CPTAC Proteomic Data Processing

Olivier Gevaert
Department Medicine
Department Biomedical Data Science
Stanford University

Overview

- CPTAC data download and preprocessing issues
  - Quality controls
  - Batch correction
  - Different types of proteomics
- Integration of CPTAC proteomic data with other omics
  - ProteoMix: Example of integration with DNA methylation
- Proposed radioproteomics maps
  - Comparison with traditional radiogenomics maps.

CTPAC data download & preprocessing issues

Technical steps

Stanford MEDICINE

Clinical Proteomic Tumor Analysis Consortium (CPTAC)

Objective: Understand the molecular basis of cancer that is not fully elucidated or not possible through genomics by adding complementary functional layer of protein biology and to accelerate the translation of molecular findings into the clinical.

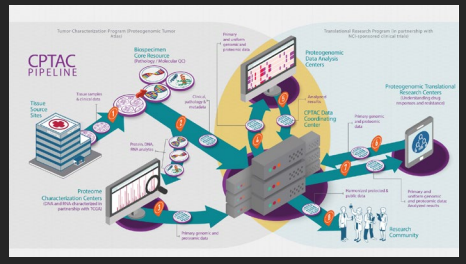
Timeline:

2006-2011 Phase I:
Process Development and Reproducibility

2011-2016 Phase II:
Proteogenomic Discovery

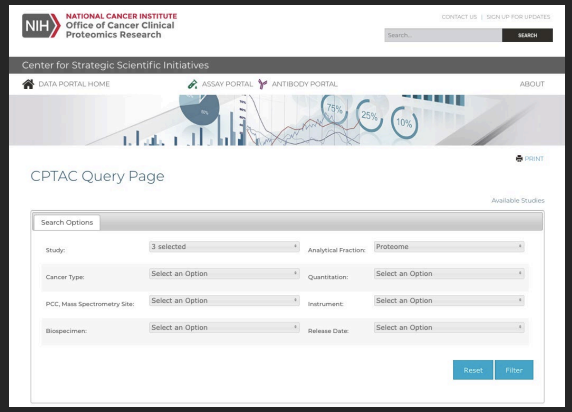
NIH
NATIONAL CANCER INSTITUTE
TCGA

2016-Present Phase III:
Expansion & Clinical Translation

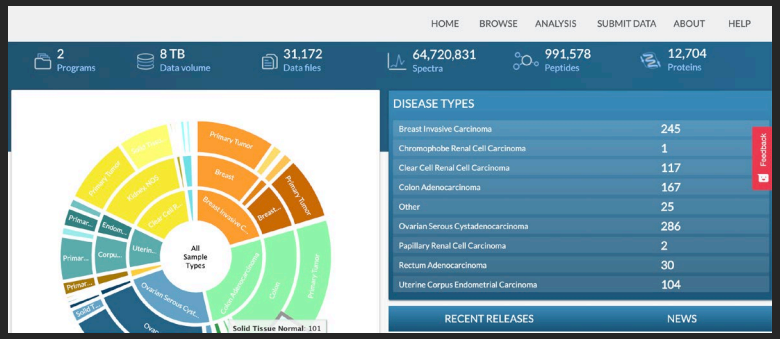


Data Download:

- Method 1: Data portal
  - Hosted by Georgetown University
  - Requires IBM Aspera Launcher
  - https://cptac-data-portal.georgetown.edu



- Method 2: NCI Proteomic Data Commons (in beta)
  - Similar to genomic data commons
  - https://pdc.esacinc.com/pdc/pdc



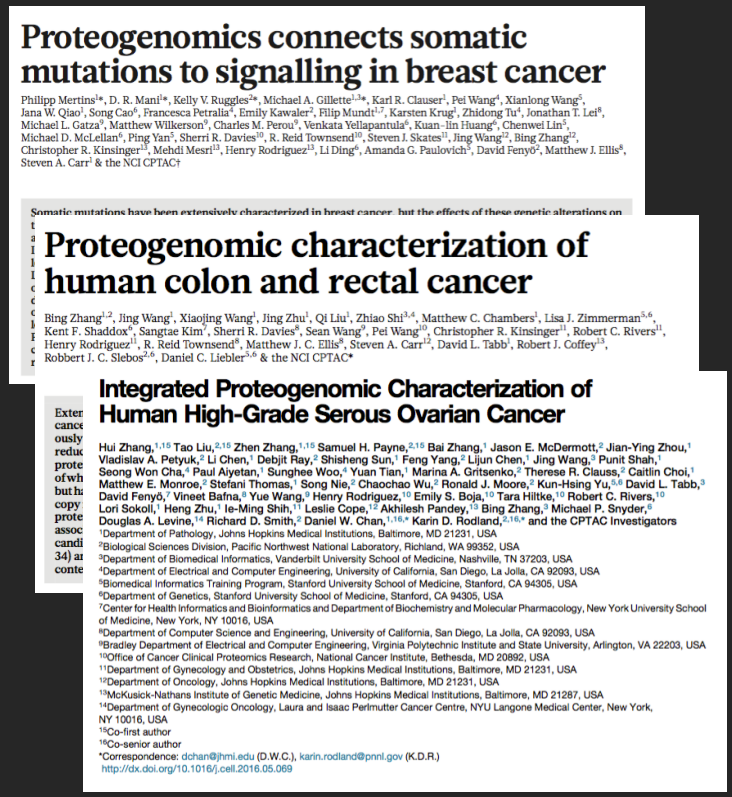
- Method 3: Programmatically
  - Direct access using Linux and command line or Python executable script
  - https://proteomics.cancer.gov/data-portal/about/faqs

CPTAC Data Overview: Phase 2 data

Generated using Common Data Analysis Pipeline (CDAP)

Dataset	Number Samples	Number Genes
BRCA – BROAD Institute	105	10624
COADREAD – Vanderbilt Uni.	95	5561
OV – Johns Hopkins Uni.	115*	8597
OV – PNNL	75*	7480

* 32 samples in common



Proteogenomics connects somatic mutations to signalling in breast cancer

Philipp Mertins[1]*, D. R. Mani[1]*, Kelly V. Ruggles[2]*, Michael A. Gillette[1,3]*, Karl R. Clauser[1], Pei Wang[4], Xianlong Wang[5], Jana W. Qiao[1], Song Cao[6], Francesca Petralia[4], Emily Kawaler[2], Filip Mundt[1,7], Karsten Krug[1], Zhidong Tu[4], Jonathan T. Lei[8], Michael L. Gatza[9], Matthew Wilkerson[9], Charles M. Perou[9], Venkata Yellapantula[6], Kuan-lin Huang[6], Chenwei Lin[5], Michael D. McLellan[6], Ping Yan[5], Sherri R. Davies[10], R. Reid Townsend[10], Steven J. Skates[11], Jing Wang[12], Bing Zhang[12], Christopher R. Kinsinger[13], Mehdi Mesri[13], Henry Rodriguez[13], Li Ding[6], Amanda G. Paulovich[3], David Fenyö[2], Matthew J. Ellis[8], Steven A. Carr[1] & the NCI CPTAC†

