiocarcinoma, high-throughput technology, integrative multi-omics analysis, molecular biomarker, disease model, translational medicine, precision oncology

BACKGROUND

The biliary system includes bile ducts and gallbladder. The main functions of bile ducts are to transfer bile from the liver and gallbladder to the small intestine to help with the digestion and absorption of dietary fats. Bile ducts can be classified into several parts based on the anatomical locations and structures. Peripheral branches of intrahepatic bile ducts drain into the right and left hepatic ducts, which then merge into a larger tube outside the liver, called the common hepatic duct. This extrahepatic bile duct further combines with the cystic duct from the gallbladder and becomes the common bile duct. Cholangiocarcinoma (CCA) is a group of heterogeneous malignancies that occurs in any part of the bile ducts. It can be further classified into three different categories based on the anatomical positions. The tumors that occur in the intrahepatic bile ducts are termed intrahepatic cholangiocarcinoma (ICC), while those located between the secondary branches of the right and left hepatic ducts and the common hepatic duct proximal to the cystic duct origin, and in the common bile duct are classified as perihilar and distal cholangiocarcinomas, respectively (Blechacz, 2017; **Figure 1**). As ICC occurs inside the liver, it is also one of the two main types of primary liver cancers besides hepatocellular carcinoma (HCC). In this review, we aim to provide a comprehensive update and novel insights on ICC, the rare type of CCA, which is known for its extraordinary complexity and heterogeneity, along with dismal prognosis.

Based on a 31-year study in the United States, ICC accounts for only 8% of all CCA cases, and is considered to be a rare disorder (DeOliveira et al., 2007). ICC occurs with the highest prevalence in Hispanic Americans (1.22 per 100,000 people) and lowest in African Americans (0.3 per 100,000 people) (McLean and Patel, 2006). By contrast, it is more common in East Asian and Southeast Asian countries. ICC has an incidence of around 10 per 100,000 people in China (males), and the highest frequency of occurrence, 71 per 100,000 people (males), is found in the northeastern part of Thailand (Shin et al., 2010a). Interestingly, the global incidence of ICC seems to have increased in recent years (Khan et al., 2012).

Risk factors of ICC include bile duct cysts, chronic biliary irritation, parasitic or viral infections, inflammatory bowel disease, abnormal bile ducts, and exposure of chemical carcinogens. Chronic inflammation caused by parasitic infection, particularly liver flukes (*Opisthorchis viverrini* and *Clonorchis sinensis*), is a well-known risk factor of ICC in northeastern Thailand (Sripa et al., 2007; Sripa and Pairojkul, 2008). Eating raw or uncooked fermented fish, a common local dish in this area, results in the high incidence of recurrent liver fluke infections, which are strongly associated with ICC. Several mechanisms have been proposed to explain the association between liver fluke infection and ICC (Sripa et al., 2007). First, when liver flukes start their parasitic life in humans, they attach themselves to the bile duct epithelia using their suckers, which cause damage to the epithelial walls of the ducts. The repeated damage-repair processes may result in the epithelial-mesenchymal transition (EMT) of cell states. Second, the inflammation reactions induced by parasites and the chemicals secreted by them, as well as mutagens from fermented food, may create more carcinogens that damage DNA and result in irreversible oncogenic mutations.

Other than parasitic infections, hepatitis B virus (HBV) and hepatitis C virus (HCV) infections are also associated with ICC. HBV and HCV nucleic acids have been found in 27% of ICC tumors in a US-based study (Perumal et al., 2006). Another study in China has shown a strong association between chronic HBV infection and ICC in a total of 317 patients, and further suggested that ICC and HCC may share a common carcinogenesis process (Zhou et al., 2010). In addition, HBV and HCV infections are proposed to be associated with increasing incidence of ICC from several case-control studies (Yamamoto et al., 2004; Fwu et al., 2011; Sempoux et al., 2011; Yu et al., 2011; Zhou et al., 2012; Li et al., 2015). Other possible risk factors of ICC include smoking, alcohol drinking, obesity and diabetes mellitus, which are mostly observed in western countries (Tyson and El-Serag, 2011). A detailed summary of established risk factors for ICC and their relative risks are shown in **Table 1**.

The most fundamental categorization of ICC is based on the macroscopic features established by the Liver Cancer Study Group of Japan in 2003 (Yamasaki, 2003). The authors described three macroscopic subtypes of ICC, namely, mass-forming type (MF), periductal-infiltrating type (PDI) and intraductal growth (IDG) type. MF type forms a definite mass in the liver parenchyma. PDI type is defined as tumors that extend longitudinally along the ducts, while the IDG type forms a papillary growth inside the lumen of intrahepatic ducts. MF subtype is the most common subtype (about 65%), whereas PDI and IDG types are less prevalent (around 5% each), and mixed-type (MF+PDI) accounts for ∼25% of the cases (Yamasaki, 2003; Sempoux et al., 2011). However, based on more recent data, a noteworthy degree of heterogeneity of ICCs in regard to their histopathological and molecular features was observed. Therefore, in addition to the traditional classifications, multiple new criteria were proposed in order to subcategorize ICC (Vijgen et al., 2017).

Serum biomarkers are usually used to help screen cancer at its earliest stages. A wide variety of markers have been tested in bile and serum with limited success. To date, disease-specific

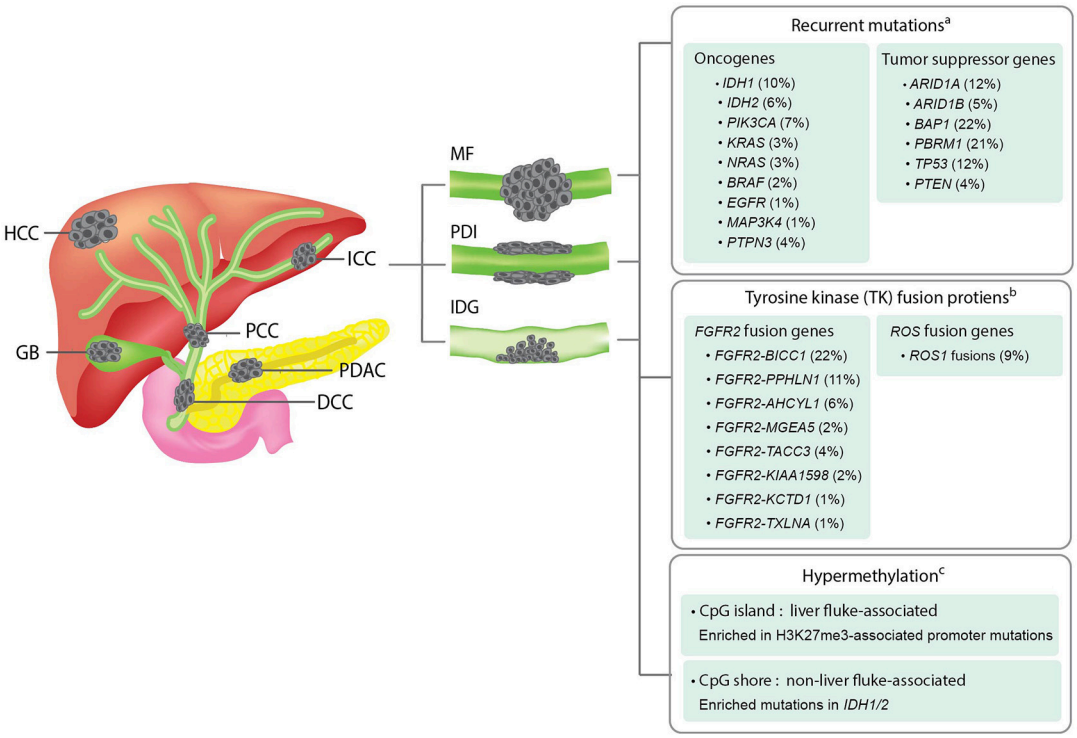


FIGURE 1 | Overview of the anatomical structures, macroscopic subtypes, and recurrent genetic alterations in ICCs. **Left panel;** an illustration showing the anatomical structures of biliary system and their associated malignancies. HCC, hepatocellular carcinoma; ICC, intrahepatic cholangiocarcinoma; PCC, perihilar cholangiocarcinoma; GB, gallbladder cancer; DCC, distal cholangiocarcinoma; PDAC, pancreatic ductal adenocarcinoma. **Middle panel;** an illustration showing the three macroscopic subtypes of ICC. MF, mass-forming type; PDI, periductal-infiltrating type; IDG, intraductal growth type. **Right panel;** a summary of recurrent genetic alterations and their reported frequencies in ICCs. [a]The mutation frequency of each gene is calculated by dividing the combined number of ICC cases presenting the mutation with the total number of ICC cases analyzed in all four cohorts included in the cBioPortal for Cancer Genomics database (www.cbioportal.org). [b]The frequency of each fusion gene were obtained from previous literatures (Nakamura et al., 2015; Moeini et al., 2016). [c]Different hypermethylation patterns of liver fluke-associated and non-liver fluke-associated ICCs and their associated alterations were summarized based on a previous study (Jusakul et al., 2017).

biomarkers for CCA have yet to be established (Valle et al., 2016) and are urgently needed. The most frequently used biomarker for diagnostic and treatment prediction in CCA patients in clinical practice is carbohydrate antigen 19–9 (CA 19-9) (Liang et al., 2015), which is the standard tumor marker for pancreatic adenocarcinoma (Ballehaninna and Chamberlain, 2012). Nevertheless, serum levels of CA 19-9 are also elevated in benign cholestasis such as primary sclerosing cholangitis (PSC), complicating its usage in clinic (Lin et al., 2014). A serum CA 19-9 level >100 U/mL has quite limited sensitivity and specificity (75 and 80%, respectively) in identifying PSC patients with CCA (Chalasani et al., 2000). In ICC, a large cohort analysis by Bergquist et al reported an elevated CA 19-9 level as an independent risk factor for mortality. Elevation of CA 19-9 independently predicted increased mortality with impact similar to node-positivity, positive-margin resection, and non-receipt of chemotherapy (Bergquist et al., 2016).

Since the clinical presentation of ICC is not specific and the disease in its early stage is usually asymptomatic, the patients are often diagnosed at an advanced stage. Surgical resection, which is the only curative treatment, remains the anchor of therapy for patients with resectable ICC (Weber S. M. et al., 2015). Nevertheless, because of the late presentation of symptoms and the central hepatic location of ICC, only ∼30% of the patients are deemed eligible for resection by the time of diagnosis. This results in a low 5-year survival and high recurrent rate after resections (Hyder et al., 2013). Loco-regional therapies (LRT) including intra-arterial embolotherapy (IAT) and radiofrequency ablation have been reported as the feasible and effective palliative treatments for patients with unresectable ICC (Savic et al., 2017). Overall, systemic cytotoxic chemotherapy is still the mainstay of treatment for patients with advanced unresectable, recurrent or metastatic ICC. In a landmark phase III randomized study in patients with advanced biliary tract cancer (BTC), doublet chemotherapy (addition of cisplatin to gemcitabine) improved the response rate from 72 to 81% ($P = 0.049$) and overall survival from 8.1 to 11.7 months (hazard ratio 0.64; $P < 0.001$) (Valle et al., 2010). Thus, it has since been considered as the standard of care although the efficacy remains limited. Of note, CCA only accounts for ∼60% of all BTC patients enrolled in this study. Another well-established combination chemotherapy regimen for advanced BTC is GEMOX, which consists of gembitabine plus oxaliplatin (Sharma et al., 2010). So far, several clinical trials investigating the efficacy of targeted therapies, such as cetuximab, panitumumab, erlotinib, selumetinib, sunitinib, and bevacizumab, have failed to demonstrate the survival benefits for

TABLE 1 | Established risk factors of cholangiocarcinoma.

Risk factors	Relative risk (95% CI)	References
Liver Flukes		
Opisthorchis viverrini (OV)[a]	4.8 (2.8–8.4)	Shin et al., 2010b
Clonorchis sinensis (CS)[b]		
Viral Hepatitis		
Hepatitis C virus (HCV)	1.8–4.84	Shin et al., 2010b; Palmer and Patel, 2012
Hepatitis B virus (HBV)	2.6–5.1	
Cirrhosis	5.03–27.2	Tyson and El-Serag, 2011; Palmer and Patel, 2012
Primary Sclerosing Cholangitis (PSC)	Lifetime risk 5–35%	Tyson and El-Serag, 2011
Inflammatory bowel disease (IBD)	1.7–4.67	Tyson and El-Serag, 2011
Obesity	1.56–1.60	Jing et al., 2012; Palmer and Patel, 2012
Type II diabetes	1.43–1.89	Ren et al., 2011; Palmer and Patel, 2012
Hepatolithiasis	5.8–50.0	Tyson and El-Serag, 2011
Congenital abnormalities in biliary tract	10.7–47.1	Tyson and El-Serag, 2011
Alcohol	2.81 (1.52–5.21)	Palmer and Patel, 2012
Genetic polymorphisms[c]	0.23–5.38	Tyson and El-Serag, 2011

[a] *Endemic in Northeastern Thailand, Lao, Vietnam, Cambodia.*
[b] *Endemic in South China, Japan, Korea, Taiwan.*
[c] *HFR 677CC+TSER 2R; GSTO1*A140D; MRP2/ABCC2 variant c.3972C>T; (NKG2D rs11053781, rs2617167) +PSC; MICA5.1+PSC; CYP1A2*1A/*1A; NAT2*13,*6B,*7A; XRCCI194W; XRCC1 R280H; PYGS2 Ex10+837 (Tyson and El-Serag, 2011).*

Box 1 | Challenges in ICC treatment.

- Intrahepatic cholangiocarcinoma (ICC), a subtype of biliary tract cancer, is considered as a rare disorder with an overall incidence of 1-2 cases per 100,000 people in the US and Europe. However, ICC exhibits vastly different incidence in different parts of the world, mainly based on exposure to the specific risk factors that are common in the regions such as the Southeast Asian liver flukes. The incidence of ICC is currently increasing worldwide.
- Early stages of ICC are usually asymptomatic. The patients are usually diagnosed at advanced stages and metastases are frequently observed. Additionally, a high recurrent rate after tumor resection, which is the sole curative treatment, is also common. The 5-year survival rate for localized disease is only ∼15% (American Cancer Society, Inc., 2018).
- The existing serum tumor markers, namely carbohydrate antigen 19–9 (CA19-9) and carcinoembryonic antigen (CEA), lack sensitivity and specificity to detect ICC at an early stage. To date, efficient strategies for the screening and surveillance of ICC have not been established.
- Chemotherapy is a standard of care for advanced disease; however, the efficacy remains limited. Several targeted therapies and their predictive biomarkers have failed to demonstrate survival benefits for this group of patients. Immunotherapy such as checkpoint inhibitor may be effective only in patients with microsatellite instability (MSI), which is uncommon in ICC.
- The highly heterogeneous nature of ICC, comprising both locally advanced and metastatic disease, along with the lack of common genetic alterations and clinically actionable molecular classifications, make it difficult to design the effective clinical trials and assess the efficacy of each treatment regimen. Multiple studies focusing on integrative multi-omics analyses have recently been conducted to identify the molecular classifications of ICC that can help optimize clinical decision.

this group of patients (Zhu et al., 2010; Bekaii-Saab et al., 2011; Jensen et al., 2012; Lee et al., 2012; Yi et al., 2012; Malka et al., 2014).

Taken together, even though ICC is considered a rare cancer type, it represents an emerging health problem with increasing incidence worldwide. ICC is usually diagnosed at late stages and has poor prognosis, partly due to the complex anatomical structure of the biliary system, its various etiologies, heterogeneous subclassifications, and the lack of effective biomarkers and treatments. To date, the genetic signatures of ICC are still limitedly understood and no major driver mutations with clinical actionability have been identified. An overview of current challenges in the treatment of ICC is outlined in **Box 1**. In the next sections, we aim to provide an in-depth update on the application of recent advances in high-throughput technologies that can help expedite the translation of research discoveries in ICC and related cancers, as well as current disease models used to facilitate the development of precision oncology in ICC.

MOLECULAR FEATURES AND SUBTYPES OF ICC IDENTIFIED BY HIGH-THROUGHPUT APPROACHES

Advances in high-throughput screening methods such as next-generation sequencing (NGS) and liquid chromatography–mass spectrometry (LC-MS) have enabled broader interrogation of genetic diseases and other disorders. The so-called "omics" data can be defined and categorized according to different groups of biological molecules and regulatory processes, which provide different information of the cells. Given the advantages of broader and deeper scales of available data, different types of omics are applied widely and rapidly to study the associations between different variations and phenotypes, and also used to predict prognosis. It also helps in the classification of subtypes of a disease, which may require different treatment guidelines (Kristensen et al., 2014).

Genomics is one of the earliest to be introduced among the omics data series. Common types of somatic DNA alterations including single nucleotide variants (SNVs), insertions and deletions (INDELs), copy number alterations (CNAs), and structural variations (SVs) have all been shown to play important roles in development and progression of ICC (Zou et al., 2014). Comparative genomics of cancer and normal cells serve as an important platform to investigate molecular mechanisms of cancers; however, biological functions of oncogenes largely depend on how they are expressed (or not expressed) into functional oncoproteins and which tissues they are expressed in. Transcriptomics describes the abundance of transcribed messenger RNA (mRNA) and other non-coding RNAs. Even though most transcriptomic studies on ICC and relating cancers have been focused on mRNA (Jinawath et al., 2006), dysfunction of non-coding RNAs, particularly microRNA (miRNA) and

long non-coding (lncRNA), have recently been found to play roles in ICC as well (Wang et al., 2016; Yang et al., 2017; Zheng et al., 2017). Other than transcriptional level, transcriptomic profiling by RNA-Seq data also provides novel information on alternative splicing isoforms of a gene and confirms the expression of novel fusion gene transcripts, which is surprisingly prevalent in ICC (Arai et al., 2014; Borad et al., 2014; Ross et al., 2014; Nakamura et al., 2015; Sia et al., 2015). Transcriptional levels significantly depend on epigenetic configuration of regulatory elements targeting the oncogenes and tumor suppressor genes. It has been shown in CCA, including ICC, that DNA methylation is markedly enriched in either CpG islands or shores, which are regulatory regions enriched in cytosine and guanine nucleotides (Jusakul et al., 2017). Downstream to transcriptomes, proteomics has been widely used to quantify peptide sequences, post-translational modifications, protein abundance and interactions. Aberrant proteins secreted by cancer cells and released into various kinds of body fluids, such as blood, urine and saliva, provide good non-invasive biomarkers for early detection of cancer and the recurrent disease. A few studies have proposed potential biomarkers for CCA and HCC based on mass spectrometry analysis of cancer-specific secreted proteins (Srisomsap et al., 2010; Cao et al., 2013). Another high-throughput approach, metabolomics study, quantifies small molecules, such as amino acids, fatty acids, carbohydrates, or other compounds related to cellular metabolic functions. Metabolite levels and relative ratios reflect metabolic function, and out of normal range perturbations are often indicative of disease, as also shown in ICC (Murakami et al., 2015).

One of the most apparent applications of omic techniques on cancer research is the characterization of cancer subtypes and their signatures, which frequently leads to personalized treatments for cancer patients bearing different tumor signatures. For instance, based on a large whole exome (WES) and genome sequencing (WGS) dataset of 7,042 tumors generated from 30 primary cancer types, cancers could be categorized into 21 different molecular signatures (Alexandrov et al., 2013). Molecular signature 1, for example, has the highest prevalence in all the cancer samples (∼70%), and is mostly associated with age. Signature 3 accounts for about 10% of the prevalence and is associated with mutations in *BRCA1/2*. Therefore, combining signature 1 and 3 explains over 80% of the breast cancer cases. Even though within each cancer type, the prevalence of somatic mutations varies significantly, they can be distinguished using different combinations of signatures. In parallel, another study categorized 3,299 tumors from The Cancer Genome Atlas (TCGA) comprising 12 cancer types into two main classes, one with dominant oncogenic signatures of somatic mutations (M class), and the others with dominant signatures of CNAs (C class) (Ciriello et al., 2013). The M class tumors show primarily genomic mutations and epigenetic alterations, such as DNA hypermethylation. Conversely, the C class tumors show primarily CNAs, particularly high-level of amplifications and homozygous deletions. Targetable molecular alterations in a tumor class allow the use of class-specific combination cancer therapy. More recently, an integrated analysis of genetic alterations focusing on the 10 canonical signaling pathways in the 9,125 TCGA-profiled tumors from 33 cancer types including CCA has underlined significant representation of individual and co-occurring actionable alterations among these pathways, which suggests targeted and combination therapy opportunities (Sanchez-Vega et al., 2018). In addition, WES and transcriptome data were applied to identify molecular signatures of metastatic solid tumors from 500 adult patients (Robinson et al., 2017). Altogether, such systematic approaches can potentially be applied specifically to ICCs, where each tumor may carry different underlying genetic mechanisms and prognoses, in order to obtain more effective treatment for individual patients.

To overcome the challenges in ICC diagnosis and treatment (**Box 1**), multiple high-throughput omics studies have been performed in order to discover the underlying molecular mechanisms that can be translated into precision oncology application. In order to better understand the current progress in ICC translational research, here we review the various subclassifications of ICC with regard to its cells of origin, different etiologies and unique clinicomolecular aspects of this rare disorder. The detailed summary of the high-throughput omics studies of ICC can be seen in **Table 2**.

Cells of Origin of ICC

Primary liver cancer, which is the second leading cause of cancer-related death worldwide, is mainly composed of ICC and HCC. The molecular and clinical features of the two cancers are distinct in most cases. Many studies have shown that the two cancers may share the same driver genes, which may be due to the fact that they also share the same cells of origin; hepatocytes and cholangiocytes arise from a common progenitor, hepatoblasts. ICC usually has poorer prognosis than HCC due to the difficulties in early disease detection and poorly understood carcinogenesis mechanisms. In a small proportion of the cases, ranging from 0.4 to 14% depending on the geographical regions, the patients developed combined hepatocellular cholangiocarcinoma (CHC) (Theise et al., 2010), which was proposed to be of monoclonal origin based on a recent study (Wang et al., 2018).

Various genetically engineered mouse models have been generated to study the cellular origin of primary liver cancers; however, the results are still inconclusive. By ablation of genes in Hippo signaling pathways (Lee et al., 2010; Lu et al., 2010) or knocking out neurofibromatosis type 2 (*Nf2*) gene (Benhamouche et al., 2010) in mouse, the authors proposed that ICC and HCC may share the same progenitor cells since all surviving mice eventually developed both CCA and HCC. A similar result was achieved by performing transduction of oncogenes, i.e., *H-Ras* or *SV40LT*, in mouse primary hepatic progenitor cells, lineage-committed hepatoblasts, and differentiated adult hepatocytes. Regardless of the hepatic lineage hierarchy, transduced cells were able to give rise to a continuous spectrum of liver cancers from HCC to CCA suggesting that any hepatic lineage cell can be cell-of-origin of primary liver cancer (Holczbauer et al., 2013). Several large multi-omics studies have shown that ICC and HCC share recurrently mutated genes including *TP53*, *BAP1*, *ARID1A, ARID2* (Chaisaingmongkol et al., 2017; Farshidfar et al., 2017; Wang et al., 2018). Furthermore, ICC together with HCC can be categorized into

TABLE 2 | Subclassification of ICCs and their associated genetic alterations.

Classification	Technology	Subtype I		Subtype II		References
		Class	**Characters**	**Class**	**Characters**	
Cells of origin	RNA-Seq WES Proteomics	C1 class	- Mutations in *TP53*, *BAP1*, *ARID1A, ARID2* - Altered expression of *PLK1* and *ECT2*	C2 class	- Obesity - Bile acid metabolism - T-cell infiltration	Chaisaingmongkol et al., 2017
Anatomical structure	WGS WES RNA-Seq	ICC-specific	- *FGFR2* fusion - *IDH1/2*, *EPHA2*, *BAP1* Mutation	ICC and ECC shared	- *KRAS, SMAD4, ARID1A* and *GNAS* mutation	Nakamura et al., 2015
Liver-fluke infection	Microarray	Liver fluke positive	- Xenobiotic metabolism	Liver fluke negative	- Growth factor signaling	Jinawath et al., 2006
	WGS WES WGS		- *TP53*, *KRAS*, *SMAD4*, *MLL3*, *ROBO2*, *RNF43*, *PEG3* and *GNAS* oncogene		- *BAP1, IDH1/2*	Ong et al., 2012; Chan-On et al., 2013
	Epigenomics		- High somatic mutations - *TP53, ARID1A* and *BRCA1/2* mutations - *ERBB2* amplification - Hypermethylation in promoter CpG islands - H3K27me3-associated promoter mutations - Poorer prognosis		- High copy-number alterations - *BAP1* and *IDH1/2* mutations - Altered *PD-1/PD-L2* expression - Alterations and elevated expression of *FGFR* genes - Hypermethylation in promotor CpG shore - Better prognosis	Jusakul et al., 2017
Gene expression and copy number alterations	- Microarray - SNP array	Proliferation class	- Oncogenic pathways - Mutations in *KRAS, BRAF* and *EGFR* - Chr11q13.2 amplification - Chr14q22.1 deletion - Moderate/poorly differentiated - Poorer prognosis	Inflammation class	- Inflammatory pathways (Interleukins/chemokines), - *STAT3* activation signaling pathway - Well differentiated - Better prognosis	Sia et al., 2013a,b Sia et al., 2017
Prognosis	- Microarray	Poor prognosis	- Mutations in *KRAS* and *BRAF*	Good prognosis	- No *KRAS* mutation	Andersen et al., 2013
Mutations and copy number alterations	- WGS - RNA-seq	M class	- Recurrent mutations of *KRAS, TP53, IDH1*	C class	- Recurrent focal copy number alterations including deletions involving *CDKN2A, ROBO1/2, RUNX3* and *SMAD4*	Kim et al., 2016

C1 and C2 subtypes. ICC-C1 and HCC-C1 share similar transcriptomic patterns that are significantly different from those of ICC-C2 and HCC-C2. Interestingly, ICC-C1 and HCC-C1 are enriched for aberrant mitotic checkpoint signaling, suggesting a high rate of chromosomal instability, while C2 groups are enriched for the cell immunity-related pathways, which implies an association with inflammatory responses (Chaisaingmongkol et al., 2017). These findings indicate that ICC and HCC, while clinically treated as separate entities, share common molecular subtypes with similar actionable drivers that can be exploited to improve precision therapy.

