RESEARCH ARTICLE

The impact of DNA methylation on the cancer proteome

Majed Mohamed Magzoub[1,2], **Marcos Prunello**[3], **Kevin Brennan**[4], **Olivier Gevaert**[4,5]*

1 Biomedical Informatics Training Program, Department of Biomedical Data Science, Stanford University, Palo Alto, California, United States of America, **2** Department of Bioengineering, Stanford University, Palo Alto, California, United States of America, **3** Department of Statistics, College of Pharmaceutical and Biochemical Sciences, National University of Rosario, Rosario, Argentina, **4** Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Palo Alto, California, United States of America, **5** Department of Biomedical Data Science, Stanford University, Palo Alto, California, United States of America

* ogevaert@stanford.edu

OPEN ACCESS

Citation: Magzoub MM, Prunello M, Brennan K, Gevaert O (2019) The impact of DNA methylation on the cancer proteome. PLoS Comput Biol 15(7): e1007245. https://doi.org/10.1371/journal.pcbi.1007245

Editor: Christina S. Leslie, Memorial Sloan-Kettering Cancer Center, UNITED STATES

Received: August 15, 2018

Accepted: July 2, 2019

Published: July 29, 2019

Data Availability Statement: The data in this study are third-party data and are available from the portals of The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Specifically, we used data for three cancer sites: colorectal cancer, breast cancer, and ovarian cancer. For each cancer, we used RNA sequencing and DNA methylation data from TCGA available through the Genomic Data Commons (GDC) data portal: https://portal.gdc.cancer.gov/ We used proteomic data from the CPTAC portal here: https://cptac-data-portal.georgetown.edu/cptacPublic/. More information about the CTPAC

Abstract

Aberrant DNA methylation disrupts normal gene expression in cancer and broadly contributes to oncogenesis. We previously developed MethylMix, a model-based algorithmic approach to identify epigenetically regulated driver genes. MethylMix identifies genes where methylation likely executes a functional role by using transcriptomic data to select only methylation events that can be linked to changes in gene expression. However, given that proteins more closely link genotype to phenotype recent high-throughput proteomic data provides an opportunity to more accurately identify functionally relevant abnormal methylation events. Here we present a MethylMix analysis that refines nominations for epigenetic driver genes by leveraging quantitative high-throughput proteomic data to select only genes where DNA methylation is predictive of protein abundance. Applying our algorithm across three cancer cohorts we find that using protein abundance data narrows candidate nominations, where the effect of DNA methylation is often buffered at the protein level. Next, we find that MethylMix genes predictive of protein abundance are enriched for biological processes involved in cancer including functions involved in epithelial and mesenchymal transition. Moreover, our results are also enriched for tumor markers which are predictive of clinical features like tumor stage and we find clustering using MethylMix genes predictive of protein abundance captures cancer subtypes.

Author summary

To elucidate the molecular basis of cancer we examine the variation and dynamics characterizing the flow of information from epigenome to the transcriptome and proteome. Conducting the first genome wide analysis of epigenome-proteome associations, we present a MethylMix analysis that leverages protein abundance data taking advantage of recent high-throughput proteomic data generated using mass-spectrometry technology to elucidate the role of DNA methylation in cancer. By integrating across molecular data types,

projects for each cancer and their associated data from TCGA are in the original CPTAC and TCGA cancer site publications: breast cancer: https://www.nature.com/articles/nature18003#extended-data (CPTAC) and https://www.nature.com/articles/nature11412 (TCGA) ovarian cancer: https://linkinghub.elsevier.com/retrieve/pii/S0092-8674(16)30673-0 (CPTAC) and https://www.nature.com/articles/nature10166 (TCGA) colorectal cancer: https://www.nature.com/articles/nature13438 (CPTAC) and https://www.nature.com/articles/nature11252 (TCGA). All other data is available within the paper and Supporting Information files. The authors did not have any access privileges.

Funding: Research reported in this publication was supported by the National Institute of Dental & Craniofacial Research (NIDCR) U01 DE025188 to OG, the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIBIB), R01 EB020527 to OG, the National Cancer Institute (NCI), U01 CA217851 to OG, and the National Library of Medicine Extramural Programs Grants, under award number T15-LM-007033 (www.nih.gov) to MMM. Research also supported by the Fund for Innovation in Cancer Informatics, www.the-ici-fund.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

we confirm the benefit of using protein abundance data to provide additional insights into pathways and processes involved in oncogenesis and how they manifest as clinical phenotypes. Applying our method across three large cancer cohorts including breast cancer, ovarian cancer and colorectal cancer, MethylMix identifies key genes and describes molecular features and subtypes in these cancers.

Introduction

Genomic characterization can elucidate underlying biology, disease etiology and reveal biomarkers of cancer development and progression; however, each molecular feature is susceptible to different sources of biological and technical measurement noise and provides only one view on the cell state. Therefore, comprehensive studies are needed to understand the molecular basis of disease. Toward this end a multi-institutional consortium, The Cancer Genome Atlas (TCGA), has extensively characterized numerous cancer sites producing genome wide data for mutations, copy number alterations (CNA), RNA expression, microRNA expression, and DNA methylation [1–5]. As part of this project, the proteome was probed using protein array Reverse Phase Protein Assay (RPPA) technology. However, antibody based analysis are inherently limited because of the reduced coverage and inability to easily compare across proteins due to differential binding effects [6,7]. Transcending these limitations, recent advancements in proteomics through high sensitivity mass-spectrometry (MS) are opening new opportunities in cancer research [8]. To accelerate the uptake of proteomics the Clinical Proteomic Tumor Analysis Consortium (CPTAC) is performing proteomic analyses of TCGA tumor bio-specimens for a growing number of tissue types and establishing standardized workflows using high-throughput liquid chromatography tandem mass-spectrometry (LC-MS/MS) to capture the proteome as a whole [6,9,10].

To best leverage this new technology comparative analysis between protein abundance and RNA expression can highlight factors influencing concordance and inform how to best interpret proteomic data [11]. For example, multiple studies have proven that concordance between mRNA and protein is highly variable, such that one cannot be used to reliably predict the other. Correlation between mRNA and protein has been repeatedly shown to vary by tissue type and cancer status among other molecular features like biological function or molecular stability [7]. It was shown across multiple cancers that dynamic proteins involved in metabolism show strong agreement whereas housekeeping proteins and RNA processing proteins are weakly or negatively correlated [6,9,10]. So, although many biological functions are regulated primarily through RNA expression—producing moderate correlation between proteomic and transcriptomic data, with mean spearman rho: 0.23–0.47 –post-transcriptional mechanisms also play a significant role that cannot be overlooked.

The proteome represents the final link from genotype to molecular phenotype, so proteins are of special importance among molecular features and likely provide a more accurate depiction of cell state; this enhanced view on disease can be leveraged to assess functional effects of upstream aberrations, such as epigenetic modifications. Multi-level epigenetic features such as DNA methylation and histone modification work in concert to regulate gene expression. DNA-methylation, the covalent addition of methyl groups to CpG dinucleotides to form 5-methylcytosine (5mC), is catalyzed by DNA methyltransferases, and is influenced by both environmental and hereditary factors [12]. Previous studies have shown that DNA methylation plays a key role in health and is involved in processes of embryonic development and cellular differentiation, where changes can occur through imprinting, inheritance, or de novo events [13,14]. Furthermore, DNA methylation has been numerously cited as a potentially causative event in cancer [15,16]. Among potential DNA methylation drivers, silencing of tumor suppressors through promoter CpG island hypermethylation is best understood and linked to corresponding gene silencing [13,17,18]. Global hypo-methylation on the other hand can potentially result in genomic instability and reactivation of oncogenes [12,13,15].

To elucidate the role of DNA methylation in disease, our goal is to investigate whether linking proteomic data with DNA methylation data identifies key genes, describes molecular features and subtypes in cancer. Previously we presented MethylMix an algorithm that formalizes the identification of DNA methylation driver genes using a model-based approach [19–23]. Recognizing the complex role of the methylome in epigenetic regulation of cancer, MethylMix uses mRNA data to select only differentially methylated genes that show a downstream effect on gene expression (MethylMix-GE). This selects for likely functional aberrations with the aim of discriminating between true driver genes, and passenger events which are characteristic of genome wide dysfunction in cancer. Herein we present MethylMix-PA (Protein Abundance), a MethylMix analysis which refines candidate nominations for epigenetic driver genes by excluding aberrations that are buffered at the protein level; this likely selects for events which are functional over those which may accumulate during cancer but do not drive pathogenesis. Using proteomic data generated by MS technology from three cancer cohorts: breast invasive carcinoma, colorectal adenocarcinoma, and ovarian serous cystadenocarcinoma, we report MethylMix-PA genes, which include potential cancer progression markers and therapeutic targets. We describe MethylMix-PA's ability to elucidate key molecular and higher level disease features and evaluate MethylMix-PA performance against MethylMix-GE. In summary, our study highlights the differences between integrated epigenomic-proteomics and epigenomic-transcriptomics analyses.

Results

We applied MethylMix to identify differentially methylated genes in two ways: once with gene expression data defined as MethylMix-GE and once with protein abundance data defined as MethylMix-PA [19–22] across three cancer types (Fig 1, Table 1): breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD), and ovarian serous cystadenocarcinoma (OV). These analyses result in two lists of genes: MethylMix-GE and MethylMix-PA representing genes that are both differentially methylated compared to normal methylation, and with an significant relationship with gene expression and protein abundance respectively (S1 Table). Next, we will specifically examine the biological and clinical relevance of both analyses' output and utility for downstream analysis.

