Introduction to Proteogenomics

MBP Tech Talk
2019-12-29

Lydia Liu
lydia.liu@mail.utoronto.ca

Outline

Part 1:

- Why Proteogenomics
- What you Need for Proteogenomics
- Typically Proteogenomics Analyses

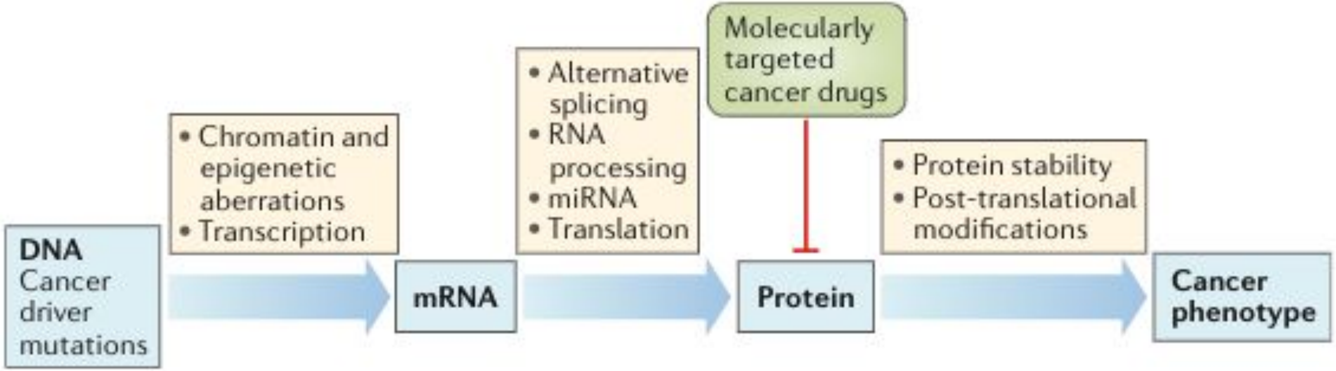
Part 2:

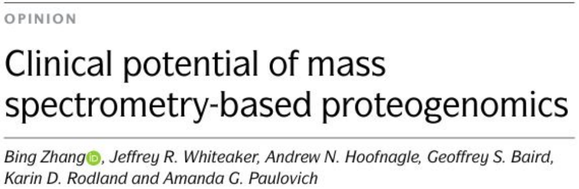
- Your questions
- What gets sweeped under the rug
- CPTAC resources

Why Proteogenomics?

Why Proteogenomics

- Mutational profiles is only one of the determinants of phenotype

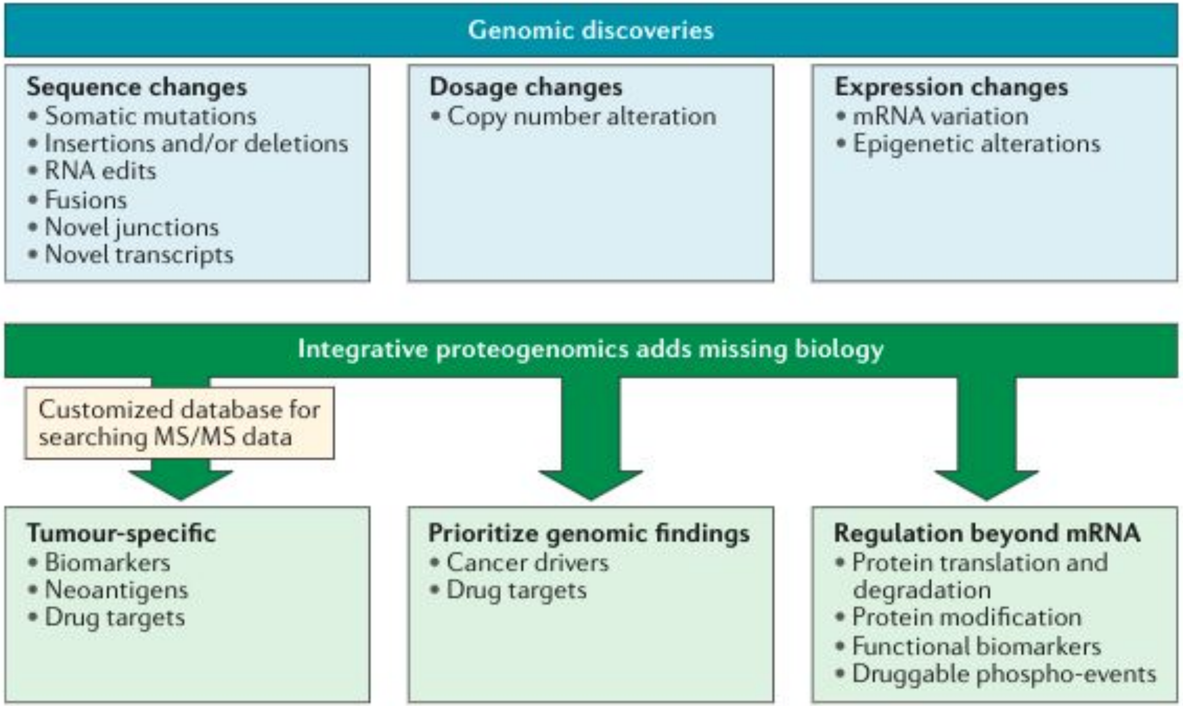



OPINION

Clinical potential of mass spectrometry-based proteogenomics

Bing Zhang, Jeffrey R. Whiteaker, Andrew N. Hoofnagle, Geoffrey S. Baird, Karin D. Rodland and Amanda G. Paulovich

Why Proteogenomics



Genomic discoveries

Sequence changes
- Somatic mutations
- Insertions and/or deletions
- RNA edits
- Fusions
- Novel junctions
- Novel transcripts

Dosage changes
- Copy number alteration

Expression changes
- mRNA variation
- Epigenetic alterations

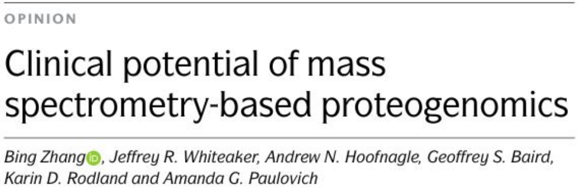
Integrative proteogenomics adds missing biology

Customized database for searching MS/MS data

Tumour-specific
- Biomarkers
- Neoantigens
- Drug targets

Prioritize genomic findings
- Cancer drivers
- Drug targets

Regulation beyond mRNA
- Protein translation and degradation
- Protein modification
- Functional biomarkers
- Druggable phospho-events

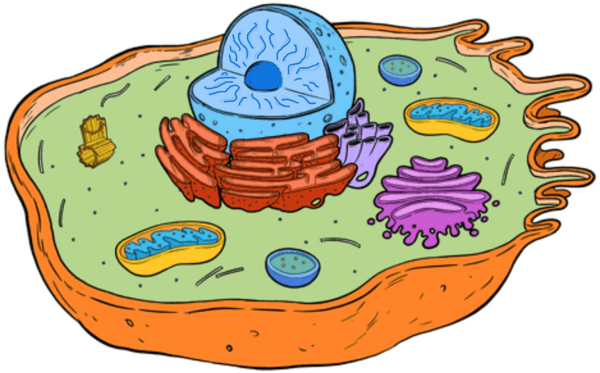


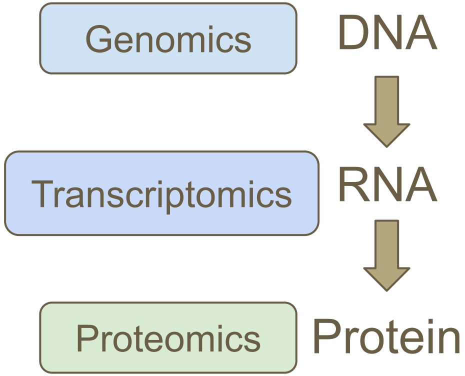
OPINION

Clinical potential of mass spectrometry-based proteogenomics

Bing Zhang, Jeffrey R. Whiteaker, Andrew N. Hoofnagle, Geoffrey S. Baird, Karin D. Rodland and Amanda G. Paulovich

Why Proteogenomics





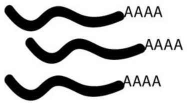
Genomics — DNA

Transcriptomics — RNA

Proteomics — Protein

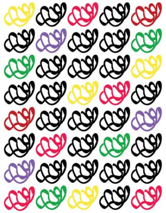
20,393 Genes

Whole-Genome Sequencing



104,763 Transcripts

RNA-sequencing



> 1,000,000
Protein Isoforms

Mass Spectrometry

Modified From: https://abeon-hosting.com/image-post/635-animal-cell-unlabeled-1.jpg.html
Sinha A. *et al.* Cancer Cell (2019)

What do you need to do proteogenomics?

What you need for proteogenomics

- Proteomics Data
- Genomics Data
- Transcriptomics Data
- Other Data
  - Clinical annotation
  - Metabolomics
  - Cytometry
  - Hi-C

- Patient sample
  - Tumour
  - Adjacent normal
  - Blood normal
- Cell line / Organoid
- Model organism
- PDX

Sinha A. et al., Cancer Cell (2019)

Proteomics Data

- Shotgun proteomics
- Phosphoproteomics
- Targeted proteomics

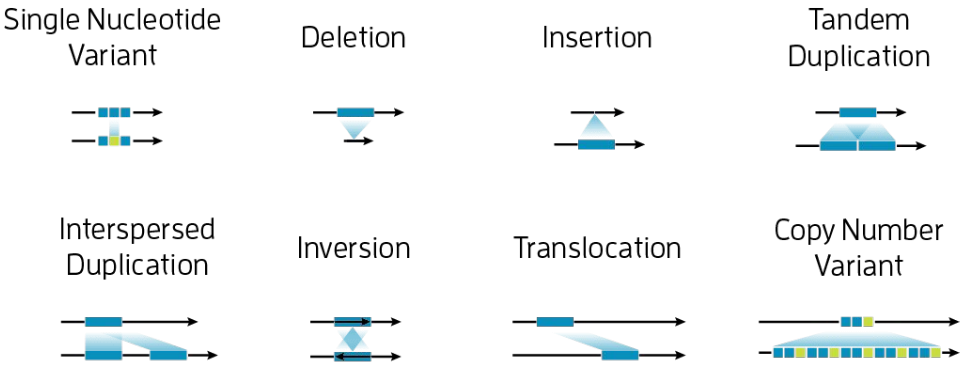
$$P = \begin{bmatrix} 880530 & 938230 & \dots & 2059600 \\ \vdots & \vdots & \ddots & \vdots \\ 1988200 & NA & \dots & 1226300 \\ \vdots & \vdots & \ddots & \vdots \\ NA & \dots & NA & 6716200 \end{bmatrix}$$

~6,000 × N

Genomics Data

- Targeted Sequencing
- Whole Exome Sequencing
- Whole Genome Sequencing

$$M = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \sim 20{,}000 \times \mathrm{N}$$



Types of Variants

- Somatic or Germline
- Coding / Noncoding
- Driver Analysis

- Chromothripsis
- Kataegis
- Variant allele frequency
- Telomere length
- Mitochondrial mutations

Transcriptomic Data

- RNA Microarray
- RNA-sequencing
- Single-cell RNA-sequencing

$$T = \begin{bmatrix} 237 & 3549 & \dots & 4583 \\ \vdots & \vdots & \ddots & \vdots \\ 1786 & 345 & \dots & 9 \\ \vdots & \vdots & \ddots & \vdots \\ 317 & \dots & 1247 & 7823 \end{bmatrix}$$

$\sim 20{,}000 \times \mathrm{N}$

- Somatic coding SNVs, Indels
- Assembled transcripts
- Fusion genes
- Circular RNAs

Other Data

- Clinical annotation
- MicroRNA
- Metabolomics
- Epigenomics
    - DNA Methylation
    - Histone Acetylation
- Cytometry
- Hi-C

What do proteogenomics studies do?

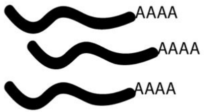
Omics Integration

Genomics

$$M = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

~20,000 × N

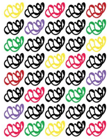
Transcriptomics



$$T = \begin{bmatrix} 237 & 3549 & \dots & 4583 \\ \vdots & \vdots & \ddots & \vdots \\ 1786 & 345 & \dots & 9 \\ \vdots & \vdots & \ddots & \vdots \\ 317 & \dots & 1247 & 7823 \end{bmatrix}$$

~20,000 × N

Proteomics



$$P = \begin{bmatrix} 880530 & 938230 & \dots & 2059600 \\ \vdots & \vdots & \ddots & \vdots \\ 1988200 & NA & \dots & 1226300 \\ \vdots & \vdots & \ddots & \vdots \\ NA & \dots & NA & 6716200 \end{bmatrix}$$

N = ~100s patients

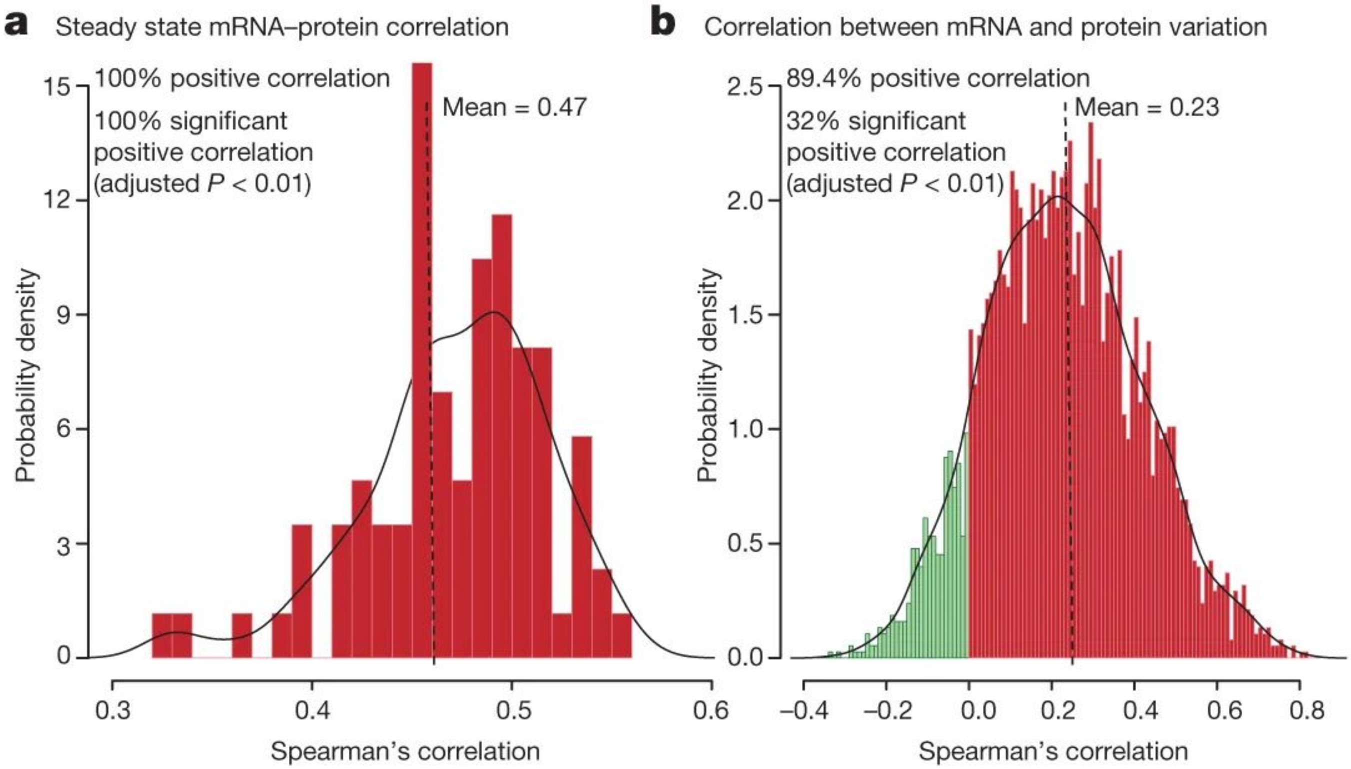
~7,000 × N

Transcriptome Proteome Correlation

- Within-sample correlation by gene
- Across-sample correlation by gene

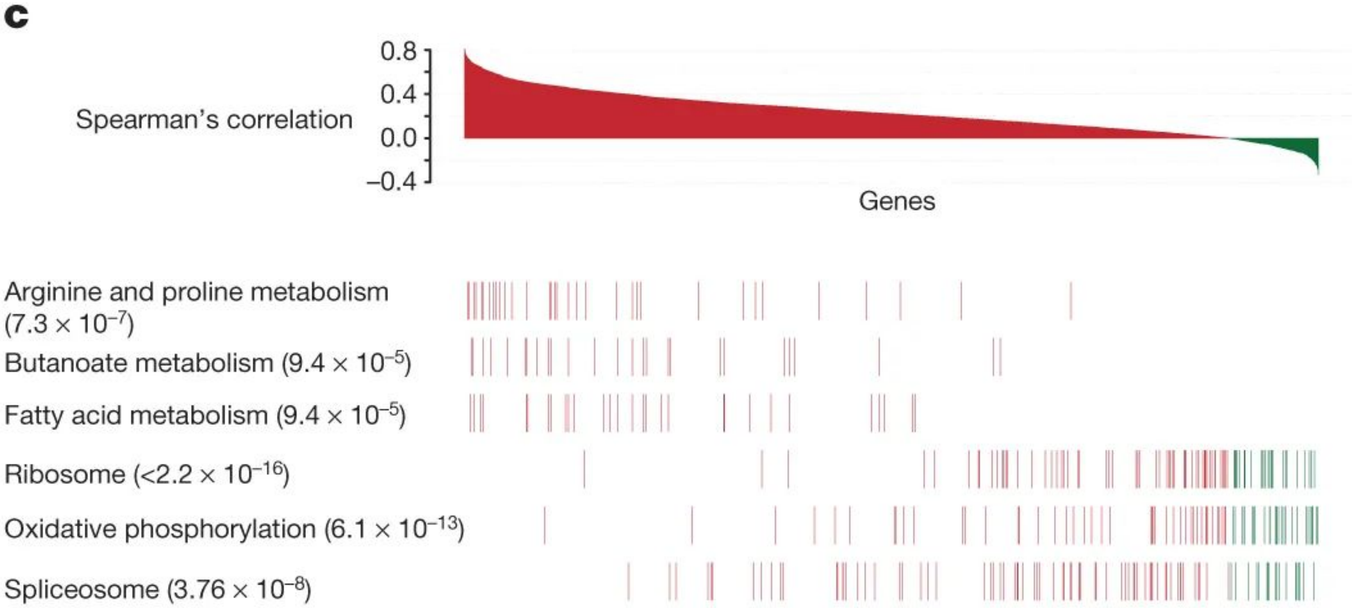
- Spearman correlation + FDR

Results from Transcriptome Proteome Correlation



Zhang, B. et al. Nature (2014)