It should be noted that ICC- or HCC-specific alterations also exist. Aberrant activation of NOTCH signaling and gain-of-function mutations in the genes encoding isocitrate dehydrogenases (*IDH1* and *IDH2*) are required for ICC development, and thus are significantly more common in ICC than in HCC (Sekiya and Suzuki, 2012; Moeini et al., 2016). In addition, activation of *KRAS* and deletion of *PTEN* in the mouse model will only generate ICC (Ikenoue et al., 2016). Multiple studies have identified different molecular features of ICC and HCC by applying large-scale high-throughput datasets. By combining metabolomics and transcriptomics data from 10 ICC and six HCC samples together with their paired normal tissues, a research team showed that ICC can be distinguished from HCC by the distinct expression patterns of 62 mRNAs, 17 miRNAs, and 14 metabolites (Murakami et al., 2015), leading to the conclusion that ICC and HCC have different oncogenic mechanisms. Recently, Farshidfar et al. conducted a meta-analysis study by combining sequencing data from a total of 458 ICC, 153 pancreatic ductal adenocarcinoma (PDAC), and 196 HCC samples from multiple studies including TCGA. They identified a distinct subtype of ICC enriched for *IDH* mutants, and found that HCC can be characterized by *CTNNB1* and *TERT* promoter mutations, which are absent in ICC (Farshidfar et al., 2017).

In conclusion, although ICC shares some molecular changes with HCC, likely because of the same cells of origin, this rare cancer also possesses its own unique differentiation and evolution pathways, as well as specific genetic alterations and distinct gene expression patterns.

Different Etiologies of ICC

Parasitic infection by liver flukes, i.e., *O. viverrini* (OV) and *C. sinensis*, is a well-known ICC risk factor, particularly in Thailand. The chronic liver fluke infection is estimated to account for 8–10% of the overall ICC incidences (Gupta and Dixon, 2017). The gene expressions studied by Jinawath et al. (2006) was one of the first reports to elucidate the different genetic mechanisms between liver fluke- and non-liver fluke-associated ICCs. Using cDNA microarray, the authors compared the two groups of ICC at the transcriptional level, and found that genes involved in xenobiotic and endobiotic metabolisms, i.e., UDP-glucuronosyltransferase (*UGT2B11*, *UGT1A10*) and sulfotransferases (*CHST4, SUT1C1*), have higher expression in liver fluke-associated ICCs comparing to non-liver fluke group. These genes are believed to play important roles in detoxification of carcinogens such as nitrosamines from preserved food and, if any, toxic substances released from the parasites or created by parasite-induced chronic inflammation. On the other hand, genes involved in growth factor signaling show higher expression in non-liver fluke ICCs.

Different causative etiologies may induce distinct somatic alterations. Recurrent infection of liver flukes, particularly OV, has been associated with different DNA mutation signatures in ICCs. A WES study demonstrated that the frequently mutated genes in OV-related ICCs comprise both known cancer genes, such as *TP53*, *KRAS* and *SMAD4*, and newly implicated cancer genes including *MLL3*, *ROBO2*, *RNF43*, *PEG3*, and *GNAS*, which are genes involved in histone methylation, genome stability, and G-protein signaling (Ong et al., 2012). Another WES study further showed that *TP53* mutations are more enriched in OV-related ICCs, while mutations in *BAP1*, *IDH1*, and *IDH2* genes are more common in non-OV-related tumors (Chan-On et al., 2013).

A recent multi-omics study analyzed the combined datasets of WGS, WES, CNAs, transcriptomes and epigenomes, and identified four CCA clusters likely driven by distinct etiologies, with separate genetic, epigenetic, and clinical features (Jusakul et al., 2017). The results showed that liver fluke infection is one of the most important classification factors and is also the factor that leads to poorer prognosis. From this study, clusters 1 and 2, which are liver fluke positive, are enriched for recurrent mutations in *TP53*, *ARID1A* and *BRCA1/2*, and *ERBB2* amplifications. In contrast, clusters 3 and 4, which comprise mostly non-liver fluke-associated tumors, are enriched for recurrent mutations in epigenetic-related genes, i.e., *BAP1* and *IDH1/2*, as well as *FGFR* rearrangements, and have high PD-1/PD-L2 expression. Additionally, DNA hypermethylation of CpG islands and high levels of mutations in H3K27me3-associated promoters were only observed in clusters 1, while cluster 4 exhibited DNA hypermethylation in CpG shores. These findings suggest different mutational pathways across all four CCA subtypes.

Other than liver fluke, hepatitis virus infection has been proposed to be associated with an increased risk of ICC as well. A meta-analysis of the combined 13 case–control studies and three cohorts of ICC patients has reported a statistically significant increased risk of ICC incidence with HBV and HCV infection (OR = 3.17, 95% CI, 1.88–5.34, and OR = 3.42, 95% CI, 1.96–5.99, respectively) (Zhou et al., 2012). To investigate whether viral hepatitis-associated ICC may harbor specific histomorphological and genetic features, Yu et al. analyzed the 170 ICC patients who were either seropositive or seronegative for HBV or HCV. The authors identified N-cadherin as an immunohistochemistry (IHC) marker for viral hepatitis-associated ICC. N-cadherin IHC positivity is also strongly associated with cholangiolar morphology, lack of CEA, high MUC2 expression, and low *KRAS* mutation frequency (Yu et al., 2011). In line with these findings, another study conducting WES in ICCs found that HBV-associated ICCs carry high *TP53* mutation loads, while mutations in *KRAS* are almost exclusively identified in tumors of HBV-seronegative patients (Zou et al., 2014). However, larger scale high-throughput studies have yet to be performed in viral hepatitis-associated ICCs.

Other Molecular and Clinical Aspects

Based on gene expression and SNP microarrays, two main subtypes of ICC, proliferation (PF) and inflammation (IF), were identified (Sia et al., 2013a). The PF subtype is more common and can be characterized by activation of oncogenic signaling pathways, DNA amplifications of 11q13.2 (including *CCND1* and *FGF19* gene loci), deletions of 14q22.1 (including *SAV1* gene locus), mutations in *KRAS* and *BRAF*, and is associated with a poor prognosis. In contrast, the IF subtype is characterized by activation of inflammatory signaling pathways, overexpression of cytokines and STAT3 activation, and is associated with a better prognosis. Another study led by Anderson et al. classified ICC patients into two subgroups based on 5-year survival rate, time to recurrence, and the absence or presence of *KRAS* mutations. Similarly, *KRAS* mutations are associated with poor clinical outcomes (Andersen et al., 2013).

As mentioned earlier, based on a large-scale TCGA study, mutational signatures can be divided into two major classes, namely M and C (Ciriello et al., 2013). By combining WES and transcriptomic data, a study showed that ICCs carry signatures of both M and C classes as well (Kim et al., 2016). ICC of C class harbors recurrent focal CNAs including deletions involving *CDKN2A, ROBO1, ROBO2*, *RUNX3*, and *SMAD4*, while those of M class harbor recurrent mutations in the genes frequently mutated in ICC, i.e., *TP53*, *KRAS*, and *IDH1*, as well as epigenetic regulators and genes in TGFβ signaling pathway.

Focusing on the genomic findings from all ICC studies discussed above, recurrent mutations of ICC are enriched in tumor suppressor genes, i.e., *ARID1A*, *ARID1B*, *BAP1*, *PBRM1*, *TP53*, *STK11*, and *PTEN*, and oncogenes, i.e., *IDH1*, *IDH2*, *KRAS*, *BRAF*, and *PIK3CA*. The frequencies of these recurrent mutations in ICC across multiple studies are summarized in **Figure 1**. The majority of these genes are associated with genome instability and epigenetic alterations, which are the common underlying mechanisms of cancer. Recurrent mutations of *BRCA2*, *MLL3*, *APC*, *NF1*, and *ELF3* tumor-suppressor genes have also been reported in ICC (Farshidfar et al., 2017). Using transcriptomic analysis, fibroblast growth factor receptor 2 (*FGFR2*) fusion genes, i.e., *FGFR2-AHCYL, FGFR2-BICC1 type1, FGFR2-BICC1 type2, FGFR2-PPHLN1, FGFR2-MGEA5, FGFR2-TACC3, FGFR2-KIAA1598, FGFR2-KCTD1, and FGFR2-TXLNA,* are found to be one of the most prevalent alterations in ICC (Jiao et al., 2013; Borad et al., 2014; Ross et al., 2014; Murakami et al., 2015; Sia et al., 2015; Farshidfar et al., 2017; **Figure 1**). Furthermore, they are reported to be exclusively present in ICC, but not ECC and gallbladder cancer (Nakamura et al., 2015). FGFR2 fusion proteins have been shown to facilitate oligomerization and FGFR kinase activation, resulting in altered cell differentiation and increased cell proliferation (Wu et al., 2013). Although the genomic and transcriptomic analyses of ICC support the use of targeted therapeutic interventions, there is currently no targeted therapy considered effective for this disorder. In order to develop a strategy to overcome this challenge, a disease model that mimics most or all biological and genetic aspects of ICC is an ideal tool for performing functional studies of the target genes or screening potential anticancer drugs. In the coming sections, we will update the recent progress and introduce new disease models that may expedite the discovery of novel treatment for ICC.

CURRENT DISEASE MODELS OF ICC

The first ICC cell line, HChol-Y1, was established in 1985. The cell line secretes very low levels of CEA and high level of CA 19-9, which are the markers of various kind of cancers (Yamaguchi et al., 1985). Since then, many more ICC cell lines originating from ICCs with different etiologies have been established around the world. PCI:SG231 (Storto et al., 1990), CC-SW-1 (Shimizu et al., 1992), CC-LP-1 (Shimizu et al., 1992) cell lines were established from patients in the US. HuH-28 (Kusaka et al., 1988), KMCH-2 (Yano et al., 1996), RBA (Enjoji et al., 1997), SSP-25 (Enjoji et al., 1997), NCC-CC1, NCC-CC3-1, NCC-CC3-2, and NCC-CC4-1 (Ojima et al., 2010) were derived from Japanese patients. SNU-1079 (Ku et al., 2002) was derived from a Korean patient, while HKGZ-CC (Ma et al., 2007), and HCCC-9810 (Liu et al., 2013) were derived from Chinese patients. In particular, HuCCA-1 was established from the tumor removed from a Thai patient with liver fluke infection (Sirisinha et al., 1991). This cell line is from epithelial cell origin and secretes a number of non-specific tumor markers including CA125 (Srisomsap et al., 2004).

Unlike most of the ICC cell lines established directly from primary tumor cells, two cell lines, namely MT-CHC01 and KKU-213L5, were established by generating xenograft, which is the growing of human primary tumor cells in the immunodeficient mice, such as nonobese diabetic (NOD)/Shi-severe combined immunodeficient (*scid*)-IL2rγ^{null} mice (NOG mice). MT-CHC01 was established from a xenograft derived from the tumor of an Italian patient. After growing primary tumor cells in NOD/Shi-*scid* mice for four generations, the xenograft was stabilized, and the tumors were resected from mice to generate xenograft-derived cell lines. MT-CHC01 retains epithelial cell markers, and shows stemness and pluripotency markers (Cavalloni et al., 2016b). After subcutaneous injection, it retains *in vivo* tumorigenicity and expresses CEA and CA19-9; *KRAS* G12D mutation is also maintained in this cell line. KKU-213L5 was recently derived from its parental cell line, KKU-213, which was established from the primary tumor of a Thai patient. KKU-213L5 was selected *in vivo* through five serial passages of pulmonary metastasized tissues via tail-vein injection into NOD/scid/Jak3 mice (Uthaisar et al., 2016). Compared to KKU-213, KKU-213L5 possesses higher metastatic behaviors, such as higher migration and invasion abilities, and also shows stem cell characteristics. The cells exhibit significantly higher expression of anterior gradient protein-2 (*AGR2*) and suppression of *KiSS-1*, which are associated with metastasis in the later stages of disease (**Figure 2A**).

Recently, the use of human tumor xenograft or patient-derived xenograft (PDX) provides a "patient-like" environment in animal models for a better study of human cancers. To generate PDX, tumor cells are transplanted into immunocompromised animals either by subcutaneous injection or by injecting into the desired organs directly. An orthotopic

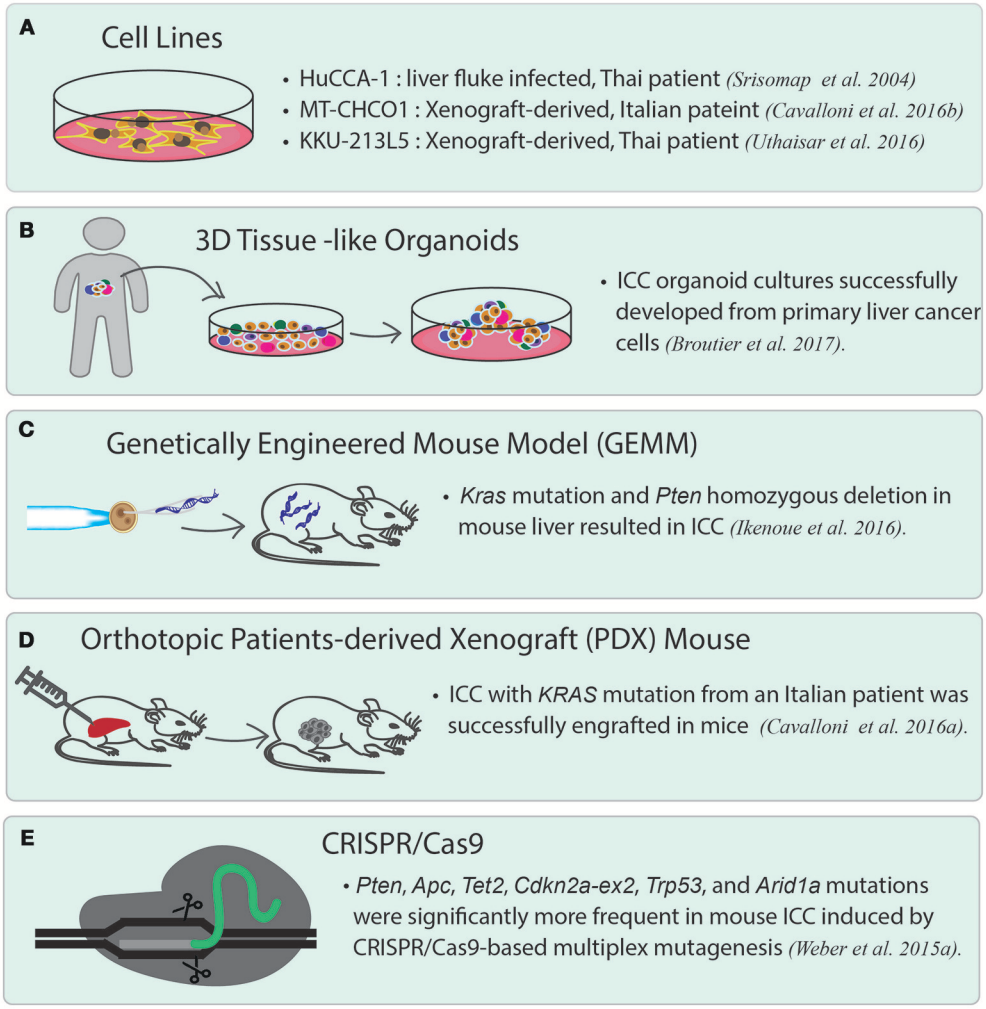


FIGURE 2 | Current disease models for studying ICC. **(A)** ICC cancer cell lines. There are many cell lines established from primary tumor cells. Three representative cell lines are listed. HuCCA-1 was derived from a Thai patient with liver fluke infection. MT-CHCO1 and KKU-213L5 were both established from patient-derived xenografts (PDX). **(B)** 3D patient-derived tissue-like organoids. Organoids preserve the properties of primary tumor cells as well as tissue heterogeneity. **(C)** Genetically engineered mouse model (GEMM). A GEMM of ICC was generated by inducing oncogenic *KRAS* mutation and homozygous *PTEN* deletion in mouse liver. **(D)** Orthotopic patient-derived xenograft (PDX). In orthotopic PDX mouse models, patient-derived tumor cells are transplanted into the same organ from which the patient's cancer originated, followed by stabilizing the tumors in the animals. **(E)** A mouse model of ICC created by CRISPR/*Cas9* gene editing. CRISPR/*Cas9* is used to introduce mutations to the selected tumor suppressor genes including *Arid1a*, *Trp53*, *Tet2*, *Pten*, *Cdkn2a, Apc, Brca1/2*, and *Smad4*, which lead to ICC in the gene-edited mice.

xenograft model is generated by either implanting or injecting human tumor cells into the equivalent organ from which the cancer originated. It is widely believed that orthotopic PDX reflects the original tumor microenvironment much better than the conventional subcutaneous xenograft models. Recently, a novel PDX model was generated from an Italian patient with ICC. This PDX shows the same biliary epithelial markers, tissue architecture, and genetic aberrations as the primary tumor (Cavalloni et al., 2016a) (**Figure 2D**). Other than PDX, a genetically engineered mouse model of ICC has been generated by inducing oncogenic *Kras* mutation and homozygous *Pten* deletion in the liver. The tumors induced in this model are exclusively ICCs and show histological phenotype similar to human ICC with cholangiocyte origin. This mouse line is suited for the development of new therapies for ICCs with an oncogenic *KRAS* mutation and the activated PI3K pathway (Ikenoue et al., 2016) (**Figure 2C**). The latest gene-editing technology, CRISPR/*Cas9* technique, has successfully been used to induce ICC in mice. A study led by Weber J. et al. (2015) introduced mutations in a set of tumor suppressor genes often altered in human ICC/HCC such as *Arid1a*, *Pten*, *Smad4*, *Trp53*, *Apc, Cdkn2a,* and in a few rarely mutated genes including *Tet2, Brca1/2,* in mice by conducting multiplex CRISPR/*Cas9* gene editing. The results showed that CRISPR/*Cas9*-induced mouse ICCs preferentially carry higher frequencies of mutations in the frequently dysregulated genes in human ICCs, especially those related to chromatin modification. However, the authors unexpectedly observed a high mutation frequency of *Tet2*,

which has never been observed in human ICCs. Although *TET2* mutations have not been reported in human ICC, *TET2* is believed to harbor tumor suppressive function linked to *IDH1/2*, which are among the commonly mutated oncogenes in ICC. The authors, therefore, brought up the importance of genetic screening in pinpointing the cancer genes that may not be mutated, but altered by other mechanisms (Weber J. et al., 2015) (**Figure 2E**).

TRANSLATIONAL CLINICAL ASPECTS AND FUTURE DIRECTIONS

Looking ahead on the future of cancer research, one of the most exciting trends is the application of patient-derived organoids, which serve as a source of expanded *in vitro* patient-derived cancer cells (**Figure 2B**). This essentially provides a 3D semi-solid tissue-like architecture that captures the real structure and heterogeneity of a solid tumor, a quality that is lacking in the commonly used immortalized cancer cell lines. Organoid, therefore, serves as a good model for studying the underlying carcinogenesis mechanisms, as well as for drug sensitivity testing and developing targeted therapies (Lancaster and Knoblich, 2014). Recently, human cholangiocytes were isolated and propagated from human extrahepatic biliary tree in the form of organoids as a proof-of-concept experiment for regenerative medicine applications (Sampaziotis et al., 2017). These extrahepatic cholangiocyte organoids can form tissue-like structures with biliary characteristics when transplanted into immunocompromised mice, and can reconstruct the gallbladder wall by repairing the biliary epithelial cells in a mouse model of injury. The results showed that bioengineered artificial ducts can functionally mimic the native common bile duct. Recently, Broutier et al. has successfully developed organoids from primary cell culture of HCC, CHC, ICC, and perihilar CCA (Broutier et al., 2017). By generating ICC organoids that reflect the heterogeneous origins and etiologies, we foresee a possibility of identifying the functions of somatic alterations in ICC by systematically conducting CRISPR/*Cas9* gene editing. In addition, one can investigate the effects of microenvironment more thoroughly (i.e., tumor-immune interactions and cell-cell communications), the cell state transition, and test the efficacy of drugs in a high-throughput manner. Ultimately, patient-derived organoids together with PDX mice may serve as two of the most important models for the development of precision medicine in ICC and other rare cancers. In **Figure 3**, we summarize the application of precision oncology through the use of high-throughput technologies and disease models to expedite translational research outcomes in ICC.

Intra-tumor heterogeneity reflects the diverse clonal evolution of tumor cells. Tumor evolution is proposed to have one of the following characteristics; hypermutability phenotypes, various mutation signature patterns, weak clonal selection, and high heterogeneity of tumor cell subclones (Schwartz and Schäffer, 2017). Extensive intra-tumor heterogeneity of ICC has lately been observed using WES, which identified branch evolution collectively shaped by parallel evolution and chromosome instability as the predominant pattern of ICC (Dong et al., 2018). As single-cell omics technologies have become more matured recently, it is now possible to characterize the reference expression patterns of individual cells in human (Nawy, 2014) in order to provide the most fundamental knowledge for understanding human health and diseases (Rozenblatt-Rosen et al., 2017). Such advanced technologies will also expedite understanding of carcinogenesis mechanisms, including those of ICC. These approaches include generating transcriptomes and epigenomes at the single-cell level (scRNA-seq and scATAC-seq, respectively), as well as spatial transcriptomes, which can be used to investigate physical relationships of each cell in a tumor mass (Ståhl et al., 2016). Single-cell genomics has also become another important tool for understanding the clonal evolution of tumor cells phylogenetically by exploring the mutating ability of cancer cells (Kim and Simon, 2014; Müller and Diaz, 2017). In the same way, single-cell genomics may help better elucidate the heterogeneity of ICC, particularly when combined with other multi-level omics data generated from either primary tumor cells or the patient-derived 3D tumor model such as organoids. A recent study by Roerink et al. has investigated the nature and extent of intra-tumor diversification at the single cell level by characterizing organoids derived from multiple single cells from three colorectal cancers and adjacent normal intestinal crypts. Interestingly, the responses to anticancer drugs between even closely related cells of the same tumor are markedly different, emphasizing the importance of studying individual cancer cells (Roerink et al., 2018).

With the current advances in NGS technology, the genomic landscapes of ICC have been largely revealed, which is critically important for the clinical development of novel drugs. In addition, the multi-omics profiles that can classify tumor types based on molecular features may be essential for the clinical success in treating the patients. Toward this direction, the clinical trials driven by biomarkers are being conducted. Many ongoing clinical trials of all types of CCA including ICC are listed in **Table 3**. Among these, targeting FGFR alterations appear to be particularly promising. A phase 2 study of BGJ398, a selective pan-FGFR inhibitor, in metastatic FGFR-altered CCA patients who failed or were intolerant to platinum-based chemotherapy demonstrated impressive anti-tumor activity (Javle et al., 2016). Among the 22 evaluable metastatic patients harboring FGFR2 fusions or other alterations, three patients achieved partial response (PR) and 15 patients had stable disease (SD). A Phase 1 study of ARQ 087, an oral pan-FGFR inhibitor, in patients harboring FGFR2 fusions demonstrated two patients with a confirmed PR and one with durable SD at $\geq$16 weeks (Papadopoulos et al., 2017). A phase 3 study of ARQ 087 is ongoing and recruiting more patients with FGFR2 fusions as well as inoperable or advanced ICC (NCT03230318). Other novel drugs targeting FGFR fusions such as INCB054828, H3B-6527, erdafitinib, and INCB062079 are in early phases of clinical development (**Table 3**).

