P-MART

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Interactive Informatics Resource for Research-driven Proteomics

CONTACT:
BOBBIE-JO WEBB-ROBERTSON
BJ@PNNL.GOV

Exploring Proteomic Data - The Need

- Wealth of proteomic cancer data available in the public domain
- Peptide level data difficult to access
- Protein level data
  - Often not reproducible
  - Doesn't offer in depth view of peptide evidence
  - Isoforms are usually lost

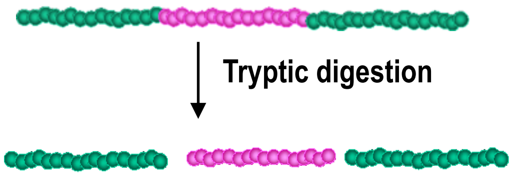
OFFICE OF CANCER CLINICAL PROTEOMICS RESEARCH
CPTAC DATA PORTAL

Early Detection Research Network
Biomarkers: the key to early detection

Anyone who wants to explore peptide and/or protein level data associated with cancer.

Typical Bottom-up Global MS-based Proteomics Workflow

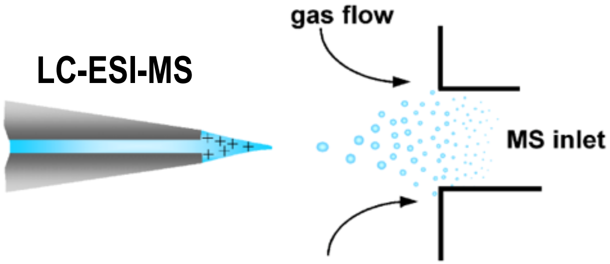
Proteins



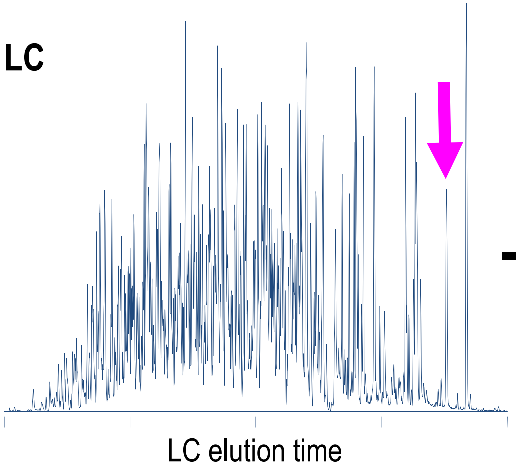
Tryptic digestion

Peptide mixture

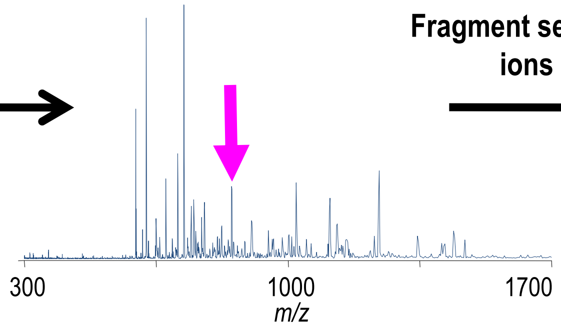
LC-ESI-MS



LC

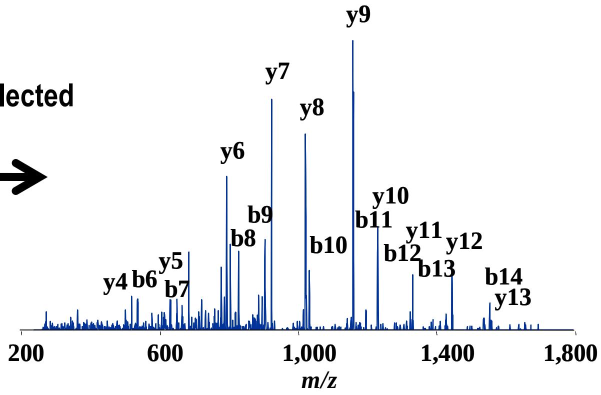


Mass spectrum (MS1)