MethylMix-PA narrows candidate nominations for epigenetically driven genes

For each cohort both models identify genes that are 1) differentially methylated when compared to normal adjacent tissue and 2) functionally predictive of downstream effects at the level of RNA expression in the case of MethylMix-GE or protein abundance in the case of MethylMix-PA (Fig 2). Among all three cancer cohorts we observe significant correlations between RNA expression and protein abundance (mean rho: 0.23–0.47), indicating that most genes are regulated at the transcript level (Fig 3, S2 Table). Therefore, it is unsurprising that MethylMix-PA shows high agreement with MethylMix, where more than 90% of MethylMix-PA genes are also identified by MethylMix. However, MethylMix-PA lists are more conservative identifying fewer candidate genes across all three cancers, where often the effect of

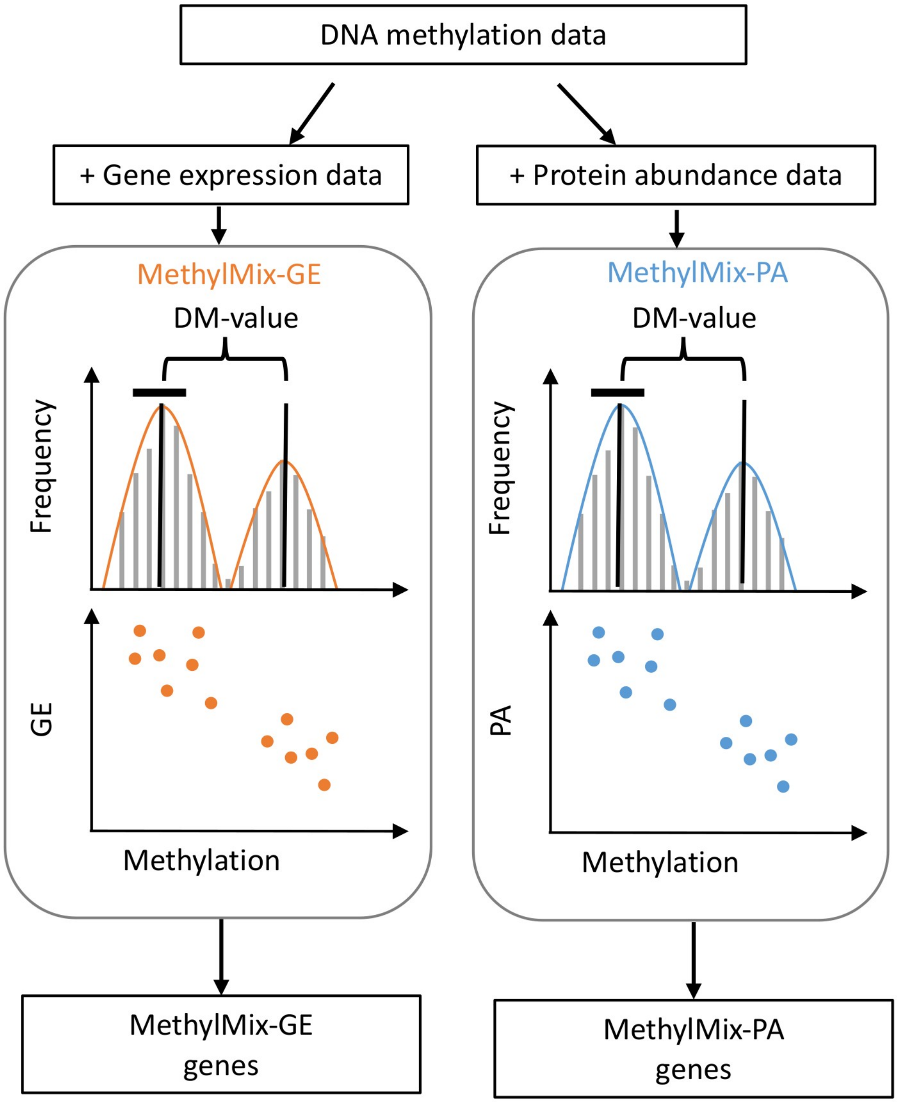

Fig 1. Workflow for MethylMix-GE and MethylMix-PA to identify genes with differential methylation between cancer and normal tissues with a significant relationship with gene expression and protein abundance respectively.

https://doi.org/10.1371/journal.pcbi.1007245.g001

methylation is present at the RNA level, but not detected at the protein level (Fig 2), likely because they are buffered at the protein due to post-transcriptional, translational, or degradation regulation. Therefore, MethylMix-PA better enriches for methylation-states that more likely execute functional roles in cancer development.

Table 1. Overview of number of genes, CpG clusters, and samples used for each TCGA cancer site analysis.

	N Genes	N CpG Clusters	N Samples: Gene Expression & Protein Abundance	N Samples: Tumor Tissue Methylation	N Samples: Normal Tissue Methylation
BRCA	6248	9221	76	972	123
COADREAD	2855	4299	85	614	78
OV	3876	5402	149	582	8

https://doi.org/10.1371/journal.pcbi.1007245.t001

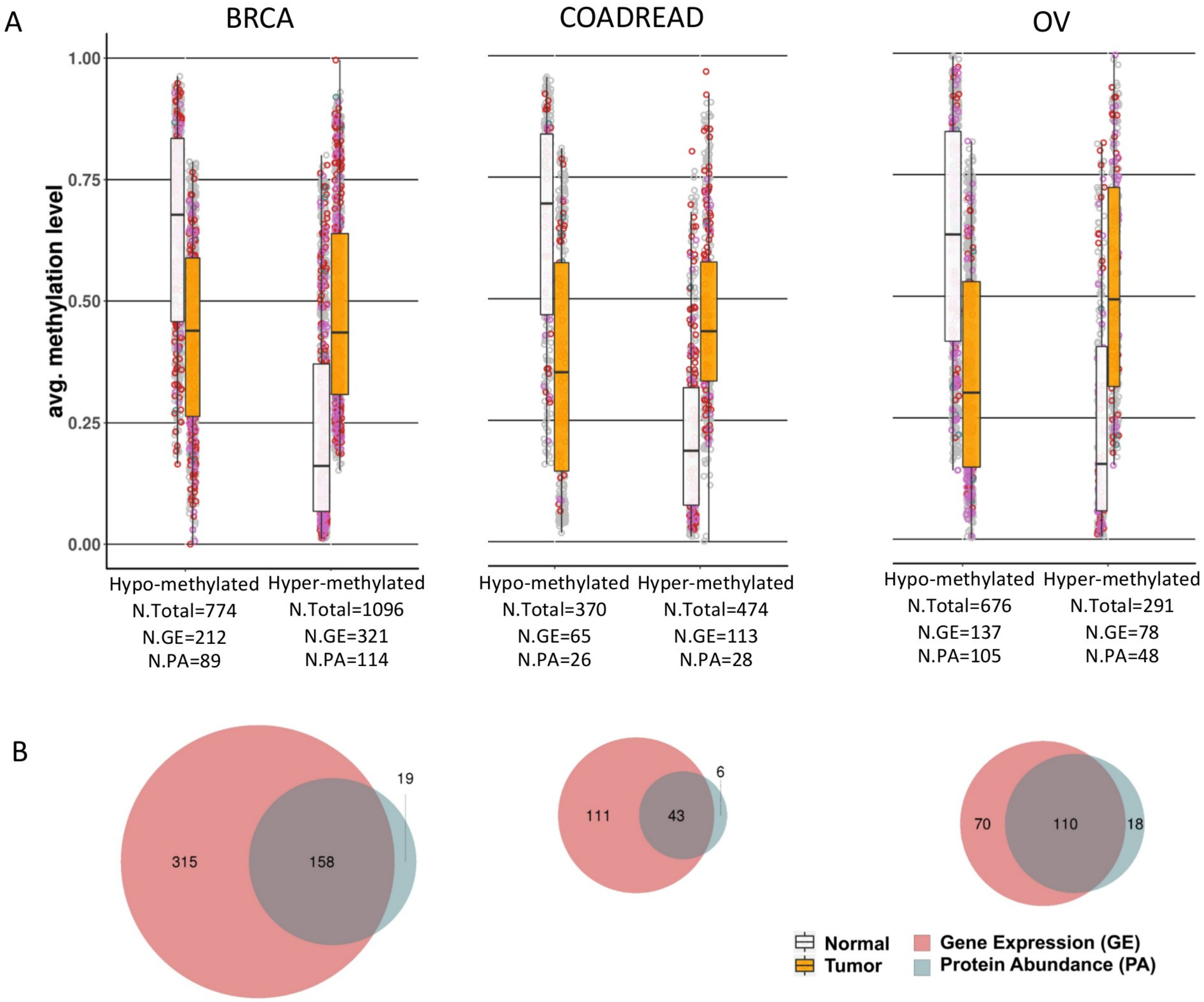

Fig 2. Visualization of MethylMix-PA and MethylMix-GE genes. A) Methylation of hyper and hypo-methylated genes for each of the three cancer sites: breast cancer (BRCA), colorectal cancer (COADREAD) and ovarian cancer (OV). Red circles: uniquely MethylMix-GE genes, blue circles: uniquely MethylMix-PA genes, and purple circles: overlapping genes between MethylMix-GE and MethylMix-PA B) Venn diagrams comparing the number of reported genes that are differentially methylated and functionally predictive for MethylMix-GE and MethylMix-PA.

https://doi.org/10.1371/journal.pcbi.1007245.g002

MethylMix-PA identifies new genes with significant methylation effects only at the protein level

For each cancer cohort using protein abundance data also identifies a few unique driver genes, the majority of which have documented roles in carcinogenesis. Explanative mechanisms by which the effect of DNA methylation may be undetected at the RNA level but functional at the protein level are further addressed below in the discussion.