Mutations of *IDH1* were reported in up to 25% of CCA (Lowery et al., 2017). AG-120, a highly selective small molecule inhibitor of mutant IDH1 protein, demonstrated a preliminary efficacy in refractory CCA patients with *IDH1* mutations. A

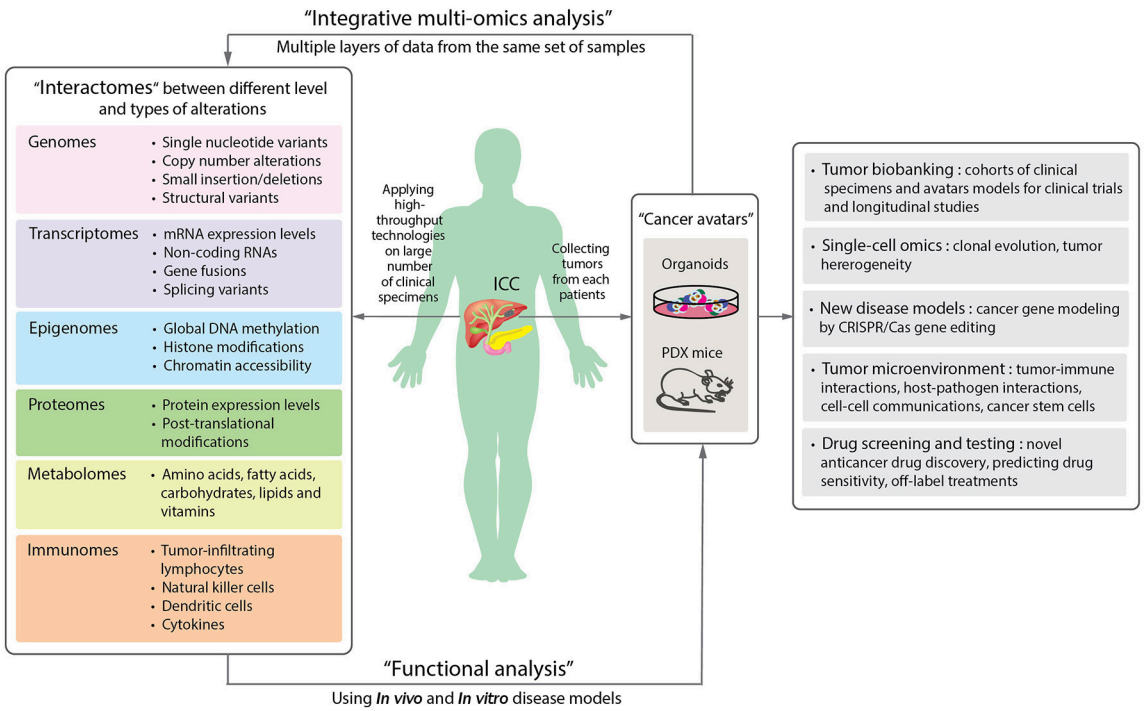


FIGURE 3 | A schematic diagram proposing the application of precision oncology in ICC through the use of high-throughput technologies and disease models. By applying high-throughput technologies on large numbers of patient samples, different levels of omics data can be obtained and provide information of the molecular changes in the tumor cells or microenvironments **(Left panel)**. Aberrant alterations identified from omics data can then be functionally validated in disease models. Organoids and patient-derived xenograft (PDX) mouse are new disease models **(Right panel)**. The two "next-generation" tumor avatars provide "patient-like" models for integrative multi-omics analyses to study the underlying mechanisms of disorders. The avatars can be used for the following studies: single cell sequencing for understanding clonal evolution and heterogeneity of tumors, disease models for gene editing, tumor microenvironments, and high-throughput systematic drug screening and testing. They can further be biobanked for future studies (**Far right panel**).

phase 1 study of AG-120 reported one patient who achieved PR and five patients with SD >6 months (Lowery et al., 2017). A phase 3 randomized placebo-controlled study of AG-120 in *IDH1* mutation-positive patients is underway (NCT02073994) (**Table 3**).

Immunotherapy such as checkpoint inhibitor may be effective only in patients with mismatch-repair deficiency (dMMR). In CCA including ICC, incidences of dMMR and/or microsatellite instability-high (MSI-H) were variously reported as quite low (Liengswangwong et al., 2003, 2006; Limpaiboon et al., 2006; Walter et al., 2017). A phase 2 non-randomized study of pembrolizumab, an anti-PD1 antibody, in 41 patients with progressive metastatic carcinoma demonstrated an immune-related objective response rate of 40, 71, and 0% for the patients who have colorectal cancer with dMMR, CCA and other cancers with dMMR, and colorectal cancer with mismatch-repair proficiency (pMMR), respectively (Le et al., 2015). In addition, WES revealed an average of 1,782 somatic mutations for each dMMR tumor compared with only 73 for a pMMR tumor ($P = 0.007$). High somatic mutation loads were also associated with prolonged progression-free survival (PFS) ($P = 0.02$). Hence, dMMR tumors with a large number of somatic mutations may be more susceptible to immune checkpoint blockade, as a result of the substantial amount of new immunogenic antigens produced. Based on these findings, US FDA (Food and Drug Administration) has granted accelerated approval to pembrolizumab in patients with unresectable or metastatic solid tumors with MSI-H or dMMR. A phase 1b study of pembrolizumab (KEYNOTE-028) with 89 advanced biliary tract cancer patients has reported a preliminary efficacy of checkpoint inhibitor (Bang et al., 2015). Overall response rate was observed in ∼17% of the patients. Several other ongoing studies of checkpoint inhibitors are being investigated in combination with other drugs including chemotherapy, targeted drugs, and other immunotherapies (**Table 3**).

CONCLUSIONS

In summary, we have described how the advances in high-throughput technologies have provided a massive amount of information in understanding the genetic mechanisms of disorders, including rare cancers, and in particular, ICC. To be able to effectively utilize such high-throughput methods in cancer research, one should take the following into consideration. First, the determination of clinical information, such as risk factor exposure or etiologies, disease stages, responsiveness to therapy, histology subtypes and anatomical locations, prior to

TABLE 3 | Ongoing clinical trials of targeted therapy in cholangiocarcinoma[a].

Drug	Targets	Phase	Combination	Trial number
DRIVER MUTATIONS				
Dasatinib	IDH1/2	II		NCT02428855
AG-120	IDH1	I, III		NCT02073994, NCT02989857
Metformin	IDH1/2	I/II	Chloroquine	NCT02496741
Varlitinib	EGFR (ErbB-1), Her-2/neu (ErbB-2)	II		NCT02609958
Leucovorin and nal-IRI	EGFR, KRAS	II	5-FU	NCT03043547
Niraparib	BAP1	II		NCT03207347
Merestinib	c-Met, HGFR	I	Gemcitabine + Cisplatin	NCT03027284
LOXO-195	NTRK1, NTRK2, NTRK3	I/II		NCT03215511
Trastuzumab Emtansine	HER2	II		NCT02999672
DKN-01	Wnt, DKK1	I	Gemcitabine + Cisplatin	NCT02375880
Copanlisib (BAY 80-6946)	PI3K signaling pathway	II	Gemcitabine + Cisplatin	NCT02631590
Panitumumab	EGF	II	Gemcitabine + Irinotecan	NCT00948935
FUSION GENE				
ARQ 087	FGFR2	I/II , II		NCT01752920, NCT03230318
BGJ398	FGFR2	II		NCT02150967
INCB054828	FGFR2	II		NCT02924376
H3B-6527	FGFR4	I		NCT02834780
Erdafitinib	FGFR	II		NCT02699606
Ceritinib (LDK378)	ROS1, ALK	II		NCT02638909, NCT02374489
INCB062079	FGFR4, FGF19	I		NCT03144661
Entrectinib	ROS1, ALK TrkA, TrkB, TrkC	II		NCT02568267
LOXO-101	NTRK fusion	II		NCT02576431
ANGIOGENESIS				
Apatinib	VEGFR-2	III		NCT03251443
Ramucirumab	VEGFR-2	II		NCT02520141
Regorafenib	VEGFR, RET, RAF-1, KIT, PDGFRB, FGFR1, TIE2, BRAF(V600E)	II		NCT02053376
Pazopanib	VEGF, PDGFR, FGFR, KIT	II	Gemcitabine	NCT01855724
	VEGFR/PDGFR/Raf MEK MAPK/ERK	I	GSK1120212	NCT01438554
CHECKPOINT INHIBITOR				
Durvalumab (MEDI 4736)	PD-L1, PD-1	I	Guadecitabine (SGI-110)	NCT03257761
Pembrolizumab	PD-1	II	Peginterferon alpha-2b (Sylatron)	NCT02982720
	PD-L1, PD-L2 HSP90	I	XL888	NCT03095781
Atezolizumab	PD-L1	II	Cobimetinib	NCT03201458
	PD-L1	I	Gemcitabine+ Cisplatin	NCT03267940
Nivolumab	PD-1, PD-L1 HDAC inhibitor	II	Entinostat	NCT03250273
	CTLA-4 PD-1	II	Ipilimumab	NCT02834013
ABBV-181	PD-1, PD-L1	I	Rovalpituzumab Tesirine	NCT03000257
ABBV-368	OX40	I	Monotherapy or combination with ABBV-181	NCT03071757
OTHER PATHWAYS				
RRx-001	G6PD	II	Gemcitabine + Cisplatin	NCT02452970
CX-4945	CK2	I/II	Gemcitabine + Cisplatin	NCT02128282
Melphalan/HDS	Induce covalent guanine N7-N7 intra- and inter-crosslinks and alkylation of adenine N3 of DNA.	II/III	Gemcitabine + Cisplatin	NCT03086993
BBI503	Cancer stem cell (CSC)	II		NCT02232633
Acelarin (NUC-1031)	dFdCDP, dFdCTP	I	Cisplatin	NCT02351765
CX-2009	Tumor-associated antigen (TAA) CD166	I/II		NCT03149549

[a] *Information acquired from Clinicaltrials.gov.*

inclusion of the clinical samples is crucial, as it may affect the overall success of downstream analyses. For ICC, liver fluke and hepatitis virus infections are both strongly associated with the disease. Hence, additional information on whether the patients are seropositive for these infections may help better characterize the sample subgroups. Furthermore, ICC can also be subcategorized by macroscopic features, i.e., MF, IDG, and PDI, which rely on accurate pathological determination of the tumor sections. Secondly, insufficient sample size is one of the greatest challenges in studying ICC and other rare cancer types. This cancer in particular is prevalent in certain regions in Asia, such as northeastern Thailand, where most patients are believed to be associated with liver fluke infection. Finding a suitable ICC cohort with adequate sample size is difficult. Earlier studies have combined patients from different countries/geographical regions as well as other different types of BTC e.g., ECC and gallbladder cancer, in order to elucidate the molecular mechanisms and treatment responses. These cohorts, particularly in the form of clinical trials, are consisted of patients and tumors with different genetic backgrounds, which may have resulted in therapeutic failure due to the confounding factors and selection biases. Lastly, the small amount or low quality of source clinical materials limit the comprehensive applications of true "multi-omics" approaches. The majority of previous studies relied on obtaining multiple levels of omics information from different sets of ICC patients. The restricted amount of biological materials from one patient is the main hindrance of performing multiple omics analyses at once to comprehensively investigate the correlation and connections between multiple regulatory processes. Therefore, in addition to a good systematic longitudinal collection of clinical specimens from cancer patients in a tumor biobank, having organoids or PDX mouse models as "cancer avatars" would, at least in part, solve the problem of sample limitation, and should contribute to better omics study design and more effective translational outcomes for rare cancer patients.

AUTHOR CONTRIBUTIONS

KC, M-SS, and NJ conceived the concept of the review and figures. KC, M-SS, VC, NN, and NJ wrote the manuscript. KC, M-SS, NN, and NJ prepared the tables and figures. All the authors read, reviewed, and approved the final manuscript.

FUNDING

NJ is a recipient of TRF Research Scholar Fund (RSA5780065), Government Fiscal Year Budget Funds administered through Mahidol University-National Research Council of Thailand and the Research University Network (RUN) Program, and the research grant from the Ramathibodi Cancer Center. NN acknowledges the Talent Management Program, Mahidol University. VC acknowledges the TRF Grant for New Researcher (MRG6080235), Newton Advanced Fellowship through TRF (DBG60800003) and Royal Society (NA160153), and Faculty of Science, Mahidol University. The NJ and VC laboratories are supported by the Crown Property Bureau Foundation through Integrative Computational BioScience (ICBS) Center, Mahidol University.

ACKNOWLEDGMENTS

The authors thank Mr. Jeffrey Makin for very helpful comments on the manuscript.

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421. doi: 10.1038/nature12477

American Cancer Society, Inc. (2018). *Data from: Bile Duct Cancer; Early Detection, Diagnosis, and Staging; Survival Rates for Bile Duct Cancer*. Available online at: https://www.cancer.org/content/cancer/en/cancer/bile-duct-cancer/detection-diagnosis-staging/survival-by-stage.html

Andersen, J. B., Spee, B., Blechacz, B. R., Avital, I., Komuta, M., Barbour, A., et al. (2013). Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. *Gastroenterology* 142, 1021.e1015–1031.e1015. doi: 10.1053/j.gastro.2011.12.005

Arai, Y., Totoki, Y., Hosoda, F., Shirota, T., Hama, N., Nakamura, H., et al. (2014). Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma. *Hepatology* 59, 1427–1434. doi: 10.1002/hep.26890

Ballehaninna, U. K., and Chamberlain, R. S. (2012). The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: an evidence based appraisal. *J. Gastrointest. Oncol.* 3, 105–119. doi: 10.3978/j.issn.2078-6891.2011.021

Bang, Y. J., Doi, T., Braud, F. D., Piha-Paul, S., Hollebecque, A., Razak, A. R. A., et al. (2015). 525 Safety and efficacy of pembrolizumab (MK-3475) in patients (pts) with advanced biliary tract cancer: interim results of KEYNOTE-028. *Eur. J. Cancer* 51:S112. doi: 10.1016/S0959-8049(16)30326-4

Bekaii-Saab, T., Phelps, M. A., Li, X., Saji, M., Goff, L., Kauh, J. S., et al. (2011). Multi-institutional phase II study of selumetinib in patients with metastatic biliary cancers. *J. Clin. Oncol.* 29, 2357–2363. doi: 10.1200/JCO.2010.33.9473

Benhamouche, S., Curto, M., Saotome, I., Gladden, A. B., Liu, C. H., Giovannini, M., et al. (2010). Nf2/Merlin controls progenitor homeostasis and tumorigenesis in the liver. *Genes Dev.* 24, 1718–1730. doi: 10.1101/gad.1938710

Bergquist, J. R., Ivanics, T., Storlie, C. B., Groeschl, R. T., Tee, M. C., Habermann, E. B., et al. (2016). Implications of CA19-9 elevation for survival, staging, and treatment sequencing in intrahepatic cholangiocarcinoma: a national cohort analysis. *J. Surg. Oncol.* 114, 475–482. doi: 10.1002/jso.24381

Blechacz, B. (2017). Cholangiocarcinoma: current knowledge and new developments. *Gut Liver* 11, 13–26. doi: 10.5009/gnl15568

Borad, M. J., Champion, M. D., Egan, J. B., Liang, W. S., Fonseca, R., Bryce, A. H., et al. (2014). Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma. *PLoS Genet.* 10:e1004135. doi: 10.1371/journal.pgen.1004135

Broutier, L., Mastrogiovanni, G., Verstegen, M. M. A., Francies, H. E., Gavarró, L. M., Bradshaw, C. R., et al. (2017). Human primary liver cancer -derived organoid cultures for disease modelling and drug screening. *Nat. Med.* 23, 1424–1435. doi: 10.1038/nm.4438

Cao, X. L., Li, H., Yu, X. L., Liang, P., Dong, B. W., Fan, J., et al. (2013). Predicting early intrahepatic recurrence of hepatocellular carcinoma after microwave ablation using SELDI-TOF proteomic signature. *PLoS ONE* 8:e82448. doi: 10.1371/journal.pone.0082448

Cavalloni, G., Peraldo-Neia, C., Sassi, F., Chiorino, G., Sarotto, I., Aglietta, M., et al. (2016a). Establishment of a patient-derived intrahepatic cholangiocarcinoma xenograft model with KRAS mutation. *BMC Cancer* 16:90. doi: 10.1186/s12885-016-2136-1

Cavalloni, G., Peraldo-Neia, C., Varamo, C., Casorzo, L., Dell'Aglio, C., Bernabei, P., et al. (2016b). Establishment and characterization of a human intrahepatic cholangiocarcinoma cell line derived from an Italian patient. *Tumour Biol.* 37, 4041–4052. doi: 10.1007/s13277-015-4215-3

Chaisaingmongkol, J., Budhu, A., Dang, H., Rabibhadana, S., Pupacdi, B., Kwon, S. M., et al. (2017). Common molecular subtypes among asian hepatocellular carcinoma and cholangiocarcinoma. *Cancer Cell* 32, 57.e3–70.e3. doi: 10.1016/j.ccell.2017.05.009

Chalasani, N., Baluyut, A., Ismail, A., Zaman, A., Sood, G., Ghalib, R., et al. (2000). Cholangiocarcinoma in patients with primary sclerosing cholangitis: a multicenter case-control study. *Hepatology* 31, 7–11. doi: 10.1002/hep.510310103

Chan-On, W., Nairismagi, M.-L., Ong, C. K., Lim, W. K., Dima, S., Pairojkul, C., et al. (2013). Exome sequencing identifies distinct mutational patterns in liver fluke-related and non-infection-related bile duct cancers. *Nat. Genet.* 45, 1474–1478. doi: 10.1038/ng.2806

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

DeOliveira, M. L., Cunningham, S. C., Cameron, J. L., Kamangar, F., Winter, J. M., Lillemoe, K. D., et al. (2007). Cholangiocarcinoma: thirty-one-year experience with 564 patients at a single institution. *Ann. Surg.* 245, 755–762. doi: 10.1097/01.sla.0000251366.62632.d3

Dong, L. Q., Shi, Y., Ma, L. J., Yang, L. X., Wang, X. Y., Zhang, S., et al. (2018). Spatial and temporal clonal evolution of intrahepatic cholangiocarcinoma. *J. Hepatol.* 69, 89–98. doi: 10.1016/j.jhep.2018.02.029

Enjoji, M., Sakai, H., Nawata, H., Kajiyama, K., and Tsuneyoshi, M. (1997). Sarcomatous and adenocarcinoma cell lines from the same nodule of cholangiocarcinoma. *In Vitro Cell. Dev. Biol. Anim.* 33, 681–683. doi: 10.1007/s11626-997-0125-z

Farshidfar, F., Zheng, S., Gingras, M.-C., Newton, Y., Shih, J., Robertson, A. G., et al. (2017). Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-mutant molecular profiles. *Cell Rep.* 18, 2780–2794. doi: 10.1016/j.celrep.2017.02.033

Fwu, C. W., Chien, Y. C., You, S. L., Nelson, K. E., Kirk, G. D., Kuo, H. S., et al. (2011). Hepatitis B virus infection and risk of intrahepatic cholangiocarcinoma and non-hodgkin lymphoma: a cohort study of parous women in Taiwan. *Hepatology* 53, 1217–1225. doi: 10.1002/hep.24150

Gupta, A., and Dixon, E. (2017). Epidemiology and risk factors: intrahepatic cholangiocarcinoma. *Hepatobiliary Surg. Nutr.* 6, 101–104. doi: 10.21037/hbsn.2017.01.02

Holczbauer, A., Factor, V. M., Andersen, J. B., Marquardt, J. U., Kleiner, D. E., Raggi, C., et al. (2013). Modeling pathogenesis of primary liver cancer in lineage-specific mouse cell types. *Gastroenterology* 145, 221–231. doi: 10.1053/j.gastro.2013.03.013

Hyder, O., Hatzaras, I., Sotiropoulos, G. C., Paul, A., Alexandrescu, S., Marques, H., et al. (2013). Recurrence after operative management of intrahepatic cholangiocarcinoma. *Surgery* 153, 811–818. doi: 10.1016/j.surg.2012.12.005

Ikenoue, T., Terakado, Y., Nakagawa, H., Hikiba, Y., Fujii, T., Matsubara, D., et al. (2016). A novel mouse model of intrahepatic cholangiocarcinoma induced by liver-specific Kras activation and Pten deletion. *Sci. Rep.* 6:23899. doi: 10.1038/srep23899

Javle, M. M., Shroff, R. T., Zhu, A., Sadeghi, S., Choo, S., Borad, M. J., et al. (2016). A phase 2 study of BGJ398 in patients (pts) with advanced or metastatic FGFR-altered cholangiocarcinoma (CCA) who failed or are intolerant to platinum-based chemotherapy. *J. Clin. Oncol.* 34(Suppl. 4), 335–335. doi: 10.1200/jco.2016.34.4_suppl.335

Jensen, L. H., Lindebjerg, J., Ploen, J., Hansen, T. F., and Jakobsen, A. (2012). Phase II marker-driven trial of panitumumab and chemotherapy in KRAS wild-type biliary tract cancer. *Ann. Oncol.* 23, 2341–2346. doi: 10.1093/annonc/mds008

Jiao, Y., Pawlik, T. M., Anders, R. A., Selaru, F. M., Streppel, M. M., Lucas, D. J., et al. (2013). Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. *Nat. Genet.* 45, 1470–1473. doi: 10.1038/ng.2813

Jinawath, N., Chamgramol, Y., Furukawa, Y., Obama, K., Tsunoda, T., Sripa, B., et al. (2006). Comparison of gene expression profiles between *Opisthorchis viverrini* and non-Opisthorchis viverrini associated human intrahepatic cholangiocarcinoma. *Hepatology* 44, 1025–1038. doi: 10.1002/hep.21330

Jing, W., Jin, G., Zhou, X., Zhou, Y., Zhang, Y., Shao, C., et al. (2012). Diabetes mellitus and increased risk of cholangiocarcinoma: a meta-analysis. *Eur. J. Cancer Prev.* 21, 24–31. doi: 10.1097/CEJ.0b013e3283481d89

Jusakul, A., Cutcutache, I., Yong, C. H., Lim, J. Q., Huang, M. N., Padmanabhan, N., et al. (2017). Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. *Cancer Discov.* 7, 1116–1135. doi: 10.1158/2159-8290.CD-17-0368

Khan, S. A., Emadossadaty, S., Ladep, N. G., Thomas, H. C., Elliott, P., Taylor-Robinson, S. D., et al. (2012). Rising trends in cholangiocarcinoma: is the ICD classification system misleading us? *J. Hepatol.* 56, 848–854. doi: 10.1016/j.jhep.2011.11.015

Kim, K. I., and Simon, R. (2014). Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* 15:27. doi: 10.1186/1471-2105-15-27

Kim, Y. H., Hong, E. K., Kong, S. Y., Han, S. S., Kim, S. H., Rhee, J. K., et al. (2016). Two classes of intrahepatic cholangiocarcinoma defined by relative abundance of mutations and copy number alterations. *Oncotarget* 7, 23825–23836. doi: 10.18632/oncotarget.8183

Kristensen, V. N., Lingjaerde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Borresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Ku, J. L., Yoon, K. A., Kim, I. J., Kim, W. H., Jang, J. Y., Suh, K. S., et al. (2002). Establishment and characterisation of six human biliary tract cancer cell lines. *Br. J. Cancer* 87, 187–193. doi: 10.1038/sj.bjc.6600440

Kusaka, Y., Tokiwa, T., and Sato, J. (1988). Establishment and characterization of a cell line from a human cholangiocellular carcinoma. *Res. Exp. Med.* 188, 367–375. doi: 10.1007/BF01851205

Lancaster, M. A., and Knoblich, J. A. (2014). Organogenesis in a dish: modeling development and disease using organoid technologies. *Science* 345:1247125. doi: 10.1126/science.1247125

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., et al. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* 372, 2509–2520. doi: 10.1056/NEJMoa1500596

Lee, J., Park, S. H., Chang, H. M., Kim, J. S., Choi, H. J., Lee, M. A., et al. (2012). Gemcitabine and oxaliplatin with or without erlotinib in advanced biliary-tract cancer: a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol.* 13, 181–188. doi: 10.1016/S1470-2045(11)70301-1

Lee, K. P., Lee, J. H., Kim, T. S., Kim, T. H., Park, H. D., Byun, J. S., et al. (2010). The Hippo-salvador pathway restrains hepatic oval cell proliferation, liver size, and liver tumorigenesis. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8248–8253. doi: 10.1073/pnas.0912203107

Li, H., Hu, B., Zhou, Z. Q., Guan, J., Zhang, Z. Y., and Zhou, G. W. (2015). Hepatitis C virus infection and the risk of intrahepatic cholangiocarcinoma and extrahepatic cholangiocarcinoma: evidence from a systematic review and meta-analysis of 16 case-control studies. *World J. Surg. Oncol.* 13:161. doi: 10.1186/s12957-015-0583-9

Liang, B., Zhong, L., He, Q., Wang, S., Pan, Z., Wang, T., et al. (2015). Diagnostic accuracy of serum CA19-9 in patients with cholangiocarcinoma: a systematic review and meta-analysis. *Med. Sci. Monit.* 21, 3555–3563. doi: 10.12659/MSM.895040

Liengswangwong, U., Karalak, A., Morishita, Y., Noguchi, M., Khuhaprema, T., Srivatanakul, P., et al. (2006). Immunohistochemical expression of mismatch repair genes: a screening tool for predicting mutator phenotype in liver fluke infection-associated intrahepatic cholangiocarcinoma. *World J. Gastroenterol.* 12, 3740–3745. doi: 10.3748/wjg.v12.i23.3740

Liengswangwong, U., Nitta, T., Kashiwagi, H., Kikukawa, H., Kawamoto, T., Todoroki, T., et al. (2003). Infrequent microsatellite instability in liver fluke infection-associated intrahepatic cholangiocarcinomas from Thailand. *Int. J. Cancer* 107, 375–380. doi: 10.1002/ijc.11380

Limpaiboon, T., Tapdara, S., Jearanaikoon, P., Sripa, B., and Bhudhisawasdi, V. (2006). Prognostic significance of microsatellite alterations at 1p36 in cholangiocarcinoma. *World J. Gastroenterol.* 12, 4377–4382. doi: 10.3748/wjg.v12.i27.4377

Lin, M.-S., Huang, J.-X., and Yu, H. (2014). Elevated serum level of carbohydrate antigen 19-9 in benign biliary stricture diseases can reduce its value as a tumor marker. *Int. J. Clin. Exp. Med.* 7, 744–750.