Proteogenomic characterization of human colon and rectal cancer

Bing Zhang[1,2], Jing Wang[1], Xiaojing Wang[1], Jing Zhu[1], Qi Liu[1], Zhiao Shi[3,4], Matthew C. Chambers[1], Lisa J. Zimmerman[5,6], Kent F. Shaddox[6], Sangtae Kim[7], Sherri R. Davies[8], Sean Wang[9], Pei Wang[10], Christopher R. Kinsinger[11], Robert C. Rivers[11], Henry Rodriguez[11], R. Reid Townsend[8], Matthew J. C. Ellis[8], Steven A. Carr[12], David L. Tabb[1], Robert J. Coffey[13], Robbert J. C. Slebos[5,6], Daniel C. Liebler[5,6] & the NCI CPTAC*

Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer

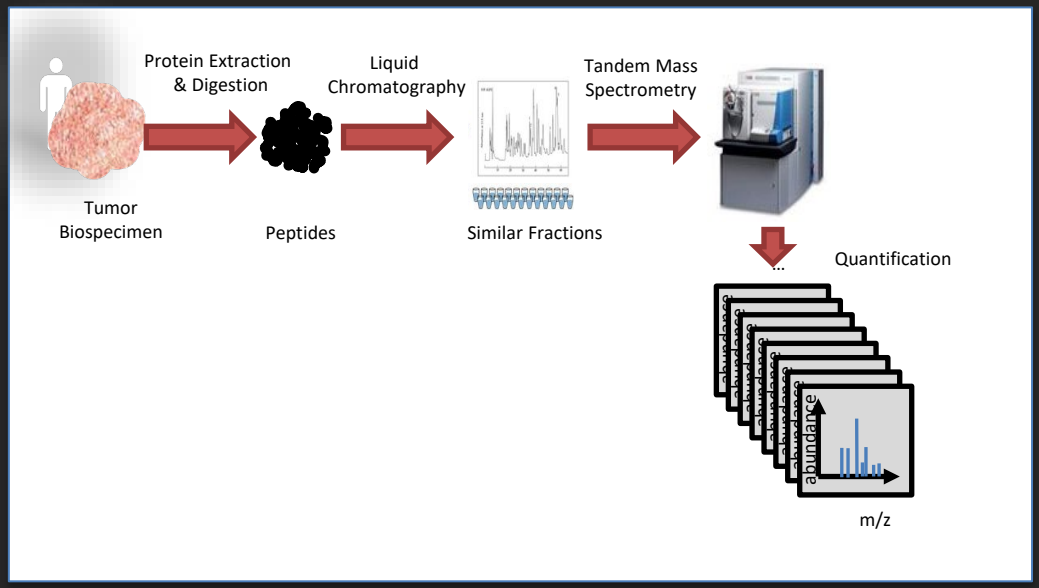
Hui Zhang,[1,15] Tao Liu,[2,15] Zhen Zhang,[1,15] Samuel H. Payne,[2,15] Bai Zhang,[1] Jason E. McDermott,[2] Jian-Ying Zhou,[1] Vladislav A. Petyuk,[2] Li Chen,[1] Debjit Ray,[2] Shisheng Sun,[1] Feng Yang,[2] Lijun Chen,[1] Jing Wang,[3] Punit Shah,[1] Seong Won Cha,[4] Paul Aiyetan,[1] Sunghee Woo,[4] Yuan Tian,[1] Marina A. Gritsenko,[2] Therese R. Clauss,[2] Caitlin Choi,[1] Matthew E. Monroe,[2] Stefani Thomas,[1] Song Nie,[2] Chaochao Wu,[2] Ronald J. Moore,[2] Kun-Hsing Yu,[5,6] David L. Tabb,[3] David Fenyö,[7] Vineet Bafna,[8] Yue Wang,[9] Henry Rodriguez,[10] Emily S. Boja,[10] Tara Hiltke,[10] Robert C. Rivers,[10] Lori Sokoll,[1] Heng Zhu,[1] Ie-Ming Shih,[11] Leslie Cope,[12] Akhilesh Pandey,[13] Bing Zhang,[3] Michael P. Snyder,[6] Douglas A. Levine,[14] Richard D. Smith,[2] Daniel W. Chan,[1,16,*] Karin D. Rodland,[2,16,*] and the CPTAC Investigators

[1]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA
[2]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA
[3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA
[4]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[5]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA 94305, USA
[6]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[7]Center for Health Informatics and Bioinformatics and Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY 10016, USA
[8]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA
[9]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA
[10]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Bethesda, MD 20892, USA
[11]Department of Gynecology and Obstetrics, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA
[12]Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA
[13]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21287, USA
[14]Department of Gynecologic Oncology, Laura and Isaac Perlmutter Cancer Centre, NYU Langone Medical Center, New York, NY 10016, USA
[15]Co-first author
[16]Co-senior author
*Correspondence: dchan@jhmi.edu (D.W.C.), karin.rodland@pnnl.gov (K.D.R.)
http://dx.doi.org/10.1016/j.cell.2016.05.069

Ongoing phase 3

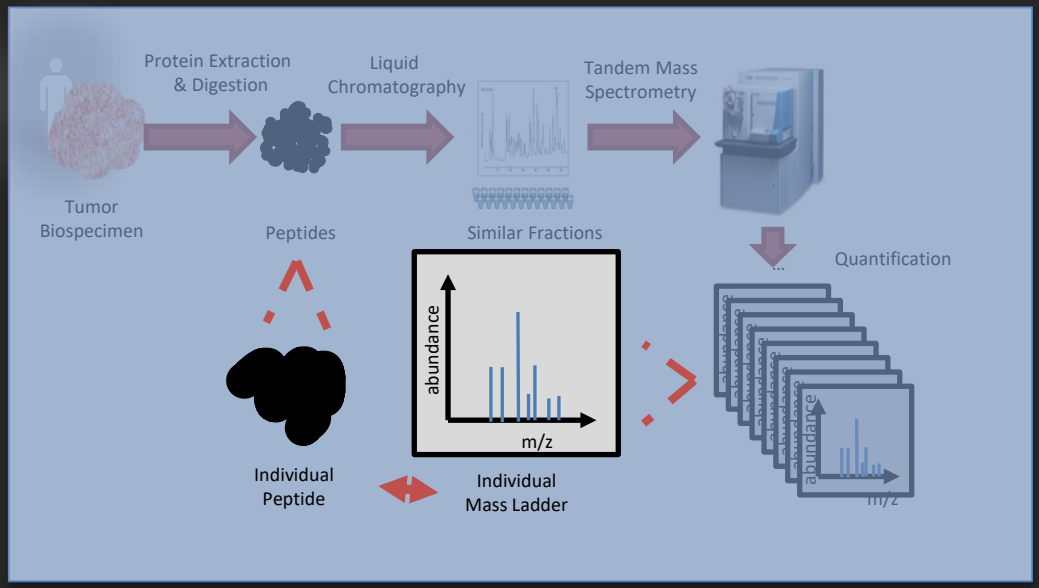
- 10 new cancer sites
  - AML
  - CCRCC
  - Cutaneous Melanoma
  - GBM
  - HNSCC
  - LSCC
  - LUAD
  - Ductal Adenocarcinoma
  - Sarcomas
  - UCEC