In breast cancer we discovered 19 novel differentially methylated genes of diverse biological functions. MethylMix-PA identified hyper-methylation of *FSTL1*, an autoantigen that promotes immune response. This candidate tumor suppressor, *FSTL1*, has also been shown to mediate tumor immune evasion in nasopharyngeal cancer through hyper-methylation silencing [24]. MethylMix-PA also found hyper-methylation of *DHX40* which has an unclear link to cancer; although it is of note that RNA splicing proteins—like *DHX40* –are highly stable, perhaps explaining the particularly stronger effect of DNA methylation on protein abundance

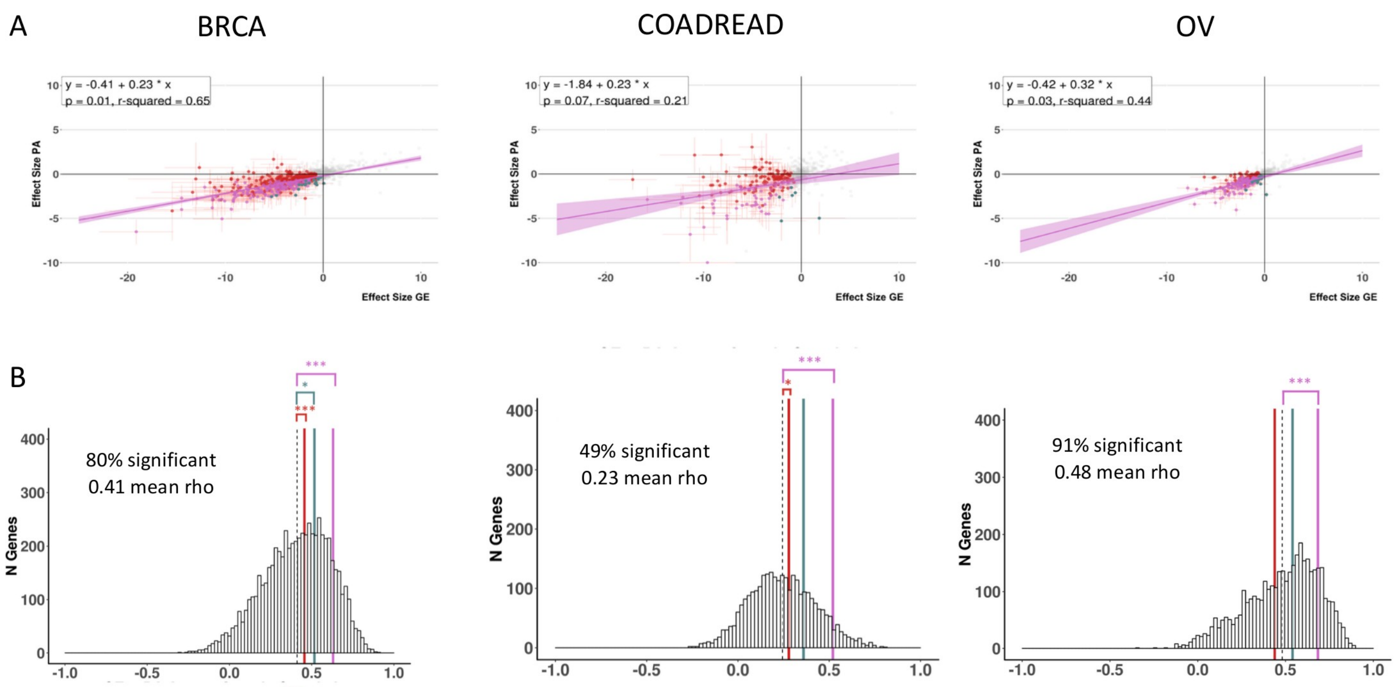

Fig 3. Correlation analysis between gene expression and protein abundance for the three cancer sites: Breast cancer (BRCA), colorectal cancer (COADREAD) and ovarian cancer (OV). A) Regression analysis between gene expression and protein abundance. Regression line in purple with confidence interval. Red circles: uniquely MethylMix-GE genes, blue circles: uniquely MethylMix-PA genes, and purple circles: overlapping genes between MethylMix-GE and MethylMix-PA. B) histogram of correlation between gene expression and protein abundance across samples. Red line: MethylMix-GE genes, blue line: MethylMix-PA, and purple line: overlapping genes between MethylMix-GE and MethylMix-PA, showing that MethylMix-PA genes have higher correlation.

https://doi.org/10.1371/journal.pcbi.1007245.g003

than mRNA [25] (S1 Table). Next, MethylMix-PA identified hypo-methylation of CEACAM5 (also known as CEA), a cell surface glycoprotein that is used as a clinical biomarker for gastrointestinal cancers and may promote tumor development through its role as a cell adhesion molecule. High levels of CEACAM5 have been associated with operable early breast cancer [26,27]. Next, MethylMix-PA also identified hyper-methylation of FOXO1, a transcritionf factor where low expression has been associated with cancer [28].

In colorectal cancer the MethylMix-PA analysis uniquely recovers several genes associated with immune function and inflammation, which is known to play a key role in pathogenesis. We found that MethylMix-PA identifies a functional effect of UTR hypo-methylation of the *PTPRC* gene. *PTPRC* belongs to a family of protein tyrosine phosphatase which contains oncogenes regulating cell growth and differentiation. *PTPRC* is also related to tumor necrosis and disrupts normal T- and B-cell signaling through SRC kinase pathways—which are separately implicated in colorectal cancer through amplification [9,29]. Next, MethylMix-PA identified upregulation of *S100A9* through promoter hypo-methylation. Of note, elevated *S100A9* mRNA and protein levels are commonly observed in many conditions associated with inflammation [30]; additionally in hydropharangeal cancer where knockdown inhibited cell growth and invasion, *S100A9* is also prognostic of worse outcome and indications like metastasis [31]. Of note MethylMix-PA filtered out functional effects of a UTR hypo-methylation in *S100A9* previously detected by MethylMix-GE. Next, MethylMix-PA identified hyper-methylation across the promoter region of *LTF*, a likely tumor suppressor which is produced by neutrophils to regulate growth and differentiation. In the context of colorectal tissue *LTF* has been shown to restrict inflammation by regulating T cell interaction [32]. Additionally, gene expression of *LTF* has previously been shown to correlate with tumor size and survival in breast cancer [33].

MethylMix-PA picks up hypo-methylation states in 18 unique genes in ovarian cancer related to processes of invasion and proliferation. MethylMix-PA uniquely identifies hypo-methylation in the promoter region of *EVL* a key regulator of the actin cytoskeleton, associated with invasion and metastasis. Overexpression of *EVL* is also indicative of advanced stage in breast cancer [34] and has been implicated in malignancies due to inappropriate recombination [35]. Next, MethylMix-PA identified hypo-methylation of CTSZ, also known as cathepsin Z, a a lysosomal cysteine proteinase that has been shown to be involved in many primary tumors. For example, high levels of CTSZ promote epithelial to mesenchymal transition and are associated with the mesenchymal-like cell phenotype [36]. We also found hypermethylation of GSTM2, a gene that is normally high expressed in ovary, but has been shown to be a hypermethylated in lung cancers [37] and colorectal cancers [38], suggesting a tumor suppressor role for GSTM2 across tissues. Lastly, MethylMix-PA also identifies hypo-methylation in the mitochondrial genes *SPG7*, speculatively linked to cancer through metabolic function [39].

MethylMix-PA genes are enriched for biological processes involved in cancer

We conducted enrichment analysis to identify biological processes that are overrepresented in MethylMix-PA and MethylMix-GE genes (Table 2, S3 Table). Given the large proportion of common genes, across all three cancers both models capture many of the same annotations. However, comparing enrichments found for each cancer site, we find that broadly MethylMix-PA results include more significant enrichments for functions associated with cell adhesion and migration of epithelial and endothelial cells; these processes increase cell motility and invasiveness and are indicative of epithelial to mesenchymal transition (EMT) which is key to cancer development. Additionally, we observed that enrichment for immune functions are highly variable between each model's results.

Comparing unique annotations among breast cancer genes, MethylMix-PA includes enrichments for cell-cell adhesion, STAT signaling, response to interferon-gamma and immune cell functions, whereas MethylMix-GE similar pathways, but is also enriched for several other functions with less relevance to cancer such as homeostasis, muscle cell proliferation and skin development. In colorectal cancer, the MethylMix-PA gene list is shorter as fewer MethylMix-PA genes have been identified (Fig 2). These genes are only enriched in cell-cell adhesion (Table 2). The MethylMix-GE list for colorectal cancer is also enriched in cell-cell adhesion but also includes seemingly irrelevant enrichments for humoral immune response and detection of stimulus involved in sensory perception (S3 Table). For ovarian cancer, the MethylMix-PA enrichment mirrors the MethylMix-GE enrichment almost exactly with enrichments for metabolic processes, NF-kappa-beta signaling and interleukin-1 production.

