Cancer Causes & Control (2022) 33:1107–1120
https://doi.org/10.1007/s10552-022-01594-7

ORIGINAL PAPER

Proceedings of the fifth international Molecular Pathological Epidemiology (MPE) meeting

Song Yao[1] · Peter T. Campbell[2] · Tomotaka Ugai[3] · Gretchen Gierach[4] · Mustapha Abubakar[4] · Viktor Adalsteinsson[5] · Jonas Almeida[4] · Paul Brennan[6] · Stephen Chanock[4] · Todd Golub[5,7] · Samir Hanash[8] · Curtis Harris[9] · Cassandra A. Hathaway[10] · Karl Kelsey[11] · Maria Teresa Landi[4] · Faisal Mahmood[12] · Christina Newton[13] · John Quackenbush[14,15] · Scott Rodig[12] · Nikolaus Schultz[16] · Guillermo Tearney[17] · Shelley S. Tworoger[10] · Molin Wang[3,14,15] · Xuehong Zhang[14,18] · Montserrat Garcia-Closas[4] · Timothy R. Rebbeck[19] · Christine B. Ambrosone[1] · Shuji Ogino[3,5,12]

Received: 15 October 2021 / Accepted: 20 May 2022 / Published online: 27 June 2022

Abstract

Cancer heterogeneities hold the key to a deeper understanding of cancer etiology and progression and the discovery of more precise cancer therapy. Modern pathological and molecular technologies offer a powerful set of tools to profile tumor heterogeneities at multiple levels in large patient populations, from DNA to RNA, protein and epigenetics, and from tumor tissues to tumor microenvironment and liquid biopsy. When coupled with well-validated epidemiologic methodology and well-characterized epidemiologic resources, the rich tumor pathological and molecular tumor information provide new research opportunities at an unprecedented breadth and depth. This is the research space where Molecular Pathological Epidemiology (MPE) emerged over a decade ago and has been thriving since then. As a truly multidisciplinary field, MPE embraces collaborations from diverse fields including epidemiology, pathology, immunology, genetics, biostatistics, bioinformatics, and data science. Since first convened in 2013, the International MPE Meeting series has grown into a dynamic and dedicated platform for experts from these disciplines to communicate novel findings, discuss new research opportunities and challenges, build professional networks, and educate the next-generation scientists. Herein, we share the proceedings of the Fifth International MPE meeting, held virtually online, on May 24 and 25, 2021. The meeting consisted of 21 presentations organized into the three main themes, which were recent integrative MPE studies, novel cancer profiling technologies, and new statistical and data science approaches. Looking forward to the near future, the meeting attendees anticipated continuous expansion and fruition of MPE research in many research fronts, particularly immune-epidemiology, mutational signatures, liquid biopsy, and health disparities.

Keywords Molecular pathological epidemiology · Meeting report · Meeting proceedings · Meeting summary · Immuno-epidemiology

Song Yao, Peter T. Campbell, Tomotaka Ugai, and Gretchen Gierach have contributed equally to this work as co-first authors.

Montserrat Garcia-Closas, Timothy R. Rebbeck, Christine B. Ambrosone, and Shuji Ogino have contributed equally to this work as co-senior authors.

Peter T. Campbell, Montserrat Garcia-Closas, Timothy R. Rebbeck, Christine B. Ambrosone, and Shuji Ogino have contributed equally to this work as the Co-Chairs of this meeting.

Extended author information available on the last page of the article

Abbreviations

AI	Artificial intelligence
AIPW	Augmented inverse probability weighting
BBD	Benign breast disease
BMI	Body mass index
cfDNA	Cell-free DNA
CFR	Cancer family registry
CLAM	Clustering-constrained-attention multiple-instance learning
COSMIC	Catalogue of somatic mutations in cancer
DepMap	Dependency map
DμOCT	Dynamic micro-optical coherence tomography

ER	Estrogen receptor
ESCC	Esophageal squamous cell carcinoma
GECCO	Genetics and Epidemiology of Colorectal Cancer Consortium
GWAS	Genome-wide association study
HRS	Hodgkin Reed-Sternberg
H&E	Hematoxylin and eosin
LCINS	Lung cancer in never smokers
LIONESS	Linear interpolation to obtain network estimates for single samples
MAR	Missing at random
MDR	Multi-drug resistance
MetMap	Metastasis map
MHC	Major histocompatibility complex
MPE	Molecular pathological epidemiology
MSI	Microsatellite instability
NGS	Next-generation sequencing
PANDA	Passing attributes between networks for data assimilation
PD-1	Programmed cell death 1
PDAC	Pancreatic ductal adenocarcinoma
PD-L1	Programmed cell death 1 ligand 1
PD-L2	Programmed cell death 1 ligand 2
PTC	Papillary thyroid carcinoma
RTK	Receptor tyrosine kinase
SBS	Single base substitution
SIRE	Self-identified race and ethnicity
TMA	Tissue microarray
TOAD	Tumor origin assessment via deep learning

Introduction

Cancer epidemiology has a long history of success in establishing relationships between exposures and cancer development in population-based settings. The statistical associations initially identified through epidemiologic studies often rely on laboratory-based experimental studies to illustrate the underlying biological mechanisms. In the last few decades, however, the integration of biospecimens into epidemiologic studies and the advent of high-throughput genomic profiling technologies, have enabled molecular epidemiologists to increasingly take a "driver seat" in conducting mechanistic studies at an ever-expanding scope and ever-refining biological depth. The boundaries across epidemiology, genetics, statistics, and clinical and translational research have come down, giving rise to molecular epidemiology as a multidisciplinary tool to lead the forefront of our exploration into the complexity of cancer etiology and outcomes [1]. It was under this "melting pot" backdrop that the concept of molecular pathological epidemiology (MPE) was first proposed to reflect the focus on the heterogeneity of the cancer pathology and genomics [2–7].

There are two main types of cancer heterogeneity on which MPE research currently focuses. The first is etiological heterogeneity, i.e., disease subgroups arising through distinct causal pathways driven by external environmental factors and internal host genetic background. The inter-individual diversity in exposures, biological processes to internalize and respond to such exposures, plus stochastic variations during cell division and proliferation, dictate that no two cancers arise following the exact same pathway. Different tumorigenic processes leave unique characteristic imprints on cancer genome and pathology, which can be profiled in tumor tissues, and subsequently grouped based on shared similarities and related to exposome data for etiological inference. The second type is prognostic heterogeneity, i.e., disease subgroups behaving distinctively after diagnosis in cancer progression or response to cancer therapy, which again is likely shaped by a multitude of factors and the interplay between tumor and host and between tumor cells and their local microenvironment including infiltrating immune cells, the importance of which has been increasingly appreciated thanks to the recent successes of cancer immunotherapy.

Our understanding of cancer heterogeneity is driven largely by the advances in pathological and molecular profiling technologies, beginning with morphological evaluation and immunohistochemical staining, followed by high-density microarrays, and then more recently next-generation sequencing (NGS) approaches and spatial imaging analysis and others. The rapid emergence and evolution of new technologies open doors to numerous unprecedented research opportunities for cancer epidemiologists. The transdisciplinary nature of MPE research makes it a quintessential team science that benefits from a broad, vibrant, and engaging community consisting of investigators from a wide range of disciplines. The International MPE Meeting Series provides a dedicated platform for this community to convene and exchange ideas, to communicate discoveries and challenges, and to network [8–10].