Fragment selected ions

MS/MS fragmentation spectrum

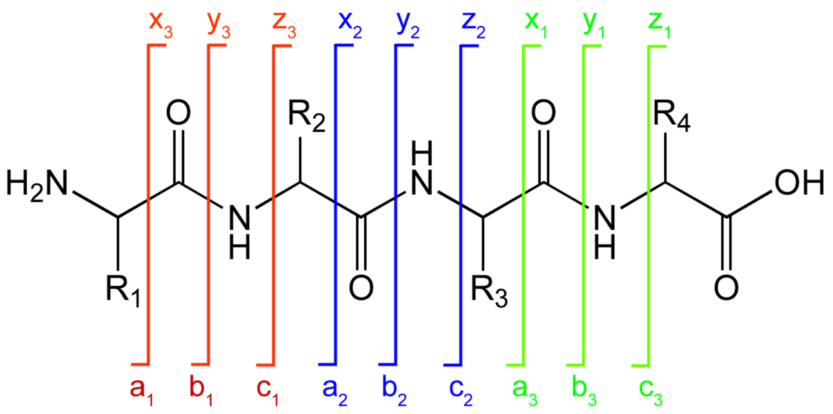


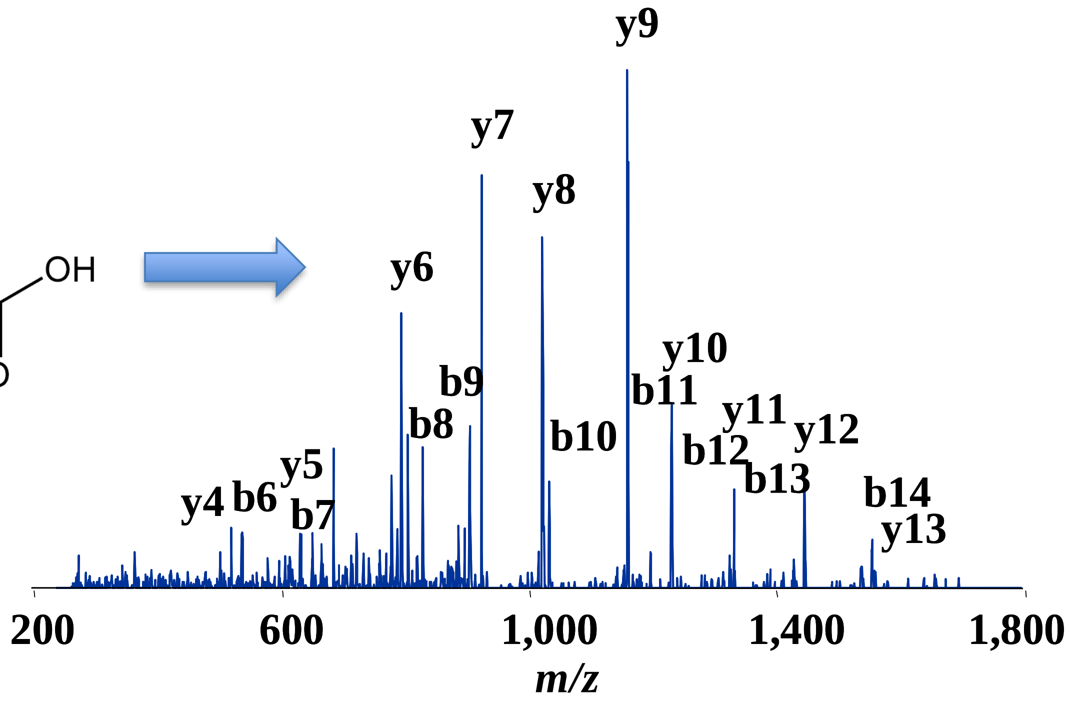
Quantification

Peptide identifications

Peptide Identification

In theory the peptide should fragment into a MS/MS spectra based on measurable mass shifts





Eng et al, J Am Soc Mass Spectrom, 1994
Perkins et al, Electrophoresis, 1999
Kim et al, J Proteome Res, 2008

Peptides are "Initially" Identified via a Database Search

Protein Database

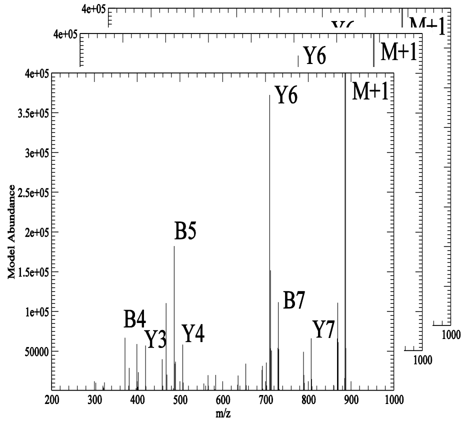
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]

LCLYTHIGRNIYYGSYLYSETWNKGIMLLLITMATAFMGYVLPWKQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY

SEQUEST

MS-GF+ X!

LCLYTHIGR•N
R •NIYYGSYLYSETWNK•G
…….

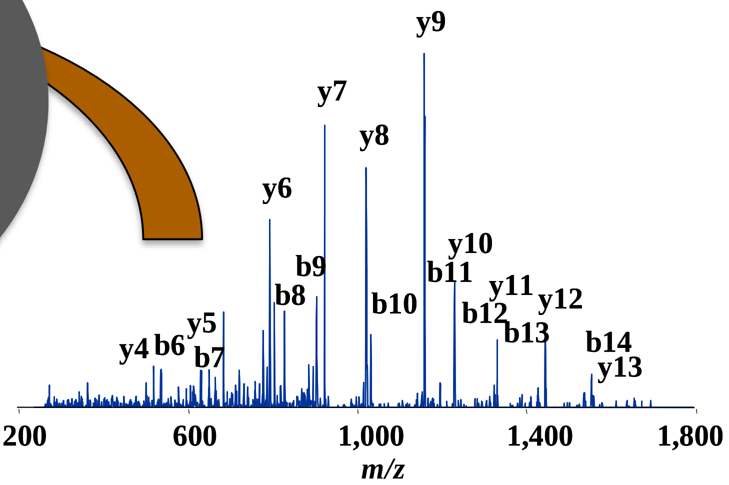


Does the Experimental Match a Theoretical Spectra?

Peptide Match (scoring threshold)