Proteomic workflow
Experimental Quantification



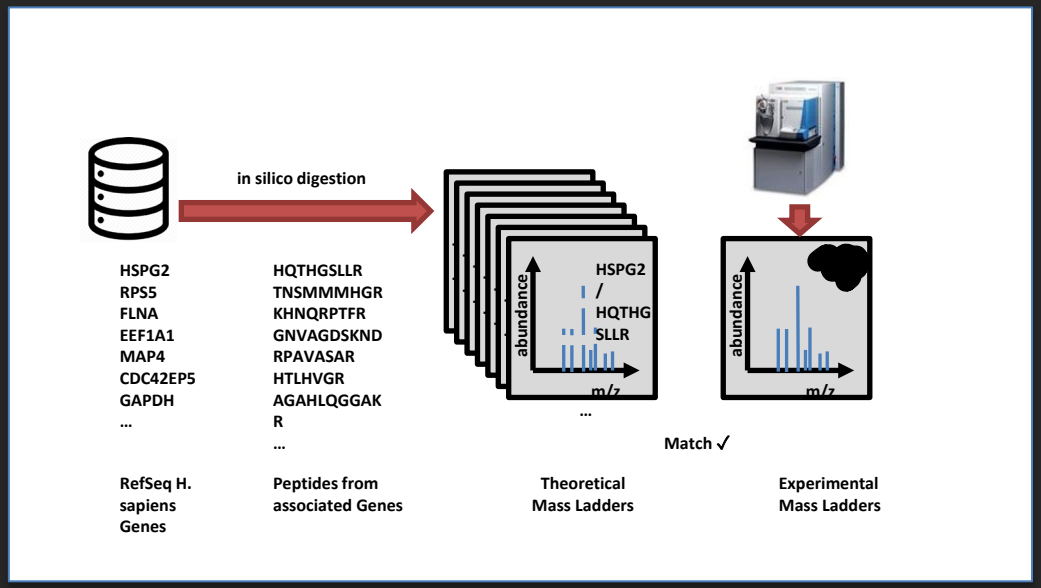
- Proteins extracted from tumor biospecimens from matched TCGA samples and proteins were tryptically digested into peptides, small segments of 7-30 amino acids
- Multi-stage high performance liquid chromatography produced homogenous fractions
- High resolution tandem mass spectrometry measures individual peptides at a time

Experimental Quantification



- Processing raw data produces a mass ladder for each peptide or group of few peptides
- Each peak corresponds to a sub-peptide
- These mass ladders are compared to theoretical mass ladders to identify peptides & proteins using the RefSeq database

Bioinformatics: Peptide Matching



- Before using abundance measurements for genomic analysis mass ladders must be linked to associated peptides and genes
- Mass ladders are identified using the RefSeq Database
  - Sequences for individual genes were fragmented in silico into peptides
  - Composition of peptides used to generate theoretical mass ladders
  - Mass ladders from experimental samples were matched by searching against theoretical mass ladders

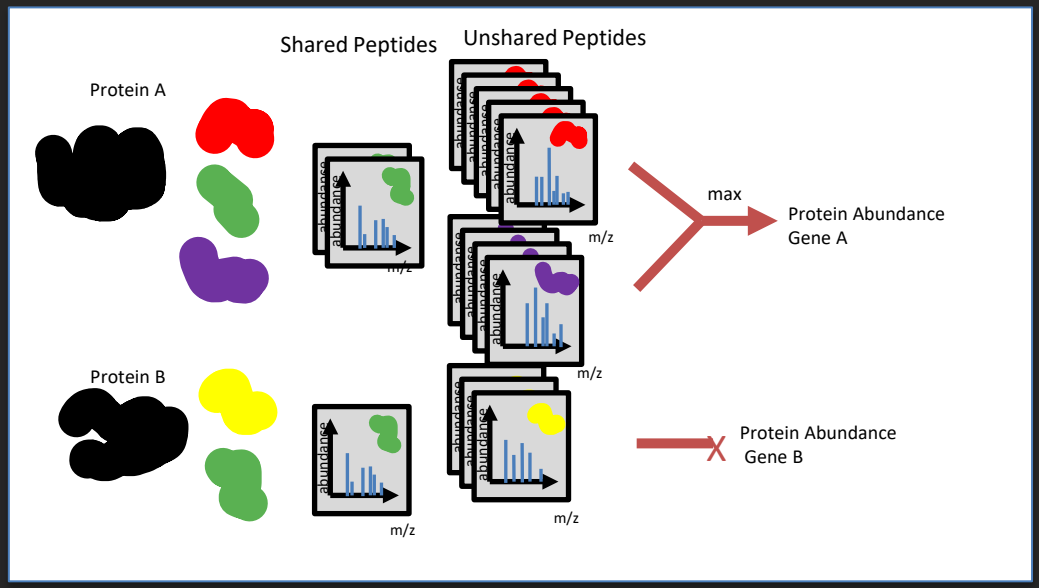
CDAP: Common Data Analysis Pipeline

	CDAP	Broad	JHU	PNNL	Vanderbilt
FASTA	RefSeq-Human-v37-Trypsin.fasta (32,800 entries)	Same as CDAP	Same as CDAP	Same as CDAP	humanRefSeq_v54_trypsin.fasta (34,589 entries)
Search Engine(s)	MS-GF+ (v9733)	SpectrumMill 4.0 (Beta)	MS-GF+ (v9146)	1. MS-GF+ v9324 (2/27/2013) v9358 (3/05/2013) v9593 (05/06/2013) v9699 (07/26/2013) v9736 (09/16/2013)	1. Pepitome 1.0.42 (library) 2. MyriMatch 2.1.87 3. MS-GF+ (v9176)
Ambiguous matches flagged?	Yes	No	No	Yes	Yes
Variable Protein Mods searched	MetOx(+16) Deamidation(+1)	MetOx(+16) Glu->pyro-Glu(-18) Gln->pyro-Glu(-17) Deamidation (N)(+1)	MetOx(+16)	MetOx(+16)	MetOx(+16) Glu->pyro-Glu(-18) Gln->pyro-Glu(-17) Acetylation (+42)
Semi-tryptic searched	Yes	No	Yes	Yes	Yes
Precursor tolerance	20 ppm	20 ppm	10 ppm	10 ppm (post-DTARefinery)	20 ppm
Missed Cleavages	No limit	<5	<2 post search	No limit	By search engine
False Discovery Rate	1% PSM	1% PSM	1 % peptide	1% Peptide	1% PSM

1

- CDAP is a standard for analyzing proteomic data proposed by CPTAC
- Some historical data from previous phases deviates from CDAP
  - Individual research institutions were free to select different analysis methods for quantifications based on specific needs
- However the Common Data Analysis Pipeline (CDAP) methods are the standard moving forward

Bioinformatics: Gene Level Assembly



- Peptide identifications used to map abundance values to genes; multiple peptides per gene must be considered
- Peptides sequences which are not unique to single gene, shared peptides, can be excluded depending on data analysis choices
- Remaining peptides (min 2) aggregated at gene level using max abundance per peptide