The meeting series has grown from the inaugural local meeting of 10 investigators at the Harvard School of Public Health in April 2013 to more than 200 attendees from 16 countries in its fourth meeting in May/June 2018 held at Dana-Farber Cancer Institute in Boston, MA, USA. Due to the COVID-19 pandemic, the Fifth International MPE Meeting originally scheduled in 2020 was postponed to May 2021 and the meeting was held virtually through a teleconference platform. The meeting continued its tradition of being open and free to the research community and attracted more than 490 registrants from 21 countries around the world. Herein, we share the proceedings from this two-day meeting with 21 presentations including two keynote lectures. Moreover, three Meet-the-Experts sessions were also held virtually to provide opportunities for the meeting attendees to communicate directly with the speakers; these sessions were well attended with live discussions. Consistent with our intent to provide an open forum for the broad research community to communicate and discuss a wide-range of studies in which MPE principles may apply, the proceedings are organized into three broad themes, consisting of integrative MPE studies, novel cancer profiling technologies, and new statistical and data science approaches. A list of speakers and presentation titles are provided in Table 1. For clarity and unambiguous communications, we use HUGO Gene Nomenclature Committee-approved symbols for genes and gene products (proteins) along with common colloquial names in parenthesis if appropriate, following the standardized nomenclature recommended by an expert panel [11].

Theme 1: Integrative MPE studies in the "driver seat" to advance our understanding of cancer etiologies and racial disparities

Dr. Peter Campbell opened the meeting with a presentation that summarized efforts aimed at integrating genetic and MPE approaches toward a better understanding of the connection between obesity and colorectal cancer risk. High body mass index (BMI) has been an established risk factor for colorectal cancer for more than 15 years; however, associations are often higher for tumors that occur in the colon than the rectum and associations are also higher in men than in women. This concept of etiologic heterogeneity led he and colleagues to investigate potential heterogeneity between BMI and colorectal cancer risk and prognosis in the Colorectal Cancer Family Registry (Colon-CFR) according to tumor microsatellite instability (MSI) status [12, 13]. In

Table 1 Presentations at the Fifth International Molecular Pathological Epidemiology (MPE) Meeting

Speaker	Presentation title
Theme 1: Integrative MPE studies	
Peter Campbell	Obesity and colorectal cancer: new insights from genetic and molecular epidemiology
Shuji Ogino	Integration of immunology into molecular pathological epidemiology
Shelley Tworoger	Integration of epidemiological factors and immunological markers: ovarian cancer as a use case
Mustapha Abubakar	Computational pathology of stromal morphology in relation to breast cancer risk
Christine Ambrosone	Molecular mechanisms underlying relationships between parity, breastfeeding and aggressive breast cancer: part II
Timothy Rebbeck	Is there a biological basis for cancer disparities?
Paul Brennan	Can we identify novel causes of esophageal squamous cell cancer by studying mutational signatures
Maria Teresa Landi	Genomic and evolutionary classification of lung cancer in never smokers
Stephen Chanock	Radiation-related genomic profile of papillary thyroid cancer after the chernobyl accident
Theme 2: Novel cancer profiling technologies	
Todd Golub	Perspectives on cancer precision medicine
Guillermo Tearney	Dynamic micro-optical coherence tomography for cellular tissue phenotyping
Scott Rodig	Classic hodgkin lymphoma as a model system to study immune evasion in cancer
Faisal Mahmood	Data-efficient and multimodal computational pathology
Curtis Harris	Metabolome of lung cancer
Viktor Adalsteinsson	Tracing tumor signatures from plasma cell-free DNA
Samir Hanash	How liquid biopsy can inform tumor development and progression for pancreatic cancer
Karl Kelsey	Defining the methylation profile of lymphocyte memory yields an enhanced library for deconvolution of peripheral blood
Theme 3: New statistical and data science approaches	
Nikolaus Schultz	Interpreting genomic alterations and therapeutic implications using oncokb and the cbioportal for cancer genomics
Jonas Almeida	Integrating data, from digital to molecular pathology—let the code do the travelling
John Quackenbush	Why bother with networks
Molin Wang	New development in methods for dealing with missing subtype data

a 2010 publication, they showed that high BMI was associated with the more-common non-MSI-high tumors, but not with MSI-high tumors [12]. Given that the Colon-CFR is enriched for patients with Lynch syndrome, these results suggested that BMI is not a risk factor for MSI-high colorectal cancers due to Lynch syndrome but may not be generalizable to MSI-high tumors due to methylation of *MLH1*. Dr. Campbell then presented more recent, unpublished work, from the large Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) of more than 10,000 cases with findings that BMI was indeed associated with non-MSI-high and MSI-high tumors due to methylation, but not with tumors consistent with Lynch syndrome. He also presented additional unpublished work from the GECCO consortium in regard to high BMI and associations with specific mutations in tumor tissues, including three examples in which BMI was more convincingly associated with tumors with or without specific mutations in genes or pathways known to regulate energy balance or metabolism. In the final part of his talk, Dr. Campbell presented results from a recent genome-wide association study (GWAS) x BMI publication where BMI was more strongly associated with colorectal cancer risk among women with certain variants at a common *SMAD7* locus [14]. Since SMAD7 is known to influence other bowel diseases, the authors would like to use an MPE approach in the future to identify any potential Gene x environment x tumor molecular phenotype associations in this context. Such an approach will necessitate considerable resources and broad collaborations to amass the number of cases and the amount of data necessary for this endeavor.

Dr. Shuji Ogino's lecture provided another high-level overview of the status of MPE research, highlighting the integration of immunology into MPE as a major development in this field. Cancer immunology has been rapidly advancing, and immunoprevention and immunotherapy strategies have a great potential to reduce the cancer burden. Neoplasms and cancers represent heterogeneous pathological processes due to interactive influences of the exposome (including the microbiome), the immune system, and neoplastic cells. To address this, investigators can examine the influences of exposures on tumor-immune interactions using the MPE research framework that can link the exposures with tumor pathological signatures. This is the so-called "immunology-MPE" approach [15]. Using archival tumor tissue of over 1,500 colorectal cancers in the Nurses' Health Study and the Health Professionals Follow-up Study, Dr. Ogino and colleagues conducted several proof-of-principle studies to provide evidence for influences of smoking [16], aspirin [17], vitamin D [18], inflammatory diet [19], and marine omega-3 fatty acids [20] on tumor immunity and cancer incidence. Recently, Dr. Ogino and collaborators further developed and validated multiplex immunofluorescence assays to in-depth phenotype immune cells [21, 22], as well as microbial assays for putative cancer pathogens. The immunology-MPE approach is also expected to advance research on early-onset cancers [23]. The new research paradigms to integrate microbial and immune assays on archival tumor tissue in large-scale population studies can provide possible paths for precision prevention and public health.

Dr. Shelley Tworoger spoke about her group's efforts to evaluate the risk factors that predict tumor immunological response in ovarian cancer, using tumor tissue microarrays (TMA), which are well suited for large epidemiologic studies. Dr. Tworoger discussed several operational issues that researchers face. First, she spoke about missing or folded tissue in the TMA cores, which impacts accounting for total evaluable area and thus the density of immune markers. Special consideration is needed when preparing samples and the image analysis (e.g., using semi-automated segmentation) to ensure this is minimized and post hoc analyses are necessary to adjust for area of the sample. Additionally, immunological markers can often be sparse or only observed in a small proportion of tumors; as such, many markers do not follow a normal distribution. This leads to statistical challenges that can be mitigated by using categories, present/absent values, and specialized models (e.g., beta-binomial models). Further, correlations between TMA cores within the same person are lower for immune markers that are rarer, possibly necessitating use of full slides for some cell types. Finally, batch effects across TMAs may be present, which can be attributable to differences in antigenicity of the sample and is often lower in older samples. Such batch effects may require imaging reanalysis and statistical methods for batch correction. When thinking about the relationship between host exposures and tumor-immune response, Dr. Tworoger discussed the importance of evaluating factors that may not be necessarily associated with cancer risk, as such factors that could influence tumor-immune response independent of tumorigenesis. For example, in preliminary data, her team saw no association of early life abuse and ovarian cancer risk, but those who experienced early life abuse versus not had suggestively decreased helper T cells and B cells in their ovarian tumors. Overall, there is a need to understand exposures that affect the tumor-immune microenvironment to increase our understanding of cancer development and progression. To do this, novel epidemiological methods need to be developed to evaluate the tumor-immune milieu especially considering sample preparation, batch variation, and marker distributions.