K •IRYQVTSVSNK•G

MS/MS Experimental fragmentation spectrum



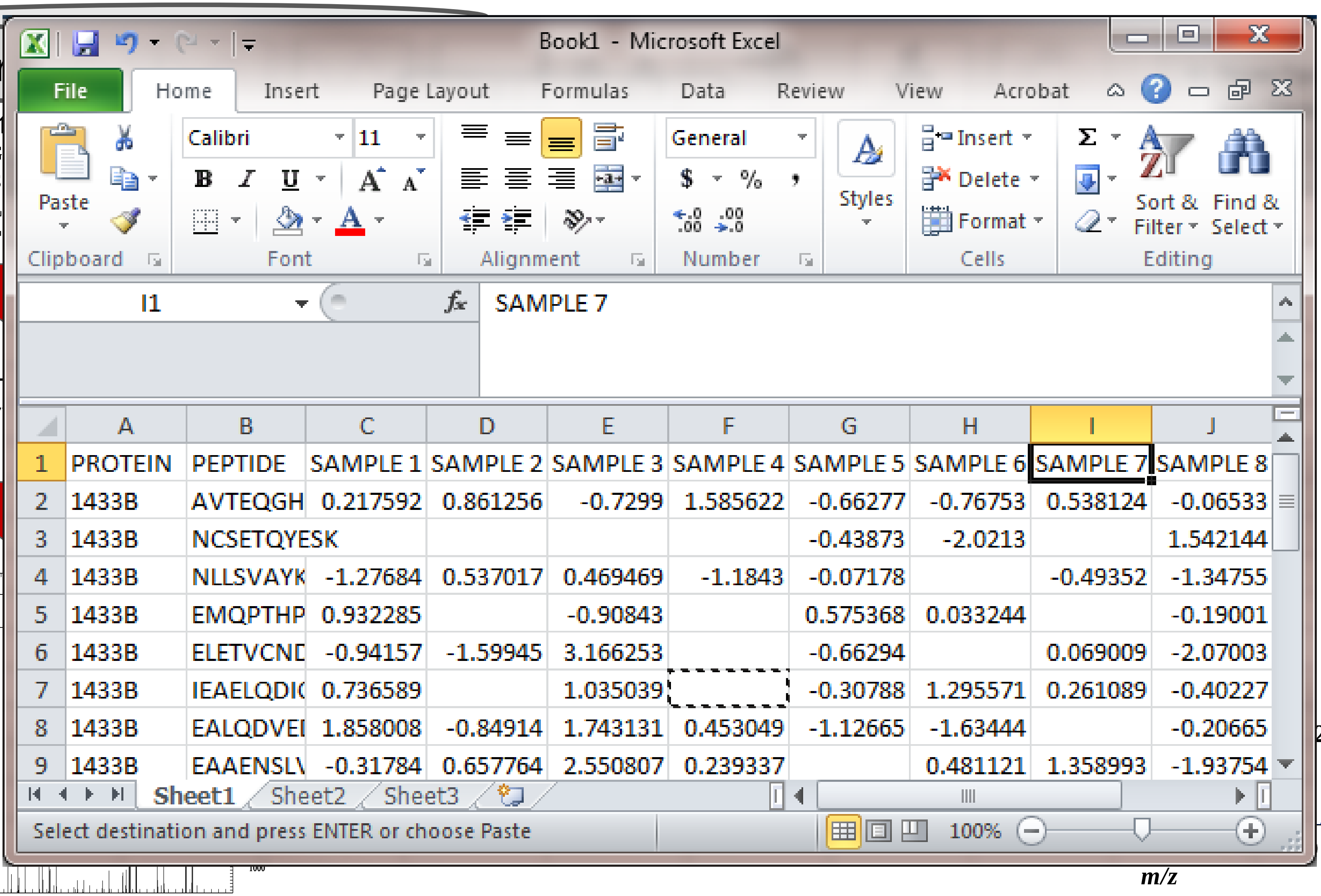
Peptides are “Initially” Identified via a Database Search

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Pr

>gi|552421
LCLYTHIG
EWIWGGFS
LLILILLL
GLMPFLHT
IENY

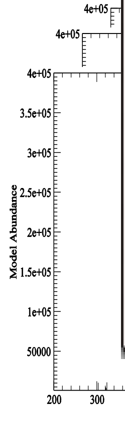
LCLYTH
R •NIY
....

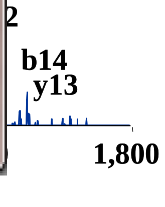


SAMPLE 7

	A	B	C	D	E	F	G	H	I	J
1	PROTEIN	PEPTIDE	SAMPLE 1	SAMPLE 2	SAMPLE 3	SAMPLE 4	SAMPLE 5	SAMPLE 6	SAMPLE 7	SAMPLE 8
2	1433B	AVTEQGH	0.217592	0.861256	-0.7299	1.585622	-0.66277	-0.76753	0.538124	-0.06533
3	1433B	NCSETQYESK					-0.43873	-2.0213		1.542144
4	1433B	NLLSVAYK	-1.27684	0.537017	0.469469	-1.1843	-0.07178		-0.49352	-1.34755
5	1433B	EMQPTHP	0.932285		-0.90843		0.575368	0.033244		-0.19001
6	1433B	ELETVCND	-0.94157	-1.59945	3.166253		-0.66294		0.069009	-2.07003
7	1433B	IEAELQDI(	0.736589		1.035039		-0.30788	1.295571	0.261089	-0.40227
8	1433B	EALQDVEI	1.858008	-0.84914	1.743131	0.453049	-1.12665	-1.63444		-0.20665
9	1433B	EAAENSLV	-0.31784	0.657764	2.550807	0.239337		0.481121	1.358993	-1.93754

Select destination and press ENTER or choose Paste





The challenges for proteomics compared to genomics?