Results from Transcriptome Proteome Correlation



Zhang, B. et al. Nature (2014)

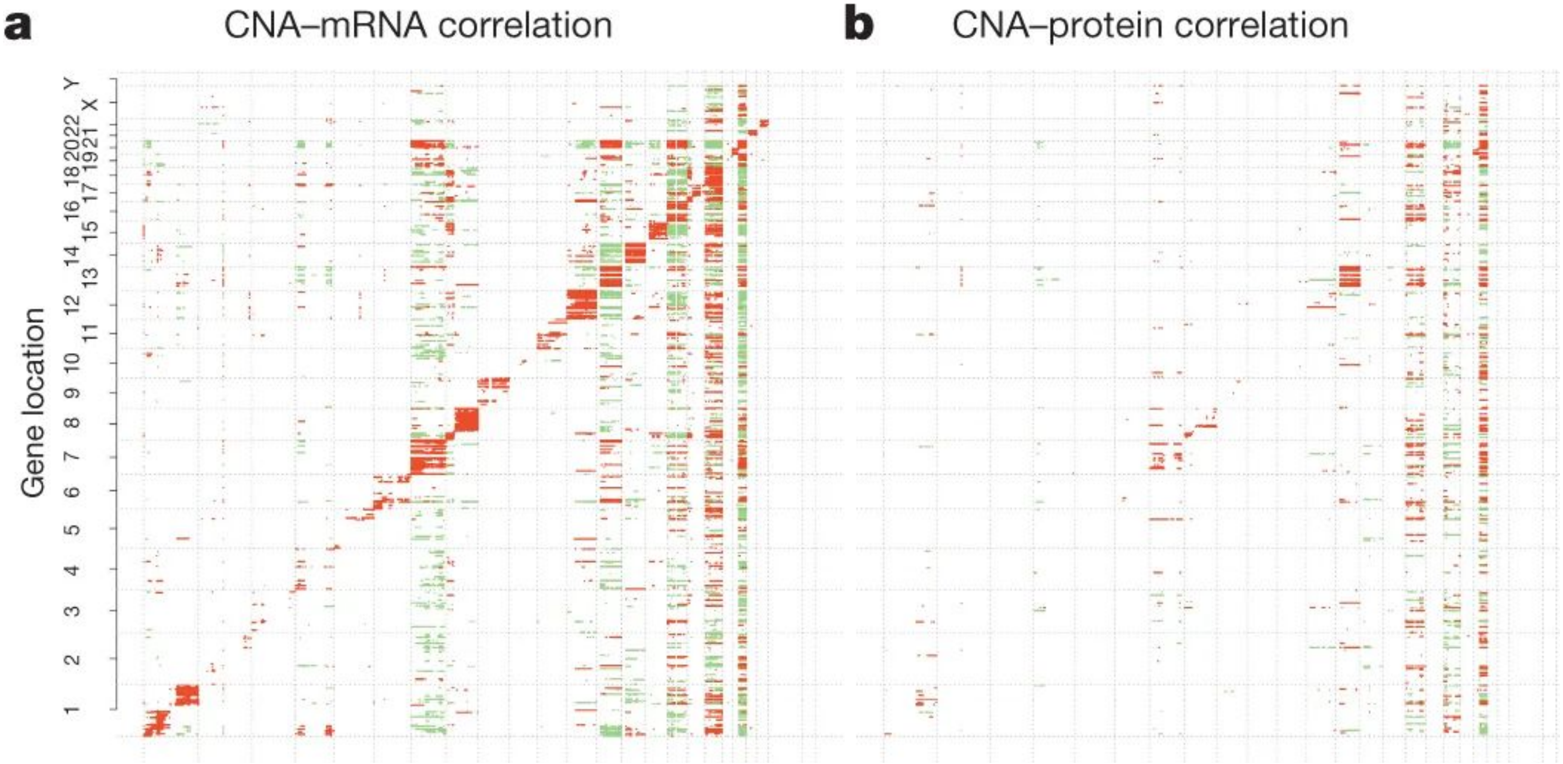
Copy Number Cis Trans Effects

- Correlate copy number changes with mRNA and protein abundance
- Genes directly affected by the CNA

OR

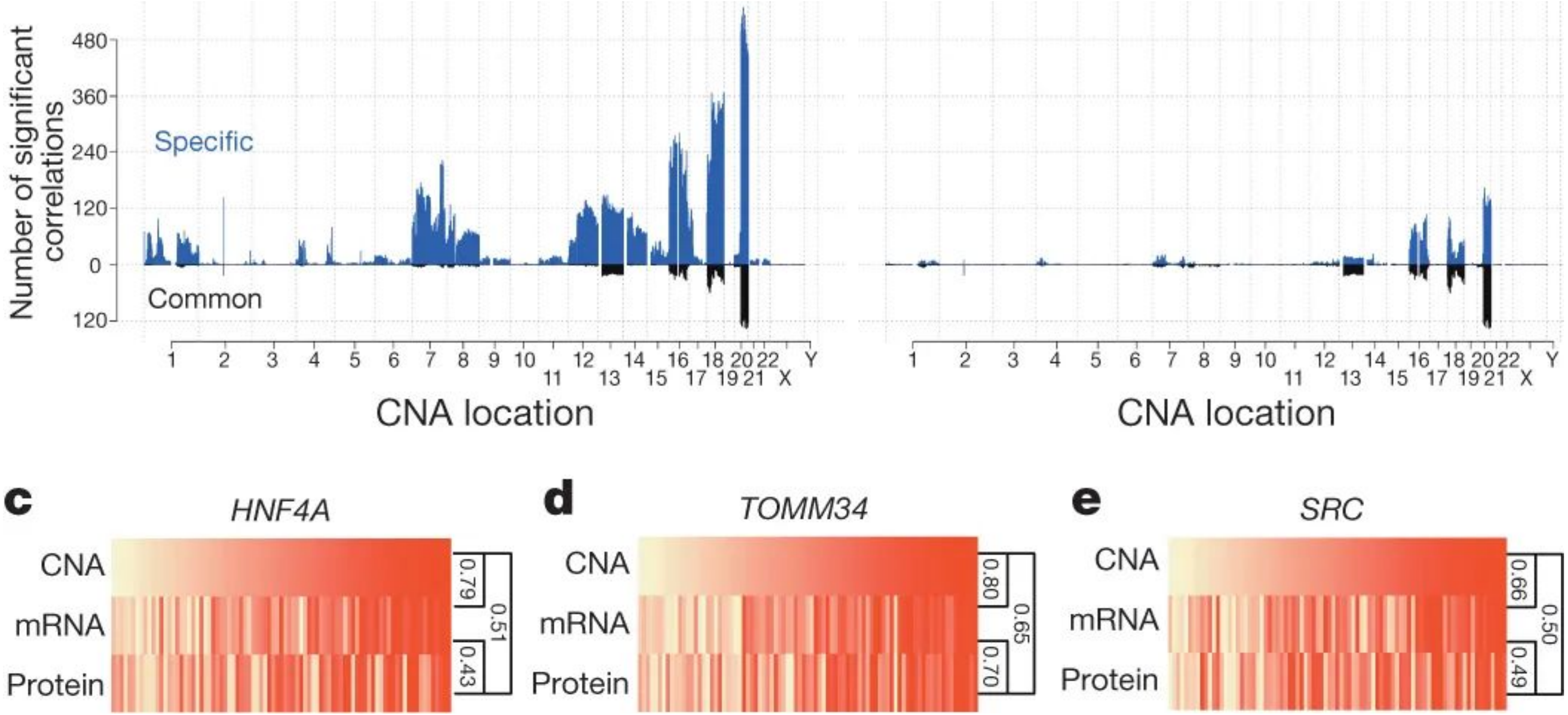
- Genes indirectly affected by the CNA

Results from Copy Number Cis Trans Effects



Zhang, B. et al. Nature (2014)

Results from Copy Number Cis Trans Effects

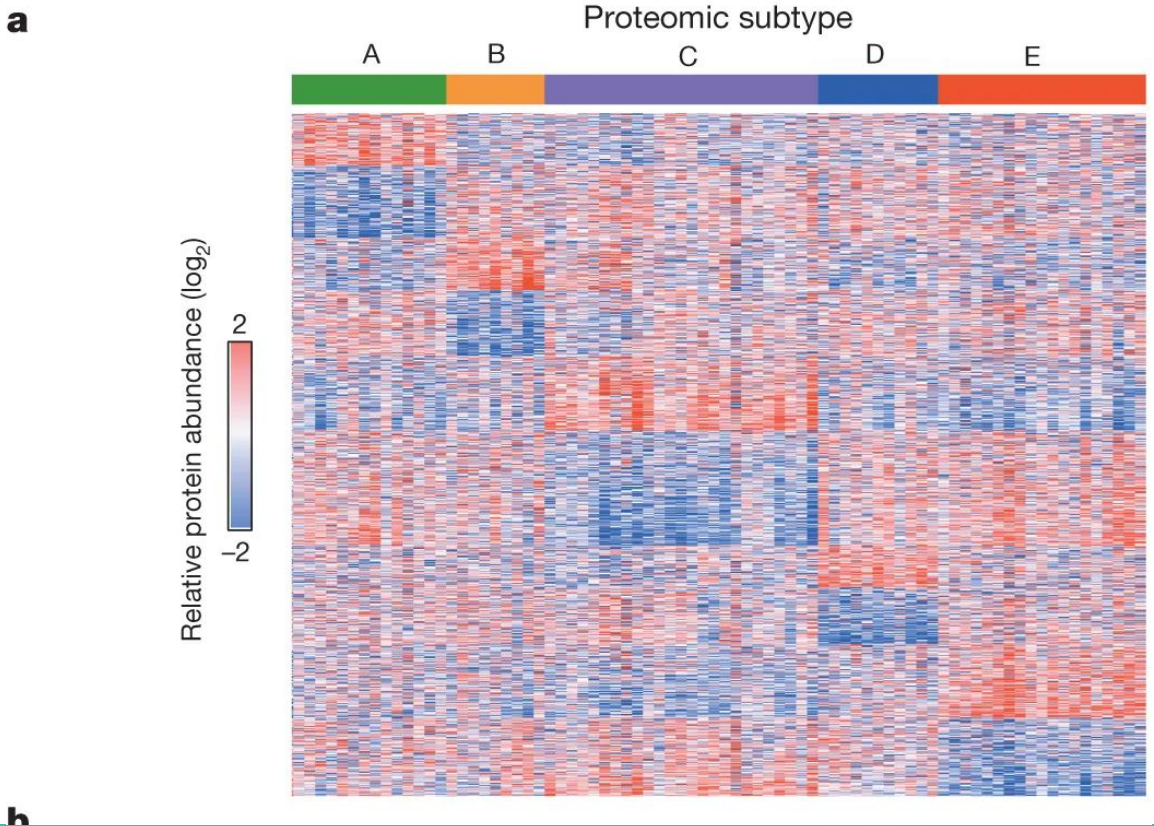


Zhang, B. et al. Nature (2014)

Proteogenomics patient subtyping

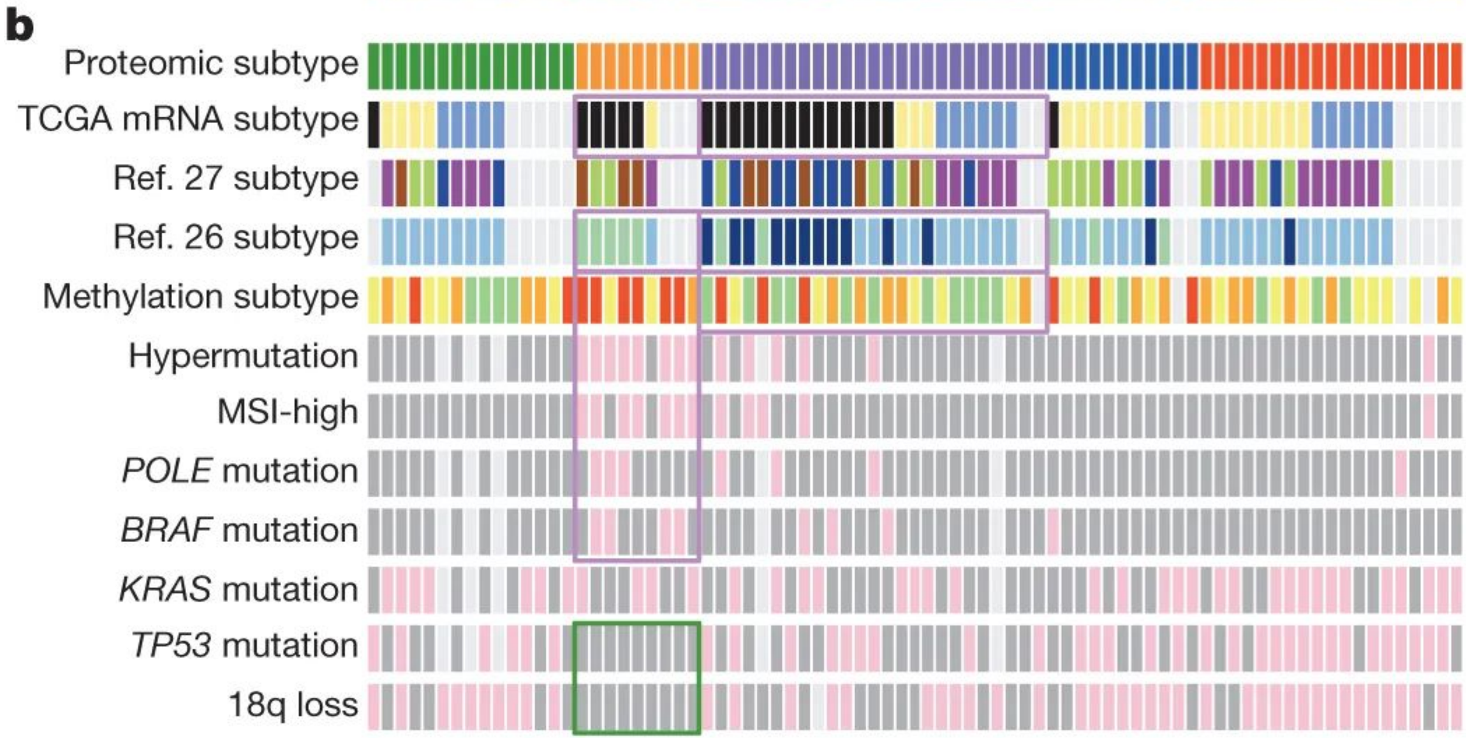
- Cluster patients based on proteomics profiles
- Compare to established genomic / transcriptomic based clusters