Dr. Mustapha Abubakar's presentation demonstrated how MPE strategies could be adapted to identify novel tissue biomarkers for predicting risk of future invasive breast cancer development among women with pre-cancerous lesions. By integrating computational pathology and epidemiology, Dr. Abubakar's group conducted a case–control study nested in a large cohort of women who were biopsied for benign breast disease (BBD) at Kaiser Permanente Northwest (1971–2006) and followed through mid-2015 [24]. Patients who developed incident invasive breast cancer at least one year after BBD diagnosis and those who did not were matched on BBD diagnosis age and plan membership duration. By applying supervised machine learning algorithms to digitized H&E-stained slides, they generated quantitative tissue composition metrics, including epithelium, stroma, and adipose tissue, and determined their association with future invasive breast cancer diagnosis, overall and by BBD histological classification. They found that increasing epithelial area on BBD biopsy was associated with increased breast cancer risk, irrespective of BBD histological classification [25]. Conversely, increasing stroma area was associated with decreased risk in non-proliferative disease (NPD) but with increased risk in proliferative disease (PD), supporting a context-dependent role of the stroma to either prevent or promote tumor formation. A metric of the proportion of fibroglandular tissue that is epithelium, relative to stroma, i.e., epithelium-to-stroma proportion (ESP) was independently and strongly predictive of increased breast cancer risk. In combination with mammographic breast density (MBD), women with high ESP and high MBD had substantially higher breast cancer risk than those with low ESP and low MBD. The findings were particularly striking for women with NPD (comprising approximately 70% of all BBD patients), for whom relevant predictive biomarkers of subsequent breast cancer development are lacking. These findings could thus have important implications for risk stratification and clinical management of women with NPD upon breast biopsy.

Epigenetic alterations, including DNA methylation, are widespread in tumor genomes. Given the dynamics and versatility of DNA methylation regulation in response to changes in internal and external environment, it may provide a window into the biological effects of etiological factors on cancer genome. Dr. Christine Ambrosone presented the newest work from her group, following up on findings presented at the MPE meeting in 2018, which showed differential DNA methylation in breast tumors from Black and White women, particularly in *ESR1* (estrogen receptor 1, ER)-negative cancer in Black women. Because one of the top differentially methylated loci*, FOXA1*, was highly methylated in women who had children but did not breastfeed, and is important to differentiation of luminal cell progenitors [26], Dr. Ambrosone's group sought to verify findings using a number of approaches. Using immunohistochemistry for FOXA1 protein, 1,329 breast tumors were stained and results showed that FOXA1 expression was lower in ESR1 (ER)-negative tumors and was lowest in parous women, with the relationship attenuated among women who breastfed [27]. In another study based on women without breast cancer from the Komen Tissue Bank, breast tissues from 52 Black women who did not have children, 53 who were parous and did not breast feed, and 51 who breastfed their children were subjected to a targeted bisulfite sequencing approach using SureSelect Methyl-Seq. Results showed similar trends as those in the breast tumors, with lowest methylation in nulliparous women and higher in parous women who did not breastfeed. To further address how methylation or downregulation of FOXA1 may affect progenitor cell pools in the breast, the team used a *Foxa1*-knockout mouse and created strains to obtain experimental genotypes, dissecting the mammary glands and using flow cytometry to separate epithelial cell populations. Depletion of *Foxa1* led to dramatic changes in proportions of mammary gland epithelial cell populations—with abnormal accumulation of differentiation-arrested luminal progenitors and marked decreases in the number of ESR1 (ER)-positive cells [28]. Mouse models were also developed to mimic the human reproductive scenarios. For the virgin mice, luminal progenitors comprised 40% of the cell composition, which was increased to 50% in the parous mice. This increase in luminal progenitors was reduced when mice were allowed to nurse their pups, consistent with the hypothesis that parity results in increases in luminal progenitors, which are lowered with breastfeeding. Together, these findings could provide an epigenetic mechanism for the higher prevalence of ESR1 (ER)-negative breast cancer in Black women.

Cancer health disparities have been well described in the literature, which may have multi-level causes with both biological and non-biological factors, as well as the interactions between the two, at play. Dr. Rebbeck posed the question "is there a biological basis for cancer disparities?". He noted that Self-Identified Race and Ethnicity (SIRE) is a social, not biological, construct. SIRE is a manifestation of numerous underlying complex correlates including social factors such as culture, behavior, and environment. These factors are driven by historic systemic racism and other sociopolitical features. Ancestry is also correlated with SIRE, and includes genomic architecture and phenotypes determined by continental origin. Genomic architecture overall and the distribution of disease-associated genetic variation is also known to vary across SIRE and ancestral groups. In African Americans, these associations are complex, and are driven by distant evolutionary and population genetic forces as well as more recent admixture due to the transatlantic slave trade. Dr. Rebbeck presented data that suggests genomic architecture of cancer susceptibility and molecular signatures in tumors vary by SIRE. While this information has important implications for our understanding of the etiology of disease, risk assessment, and applications of precision medicine, these data do not imply that SIRE is a biological construct. Instead, diversity in etiology, prevention and treatment by SIRE can both inform our understanding of cancer etiology and lead to improved application of genomics into clinical and public health practice.

Three presentations at the meeting provided a timely update on the burgeoning field of mutational signatures, which are a definite number of patterns of nucleotide substitution within trinucleotide contexts that can be deciphered mathematically from tumor mutation data. Some of those signatures have been linked to known cancer risk factors and endogenous biological processes, thus providing a "forensic" tool to detect cancer causes not only at an aggregated population level but potentially at an individual level.

Dr. Paul Brennan presented on the findings of esophageal squamous cell carcinoma (ESCC) from the Mutographs project. ESCC has a remarkable geographic variation with high incidence in regions of Asia, Africa, and South America, yet the variation cannot be fully explained by known lifestyle and environmental risk factors. A total of 552 ESCC cases from eight countries with varying incidence rates were whole-genome sequenced, which revealed similar mutational profiles across all countries studied. Eight single base substitution (SBS) mutational signatures dominate within each country, led by APOBEC associated mutational signatures SBS2 and SBS13 found in 88% and 91% of cases, respectively. Several etiologic associations were identified, including SBS3 with deleterious BRCA1/BRCA2 variants, SBS16 with alcohol consumption, a novel T > C signature (SBS288J) with long term opium use, all of which had modest impact on mutation burden; yet no association was found between any of the mutational signatures with other major risk factors for ESCC, including hot drinks, indoor air pollution, and poor diet. As a result, no mutational signatures linked with an exogenous exposure could explain the geographic variation in ESCC incidence. These findings suggest that not all carcinogens generate distinct mutational signatures or increase mutation burden, whereas most mutations arise from tissue specific endogenous processes.