http://proteomics.cancer.gov/whatisproteomics

Exact quote from the cited website

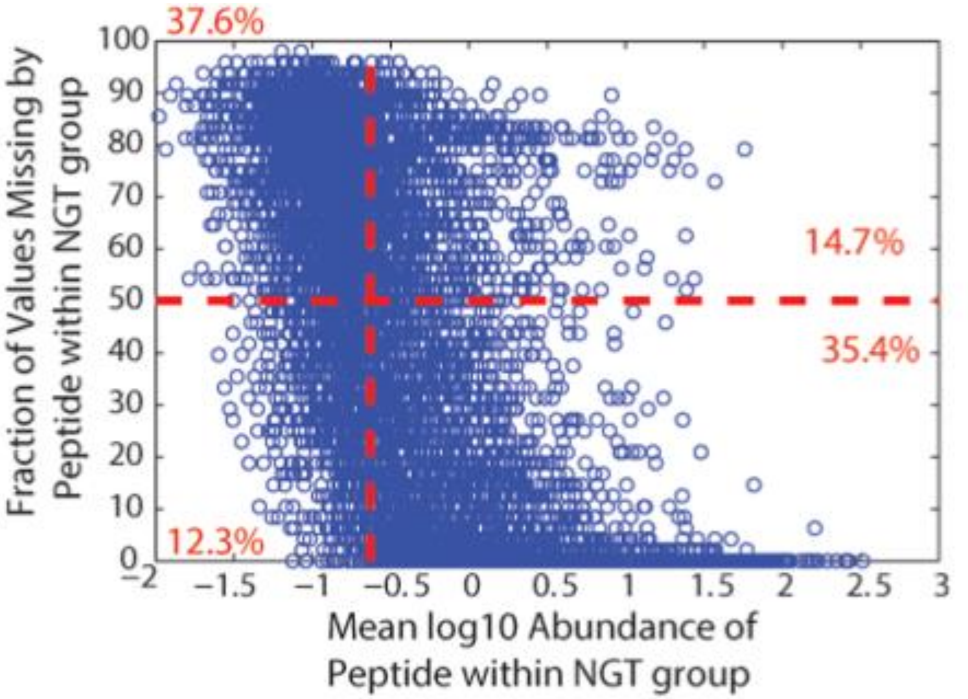
- One gene can encode more than one protein (even up to 1,000). The human genome contains about 21,000 protein-encoding genes, but the total number of proteins in human cells is estimated to be between 250,000 to one million.
- Proteins are dynamic. Proteins are continually undergoing changes, e.g., binding to the cell membrane, partnering with other proteins to form complexes, or undergoing synthesis and degradation. The genome, on the other hand, is relatively static.
- Proteins are co- and post-translationally modified. As a result, the types of proteins measured can vary considerably from one person to another under different environmental conditions, or even within the same person at different ages or states of health.

  Additionally, certain modifications can regulate the dynamics of proteins.
- Proteins exist in a wide range of concentrations in the body. For example, the concentration of the protein albumin in blood is more than a billion times greater than that of interleukin-6, making it extremely difficult to detect the low abundance proteins in a complex biological matrix such as blood. Scientists believe that the most important proteins for cancer may be those found in the lowest concentrations.

The challenges of statistical analysis of proteomics data

- Missing data
  - There is a lot of it
  - The mechanisms by which it is missing is complex

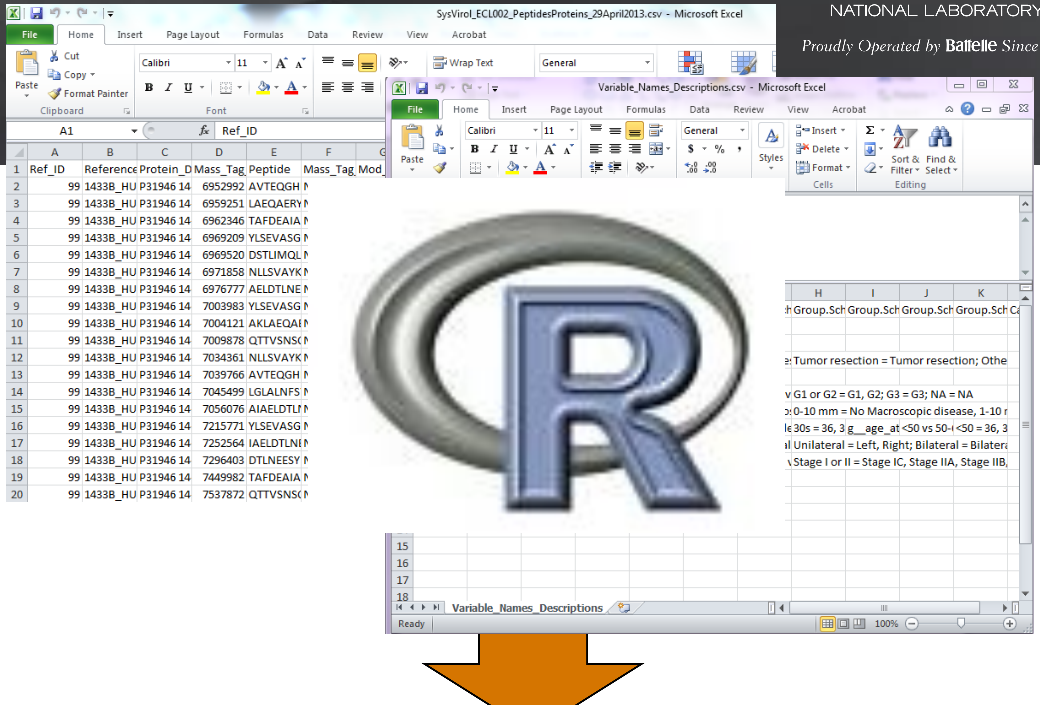
E.g., In CPTAC ovarian cancer data there is not a single peptide that has 100% coverage across all biological samples.