Results from Proteogenomics patient subtyping



Zhang, B. et al. Nature (2014)

Results from Proteogenomics patient subtyping

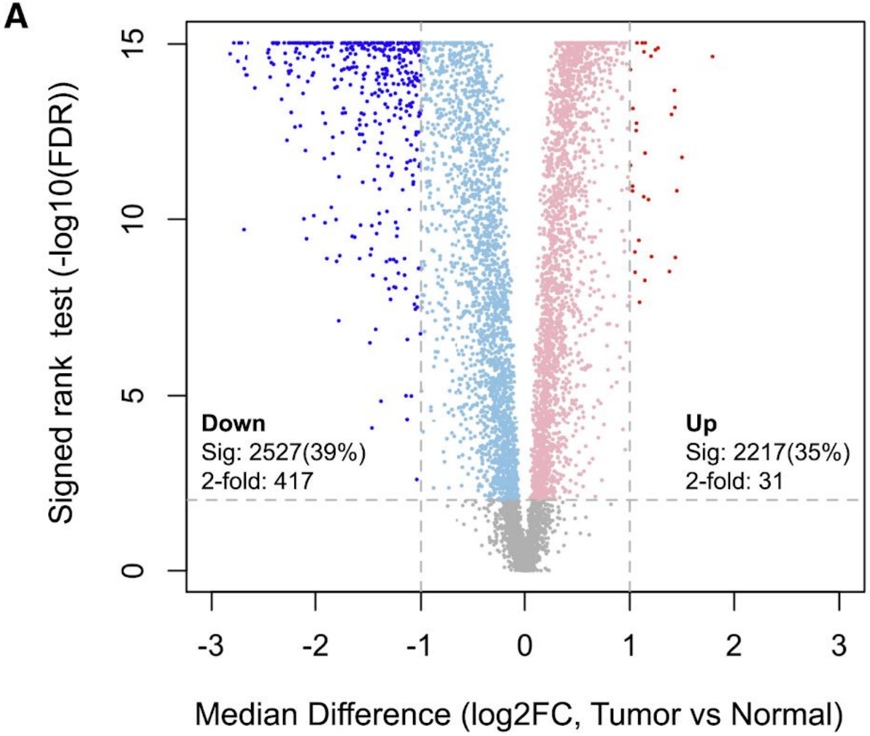


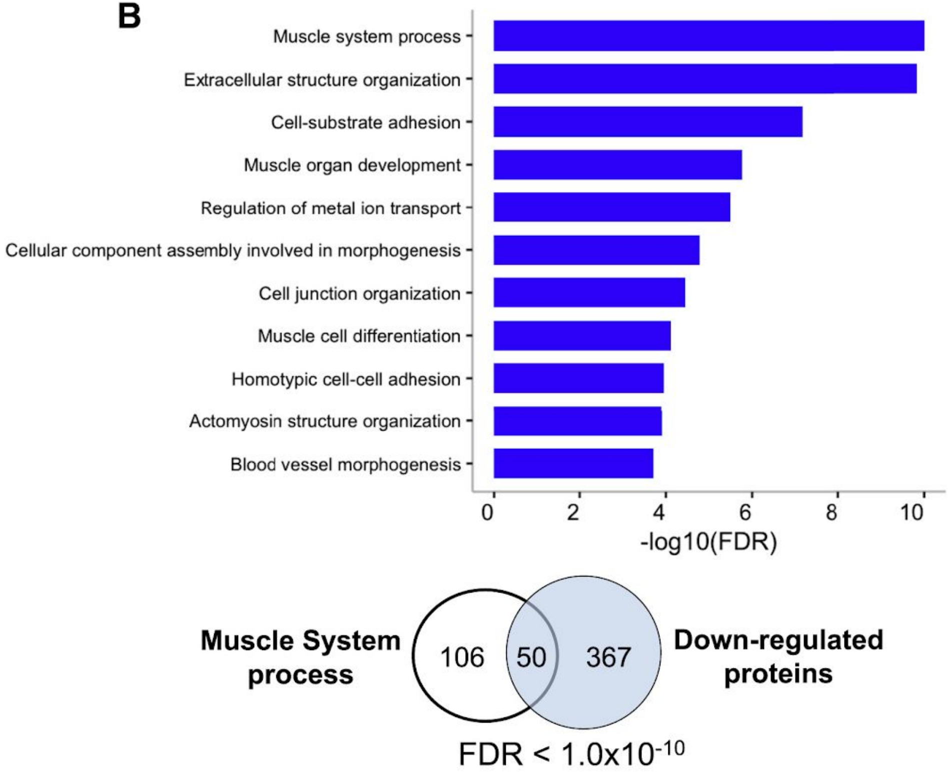
Zhang, B. et al. Nature (2014)

Cancer Associated Expression Changes

- Differential expression analysis of mRNA and protein abundance
- Between tumour tissue and adjacent normal tissue

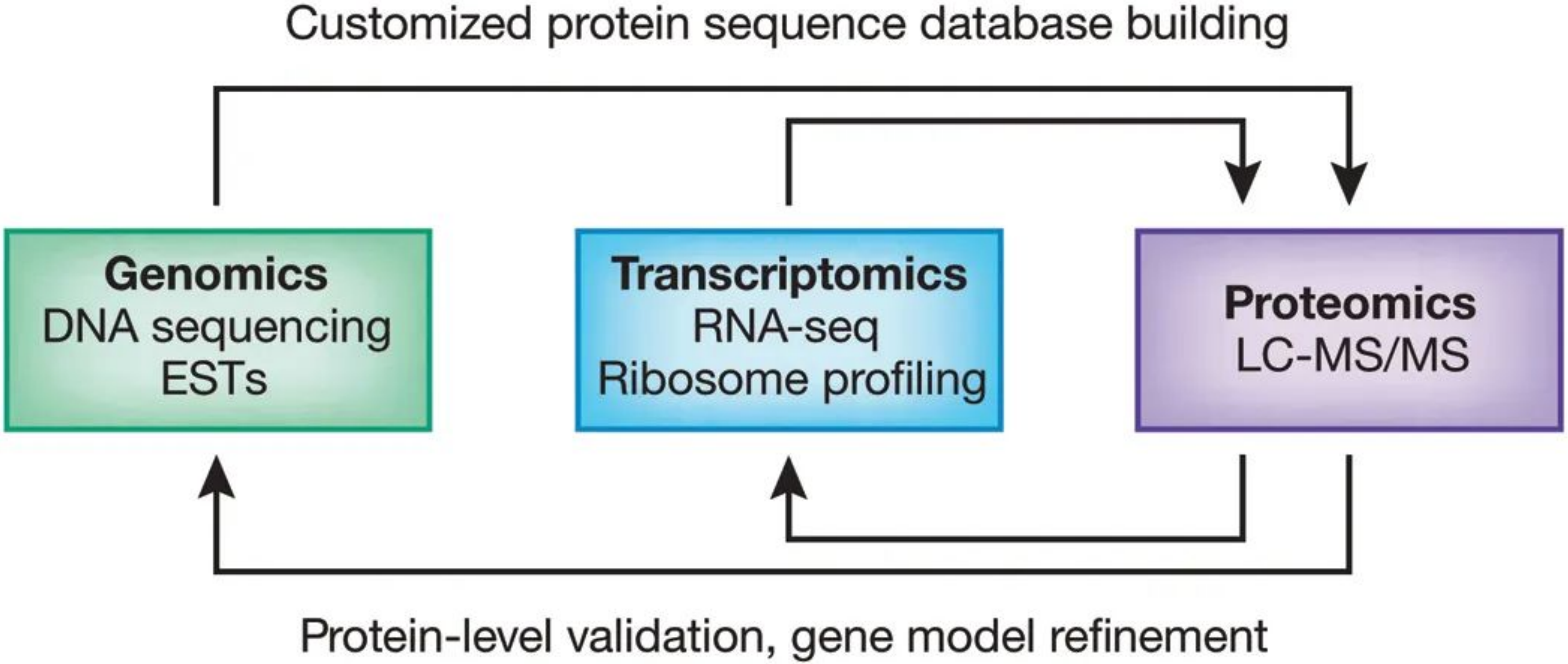
Results from Cancer Associated Changes





Vasaikar, S. et al., Cell (2019)

Custom Database Construction

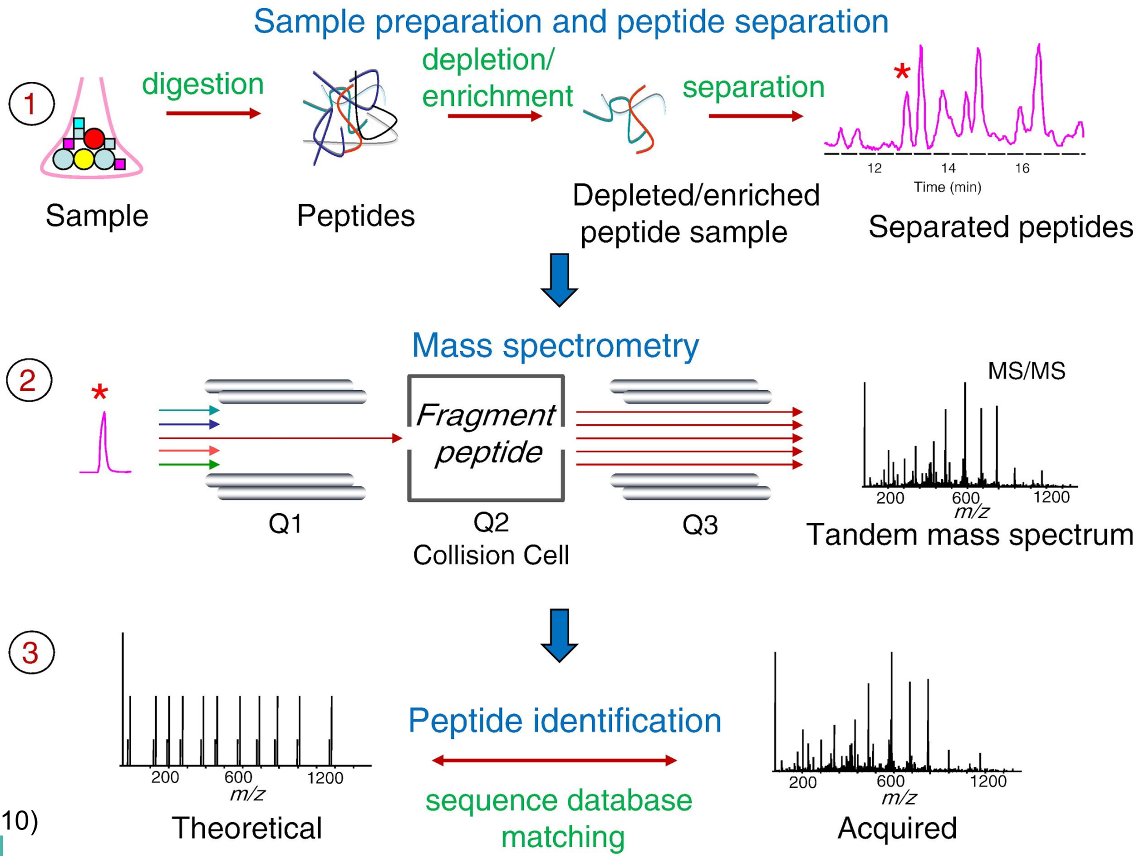


Customized protein sequence database building

Protein-level validation, gene model refinement

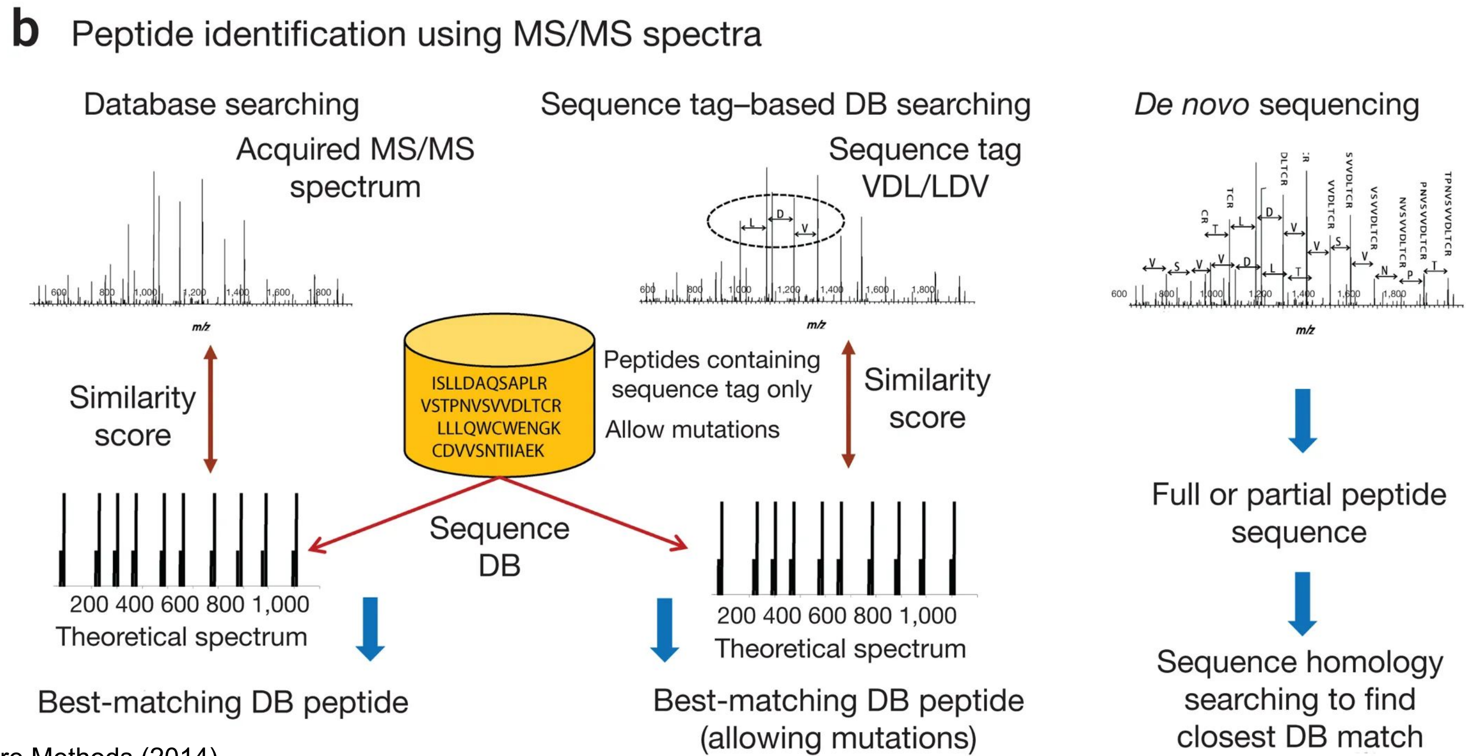
Nesvizhskii, Nature Methods (2014)

Why Custom Database



Nesvizhskii, Journal of Proteomics (2010)

Why Custom Database



b Peptide identification using MS/MS spectra

Database searching

Acquired MS/MS spectrum

Similarity score

Theoretical spectrum

Best-matching DB peptide

Sequence DB

ISLLDAQSAPLR
VSTPNVSVVDLTCR
LLLQWCWENGK
CDVVSNTIIAEK

Sequence tag–based DB searching

Sequence tag VDL/LDV

Peptides containing sequence tag only

Allow mutations

Similarity score

Theoretical spectrum

Best-matching DB peptide (allowing mutations)

De novo sequencing

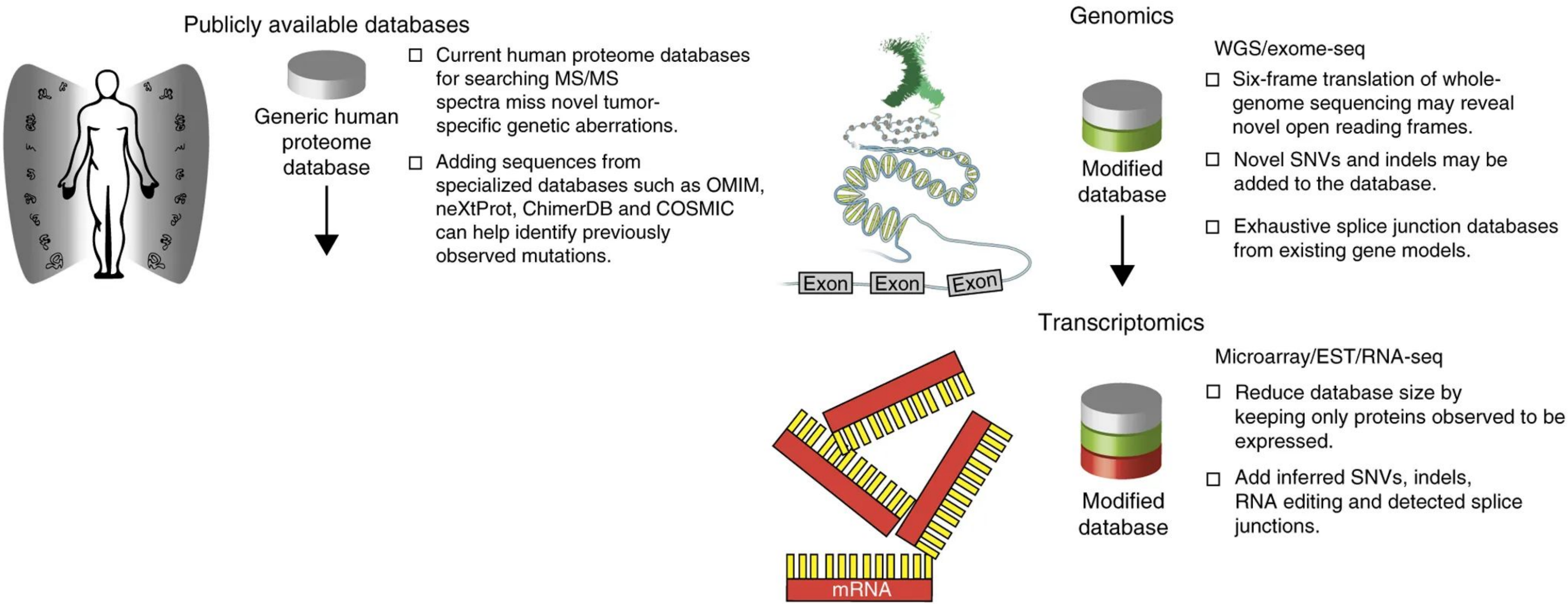
Full or partial peptide sequence

Sequence homology searching to find closest DB match

600 800 1,000 1,200 1,400 1,600 1,800
m/z
200 400 600 800 1,000
L D V
CR TCR DLTCR VVDLTCR SVVDLTCR VSVVDLTCR NVSVVDLTCR PNVSVVDLTCR TPNVSVVDLTCR
V S V V D L T V S V N P T

Nesvizhskii, Nature Methods (2014)

Why Custom Database



Publicly available databases

Generic human proteome database

- □ Current human proteome databases for searching MS/MS spectra miss novel tumor-specific genetic aberrations.
- □ Adding sequences from specialized databases such as OMIM, neXtProt, ChimerDB and COSMIC can help identify previously observed mutations.