MethylMix-PA genes are enriched for tumor progression markers

Taking an orthogonal approach, we identified putative biomarker of disease progression based on correlations between gene expression and clinical features (Table 3). Although MethylMix-PA gene lists contain much fewer identifications, we find that across all three cancers MethylMix-PA lists include a larger proportion of markers of tumor stage and size and show stronger odds of containing such genes (Table 3). The greatest difference in frequency of tumor stage marker is observed in breast cancer where 12% versus 7% of genes show correlation in MethylMix-PA and MethylMix-GE gene lists respectively. The most significant associations however are observed in colorectal cancer where 15% of MethylMix-PA genes show correlation

Table 2. Gene set enrichment analysis results for each cancer site including MethylMix-GE and MethylMix-PA genes, showing only results where the MethylMix-PA adjusted P-value<0.10. Complete results are in S3 Table. Genes in bold are specific to the MethylMix-PA analysis.

Gene Ontology term	Genes	MethylMix gene list	Nr of genes overlap	Adjusted P Value
	BRCA			
regulation of leukocyte activation	BCL2, CARD11, CBFB, **CCL5**, CD40, CGAS, DOCK8, FES, **GRAP2**, HLA-DPB1, IDO1, IKZF3, **IL18**, **NCKAP1L**, PRNP, PTPN6, **PTPRC**, PYCARD, **RIPK3**, **RUNX3**, SFRP1, STAT6, STXBP2, TBC1D10C, **TNFAIP8L2**, **VTCN1**, ZBTB7B	pa	9	0.10
		ge	27	0.08
fatty acid metabolic process	ACAA2, **ACADS**, ACOT2, ALOX15B, AOAH, **CBR1**, **CRAT**, **CROT**, ECI2, **FADS2**, GPX1, **GSTM2**, **GSTP1**, HSD17B12, **HSD17B4**, **HSD17B8**, **IVD**, LONP2, **LPIN2**, MGST1, PAM, PDK4, PON1, **PON3**, **PTGR1**, **RGN**, THNSL2	pa	14	0.09
		ge	27	0.01
leukocyte cell-cell adhesion	CARD11, CBFB, **CCL5**, CD44, DOCK8, ETS1, **GRAP2**, HLA-DPB1, IDO1, **IL18**, **ITGAL**, **NCKAP1L**, NFAT5, PRNP, PTPN6, **PTPRC**, PYCARD, **RUNX3**, SKAP1, **TNFAIP8L2**, **VTCN1**, ZBTB7B	pa	9	0.03
		ge	22	0.05
interleukin-2 production	CARD11, **CARD9**, **IL18**, PRNP, **PTPRC**, **VTCN1**	pa	4	0.10
		ge	6	0.37
response to interferon-gamma	**BST2**, **CCL5**, CD40, CD44, **DAPK1**, DAPK3, **EVL**, **GBP4**, HLA-DPB1, HPX, IRF6, IRF7, OAS1, STXBP2, **TRIM22**, TRIM38, VIM	pa	6	0.09
		ge	16	0.04
T cell activation	BCL2, CARD11, CASP8, CBFB, **CCL5**, CGAS, **CTPS1**, **CXADR**, DOCK2, DOCK8, **GRAP2**, HLA-DPB1, IDO1, **IL18**, **ITGAL**, LCP1, **NCKAP1L**, **PIK3CD**, PIK3CG, **PREX1**, PRNP, PTPN6, **PTPRC**, **PYCARD**, **RIPK3**, **RUNX3**, SP3, STAT6, **TNFAIP8L2**, **VTCN1**, ZBTB7B	pa	15	0.10
		ge	31	0.01
small molecule catabolic process	ACAA2, **ACADS**, ADHFE1, ALDH2, **ALDOC**, APOBEC3C, **ASRGL1**, CD44, CRABP1, **CRAT**, CROT, CYP24A1, ECI2, **FGF2**, **GALM**, GALT, GLDC, **GPT**, HAAO, **HSD17B4**, HSD3B7, IDO1, INPP5B, **IVD**, LDHD, LONP2, LPIN2, **MAT1A**, **MGAT1**, MTAP, PFKP, PON1, **PON3**, **PRTFDC1**, THNSL2	pa	13	0.08
		ge	35	0.01
positive regulation of cell activation	CBFB, **CCL5**, DOCK8, **GRAP2**, HLA-DPB1, **IL18**, **NCKAP1L**, **PLEK**, PTPN6, **PTPRC**, **RUNX3**, **VTCN1**, ZBTB7B	pa	8	0.06
		ge	13	0.03
leukocyte migration	APOD, **CCL5**, **CD99L2**, **CEACAM5**, **CXADR**, DOCK8, **DOK2**, **ECM1**, F7, **ITGAL**, **NCKAP1L**, **PIK3CD**, PIK3CG, **PLCB1**, **PREX1**, PTPN6, **PYCARD**, **RIPK3**, **S100A14**, **TGFB2**	pa	15	0.09
		ge	14	0.07
regulation of peptidase activity	**A2ML1**, AIFM1, **BST2**, CASP8, **DAPK1**, DHCR24, **ECM1**, **PSMB8**, SERPINA3, SERPINA5, SERPINA6, **SFN**, **SLPI**, **TFAP2B**, **TNFAIP8**	pa	9	0.09
		ge	15	0.07
STAT cascade	AKR1B1, **CCL5**, CD40, **HCLS1**, HPX, **IL18**, **PTK6**, **PTPRC**, STAT5A	pa	5	0.06
		ge	9	0.16
cell-cell adhesion via plasma-membrane adhesion molecules	APOA1, CDH13, CDH5, **CEACAM5**, **CLSTN2**, **CXADR**, **ITGAL**, **NECTIN4**, PTPRM, TGFB2	pa	5	0.09
		ge	9	0.26
	COADREAD			
cell-cell adhesion via plasma-membrane adhesion molecules	APOA1, **CDH17**, CEACAM1, **CEACAM5**, **CEACAM6**, **ITGAM**	pa	4	0.03
		ge	6	0.01
	OV			
cellular modified amino acid metabolic process	**CRAT**, **FOLR1**, **GOT2**, **GSTM1**, **GSTM2**, **GSTP1**, **GSTT1**, **PAX8**, **TMLHE**	pa	9	0.05
		ge	7	0.43
I-kappaB kinase/NF-kappaB signaling	**BST2**, **CASP1**, **CASP10**, **CASP8**, **GSTP1**, HMOX1, LTF, **MYD88**, **PYCARD**, **S100A13**, **SLC44A2**, **TRIM22**	pa	10	0.05
		ge	12	0.05
interleukin-1 production	APOA1, **CASP1**, **GSTP1**, **NLRP2**, **PYCARD**, **S100A13**	pa	5	0.05
		ge	6	0.01
response to type I interferon	**BST2**, **IFI35**, **MX2**, **MYD88**, **PSMB8**, **SP100**	pa	6	0.09
		ge	6	0.14
benzene-containing compound metabolic process	**GOT2**, **GSTM1**, **GSTM2**, **IDO1**	pa	4	0.03
		ge	2	0.47

https://doi.org/10.1371/journal.pcbi.1007245.t002

Table 3. Report of overlap between MethylMix-GE and MethylMix-PA genes with tumor progression markers produced using a fisher exact test.

	Cancer	model	met drivers	overlap	percentage	p.value
A. Tumor Stage	BRCA	MethylMix-GE	148	10	7.5%	9.83E-04
		MethylMix-PA	46	5	12.2%	2.38E-03
	COADREAD	MethylMix-GE	125	15	12.9%	1.74E-06
		MethylMix-PA	28	4	14.8%	3.86E-04
	OV	MethylMix-GE	70	2	3.3%	2.42E-02
		MethylMix-PA	52	2	4.4%	2.43E-02
B. Tumor Size	BRCA	MethylMix-GE	148	28	21.1%	2.13E-11
		MethylMix-PA	46	12	29.3%	1.55E-08
	COADREAD	MethylMix-GE	125	3	2.6%	1.69E-01
		MethylMix-PA	28	2	7.4%	4.25E-02
	OV	MethylMix-GE	70	4	6.6%	3.51E-01
		MethylMix-PA	52	4	8.9%	1.76E-01

https://doi.org/10.1371/journal.pcbi.1007245.t003

between gene expression and tumor stage, this includes LTF which is mentioned among unique MethylMix-PA genes (Table 3A, S1 Table). The same trend applies when correlating gene expression with tumor size where the largest difference in enrichment can be seen in colorectal cancer where 7% versus 3% of genes correlate with size when comparing models. However, the enrichment is much stronger for breast cancer where 29% of genes correlate with tumor size compared to 21% of MethylMix-GE genes (Table 3B).

Clustering on MethylMix-PA genes captures cancer subtypes

Clustering on methylation has been shown to stratify patients into clinically relevant subgroups [2,20,21,23]. We performed consensus clustering using the DM values for MethylMix-PA and MethylMix-GE genes evaluating clusters sizes from two to six (Table 4); for clarity we discuss clusters at K = 2, examining the gross differences between MethylMix-GE and MethylMix-PA. We evaluated if these epigenetically defined subgroups correspond to previously

Table 4. Summary statistics from consensus clustering analysis across K = 2–6 for each cancer; we report inter- and intra-cluster scores along with PAC score.