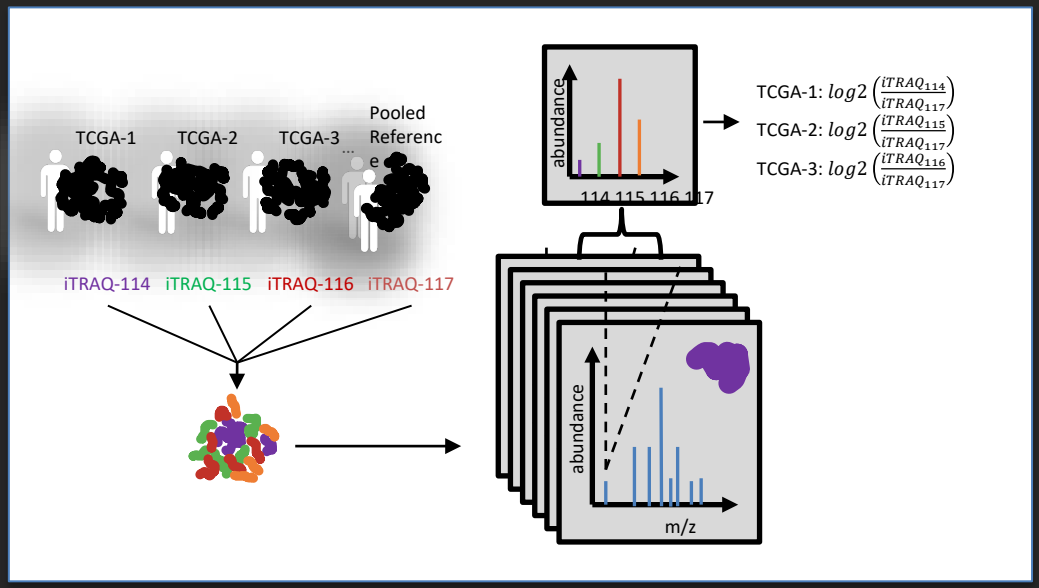
CDAP more info

- https://pdc.esacinc.com/data-dictionary/harmonization.html
- https://www.ncbi.nlm.nih.gov/pubmed/26860878

Proteomic technologies

- Two versions:
  - Label free
  - iTRAQ

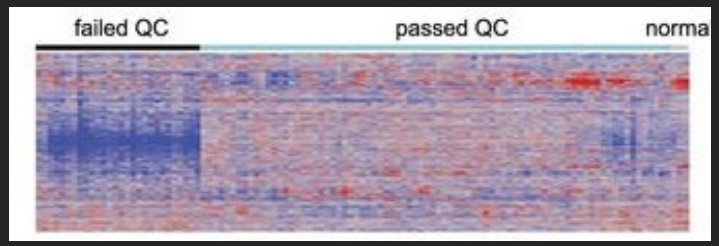
iTRAQ Data Format (BRCA & OV)



- iTRAQ: Isobaric tags for relative and absolute quantitation – low molecular weight ions are used to tag peptides from each sample
  → tighter quantification
- 4-plex measurements are made: where 3 samples are compared against pooled reference from 40 tumors
- Protein abundance reported as relative log2 ratio between sample versus reference

iTRAQ Processing: Quality Control issue

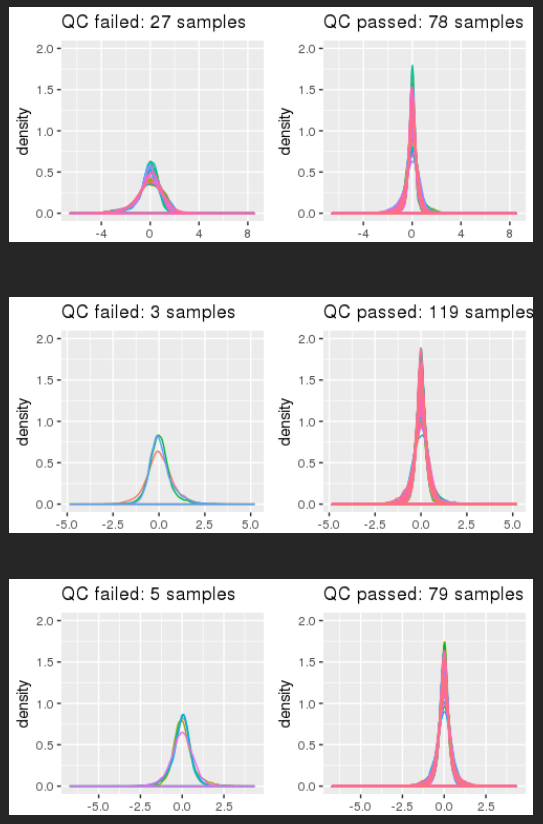
- Mertins et al. identified compromised samples with excessive low abundance proteins due to protein degradation



→ Bimodal or skewed protein abundance distribution

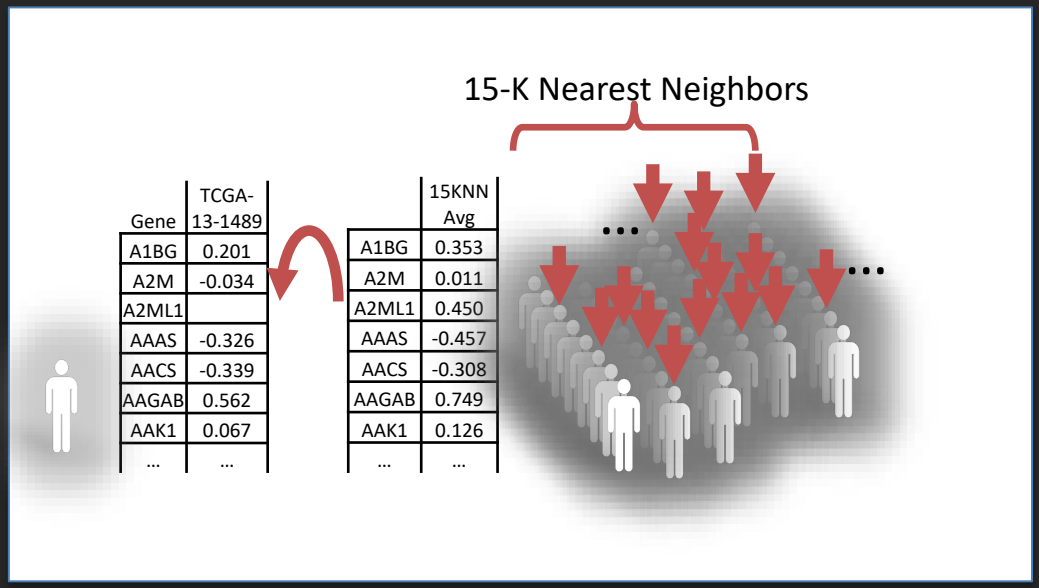
→ Std Dev is a natural QC statistic

- Two component Gaussian mixture model using Std Dev per sample to identify compromised samples



Mertins et al. Nature 2016

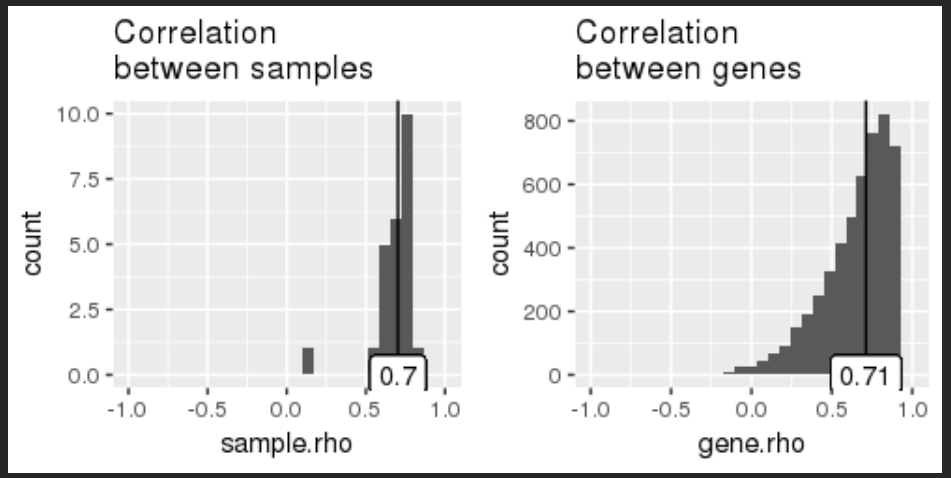
iTRAQ Processing: Missing Data



- iTRAQ measurement produces quality measures to remove low confidence measurements → missing values
- Filtering: samples and genes with >25% missing data were removed from analysis
- Imputation
  - To impute missing values each sample is matched to 15 patients with the most similar protein abundances across all measured genes (Euclidian Distance)
  - Use the average abundance among neighbors to impute missing value