Genomics

Modified database

WGS/exome-seq

- □ Six-frame translation of whole-genome sequencing may reveal novel open reading frames.
- □ Novel SNVs and indels may be added to the database.
- □ Exhaustive splice junction databases from existing gene models.

Transcriptomics

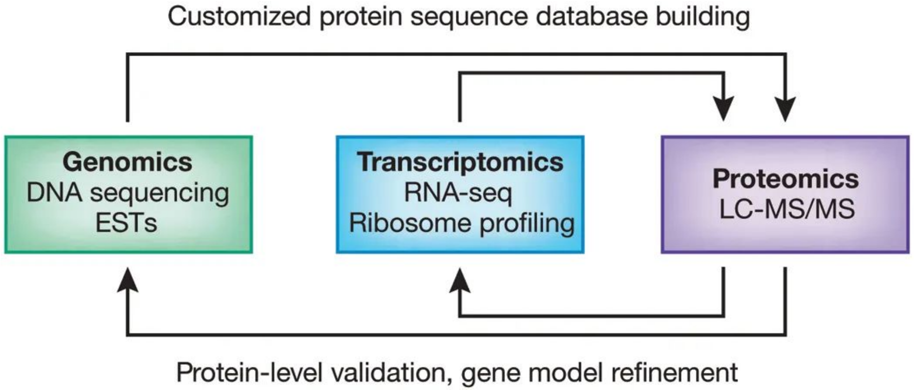
Modified database

Microarray/EST/RNA-seq

- □ Reduce database size by keeping only proteins observed to be expressed.
- □ Add inferred SNVs, indels, RNA editing and detected splice junctions.

Alfaro et al., Nature Methods (2014)

Custom Database Construction


Customized protein sequence database building
Genomics
DNA sequencing
ESTs
Transcriptomics
RNA-seq
Ribosome profiling
Proteomics
LC-MS/MS
Protein-level validation, gene model refinement

- Somatic SNV
- Germline SNV

- Indels
- Splice variants

- SNVs
- Indels

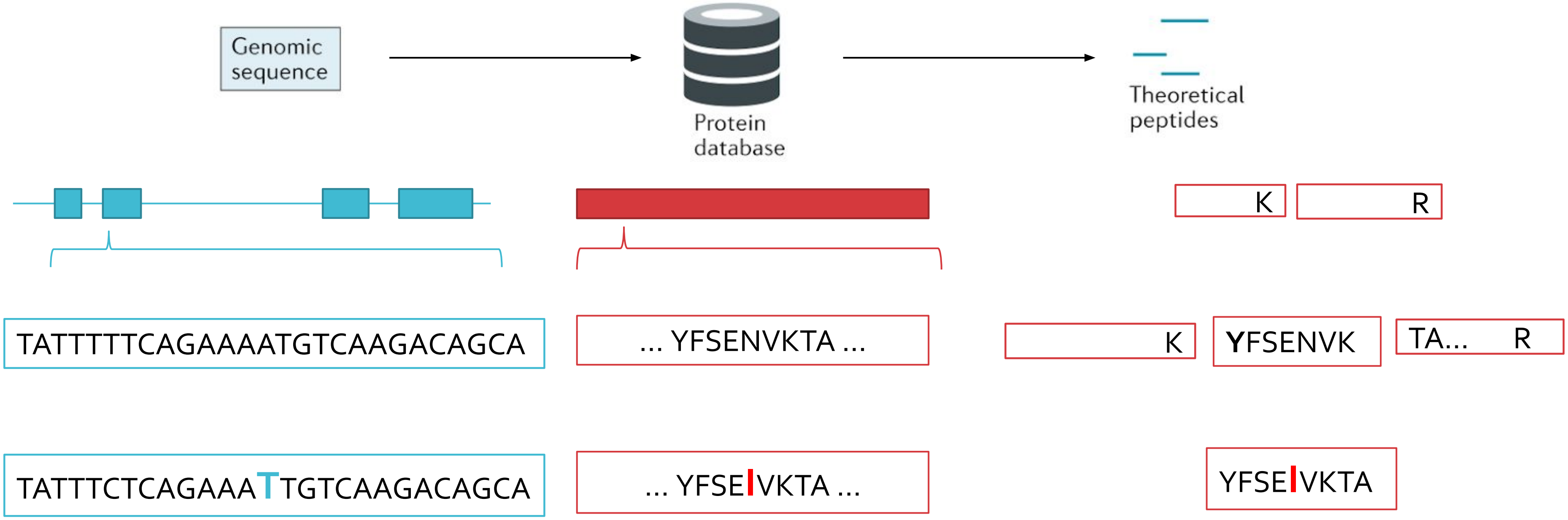
- Alternative transcripts
- Noncoding transcripts
-

Nesvizhskii, Nature Methods (2014)

Tools for Custom Database Construction

- customProDB
- QUILTS

Mutant Peptide Database Creation



Genomic sequence

Protein database

Theoretical peptides

K R

TATTTTTCAGAAAATGTCAAGACAGCA

... YFSENVKTA ...

K YFSENVK TA... R

TATTTCTCAGAAATTGTCAAGACAGCA

... YFSEIVKTA ...

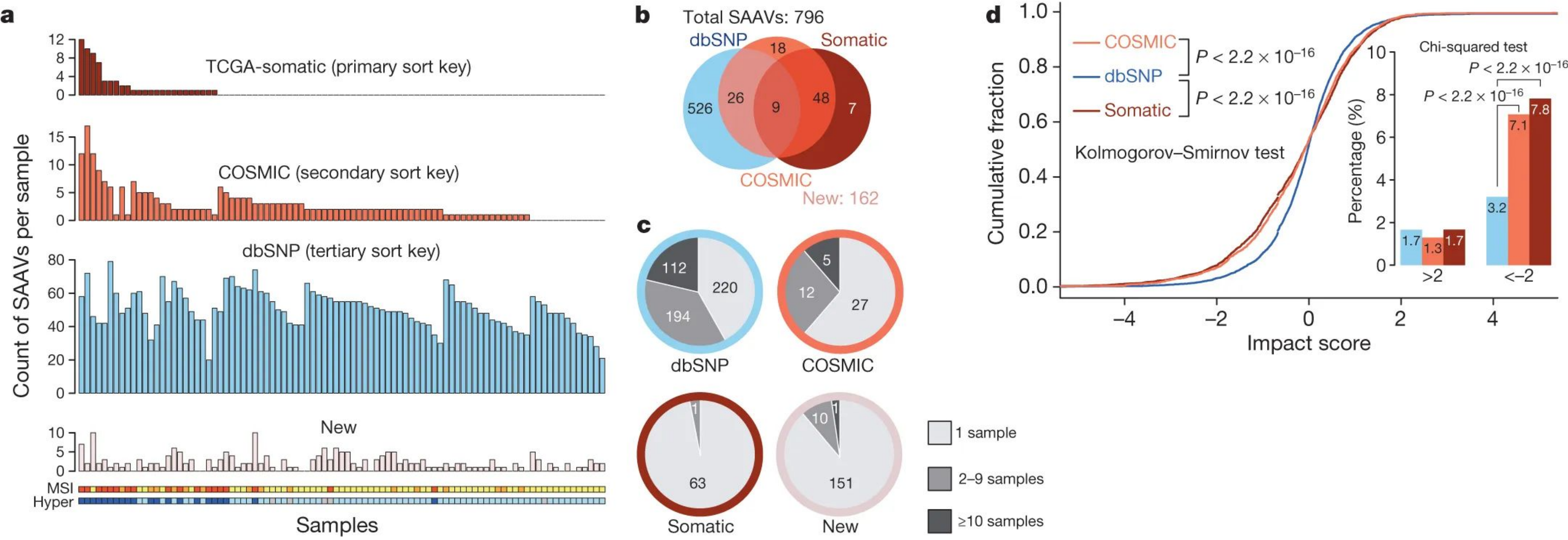
YFSEIVKTA

Missense DNA Mutation

Altered Amino Acid Sequence

Mutant Peptide Sequence

Results from SNV Search



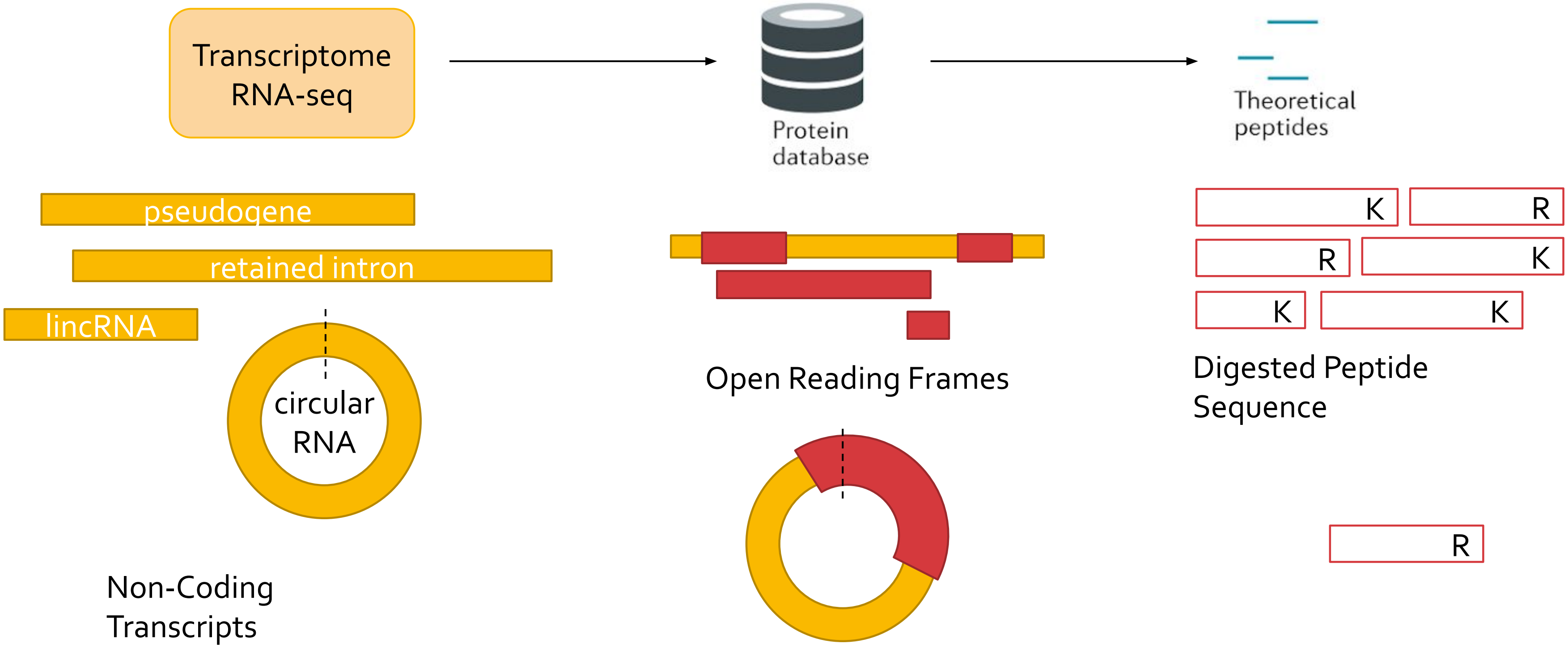
Zhang, B. et al. Nature (2014)

SNV Impact

$\text{Impact} = (\text{Exp} - \text{Median}_{\text{non-mutant}}) / \text{MAD}_{\text{non-mutant}}$

Zhang, B. et al. Nature (2014)