Dr. Maria Teresa Landi spoke about her group's efforts analyzing the pilot dataset in the Sherlock-Lung study, aiming to elucidate the mutational landscape and etiology of lung cancer in never smokers (LCINS) [29]. Lung cancer is the leading cause of cancer death, with LCINS accounts for 10–25% of the disease burden. Despite its prevalence, the genomic landscape of LCINS is not well-characterized. The analyses of high-coverage whole-genome sequencing of 232 LCINS revealed three molecular subtypes defined by distinct copy number aberrations. While the dominant subtype ('*piano*') is characterized by quiet copy number profiles, the other subtypes are associated with specific arm-level amplifications and *EGFR* mutations ('*mezzo-forte*'), and whole-genome doubling ('*forte*'). *Piano* tumors are characterized by *UBA1* mutations, germline *AR* variants, and stem-cell like features including low mutational burden, depleted *TP53* alterations, high intra-tumor heterogeneity, long telomeres, and slow growth with cancer driver genes acquired several years prior to tumor diagnosis. In contrast, driver mutations in *mezzo-forte* and *forte* tumors are generally late clonal events acquired close to tumor diagnosis, thus potentially facilitating target identification with a single biopsy. Future studies utilizing single-cell RNA-sequencing and genome-wide DNA methylation will be required to verify *piano* tumors stem cell-like state and identify cancer initiating events in those with no apparent drivers. Strong tobacco smoking signatures were not detected in LCINS, even in cases with exposure to second-hand tobacco smoke. Patients with *piano* subtypes overall displayed better survival, particularly those with carcinoids or *SETD2* mutations. While mutations in genes within the receptor tyrosine kinase (RTK)-RAS pathway have various impacts on survival, mutations in *TP53, CHEK2, EGFR*, or loss of 15q or 22q are associated with increased mortality. These genomic alterations in *forte* and *mezzo-forte* subtypes, and stem cell-like features in *piano* tumors create avenues for personalized therapeutic strategies in LCINS.

The keynote lecture delivered by Dr. Stephen Chanock also delved into the newest effort to study tumor somatic changes in relation to exogenous exposures, in this case radioactive contaminants after the Chernobyl accident and incident papillary thyroid carcinoma (PTC) [30]. The radioactive fallout as a potential carcinogenic exposure led to increased PTC incidence in contaminated regions. The established etiological causality herein provides a rare opportunity to investigate the impact of the environmental exposure on PTC genomics. Tumor samples from a total of 440 PTC patients from Ukraine, including 359 exposed to radioactive ^{131}I during childhood or in utero and 81 unexposed children, were profiled using multi-omic platforms. With increased estimated radiation dose, there were an increased number of small deletions and simple structural variants, which are hallmarks of nonhomologous end-joining repair, suggesting DNA double-strand breaks as early events in radiation-caused PTC. Moreover, an estimated 94% of PTCs were driven by alterations in the mitogen-activated protein kinase (MAPK) pathway, with a radiation dose-dependent enrichment of fusion versus mutation drivers. The effects on small deletions, simple structural variants, and fusions were most prominent among patients with radiation exposure at younger age. In mutational signature analysis, 7 COSMIC SBS signatures and 6 insertion/deletion signatures were identified, a majority of which were attributable to two clock-like signatures, yet none were correlated with environmental radiation exposure. Analyses of de novo mutational signatures revealed no novel signatures specific for radiation, either.

The above three studies represent to date some of the most comprehensive and sophisticated endeavors to identify etiological causes of tumor somatic changes. The numerous discoveries that emerged from these studies are remarkable. The lack of a clear explanation to either the geographic variation in ESCC incidence or to the elevated PTC incidence following environmental radiation exposure by mutational signatures is similarly intriguing. A larger sample size of tumor genomic data with accurate and in-depth epidemiological annotation of exposome may be necessary in future studies.

Theme 2: Innovative technologies for in-depth tumor profiling

Modern pathological and genomic technologies have fundamentally advanced our understanding of cancer. No longer considered as one homologous disease entity, cancers manifest intra- and inter-tumor heterogeneities in almost every way they are dissected, by a scalpel, a microscope, or a DNA sequencer. When aggregated at a population-level in an epidemiological setting, these heterogeneities can be grouped based on shared features that provide a potential new understanding of cancer etiology and prognosis. The pace of the development and improvement of new technologies is impressive. As a result, we now have an unprecedented variety of tumor profiling tools at our disposal, and even newer ones are emerging. At this meeting, several speakers shared some of the most recent developments in technologies that may be adapted toward future MPE studies.

In a keynote lecture, Dr. Todd Golub presented his group's work on forging a path toward cancer precision medicine. He discussed the development of the Cancer Dependency Map (DepMap), a large-scale effort of the Broad Institute to systematically perturb human cancer models genetically and pharmacologically, thereby defining the vulnerabilities of each cell line, and predictive biomarkers of such vulnerabilities. Over 800 cancer cell lines have been comprehensively characterized at the DNA-level, RNA-level, protein-level, and metabolite-level. The cell lines were also subjected to genome-wide CRISPR/Cas9 loss of function screening, thereby identifying genetic dependencies of each model. Dr. Golub discussed the follow-up of two DepMap findings, including the dependency of ovarian cancer cells on the phosphate exporter XPR1, and the sensitivity of cells over-expressing the multi-drug resistance gene MDR to the compound tepoxalin, whose mechanism of cancer cell killing remains unknown. Moreover, Dr. Golub described his group's development of the PRISM barcoding method, whereby each cell line is molecularly barcoded with a unique 24 nucleotide sequence, thereby allowing for cell lines to be pooled together. Barcode abundance is then measured, either by sequencing or by hybridization of Luminex beads coupled to anti-barcode tags, before and after small molecule treatment. The PRISM method was used to characterize the response of more than 500 cell lines to each of more than 4,000 drugs from the Broad's drug repurposing library. Lastly, Dr. Golub described his group's efforts to extend the PRISM method to the in vivo setting, where the metastatic potential of 488 DepMap cell lines was characterized in immunodeficient mice, thereby creating a Metastasis Map (MetMap) with potential to reveal for the first time at scale, tumor-microenvironment interactions.

Dr. Guillermo Tearney discussed his group's most recent efforts developing novel technologies for cellular tissue phenotyping [31]. Living cells exhibit active intracellular molecular motion that reflect their functional states. Traditional microscopy techniques that solely capture high resolution static images of cells miss the opportunity to capture the wealth of information provided by complex intracellular activity. Dr. Tearney's group recently developed dynamic micro-optical coherence tomography (DμOCT), an extension of μOCT that achieves near-isotropic sub-cellular resolution in all three dimensions (2 μm lateral × 1 μm axial) for assessing the metabolism of cells in cross-sectional and 3D tissues. DμOCT substantially enhanced the contrast of cells and organelles while revealing stratified, depth-dependent dynamics in the epithelial layers by acquiring a time series of μOCT images and conducting power frequency analysis of the temporal fluctuations that arise from intracellular motion on a pixel-per-pixel basis. His group has expanded the application DμOCT to encompass imaging of human skin in vivo, evaluating responses to drugs delivered via implantable microdevice, and determining dose response of melanoma spheroids to anti-cancer drugs. The results demonstrated potential utility of DμOCT for cellular phenotyping across a wide range of tissue types and for diverse bioscience and biomedical applications. Future work will focus on validating DμOCT for discriminating disease, cell activation state, and response to therapy, as well as developing technology for conducting DμOCT in vivo inside the human body.