Liu, Y., Devescovi, V., Chen, S., and Nardini, C. (2013). Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Syst Biol* 7:14. doi: 10.1186/1752-0509-7-14

Lowery, M. A., Abou-Alfa, G. K., Burris, H. A., Janku, F., Shroff, R. T., Cleary, J. M., et al. (2017). Phase I study of AG-120, an IDH1 mutant enzyme inhibitor: Results from the cholangiocarcinoma dose escalation and expansion cohorts. *J. Clin. Oncol.* 35, 4015–4015. doi: 10.1200/JCO.2017.35.15_suppl.4015

Lu, L., Li, Y., Kim, S. M., Bossuyt, W., Liu, P., Qiu, Q., et al. (2010). Hippo signaling is a potent *in vivo* growth and tumor suppressor pathway in the mammalian liver. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1437–1442. doi: 10.1073/pnas.0911427107

Ma, S., Hu, L., Huang, X. H., Cao, L. Q., Chan, K. W., Wang, Q., et al. (2007). Establishment and characterization of a human cholangiocarcinoma cell line. *Oncol. Rep.* 18, 1195–1200. doi: 10.3892/or.18.5.1195

Malka, D., Cervera, P., Foulon, S., Trarbach, T., de la Fouchardiere, C., Boucher, E., et al. (2014). Gemcitabine and oxaliplatin with or without cetuximab in advanced biliary-tract cancer (BINGO): a randomised, open-label, non-comparative phase 2 trial. *Lancet Oncol.* 15, 819–828. doi: 10.1016/S1470-2045(14)70212-8

McLean, L., and Patel, T. (2006). Racial and ethnic variations in the epidemiology of intrahepatic cholangiocarcinoma in the United States. *Liver Int.* 26, 1047–1053. doi: 10.1111/j.1478-3231.2006.01350.x

Moeini, A., Sia, D., Bardeesy, N., Mazzaferro, V., and Llovet, J. M. (2016). Molecular pathogenesis and targeted therapies for intrahepatic cholangiocarcinoma. *Clin. Cancer Res.* 22, 291–300. doi: 10.1158/1078-0432.CCR-14-3296

Müller, S., and Diaz, A. (2017). Single-cell mRNA sequencing in cancer research: integrating the genomic fingerprint. *Front. Genet.* 8:73. doi: 10.3389/fgene.2017.00073

Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., et al. (2015). Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. *Sci. Rep.* 5:16294. doi: 10.1038/srep16294

Nakamura, H., Arai, Y., Totoki, Y., Shirota, T., Elzawahry, A., Kato, M., et al. (2015). Genomic spectra of biliary tract cancer. *Nat. Genet.* 47, 1003–1010. doi: 10.1038/ng.3375

Nawy, T. (2014). Single-cell sequencing. *Nat. Methods* 11:18. doi: 10.1038/nmeth.2771

Ojima, H., Yoshikawa, D., Ino, Y., Shimizu, H., Miyamoto, M., Kokubu, A., et al. (2010). Establishment of six new human biliary tract carcinoma cell lines and identification of MAGEH1 as a candidate biomarker for predicting the efficacy of gemcitabine treatment. *Cancer Sci.* 101, 882–888. doi: 10.1111/j.1349-7006.2009.01462.x

Ong, C. K., Subimerb, C., Pairojkul, C., Wongkham, S., Cutcutache, I., Yu, W., et al. (2012). Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nat. Genet.* 44, 690–693. doi: 10.1038/ng.2273

Palmer, W. C., and Patel, T. (2012). Are common factors involved in the pathogenesis of primary liver cancers? a meta-analysis of risk factors for intrahepatic cholangiocarcinoma. *J. Hepatol.* 57, 69–76. doi: 10.1016/j.jhep.2012.02.022

Papadopoulos, K. P., El-Rayes, B. F., Tolcher, A. W., Patnaik, A., Rasco, D. W., Harvey, R. D., et al. (2017). A phase 1 study of ARQ 087, an oral pan-FGFR inhibitor in patients with advanced solid tumours. *Br. J. Cancer*. 117, 1592–1599. doi: 10.1038/bjc.2017.330

Perumal, V., Wang, J., Thuluvath, P., Choti, M., and Torbenson, M. (2006). Hepatitis, C., and hepatitis B nucleic acids are present in intrahepatic cholangiocarcinomas from the United States. *Hum. Pathol.* 37, 1211–1216. doi: 10.1016/j.humpath.2006.04.012

Ren, H. B., Yu, T., Liu, C., and Li, Y. Q. (2011). Diabetes mellitus and increased risk of biliary tract cancer: systematic review and meta-analysis. *Cancer Causes Control* 22, 837–847. doi: 10.1007/s10552-011-9754-3

Robinson, D. R., Wu, Y. M., Lonigro, R. J., Vats, P., Cobain, E., Everett, J., et al. (2017). Integrative clinical genomics of metastatic cancer. *Nature* 548, 297–303. doi: 10.1038/nature23306

Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., et al. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 556, 457–462. doi: 10.1038/s41586-018-0024-3

Ross, J. S., Wang, K., Gay, L., Al-Rohil, R., Rand, J. V., Jones, D. M., et al. (2014). New routes to targeted therapy of intrahepatic cholangiocarcinomas revealed by next-generation sequencing. *Oncologist* 19, 235–242. doi: 10.1634/theoncologist.2013-0352

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: from vision to reality. *Nature* 550, 451–453. doi: 10.1038/550451a

Sampaziotis, F., Justin, A. W., Tysoe, O. C., Sawiak, S., Godfrey, E. M., Upponi, S. S., et al. (2017). Reconstruction of the mouse extrahepatic biliary tree using primary human extrahepatic cholangiocyte organoids. *Nat. Med.* 23, 954–963. doi: 10.1038/nm.4360

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321.e310–337.e310. doi: 10.1016/j.cell.2018.03.035

Savic, L. J., Chapiro, J., and Geschwind, J.-F. H. (2017). Intra-arterial embolotherapy for intrahepatic cholangiocarcinoma: update and future prospects. *Hepatobiliary Surg. Nutr.* 6, 7–21. doi: 10.21037/hbsn.2016.11.02

Schwartz, R., and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18, 213–229. doi: 10.1038/nrg.2016.170

Sekiya, S., and Suzuki, A. (2012). Intrahepatic cholangiocarcinoma can arise from Notch-mediated conversion of hepatocytes. *J. Clin. Invest.* 122, 3914–3918. doi: 10.1172/JCI63065

Sempoux, C., Jibara, G., Ward, S. C., Fan, C., Qin, L., Roayaie, S., et al. (2011). Intrahepatic cholangiocarcinoma: new insights in pathology. *Semin. Liver Dis.* 31, 49–60. doi: 10.1055/s-0031-1272839

Sharma, A., Dwary, A. D., Mohanti, B. K., Deo, S. V., Pal, S., Sreenivas, V., et al. (2010). Best supportive care compared with chemotherapy for unresectable gall bladder cancer: a randomized controlled study. *J. Clin. Oncol.* 28, 4581–4586. doi: 10.1200/JCO.2010.29.3605

Shimizu, Y., Demetris, A. J., Gollin, S. M., Storto, P. D., Bedford, H. M., Altarac, S., et al. (1992). Two new human cholangiocarcinoma cell lines and their cytogenetics and responses to growth factors, hormones, cytokines or immunologic effector cells. *Int. J. Cancer* 52, 252–260. doi: 10.1002/ijc.2910520217

Shin, H. R., Oh, J. K., Masuyer, E., Curado, M. P., Bouvard, V., Fang, Y., et al. (2010a). Comparison of incidence of intrahepatic and extrahepatic cholangiocarcinoma–focus on East and South-Eastern Asia. *Asian Pac. J. Cancer Prev.* 11, 1159–1166.

Shin, H. R., Oh, J. K., Masuyer, E., Curado, M. P., Bouvard, V., Fang, Y. Y., et al. (2010b). Epidemiology of cholangiocarcinoma: an update focusing on risk factors. *Cancer Sci.* 101, 579–585. doi: 10.1111/j.1349-7006.2009.01458.x

Sia, D., Hoshida, Y., Villanueva, A., Roayaie, S., Ferrer, J., Tabak, B., et al. (2013a). Integrative molecular analysis of intrahepatic cholangiocarcinoma reveals 2 classes that have different outcomes. *Gastroenterology* 144, 829–840. doi: 10.1053/j.gastro.2013.01.00

Sia, D., Losic, B., Moeini, A., Cabellos, L., Hao, K., Revill, K., et al. (2015). Massive parallel sequencing uncovers actionable FGFR2–PPHLN1 fusion and ARAF mutations in intrahepatic cholangiocarcinoma. *Nat. Commun.* 6:6087. doi: 10.1038/ncomms7087

Sia, D., Tovar, V., Moeini, A., and Llovet, J. M. (2013b). Intrahepatic cholangiocarcinoma: pathogenesis and rationale for molecular therapies. *Oncogene* 32, 4861–4870. doi: 10.1038/onc.2012.617

Sia, D., Villanueva, A., Friedman, S. L., and Llovet, J. M. (2017). Liver cancer cell of origin, molecular class, and effects on patient prognosis. *Gastroenterology* 152, 745–761. doi: 10.1053/j.gastro.2016.11.048

Sirisinha, S., Tengchaisri, T., Boonpucknavig, S., Prempracha, N., Ratanarapee, S., and Pausawasdi, A. (1991). Establishment and characterization of a cholangiocarcinoma cell line from a Thai patient with intrahepatic bile duct cancer. *Asian Pac. J. Allergy Immunol.* 9, 153–157.

Sripa, B., Kaewkes, S., Sithithaworn, P., Mairiang, E., Laha, T., Smout, M., et al. (2007). Liver fluke induces cholangiocarcinoma. *PLoS Med.* 4:e201. doi: 10.1371/journal.pmed.0040201

Sripa, B., and Pairojkul, C. (2008). Cholangiocarcinoma: lessons from Thailand. *Curr. Opin. Gastroenterol.* 24, 349–356. doi: 10.1097/MOG.0b013e3282fbf9b3

Srisomsap, C., Sawangareetrakul, P., Subhasitanont, P., Chokchaichamnankit, D., Chiablaem, K., Bhudhisawasdi, V., et al. (2010). Proteomic studies of cholangiocarcinoma and hepatocellular carcinoma cell secretomes. *J. Biomed. Biotechnol.* 2010:437143. doi: 10.1155/2010/437143

Srisomsap, C., Sawangareetrakul, P., Subhasitanont, P., Panichakul, T., Keeratichamroen, S., Lirdprapamongkol, K., et al. (2004). Proteomic analysis of cholangiocarcinoma cell line. *Proteomics* 4, 1135–1144. doi: 10.1002/pmic.200300651

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi: 10.1126/science.aaf2403

Storto, P. D., Saidman, S. L., Demetris, A. J., Letessier, E., Whiteside, T. L., and Gollin, S. M. (1990). Chromosomal breakpoints in cholangiocarcinoma cell lines. *Genes Chromosomes Cancer* 2, 300–310. doi: 10.1002/gcc.2870020408

Theise, N. D., Nakashima,O., Park, Y. N., and Nakanuma, Y. (2010). "Combined hepatocellular-cholangiocarcinoma," in *Who Classification of Tumours of the Digestive System,* eds F. T. Bosman, F. Garneiro, R. H. Hruban, N. D. Theise (Lyon: IARC), 225–227.

Tyson, G. L., and El-Serag, H. B. (2011). Risk factors of cholangiocarcinoma. *Hepatology (Baltimore, Md.)* 54, 173–184. doi: 10.1002/hep.24351

Uthaisar, K., Vaeteewoottacharn, K., Seubwai, W., Talabnin, C., Sawanyawisuth, K., Obchoei, S., et al. (2016). Establishment and characterization of a novel human cholangiocarcinoma cell line with high metastatic activity. *Oncol. Rep.* 36, 1435–1446. doi: 10.3892/or.2016.4974

Valle, J., Wasan, H., Palmer, D. H., Cunningham, D., Anthoney, A., Maraveyas, A., et al. (2010). Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *New Engl. J. Med.* 362, 1273–1281. doi: 10.1056/NEJMoa0908721

Valle, J. W., Borbath, I., Khan, S. A., Huguet, F., Gruenberger, T., Arnold, D., and ESMO Guidelines Committee. (2016). Biliary cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 27(Suppl. 5), v28–v37. doi: 10.1093/annonc/mdw324

Vijgen, S., Terris, B., and Rubbia-Brandt, L. (2017). Pathology of intrahepatic cholangiocarcinoma. *Hepatobiliary Surg. Nutr.* 6, 22–34. doi: 10.21037/hbsn.2016.11.04

Walter, D., Döring, C., Feldhahn, M., Battke, F., Hartmann, S., Winkelmann, R., et al. (2017). Intratumoral heterogeneity of intrahepatic cholangiocarcinoma. *Oncotarget* 8, 14957–14968. doi: 10.18632/oncotarget.14844

Wang, A., Wu, L., Lin, J., Han, L., Bian, J., Wu, Y., et al. (2018). Whole-exome sequencing reveals the origin and evolution of hepato-cholangiocarcinoma. *Nat. Commun.* 9:894. doi: 10.1038/s41467-018-03276-y

Wang, J., Xie, H., Ling, Q., Lu, D., Lv, Z., Zhuang, R., et al. (2016). Coding-noncoding gene expression in intrahepatic cholangiocarcinoma. *Transl. Res.* 168, 107–121. doi: 10.1016/j.trsl.2015.07.007

Weber, J., Öllinger, R., Friedrich, M., Ehmer, U., Barenboim, M., Steiger, K., et al. (2015). CRISPR/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13982–13987. doi: 10.1073/pnas.1512392112

Weber, S. M., Ribero, D., O'Reilly, E. M., Kokudo, N., Miyazaki, M., and Pawlik, T. M. (2015). Intrahepatic cholangiocarcinoma: expert consensus statement. *HPB* 17, 669–680. doi: 10.1111/hpb.12441

Wu, Y.-M., Su, F., Kalyana-Sundaram, S., Khazanov, N., Ateeq, B., Cao, X., et al. (2013). Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discov.* 3, 636–647. doi: 10.1158/2159-8290.CD-13-0050

Yamaguchi, N., Morioka, H., Ohkura, H., Hirohashi, S., and Kawai, K. (1985). Establishment and characterization of the human cholangiocarcinoma cell line HChol-Y1 in a serum-free, chemically defined medium. *J. Natl. Cancer Inst.* 75, 29–35.

Yamamoto, S., Kubo, S., Hai, S., Uenishi, T., Yamamoto, T., Shuto, T., et al. (2004). Hepatitis C virus infection as a likely etiology of intrahepatic cholangiocarcinoma. *Cancer Sci.* 95, 592–595. doi: 10.1111/j.1349-7006.2004.tb02492.x

Yamasaki, S. (2003). Intrahepatic cholangiocarcinoma: macroscopic type and stage classification. *J. Hepatobiliary. Pancreat. Surg.* 10, 288–291. doi: 10.1007/s00534-002-0732-8

Yang, W., Li, Y., Song, X., Xu, J., and Xie, J. (2017). Genome-wide analysis of long noncoding RNA and mRNA co-expression profile in intrahepatic cholangiocarcinoma tissue by RNA sequencing. *Oncotarget* 8, 26591–26599. doi: 10.18632/oncotarget.15721

Yano, H., Iemura, A., Haramaki, M., Momosaki, S., Sachiko, O., Koichi, H., et al. (1996). A human combined hepatocellular and cholangiocarcinoma cell line (KMCH-2) that shows the features of hepatocellular carcinoma or cholangiocarcinoma under different growth conditions. *J. Hepatol.* 24, 413–422. doi: 10.1016/S0168-8278(96)80161-9

Yi, J. H., Thongprasert, S., Lee, J., Doval, D. C., Park, S. H., Park, J. O., et al. (2012). A phase II study of sunitinib as a second-line treatment in advanced biliary tract carcinoma: a multicentre, multinational study. *Eur. J. Cancer* 48, 196–201. doi: 10.1016/j.ejca.2011.11.017

Yu, T. H., Yuan, R. H., Chen, Y. L., Yang, W. C., Hsu, H. C., and Jeng, Y. M. (2011). Viral hepatitis is associated with intrahepatic cholangiocarcinoma with cholangiolar differentiation and N-cadherin expression. *Mod. Pathol.* 24, 810–819. doi: 10.1038/modpathol.2011.41

Zheng, S., Zhu, Y., Zhao, Z., Wu, Z., Okanurak, K., and Lv, Z. (2017). Liver fluke infection and cholangiocarcinoma: a review. *Parasitol. Res.* 116, 11–19. doi: 10.1007/s00436-016-5276-y

Zhou, H., Wang, H., Zhou, D., Wang, H., Wang, Q., Zou, S., et al. (2010). Hepatitis B virus-associated intrahepatic cholangiocarcinoma and hepatocellular carcinoma may hold common disease process for carcinogenesis. *Eur. J. Cancer* 46, 1056–1061. doi: 10.1016/j.ejca.2010.02.005

Zhou, Y., Zhao, Y., Li, B., Huang, J., Wu, L., Xu, D., et al. (2012). Hepatitis viruses infection and risk of intrahepatic cholangiocarcinoma: evidence from a meta-analysis. *BMC Cancer* 12:289. doi: 10.1186/1471-2407-12-289

Zhu, A. X., Meyerhardt, J. A., Blaszkowsky, L. S., Kambadakone, A. R., Muzikansky, A., Zheng, H., et al. (2010). Efficacy and safety of gemcitabine, oxaliplatin, and bevacizumab in advanced biliary-tract cancers and correlation of changes in 18-fluorodeoxyglucose PET with clinical outcome: a phase 2 study. *Lancet Oncol.* 11, 48–54. doi: 10.1016/S1470-2045(09)70333-X

Zou, S., Li, J., Zhou, H., Frech, C., Jiang, X., Chu, J. S., et al. (2014). Mutational landscape of intrahepatic cholangiocarcinoma. *Nat. Commun.* 5:5696. doi: 10.1038/ncomms6696

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 16 August 2016 | Revised: 13 January 2017 | Accepted: 16 January 2017

DOI 10.1002/gcc.22447

RESEARCH ARTICLE

WILEY

Alterations of type II classical cadherin, cadherin-10 (CDH10), is associated with pancreatic ductal adenocarcinomas

Natini Jinawath[1,2*] | Meng-Shin Shiao[2] | Alexis Norris[3†] | Kathleen Murphy[3‡] | Alison P. Klein[3,4] | Raluca Yonescu[3] | Christine Iacobuzio-Donahue[3] | Alan Meeker[3] | Artit Jinawath[5] | Charles J. Yeo[6] | James R. Eshleman[3,4] | Ralph H. Hruban[3,4] | Jonathan R. Brody[6] | Constance A. Griffin[1,2,4] | Shuko Harada[3]

[1]Institute of Genetic Medicine, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA

[2]Research Center, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[3]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA

[4]Sidney Kimmel Cancer Center Johns Hopkins Medical Institutions, Baltimore, Maryland, USA

[5]Department of Pathology, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[6]Department of Surgery, Jefferson Center for Pancreatic, Biliary and Related Cancers, Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

Correspondence
Shuko Harada, Department of Pathology, University of Alabama at Birmingham, 619 19th St South, NP3540, Birmingham, AL 35249, USA.
Email: sharada@uabmc.edu

Abstract

Pancreatic ductal adenocarcinoma (PDAC), either sporadic or familial, has a dismal prognosis and finding candidate genes involved in development of the cancer is crucial for the patient care. First, we identified two patients with germline alterations in or adjacent to *CDH10* by chromosome studies and sequencing analyses in 41 familial pancreatic cancer (FPC) cases. One patient had a balanced translocation between chromosome 5 and 20. The breakpoint on chromosome band 5p14.2 was ~810 Kb upstream of *CDH10*, while that on chromosome arm 20p was in the pericentromeric region which might result in inactivation of one copy of the gene leading to reduced expression of *CDH10*. This interpretation was supported by loss of heterozygosity (LOH) seen in this region as determined by short tandem repeat analyses. Another patient had a single nucleotide variant in exon 12 (p.Arg688Gln) of *CDH10*. This amino acid was conserved among vertebrates and the mutation was predicted to have a pathogenic effect on the protein by several prediction algorithms. Next, we analyzed LOH status in the *CDH10* region in sporadic PDAC and at least 24% of tumors had evidence of LOH. Immunohistochemical stains with CDH10 antibody showed a different staining pattern between normal pancreatic ducts and PDAC. Taken together, our data supports the notion that *CDH10* is involved in sporadic pancreatic carcinogenesis, and might have a role in rare cases of FPC. Further functional studies are needed to elucidate the tumor suppressive role of *CDH10* in pancreatic carcinogenesis.

1 | INTRODUCTION

Pancreatic cancer has a dismal prognosis. The most common type of pancreatic cancer is pancreatic ductal adenocarcinoma (PDAC), which is generally sporadic in origin and accounts for >90% of pancreatic cancers. Up to 10% of pancreatic cancer patients are designated familial pancreatic cancer (FPC),[1] which is defined as a kindred with at least two first-degree relatives affected by pancreatic cancers.[2]

Large-scale genome-wide screening, for example, next generation sequencing and SNP array analysis, provides an extensive and unbiased way to search for susceptibility genes in cancers. Using these approaches, genomic regions showing copy number variations and mutations are often found in PDAC.[3–5] Jones et al.[4] showed that point mutations contributed the most to the genetic alterations in PDAC by analyzing sequences of ~23,000 transcripts in 24 pancreatic cancers using Sanger sequencing. These mutations may change expression of

*Current address: Program in Translational Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand; and Integrative Computational Bioscience Center, Mahidol University, Nakhon Pathom, Thailand

†Current address: Kennedy Krieger Institute, Baltimore, MD, USA

‡Current address: ProPath Services, Dallas, Texas, USA

genes involving in core signaling pathway and regulatory processes, such as DNA damage control (*TP53*), KRAS signaling (*KRAS*), and homophilic cell adhesion (*CDH* gene family). Further whole-exome or whole-genome sequencing studies help identify more genomic aberrations in PDAC, which can be further categorized into four different subtypes—stable, locally rearranged, scattered, and unstable—with potential clinical utility.[3,5]

Although relatively rare, several genes associated with FPC such as *BRCA2*, *BRCA1*, *STK11*, *CDKN2A*, *PALB2*, *ATM*, and mismatch repair genes[6–9] have been identified. Conducting whole-genome sequencing of 638 patients with FPC and exome sequencing of 39 FPC tumor tissues, a recent study further identified more genes carrying private heterozygous premature truncations or deleterious mutations in FPC.[8] However, variants identified in the FPC kindreds are highly heterogeneous—more than 60% of the genes identified only appear in one single FPC kindred. As the risk of developing pancreatic cancer is significantly higher—4.6 to 32 fold—in members of FPC kindreds,[2,10] it remains an important task to understand the genetic underpinning of FPC.

Either familial or sporadic, finding candidate genes that are involved in PDAC tumorigenesis should contribute to the development of early detection biomarkers, offer the opportunities of preventive medicine.

In this study, we first performed chromosome studies in 41 FPC cases followed by direct sequencing of cadherin-10 (*CDH10*) gene, a lesser-known gene that encodes a type II classical cadherin,[11] in all cases. We further carried out the loss of heterozygosity (LOH) analysis of *CDH10* in 50 sporadic PDAC tissues. Our results found alterations of *CDH10* genes and LOH of *CDH10* regions in PDAC tissues, suggesting that *CDH10* may be involved in pancreatic carcinogenesis.

2 | MATERIALS AND METHODS

2.1 | Materials

This study was approved by Johns Hopkins University Institutional Review Board. For FPC cases, DNA from lymphoblastoid cell lines created from individuals in the National Familial Pancreas Tumor Registry (NFPTR; www.nfptr.org) was used. Individuals studied had pancreatic cancer documented by review of the pathology report with the first degree member of the family also having pancreatic cancer (Supporting Information Table S1). Individuals ($n = 39$) or other family members ($n = 2$) had been tested for mutations in *BRCA2* and *PALB2* and they were negative (Supporting Information Table S1).

For sporadic pancreatic carcinoma cases, formalin-fixed, paraffin-embedded (FFPE) tissue from 28 individuals at Johns Hopkins Hospital and snap-frozen fresh tissue from 22 individuals at Thomas Jefferson University were used. Pancreas cancer cell lines including Hs766T, PL45, MIA PaCa-2, Capan-1, and BxPC-3 were used in this study.

Anonymized samples of bone marrow or peripheral blood from individuals serving as bone marrow transplantation donors were used as controls for sequencing.