Novel Peptide Database Creation



Transcriptome RNA-seq

Protein database

Theoretical peptides

pseudogene

retained intron

lincRNA

circular RNA

K R R K K K

R

Non-Coding Transcripts

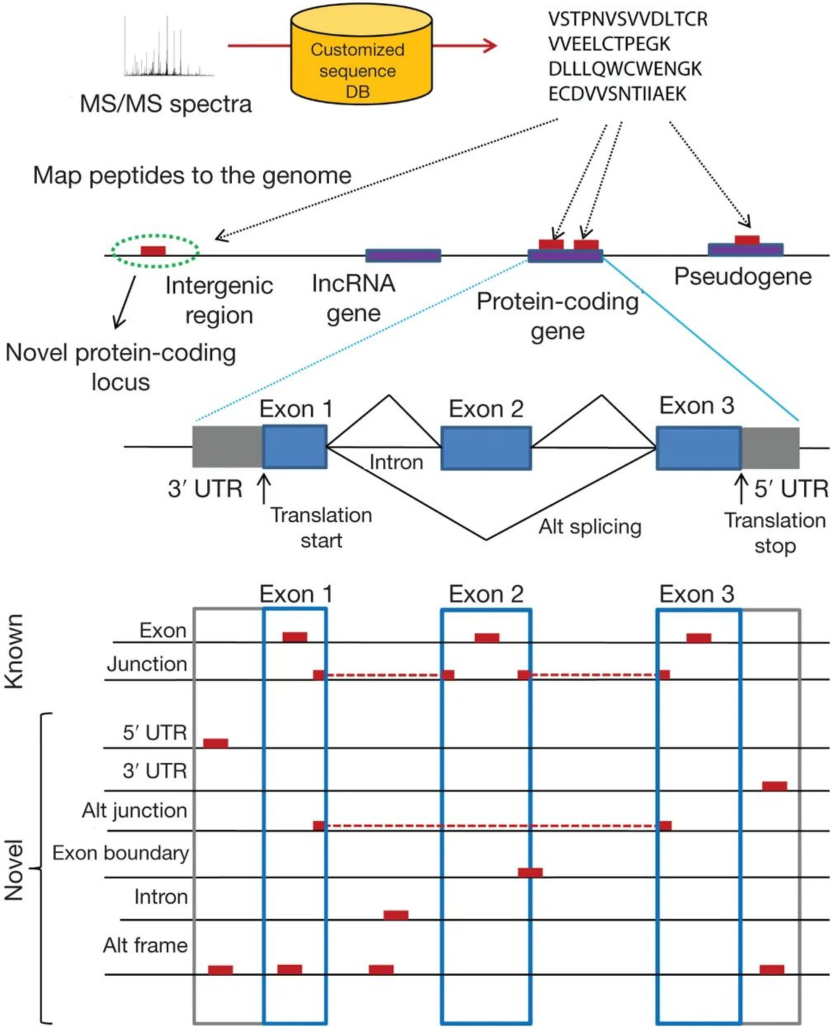
Open Reading Frames

Digested Peptide Sequence

Junction Spanning ORFs

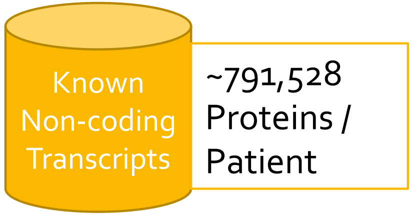
Junction Spanning Peptide Sequence

Type of peptides



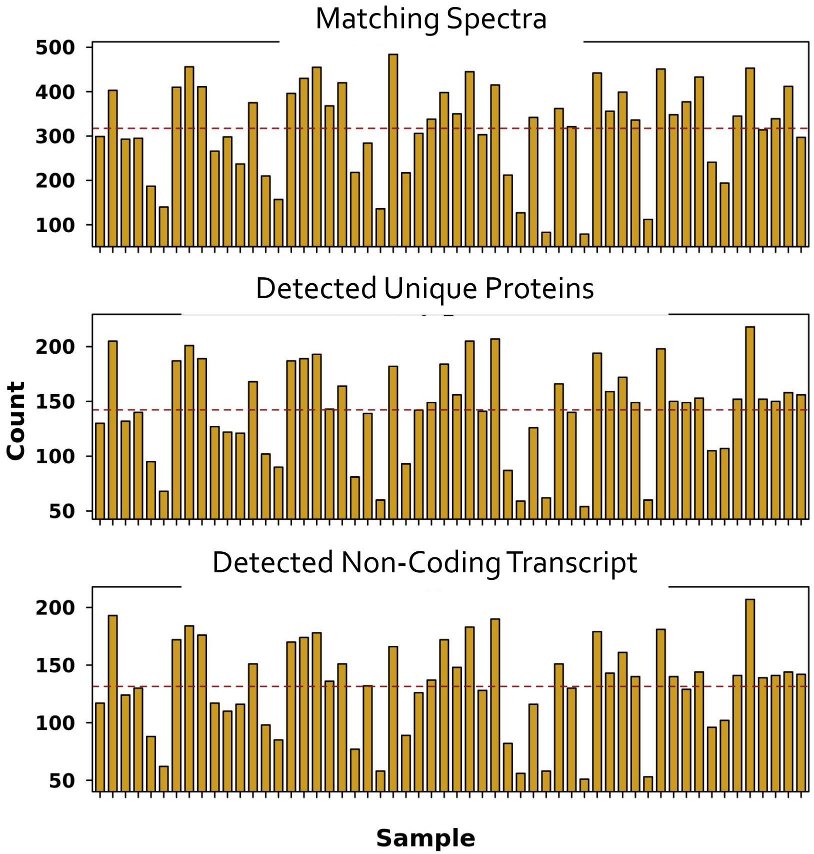
Nesvizhskii, Nature Methods (2014)

Personal Novel Peptides Search Results



Known Non-coding Transcripts

~791,528 Proteins / Patient

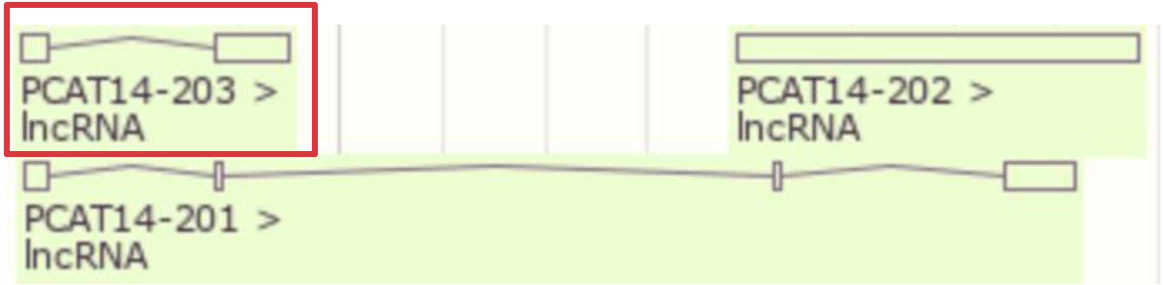


~317 spectra

~142 non-coding transcript derived proteins

~132 non-coding transcripts

Prostate Cancer Associated Transcript-14



Transcript levels of

- PCAT14-202
- PCAT14-203

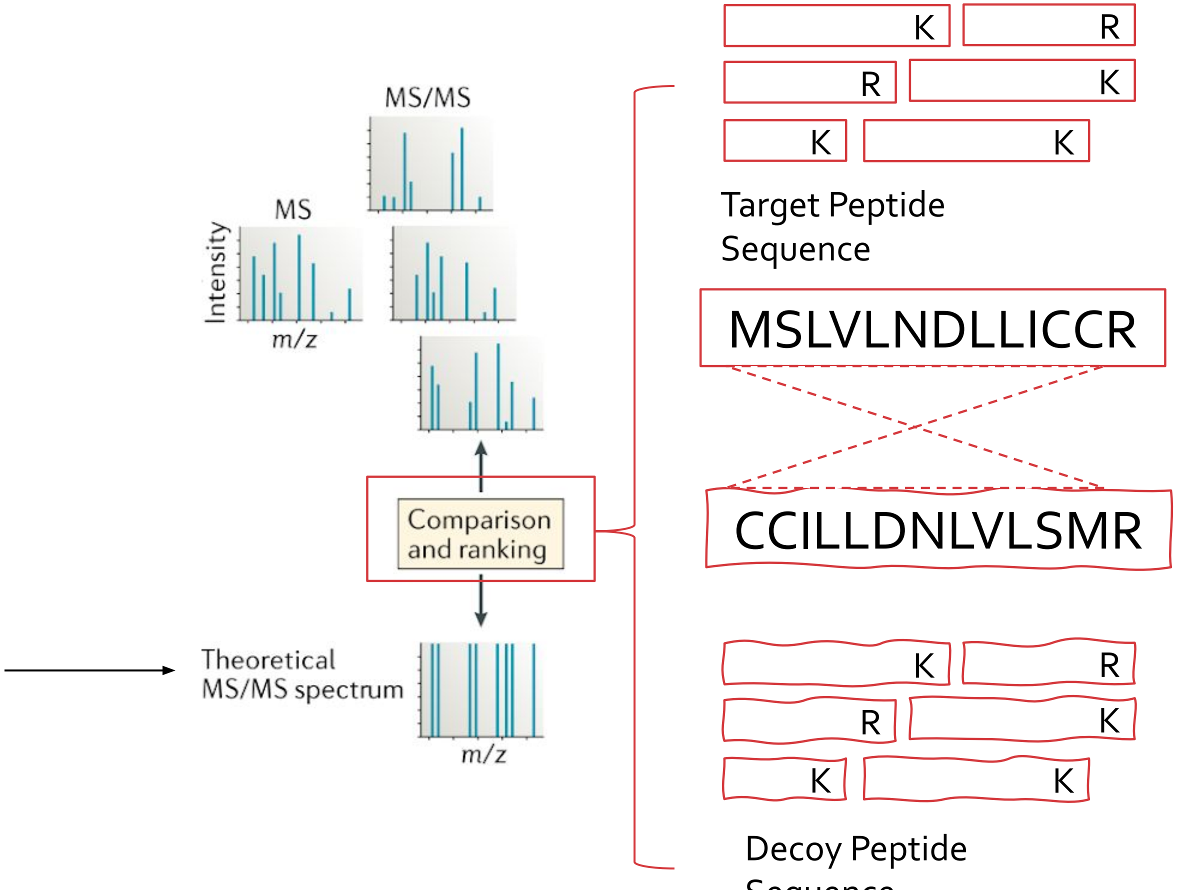
are univariately predictive of biochemical recurrence

PCAT14-203 : 370-975

MGQTESK**YASYLSFIK**ILLRRGGVRASTENLITLFQTIEQFCPWF
PEQGTLDLKDWEKIGKELKQANREGK**IIPLTVWNDWAIIKA**
TLEPFQTGEDIVSVSDAPKSCVTDCEEEAGTESQ
QGTESSHCKYVAESVMAQSTQNVDYSQLQEIIYPESSKLGEG
GPESLGPSEPKPRSPSTPPPVVQMPVTLQPQTQVRQAQTP

White et al., European Urology (2017)

Target Decoy Database Search



Target Peptide Sequence

MSLVLNDLLICCR

CCILLDNLVLSMR

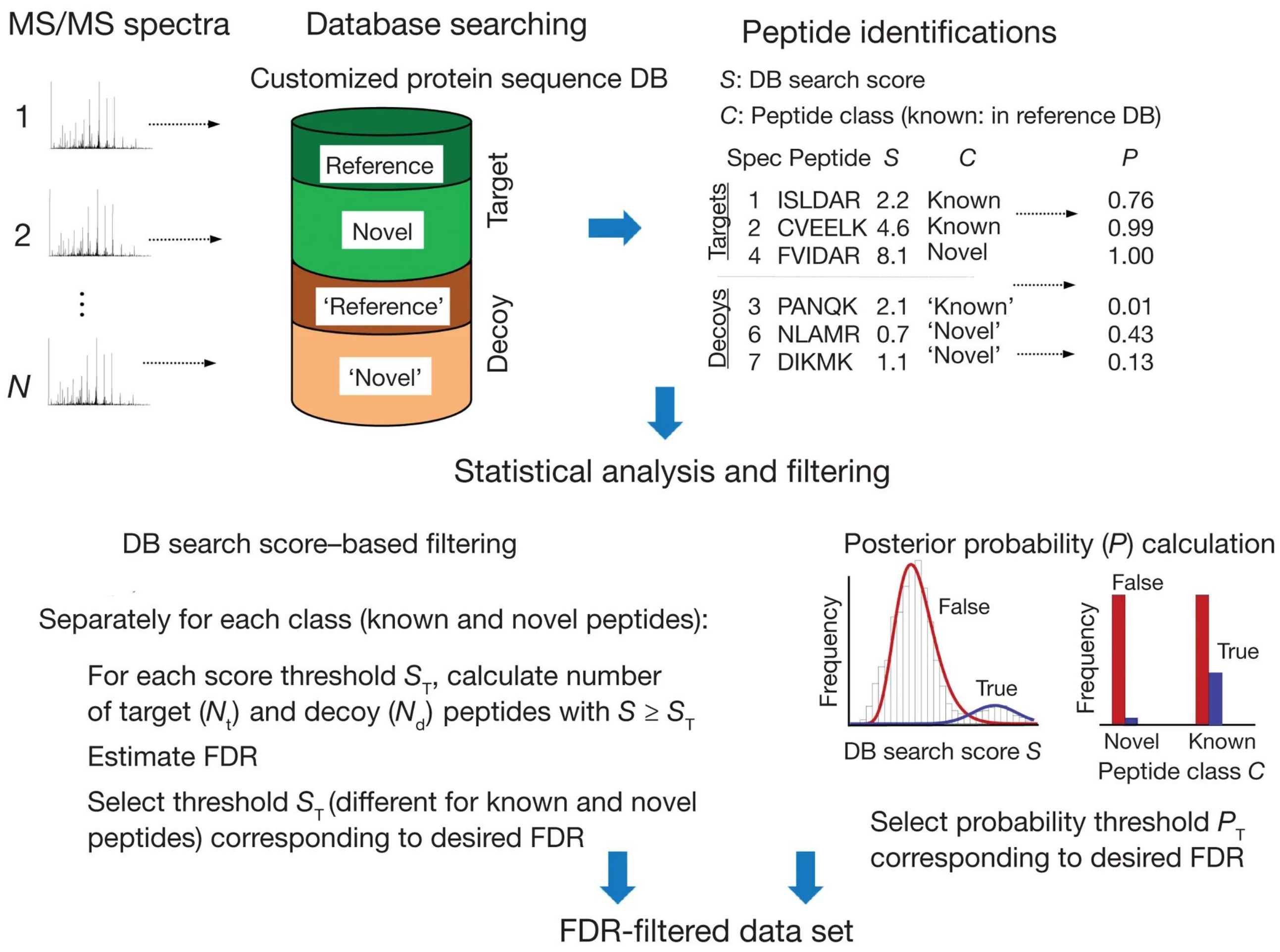
Decoy Peptide Sequence

$$FDR\ (t)$$

$$= \frac{\#\ (Target\ Peptides \geq t)}{\#\ (Decoy\ Petpdies \geq t)}$$

Select (t) corresponding to FDR = 0.01 → Peptide Detection

FDR Correction



Customized protein sequence DB

Peptide identifications

S: DB search score

C: Peptide class (known: in reference DB)

	Spec	Peptide	S	C	P
Targets	1	ISLDAR	2.2	Known	0.76
	2	CVEELK	4.6	Known	0.99
	4	FVIDAR	8.1	Novel	1.00
Decoys	3	PANQK	2.1	'Known'	0.01
	6	NLAMR	0.7	'Novel'	0.43
	7	DIKMK	1.1	'Novel'	0.13

Statistical analysis and filtering

DB search score–based filtering

Separately for each class (known and novel peptides):

For each score threshold S_T, calculate number of target (N_t) and decoy (N_d) peptides with $S \geq S_T$

Estimate FDR

Select threshold S_T (different for known and novel peptides) corresponding to desired FDR

Posterior probability (P) calculation

Select probability threshold P_T corresponding to desired FDR

FDR-filtered data set

Break!

Questions?

What gets sweeped under the rug?

Which samples goes into the analysis

- XX number of proteins detected
- Protein abundance distributions similar to other samples
- Normal / Tumour contamination
- Expected genomic / transcriptomic features

“Extensive analyses concluded that 28 of the 105 samples were compromised by protein degradation. “

How to deal with technical replicates

- Binary measurement: protein detected in any replicate
- Abundance measurement: average ignoring zero

Which genes goes into the analysis

- Protein detected in >XX% of samples

OR

- Protein detected with minimal average of X

Copy Number of a Gene

- Copy Number assigned to 1Mbp bins
- Copy Number assigned to each nucleotide base

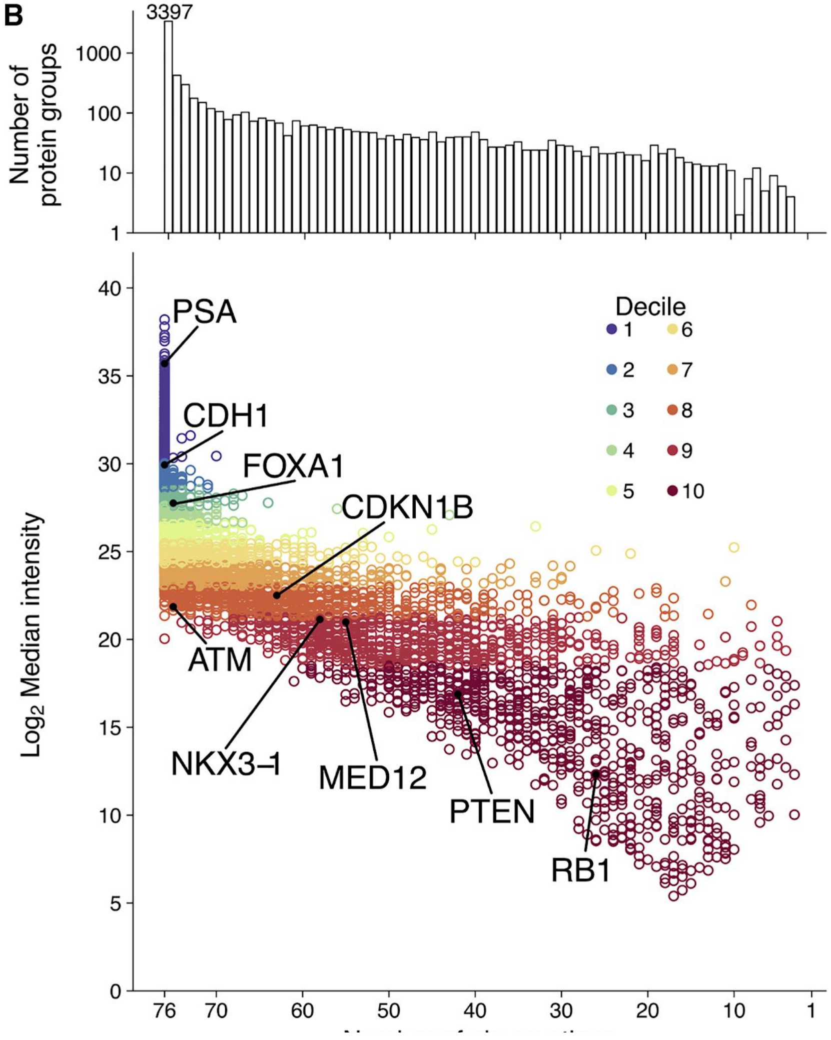
- Gene completely overlapping copy number aberration region
- Partial overlap with gene

Data Missingness

- Proteomics data is notorious for having missing values

Level of Missingness

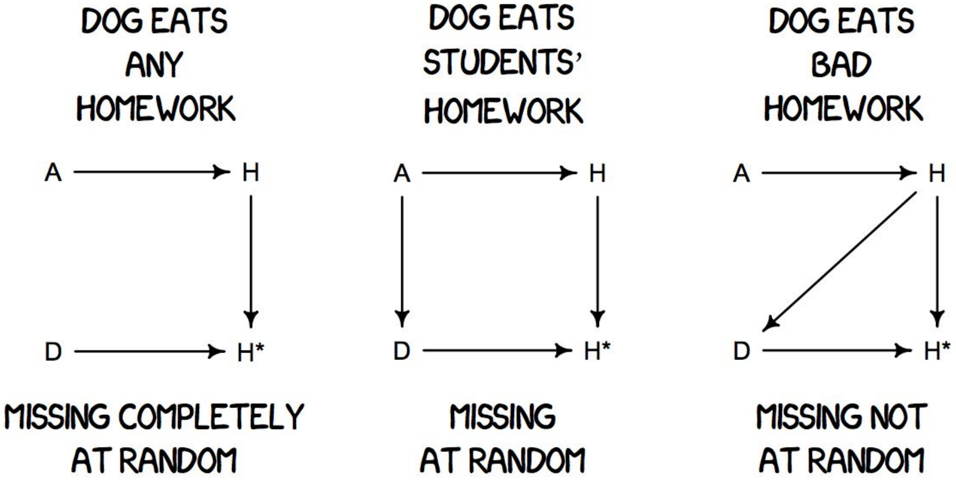
- 7054 protein groups
- 6924 protein coding genes
- 3,397 in all 76 patients



Types of Missingness

- MCAR: missing completely at random
- MAR: missing at random
- MNAR: missing not at random

H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)



Sources of Missingness

- MCAR = MAR
    - Stochastic fluctuations, not dependent on abundance
    - Protein present but not detected / incorrectly detected
- MNAR: missing not at random
    - Left-censored: protein present but below instrument detection limits
    - Negative correlation between missingness and peptide abundance
- MCAR / MNAR = ???