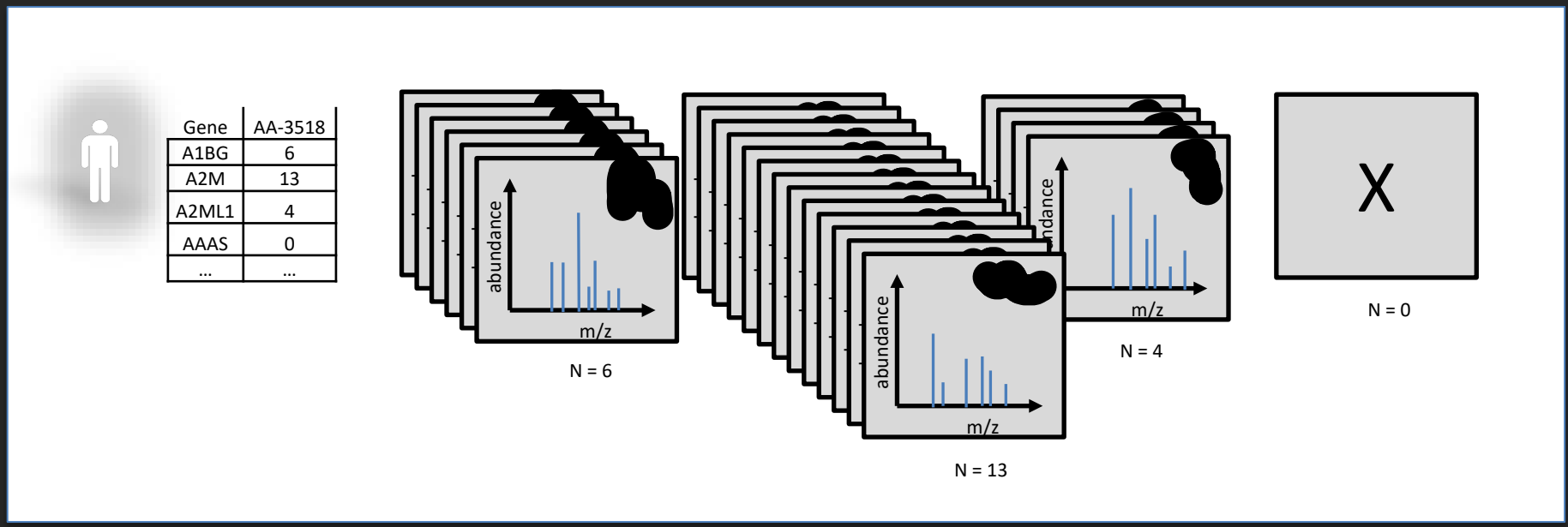
iTRAQ Processing: Merging Multi-Site Datasets

- Ovarian Cancer samples measured at two sites Johns Hopkins University and Pacific Northwest National Laboratory
- 32 samples measured at both sites show high correspondence :
  - median Spearman Rho of 0.70 for intra-tumor variation
  - median Spearman Rho of 0.71 for inter-tumor variation
- To unify datasets
  - Remove duplicate samples from JHU
  - Combine datasets and remove remaining bias using ComBat batch correction.



Johnson et al. Biostatistics 2007
https://www.ncbi.nlm.nih.gov/pubmed/16632515

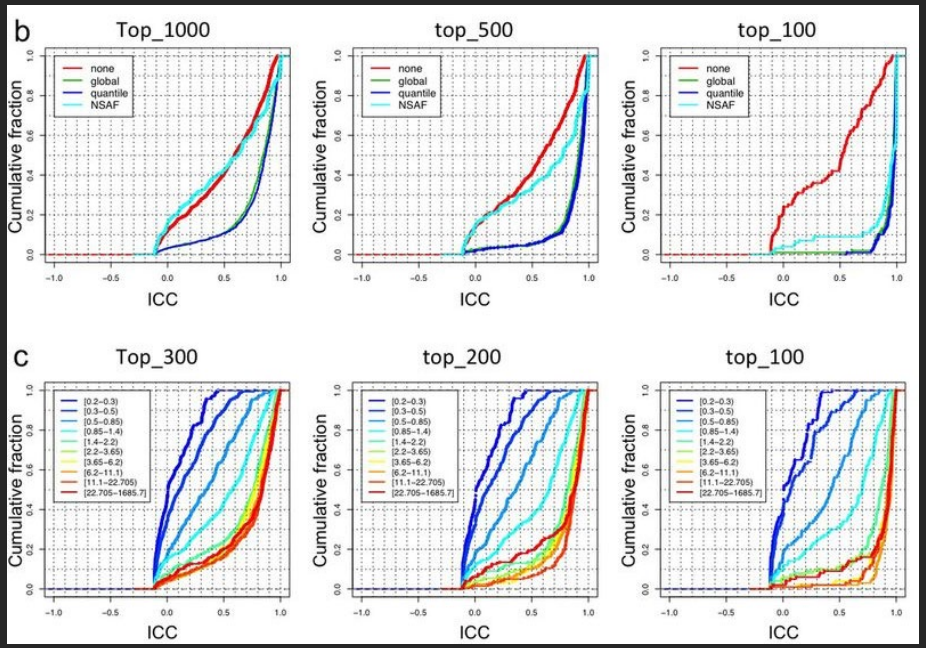
Label-free Data Format (COADREAD)



- Unlike iTRAQ, label-free quantification only provides absolute values per sample
- Protein abundance is quantified using the number of spectra measured, spectral counts

Label-free Processing: Zhang et al. Normalization

- To guide analysis Zhang et al. (COADREAD) quality control data set and examined groups of peptides which map to same protein and assessed quality using intraclass consensus (ICC)
- Avg value of 1.4 across samples set as min threshold for low abundance peptides
- Quantile normalization used to make distribution of protein abundance in each sample comparable
- Last values are log2 transformed



Zhang et al. Nature 2014

Processed Data Overview

- CDAP preprocessing & additional QC
  - Mapping mass ladders to peptides
  - Normalization
  - QC
  - Log2 transform
  - Missing value estimation
  - Batch correction
- Resulting data set can be processed similarly to other genomic data using similar statistical tools

Dataset	Number Samples	Number Proteins
BRCA – BROAD Institute	78	8662
OV – JHU & PNNL	150	5233
COADREAD – Vanderbilt University	95	2889

CTPAC data analysis

Linking with other omics data

Stanford MEDICINE

CPTAC multi-omics data fusion

- Same samples also have (phase 2 & 3)
  - RNA sequencing
  - DNA methylation
  - DNA copy number
  - Etc.
- Example from our work
  - Linking DNA methylation with Proteomic data

Magzoub et al. PLoS Comp Biol 2019, In Press

CPTAC data: overlapping samples & genes/proteins

	Nr Genes	Samples with mRNA & protein expression		
BRCA	2514	78		
COADREAD	2848	85		
OV	1896	168		

- Only focus on genes with both mRNA & protein expression

CPTAC data

	Nr Genes	Samples with mRNA & protein expression	Samples with methylation data	Sample with normal methylation data
BRCA	2514	78	972	123
COADREAD	2848	85	614	78
OV	1896	168	582	8