2.2 | Copy number and SNP array analysis of T(5;20) somatic cell hybrids

Genomic DNAs for SNP array analysis were isolated from two t(5;20) human/mouse somatic cell hybrid clones (human der(5)- and der(20)-retaining clones). Copy number variation was evaluated using Illumina HumanHap 550 arrays. SNP genotyping was performed at the SNP Center of The Johns Hopkins University Genetic Resources Facility. Samples were processed and analyzed according to the Infinium II Assay protocol (Illumina, San Diego, CA) and the processed BeadChips were imaged on an Illumina BeadArray reader. The signal intensity (log *R* ratio) and allelic composition (B allele frequency) of human der(5) and der(20) chromosomes were analyzed by direct observation of the scan data using BeadStudio v.3.0.27. The minimal size of detected aberrant findings was calculated from the base position of the proximal and distal aberrant SNPs based on the UCSC Genome Browser, Human Dec. 2013 assembly (hg38). It is estimated that $<3\%$ of human SNP probes on Illumina array are potentially conserved in rodent. Therefore, the human DNA signals can be clearly separated from those of mouse, and balanced translocation breakpoints can be mapped with high resolution using this method.

2.3 | Genomic DNA and total nucleic acid isolation

Genomic DNAs were isolated from peripheral blood, bone marrow, somatic cell hybrid clones or transformed lymphoblast cells with the Gentra PureGene DNA isolation kit or QIAGEN QIAamp Midi kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. Total nucleic acid was isolated from the FFPE tumor samples using the Agencourt FormaPure kit (Agencourt Bioscience Corp., Beverly, MA) according to the manufacturer's protocol. Isolated DNA and total nucleic acid were quantified by Nanodrop and examined for size by agarose gel electrophoresis.

2.4 | Polymerase chain reaction (PCR) and direct sequencing of CDH10

PCR reactions using primers for *CDH10* gene (Supporting Information Table S2) were carried out in a reaction containing 1× PCR buffer, 0.2 mM dNTP, 1.5 mM $MgCl_2$, 0.25 μM each of forward and reverse primers, 1.25 units Platinum Taq DNA polymerase (Invitrogen, Carlsbad, CA), and 20 ng of DNA in a 20-μL reaction volume. PCR amplification was performed using the ABI9700 and touchdown thermal cycling conditions as follows: 94°C for 2 min; 3 cycles of 94°C for 30 s, 64°C for 30 s, 72°C for 30 s; 3 cycles of 94°C for 30 s, 61°C for 30 s, 72°C for 30 s; 3 cycles of 94°C for 30 s, 58°C for 30 s, 72°C for 30 s; 35 cycles of 94°C for 30 s, 57°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 7 min.

PCR products were purified using QiaQuick reagents (Qiagen, Valencia, CA) or ExoSAP-IT (USB Corp., Cleveland, OH) and were cycle sequenced using Big Dye v3.1 reagents (Applied Biosystems, Foster City, CA) and the standard M13F or M13R sequencing primers according to the manufacturer's protocol. Sequencing products were purified with CleanSEQ Sequencing Purification System (Agencourt Bioscience

Corp., Beverly, MA), and automated sequencing was performed by capillary electrophoresis (CE) on an ABI3700 (Applied Biosystems, Foster City, CA). Sequences were aligned and examined using Sequencher software (Gene Codes, Ann Arbor, MI).

2.5 | Short tandem repeat (STR) analysis

The following six STRs around *CDH10* on chromosome 5 were analyzed: *D5S2845* (5p14.3), *D5S1473* (5p14.2), *D5S813* (5p14.2), *D5S648* (5p14.1), *D5S814* (5p14.1), *D5S419* (5p14.1) (Supporting Information Table S3). No STRs are described within *CDH10* itself. Reactions were individually prepared and thermalcycled according to the PCR protocol described above. After amplification, 2 μL of each PCR product was mixed with 8 μL of deionized formamide/GeneScan 500 [ROX] size standard (Applied Biosystems, Foster City, CA). Samples were denatured at 95°C for 2 min and placed on ice for at least 1 minbefore analyzing on the ABI3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). CE data from the tumor samples and from nontumor control samples were analyzed to identify alleles at each locus and determine the allelic ratios.

2.6 | CDH10 immunohistochemistry (IHC)

Unconjugated Rabbit Anti-Human CDH10, C-Terminus polyclonal antibody was obtained from Abgent, San Diego CA(Cat.# AP1482b). Chromogenic IHC labeling for CDH10 was performed as follows: 5 μm tissue sections of FFPE tissues on charged slides were deparaffinized and rehydrated by sequential 10 min room temperature incubations in xylene, 100% ethanol, 95% ethanol, 70% ethanol, and distilled water. A 60-s immersion in distilled water containing 1% Tween-20 detergent (Sigma-Aldrich, St. Louis, MO, Cat. # P-7949) was followed by heat-induced antigen retrieval. The slides were immersed in an EDTA target retrieval buffer (Invitrogen, Carlsbad, CA, #00-5500) and steamed in a vegetable steamer (Black and Decker Handy Steamer Plus, Black and Decker, Towson, MD) for 45 min. Endogenous peroxidase activity was blocked by 10 min treatment with peroxidase blocking reagent (Dako, Carpinteria, CA, Cat. # S2001). The primary antibody was applied at a dilution of 1:50 diluted in antibody dilution buffer (ChemMate Cat. # ADB250) and incubated 14 h at 4°C. The primary antibody was detected using the Power Vision Plus HRP-polymer detection system (Leica Cat. # PV6119) per manufacturer's instructions. All washing steps utilized Tris Buffered Saline with Tween (TBST; Sigma-Aldrich, Cat. # T-9039). DAB chromogen (Sigma-Aldrich, Cat. # D4293) was applied to develop the secondary detection reagent. Slides were then counter stained with Mayer's hematoxylin (Dako Cat. # S3309), dehydrated and cover slips were mounted.

2.7 | Databases

Single Nucleotide Polymorphism Database (dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/) build 147, 1000 Genomes May 2013 release (http://browser.1000genomes.org/), and the Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org/) were used to identify the minor allele frequency of *CDH10*'s missense mutations. Expression level of *CDH10* in various tissue types was obtained from the Unigene's expressed sequence tag (EST) database, and the Genotype-Tissue Expression project (GTEx Portal, http://www.gtexportal.org/home/).

3 | RESULTS

3.1 | Germline alterations of CDH10 identified in FPC

Through routine karyotyping of lymphoblastoid cell lines of NFPTR enrollees, we identified one individual with a typical pancreatic adenocarcinoma showing a balanced constitutional translocation t(5;20)(p14;p11.1) in all metaphases (Figure 1A). Metaphase fluorescent in-situ hybridization (FISH) using bacterial artificial chromosomes (BACs) obtained from the BACPAC Resource Center (Children's Hospital Oakland, Oakland, CA) were used to further define the breakpoint. Several primer sets (Figure 1D) were designed in the regions covered by the BAC clone that straddled the breakpoint on chromosome 5 (Figure 1B). We further conducted Illumina HumanHap 550 array analysis of the human-mouse somatic cell hybrid clones containing only the derivative chromosomes to identify the translocation breakpoint on chromosome 5 (Figure 1C).

Interestingly, the breakpoint is at 810 kb upstream of (*CDH10*), which is the nearest gene without any microdeletions or duplications in the putative breakpoint regions (Figure 1D). Subsequent sequencing analysis identified no germline *CDH10* mutation in this individual. We further analyzed the genomic stability of the patient's tumor using 6 STR markers on chromosome band 5p14.2. The results demonstrated a region of LOH covering *CDH10* (Figure 3, Familial t(5;20) case), which is commonly seen in cancers. This indicated that *CDH10* might be associated with FPC. However, limited tumor tissue was available from a biopsy of this patient's tumor and no additional materials were available for *CDH10* expression analysis by RT-PCR or IHC.

We further extended our studies by performing karyotyping and all-exon sequencing of germline *CDH10* in 41 FPC cases. *CDH10* germline sequence changes were also analyzed in 106 deidentified normal bone marrow donors. *CDH10* polymorphisms were checked against public variant databases (dbSNP build 147, 1000 Genomes May 2013 release) before reporting as a possible novel variant.

Among 41 patients, we observed a patient who demonstrated a germline G > A transition base substitution at coding DNA position 2063 resulting in an amino acid arginine (R) to glutamine (Q) substitution at amino acid codon 688 (p.R688Q) (Figure 2A). This alteration occurs in a highly conserved cadherin cytoplasmic domain in exon 12. The amino acid is conserved among vertebrates (Figure 2B). To be noted, this patient had two first degree relatives who had developed pancreatic cancer. Tumor tissue was not available from this individual for analysis. In contrast, no nonsynonymous changes in exon 12 were found among the 106 control samples or in public databases (dbSNP build 147, 1000 Genomes May 2013 release; see more detail in discussion). Although one synonymous alteration in coding sequence of exon 12 that had not been previously described was found in two control individuals, this variant was less likely to cause any deleterious effect

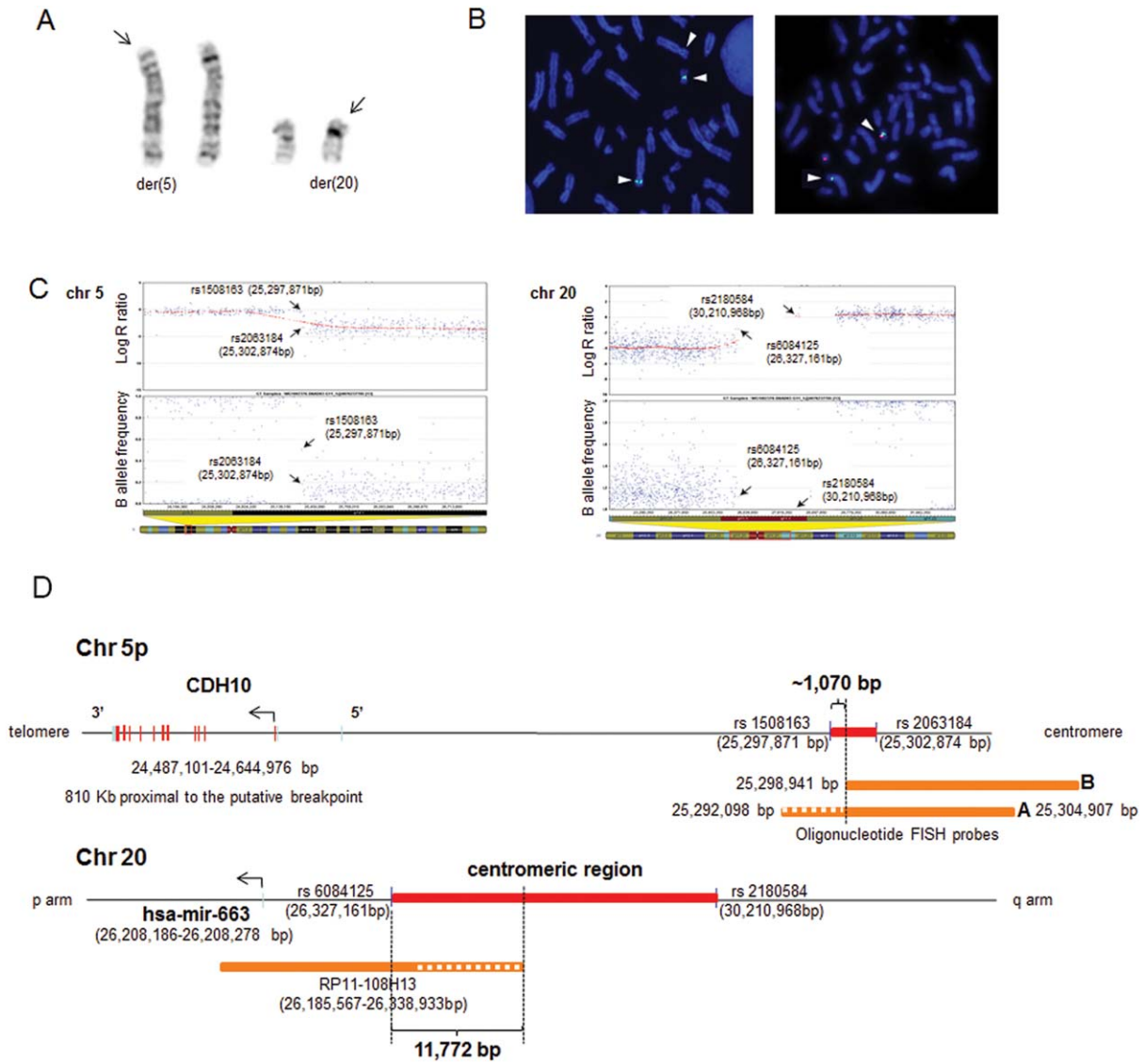


FIGURE 1 Identification of the translocation breakpoint in a FPC patient with t(5;20)(p14;p11.1). **(A)** The G-banded chromosomes 5 and 20 from a lymphoblastoid cell line of the FPC patient showing an apparently balanced translocation t(5;20)(p14;p11.1) (black arrows). der(5), derivative chromosome 5; der(20), derivative chromosome 20. **(B)** FISH analysis of the translocation breakpoint. Left panel: Oligonucleotide probes PCR-amplified from RP11-184E9 (chromosome band 5p14) shows green split signals (white arrows) between chromosome arms 5p and 20p. Right panel: BAC FISH analysis showing the split signals (white arrows) of RP11-108H13 (chromosome band 20q11.1; green) on 20p and 5p, while those of RP11-348i14 (chromosome band 20q11.1; orange) are intact. **(C)** Illumina HumanHap550 genotyping beadchip analysis of the human-mouse somatic cell hybrids containing a copy of der(20). Log R ratio and B allele frequency plots demonstrating the translocation breakpoints on chromosomes 5 (left panel) and 20 (right panel). The breakpoint flanking SNPs are also shown (black arrows). Both der(5) and der(20)-containing somatic hybrids were analyzed with SNP array, but this figure only shows der(20) cell results. **(D)** A schema illustrating the translocation breakpoint regions on chromosomes 5 and 20 identified using a combination of FISH, somatic cell hybrid coupled with SNP array analysis. The nearest genes to the putative breakpoints on chromosome bands 5p14 and 20p11.1 are *CDH10* and *hsa-miR-663*, respectively. Genomic locations shown are based on the UCSC genome browser (hg38). [Color figure can be viewed at wileyonlinelibrary.com]

on *CDH10* function (DNA position 2019 C > T). These results further suggest that alteration of *CDH10* gene may be related to small portion of FPC cases.

To investigate whether the amino acid change at the position 688 affects the protein function, we performed predictions of the pathogenic effect of this variant using MutationTaster2,[12] PolyPhen-2,[13] and SIFT.[14] As expected, all three algorithms suggested pathogenic effect of this variant with disease causing probabilities 1.00 (MutationTaster2), 0.982 (PolyPhen-2), and 100% (SIFT).

3.2 | Somatic alterations in *CDH10* gene region

We further examined if somatic alteration in *CDH10* gene region is present in sporadic PDAC, and analyzed LOH status using six microsatellite

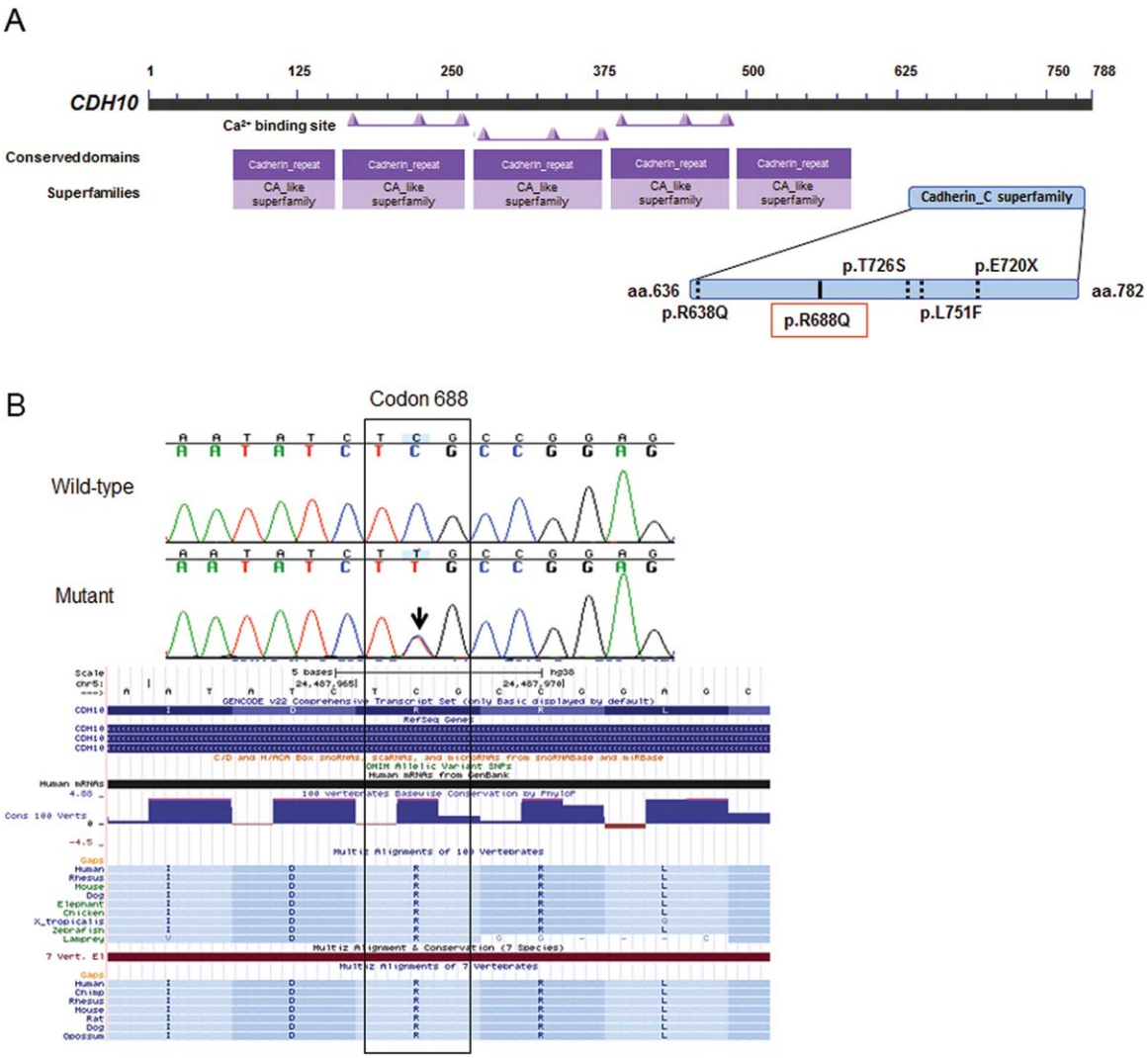


FIGURE 2 *CDH10* single nucleotide changes identified in the FPC and sporadic PDAC cases. (A) A diagram of CDH10 protein showing the locations of conserved domains and related domain superfamily as predicted by NCBI conserved domains. All previously reported *CDH10* somatic mutations (dotted line) and a novel germline variant (p.R688Q; solid line) identified in our study are clustered in exon 12 (Cadherin_C superfamily), which corresponds to the predicted cadherin cytoplasmic superfamily region. **(B)** Comparison between conserved amino acids of the existing cadherin cytoplasmic domains across vertebrate species. R688 is among the highly conserved amino acid component of this domain. Genomic locations shown are based on the UCSC genome browser (hg38). [Color figure can be viewed at wileyonlinelibrary.com]

markers surrounding *CDH10*. DNA isolated from 28 microdissected FFPE PDACs and 22 grossly dissected fresh frozen PDACs were analyzed to identify somatic alterations. For LOH analysis, we included one sample of familial tumor for comparisons. One tumor (1/51, 2%) demonstrated an A > T transition at 2176, which resulted in amino acid threonine (T) to serine (S) substitution at codon 726 (p.T726S; mutation id: COSM84892). We subsequently confirmed that this mutation was identified in the tumor cell line from the same individual.[4] Additional mutations were not identified.

Next, we examined LOH status of a total of 51 samples including 50 sporadic tumors and one familial tumor, which was included for comparisons. Twelve cases (24%) including 11 FFPE samples and 1 fresh PDAC sample demonstrated LOH at one or more of the markers most proximal to *CDH10*, that is, *D5S813* or *D5S648* (Figure 3 and Table 1). The remaining 38 sporadic cases (76%) demonstrated no definitive evidence of LOH at a locus adjacent to *CDH10*, whereas the familial tumor from the patient with a constitutional t(5;20) demonstrated LOH at a locus adjacent to *CDH10*. Among the 38 cases without evidence of LOH in the two loci, 7 of them showed LOH in at least one of the 4 loci analyzed (14%; Locus *D5S2845*, *D5S1473*, *D5S814*, and *D5S419*).

We noted that STR marker *D5S648*, the closest upstream marker to *CDH10*, was only informative in 13 out of the 50 tumors tested (26%; Figure 3). The heterozygosity scores for this STR marker were shown to be similar to *D5S813*, which are 0.74 and 0.75, respectively. However, STR *D5S813* was informative (heterozygous) in 64% of the tumors (32/50).

In addition, DNA microdissected from FFPE specimens demonstrated a significantly higher frequency of LOH compared to DNA isolated from

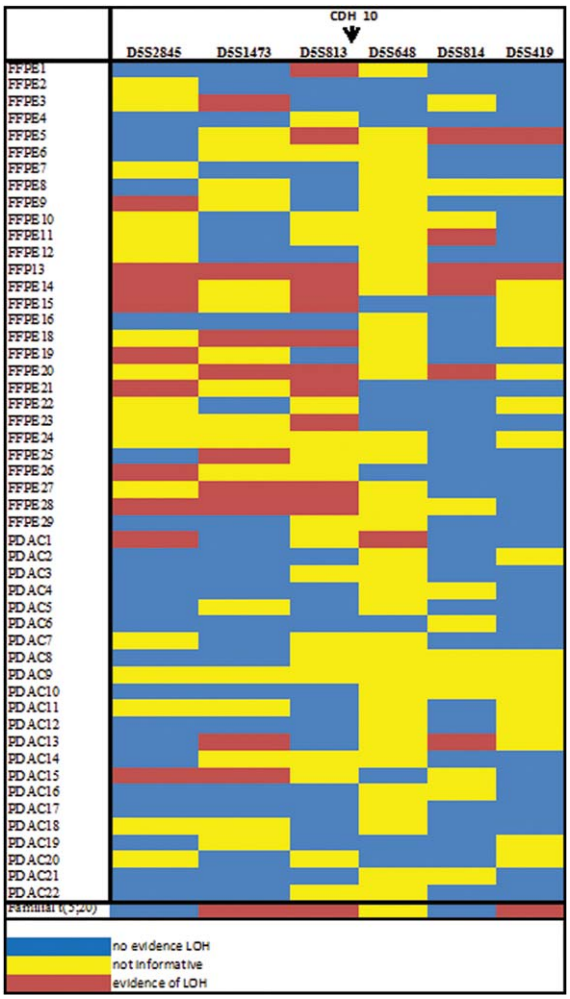


FIGURE 3 Summary of LOH analysis of *CDH10* gene region in sporadic PDACs. LOH analysis using 6 STR markers on chromosome 5p14.1-5p14.3 in 28 FFPE samples from Johns Hopkins Hospital and 22 fresh frozen tumor specimens from Thomas Jefferson University. Additionally, one FPC with t(5;20) was also analyzed. *CDH10* gene is located between *D5S813* and *D5S648* (indicated by arrow). Red = LOH, Blue = not compatible with LOH, NI = not informative. [Color figure can be viewed at wileyonlinelibrary.com]

fresh tumor specimens (11/28, 39.3% vs. 1/22, 4.5%, respectively). This raises the possibility that some of the DNA isolated from fresh frozen tumors may have contained a low percentage of tumor cells, which made it inadequate for LOH detection.

Because of this concern, we analyzed DNA from each tumor specimen for oncogene *KRAS* codons 12 and 13 mutations as evidence of tumor content. DNA isolated from FFPE has a significantly higher mutant allele percentage compared to DNA isolated from fresh tissue [39.68 ± 4.20% and 21.12 ± 3.26% (mean ± SEM), respectively, $P = .026$; Table 1). In general, a tumor percentage >30% is needed to identify LOH. Although a low or negative *KRAS* result is not definitive for a low percentage of tumor DNA present in the sample, our results suggest that some of our DNA specimens may have had an inadequate percentage of tumor cells to detect LOH (or mutations) and actual frequency of LOH in this region may be higher.

3.3 | Localization of CDH10 protein in PDAC

Lastly, we examined CDH10 protein expression in PDACs in order to determine whether CDH10 expression and distribution is altered. Immunohistochemical stains (IHC) with anti-CDH10 were performed on 31 sporadic PDACs on a tissue microarray with two normal pancreatic tissues as a control. The polyclonal antibody we used targets the C-terminal region of CDH10 corresponding to the cytoplasmic domain region. The two normal pancreas tissue samples demonstrated a staining pattern consistent with localization of the protein to gap/tight junctions (Figure 4A,B). In contrast, very weak cytoplasmic staining was observed with CDH10 in PDAC specimens (Figure 4C,D). Expression of CDH10 in PDAC specimens was greatly reduced compared to that in normal pancreas. IHC was also attempted with another polyclonal antibody to CDH10 (Sigma HPA010651) but this antibody failed to stain normal controls. These results suggest that decreased CDH10 expression in PDAC possibly correlates with LOH in *CDH10* gene region.