Types of Imputation Algorithms

- Single digit replacement
  - Mean - not recommended
  - Minimum
  - Probabilistic minimum
- Imputing around the limit of detection
  - Underestimate biological variation
  - More suitable for values Missing Not At Random

Types of Imputation Algorithms

- Impute by local structure
  - K-nearest neighbors
  - local least squares (LLS)
  - Maximum Likelihood estimation
  - Single value deposition
- Impute by Global structure
  - Probabilistic PCA
  - Bayesian PCA
  - Single value deposition
- More suitable for Missing At Random data
  - In general cases work better than the previous class

General Guidelines

- Impute at the peptide level
  - Aggregative to the protein level has implement imputation rules
- If don’t know about MCAR / MNAR ratio
  - Use MCAR suitable methods
- Could consider hybrid strategies

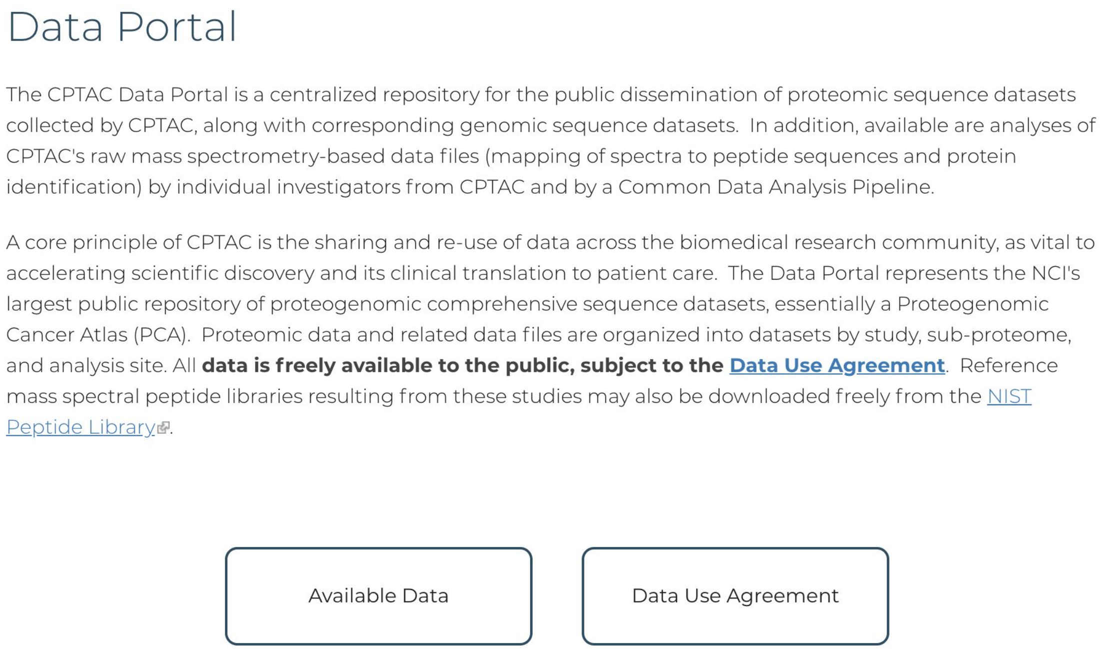
Where to find proteogenomics datasets for fun?

CPTAC

The National Cancer Institute’s Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics.

CPTAC Data Portal

https://proteomics.cancer.gov/data-portal



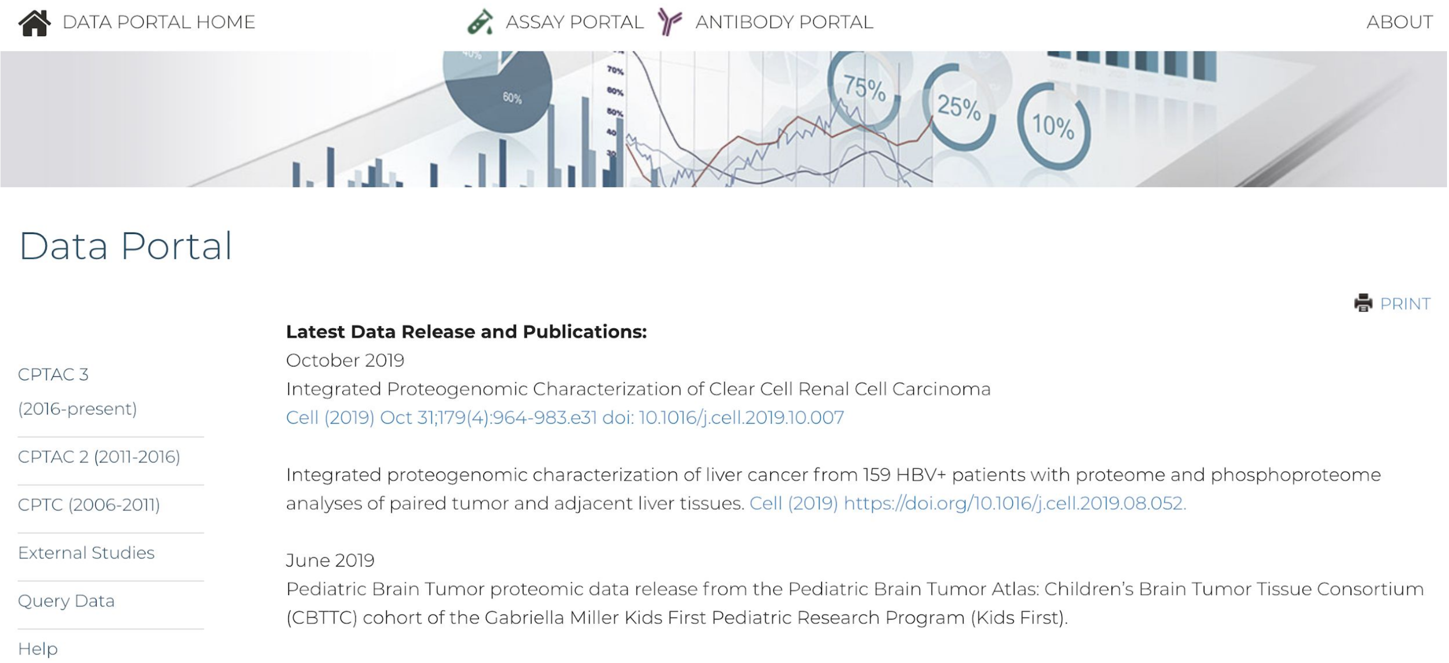
Data Portal

The CPTAC Data Portal is a centralized repository for the public dissemination of proteomic sequence datasets collected by CPTAC, along with corresponding genomic sequence datasets. In addition, available are analyses of CPTAC's raw mass spectrometry-based data files (mapping of spectra to peptide sequences and protein identification) by individual investigators from CPTAC and by a Common Data Analysis Pipeline.

A core principle of CPTAC is the sharing and re-use of data across the biomedical research community, as vital to accelerating scientific discovery and its clinical translation to patient care. The Data Portal represents the NCI's largest public repository of proteogenomic comprehensive sequence datasets, essentially a Proteogenomic Cancer Atlas (PCA). Proteomic data and related data files are organized into datasets by study, sub-proteome, and analysis site. All **data is freely available to the public, subject to the Data Use Agreement**. Reference mass spectral peptide libraries resulting from these studies may also be downloaded freely from the NIST Peptide Library.

CPTAC Data Portal

https://cptac-data-portal.georgetown.edu/cptacPublic/



Data Portal

Latest Data Release and Publications:

October 2019

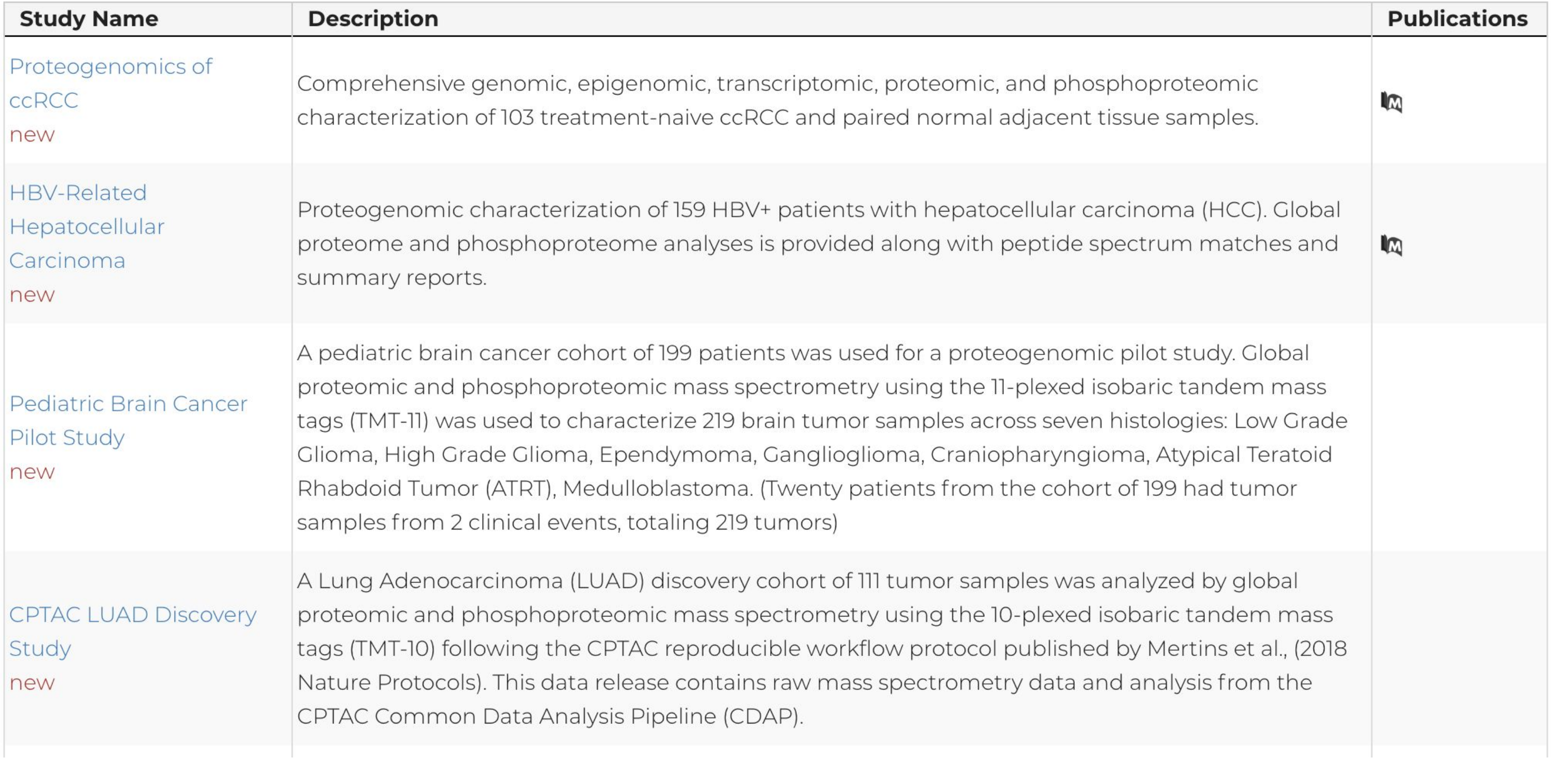
Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma
Cell (2019) Oct 31;179(4):964-983.e31 doi: 10.1016/j.cell.2019.10.007

Integrated proteogenomic characterization of liver cancer from 159 HBV+ patients with proteome and phosphoproteome analyses of paired tumor and adjacent liver tissues. Cell (2019) https://doi.org/10.1016/j.cell.2019.08.052.

June 2019

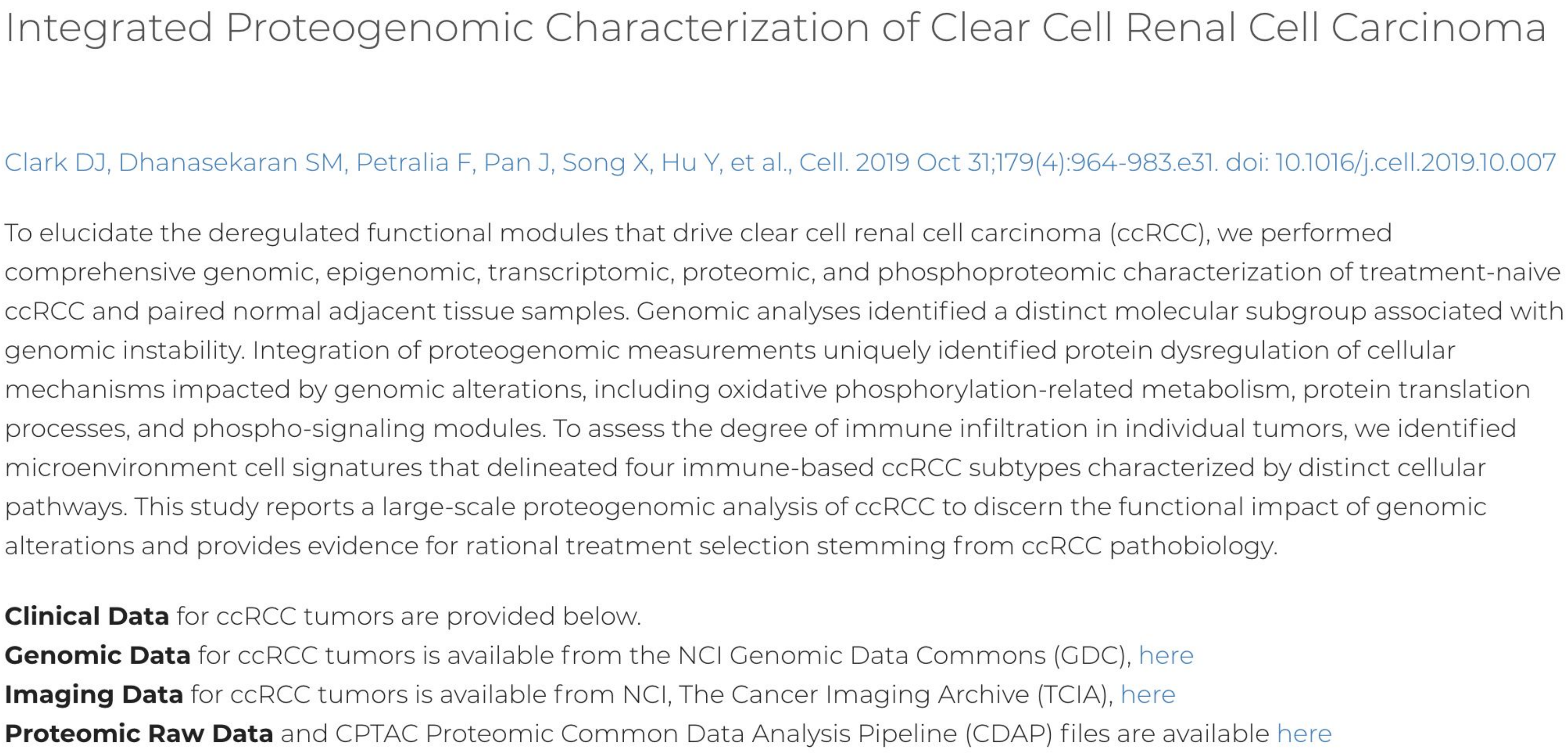
Pediatric Brain Tumor proteomic data release from the Pediatric Brain Tumor Atlas: Children's Brain Tumor Tissue Consortium (CBTTC) cohort of the Gabriella Miller Kids First Pediatric Research Program (Kids First).