- Mis-identifications
- Variability (technical and biological

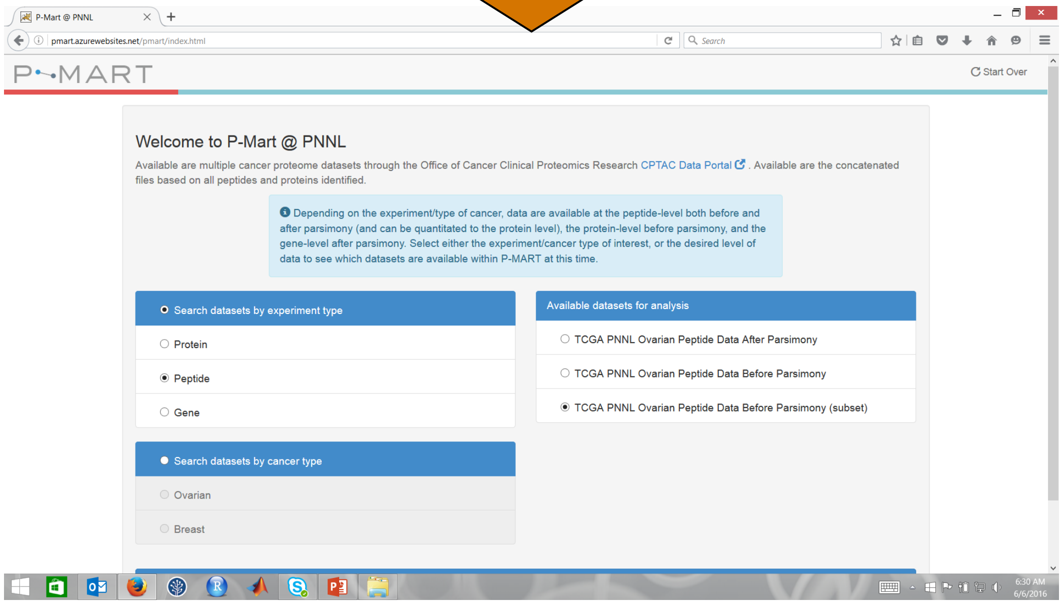
P-Mart Goal

Enable Biomarker Discovery from large complex proteomics datasets by facilitating reproducible statistical processing and complex exploration of high-throughput peptide and protein level data generated through MS robust to missing values.



	A	B	C	D	E	F	G
1	Ref_ID	Reference	Protein_D	Mass_Tag	Peptide	Mass_Tag	Mod
2	99	1433B_HU	P31946 14	6952992	AVTEQGH		
3	99	1433B_HU	P31946 14	6959251	LAEQAERY		
4	99	1433B_HU	P31946 14	6962346	TAFDEAIA		
5	99	1433B_HU	P31946 14	6969209	YLSEVASG		
6	99	1433B_HU	P31946 14	6969520	DSTLIMQL		
7	99	1433B_HU	P31946 14	6971858	NLLSVAYK		
8	99	1433B_HU	P31946 14	6976777	AELDTLNE		
9	99	1433B_HU	P31946 14	7003983	YLSEVASG		
10	99	1433B_HU	P31946 14	7004121	AKLAEQAE		
11	99	1433B_HU	P31946 14	7009878	QTTVSNS(		
12	99	1433B_HU	P31946 14	7034361	NLLSVAYK		
13	99	1433B_HU	P31946 14	7039766	AVTEQGH		
14	99	1433B_HU	P31946 14	7045499	LGLALNFS		
15	99	1433B_HU	P31946 14	7056076	AIAELDTLI		
16	99	1433B_HU	P31946 14	7215771	YLSEVASG		
17	99	1433B_HU	P31946 14	7252564	IAELDTLNI		
18	99	1433B_HU	P31946 14	7296403	DTLNEESY		
19	99	1433B_HU	P31946 14	7449982	TAFDEAIA		
20	99	1433B_HU	P31946 14	7537872	QTTVSNS(		

H	I	J	K
Group.Sch	Group.Sch	Group.Sch	Group.Sch
Tumor resection = Tumor resection; Othe			
G1 or G2 = G1, G2; G3 = G3; NA = NA			
0-10 mm = No Macroscopic disease, 1-10 r			
30s = 36, 3	g__age_at	<50 vs 50-(	<50 = 36, 3
Unilateral = Left, Right; Bilateral = Bilatera			
Stage I or II = Stage IC, Stage IIA, Stage IIB,			



Welcome to P-Mart @ PNNL

Available are multiple cancer proteome datasets through the Office of Cancer Clinical Proteomics Research CPTAC Data Portal. Available are the concatenated files based on all peptides and proteins identified.

Depending on the experiment/type of cancer, data are available at the peptide-level both before and after parsimony (and can be quantitated to the protein level), the protein-level before parsimony, and the gene-level after parsimony. Select either the experiment/cancer type of interest, or the desired level of data to see which datasets are available within P-MART at this time.