		MethylMix—GE			MethylMix—PA		
Cancer	K	Intra	Inter	PAC	Intra	Inter	PAC
BRCA	2	99.6	28.6	0.013	99.8	27.6	0.010
	3	94.5	21.8	0.326	90.0	24.2	0.355
	4	92.3	15.4	0.190	88.9	16.0	0.222
	5	84.1	13.5	0.233	81.5	14.2	0.230
	6	81.7	11.8	0.235	79.5	11.9	0.252
COADREAD	2	97.0	27.4	0.091	97.3	26.6	0.081
	3	71.0	26.1	0.554	85.3	21.4	0.410
	4	77.2	18.1	0.387	68.5	19.0	0.405
	5	73.3	14.9	0.358	70.6	15.6	0.366
	6	72.1	12.3	0.291	72.0	12.5	0.333
OV	2	94.1	28.7	0.150	93.3	29.3	0.205
	3	86.4	21.4	0.291	82.3	22.1	0.445
	4	61.1	20.1	0.497	68.3	18.6	0.470
	5	60.2	16.5	0.454	57.5	16.1	0.461
	6	62.0	13.5	0.395	54.1	13.7	0.402

https://doi.org/10.1371/journal.pcbi.1007245.t004

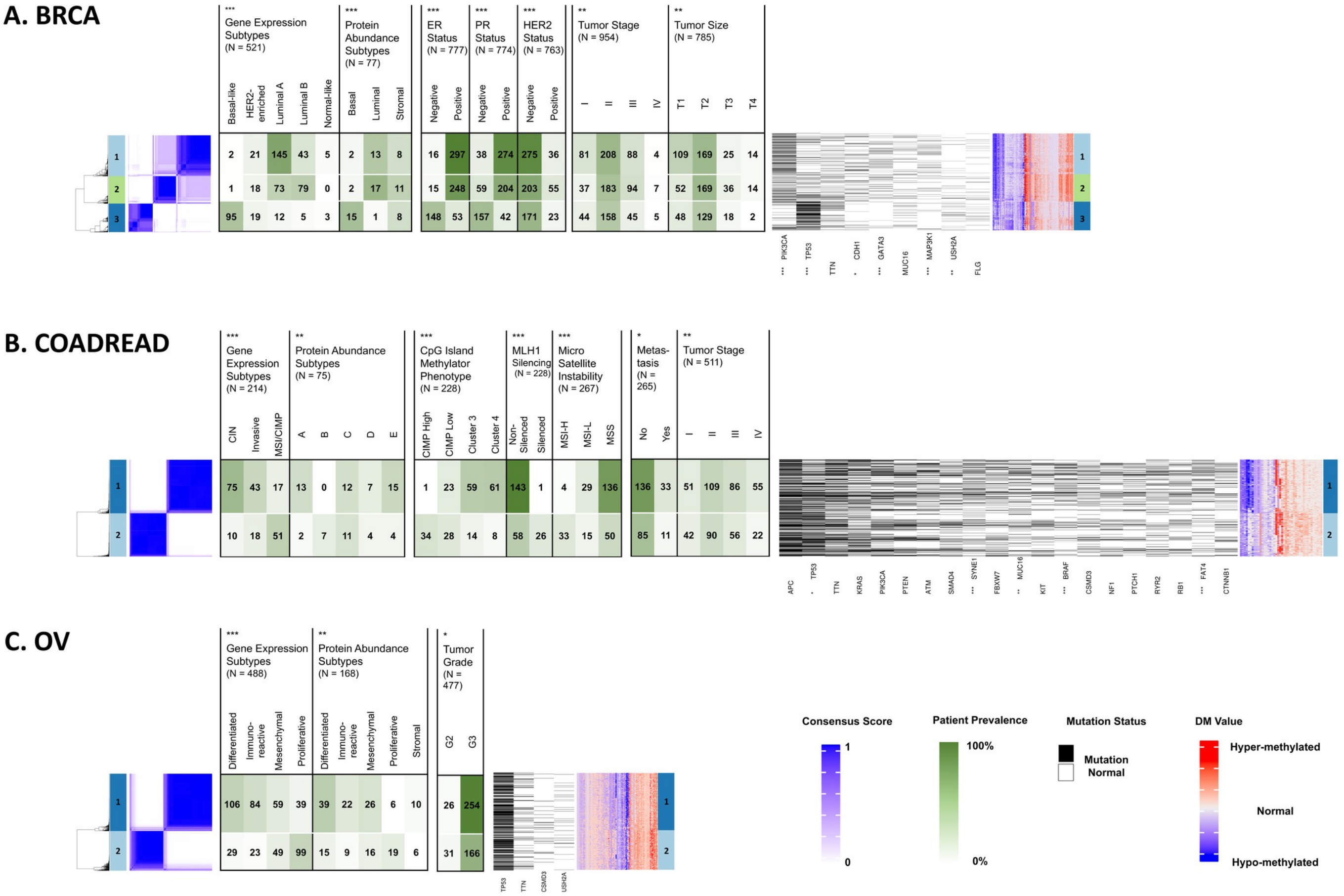

Fig 4. Consensus clustering and methylation profiles for three cancer sites at K = 2. (A) breast cancer (BRCA); (B) colorectal cancer (COADREAD); (C) ovarian cancer (OV). For each cancer site: Left panel: visualization of the consensus clustering with blue indicating high consensus and white indicating low consensus. Middle panels: association with published gene expression and protein abundance subtypes, and additional molecular and clinical features for each cancer, followed by association with somatic mutation status. Statistically significant overlaps, found using Chi-squared and Kruskal-Wallis tests, are marked with asterisks.

https://doi.org/10.1371/journal.pcbi.1007245.g004

published subtypes and clinical and genetic features and found that MethylMix-PA identifies subgroups of patients that enriched for specific cancer subtypes and other molecular features and performs similarly to MethylMix-GE (Fig 4, S4 Table).

In breast cancer MethylMix-PA clusters significantly correlate with molecular subtypes and other molecular features such as Progesterone and Estrogen Receptor (PR, ER) status (Fig 4A). Similar to other studies our clusters differentiate between canonical breast cancer molecular subtypes: Cluster-1 and Cluster-2 containt the majority of patients with Luminal A/B type tumors while Cluster-3 contains the majority of patients with Basal-like tumors and as expected is enriched for samples negative for ER, PR, or HER2. HER2 and Normal subtypes are less clearly distinguished in MethylMix-PA clusters.

Among colorectal samples we are able to confirm the CpG island methylator phenotype (CIMP) (Fig 4B). Cluster-2 contains all but one of the patients labeled CIMP-High using methylation signatures and 75% of patients labeled Microsatellite Instable/CIMP using gene-expression signatures. The CIMP subtype has known association with MLH1 silencing through hyper-methylation, which is reflected in our MethylMix-PA subtypes where we find cluster-1 to include the majority of samples with non-silenced MLH1. MethylMix-PA subtypes also significantly correlate with Microsatellite Instability where samples labeled as Microsatellite Instability-Low (MSI-L) or Microsatellite Stable (MSS) are found by majority in cluster-1.

Examining subtypes in ovarian cancer our MethylMix-PA clusters agree well with molecular subtypes and are significantly correlated (Fig 4C). Cluster-1 contains 78% of Immunoreactive subtype and 78% of Differentiated subtype patients, while about half of cluster-2 is comprised of patients labeled as Proliferative. Lastly Mesenchymal subtype patients can be found with relatively equal frequencies in each cluster [40–42]. MethylMix-PA clusters also significantly correlate with tumor features, where cluster-2 and cluster-1 roughly correspond patients with lower-grade and higher-grade tumors.

Discussion

Epigenetic aberrations contribute to oncogenesis, where DNA hypermethylation inactivates tumor suppressor genes, while hypomethylation is known to promote genomic instability and activate oncogenes [12,20]. Therefore, DNA methylation has potential to inform patient treatment and improve patient outcomes through new diagnostics and therapeutics. When identifying epigenetically driven cancer genes, it is of note that most biological functions—subject to genomic and epigenomic dysregulation—are ultimately executed at the protein level, so we can expect neutralization of non-functional upstream effects at—or before—the proteome. Herein we confirm the potential of using proteomic data to elucidate functional DNA methylation events by conducting the first genome wide analysis of epigenome-proteome relationships across three large human cancer cohorts. We present MethylMix-PA, a method that formalizes the identification of abnormally methylated genes that are predictive of protein abundance, like MethylMix-GE, and uses a model-based approach, negating the use of arbitrary user-defined thresholds for abnormal DNA methylation, and identifies subpopulations of hypo or hypermethylated samples within a heterogeneous population. By integrating DNA methylation array and quantitative MS technologies, MethylMix-PA identifies candidate epigenetic driver genes with clinical value as potential therapeutic targets and protein biomarkers for assessing prognosis and treatment stratification. MethylMix-PA builds on our model MethylMix and addresses the potential limited predictive value of mRNA as proxy for phenotype due to the role of post-transcriptional mechanisms.