CPTAC Data Portal



Study Name	Description	Publications
Proteogenomics of ccRCC new	Comprehensive genomic, epigenomic, transcriptomic, proteomic, and phosphoproteomic characterization of 103 treatment-naive ccRCC and paired normal adjacent tissue samples.	📖
HBV-Related Hepatocellular Carcinoma new	Proteogenomic characterization of 159 HBV+ patients with hepatocellular carcinoma (HCC). Global proteome and phosphoproteome analyses is provided along with peptide spectrum matches and summary reports.	📖
Pediatric Brain Cancer Pilot Study new	A pediatric brain cancer cohort of 199 patients was used for a proteogenomic pilot study. Global proteomic and phosphoproteomic mass spectrometry using the 11-plexed isobaric tandem mass tags (TMT-11) was used to characterize 219 brain tumor samples across seven histologies: Low Grade Glioma, High Grade Glioma, Ependymoma, Ganglioglioma, Craniopharyngioma, Atypical Teratoid Rhabdoid Tumor (ATRT), Medulloblastoma. (Twenty patients from the cohort of 199 had tumor samples from 2 clinical events, totaling 219 tumors)	
CPTAC LUAD Discovery Study new	A Lung Adenocarcinoma (LUAD) discovery cohort of 111 tumor samples was analyzed by global proteomic and phosphoproteomic mass spectrometry using the 10-plexed isobaric tandem mass tags (TMT-10) following the CPTAC reproducible workflow protocol published by Mertins et al., (2018 Nature Protocols). This data release contains raw mass spectrometry data and analysis from the CPTAC Common Data Analysis Pipeline (CDAP).	

CPTAC Data Portal



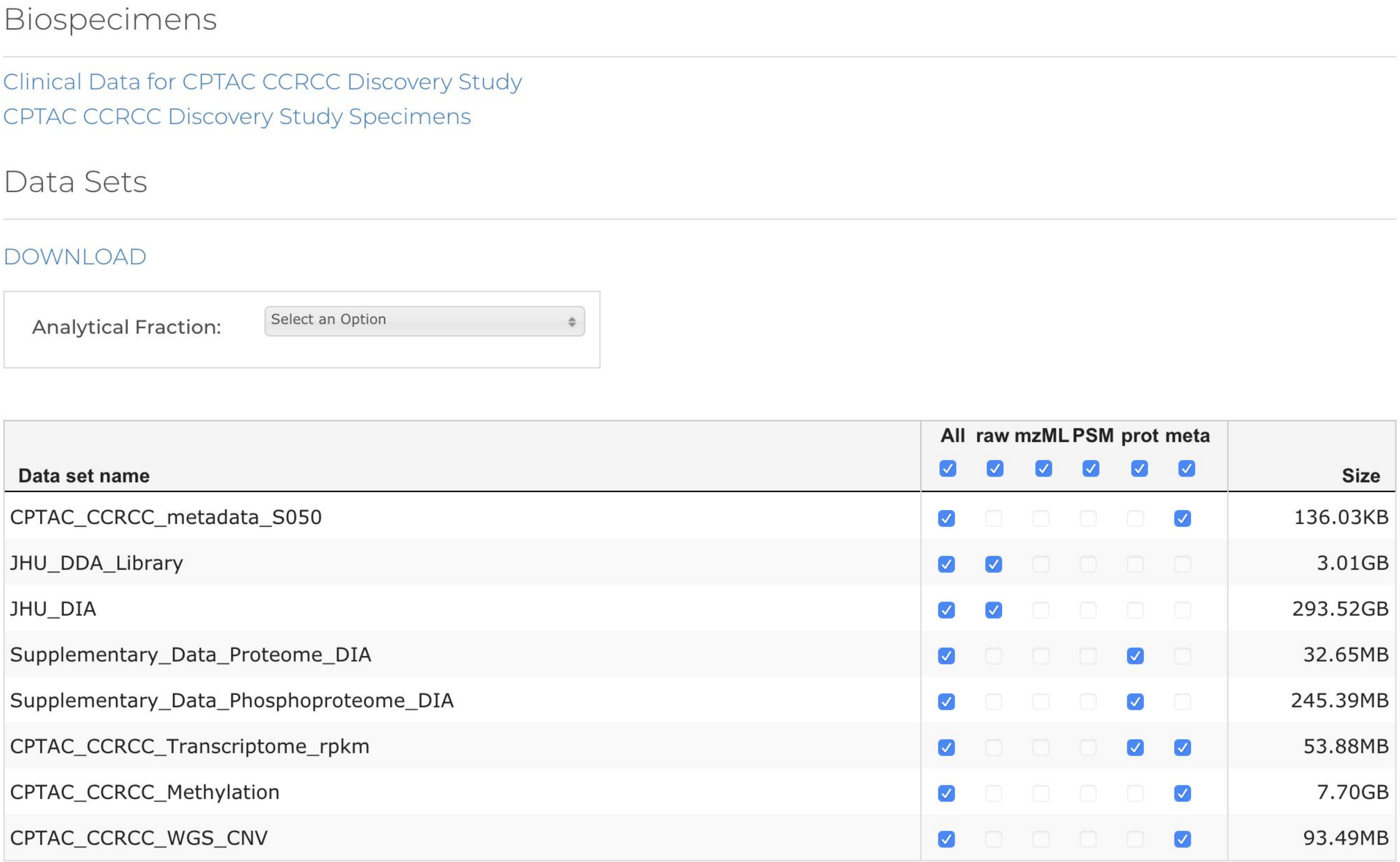
Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma

Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al., Cell. 2019 Oct 31;179(4):964-983.e31. doi: 10.1016/j.cell.2019.10.007

To elucidate the deregulated functional modules that drive clear cell renal cell carcinoma (ccRCC), we performed comprehensive genomic, epigenomic, transcriptomic, proteomic, and phosphoproteomic characterization of treatment-naive ccRCC and paired normal adjacent tissue samples. Genomic analyses identified a distinct molecular subgroup associated with genomic instability. Integration of proteogenomic measurements uniquely identified protein dysregulation of cellular mechanisms impacted by genomic alterations, including oxidative phosphorylation-related metabolism, protein translation processes, and phospho-signaling modules. To assess the degree of immune infiltration in individual tumors, we identified microenvironment cell signatures that delineated four immune-based ccRCC subtypes characterized by distinct cellular pathways. This study reports a large-scale proteogenomic analysis of ccRCC to discern the functional impact of genomic alterations and provides evidence for rational treatment selection stemming from ccRCC pathobiology.

Clinical Data for ccRCC tumors are provided below.
Genomic Data for ccRCC tumors is available from the NCI Genomic Data Commons (GDC), here
Imaging Data for ccRCC tumors is available from NCI, The Cancer Imaging Archive (TCIA), here
Proteomic Raw Data and CPTAC Proteomic Common Data Analysis Pipeline (CDAP) files are available here

Clinical



Biospecimens

Clinical Data for CPTAC CCRCC Discovery Study

CPTAC CCRCC Discovery Study Specimens

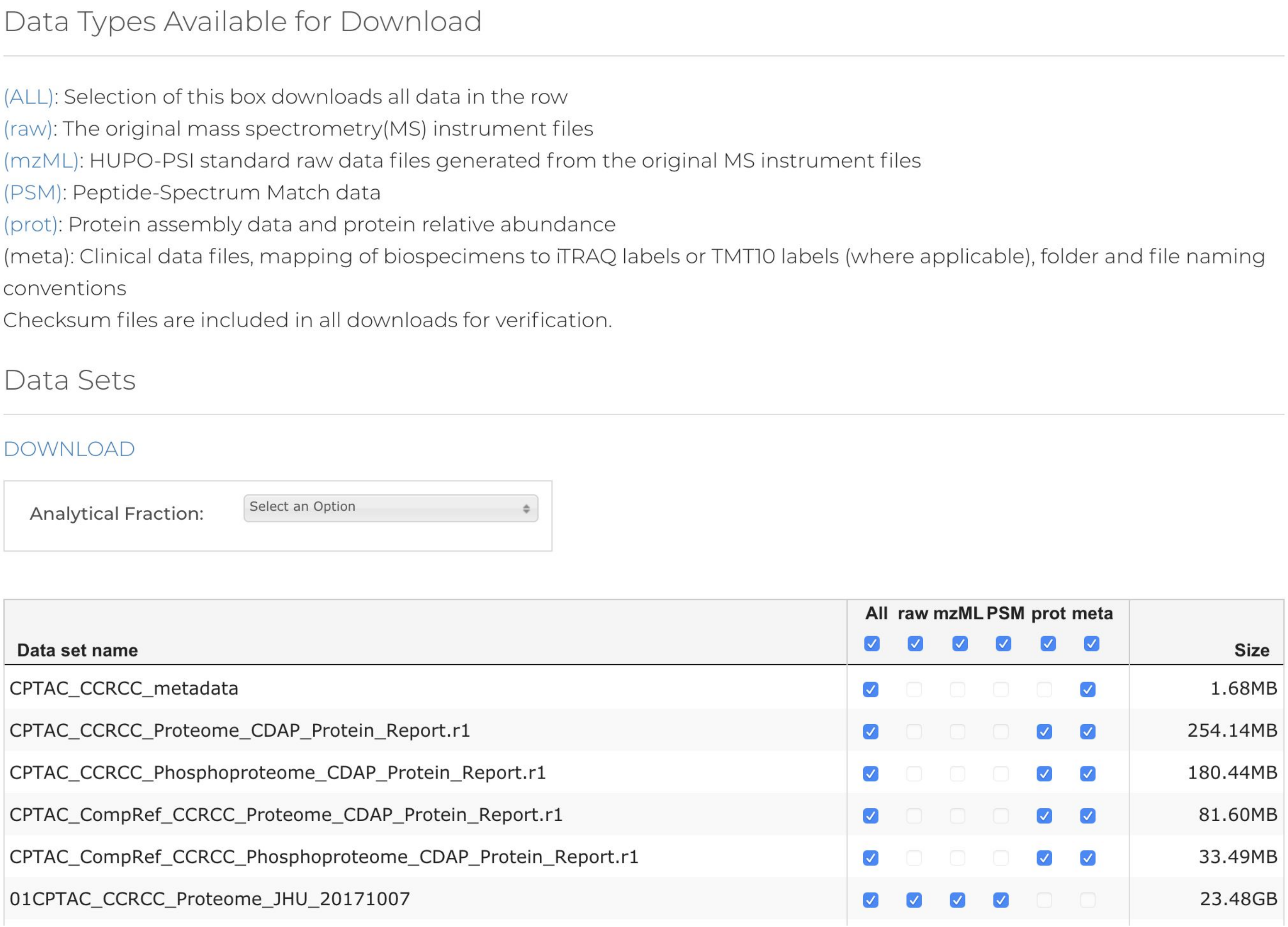
Data Sets

DOWNLOAD

Analytical Fraction: Select an Option

Data set name	All	raw	mzML	PSM	prot	meta	Size
	☑	☑	☑	☑	☑	☑	
CPTAC_CCRCC_metadata_S050	☑	□	□	□	□	☑	136.03KB
JHU_DDA_Library	☑	☑	□	□	□	□	3.01GB
JHU_DIA	☑	☑	□	□	□	□	293.52GB
Supplementary_Data_Proteome_DIA	☑	□	□	□	☑	□	32.65MB
Supplementary_Data_Phosphoproteome_DIA	☑	□	□	□	☑	□	245.39MB
CPTAC_CCRCC_Transcriptome_rpkm	☑	□	□	□	☑	☑	53.88MB
CPTAC_CCRCC_Methylation	☑	□	□	□	□	☑	7.70GB
CPTAC_CCRCC_WGS_CNV	☑	□	□	□	□	☑	93.49MB

Proteomics



Data Types Available for Download

(ALL): Selection of this box downloads all data in the row
(raw): The original mass spectrometry(MS) instrument files
(mzML): HUPO-PSI standard raw data files generated from the original MS instrument files
(PSM): Peptide-Spectrum Match data
(prot): Protein assembly data and protein relative abundance
(meta): Clinical data files, mapping of biospecimens to iTRAQ labels or TMT10 labels (where applicable), folder and file naming conventions
Checksum files are included in all downloads for verification.

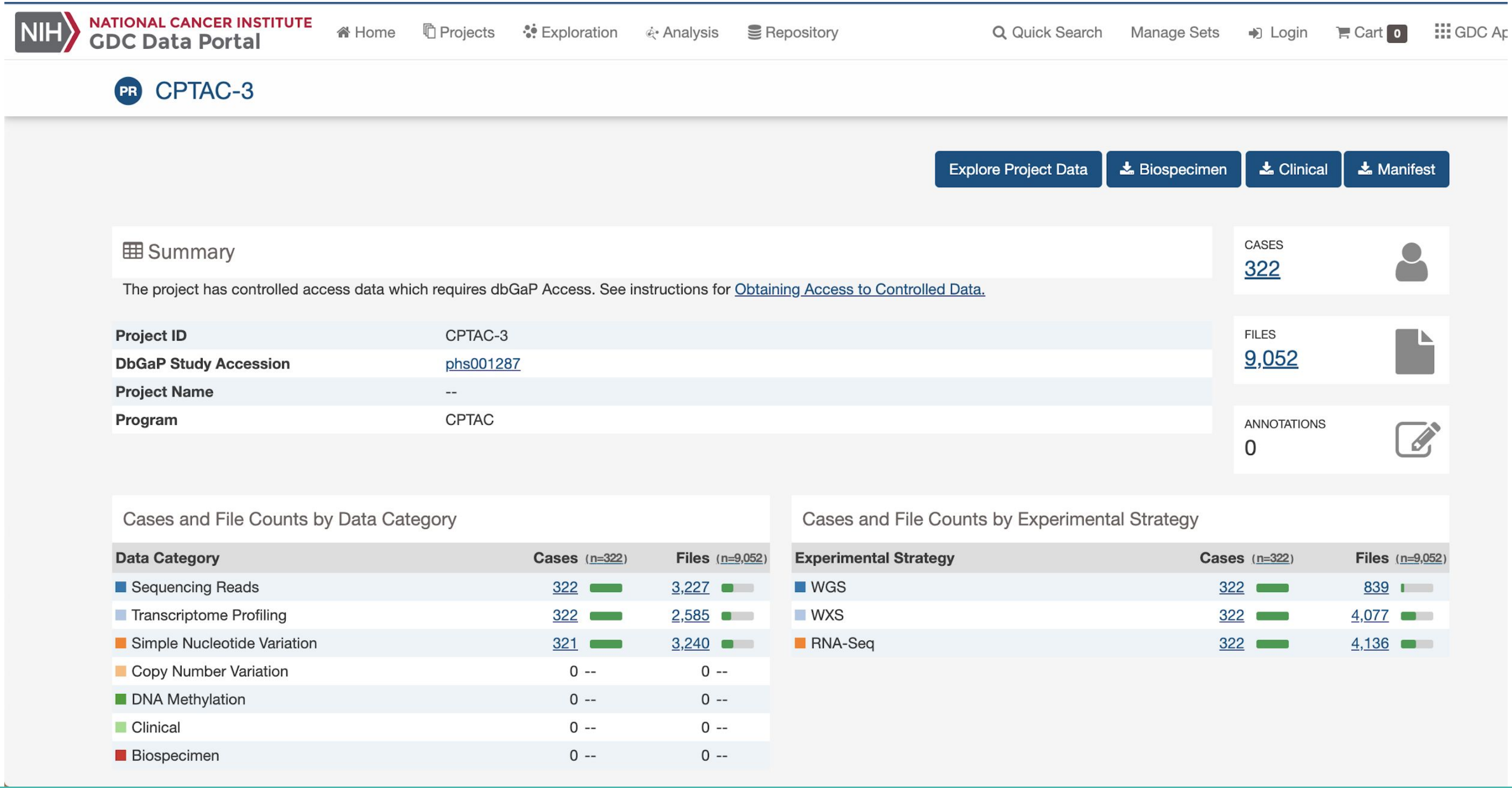
Data Sets

DOWNLOAD

Analytical Fraction: Select an Option

Data set name	All	raw	mzML	PSM	prot	meta	Size
	☑	☑	☑	☑	☑	☑	
CPTAC_CCRCC_metadata	☑	☐	☐	☐	☐	☑	1.68MB
CPTAC_CCRCC_Proteome_CDAP_Protein_Report.r1	☑	☐	☐	☐	☑	☑	254.14MB
CPTAC_CCRCC_Phosphoproteome_CDAP_Protein_Report.r1	☑	☐	☐	☐	☑	☑	180.44MB
CPTAC_CompRef_CCRCC_Proteome_CDAP_Protein_Report.r1	☑	☐	☐	☐	☑	☑	81.60MB
CPTAC_CompRef_CCRCC_Phosphoproteome_CDAP_Protein_Report.r1	☑	☐	☐	☐	☑	☑	33.49MB
01CPTAC_CCRCC_Proteome_JHU_20171007	☑	☑	☑	☑	☐	☐	23.48GB

Genomics



PR CPTAC-3

Summary

The project has controlled access data which requires dbGaP Access. See instructions for Obtaining Access to Controlled Data.