4 | DISCUSSION

Abnormal cadherin expression has been associated with a large spectrum of disease, including metastatic cancer (Berx and van Roy, 2009). Members of the cadherin superfamily are increasingly shown to have a defining role in cancer. Its best-known member, E-cadherin, has been shown to suppress invasion and metastasis, and germline mutations in this gene causes an autosomal dominant predisposition to diffuse gastric cancer and early onset breast cancer.[15,16] *CDH10* was found to be one of a number of genetically altered cadherins involved in the homophilic cell adhesion pathway in pancreatic cancer, where it was considered a driver mutation.[4] We hypothesized that *CDH10* may play a role in the development of pancreatic cancer. Here we identified two novel germline alterations in individuals with FPC, who were negative for mutations in other known pancreatic cancer risk/causative genes (*BRCA2* and *PALB2*). Of note, one of two cases had variant of uncertain significance (VUS) in *BRCA1* and *FANCA* genes (data not shown), and significance of these alterations in pathogenesis is not known. Roberts

TABLE 1 Summary of LOH and *KRAS* mutation analysis in sporadic pancreatic cancer cases from formalin fixed paraffin embedded tissue (FFPE PDAC) and fresh tissue (Fresh PDAC)

	FFPE PDAC	Fresh PDAC	Total
Number of cases	28	22	50
LOH at loci adjacent to CDH10 (*D5S813* and/or *D5S648*)	11 (39.5%)	1 (4.5%)	12 (24%)
Suspicious for LOH	5 (10.7%)	2 (18.2%)	5 (10%)
KRAS mutation	24 (85.7%)	17 (77.3%)	41 (82%)
% of mutant allele (mean ± SEM)	39.68 ± 4.2	21.12 ± 3.3	32.2 ± 3.12

Suspicious for LOH: at least one loci showing LOH in another four loci if both *D5S813* and *D5S648* are not informative.

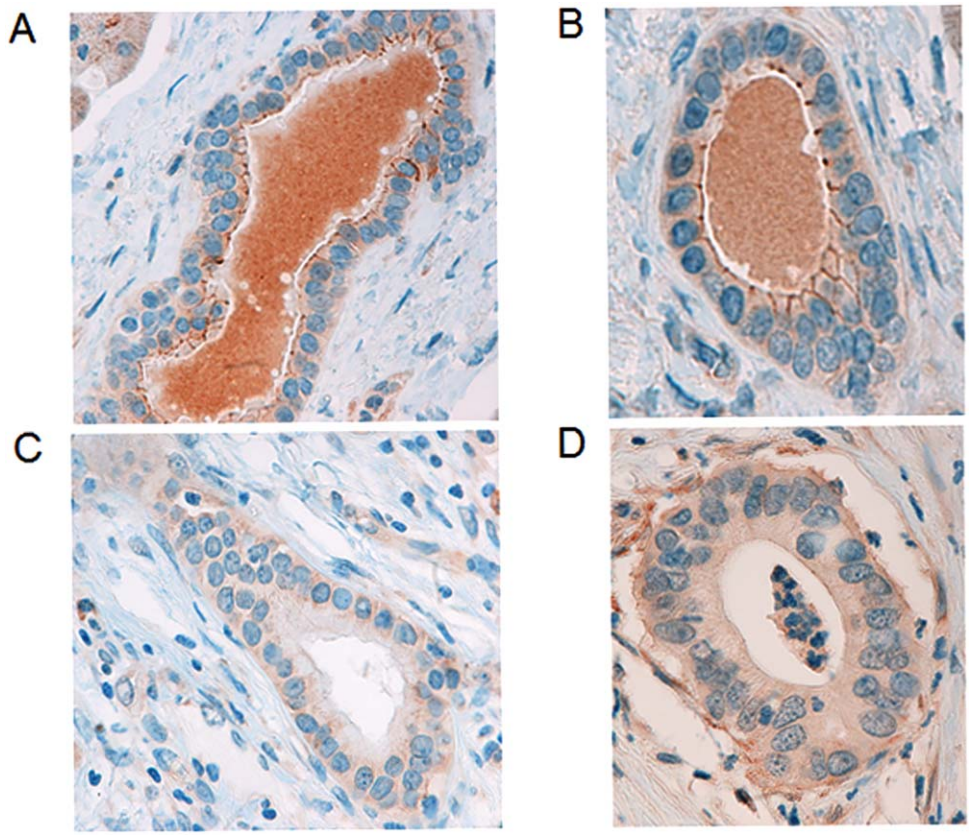


FIGURE 4 Immunohistochemical stains with CDH10 antibody on normal pancreatic ducts (A and B) and PDACs (C and D). Images were taken at 200× (A and C) and 400× (B and D) original magnification. [Color figure can be viewed at wileyonlinelibrary.com]

et al.[8] showed that more than half of the germline alterations in FPC were only found in one kindred, indicating that most FPC-associated variants are very heterogeneous and only contribute to a small number of cases. It is known that whole genome sequencing approach is more technically challenging to identify structural variation. In addition, discovering structural variants especially in primary tumors is difficult due to the artifact created by contamination of normal stromal cells and lymphocytes.

CDH10 has been proposed to be one of the driver mutation genes in sporadic PDACs.[4] Additionally, cadherins are important in cell-cell adhesion, and are known to function in cell recognition, coordinated cell movement, and inducing and maintaining both structural and functional cell and tissue polarity. Abnormal expression of cadherins, such as the most well-known member E-cadherin, often results in increased tumor cell invasion, which ultimately leads to metastasis of tumors.[17] *CDH10* is a lesser known gene that encodes a type II classical cadherin, which is defined based on the lack of a HAV cell adhesion recognition sequence specific to type I cadherins.[11] It is predominantly expressed in brain and is putatively involved in synaptic adhesions and in axon outgrowth and guidance.[18,19] Recent studies further suggest that somatic mutations in *CDH10* are associated with colorectal cancer, gastric cancer, and lung cancer.[20–22] Therefore, it is of interest to clarify the association between genetic alterations in *CDH10* and pancreatic cancer.

In our study, germline alterations in or nearby *CDH10* were observed in 2 out of 41 FPC patients using a combination of classical chromosome analysis and direct sequencing in order to precisely capture the multifaceted genetic alterations. First, we identified a patient with FPC who has a balanced translocation between chromosome 5 and 20 by analyzing karyotypes of participants in National Familial Pancreas Tumor Registry (NFPTR). We narrowed the breakpoints of the constitutional t(5;20)(p14.2p11.1) using FISH with BAC and long-range PCR probes, somatic cell hybrids and SNP array to a region 810 Kb from *CDH10* on chromosome 5 and in the pericentromeric region of chromosome 20. We postulate that *CDH10* expression has been affected by proximity to the pericentromeric region of chromosome 20. Genomic rearrangements that put euchromatic genes near the heterochromatin of a centromere often result in gene inactivation due to the proximity of heterochromatin.[23] This effect, known as position effect variegation, has been recognized since 1930, although direct demonstration of this effect in humans has been limited, possibly due to the difficulty of cloning breakpoints that involve heterochromatin. There are identified examples in humans of cis-acting elements that mediate their effect on gene regulation over large genomic distances. Individuals with campomelic dysplasia without a mutation in the coding region of *SOX9*, a gene known to cause the syndrome, have been found; several have translocations with breakpoints up to 932 kb upstream of the gene.[24] Other examples include a noncoding sequence

~10 kb downstream of the promoter of *RET* in Hirschsprung disease, and a SNP within a conserved noncoding sequence 10 kb upstream from the promoter of *IRF6* in Van der Woude syndrome (reviewed in Ref. 25).

The second germline alteration, a missense variant, has previously been reported as somatic mutation in lung adenocarcinoma.[26,27] This missense variant (p.R688Q; mutation id: COSM738261) occurs in the predicted cadherin cytoplasmic superfamily region. Therefore, while it is difficult to predict exactly what effect this missense variant would exert on the function of CDH10, comparison between the conserved amino acids of the existing cadherin cytoplasmic domains in various cadherin genes shows that R688 is among the highly conserved amino acid sequences of this domain. This suggests that the amino acid substitution may be significant. Although further functional analysis of CDH10 is required to elucidate this, all three predictions algorithms suggested the variant likely has a pathogenic effect on the protein. Additionally, the allele frequency of this missense variant is reported to be 0.00000826 in ExAC database version 0.3.1 (http://exac.broadinstitute.org/variant/5-24488076-C-T) suggesting that it may be a rare pathogenic variant. However, a recent whole genome sequencing study by Roberts et al. on 638 FPC cases and 967 controls identified only one premature truncating variant in *CDH10*. Hence, *CDH10* alterations we identified may be an uncommon cause of FPC.

Next, we studied a possible role of somatic alterations of *CDH10* in sporadic pancreatic tumors. We identified one mutation and a significant number of tumors with LOH in the region. Our analysis may underestimate LOH for several reasons. First, there may not have been enough tumor cell percentage in the specimens analyzed to accurately identify LOH. Second, the marker closest to *CDH10* (*D5S648*) had a much lower heterozygosity than expected. Since normal tissue from these tumors was not available for comparison, it is possible that we have underestimated the frequency of LOH at the *D5S648* locus. This speculation is supported by the result of LOH status of PDAC cell lines: 4 out of the 5 cell lines showed noninformative in locus *D5S648* (Supporting Information Figure S1). Since cell lines are pure populations of tumor cells, there is high frequency of "noninformative" allele, which could be a result of LOH. Third, FFPE tissues have degraded DNA, and 3 of six markers (*D5S1473*, *D5S813*, and *D5S419*) had relatively long allele lengths (>200 bp), causing variations in the ratios we observed in our normals and therefore a relatively wide range of ratios consistent with no loss. Nevertheless, our data indicate that *CDH10* alteration is seen in both familial and sporadic pancreatic cancer patients, which is consistent with recent observations showing FPC undergoes a similar somatic molecular pathogenesis as sporadic PDAC.[1]

Unigene's EST profiles show that *CDH10* is highly expressed in brain (23/1,092,688 total ESTs), while lower expression has been detected in pancreas (2/213,440 total ESTs). GTEx (http://www.gtexportal.org/home/), which contains gene expression data from microarray and RNA-Seq platforms, shows that *CDH10* has the highest expression in brain, particularly cerebellar hemisphere. Medium Reads Per Kilobase of transcript per Million mapped reads of the tissue is around 15, while it is very low or not detectable in pancreases. Our data indicate that CDH10 protein expression is altered in PDAC. CDH10 expression was previously shown in normal human prostate luminal epithelial cells but was absent in prostate cancer.[28] The authors developed their own antisera to CDH10, and concluded that expression of CDH10 was involved in a specific role of secretory cell terminal differentiation. Using commercially available antisera, we preliminarily demonstrated a difference in expression and localization in normal pancreas as compared with normal prostate.

Mature cadherin proteins are composed of a large N-terminal extracellular domain, a single membrane-spanning domain, and a small, highly conserved C-terminal cytoplasmic domain. The extracellular domain consists of five subdomains, each containing a cadherin motif, and appears to determine the specificity of the protein's homophilic cell adhesion activity. Relatively little is known about CDH10. It was first discovered in 1999[11] and was found to be expressed in brain, where it was shown to be involved in blood-brain barrier "synaptic adhesions, axon outgrowth and guidance."[19] *CDH10* spans 157.7kb on chromosome 5 (chr5: 24,487,209-24,645,087, hg38), and produces two transcript variants. It is a Type II (atypical) cadherin, that is, it lacks a HAV cell adhesion recognition sequence specific to type I cadherins.

Previous literature suggests that sequences near *CDH10* might be involved in synaptic adhesion, axon outgrowth and guidance and genetic variations nearby *CDH10* are associated autism.[29,30] In addition, Biankin et al.[3] found aberrations in axon guidance pathway genes in pancreatic cancer genomes. Furthermore, mutations in *CDH10* were identified in colorectal cancer and lung cancer recently.[20–22] These findings are in line with our studies that variations in *CDH10* coding or nearby genomic regions may play roles in pancreatic cancer.

In summary, we have found germline alterations in and adjacent to *CDH10* in 2 of 41 individuals with FPC suggesting *CDH10* genomic alterations may play an as yet undefined role in predisposition of selected individuals for development of pancreatic cancer. The finding of LOH at the region most proximal to *CDH10* gene in at least 24% of sporadic pancreas cancer confirms and extends the report of Jones et al.[4] Our data supports the notion that *CDH10* is involved in sporadic pancreatic carcinogenesis, and might have some roles in rare cases of FPC. Further evaluation of the function of *CDH10* in epithelial neoplasms is warranted.

ACKNOWLEDGMENTS

This study was supported by the Thailand Research Fund (http://www.trf.or.th) and Mahidol University (TRF-MU grant number RSA5780065), Ramathibodi Cancer Center Grant (NJ); the NIH Specialized Programs of Research Excellence P50CA0062924 (APK); Lustgarten Foundation, Dorothea S. Corwin fund and Sol Goldman center (CAG).

We would like to dedicate this manuscript to Dr. Connie A. Griffin, who passed away from pancreatic cancer in the middle of this study.

REFERENCES

[1] Norris AL, Roberts NJ, Jones S, et al. Familial and sporadic pancreatic cancer share the same molecular pathogenesis. *Fam Cancer*. 2015;14:95–103.

[2] Klein AP, Brune KA, Petersen GM, et al. Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. *Cancer Res*. 2004; 64:2634–2638.

[3] Biankin AV, Waddell N, Kassahn KS, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*. 2012;491:399–405.

[4] Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*. 2008;321:1801–1806.

[5] Waddell N, Pajic M, Patch AM, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. 2015;518:495–501.

[6] Jones S, Hruban RH, Kamiyama M, et al. Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science*. 2009; 324:217.

[7] Roberts NJ, Klein AP. Genome-wide sequencing to identify the cause of hereditary cancer syndromes: with examples from familial pancreatic cancer. *Cancer Lett*. 2013;340:227–233.

[8] Roberts NJ, Norris AL, Petersen GM, et al. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discov*. 2016;6:166–175.

[9] Slater EP, Langer P, Niemczyk E, et al. *PALB2* mutations in European familial pancreatic cancer families. *Clin Genet*. 2010;78:490–494.

[10] Hruban RH, Canto MI, Goggins M, Schulick R, Klein AP. Update on familial pancreatic cancer. *Adv Surg*. 2010;44:293–311.

[11] Kools P, Vanhalst K, Van den Eynde E, van Roy F. The human cadherin-10 gene: complete coding sequence, predominant expression in the brain, and mapping on chromosome 5p13-14. *FEBS Lett*. 1999;452:328–334.

[12] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11:361–362.

[13] Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–249.

[14] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012; 7:e46688.

[15] Benusiglio PR, Malka D, Rouleau E, et al. *CDH1* germline mutations and the hereditary diffuse gastric and lobular breast cancer syndrome: a multicentre study. *J Med Genet*. 2013;50:486–489.

[16] Hansford S, Kaurah P, Li-Chang H. Hereditary diffuse gastric cancer syndrome: *CDH1* mutations and beyond. *JAMA Oncol*. 2015;1: 23–32.

[17] Paredes J, Figueiredo J, Albergaria A, et al. Epithelial E- and P-cadherins: role and clinical significance in cancer. *Biochim Biophys Acta*. 2012;1826:297–311.

[18] Oeschger FM, Wang WZ, Lee S, et al. Gene expression analysis of the embryonic subplate. *Cereb Cortex*. 2012;22:1343–1359.

[19] Williams MJ, Lowrie MB, Bennett JP, Firth JA, Clark P. Cadherin-10 is a novel blood-brain barrier adhesion molecule in human and mouse. *Brain Res*. 2005;1058:62–72.

[20] An CH, Je EM, Yoo NJ, Lee SH. Frameshift mutations of cadherin genes *DCHS2*, *CDH10* and *CDH24* genes in gastric and colorectal cancers with high microsatellite instability. *Pathol Oncol Res*. 2015; 21:181–185.

[21] Li C, Gao Z, Li F, et al. Whole exome sequencing identifies frequent somatic mutations in cell-cell adhesion genes in Chinese patients with lung squamous cell carcinoma. *Sci Rep*. 2015;5:14237.

[22] Yu J, Wu WK, Li X, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut*. 2015;64: 636–645.

[23] Girton JR, Johansen KM. Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. *Adv Genet*. 2008;61:1–43.

[24] Bien-Willner GA, Stankiewicz P, Lupski JR. SOX9cre1, a cis-acting regulatory element located 1.1 Mb upstream of SOX9, mediates its enhancement through the SHH pathway. *Hum Mol Genet*. 2007;16: 1143–1156.

[25] Noonan JP, McCallion AS. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet*. 2010;11:1–23.

[26] Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43:D805–D811.

[27] Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012;150:1107–1120.

[28] Walker MM, Ellis SM, Auza MJ, Patel A, Clark P. The intercellular adhesion molecule, cadherin-10, is a marker for human prostate luminal epithelial cells that is not expressed in prostate cancer. *Mod Pathol*. 2008;21:85–95.

[29] Ma D, Salyakina D, Jaworski JM, et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet*. 2009;73(Pt 3):263–273.

[30] Wang K, Zhang H, Ma D, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*. 2009; 459:528–533.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

How to cite this article: Jinawath N, Shiao M-S, Norris A, et al. Alterations of type II classical cadherin, cadherin-10 (CDH10), is associated with pancreatic ductal adenocarcinomas. *Genes Chromosomes Cancer*. 2017;56:427–435. https://doi.org/10.1002/gcc.22447

Jinawath *et al. J Transl Med (2016) 14:324*
DOI 10.1186/s12967-016-1078-3

Journal of Translational Medicine

REVIEW Open Access

Bridging the gap between clinicians and systems biologists: from network biology to translational biomedical research

Natini Jinawath[1,2†], Sacarin Bunbanjerdsuk[2†], Maneerat Chayanupatkul[3,4], Nuttapong Ngamphaiboon[5], Nithi Asavapanumas[6], Jisnuson Svasti[1,7,8] and Varodom Charoensawan[1,7,9*]

Abstract

With the wealth of data accumulated from completely sequenced genomes and other high-throughput experiments, global studies of biological systems, by simultaneously investigating multiple biological entities (e.g. genes, transcripts, proteins), has become a routine. Network representation is frequently used to capture the presence of these molecules as well as their relationship. Network biology has been widely used in molecular biology and genetics, where several network properties have been shown to be functionally important. Here, we discuss how such methodology can be useful to translational biomedical research, where scientists traditionally focus on one or a small set of genes, diseases, and drug candidates at any one time. We first give an overview of network representation frequently used in biology: what nodes and edges represent, and review its application in preclinical research to date. Using cancer as an example, we review how network biology can facilitate system-wide approaches to identify targeted small molecule inhibitors. These types of inhibitors have the potential to be more specific, resulting in high efficacy treatments with less side effects, compared to the conventional treatments such as chemotherapy. Global analysis may provide better insight into the overall picture of human diseases, as well as identify previously overlooked problems, leading to rapid advances in medicine. From the clinicians' point of view, it is necessary to bridge the gap between theoretical network biology and practical biomedical research, in order to improve the diagnosis, prevention, and treatment of the world's major diseases.

Keywords: Network biology, Systems biology, Biomedical research, Cancers, Personalized therapy

Background

Next-generation sequencing (NGS) and other high-throughput experiments highlight one of the most significant advances in molecular biology over the past decade. Such technological improvements enable a large number of molecules, including genes, transcripts, and proteins to be simultaneously measured in different conditions over time. This rapid generation of data has transformed molecular biology from a "data poor" to "data rich" discipline, leading to the emergence of systems biology [1–4]. The key challenges and bottlenecks of the modern-day molecular biology have shifted from simply gathering information to the analysis and interpretation of large quantities of data that can now be obtained.

Network representations have been widely used in physics and social science for decades, and are now among the most frequently used tools in systems biology. This technique provides not only a systematic representation of both the presence and abundance of biological molecules, but also displays the relationships or interactions between them. Networks have been used to represent the interactions between different types of biological molecules, e.g. protein–protein interactions [5–8], and in various biological systems including transcriptional

*Correspondence: varodom.cha@mahidol.ac.th
†Natini Jinawath and Sacarin Bunbanjerdsuk contributed equally to this work
[7] Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok, Thailand
Full list of author information is available at the end of the article

regulation [9–11], signaling [12–14], and metabolic pathways [15, 16]. Analyses of network sub-structures have revealed fundamental insights into how biological molecules are organized [17–20], which would not have been possible by studying individual genes or proteins.

Network representation and analysis has been successfully applied to study many systems in molecular biology [21]; however, the use of these tools in translational medicine and drug discovery is relatively new [22–24]. This might be due in part to the knowledge and understanding gaps between clinicians and systems biologists. By convention, clinicians typically focus on specific sets of key genetic markers associated with diseases, to identify the most probable drug targets. In contrast, systems biologists have strong computational and analytical skills, but frequently lack hands-on experimental experience. The lack of interaction of systems biologists with patients can prevent a full appreciation of the complexity of the problems and hindrances in biomedical research [25, 26]. In this review, we aim to improve the understanding of challenges in biomedical research and establish a common ground between clinicians and systems biologists to further promote the application of network biology in translational medicine.

Network biology in a nutshell

What are networks; what do they represent?

We first outline the fundamental concepts of a network representation. In general, a network represents the presence of objects or entities in a system as "nodes", and the relationships or interactions among the nodes are called "edges" (Fig. 1). In biology, nodes can represent biological molecules such as genes, proteins, and ligands, or even larger entities such as cells or individual humans. Edges represent physical interactions or contacts between biological molecules, biochemical processes between substrates and products, genetic interactions between genes, and in some cases, interactions between cells or individual organisms.

Biological information described in a network is not restricted to the presence of nodes and their relationships. The size of node, for instance, can reflect abundance of biological molecules (e.g. gene expression levels). Nodes can also be drawn in different shapes and/or colors according to the classification of interest (e.g. gene/protein family). Likewise, the thickness of an edge or the distance between nodes may represent the frequency or strength of pairwise interaction (e.g. affinity of protein–protein interaction); whereas colors can indicate different types of interactions (e.g. physical or genetic interaction). In addition, edges can be directional or non-directional, solid or dotted, depending on the types of interactions. Thus, networks are information-rich representations, which are widely used to summarize, visualize, and analyze large-scale datasets obtained from high-throughput experiments. To give an overview of the current application of networks in biomedical-related fields, here we review two major types of biological networks.

Interaction networks

We first illustrate the components of interaction networks, where the edges represent a "direct" relationship between nodes (Fig. 1, left). For instance, protein interaction networks, i.e. interactomes, describe physical interactions between proteins, usually obtained from high-throughput screening techniques such as yeast-two hybrid [6, 27], or affinity purification followed by mass spectrometry [5, 28]. In humans, analyses of protein–protein interaction networks have shown that dysfunctional interactions can lead to several diseases including neurological disorders such as ataxias [29], autism [30], several types of cancers including breast [31] and colorectal cancers [32], acute lymphoblastic leukemia [33], as well as other inheritable genetic diseases [34–37].

Transcriptional regulation networks (also known as Gene Regulatory Networks, GRNs) are widely used to illustrate the binding events of regulatory proteins, such as transcription factors, to the promoters of targeted genes, and this technique has been employed in the analysis of bacteria [38], budding yeasts [9], worms [39], and embryonic stem cells [40, 41]. GRNs are directional, and the relationship between two nodes is represented by an arrow starting from a regulator and pointing toward a targeted gene. Mis-regulation of gene expression leads to various diseases especially cancers, as seen in the genome-wide transcription network of the vertebrate transcription factor SOX4 [42], and the androgen receptor, a transcription factor that regulates the onset and progression of prostate cancer [43].

Interaction networks have also been used to describe the binding and affinity of ligands or small molecules to targeted proteins. As seen in a drug-target network [44], a list of drugs approved by the Food and Drug Administration (FDA) were linked to proteins according to drug-target binary associations. The analysis of these networks revealed that many drugs have overlapping but not identical sets of targets. In addition, the network analysis indicated that new drugs tend to be, at least partly, linked to well-characterized proteins already targeted by previously developed drugs. This suggests that the pharmaceutical industry might be shifting toward polypharmacology, to systematically address complex diseases using multiple drugs aimed at multiple specific targets in related pathways to improve treatment efficacy [45, 46].