MethylMix-PA identifies oncogenes and tumor suppressors and—with the exception of a few genes—returns a subset of MethylMix identifications, where often the effect of DNA Methylation does not propagate to the proteome (Fig 1, S1 Table). In other cancer studies similar buffering has been observed in both cis and trans CNA effects, suggesting that many detectable aberrations in cancer do not manifest in expression changes at the protein level [6,10]. Otherwise put, many abnormally methylated genes are likely only passengers and do not functionally contribute to cancer development. Identification of a reduced set of genes in our study has pragmatic benefits for cancer research, where narrowing nominations to fewer high-quality candidates increases the likelihood of finding true targets; strongest candidates include genes identified by both models that show negative correlation between DNA methylation and both gene expression and protein abundance, and therefore have clear biological interpretations amenable to validation in the laboratory. Similar methods to identify true targets have been described, where genes that show correlation between mRNA and protein are more likely to have tumor promoting effects [10]. Conversely, novel MethylMix-PA genes should be taken with due consideration given the lack of clear mechanisms explaining how changes in DNA methylation may alter protein levels, but be undetectable at the transcript level—plausible explanations that remain to be tested include erroneous or noisy gene expression data, low mRNA stability or alternative splicing confounding expression at the RNA level.

Nevertheless, most new identifications are well supported to have tumor promoting effects and therefore warrant further investigation to uncover how DNA-methylation may influence regulation of genes like *EHF*, *FSTL1*, *PTPRC*, *S100A9*, *LTF*, *EVL*, and *TSTA3*. Importantly, in all these cases the type of DNA methylation is consistent with gene function, where known tumor-suppressors are hyper-methylated and oncogenes are hypo-methylated at regions where DNA methylation negatively regulates transcription.

Taken together MethylMix-PA genes highlight important features in cancer related to tumor features and subtypes. MethylMix-PA genes also capture oncogenic biological processes based on enrichment analysis showing key aspects of cancer development such as processes related to EMT, immune function, and proliferative signaling (Table 2, S3 Table). MethylMix-PA also elucidates more shared annotations between cancer types, and thus a greater ability to identify genes of core cancer pathways that are shared across cancer sites. Next, using a completely orthogonal approach we also find that MethylMix-PA is more descriptive of tumor progression; although this new analysis produces a reduced number of identifications, MethylMix-PA genes are more likely to correlate in expression with disease features such as tumor stage and size (Table 3). Lastly, we find MethylMix-PA performs reasonably well for patient clustering recapitulating established molecular subtypes. Given the limitations of our study, we expect our clustering to have reduced discriminative power, since we limit our observations to genes for which we have both matched gene expression and protein abundance measurements in our analysis. This significantly diminishes the feature space we used for learning. Nevertheless, we find that MethylMix-PA performs similarly to MethylMix-GE in identifying cancer subtypes such as luminal and basal types of breast cancer, the CIMP type in colorectal cancer and all subtypes in ovarian cancer, with the exception of the mesenchymal subtype which is the least clearly defined subtype [40] (Fig 2, S4 Table). These findings suggest the reduced number MethylMix-PA genes capture the major sources of variation in each cancer cohort and facilitate translatability into feasible panels for testing.

Overall MethylMix-PA shows practical utility for improving nominations of cancer driver genes and elucidating new mechanisms of cancer development missed by our previous model. More broadly our study supports using proteomic data to better understand how epigenetic deregulation promotes cancer. Similar approaches have been applied and found to potentially improve aspects of patient care. For example, a retrospective analysis of outcomes in an oncology trial for glioblastoma—which tested efficacy of different temozolomide regiments—found that updating the clustering model to incorporate MGMT protein expression and c-MET protein abundance provided better separation of overall survival prognostic groups than incorporating MGMT promoter methylation alone [43]. These findings and ours support the claim that protein data combined with DNA methylation is a better way to stratify patients and understand cancer features then using DNA methylation alone.

Although milestone initiatives like TCGA and CPTAC provide valuable date for the acceleration of discovery and research in cancer, we acknowledge the limitations of this study and further work required. A barrier to translation, the number of specimens used here is insufficient to draw conclusive clinical correlations and require replication of these results by independent studies. Importantly molecular measurements used here are also subject to sources of technical and biological bias. For example, it is known that bulk measurements obscure the complex nature of tumor microenvironment which includes many cell types including vascular, lymphatic, and immune cells. This confounding effect is compounded considering that each molecular feature was measured using different tumor fragments, which may have very different cellular compositions due to intra-tumor heterogeneity. Additionally, we recognize further characterization of genome wide proteomic studies is required to fully understand possible biases, such as worse detection of highly hydrophobic and hydrophilic peptides, or low-abundance peptides co-eluting with very high-abundance peptide [9]. Moreover, early proteomic techniques such as those utilized in CPTAC's Common Data Analysis Pipeline have not yet reached the genome level resolution of other omic measurements; these methods require refinement to address low coverage due to inherent limitations of proteolytic measurements such immeasurable peptides that are excessively large or small tryptic fragments and the inability to distinguish some amino acids [9]. This reduced coverage to a few thousand genes in our study excludes many genes with possible roles in cancer.

The complex nature of disease development and interplay between interacting biological aberrations—genetic, epigenetic, somatic or germline—often makes it difficult to elucidate causal mechanisms of cancer development. Furthermore, there is still much work in multi-omics to elucidate causal flows of information influencing cellular physiology and pathology and to discriminate how separate phenomena are linked to create cancer [3,5,42,44]. However, integrated multi-omic approaches like MethylMix-PA can provide additional insights into pathways and processes involved in oncogenesis and how they manifest as clinical phenotypes. As CPTAC moves into its second phase and characterizes more samples across more cancer types, models such as MethylMix-PA may leverage this valuable data to improve understanding of the molecular basis of cancer.

Methods

Ethics statement

All data used in this study is third party data, and is available from the following articles [6,9,10,45–47]. All other data is available within the paper and Supporting Information files.

Data processing

Molecular data were produced from tissue bio-specimens from three cancer cohorts: breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD), and ovarian serous cystadenocarcinoma (OV) (Table 1).

DNA methylation. CpG site methylation levels/percentages were measured using Illumina Infinium Human Methylation 27k and 450k BeadChip Platforms [45–47]. We limit our observations to overlapping probes or CpG sites for cancer tissues measured using both platforms, otherwise we use all available probes. The methylation level is recorded as a beta value representing a ratio of the signal/intensity from the methylated probe over the sum of both the methylated probe and the unmethylated probes. Values close to 0 indicate low levels of DNA methylation and values close to 1 indicate high levels of DNA methylation. We removed CpG sites with more than 10% missing entries across all samples and we applied 15-K Nearest Neighbor (KNN) to impute the remaining missing values, this procedure was replicated for all molecular data types. We observed significant technical sources of variation among tissue samples processed in batches using a one-way analysis, which we corrected using COMBAT [48]. To reduce dimensionality of the CpG data we applied hierarchical clustering with complete linkage and a minimum average Pearson correlation of 0.4 between values. Last, we matched clusters to their corresponding genes by mapping to the closest transcriptional start sites, where one gene may relate to many CpG clusters but each CpG cluster only maps to one gene. Therefore, we limit our analysis to DNA methylation states with cis regulation effects.

RNA expression. We used transcriptomic data in MethylMix produced by RNA sequencing technology [45–47]. We log-transformed the RNAseq counts and replaced infinities with a non-zero low value. Similar to our DNA methylation data processing, we estimated missing values using 15-KNN and used COMBAT to correct for batch effects [48].

Protein abundance. Proteomic data used in MethylMix-PA was provided by CPTAC [6,9,10]. Participating research institutions used the following Common Data Analysis Pipeline to produce protein level measurements: First tissue samples were enzymatically digested, cutting large proteins in a sequence specific manner into smaller peptides. Peptides were fractioned using liquid chromatography to improve downstream quality before measurements using Thermo Fisher high-resolution tandem mass spectrometry (LC-MS/MS). Next, the resultant mass ladders were matched to theoretical mass ladders in the FASTA database and subsequently assigned to peptide spectra using software tools and The Reference Sequence Database. The data was then filtered to exclude peptide fragments common to more than one protein and to only include protein-identifying or unshared peptides i.e. fragments with unique sequences. Lastly peptides were mapped to a parsimonious set of genes.

The BRCA and OV workflows used iTRAQ-labeling to increase throughput, where 3 patient samples are isotopically labelled and analyzed against a common reference standard and describe relative ion intensities. Quantities are recorded after taking the log2 ratio of the abundances. Alternatively, measurement of COADREAD samples used label free MS technology and are reported as absolute counts, which were transformed to relative quantities by taking the log2 of quantile normalized values using the limma R package [10]. OV samples collected from Pacific Northwest National Laboratory and John Hopkins University were merged and corrected for batch effects using COMBAT [48].

To remove samples compromised by protein degradation we filtered samples using the QC method described by Mertins et al. [6]: we calculated the standard deviation of non-normalized protein measurements across all genes for each sample and segmented samples into groups using a two component Gaussian mixture model. Samples belong to the poor-quality group i.e. higher mean standard deviation were excluded from study. Applying this method we discarded 28 BRCA and 5 OV samples. Finally, for each cancer we removed samples with greater than 75% missing values, estimating the remaining missing values using 15-KNN algorithm [49].

Algorithm

Step 1: Fit mixture model to methylation data. As described earlier methylation levels are recorded as beta values or values ranging from 0 to 1 representing the percentage of methylation and therefore gene values across all samples are beta distributed. MethylMix identifies subgroups of patients with a distinct methylation pattern or state by finding the beta mixture model with the number of components that best describe the data. To map samples to subgroups we iteratively add components requiring that each additional component improve the Bayesian Information Criterion (BIC) to avoid overfitting. To define the most descriptive subgroups we include methylation measurements across all samples, however our model integrates epigenetic data with proteomic and transcriptomic data using only the subset of these samples with available matched data (Table 1).