In the last several years, the rapid uptake of cancer immunotherapy in the clinical management of many cancers highlights the importance of continuous research efforts to deepen our understanding of tumor-immune interactions and heterogeneities in tumor microenvironment. Dr. Scott Rodig presented his group's findings using classical Hodgkin lymphoma as a model system to study immune evasion in cancer. Classical Hodgkin lymphoma is a malignancy affecting mostly young adults and the elderly. The disease-defining feature of classical Hodgkin lymphoma is the presence of Hodgkin Reed-Sternberg (HRS) cells, which have a very high tumor mutation burden and are highly immunogenic. However, these cells reside in a specialized immunosuppressive microenvironmental niche, characteristic of high expression of CD274 (PD-L1)/PDCD1LG2 (PD-L2) to suppress T-cell activation and loss of MHC class I/ B2M and/

or MHC class II to suppress antigen presentation. Using multiplex immunofluorescence and digital image analysis, Dr. Rodig's group found abundant CD274 (PD-L1)-positive tumor-associated macrophages that colocalized with CD274 (PD-L1)-positive HRS cells in tumor microenvironment. Further, CD274 (PD-L1)-positive TAMs were more likely to be in contact with T cells, and CD274 (PD-L1)-positive HRS cells were more likely to be in contact with CD4$^+$ T cells, a subset of which were positive for PDCD1 (PD-1) [32]. In another study from Dr. Rodig's group using multiplex immunofluorescence, enriched CTLA4-positive T cells that were in contact with HRS cells outnumbered PDCD1 (PD-1)-positive and LAG3-positive T cells. Moreover, in recurrent classical Hodgkin lymphomas despite therapy with PDCD1 (PD-1) blockade, CTLA4-positive cells were found to present and focally contact HRS cells, suggesting that patients refractory to PDCD1 (PD-1) blockade might benefit from CTLA4 blockade [33]. Similar findings were also made in T-cell/histiocyte-rich large B-cell lymphoma, which is an aggressive rare malignant B cells within a robust but ineffective immune cell infiltrate. Unbiased clustering of spatially resolved immune signatures revealed increased CD274 (PD-L1) expressing macrophages and PDCD1 (PD-1)$^+$ T cells in tumor-immune "neighborhoods" in T-cell/histiocyte-rich large B-cell lymphoma, which could be used to distinguish from related subtypes of B-cell lymphoma [34]. Lastly, Dr. Rodig also described a new workflow based on multiplex ion beam imaging (MIBI) for assessing the intact tumor microenvironment in diffuse large B-cell lymphoma, which has recapitulated their prior work and knowledge in diffuse large B-cell lymphoma not otherwise specified and T-cell/histiocyte-rich large B-cell lymphoma.

Dr. Faisal Mahmood's lecture focused on novel deep learning-based computational pathology methods his group developed and their applications in clinical research. Clustering-constrained-attention multiple-instance learning (CLAM) is a data-efficient and weakly supervised computational pathology method on whole-slide images that was developed to solve challenges often seen in deep-learning methods for pathology image analysis, including requirements for extensive annotation data from large datasets of WSIs and poor domain adaptation and interpretability. CLAM used attention-based learning to identify subregions of high diagnostic value to accurately classify whole slides and instance-level clustering over the identified representative regions to constrain and refine the feature space. In several applications of CLAM, Dr. Mahmood demonstrated superior performance for the subtyping of renal cell carcinoma and non-small-cell lung cancer, as well as for the detection of lymph node metastasis in breast cancer, in comparison to standard weakly supervised classification algorithms. CLAM was readily adaptable to independent test cohorts, varying tissue content, smartphone microscopy and standalone 3D printed microscopy [35]. To predict the origins for cancers of unknown primary, Dr. Mahmood's group developed another computational pathology algorithm, Tumor Origin Assessment via Deep Learning (TOAD), using routinely acquired histology slides. The model was first trained with whole-slide images of tumors with known primary origins to simultaneously identify the tumor as primary versus metastatic and to predict its origin. In the subsequent testing with tumors of known primary origins, TOAD achieved top-1 and top-3 accuracy as high as 83% and 96%, respectively. In another test with cancer cases of unknown origins where a differential diagnosis was assigned, predictions from the algorithm resulted in concordance for 61% of cases and a top-3 agreement of 82% [36]. In unpublished work from Dr. Mahmood's group, the algorithm was also successfully applied to assess endomyocardial biopsy, which achieved an accuracy comparable to human experts. These artificial intelligence (AI)-based computational pathology tools have the potential to be used in conjunction with or in lieu of ancillary tests for more accurate and efficient diagnosis.

In the last decade, there has been increasing research attention to the concept of liquid biopsy, which uses non-invasive collection of various types of body fluids, usually peripheral blood, for the detection, profiling, selecting treatment, and monitoring treatment response for certain types of cancer. Liquid biopsy has the potential to inform about tumor development and progression through the release of a multitude of cell derived materials from tumor cells, the tumor microenvironment and from a systemic response to tumor development and progression. Some of the liquid biopsy-based tests have already been clinically adapted. At this meeting, several speakers discussed findings of research on liquid biopsy from their groups.

Dr. Curtis Harris lectured on cancer-derived liquid biopsy metabolomics, which provided a non-invasive means of cancer screening. Correlation of the identified metabolites with specific cancers created biomarker profiles that can be utilized for diagnostic and prognostic evaluation of many types of human cancer. Through metabolomic analyses, creatine riboside has been identified and shown to be a companion-diagnostic biomarker in multiple types of human cancer including lung, liver, pancreas, breast, and brain cancers. In lung cancer, creatine riboside was increased in tumor tissues, and the urine levels were highly positively correlated with the levels of the metabolite in tumors and in blood. When paired with other identified urinary metabolite biomarkers, such as N-acetylneuraminic acid, creatine riboside was demonstrated with improved diagnostic capability and reliability for prognosis in lung cancer [37]. These foundational studies have validated the use of urinary metabolite screening that has led to further investigation into liquid biopsy-based biomarker in association with human cancer.

Dr. Viktor Adalsteinsson spoke about his group's efforts to enhance the sensitivity of liquid biopsies to detect minimal residual disease after cancer therapy. Firstly, he showed that minimal residual disease could be detected after curative intent treatment for breast cancer with up to 39 months of lead time to clinical recurrence by tracking up to hundreds of patient-specific tumor mutations in cell-free DNA (cfDNA) [38]. Then, he described three new technologies, including Duplex-Repair [39], MAESTRO [40], and CODEC [41], to maximize the accuracy and efficiency of mutation tracking in cfDNA and improve detection of MRD. Duplex-Repair addresses the challenge that existing sequencing preparation methods may copy base damage errors from one strand of a DNA duplex to both strands, rendering them indistinguishable from true mutations on both strands [39]. MAESTRO overcomes the high cost of rare mutation detection by converting low abundance mutations into high abundance mutations prior to sequencing [40]. CODEC converts existing NGS instruments into massively parallel 'single duplex' sequencers which can read both strands of each DNA duplex at 100-fold lower cost [41]. Lastly, he showed that by tracking one thousand individualized mutations, it is possible to resolve one part-per-million tumor DNA in cfDNA, which is up to 100-fold more sensitive than prior liquid biopsy tests. His team is now working to put these technologies together and apply them to larger clinical studies to determine whether enhanced analytical sensitivity translates into better detection of minimal residual disease, longer lead time to recurrence, and more precise therapy and improved outcomes for patients.

Dr. Samir Hanash's presentation focused on the use of liquid biopsy to study the dynamic changes associated with the development and progression of pancreatic cancer by implementing a mouse-to-human approach. Genetically engineered mouse models complement the use of human biospecimens as they overcome the substantial heterogeneity of human subjects that is unrelated to disease and the potential biases in collection of human samples. Moreover, mouse models allow sampling of mice at defined stages of tumor development and identification of markers linked to pathways for tumor initiation and progression that are turned on in the genetically engineered mouse models. This work led to the identification of circulating proteins that are released at the earliest PanIn stages of pancreatic cancer development [42]. Early on, his group identified 51 potential markers derived from mouse model studies, mass spectrometry studies of early-stage pancreatic cancer plasmas, and from the literature that were further screened to identify the most informative markers that are upregulated at early stages of pancreatic ductal adenocarcinoma (PDAC). Subsequent validation studies were focused on three markers consisting of CA19-9, LRG1, and TIMP1 to determine their lead-time trajectory for PDAC early detection using prospectively collected samples from the NCI Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial cohort [43]. Increases in marker levels compared to baseline were observed starting two years before diagnosis with a steep rise closer to diagnosis, pointing to the merit of monitoring biomarker levels in subjects at risk to detect PDAC at the earliest stages.