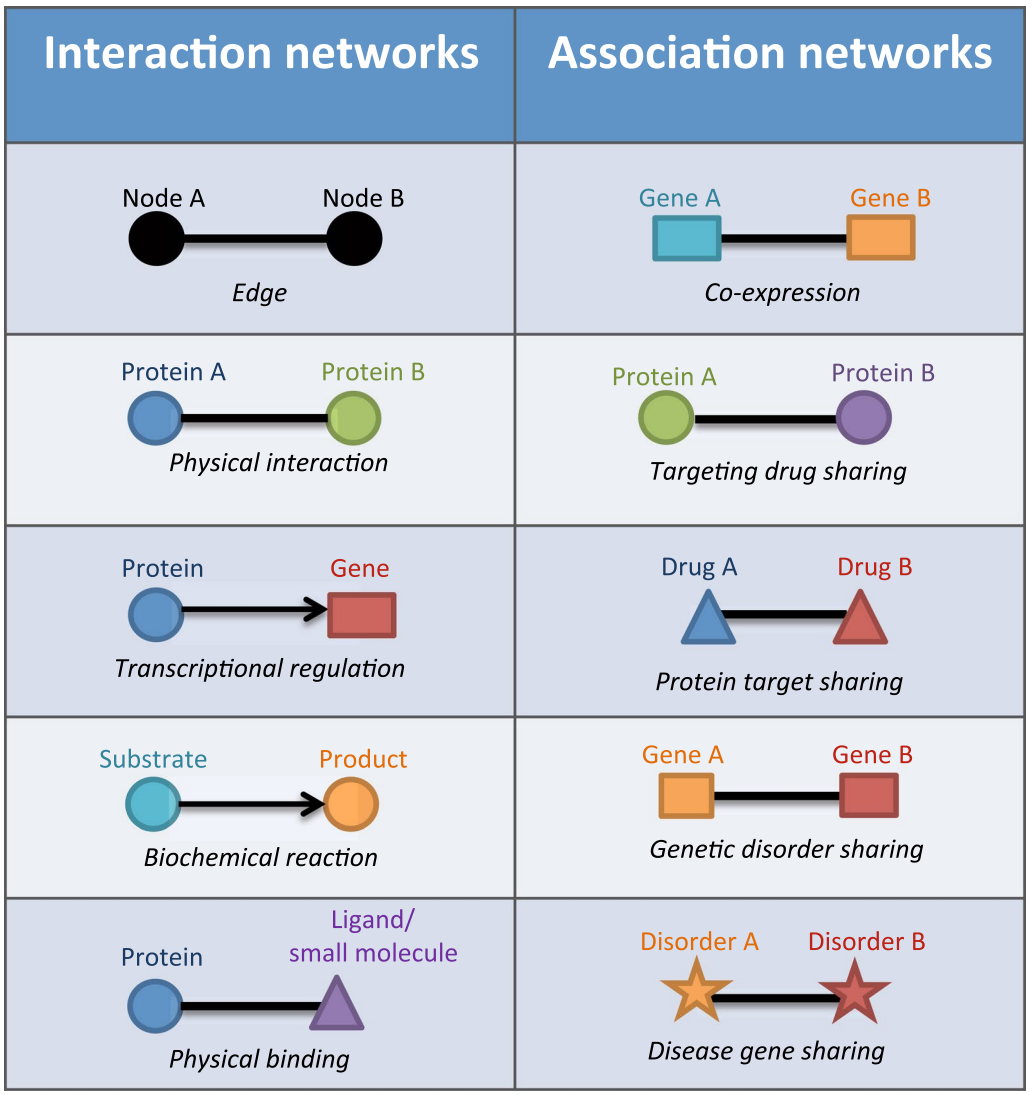


Fig. 1 Interaction networks (*Left*) represent direct interactions between biological molecules (e.g. transcripts, proteins, and ligands). The interactions represented include direct physical interaction (e.g. protein–protein, and gene regulatory networks) or transition (e.g. metabolic network). Association networks (*Right*) represent biological molecules that are linked based on their shared and/or common properties (e.g. co-expression)

Metabolic networks differs from other networks described earlier in the sense that the edges between two nodes (metabolites) do not represent physical contacts, but instead biochemical reactions that convert one metabolite to another. Recent studies have reconstructed and explored genome-scale metabolic networks in pathogenic microbes including *Staphylococcus aureus* [47], *M. tuberculosis* [48], as well as in human hosts [49]. These analyses may lead to a better understanding of host-pathogen interactions, and could aid in the design of drugs that specifically target the metabolic pathways of microbes and cause minimal interference with those of the hosts.

Association networks

Networks can also be used to visualize and summarize the overlap in expression profiles for thousands of transcripts/proteins obtained from high-throughput methods, such as expression microarray, RNA-seq, or short-gun proteomics [50]. In co-expression networks, two or more genes are linked if their products (mRNAs or proteins) exhibit similar expression profiles, with the strength/thickness of the edges proportional to how often the two transcripts are expressed at the same time and/or place [51, 52]. Co-expression networks are widely used as a starting point for inferring the cellular functions of uncharacterized genes, as in many cases, genes with

related functions show overlapping expression patterns [53]. New disease markers can be discovered from clusters of genes that are co-expressed with known disease-associated genes, as they frequently show differential expression between the normal and diseased populations [54–57].

Other association networks include drug target-protein networks [44], where each node is a protein and two proteins are linked if they are targeted by the same compounds. These networks can be computationally derived from the drug-target network described in the previous section. It provides a complementary protein-centric view by focusing on the proteins that are often co-targeted, and might be involved in related pathways. Conversely, two or more drugs can be linked in a network based on common properties, such as targeting specific proteins or side effects. It has been shown that documented adverse side effects could be used to infer molecular drug-target interactions [58]. This type of network has the potential to predict whether or not existing and routinely used drugs have additional unknown off-targets, allowing for these drugs to be candidates for additional, distinct therapeutic categories. Illustrations of the potential of alternative uses for current drugs are sildenafil, losartan, and fenofibrate. Sildenafil (e.g. Viagra®, Pfizer Incorporated) was initially developed to treat angina, but a side effect (prolong penile erection) discovered during clinical trial has become its main use. The antihypertensive drug losartan blocks angiotensin II type 1, and is now a candidate drug for preventing aortic aneurysm complications in Marfan syndrome patients, through reduction of TGF-β activitiy [59, 60]. Fenofibrate, a drug mainly used for controlling cholesterol levels in cardiovascular patients, has also been shown to suppress growth of hepatocellular carcinoma [61].

Global disease networks offer a useful insight into how human disorders are related. In the "human disease network" [62], disease nodes are connected if they share at least one gene with mutations associated with both diseases. Complementarily, the gene-centric version of this network comprises nodes of disease genes, linked if they are associated with the same disorders. Such networks not only represent a framework to visualize all known disease genotype-phenotype associations, but also reveal that human diseases are much more genetically related than previously appreciated [63]. This is highlighted by a gigantic network comprising over 500 interconnected human diseases [7].

What can we learn from networks and their properties?

In addition to being a framework for visualizing and documenting all the known relationships between nodes, earlier analyses of large-scale networks from high-throughput studies have revealed many interesting biologically relevant properties, which cannot be obtained by studying genes and proteins individually [64–66]. One of the most frequently observed properties of biological networks is the connectivity distribution that follows a power-law distribution, known as "scale-free networks". This pattern of connections, also known as the "small world property", has also been extensively studied for their statistical features in different types of networks, including social networks, scientific collaboration networks, and the World Wide Web [67–72]. In brief, a scale-free network consists of a small number of "hubs", i.e. nodes that are connected to a larger number of other nodes, through different types of interactions aforementioned. In contrast to hubs, the majority of nodes in the network have much fewer connections. Several studies have documented similar observation for biological networks, including protein–protein interaction networks [6, 17, 73] and metabolic networks [15, 74].

Because of their connectivity distribution, scale-free networks are robust against random deletion of nodes. That is, the connections between a node and most other nodes remain intact, if nodes are removed randomly. In contrast, scale-free networks quickly become non-functional if hubs are targeted. Earlier studies have shown that many pathogenic organisms have evolved to target the central components (i.e. hubs) of a human protein interaction network, and quickly disrupt various cellular functions, including the immune response [75, 76]. Similarly, one would expect drugs that specifically inhibit the central components of the regulatory circuits in a pathogen will rapidly disrupt their homeostatic processes, and thus efficiently eliminate them. As a result, these hubs from pathogenic organisms could be promising candidates for novel drugs. Network connectivity distribution is one of the better-studied areas, and a number of insightful reviews and analyses are available [77, 78].

Another interesting example of biological network properties are the network motifs, which are sets of well-defined interconnection patterns between nodes [19]. These connectivity patterns, or network sub-circuits, recur in biological networks at a frequency significantly higher than in randomized networks [79–81], signifying their important roles as building blocks for the large-scale organization of interactions. The patterns and proportions of sub-circuits used in different networks are distinct, depending on the functionality required under different conditions. Interestingly, it has been shown in a yeast transcription regulatory network that sub-network structures, facilitating fast signal propagation (e.g. single-inputs), are more frequently employed to respond to external stressors and sudden environmental changes (e.g. DNA damage or diauxic shift), because a rapid

response is required against the stressors. In contrast, motifs that buffer spurious inputs or only respond to persistent signals (e.g. feed-forward-loops) are more suitable for analysis of normal growth stages (e.g. sporulation) [18, 82].

Applications of network biology in translational medicine

Disease network and drug discovery

Using a transistor radio as an analog of a biological system, Yuri Lazebnik described how a biologist would fix a broken radio, assuming no prior knowledge of how the radio components were wired together [83]. A traditional biological approach would involve removing (gene knockout, mutagenesis) each part of a functioning radio and track the changes in performance (phenotype). However, the human "radios" are different and repeating this process on all the components would generate an enormous amount of data, some of which may be redundant or contradictory. In contrast, a typical engineering approach would involve systematic reconstruction of a component diagram from a normal radio (e.g. regulatory network), and compare the broken radios with the normal reference. Can a similar problem-solving mindset help expedite advances in biomedical research?

If regulatory circuits that control biological activities in a human body can be represented using a complex network, then a diseased state would be expected to occur when the normal state of the network is perturbed. Failure of key components (e.g. mutations in hub genes in genetic diseases) or external stimuli (e.g. invasion of pathogens in infectious diseases) would lead to loss of network integrity. Diseased perturbations can occur at different regulatory levels, as illustrated in Fig. 2. Firstly, the absence or malfunction in important network components can lead to diseases, such as the loss of a particular gene. The absence of *TBX1*, in 22q11.2 deletion syndrome (DiGeorge syndrome) is responsible for the majority of characteristic features of this disease [84] (Fig. 2a, the absence of node is illustrated in red). Similarly, inappropriate levels of gene expression can cause disorders (Fig. 2b, altered node size). For example, specific mutations in the *FGFR3* gene result in an overactive receptor and lead to the short stature phenotype observed in achondroplasia [85]. Some diseased states can be explained by mis-regulation of the interactions between key components of the network (Fig. 2c, missing edge), as well as mis-direction (Fig. 2d, mis-directed edge) or strength (Fig. 2e, altered edge's thickness) of interactions. The diseases that can be linked to erroneous interactions include neurodegenerative and neurodevelopmental diseases, genetic disorders, and cancers. In these cases, mutations in multiple relevant genes lead to abnormal protein interactions, and disrupt networks (see [29, 30, 36, 37] for details).

Some of the long-standing challenges in drug discovery are lack of specificity, high incidence of adverse effects, and unpredicted toxicities of new therapeutic compounds [86]. As a result, modern-day drug discovery employs more targeted approaches, such as virtual screening and structure-based drug design to complement conventional in vitro high-throughput screening [46, 87]. These new approaches rely on an accurate global understanding of the mechanisms of diseases. Comprehensive understanding of the network and regulatory circuit for a particular disease process would help to identify network hubs with the potential to be novel drug targets.

A network model of cancers

In the past decades, chemotherapy had been the backbone for systemic treatment of cancers. When administered to patients, these drugs target rapidly dividing cells but lack specificity. Survival of both cancer cells and normal, rapidly growing cells are impaired, resulting in side effects such as bone marrow suppression and hair loss, due to toxicity toward bone marrow cells and hair follicles, respectively. With recent advances in molecular biology and genetics, several genetic mutations and other alterations have been described for various cancers, and these changes specific to cancer cells have become an attractive target for novel therapies. The concept of "driver" and "passenger" mutations in carcinogenesis is comparable to hubs and peripheral nodes in a network, whereby a subset of somatic alterations present in each tumor is a driver of the oncogenic process [88]. Acting as a complex network hub, these driver mutations promote cancer cell survival, resistance to apoptosis, and lead to carcinogenesis (so-called "oncogene addiction"). This idea is supported by successful identification of new cancer fusion drivers from the network hubs and their partners, as the fusion mutation can lead to functional de-regulation of multiple genes and pathways [89]. Inhibition of the driver mutation has the potential to induce cell death, and thus becomes a strong candidate for targeted therapy [90]. As cancer cells are addicted to this driver mutation, specifically blocking these hubs would theoretically be more effective and less toxic compared to conventional chemotherapy.

To date, many targeted therapies have been approved as a standard of care in various cancers with additional clinical studies underway. Identification of a true driver; however, remains one of the biggest challenges. Pathogenesis of cancer development is usually complex and involves several molecules and pathways. Therefore, targeting one particular molecule or pathway might not be effective, as cancer cells may utilize alternative pathways

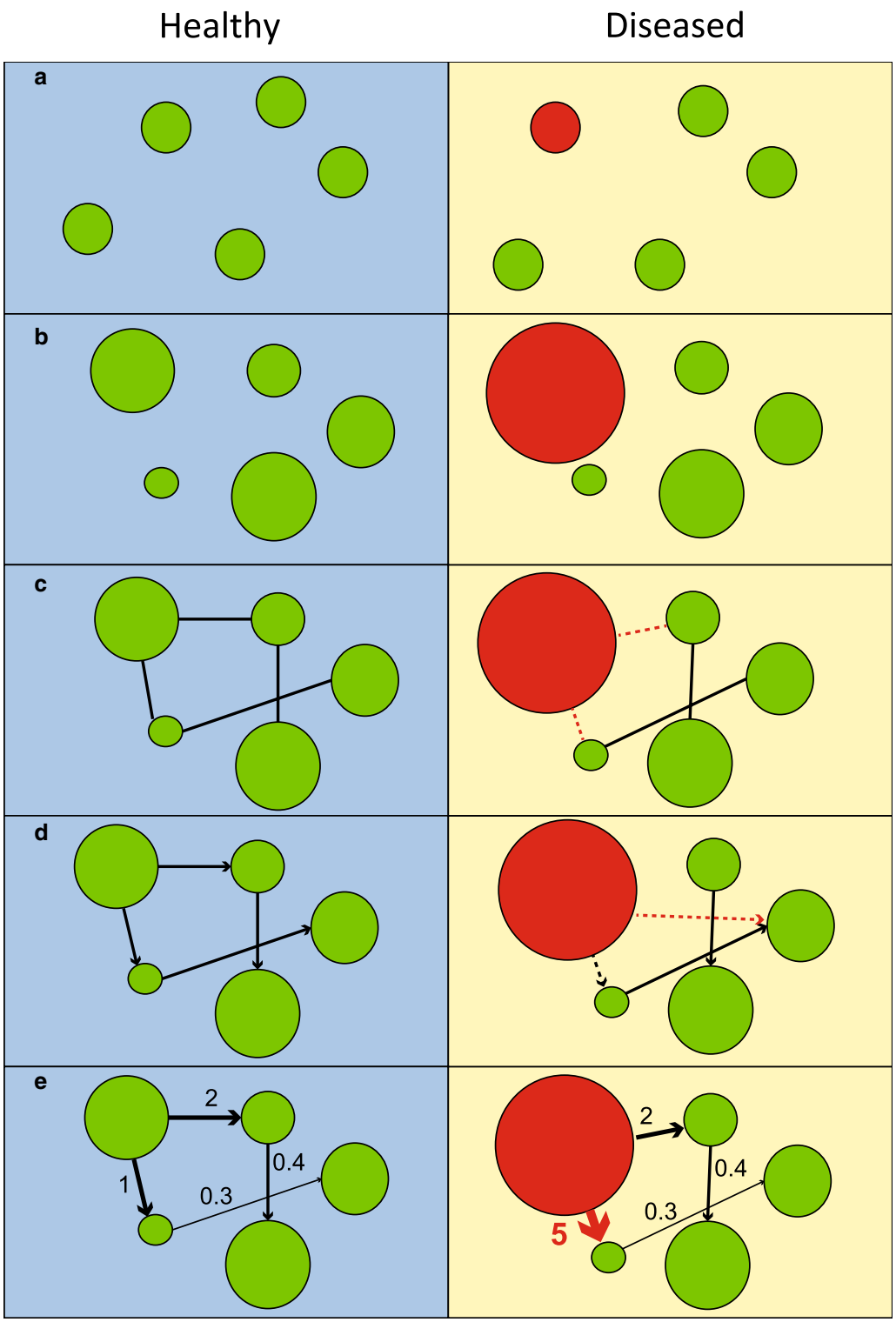


Fig. 2 Biological networks of healthy (*left panel*) and diseased (*right panel*) individuals. Biological components in healthy individuals are represented as *green nodes* in a network. Pathological perturbation, represented by *red nodes* that lead to morbidity, can occur at different stages of the regulation of key components: **a** presence and absence of key component (*green* for presence and *red* for absence), **b** mis-regulated gene expression, leading to over- or under-expression (*node sizes* represent expression levels), **c** absence or erroneous interactions with interacting partners (*dotted lines* represent erroneous interactions), **d** mis-regulated directions (*mis-directed arrows*), or **e** strengths of interactions (*thicknesses of arrows* and *accompanying numbers* denote interaction strengths)

to promote cell survival. Additionally, with the advent of next-generation sequencing, the previously well-accepted but unproven concept of tumor genetic heterogeneity has been solidly confirmed [91]. Sequential use of more than one targeted cancer therapy to finish off resistant clones, such as in the case of tumor recurrence, is likely to become a trend in cancer genomic medicine [92].

Breast cancer network: mechanisms of resistance

The regulatory network in breast cancer is a particularly interesting case study, due to its heterogeneous histological and molecular features, and clinical manifestations that lead to multiple molecular sub-types. Based on gene expression profiling, breast cancer can be categorized into four main molecular sub-types: (i) basal-like breast cancer (mainly estrogen-receptor (ER)-negative, progesterone-receptor (PR)-negative, and human epidermal growth factor receptor 2 (HER2)-negative); (ii) luminal-A cancer (ER-positive or ER+, and histologically low-grade); (iii) luminal-B cancer (ER+ and histologically high-grade); and (iv) HER2-positive (HER2+) cancer (over-expression and/or amplification of HER2). Each molecular sub-type has a distinct course of disease progression and responds differently to specific treatments, including endocrine therapy, anti-HER2 drugs and cytotoxic chemotherapy [93].

As shown in Fig. 3, ER and HER2 can be considered as hubs of the breast cancer network. The ER+ breast cancer cells depend on activation of ER by estrogen, a sex steroid hormone. ER acts as a transcription factor in the nucleus when bound by estrogen in the genomic (nuclear) pathway, resulting in tumor cell proliferation [94]. The signal can also be activated through the non-genomic (non-nuclear) pathway, where estrogen binds to membrane-associated ER. Endocrine therapy against the ER hubs is one of the cornerstones of treatment for ER+/HER2- breast cancers (luminal-A and B) [95]. The predominant endocrine therapies are a selective ER modulator (SERM), an aromatase inhibitor (AI), and selective ER down-regulators (SERD), such as tamoxifen, anastrozole, and fulvestrant [96].

HER2, a member of the epidermal growth factor receptor tyrosine kinase family, is a hub in the HER2+ breast cancer network. Over-expressed and/or amplified *HER2* is found in approximately 20–30% of invasive breast cancers [97]. HER2 activates intracellular signaling cascades, leading to tumor cell proliferation. Inhibition of HER2 through the use of anti-HER2 drugs significantly prolongs survival in HER2+ breast cancer patients. Currently, several anti-HER2 drugs are FDA-approved for HER2+ breast cancer, including trastuzumab, lapatinib, pertuzumab, and trastuzumab emtansine (T-DM1). Resistance to each of these specific treatments has been observed, as well as interactions between the ER and HER2 hubs (Fig. 3) [94, 98]. Since ER+/HER2+ tumor cells depend on both hubs, endocrine therapy alone cannot completely inhibit signals with tumor cell proliferation continuing to be activated through HER2 (so-called "cross-talk"). This has been identified as a primary mechanism of resistance in ER+/HER2+ breast cancer patients with a low response to endocrine therapy. With a better understanding of global gene regulation networks and the interplay between the two hubs, a combined treatment of endocrine therapy and anti-HER2 drugs was proposed. Several phase 3 clinical studies have already demonstrated increased efficacy of endocrine therapy in the ER+/HER2+ breast cancer when combined with anti-HER2 drugs [99–101].

On the other hand, ER+/HER− breast cancer does not depend on the HER2 hub, and is thus usually responsive to the first line endocrine therapy. However, resistance can still occur leading to less effective endocrine therapy. Blocking the ER hub with any endocrine therapy would inhibit only the genomic pathway, but not the non-genomic pathway where abnormal activation of the PI3K/Akt/mTOR pathway by somatic mutations can result in either *de novo* or acquired endocrine therapy resistance [102, 103]. Understanding this relationship has led to a second line of endocrine therapy using mTOR inhibitors. A large phase 3 clinical study of metastatic ER+/HER2− breast cancer patients, who failed the first line AI treatment, reported longer progression-free survival in a group treated with a combination of an mTOR inhibitor and another different AI [104, 105].

Having a comprehensive understanding of the interactions between network components of specific disease should lead to improved efficacy in treatments, similar to those elucidated using the breast cancer model above. Indeed, a number of groups have already begun utilizing network biology to address different aspects of cancers with the goal to improve diagnosis and treatment. A model to identify genes potentially associated with high risks of breast cancer has been developed by integrating data from co-expression, biochemical, and protein interaction networks. Using this model, Pujana and coworkers successfully identified Hyaluronan Mediated Motility Receptor (*HMMR*), a hub of the integrated network, as a novel high risk associated locus [31]. The gene regulatory network for breast cancer has also been constructed [106]. Taylor and colleagues merged spatial gene expression information with the protein interaction network to highlight the interactions that are active in specific tissues, where the interacting partners are also co-expressed [107]. This work also revealed the loss of key interactions between the network hubs, such as *BRCA1* and their binding partners, in patients who died of breast cancer due to mis-regulation

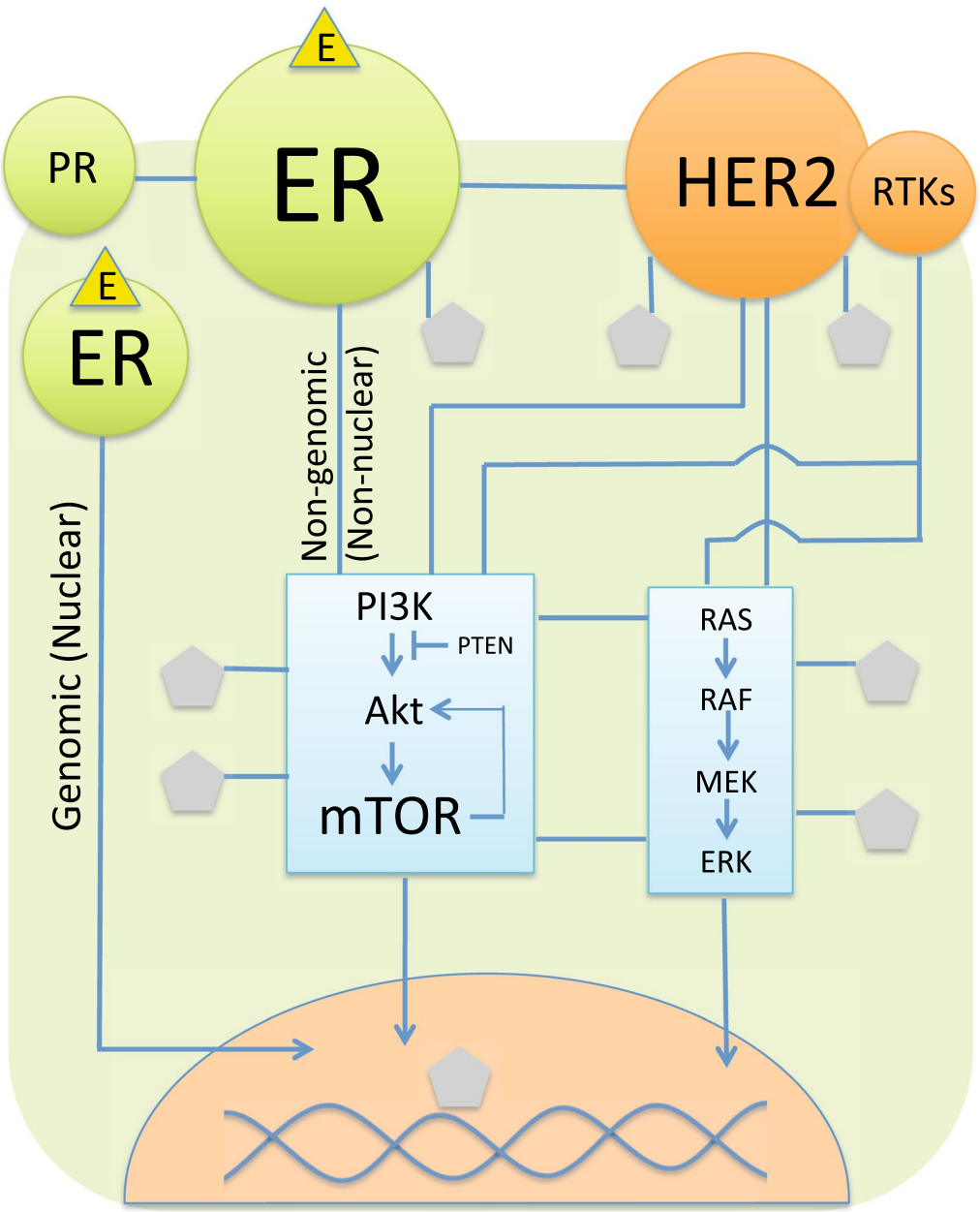


Fig. 3 A simplified diagram of the therapeutic breast cancer network. The main targetable hubs are ER and HER2 receptor. The PI3K/Akt/mTOR hub was relatively recently identified to be the common mechanism of targeted therapy resistance. *Circles* and *rectangles* represent cellular receptors and signaling pathways, respectively. The *pentagons* represent other unspecified molecules interacted with the hubs. *Arrows* represent the directions of signals. (*E* estrogen, *ER* estrogen receptor, *PR* progesterone receptor, *HER2* HER2 receptor, *RTKs* receptor tyrosine kinases)

of the partner proteins. In contrast, the expression of hubs and their partners were strongly correlated in surviving patients. The complexity of the disease network is not only restricted to the gene–gene and gene-drug interactions, but also hinges upon the interactions between disease/drug and the host (i.e. genetic background of the patients), as we discuss in the next section.