Step 2: Compare methylation to normal tissue. To identify differentially methylated CpG clusters we compare the mean methylation level—the mean value of the beta mixture component—to the mean methylation level of normal samples. To measure if an observed difference is significant we perform a Wilcoxon rank sum test with a Q-value cutoff of 0.05, using both p-value multiple testing correction with False Discover Rate (FDR). As an additional measure, we require a minimum difference of 0.10 based on the platform sensitivity [50]. If significant, the difference in methylation level between the mode and normal is recorded as the Differential Methylation value or DM value for each methylation state.

Step 3: Select for functionally predictive genes. Next, we filter our set of genes, requiring that genes be not only differentially methylated when compared to normal but also predictive of gene expression or protein abundance. Hyper-methylation should lower gene expression and corresponding protein abundance when compared to the normally methylated samples, therefore we only accept genes that have a negative correlation between methylation level and downstream gene products. Note this assumption is only explanatory of methylation at promoter regions and does not necessarily apply to methylation at the gene-body or 3' and 5' untranslated regions (UTRs). To assess the likelihood that methylation events are functional, MethylMix-GE uses the relationship between methylation and gene expression, whereas MethylMix-PA examines the effect on protein abundance. In both cases, we perform a linear regression between methylation levels and RNA expression or protein abundance data respectively. We use the R-square statistic to estimate the magnitude of the correlation and used cutoffs at R-square greater than 0.05 and a Q-value of 0.01 using FDR multiple testing correction.

Applying the procedures outlined above for MethylMix-GE and MethylMix-PA each produces a list of candidate cancer drivers (referenced as MethylMix-GE and MethylMix-PA genes) and a corresponding matrix of DM-values for identified CpG clusters across all samples. All MethylMix genes have the following statistical properties: (i) a DNA methylation difference based on the beta mixture model that is significantly different from normal that is > = 0.1 based on the platform sensitivity [50], and (ii) an R-square statistic greater than 0.05 with a Q-value less than 0.01 for correlation with gene expression and protein abundance for MethylMix-GE and MethylMix PA respectively.

To assess the validity of each list we used orthogonal clinical and biological data to assess utility for downstream analysis and relevance to disease state.

Evaluation

GO term enrichment. To describe the underlying biological processes captured by each model, we tested for enrichment of Gene Ontology (GO) terms in MethylMix-GE and MethylMix-PA genes. This analysis was implemented using the PANTHER Classification System's statistical overrepresentation tool [51] with the following settings: Homo-sapiens for organism, the background set to include all genes with matching protein and RNA data, and either MethylMix-PA or MethylMix-GE genes for input. Enrichment was calculated using fisher's exact test. For each gene list we rank terms using significance of the test statistic with a minimum p-value of 0.10.

Methylation subtypes. With the matrices of DM values for our CpG clusters we performed consensus clustering to identify robust groupings of patients based on epigenetic signatures [52]. Our analysis for each cancer cohort used the following parameters: maximum number of clusters is 6, number of bootstrap subsamples is 500 with 0.8 the proportion of the subsample, and our method uses k-means cluster algorithm and Euclidean distance. To identify the optimal number of clusters we inspected the proportion of ambiguous classification (PAC Score) [53,54], and the consensus heatmap and values, where the score/index between two samples is the proportion of clustering runs in which the two items are clustered together. We define the intra cluster consensus as the mean of all pairwise consensus scores between samples clustered in the same group, and inter cluster consensus as the mean of all consensus indexes between a sample and all the other samples clustered in different groups. A robust clustering result ideally shows high intra cluster consensus and low inter cluster consensus. We tested for association between cluster assignments and several disease features, using a Chi-squared test for categorical variables such as molecular subtype labels or a Kruskal-Wallis test for ordinal values such as tumor grade. Our analysis includes genetic, molecular, and clinical annotations, which were collected from supplementary tables from the original TCGA publications [45–47] in addition to annotations downloaded using the TCGAbiolinks R package [55].

Enrichment for putative tumor markers. We compared MethylMix-GE and MethylMix-PA genes by investigating their enrichment in genes related to disease progression. We used correlation of gene expression with cancer stage and tumor size to identify potential genes capturing disease progression. We took the spearman correlation between gene expression levels and these clinical variables using all available samples. We selected genes using a p-value cutoff of 0.05 and biased for genes with greater likelihood of relevance by taking top 50th quantile in sample variance. Next, we filtered for only relationships that can be explained by methylation, such that genes identified as hyper-methylated in cancer tissue were required to show a negative correlation between gene expression and disease progression (tumor-suppressor genes) and hypo-methylated genes positively correlated (oncogenes). To assess each models' likelihood in picking up genes related to disease progression we examined the overlap between these genes and the MethylMix-PA and MethylMix-GE genes, using Fisher's exact test to evaluate significance.

Supporting information

S1 Table. Gene level results from MethylMix-GE and MethylMix-PA for each cancer site and summary statistics from linear regression taken between DNA-methylation beta values and gene-expression or protein abundance values, respectively.
(XLSX)

S2 Table. Results for spearman correlations taken for each gene mRNA-protein pair.
(XLSX)

S3 Table. Gene Ontology term enrichments found for each cancer site from MethylMix-GE and MethylMix-PA gene lists.
(XLSX)

S4 Table. Enrichment found between cluster assignments and various clinical and molecular features, taken using Chi-squared and Kruskal-Wallis tests.
(XLSX)

Author Contributions

Conceptualization: Majed Mohamed Magzoub, Marcos Prunello, Kevin Brennan, Olivier Gevaert.

Data curation: Majed Mohamed Magzoub, Kevin Brennan, Olivier Gevaert.

Formal analysis: Majed Mohamed Magzoub, Marcos Prunello, Kevin Brennan, Olivier Gevaert.

Funding acquisition: Olivier Gevaert.

Investigation: Majed Mohamed Magzoub, Olivier Gevaert.

Methodology: Majed Mohamed Magzoub, Marcos Prunello, Olivier Gevaert.

Project administration: Olivier Gevaert.

Resources: Kevin Brennan, Olivier Gevaert.

Software: Majed Mohamed Magzoub, Olivier Gevaert.

Supervision: Olivier Gevaert.

Validation: Majed Mohamed Magzoub, Olivier Gevaert.

Visualization: Kevin Brennan, Olivier Gevaert.

Writing – original draft: Majed Mohamed Magzoub, Olivier Gevaert.

Writing – review & editing: Majed Mohamed Magzoub, Marcos Prunello, Kevin Brennan, Olivier Gevaert.

References

1. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018; 173: 338–354. e15. https://doi.org/10.1016/j.cell.2018.03.034 PMID: 29625051
2. Campbell JD, Yau C, Bowlby R, Liu Y, Brennan K, Fan H, et al. Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. Cell Rep. 2018; 23: 194–212.e6. https://doi.org/10.1016/j.celrep.2018.03.063 PMID: 29617660
3. Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. EBioMedicine. 2018; 27: 156–166. https://doi.org/10.1016/j.ebiom.2017.11.028 PMID: 29331675
4. Cheerla N, Gevaert O. MicroRNA based pan-cancer diagnosis and treatment recommendation. BMC Bioinformatics. 2017; https://doi.org/10.1186/s12859-016-1421-y PMID: 28086747
5. Manolakos A, Ochoa I, Venkat K, Goldsmith AJ, Gevaert O. CaMoDi: A new method for cancer module discovery. BMC Genomics. 2014; https://doi.org/10.1186/1471-2164-15-S10-S8 PMID: 25560933
6. Mertins P, Mani DR, Ruggles K V, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016; 534: 55–62. https://doi.org/10.1038/nature18003 PMID: 27251275
7. Kosti I, Jain N, Aran D, Butte AJ, Sirota M. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. Sci Rep. 2016; 6: 24799. https://doi.org/10.1038/srep24799 PMID: 27142790
8. Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi D V, et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. J Proteome Res. 2016; 15: 1023–32. https://doi.org/10.1021/acs.jproteome.5b01091 PMID: 26860878
9. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell. 2016; 166: 755–765. https://doi.org/10.1016/j.cell.2016.05.069 PMID: 27372738
10. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014; 513: 382–7. https://doi.org/10.1038/nature13438 PMID: 25043054
11. Ning K, Fermin D, Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J Proteome Res. 2012; 11: 2261–71. https://doi.org/10.1021/pr201052x PMID: 22329341
12. Paska AV, Hudler P. Aberrant methylation patterns in cancer: a clinical view. Biochem medica. 2015; 25: 161–76. https://doi.org/10.11613/BM.2015.017 PMID: 26110029
13. Weisenberger DJ. Characterizing DNA methylation alterations from The Cancer Genome Atlas. J Clin Invest. 2014; 124: 17–23. https://doi.org/10.1172/JCI69740 PMID: 24382385
14. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012; 13: 484–92. https://doi.org/10.1038/nrg3230 PMID: 22641018
15. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002; 3: 415–28. https://doi.org/10.1038/nrg816 PMID: 12042769
16. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, et al. A DNA methylation fingerprint of 1628 human samples. Genome Res. 2012; 22: 407–19. https://doi.org/10.1101/gr.119867.110 PMID: 21613409
17. Litovkin K, Van Eynde A, Joniau S, Lerut E, Laenen A, Gevaert T, et al. DNA methylation-guided prediction of clinical failure in high-risk prostate cancer. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0130651 PMID: 26086362