Dr. Karl Kelsey presented novel data defining the DNA methylation profile of lymphocyte memory, noting that this yields an enhanced library for deconvolution of peripheral blood. Epigenetic mechanisms, including DNA methylation, are critical drivers of immune cell lineage differentiation and activation. His group scanned genome-wide differences in DNA methylation among CD4 and CD8 naïve and memory cell states and combined this data with similar data on naïve and memory B-cell states. Overall, their findings were consistent with the literature describing the DNA methylome as a major driver of individual central and effector T-cell memory states as well as in memory B cells. Dr. Kelsey observed unusually large differences in DNA methylation in thousands of CpG sites in hundreds of genes associated with the development of memory in each lineage. The data similarly describe considerable overlap in genes with altered DNA methylation in the T-cell lineage, with primarily a loss of DNA methylation in over 125,000 CpGs significantly associated with the generation of central memory in both CD4 and CD8 cells. Furthermore, their analyses revealed specific CpG dinucleotides in both CD4 and CD8 cells whose methylation pattern is consistent with the circular model of memory generation. As evidence of common pathways in the generation of immune memory, they highlighted 22 gene loci, including several within the promoter region of the *AIM2* gene, with dramatically altered DNA methylation in all three memory lineages. The description of the immune memory profile also allowed for the enhancement of reference-based deconvolution of blood DNA methylation to include 12 leukocyte subtypes (neutrophils, eosinophils, basophils, monocytes, B cells, CD4$^+$ and CD8$^+$ naïve and memory cells, natural killer, and T regulatory cells). Including derived variables, this enhanced method provided up to 56 immune profile variables [44]. The IDOL (IDentifying Optimal Libraries) algorithm was used to identify libraries for deconvolution of DNA methylation data both for current and retrospective platforms [45]. The accuracy of deconvolution estimates obtained using these enhanced libraries was validated using artificial mixtures, and whole-blood DNA methylation with known cellular composition from flow cytometry. This pioneering work enabled a more detailed understanding of lymphocyte memory, as well as enhanced representation of immune-cell profiles in blood, using only DNA and facilitates a standardized, thorough investigation of the immune system in human health and disease.

Theme 3: Novel statistical and data science approaches for MPE research

One of the transformative changes in MPE research is the integration of multi-level, high-dimension molecular and pathological data into the framework of cancer epidemiology, which requires a full embrace of biostatistics, bioinformatics, and data sciences. Several speakers at this meeting shared novel methodological and resource development in this space.

Dr. Nikolaus Schultz discussed his group's efforts to develop methods and resources for the interpretation of genomic variants in cancer. With prospective clinical sequencing of tumors emerging as a mainstay in cancer care, there is an urgent need for clinical support tools that identify the clinical implications associated with specific mutation events. To this end, his group has developed several tools for the interpretation and visualization of cancer variants, enabling researchers and clinicians to make discoveries and treatment decisions: (1) Cancer Hotspots is a method and resource that identifies recurrently mutated cancer genes resulting in amino acid substitutions [46]. These variants, so-called hotspots, are more likely to be drivers in cancer. (2) OncoKB is a precision oncology knowledge base that annotates the biologic and oncogenic effects as well as prognostic and predictive significance of somatic molecular alterations [47]. Potential treatment implications are stratified by the level of evidence that a specific molecular alteration is predictive of drug. (3) The cBioPortal for Cancer Genomics is a web-based analysis tool for the visualization and analysis of cancer variants [48]. Through its intuitive interface it makes complex cancer genomics data easily accessible by researchers and clinicians without bioinformatics experience. It integrates information from Cancer Hotspots and OncoKB to enable the identification of potential driver mutations and therapeutic options. These resources are used routinely at Memorial Sloan Kettering Cancer Center in clinical sequencing and by countless cancer genomic researchers and clinicians around the globe.

Dr. Jonas Almeida described distributed computational systems designed for data integration, from Digital to Molecular Pathology, by having the code, rather than the data do the traveling between sensitive, user-governed, sub-systems. Primarily, these integrative designs were developed to address the logistics of reusable analytical workflows satisfying principles of FAIR stewardship of scientific data [49]. Just as important, however, they do so with the data remaining under the control, and compliance, of various stake holders. The integration of molecular pathology data with whole-slide images was illustrated with hands-on demonstrations of AI tools applied to TCGA data at https://mathbiol.github.io/tcgatil [50], and segmentation by gene mutation [51]. Finally, this approach was shown to be applicable to a variety of data integration systems developed as Data Commons, as illustrated by an application to real-time tracking of mortality by COVID-19 at https://episphere.github.io/mortalitytracker [52]. In conclusion, Dr Almeida argued that data science infrastructure is becoming available to many research organizations as part of cloud computing platforms, such as those made available through NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative, promising far more scalable, governable, and user-friendly approaches to the development of integrative "AI-first" pathology data commons.

Dr. John Quackenbush discussed the importance of studying gene regulatory networks in addition to analyzing gene expression and described his group's development and application of a suite of system biology tools to identify drivers of disease and therapeutic targets. Although differential expression and co-expression analysis is often used to identify genes associated with disease states, Dr. Quackenbush argued that these do not explain what processes drive the observed expression differences and the disease itself—factors that might be missed in analysis focusing on single genes. Instead, by identifying gene regulatory networks linking transcription factors (or other regulators) to their target genes, one can find key regulators that are central to the diverse processes active in disease development, progression, and response to therapy. He presented several network inference algorithms from his group, including Passing Attributes between Networks for Data Assimilation (PANDA) [53] and Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) [54]. As an example of network inference analysis, when PANDA and LIONESS were compared to differential expression and co-expression methods to analyze transcriptome data of pancreatic cancer in TCGA, several major biological processes, including immune-related processes, epigenetic, and cell cycle process were identified only when using the network inference algorithms. Dr. Quackenbush also described applications of network analysis to investigate sexual dimorphism in multiple tissues and colorectal cancer and an analysis of glioblastoma multiforme that found differential regulation of processes including PDCD1 (PD-1) signaling that were associated with survival. Lastly, Dr. Quackenbush described netZoo, an integrated collection of gene regulatory network inference and analysis tools that he and his group developed. A collection of more than 180,000 gene regulatory networks generated using the netZoo tools to analyze human tissues, cancers, cell lines, and small molecule drug perturbation are curated in an online database, GRAND. Included in GRAND are a variety of search tools and analyses, including tools to identify drug that alter specific regulatory processes [55].

Dr. Molin Wang spoke about her group's efforts in developing analytical methods for addressing the selection bias problem caused by missing tumor marker data in MPE analyses. The high percentage of cases with tumor maker data missing is often due to tissue unavailability or insufficient quality of tumor tissues. Standard data analysis methods using only complete data can lead to biased estimates and misleading scientific findings. When disease subtypes are classified by multiple markers, Wang's group has developed an augmented inverse probability weighting (AIPW) Cox proportional hazard model method to evaluate the effect of exposures on disease subtypes in the presence of partially or completely missing biomarkers. The AIPW method is valid under the missing at random (MAR) assumption, is typically more efficient than the IPW method, and enjoys the double robustness property, which means that the method leads to valid estimates and inferences even if one of the missingness probability model or the tumor marker model is mis-specified. The MAR assumption may often be achieved by including auxiliary tumor variables, such as tumor stage, and tumor location, in the missingness probability and tumor maker models. However, the MAR assumption cannot be verified empirically using the observed data. To address the potential issue of missingness not at random, Wang's group has developed a partial likelihood-based method to obtain valid estimates for the effect of exposures on disease subtypes. This method can often be used as a sensitivity analysis. R functions implementing the methods are available at Dr. Wang's software page (https://www.hsph.harvard.edu/molin-wang/software/).