From individual network to personalized medicine

As we are approaching the so-called personalized and precision medicine era, where does network biology fit in the picture? Figure 4 depicts our view on how networks can be an important tool to help clinicians understand the physiological complexity of individual humans, predict possible failure of certain components that may lead

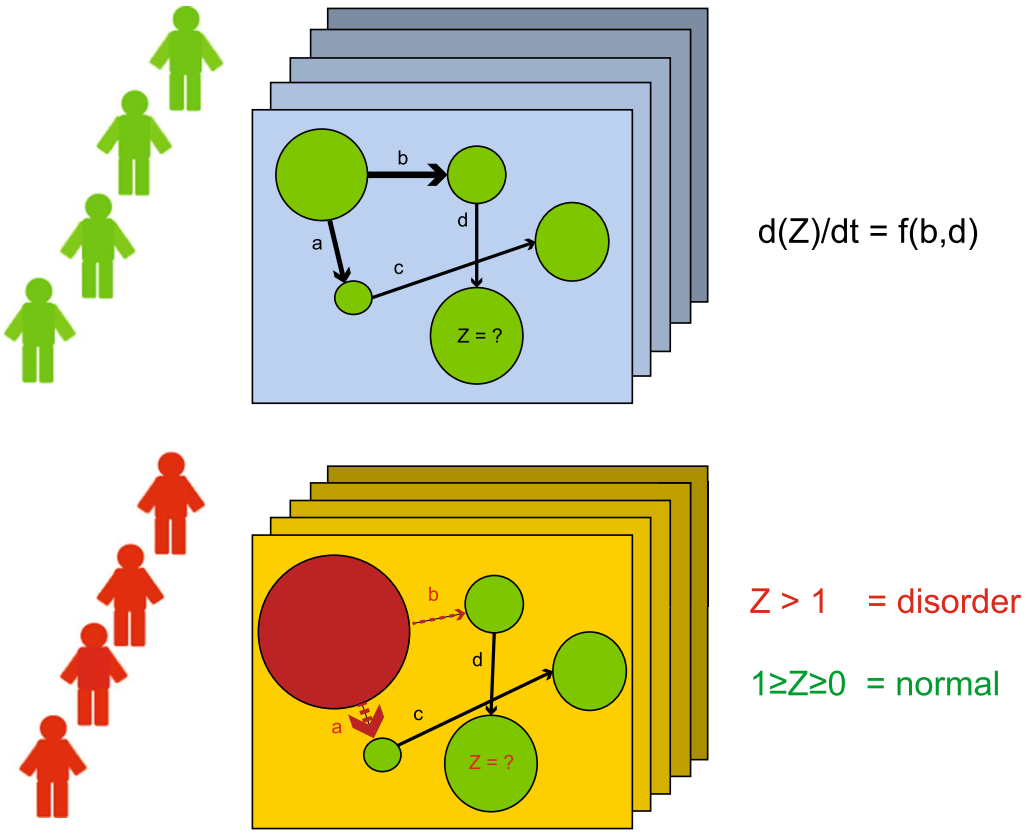


Fig. 4 Healthy (*top panel*) and diseased (*bottom panel*) individual networks. Healthy individuals might show slight variations in their individual networks, which also differ over time. However, diseased networks are expected to show greater disparity than that between healthy individuals. In the example shown, the network component *Z* is controlled by its upstream components through the interactions of *b* and *d* (the molecule Z is a function of *b* and *d*). If the expression of *Z* is greater than a defined limit (e.g. 1 in this case), morbidity can be predicted (d(Z)/dt: change of expression level of molecule Z over time)

to morbidity, and deduce the most suitable preventative and treatment plans for individual patients. Genetic variation between human individuals is estimated to be less than 1% of the human genome, but through sophisticated regulation of genes and other genetic elements, this small amount of genetic variation accounts for much greater differences in terms of our appearance, intellect, and health [108]. On top of genomes, which encode individual sets of gene products (e.g. proteins, mRNA), individual networks represent the unique interplay between different components in each patient. Understanding the extent of variations between individual networks may allow clinicians to statistically and quantitatively distinguish normal variations in healthy individuals (Fig. 4, upper panel) from critical perturbations that lead to diseases and disorders (Fig. 4, lower panel). Network biology enables researchers to assess multiple components that do not show distinguishable differences between healthy individuals and those with cancers, but are collectively dysfunctional in cancers. A sub-network in which overall activity can be discriminated between patients versus controls has been shown to be a more reproducible prognostic marker of diseases than individual genes in the sub-network, which are not significantly differentially expressed [109, 110].

Single nucleotide polymorphisms (SNPs) and other genetic variations add another dimension of disease-host interaction to disease networks. SNPs can provide clinicians with a good indication on how likely an individual might be to develop certain genetic diseases, assuming that all genetic elements associated with diseases are eventually identified. In addition, networks of individuals can, in part, aid pharmacogenomic progress by explaining why the efficacy and toxicity profiles for the same drug may differ in each patient. For instance, tamoxifen is metabolized by *CYP2D6* and variations in this gene among individuals may affect the response to the drug [111].

No matter how comprehensive, a genetic map cannot capture environmental factors (e.g. lifestyle, contact with

pathogens) that heavily influence biochemical stages. Thus, outcomes for the interplay between genetics and environment may be absent in the analysis. Having a network that combines both the genetic variations and measurable biochemical outcomes, such as gene expression, should assist in turning conceptual ideas into more quantitative models, which in turn would enhance the accuracy of prognosis and predictions of disease progression in each patient (as demonstrated in Fig. 4). Such a complete individual network may not be possible in the near future; however, we start to see that the integration of genetic variations and biochemical outcomes (gene expression and protein interaction profiles) has utility in helping identify new disease-associated marker genes [110, 112, 113].

Thanks to considerable effort and resources the community has put into developing computational tools for biological network analysis, we are now well-equipped with a range of user-friendly software that can be employed to handle, visualize, and analyze large-scale datasets. Importantly, the tools that will be particularly useful for translational medical research need to be able to combine multiple layer datasets (e.g. genomics, transcriptomics, proteomics, and metabolomics) and/or heterogeneous datasets (e.g. from different platforms or formats) [3]. The most commonly known network analysis tools currently available are Cytoscape [114], NAViGaTOR [115], VisANT [116], CellDesigner [117], and the commercial software Ingenuity IPA (Ingenuity Systems Inc., Redwood City, CA). More recently introduced tools include NaviCell, which has been developed for online network visualization and curation [118], and BNOmics [119], which can be used for inference and visualization of Bayesian networks of large heterogeneous data. Comprehensive guides to network biology tools, as well as detailed discussion on their key features and functionality can be found in earlier review articles [3, 120].

Conclusions

Network biology provides an opportunity to image a clear global picture of drug-disease-host interactions and the biological complexity of diseases more easily from an unprecedented top-down vantage. This will allow a better understanding of the relationships between multiple genes and other biological entities, as well as identify the missing links in our knowledge. These strategies are required to fully grasp the intricacies of diseases, which cannot be obtained by studying an individual or a smaller set of genes. The complexity of the therapeutic networks is ever-growing, and many new nodes are being discovered every day. In the future, some of these nodes may become new hubs for targeted therapy.

Abbreviations

ER: estrogen receptor; FDA: Food and Drug Administration; GRN: gene regulatory network; HER2: human epidermal growth factor receptor 2; HMMR: hyaluronan mediated motility receptor; NGS: next-generation sequencing; PR: progesterone receptor; SNP: single nucleotide polymorphisms.

Authors' contributions

NJ, NA and VC conceived the concept of the review and figures. NJ, SB, NN and VC wrote the manuscript. MC, NA and JS contributed the ideas and literature search. SB, NA and VC prepared the figures. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Author details

[1] Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom, Thailand. [2] Program in Translational Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. [3] Department of Physiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. [4] Division of Gastroenterology and Hepatology, Department of Medicine, Baylor College of Medicine, Houston, TX, USA. [5] Medical Oncology Unit, Department of Medicine Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. [6] Department of Physiology, Faculty of Science, Mahidol University, Bangkok, Thailand. [7] Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok, Thailand. [8] Laboratory of Biochemistry, Chulabhorn Research Institute, Bangkok, Thailand. [9] Systems Biology of Diseases Research Unit, Faculty of Science, Mahidol University, Bangkok, Thailand.

Acknowledgements

The authors thank Assoc. Prof. Laran Jensen, Department of Biochemistry, Mahidol University, for very helpful comments on the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

NJ is a recipient of TRF Research Scholar Fund (RSA5780065), Mahidol University-National Research Council of Thailand (Government Fiscal Year Budget) Fund, and the Research grants from the Ramathibodi Cancer Center. SB is a recipient of research assistant scholarship from Faculty of Medicine Ramathibodi Hospital and Faculty of Graduate Studies, Mahidol University. NN acknowledges the Talent Management Program, Mahidol University. VC acknowledges the TRF Grant for New Researcher (TRG5880067), and Faculty of Science, Mahidol University. The NJ and VC laboratories are supported by the Crown Property Bureau Foundation through Integrative Computational BioScience (ICBS) Center, Mahidol University.

Received: 23 August 2016 Accepted: 8 November 2016
Published online: 22 November 2016

References

1. Attur MG, Dave MN, Tsunoyama K, Akamatsu M, Kobori M, Miki J, Abramson SB, Katoh M, Amin AR. "A system biology" approach to bioinformatics and functional genomics in complex human diseases: arthritis. Curr Issues Mol Biol. 2002;4(4):129–46.
2. Ideker T, Lauffenburger D. Building with a scaffold: emerging strategies for high-to low-level cellular modeling. Trends Biotechnol. 2003;21(6):255–62.
3. Chuang HY, Hofree M, Ideker T. A decade of systems biology. Annu Rev Cell Dev Biol. 2010;26:721–44.
4. Kitano H. Systems biology: a brief overview. Science. 2002;295(5560):1662–4.
5. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440(7084):631–6.

6. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008;322(5898):104–10.
7. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011;144(6):986–98.
8. Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, et al. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–26.
9. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431(7004):99–104.
10. Kim J, Chu J, Shen X, Wang J, Orkin SH. An extended transcriptional network for pluripotency of embryonic stem cells. Cell. 2008;132(6):1049–61.
11. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512(7515):453–6.
12. Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, et al. High-throughput mapping of a dynamic signaling network in mammalian cells. Science. 2005;307(5715):1621–5.
13. Bhalla US, Ram PT, Iyengar R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. Science. 2002;297(5583):1018–23.
14. Li L, Tibiche C, Fu C, Kaneko T, Moran MF, Schiller MR, Li SS, Wang E. The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. Genome Res. 2012;22(7):1222–30.
15. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature. 2000;407(6804):651–4.
16. Oberhardt MA, Goldberg JB, Hogardt M, Papin JA. Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. J Bacteriol. 2010;192(20):5534–48.
17. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004;430(6995):88–93.
18. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature. 2004;431(7006):308–12.
19. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet. 2002;31(1):64–8.
20. Kohestani H, Giuliani A. Organization principles of biological networks: an explorative study. Biosystems. 2016;141:31–9.
21. Emmert-Streib F, Glazko GV. Network biology: a direct approach to study biological function. Wiley Interdiscip Rev Syst Biol Med. 2011;3(4):379–91.
22. Zhang H, Gustafsson M, Nestor C, Chung KF, Benson M. Targeted omics and systems medicine: personalising care. Lancet Respir Med. 2014;2(10):785–7.
23. Hood L. Systems biology and p4 medicine: past, present, and future. Rambam Maimonides Med J. 2013;4(2):e0012.
24. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
25. Clermont G, Auffray C, Moreau Y, Rocke DM, Dalevi D, Dubhashi D, Marshall DR, Raasch P, Dehne F, Provero P. Bridging the gap between systems biology and medicine. Genome Med. 2009;1(9):88.
26. Michor F, Liphardt J, Ferrari M, Widom J. What does physics have to do with cancer? Nat Rev Cancer. 2011;11(9):657–70.
27. Waaijers S, Koorman T, Kerver J, Boxem M. Identification of human protein interaction domains using an ORFeome-based yeast two-hybrid fragment library. J Proteome Res. 2013;12(7):3181–92.
28. Jirawatnotai S, Hu Y, Michowski W, Elias JE, Becks L, Bienvenu F, Zagozdzon A, Goswami T, Wang YE, Clark AB, et al. A function for cyclin D1 in DNA repair uncovered by protein interactome analyses in human cancers. Nature. 2011;474(7350):230–4.
29. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell. 2006;125(4):801–14.
30. Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, Broly M, Rodriguez M, Tam S, Trigg SA, et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat Commun. 2014;5:3650.
31. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet. 2007;39(11):1338–49.
32. Nibbe RK, Koyuturk M, Chance MR. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol. 2010;6(1):e1000639.
33. Hajingabo LJ, Daakour S, Martin M, Grausenburger R, Panzer-Grumayer R, Dequiedt F, Simonis N, Twizere JC. Predicting interactome network perturbations in human cancer: application to gene fusions in acute lymphoblastic leukemia. Mol Biol Cell. 2014;25(24):3973–85.
34. Charoensawan V, Adryan B, Martin S, Sollner C, Thisse B, Thisse C, Wright GJ, Teichmann SA. The impact of gene expression regulation on evolution of extracellular signaling pathways. Mol Cell Proteomics. 2010;9(12):2666–77.
35. Keith BP, Robertson DL, Hentges KE. Locus heterogeneity disease genes encode proteins with high interconnectivity in the human protein interaction network. Front Genet. 2014;5:434.
36. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015;161(3):647–60.
37. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabasi AL. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
38. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. BioEssays. 1998;20(5):433–40.
39. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, et al. A gene-centered *C. elegans* protein-DNA interaction network. Cell. 2006;125(6):1193–205.
40. Liu X, Huang J, Chen T, Wang Y, Xin S, Li J, Pei G, Kang J. Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. Cell Res. 2008;18(12):1177–89.
41. Huang TS, Li L, MoalimNour L, Jia D, Bai J, Yao Z, Bennett SA, Figeys D, Wang L. A regulatory network involving beta-catenin, E-cadherin, PI3K/Akt, and slug balances self-renewal and differentiation of human pluripotent stem cells in response to Wnt signaling. Stem Cells. 2015;33(5):1419–33.
42. Scharer CD, McCabe CD, Ali-Seyed M, Berger MF, Bulyk ML, Moreno CS. Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. Cancer Res. 2009;69(2):709–17.
43. Wang Q, Li W, Liu XS, Carroll JS, Jänne OA, Keeton EK, Chinnaiyan AM, Pienta KJ, Brown M. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. Mol Cell. 2007;27(3):380–92.
44. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol. 2007;25(10):1119–26.
45. Zhao S, Iyengar R. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. Annu Rev Pharmacol Toxicol. 2012;52:505–21.
46. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, Norval S, Sassano MF, Shin AI, Webster LA, et al. Automated design of ligands to polypharmacological profiles. Nature. 2012;492(7428):215–20.
47. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BO. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc Natl Acad Sci USA. 2016;113(26):E3801–9.
48. Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol. 2010;6:422.
49. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA. 2007;104(6):1777–82.

50. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2007;25(1):117–24.
51. Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 2010;4:8.
52. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–55.
53. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. Functional discovery via a compendium of expression profiles. Cell. 2000;102(1):109–26.
54. Xu Y, Duanmu H, Chang Z, Zhang S, Li Z, Liu Y, Li K, Qiu F, Li X. The application of gene co-expression network reconstruction based on CNVs and gene expression microarray data in breast cancer. Mol Biol Rep. 2012;39(2):1627–37.
55. Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, Jin R, Payne P, Huang K. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. BMC Bioinformatics. 2010;11(Suppl 9):S5.
56. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463(7279):318–25.
57. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
58. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature. 2012;486(7403):361–7.
59. Cohn RD, van Erp C, Habashi JP, Soleimani AA, Klein EC, Lisi MT, Gamradt M, ap Rhys CM, Holm TM, Loeys BL. Angiotensin II type 1 receptor blockade attenuates TGF-β–induced failure of muscle regeneration in multiple myopathic states. Nat Med. 2007;13(2):204–10.
60. Williams A, Davies S, Stuart A, Wilson D, Fraser A. Medical treatment of Marfan syndrome: a time for change. Heart. 2008;94(4):414–21.
61. Yamasaki D, Kawabe N, Nakamura H, Tachibana K, Ishimoto K, Tanaka T, Aburatani H, Sakai J, Hamakubo T, Kodama T. Fenofibrate suppresses growth of the human hepatocellular carcinoma cell via PPARα-independent mechanisms. Eur J Cell Biol. 2011;90(8):657–64.
62. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci USA. 2007;104(21):8685–90.
63. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. 2009;5(4):e1000353.
64. Cho DY, Kim YA, Przytycka TM. Network biology approach to complex diseases. PLoS Comput Biol. 2012;8(12):e1002820.
65. Furlong LI. Human diseases through the lens of network biology. Trends Genet. 2013;29(3):150–9.
66. Marcotte E, Boone C, Babu MM, Gavin A-C. Network biology editorial 2013. Mol BioSyst. 2013;9(7):1557–8.
67. Newman ME. The structure of scientific collaboration networks. Proc Natl Acad Sci USA. 2001;98(2):404–9.
68. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
69. Strogatz SH. Exploring complex networks. Nature. 2001;410(6825):268–76.
70. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–2.
71. Ipsen M, Mikhailov AS. Evolutionary reconstruction of networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2002;66(4 Pt 2):046109.
72. Milgram S. The small world problem. Psychol Today. 1967;2:60.
73. Martin S, Sollner C, Charoensawan V, Adryan B, Thisse B, Thisse C, Teichmann S, Wright GJ. Construction of a large extracellular protein interaction network and its resolution by spatiotemporal expression profiling. Mol Cell Proteomics. 2010;9(12):2654–65.
74. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci. 2001;268(1478):1803–10.
75. Brodsky IE, Medzhitov R. Targeting of immune signalling networks by bacterial pathogens. Nat Cell Biol. 2009;11(5):521–6.
76. Dyer MD, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, Murali TM, Sobral BW. The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. PLoS ONE. 2010;5(8):e12089.
77. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.
78. Stumpf MP, Porter MA. Mathematics. Critical truths about power laws. Science. 2012;335(6069):665–6.
79. Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007;8(6):450–61.
80. Davidson EH. Emerging properties of animal gene regulatory networks. Nature. 2010;468(7326):911–20.
81. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002;298(5594):824–7.
82. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol. 2004;14(3):283–91.
83. Lazebnik Y. Can a biologist fix a radio?-or, what I learned while studying apoptosis. Cancer Cell. 2002;2(3):179–82.
84. Gao S, Moreno M, Eliason S, Cao H, Li X, Yu W, Bidlack FB, Margolis HC, Baldini A, Amendt BA. TBX1 protein interactions and microRNA-96-5p regulation controls cell proliferation during craniofacial and dental development: implications for 22q11.2 deletion syndrome. Hum Mol Genet. 2015;24(8):2330–48.
85. Di Rocco F, Biosse Duplan M, Heuze Y, Kaci N, Komla-Ebri D, Munnich A, Mugniery E, Benoist-Lasselin C, Legeai-Mallet L. FGFR3 mutation causes abnormal membranous ossification in achondroplasia. Hum Mol Genet. 2014;23(11):2914–25.
86. Bantscheff M, Scholten A, Heck AJ. Revealing promiscuous drug–target interactions by chemical proteomics. Drug Discov Today. 2009;14(21):1021–9.
87. Schneider G. Virtual screening: an endless staircase? Nat Rev Drug Discov. 2010;9(4):273–6.
88. Torkamani A, Verkhivker G, Schork NJ. Cancer driver mutations in protein kinase genes. Cancer Lett. 2009;281(2):117–27.
89. Wu CC, Kannan K, Lin S, Yen L, Milosavljevic A. Identification of cancer fusion drivers using network fusion centrality. Bioinformatics. 2013;29(9):1174–81.
90. Ma WW, Adjei AA. Novel agents on the horizon for cancer therapy. CA Cancer J Clin. 2009;59(2):111–37.
91. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92.
92. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013;108(3):479–85.
93. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, Thurlimann B, Senn HJ. Tailoring therapies–improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol. 2015;26(8):1533–46.
94. Osborne CK, Schiff R. Mechanisms of endocrine resistance in breast cancer. Annu Rev Med. 2011;62:233–47.
95. Goldhirsch A, Winer E, Coates A, Gelber R, Piccart-Gebhart M, Thürlimann B, Senn H-J, Albain KS, André F, Bergh J. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206–23.
96. Wood AJ, Riggs BL, Hartmann LC. Selective estrogen-receptor modulators—mechanisms of action and application to clinical practice. N Engl J Med. 2003;348(7):618–29.
97. Hudis CA. Trastuzumab–mechanism of action and use in clinical practice. N Engl J Med. 2007;357(1):39–51.
98. Arpino G, De Angelis C, Giuliano M, Giordano A, Falato C, De Laurentiis M, De Placido S. Molecular mechanism and clinical implications of endocrine therapy resistance in breast cancer. Oncology. 2010;77(Suppl 1):23–37.

99. Argiris A, Wang CX, Whalen SG, DiGiovanna MP. Synergistic interactions between tamoxifen and trastuzumab (Herceptin). Clin Cancer Res. 2004;10(4):1409–20.
100. Johnston S, Pippen J, Pivot X, Lichinitser M, Sadeghi S, Dieras V, Gomez HL, Romieu G, Manikhas A, Kennedy MJ. Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor–positive metastatic breast cancer. J Clin Oncol. 2009;27(33):5538–46.
101. Kaufman B, Mackey JR, Clemens MR, Bapsy PP, Vaid A, Wardley A, Tjulandin S, Jahn M, Lehle M, Feyereislova A. Trastuzumab plus anastrozole versus anastrozole alone for the treatment of postmenopausal women with human epidermal growth factor receptor 2-positive, hormone receptor-positive metastatic breast cancer: results from the randomized phase III TAnDEM study. J Clin Oncol. 2009;27(33):5529–37.
102. Miller TW, Balko JM, Arteaga CL. Phosphatidylinositol 3-kinase and antiestrogen resistance in breast cancer. J Clin Oncol. 2011;29(33):4452–61.
103. Sabnis G, Goloubeva O, Jelovac D, Schayowitz A, Brodie A. Inhibition of the phosphatidylinositol 3-kinase/Akt pathway improves response of long-term estrogen-deprived breast cancer xenografts to antiestrogens. Clin Cancer Res. 2007;13(9):2751–7.
104. Wolff AC, Lazar AA, Bondarenko I, Garin AM, Brincat S, Chow L, Sun Y, Neskovic-Konstantinovic Z, Guimaraes RC, Fumoleau P. Randomized phase III placebo-controlled trial of letrozole plus oral temsirolimus as first-line endocrine therapy in postmenopausal women with locally advanced or metastatic breast cancer. J Clin Oncol. 2013;31(2):195–202.
105. Dees EC, Carey LA. Improving endocrine therapy for breast cancer: it's not that simple. J Clin Oncol. 2013;31(2):171–3.
106. Baselga J, Campone M, Piccart M, Burris HA III, Rugo HS, Sahmoud T, Noguchi S, Gnant M, Pritchard KI, Lebrun F. Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. N Engl J Med. 2012;366(6):520–9.
107. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27(2):199–204.
108. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA. The sequence of the human genome. Science. 2001;291(5507):1304–51.
109. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3(1):140.
110. Leiserson MD, Vandin F, Wu H, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106–14.
111. Westbrook K, Stearns V. Pharmacogenomics of breast cancer therapy: an update. Pharmacol Ther. 2013;139(1):1–11.
112. Schubert M, Iorio F. Exploiting combinatorial patterns in cancer genomic data for personalized therapy and new target discovery. Pharmacogenomics. 2014;15(16):1943–6.
113. Barrenas F, Chavali S, Alves AC, Coin L, Jarvelin M-R, Jornsten R, Langston MA, Ramasamy A, Rogers G, Wang H. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. Genome Biol. 2012;13(6):R46.
114. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
115. Brown KR, Otasek D, Ali M, McGuffin MJ, Xie W, Devani B, Toch IL, Jurisica I. NAViGaTOR: network analysis, visualization and graphing toronto. Bioinformatics. 2009;25(24):3327–9.
116. Hu Z, Snitkin ES, DeLisi C. VisANT: an integrative framework for networks in systems biology. Brief Bioinform. 2008;9(4):317–25.
117. Matsuoka Y, Funahashi A, Ghosh S, Kitano H. Modeling and simulation using cell designer. Methods Mol Biol. 2014;1164:121–45.
118. Bonnet E, Viara E, Kuperstein I, Calzone L, Cohen DP, Barillot E, Zinovyev A. NaviCell Web service for network-based data visualization. Nucleic Acids Res. 2015;43(W1):W560–5.
119. Gogoshin G, Boerwinkle E, Rodin AS. New algorithm and software (BNOmics) for inferring and visualizing bayesian networks from heterogeneous big biological and genetic data. J Comput Biol. 2016;23:1–17. doi:10.1089/cmb.2016.0100.
120. Thomas S, Bonchev D. A survey of current software for network analysis in molecular biology. Hum Genomics. 2010;4(5):353–60.