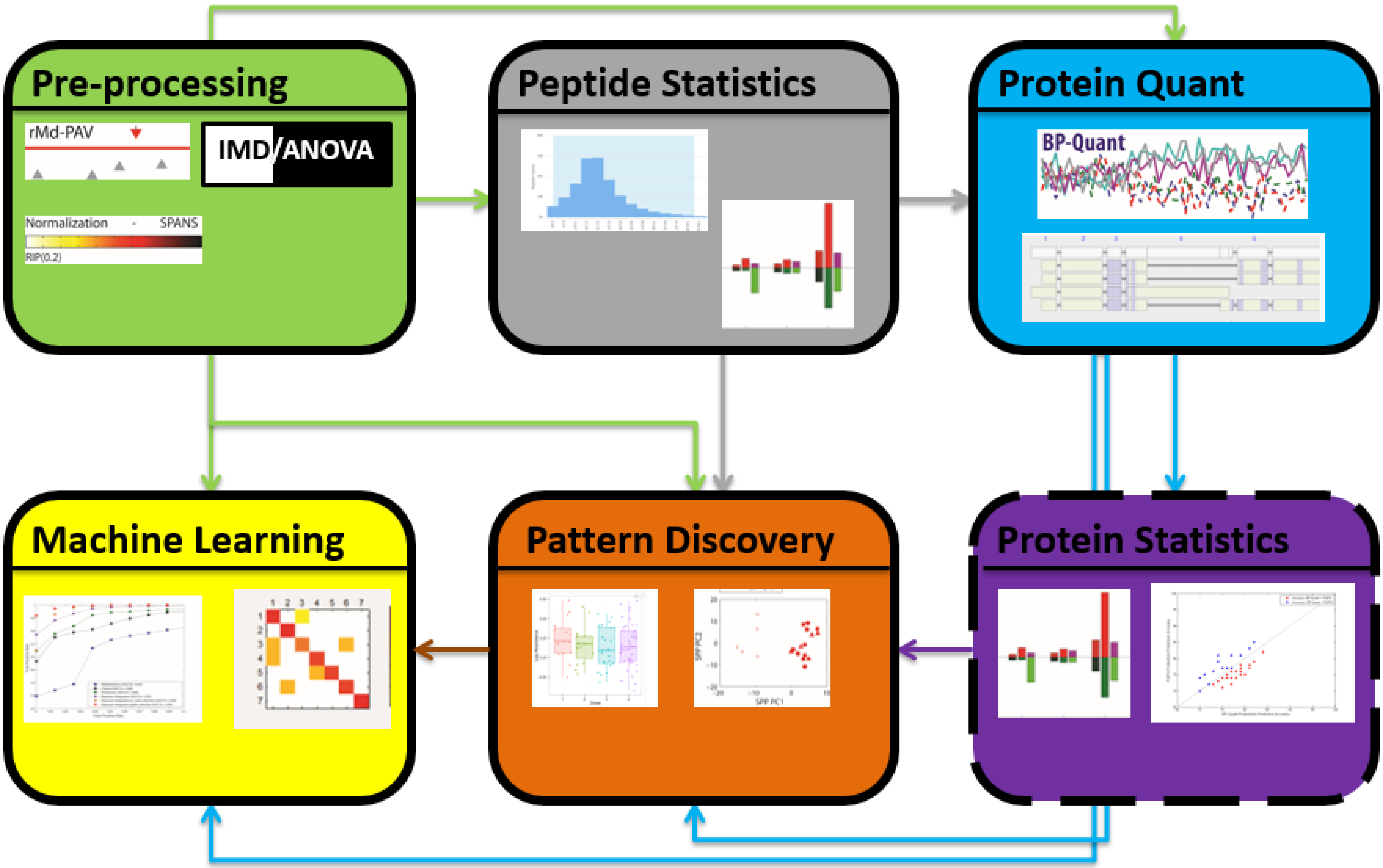
● Search datasets by experiment type
○ Protein
● Peptide
○ Gene

● Search datasets by cancer type
○ Ovarian
○ Breast

Available datasets for analysis
○ TCGA PNNL Ovarian Peptide Data After Parsimony
○ TCGA PNNL Ovarian Peptide Data Before Parsimony
● TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)