- All samples with DNA methylation data used for methylation states
- Varying # normal samples

ProteoMix: a statistical model

For each CpG site

Raw methylation data

Identify mixture components

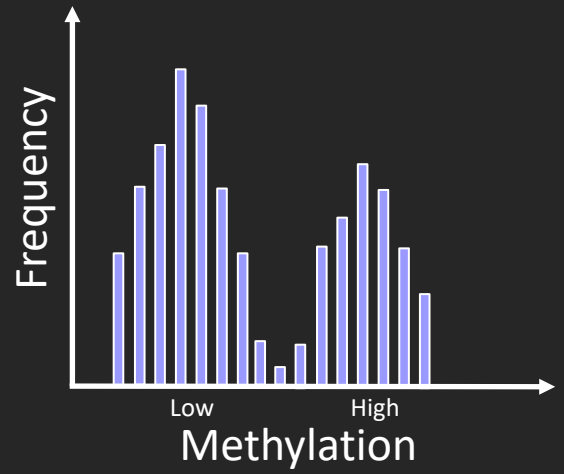
Identified methylation states

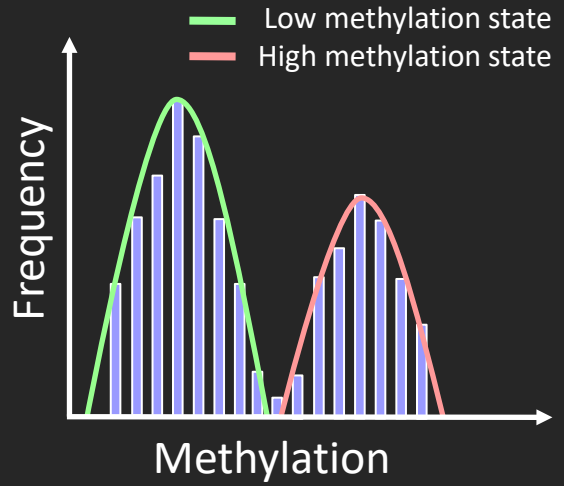
Map normal DNA methylation variation

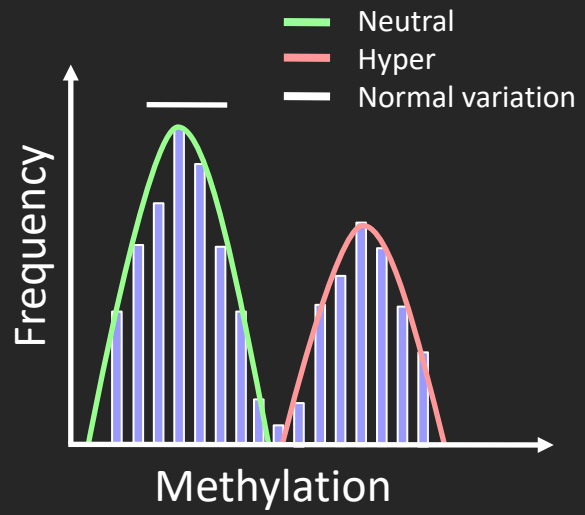
Differential methylation

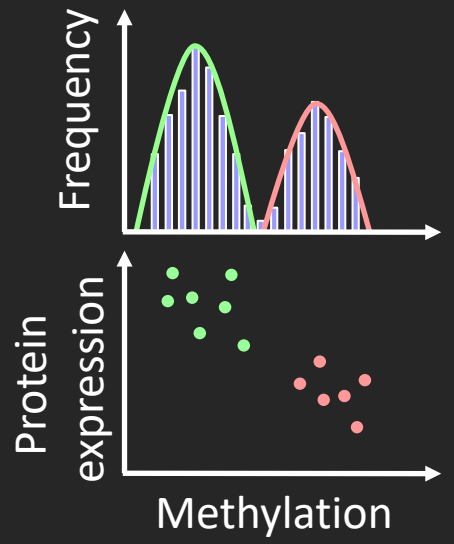
Incorporate protein expression data

Differential & Functional methylation states



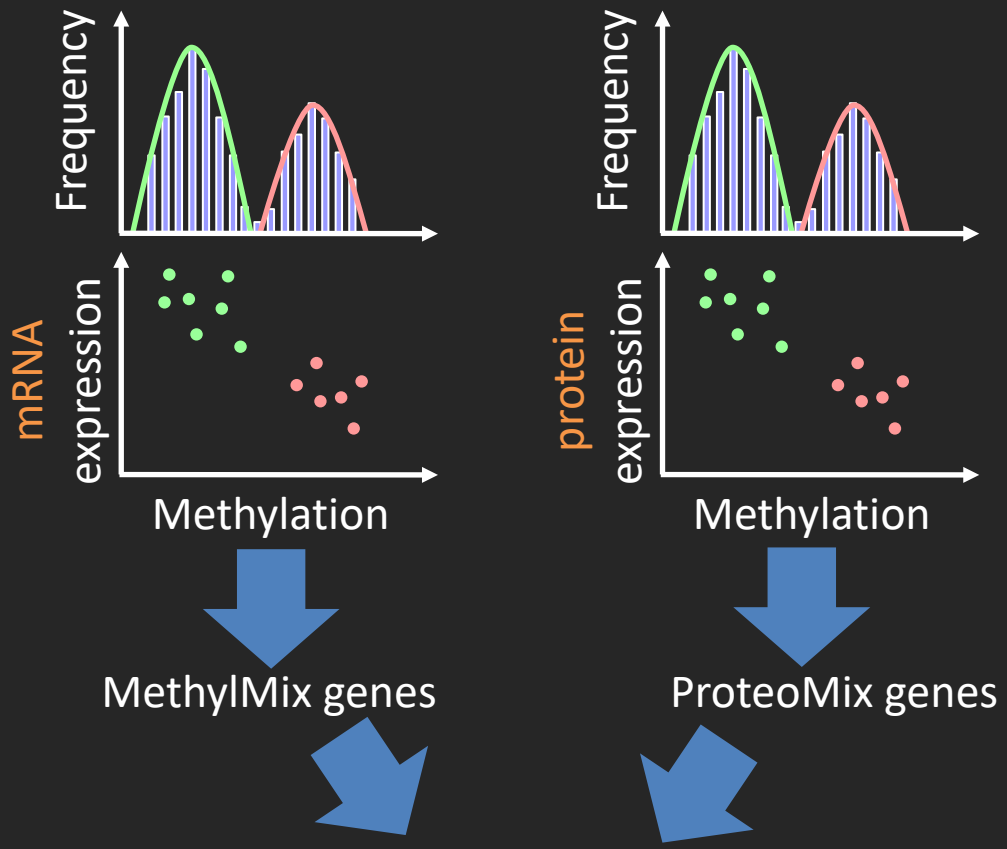






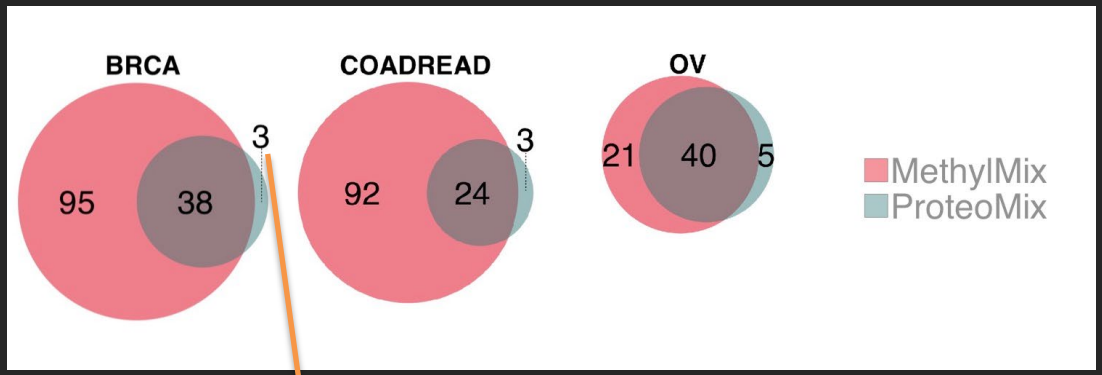
Magzoub et al. PLoS Comp Biol 2019, In Press