18. Litovkin K, Joniau S, Lerut E, Laenen A, Gevaert O, Spahn M, et al. Methylation of PITX2, HOXD3, RASSF1 and TDRD1 predicts biochemical recurrence in high-risk prostate cancer. J Cancer Res Clin Oncol. 2014; https://doi.org/10.1007/s00432-014-1738-8 PMID: 24938434

19. Gevaert O. MethylMix: An R package for identifying DNA methylation-driven genes. Bioinformatics. 2015. https://doi.org/10.1093/bioinformatics/btv020 PMID: 25609794

20. Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biol. 2015; https://doi.org/10.1186/s13059-014-0579-8 PMID: 25631659

21. Brennan K, Koenig JL, Gentles AJ, Sunwoo JB, Gevaert O. Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the CpG island methylator phenotype. EBioMedicine. 2017; 17. https://doi.org/10.1016/j.ebiom.2017.02.025 PMID: 28314692

22. Cedoz P-L, Prunello M, Brennan K, Gevaert O. MethylMix 2.0: an R package for identifying DNA methylation genes. Bioinformatics. 2018; https://doi.org/10.1093/bioinformatics/bty156 PMID: 29668835

23. Brennan K, Shin JH, Tay JK, Prunello M, Gentles AJ, Sunwoo JB, et al. NSD1 inactivation defines an immune cold, DNA hypomethylated subtype in squamous cell carcinoma. Sci Rep. 2017; 7: 17064. https://doi.org/10.1038/s41598-017-17298-x PMID: 29213088

24. Zhou X, Xiao X, Huang T, Du C, Wang S, Mo Y, et al. Epigenetic inactivation of follistatin-like 1 mediates tumor immune evasion in nasopharyngeal carcinoma. Oncotarget. 2016; 7: 16433–44. https://doi.org/10.18632/oncotarget.7654 PMID: 26918942

25. Xu J, Wu H, Zhang C, Cao Y, Wang L, Zeng L, et al. Identification of a novel human DDX40gene, a new member of the DEAH-box protein family. J Hum Genet. 2002; 47: 681–3. https://doi.org/10.1007/s100380200104 PMID: 12522690

26. Shao Y, Sun X, He Y, Liu C, Liu H. Elevated levels of serum tumor markers CEA and CA15-3 are prognostic parameters for different molecular subtypes of breast cancer. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0133830

27. Nishimukai A, Akazawa K, Miyoshi Y, Masai Y, Araki K, Michishita S, et al. Independent prognostic impact of preoperative serum carcinoembryonic antigen and cancer antigen 15–3 levels for early breast cancer subtypes. World J Surg Oncol. World Journal of Surgical Oncology; 2018; 16: 1–11. https://doi.org/10.1186/s12957-017-1299-9

28. Yang JB, Zhao Z Bin, Liu QZ, Hu TD, Long J, Yan K, et al. FoxO1 is a regulator of MHC-II expression and anti-tumor effect of tumor-associated macrophages. Oncogene. 2018; https://doi.org/10.1038/s41388-017-0048-4 PMID: 29238041

29. Cui J, Saevarsdottir S, Thomson B, Padyukov L, van der Helm-van Mil AHM, Nititham J, et al. Rheumatoid arthritis risk allele PTPRC is also associated with response to anti-tumor necrosis factor alpha therapy. Arthritis Rheum. 2010; 62: 1849–61. https://doi.org/10.1002/art.27457 PMID: 20309874

30. Myung JK, Yeo S-G, Kim KH, Baek K-S, Shin D, Kim JH, et al. Proteins that interact with calgranulin B in the human colon cancer cell line HCT-116. Oncotarget. 2017; 8: 6819–6832. https://doi.org/10.18632/oncotarget.14301 PMID: 28036279

31. Wu P, Quan H, Kang J, He J, Luo S, Xie C, et al. Downregulation of Calcium-Binding Protein S100A9 Inhibits Hypopharyngeal Cancer Cell Proliferation and Invasion Ability Through Inactivation of NF-κB Signaling. Oncol Res. 2017; 25: 1479–1488. https://doi.org/10.3727/096504017X14886420642823

32. MacManus CF, Collins CB, Nguyen TT, Alfano RW, Jedlicka P, de Zoeten EF. VEN-120, a Recombinant Human Lactoferrin, Promotes a Regulatory T Cell [Treg] Phenotype and Drives Resolution of Inflammation in Distinct Murine Models of Inflammatory Bowel Disease. J Crohns Colitis. 2017; 11: 1101–1112. https://doi.org/10.1093/ecco-jcc/jjx056 PMID: 28472424

33. Naleskina LA, Lukianova NY, Sobchenko SO, Storchai DM, Chekhun VF. Lactoferrin expression in breast cancer in relation to biologic properties of tumors and clinical features of disease. Exp Oncol. 2016; 38: 181–6. PMID: 27685526

34. Hu L-D, Zou H-F, Zhan S-X, Cao K-M. EVL (Ena/VASP-like) expression is up-regulated in human breast cancer and its relative expression level is correlated with clinical stages. Oncol Rep. 2008; 19: 1015–20. PMID: 18357390

35. Takaku M, Machida S, Hosoya N, Nakayama S, Takizawa Y, Sakane I, et al. Recombination activator function of the novel RAD51- and RAD51B-binding protein, human EVL. J Biol Chem. 2009; 284: 14326–36. https://doi.org/10.1074/jbc.M807715200 PMID: 19329439

36. Mitrović A, Pečar Fonović U, Kos J. Cysteine cathepsins B and X promote epithelial-mesenchymal transition of tumor cells. Eur J Cell Biol. 2017; https://doi.org/10.1016/j.ejcb.2017.04.003 PMID: 28495172

37. Tang SC, Wu MF, Wong RH, Liu YF, Tang LC, Lai CH, et al. Epigenetic mechanisms for silencing glutathione S-transferase m2 expression by hypermethylated specificity protein 1 binding in lung cancer. Cancer. 2011; https://doi.org/10.1002/cncr.25875 PMID: 21246532

38. Wei J, Li G, Zhang J, Zhou Y, Dang S, Chen H, et al. Integrated analysis of genome-wide DNA methylation and gene expression profiles identifies potential novel biomarkers of rectal cancer. Oncotarget. 2016; https://doi.org/10.18632/oncotarget.11534 PMID: 27566576

39. Brüggemann M, Gromes A, Poss M, Schmidt D, Klümper N, Tolkach Y, et al. Systematic Analysis of the Expression of the Mitochondrial ATP Synthase (Complex V) Subunits in Clear Cell Renal Cell Carcinoma. Transl Oncol. 2017; 10: 661–668. https://doi.org/10.1016/j.tranon.2017.06.002 PMID: 28672194

40. Planey CR, Gevaert O. CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. Genome Med. 2016; https://doi.org/10.1186/s13073-016-0281-4 PMID: 26961683

41. Willis S, Villalobos VM, Gevaert O, Abramovitz M, Williams C, Sikic BI, et al. Single gene prognostic biomarkers in ovarian cancer: A meta-analysis. PLoS One. 2016; https://doi.org/10.1371/journal.pone.0149183 PMID: 26886260

42. Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus. 2013; https://doi.org/10.1098/rsfs.2013.0013 PMID: 24511378

43. Bell EH, Pugh SL, McElroy JP, Gilbert MR, Mehta M, Klimowicz AC, et al. Molecular-Based Recursive Partitioning Analysis Model for Glioblastoma in the Temozolomide Era: A Correlative Analysis Based on NRG Oncology RTOG 0525. JAMA Oncol. 2017; 3: 784–792. https://doi.org/10.1001/jamaoncol.2016.6020 PMID: 28097324

44. Gevaert O, Plevritis S. Identifying master regulators of cancer and their downstream targets by integrating genomic and epigenomic features. Pac Symp Biocomput. 2013; https://doi.org/10.1142/9789814447973_0013

45. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490: 61–70. https://doi.org/10.1038/nature11412 PMID: 23000897

46. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487: 330–7. https://doi.org/10.1038/nature11252 PMID: 22810696

47. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474: 609–15. https://doi.org/10.1038/nature10166 PMID: 21720365

48. Parker HS, Leek JT, Favorov A V, Considine M, Xia X, Chavan S, et al. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics. 2014; 30: 2757–63. https://doi.org/10.1093/bioinformatics/btu375 PMID: 24907368

49. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17: 520–5. https://doi.org/10.1093/bioinformatics/17.6.520 PMID: 11395428

50. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. 2006; 16: 383–93. https://doi.org/10.1101/gr.4410706 PMID: 16449502

51. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013; 8: 1551–66. https://doi.org/10.1038/nprot.2013.092 PMID: 23868073

52. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010; 26: 1572–3. https://doi.org/10.1093/bioinformatics/btq170 PMID: 20427518

53. Ṣenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. Sci Rep. 2014; 4: 6207. https://doi.org/10.1038/srep06207 PMID: 25158761

54. Sweeney TE, Chen AC, Gevaert O. Combined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number, K. Sci Rep. 2015; https://doi.org/10.1038/srep16971 PMID: 26581809

55. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016; 44: e71. https://doi.org/10.1093/nar/gkv1507 PMID: 26704973