Project ID	CPTAC-3
DbGaP Study Accession	phs001287
Project Name	--
Program	CPTAC

CASES 322

FILES 9,052

ANNOTATIONS 0

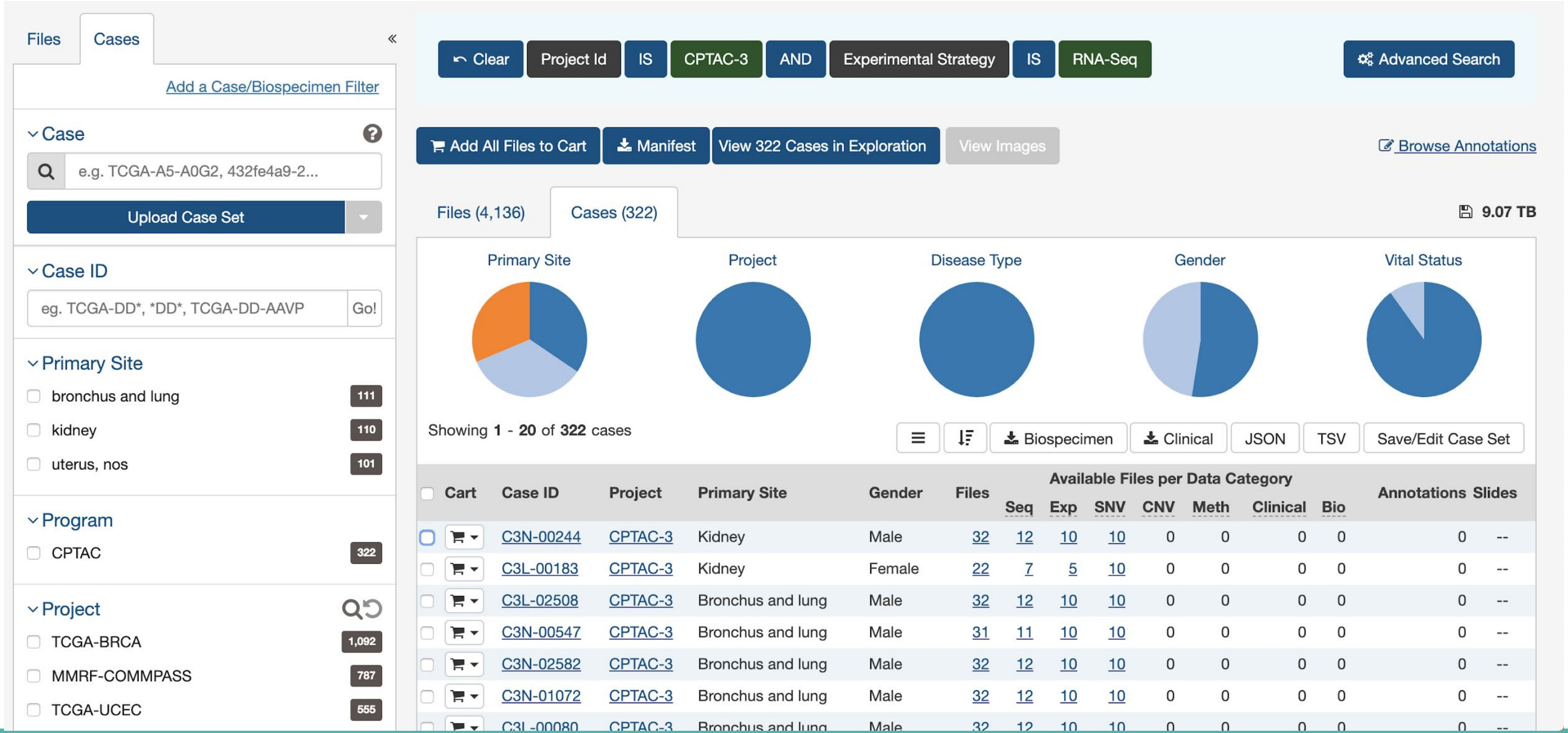
Cases and File Counts by Data Category

Data Category	Cases (n=322)	Files (n=9,052)
🟦 Sequencing Reads	322	3,227
🟦 Transcriptome Profiling	322	2,585
🟧 Simple Nucleotide Variation	321	3,240
🟧 Copy Number Variation	0 --	0 --
🟩 DNA Methylation	0 --	0 --
🟩 Clinical	0 --	0 --
🟥 Biospecimen	0 --	0 --

Cases and File Counts by Experimental Strategy

Experimental Strategy	Cases (n=322)	Files (n=9,052)
🟦 WGS	322	839
🟦 WXS	322	4,077
🟧 RNA-Seq	322	4,136

Genomics



Files | Cases | «

Add a Case/Biospecimen Filter

Case

e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Case ID

eg. TCGA-DD*, *DD*, TCGA-DD-AAVP | Go!

Primary Site

□ bronchus and lung	111
□ kidney	110
□ uterus, nos	101

Program

□ CPTAC	322

Project

□ TCGA-BRCA	1,092
□ MMRF-COMMPASS	787
□ TCGA-UCEC	555

Clear | Project Id | IS | CPTAC-3 | AND | Experimental Strategy | IS | RNA-Seq | Advanced Search

Add All Files to Cart | Manifest | View 322 Cases in Exploration | View Images | Browse Annotations

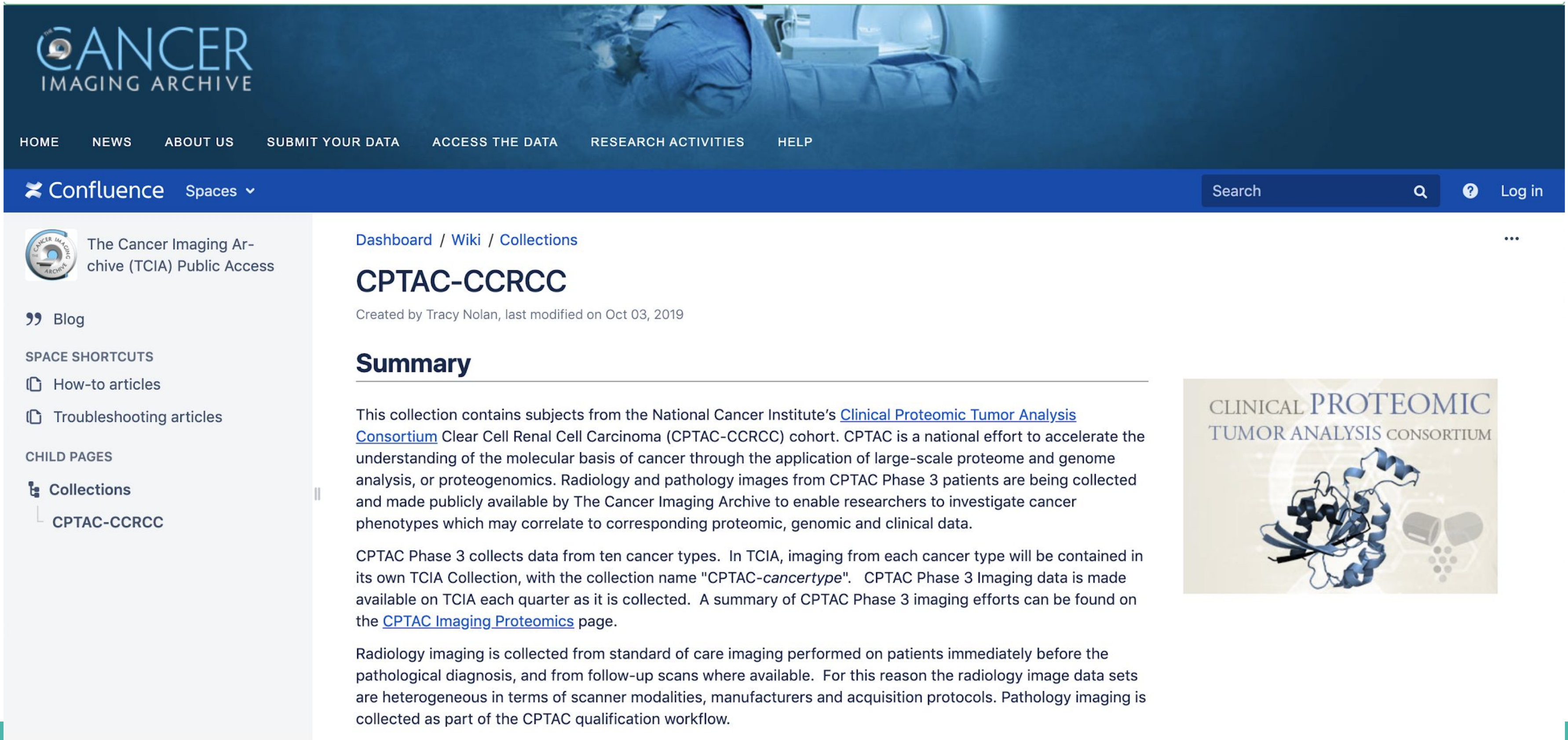
Files (4,136) | Cases (322) | 9.07 TB

Showing **1** - **20** of **322** cases

Biospecimen | Clinical | JSON | TSV | Save/Edit Case Set

□	Cart	Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							Annotations	Slides
							Seq	Exp	SNV	CNV	Meth	Clinical	Bio		
□		C3N-00244	CPTAC-3	Kidney	Male	32	12	10	10	0	0	0	0	0	--
□		C3L-00183	CPTAC-3	Kidney	Female	22	7	5	10	0	0	0	0	0	--
□		C3L-02508	CPTAC-3	Bronchus and lung	Male	32	12	10	10	0	0	0	0	0	--
□		C3N-00547	CPTAC-3	Bronchus and lung	Male	31	11	10	10	0	0	0	0	0	--
□		C3N-02582	CPTAC-3	Bronchus and lung	Male	32	12	10	10	0	0	0	0	0	--
□		C3N-01072	CPTAC-3	Bronchus and lung	Male	32	12	10	10	0	0	0	0	0	--
□		C3L-00080	CPTAC-3	Bronchus and lung	Male	32	12	10	10	0	0	0	0	0	--

Imaging



Dashboard / Wiki / Collections

CPTAC-CCRCC

Created by Tracy Nolan, last modified on Oct 03, 2019

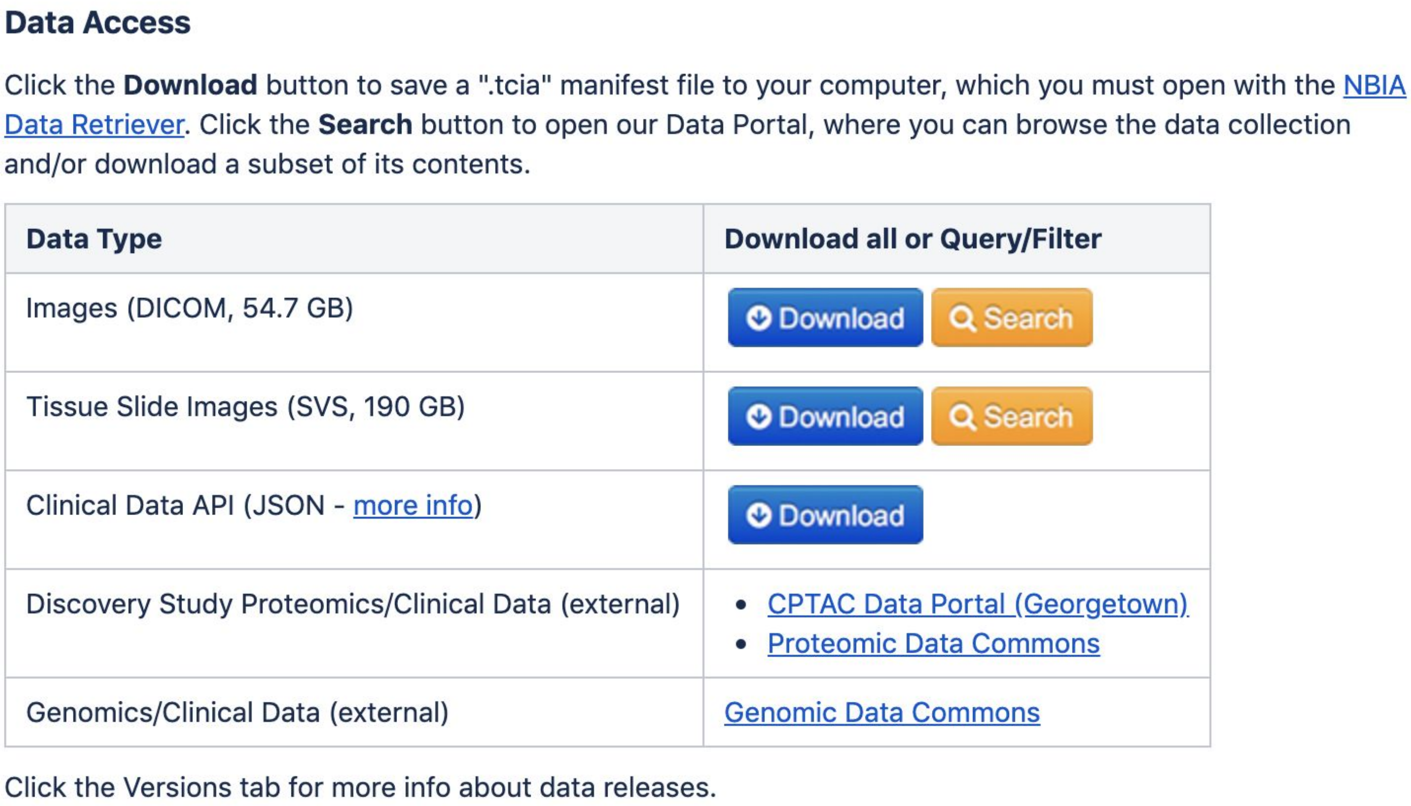
Summary

This collection contains subjects from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium Clear Cell Renal Cell Carcinoma (CPTAC-CCRCC) cohort. CPTAC is a national effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis, or proteogenomics. Radiology and pathology images from CPTAC Phase 3 patients are being collected and made publicly available by The Cancer Imaging Archive to enable researchers to investigate cancer phenotypes which may correlate to corresponding proteomic, genomic and clinical data.

CPTAC Phase 3 collects data from ten cancer types. In TCIA, imaging from each cancer type will be contained in its own TCIA Collection, with the collection name "CPTAC-*cancertype*". CPTAC Phase 3 Imaging data is made available on TCIA each quarter as it is collected. A summary of CPTAC Phase 3 imaging efforts can be found on the CPTAC Imaging Proteomics page.

Radiology imaging is collected from standard of care imaging performed on patients immediately before the pathological diagnosis, and from follow-up scans where available. For this reason the radiology image data sets are heterogeneous in terms of scanner modalities, manufacturers and acquisition protocols. Pathology imaging is collected as part of the CPTAC qualification workflow.

Imaging



Data Access

Click the **Download** button to save a ".tcia" manifest file to your computer, which you must open with the NBIA Data Retriever. Click the **Search** button to open our Data Portal, where you can browse the data collection and/or download a subset of its contents.

Data Type	Download all or Query/Filter
Images (DICOM, 54.7 GB)	Download Search
Tissue Slide Images (SVS, 190 GB)	Download Search
Clinical Data API (JSON - more info)	Download
Discovery Study Proteomics/Clinical Data (external)	• CPTAC Data Portal (Georgetown) • Proteomic Data Commons
Genomics/Clinical Data (external)	Genomic Data Commons

Click the Versions tab for more info about data releases.

Thank you

Lydia Liu

lydia.liu@mail.utoronto.ca

Appendix

For More on Proteomics

https://mbp-tech-talks.github.io/2019-2020/04-intro-proteomics/intro-proteomics_amanda-khoo.pdf