MethylMix vs. ProteoMix



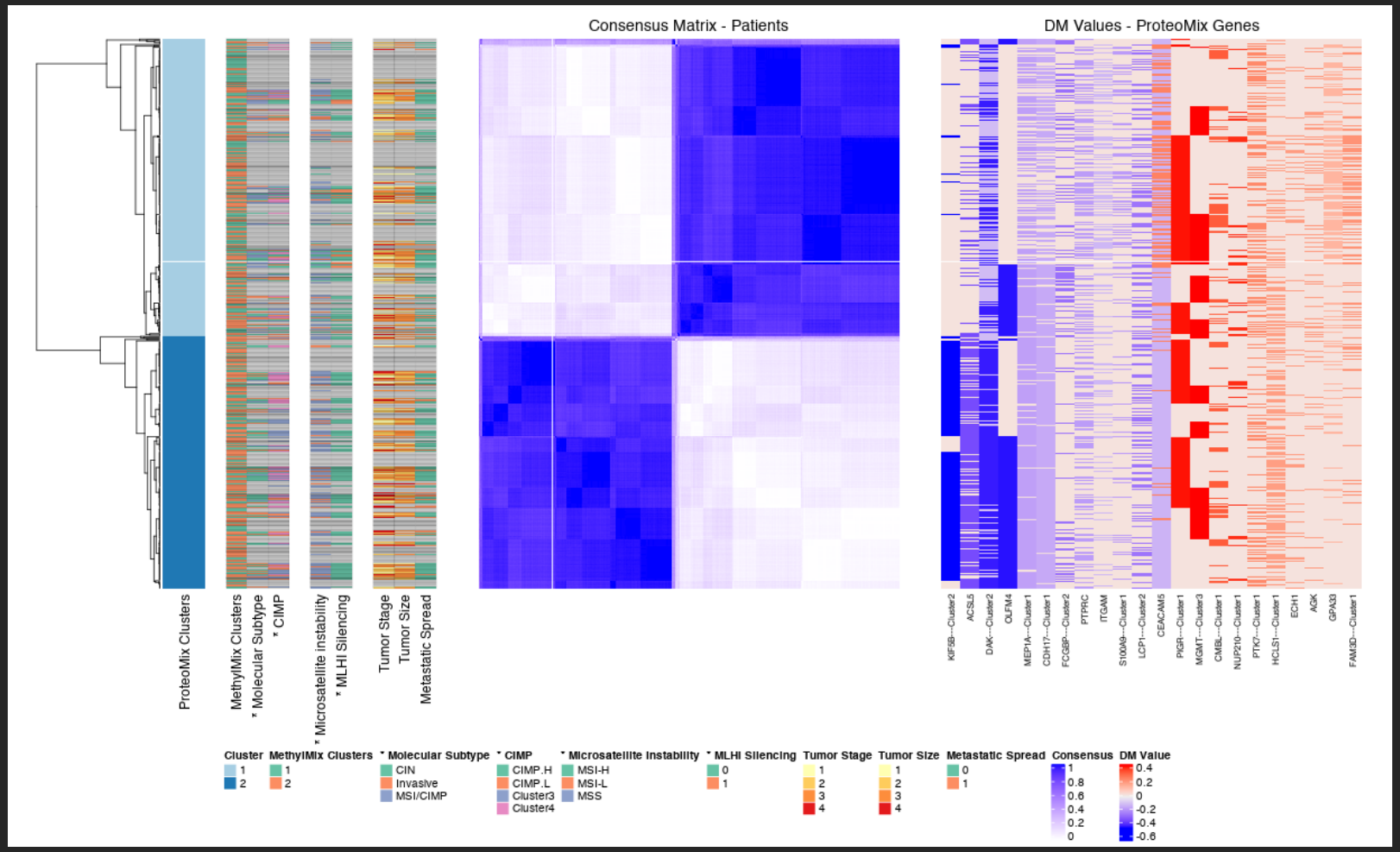
Overlap between MethylMix & ProteoMix
Gene set enrichment analysis
Overlap with cancer progression markers
Clustering

Results



- Breast cancer:
  - hypo-methylation in the UTR of EHF well-studied transcription factor involved in HER2 mediated epithelial differentiation
  - knockdown of EHF has been shown to inhibit tumor invasion and proliferation

Colorectal cancer



Significant association with CIMP vs non-CIMP

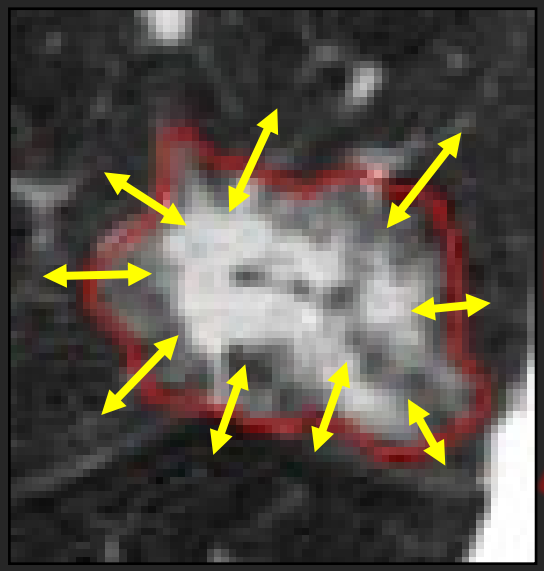
CPTAC data

Radioproteomic maps

Ongoing work

Stanford MEDICINE

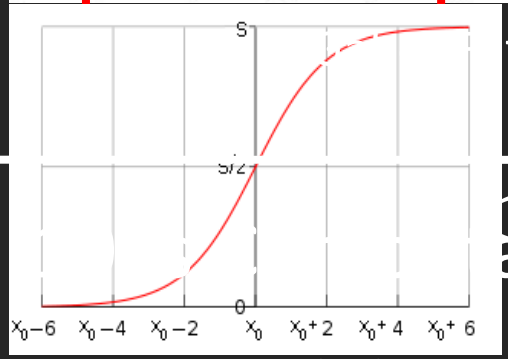
Radiomics/Quantitative imaging
Extraction of computational features



Texture Features

Shape Features

Edge Sharpness Features



Window (W)