Conclusion

After being delayed by a year and migrated to a virtual platform due to the COVID-19 pandemic, the Fifth International Molecular Pathological Epidemiology Meeting drew the largest audience in the meeting history. The feedback received from meeting attendees via a post-meeting survey was overwhelmingly positive. It was apparent from this meeting that MPE as a burgeoning research space continues to attract experts from diverse fields ranging from epidemiology, pathology, oncology to genetics, biostatistics, bioinformatics, and data sciences, who have been working collaboratively toward a shared goal to deepen our understanding of cancer heterogeneities and the implications for cancer etiology, prognosis, and treatment. The momentum of the fast-advancing research agenda and the resilience of this vibrant researcher community are equally remarkable. Looking forward in the next few years, we anticipate expansion and fruition of MPE research in many fronts, particularly immune-epidemiology, mutational signatures, liquid biopsy, and health disparities. We plan to reconvene for the Sixth International Molecular Pathological Epidemiology (MPE) Meeting tentatively scheduled for May 2023, in Buffalo, NY, USA.

Funding The work presented at the meeting and summarized in the proceedings were supported in part by the following funding sources: Drs. Mustapha Abubakar, Jonas Almeida, Stephen Chanock, Curtis Harris and Maria Teresa Landi- Intramural Research Program of the Division of Cancer Epidemiology and Genetics (DCEG) of the National Cancer Institute (NCI); Dr. Christine Ambrosone—U.S. National Institutes of Health (NIH) grant R01 CA225947 and Breast Cancer Research Foundation; Dr. Samir Hanash—NCI, the MD Anderson Moonshot Program and the PanCan Foundation; Dr. Karl Kelsey—NIH R01 CA216265, R01 CA253976, 2018 AACR-Johnson & Johnson Lung Cancer Innovation Science (18-90-52-MICH), CDMRP/Department of Defense (W81XWH-20–1-0778), NIGMS (P20 GM1044168299), P30 CA168524, P20 GM130423, P20 GM103428, R01 CA207360, P50 CA097257, and T35 HL094308; Dr. Shuji Ogino—NIH grant R35 CA197735; Dr. John Quackenbush—NIH R35 CA220523 and U24 CA231846; Dr. Scott Rodig—Bristol-Myers Squibb, Merck, KITE/Gilead, and a T32 NIH training grant provided to the Department of Pathology of Brigham and Women's Hospital; Dr. Nikolaus Schultz—NCI P30 CA008748, U24 CA220457, R01 CA204749, the Fund for Innovation in Cancer Informatics (ICI), the Robertson Foundation, a Prostate Cancer Foundation Challenge Award, and the Marie-Josée and Henry R. Kravis Center for Molecular Oncology; Dr. Guillermo Tearney—NIH R01 CA265742; Dr. Shelley Tworoger—NIH P01 CA87969, U01 CA176726, UM1 CA186107, the James and Esther King Biomedical Research Program of the Florida Department of Health 9JK02, and the Biostatistics and Bioinformatics Shared Resource at Moffitt Cancer Center.

Data Availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest Dr. Scott Rodig receives research support from Bristol-Myers Squibb, Merck, KITE/Gilead, and Affimed. Dr. Rodig is on the SAB for Immunitas Therapeutics and has stock options in the company. Dr. Guillermo Tearney received sponsored research funding from Xsphera Biosciences. Dr. Karl Kelsey is a founder of Cellintec which had no role in this research. No other authors declare conflicts of interest related to the work presented here.

References

1. Ambrosone CB, Rebbeck TR, Morgan GJ et al (2006) New developments in the epidemiology of cancer prognosis: traditional and molecular predictors of treatment response and survival. Cancer epidemiol, biomarkers prev 15:2042–2046
2. Ogino S, Nowak JA, Hamada T, Milner DA Jr, Nishihara R (2019) Insights into pathogenic interactions among environment, host, and tumor at the crossroads of molecular pathology and epidemiology. Annu Rev Pathol 14:83–103
3. Hamada T, Keum N, Nishihara R, Ogino S (2017) Molecular pathological epidemiology: new developing frontiers of big data

science to study etiologies and pathogenesis. J Gastroenterol 52:265–275

4. Ogino S, Nishihara R, VanderWeele TJ et al (2016) Review article: the role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. Epidemiology 27:602–611
5. Ogino S, King EE, Beck AH, Sherman ME, Milner DA, Giovannucci E (2012) Interdisciplinary education to integrate pathology and epidemiology: towards molecular and population-level health science. Am J Epidemiol 176:659–667
6. Ogino S, Chan AT, Fuchs CS, Giovannucci E (2011) Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut 60:397–411
7. Dai J, Nishi A, Tran N et al (2021) Revisiting social MPE: an integration of molecular pathological epidemiology and social science in the new era of precision medicine. Expert Rev Mol Diagn 21:869–886
8. Ogino S, Campbell PT, Nishihara R et al (2015) Proceedings of the second international molecular pathological epidemiology (MPE) meeting. Cancer causes control 26:959–972
9. Campbell PT, Rebbeck TR, Nishihara R et al (2017) Proceedings of the third international molecular pathological epidemiology (MPE) meeting. Cancer causes control 28:167–176
10. Campbell PT, Ambrosone CB, Nishihara R et al (2019) Proceedings of the fourth international molecular pathological epidemiology (MPE) meeting. Cancer causes control 30:799–811
11. Fujiyoshi K, Bruford EA, Mroz P, et al. (2021) Opinion: Standardizing gene product nomenclature-a call to action. Proceedings of the National Academy of Sciences of the United States of America. 118.
12. Campbell PT, Jacobs ET, Ulrich CM et al (2010) Case-control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. J Natl Cancer Inst 102:391–400
13. Campbell PT, Newton CC, Newcomb PA et al (2015) Association between body mass index and mortality for colorectal cancer survivors: overall and by tumor molecular phenotype. Cancer epidemiol, biomarkers prev 24:1229–1238
14. Campbell PT, Lin Y, Bien SA et al (2021) Association of body mass index with colorectal cancer risk by genome-wide variants. J Natl Cancer Inst 113:38–47
15. Ogino S, Nowak JA, Hamada T et al (2018) Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. Gut 67:1168–1180
16. Hamada T, Nowak JA, Masugi Y et al (2019) Smoking and risk of colorectal cancer sub-classified by tumor-infiltrating T cells. J Natl Cancer Inst 111:42–51
17. Cao Y, Nishihara R, Qian ZR et al (2016) Regular aspirin use associates with lower risk of colorectal cancers with low numbers of tumor-infiltrating lymphocytes. Gastroenterology 151:879–92 e4
18. Song M, Zhang X, Meyerhardt JA et al (2017) Marine omega-3 polyunsaturated fatty acid intake and survival after colorectal cancer diagnosis. Gut 66:1790–1796
19. Liu L, Tabung FK, Zhang X et al (2018) Diets that promote colon inflammation associate with risk of colorectal carcinomas that contain fusobacterium nucleatum. Clin Gastroenterol Hepatol 16:1622–31 e3
20. Song M, Nishihara R, Cao Y et al (2016) Marine omega-3 polyunsaturated fatty acid intake and risk of colorectal cancer characterized by tumor-infiltrating T cells. JAMA Oncol 2:1197–1206
21. Borowsky J, Haruki K, Lau MC et al (2021) Association of Fusobacterium nucleatum with Specific T-cell Subsets in the Colorectal Carcinoma Microenvironment. Clinical Cancer Res 27:2816–2826
22. Vayrynen JP, Haruki K, Lau MC et al (2021) The prognostic role of macrophage polarization in the colorectal cancer microenvironment. Cancer Immunol Res 9:8–19
23. Akimoto N, Ugai T, Zhong R et al (2021) Rising incidence of early-onset colorectal cancer—a call to action. Nat Rev Clin Oncol 18:230–243
24. Arthur R, Wang Y, Ye K et al (2017) Association between lifestyle, menstrual/reproductive history, and histological factors and risk of breast cancer in women biopsied for benign breast disease. Breast Cancer Res Treat 165:623–631