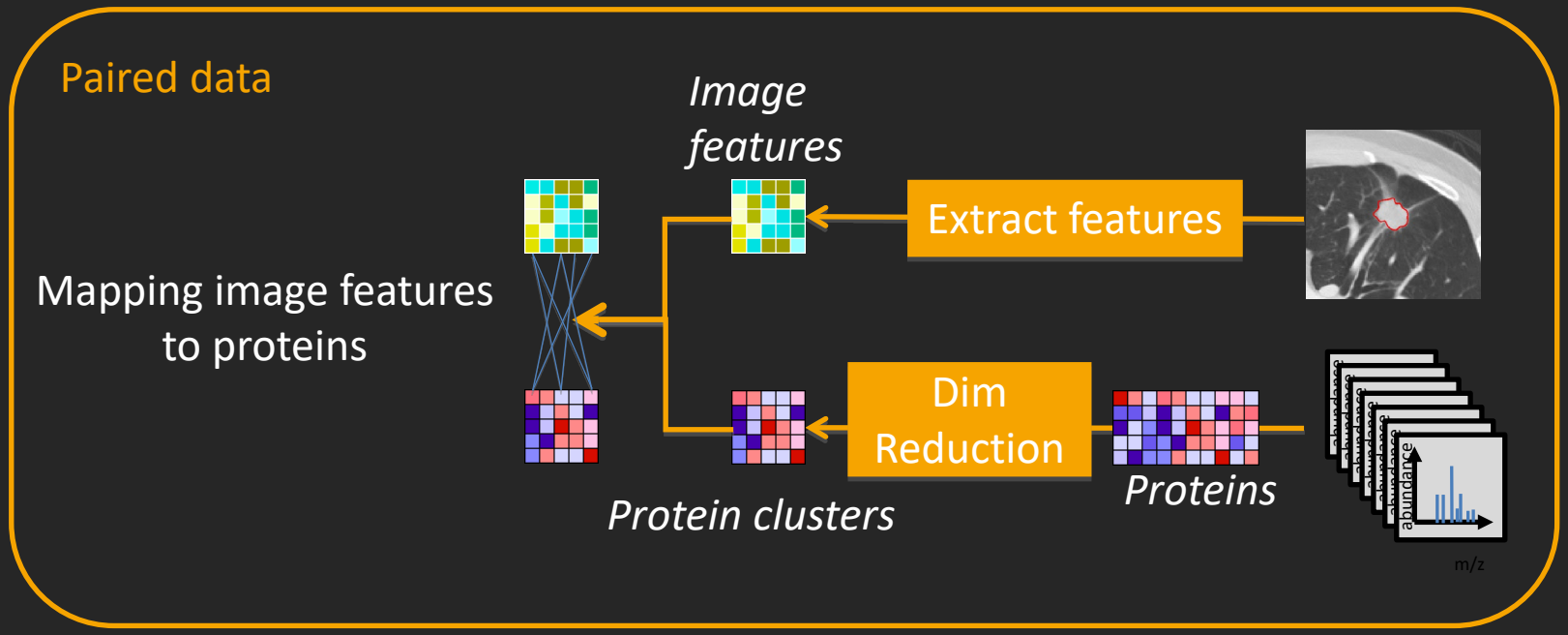
ter Bank

Scale (S

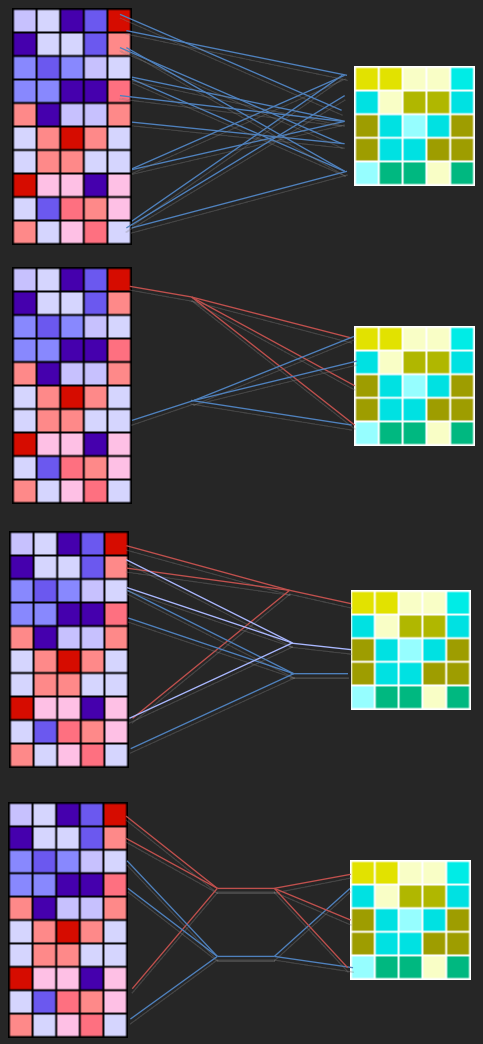
ational features

Sandy Napel

Radiogenomics mapping



Radioproteomic mapping methods



- Integration methods
  - Two-way univariate
  - Univariate – multivariate
    - Multivariate model of genes
    - Multivariate model of image features
  - Two-way multivariate

Applications

- Predict protein expression clusters from imaging
  - Non-invasive biomarkers
- Predict imaging phenotype from protein data
  - Study how pathways lead to imaging phenotypes
  - Annotate protein function
- Compare with and validate in ”traditional” radiogenomics maps built on RNA expression data

Acknowledgements

Gevaert Lab

Majed Magzoub

Kevin Brennan

Hong Zheng

Jay Shinde

Pritam Mukherjee

Lucas Patel

Radiology

Parag Mallick

Biomedical Data Science

Robert Tibshirani

Trevor Hastie

Funding

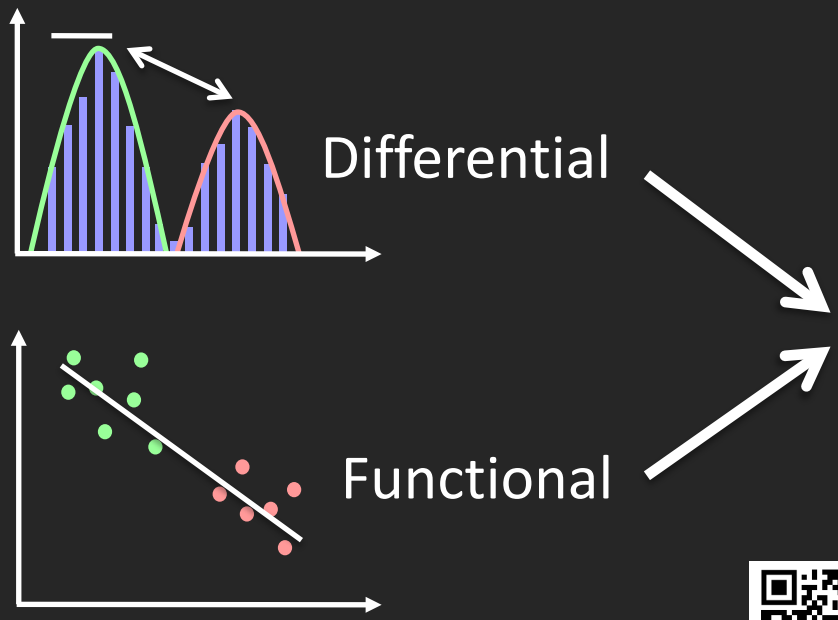
NIH/NIBIB R01 EB020527

NIH/NCI U01 DE025188

NIH/NCI R01 CA184968

NIH/NCI R01 CA176299

MethylMix & ProteoMix: identifying DNA methylation-driven genes in cancer



MethylMix & ProteoMix
R package

Available on Bioconductor & github:
https://github.com/gevaertlab

GenePattern module
in development

BioRxiv:

Stanford MEDICINE