25 Abubakar M, Fan S, Bowles EA et al (2021) Relation of quantitative histologic and radiologic breast tissue composition metrics with invasive breast cancer risk. JNCI Cancer Spectr 5:pkab015

26. Ambrosone CB, Higgins MJ (2020) Relationships between breast feeding and breast cancer subtypes: lessons learned from studies in humans and in mice. Can Res 80:4871–4877
27. Cheng TD, Yao S, Omilian AR et al (2020) FOXA1 Protein Expression in ER(+) and ER(-) Breast Cancer in Relation to Parity and Breastfeeding in Black and White Women. Cancer epidemiol, biomarkers prev 29:379–385
28. Sribenja S, Maguire O, Attwood K et al (2021) Deletion of Foxa1 in the mouse mammary gland results in abnormal accumulation of luminal progenitor cells: a link between reproductive factors and ER-/TNBC breast cancer? Am J Cancer Res 11:3263–3270
29. Zhang T, Joubert P, Ansari-Pour N et al (2021) Genomic and evolutionary classification of lung cancer in never smokers. Nat Genet 53:1348–1359

30 Morton LM, Karyadi DM, Stewart C et al (2021) Radiation-related genomic profile of papillary thyroid carcinoma after the Chernobyl accident. Science. https://doi.org/10.1126/science.abg2538

31. Leung HM, Wang ML, Osman H et al (2020) Imaging intracellular motion with dynamic micro-optical coherence tomography. Biomed Opt Express 11:2768–2778
32. Carey CD, Gusenleitner D, Lipschitz M et al (2017) Topological analysis reveals a PD-L1-associated microenvironmental niche for Reed-Sternberg cells in Hodgkin lymphoma. Blood 130:2420–2430
33. Patel SS, Weirather JL, Lipschitz M et al (2019) The microenvironmental niche in classic Hodgkin lymphoma is enriched for CTLA-4-positive T cells that are PD-1-negative. Blood 134:2059–2069
34. Griffin GK, Weirather JL, Roemer MGM et al (2021) Spatial signatures identify immune escape via PD-1 as a defining feature of T-cell/histiocyte-rich large B-cell lymphoma. Blood 137:1353–1364
35. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F (2021) Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 5:555–570
36. Lu MY, Chen TY, Williamson DFK et al (2021) AI-based pathology predicts origins for cancers of unknown primary. Nature 594:106–110
37. Patel DP, Pauly GT, Tada T et al (2020) Improved detection and precise relative quantification of the urinary cancer metabolite biomarkers—Creatine riboside, creatinine riboside, creatine and creatinine by UPLC-ESI-MS/MS: Application to the NCI-Maryland cohort population controls and lung cancer cases. J Pharm Biomed Anal 191:113596
38. Parsons HA, Rhoades J, Reed SC et al (2020) Sensitive detection of minimal residual disease in patients treated for early-stage breast cancer. Clinical Cancer Res 26:2556–2564
39. Xiong K, Shea D, Rhoades J et al (2022) Duplex-Repair enables highly accurate sequencing, despite DNA damage. Nucleic Acids Res 50:e1. https://doi.org/10.1093/nar/gkab855

40. Gydush G, Nguyen E, Bae JH et al (2021) MAESTRO affords 'breadth and depth' for mutation testing. bioRxiv. https://doi.org/10.1101/2021.01.22.427323
41. Bae JH, Liu R, Nguyen E et al (2021) CODEC enables 'single duplex' sequencing. bioRxiv. https://doi.org/10.1101/2021.06.11.448110
42. Faca VM, Song KS, Wang H et al (2008) A mouse to human search for plasma proteome changes associated with pancreatic tumor development. PLoS Med 5:e123
43. Fahrmann JF, Schmidt CM, Mao X et al (2021) Lead-time trajectory of CA19-9 as an anchor marker for pancreatic cancer early detection. Gastroenterology 160:1373–83 e6
44. Salas LA, Zhang Z, Koestler DC, et al. (2021) Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. Nature Communications. [In Press].
45. Koestler DC, Jones MJ, Usset J et al (2016) Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics 17:120
46. Chang MT, Asthana S, Gao SP et al (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 34:155–163
47 Chakravarty D, Gao J, Phillips SM et al (2017) OncoKB: a precision oncology knowledge base. JCO Precis Oncol. https://doi.org/10.1200/PO.17.00011
48. Cerami E, Gao J, Dogrusoz U et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401–404
49. Bhawsar PMS, Abubakar M, Schmidt MK et al (2021) Browser-based data annotation, active learning and real-time distribution of AI Models: from tumor tissue microarrays to COVID-19 radiology. J Pathol Inform 12:38
50. Le H, Gupta R, Hou L et al (2020) Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. Am J Pathol 190:1491–1504
51. Saltz J, Almeida J, Gao Y et al (2017) Towards generation, management, and exploration of combined radiomics and pathomics datasets for cancer research. AMIA Jt Summits Transl Sci Proc 2017:85–94
52. Almeida JS, Shiels M, Bhawsar P et al (2021) Mortality tracker: the COVID-19 case for real time web APIs as epidemiology commons. Bioinformatics 37:2073–2074
53. Glass K, Huttenhower C, Quackenbush J, Yuan GC (2013) Passing messages between biological networks to refine predicted interactions. PLoS ONE 8:e64832
54. Kuijjer ML, Hsieh PH, Quackenbush J, Glass K (2019) lionessR: single sample network inference in R. BMC Cancer 19:1003
55. Ben Guebila M, Lopes-Ramos CM, Weighill D et al (2022) GRAND: a database of gene regulatory network models across human conditions. Nucleic Acids Res 50:D610–D621. https://doi.org/10.1093/nar/gkab778

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Song Yao[1] · Peter T. Campbell[2] · Tomotaka Ugai[3] · Gretchen Gierach[4] · Mustapha Abubakar[4] · Viktor Adalsteinsson[5] · Jonas Almeida[4] · Paul Brennan[6] · Stephen Chanock[4] · Todd Golub[5,7] · Samir Hanash[8] · Curtis Harris[9] · Cassandra A. Hathaway[10] · Karl Kelsey[11] · Maria Teresa Landi[4] · Faisal Mahmood[12] · Christina Newton[13] · John Quackenbush[14,15] · Scott Rodig[12] · Nikolaus Schultz[16] · Guillermo Tearney[17] · Shelley S. Tworoger[10] · Molin Wang[3,14,15] · Xuehong Zhang[14,18] · Montserrat Garcia-Closas[4] · Timothy R. Rebbeck[19] · Christine B. Ambrosone[1] · Shuji Ogino[3,5,12]

✉ Song Yao
song.yao@roswellpark.org

✉ Shuji Ogino
sogino@bwh.harvard.edu

[1] Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Elm & Carlton Streets, Buffalo, NY 14263, USA

[2] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

[3] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[4] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

[5] Broad Institute of MIT and Harvard, Boston, MA, USA

[6] International Agency for Research On Cancer (IARC/WHO), Genomic Epidemiology Branch, Lyon, France

[7] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[8] Department of Clinical Cancer Prevention, MD Anderson Cancer Institute, Houston, TX, USA

[9] Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

[10] Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

[11] Department of Epidemiology, Brown School of Public Health, Brown University, Providence, RI, USA

[12] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[13] Department of Population Science, American Cancer Society, Atlanta, GA, USA

[14] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[15] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[16] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[17] Department of Pathology and Wellman Center for Photomedicine, Massachusetts General Hospital, Boston, MA, USA

[18] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[19] Zhu Family Center for Global Cancer Prevention, Harvard T.H. Chan School of Public Health and Dana-Farber Cancer Institute, Boston, MA, USA