P-Mart Capabilities

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Pre-processing

Peptide Statistics

Protein Quant

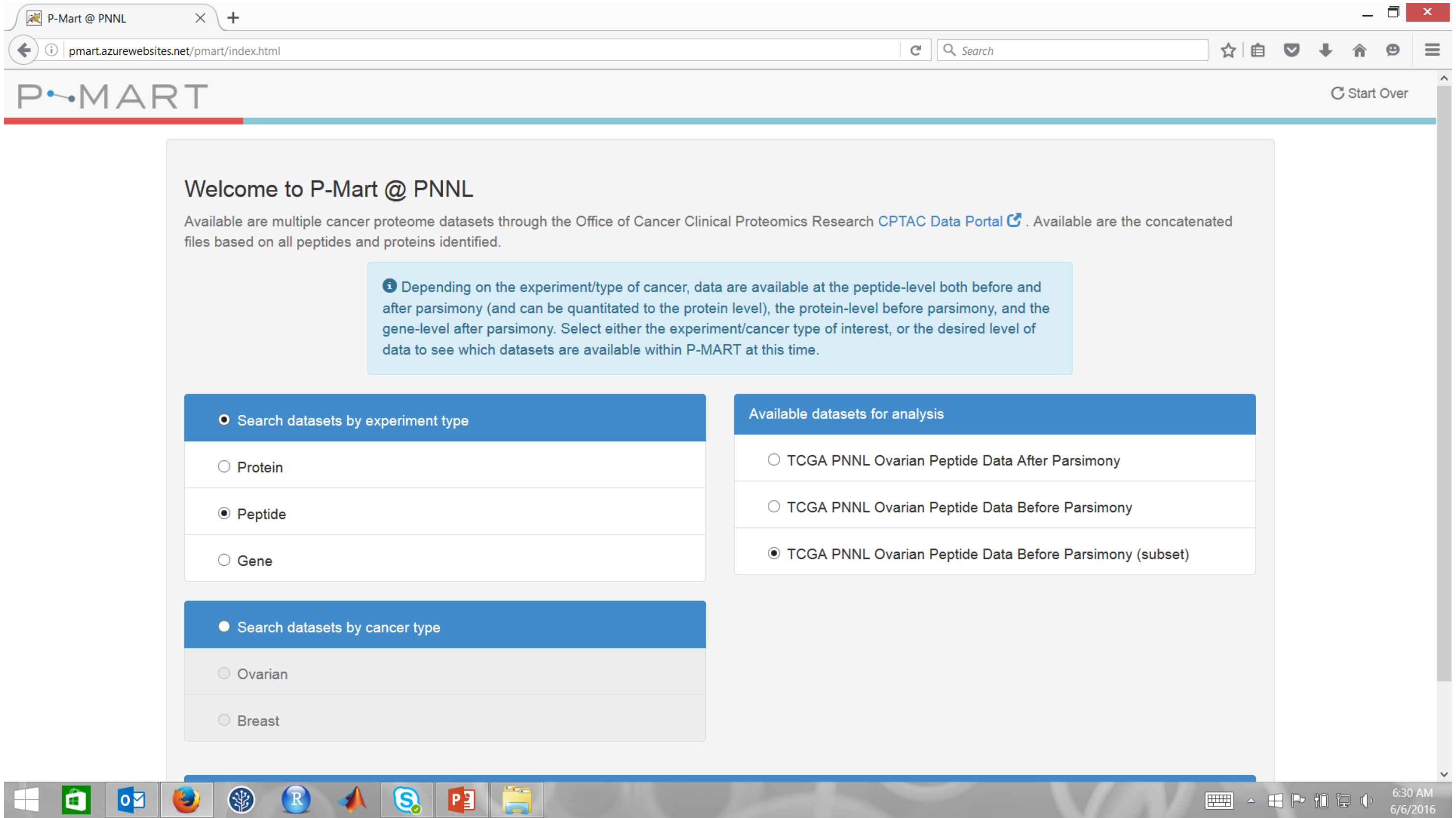
Machine Learning

Pattern Discovery

Protein Statistics

P-Mart - Data

- Access to peptide, protein or gene data
  - Peptide data is extracted from the PSM files
  - Protein and Gene data is as provided by CPTAC DCC



Welcome to P-Mart @ PNNL

Available are multiple cancer proteome datasets through the Office of Cancer Clinical Proteomics Research CPTAC Data Portal. Available are the concatenated files based on all peptides and proteins identified.

Depending on the experiment/type of cancer, data are available at the peptide-level both before and after parsimony (and can be quantitated to the protein level), the protein-level before parsimony, and the gene-level after parsimony. Select either the experiment/cancer type of interest, or the desired level of data to see which datasets are available within P-MART at this time.

● Search datasets by experiment type

○ Protein

● Peptide

○ Gene

○ Search datasets by cancer type

○ Ovarian

○ Breast

Available datasets for analysis

○ TCGA PNNL Ovarian Peptide Data After Parsimony

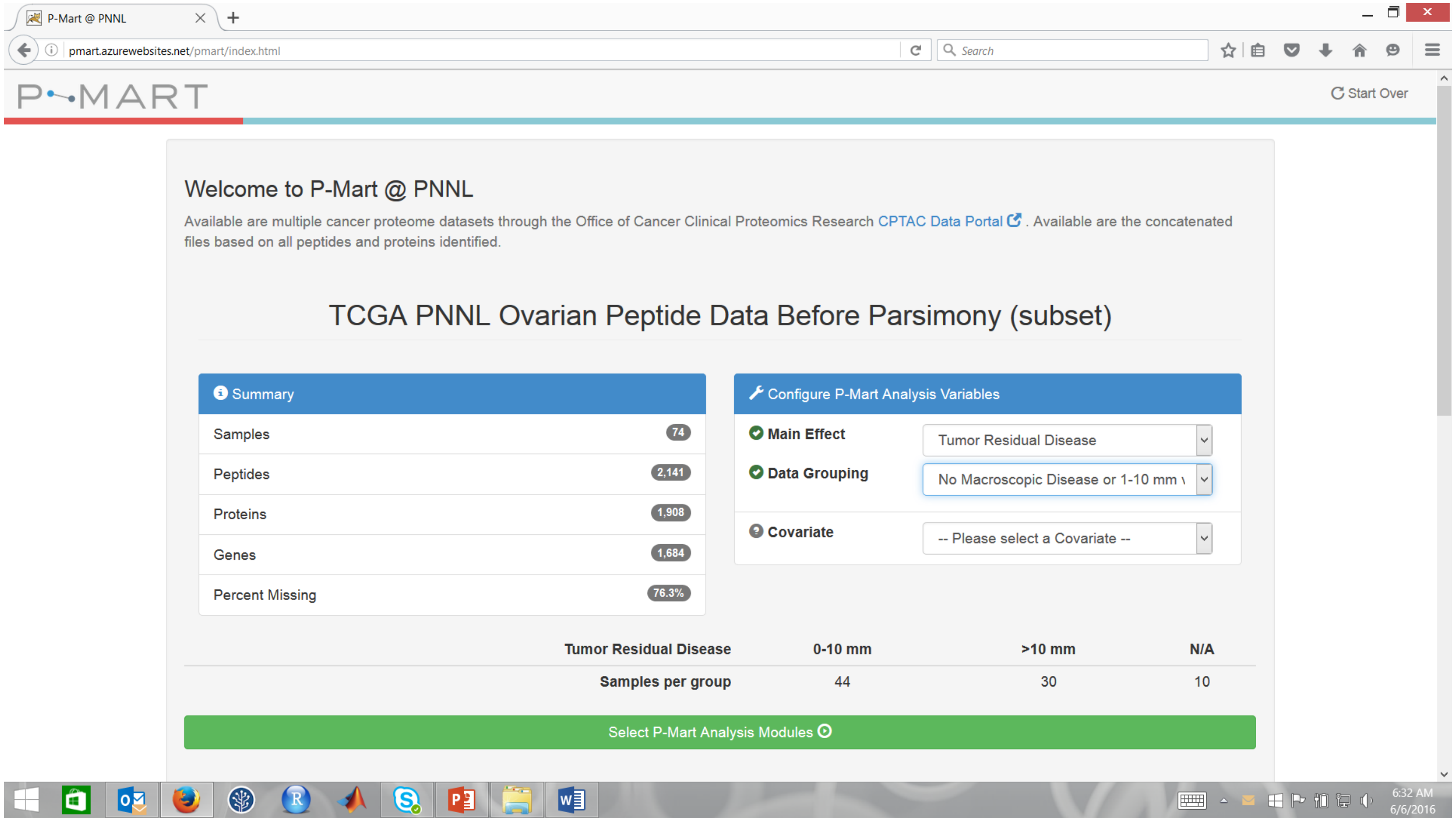
○ TCGA PNNL Ovarian Peptide Data Before Parsimony

● TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)

P-Mart - Data

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

- Summaries of the data available
- All clinical outcomes reported by CPTAC can be used as a primary factor of interest or as a covariate



Welcome to P-Mart @ PNNL

Available are multiple cancer proteome datasets through the Office of Cancer Clinical Proteomics Research CPTAC Data Portal . Available are the concatenated files based on all peptides and proteins identified.

TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)

Summary	
Samples	74
Peptides	2,141
Proteins	1,908
Genes	1,684
Percent Missing	76.3%

Configure P-Mart Analysis Variables

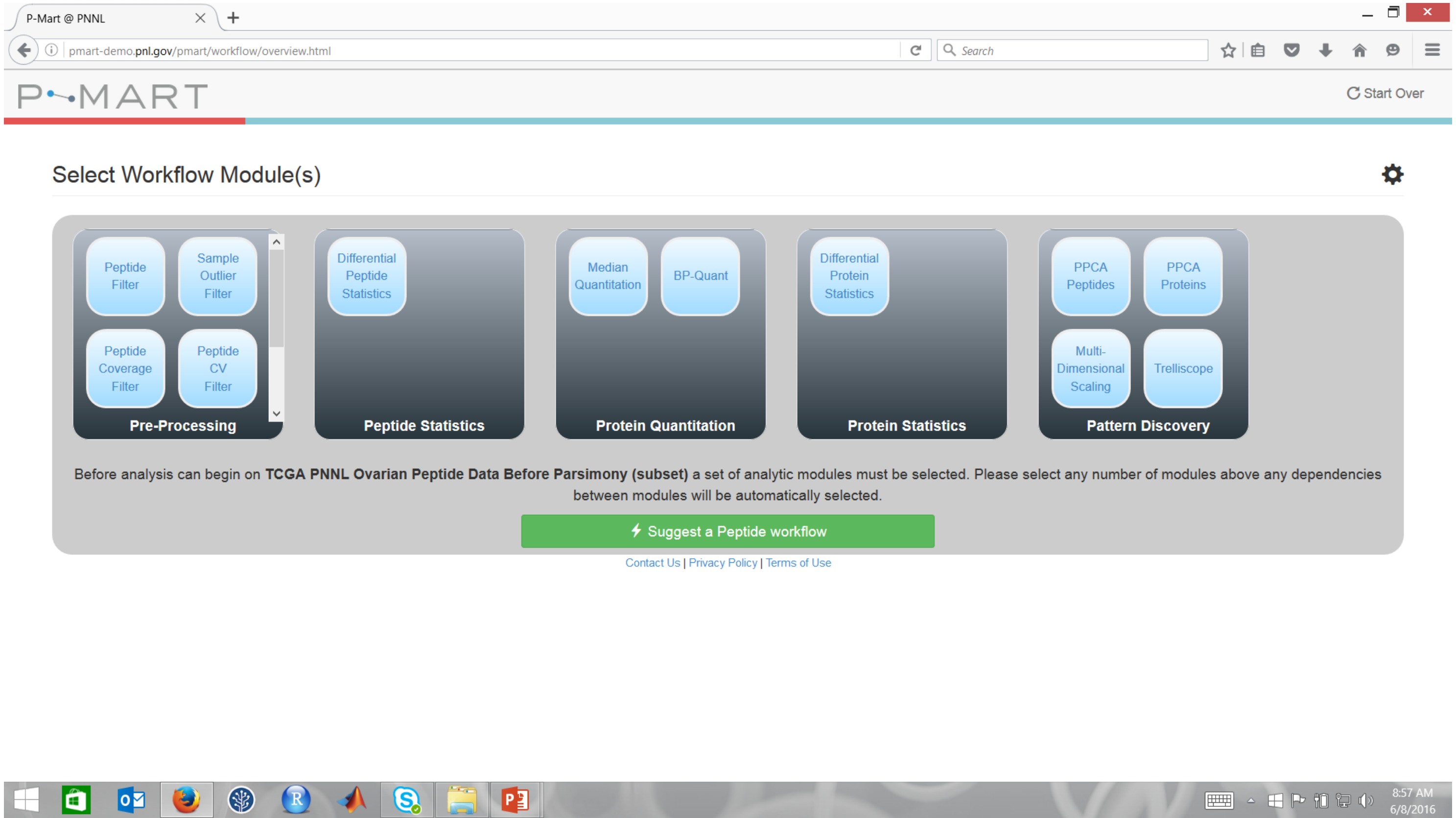
- Main Effect: Tumor Residual Disease
- Data Grouping: No Macroscopic Disease or 1-10 mm \
- Covariate: -- Please select a Covariate --

Tumor Residual Disease	0-10 mm	>10 mm	N/A
Samples per group	44	30	10

Select P-Mart Analysis Modules

P-Mart – Customized Workflows

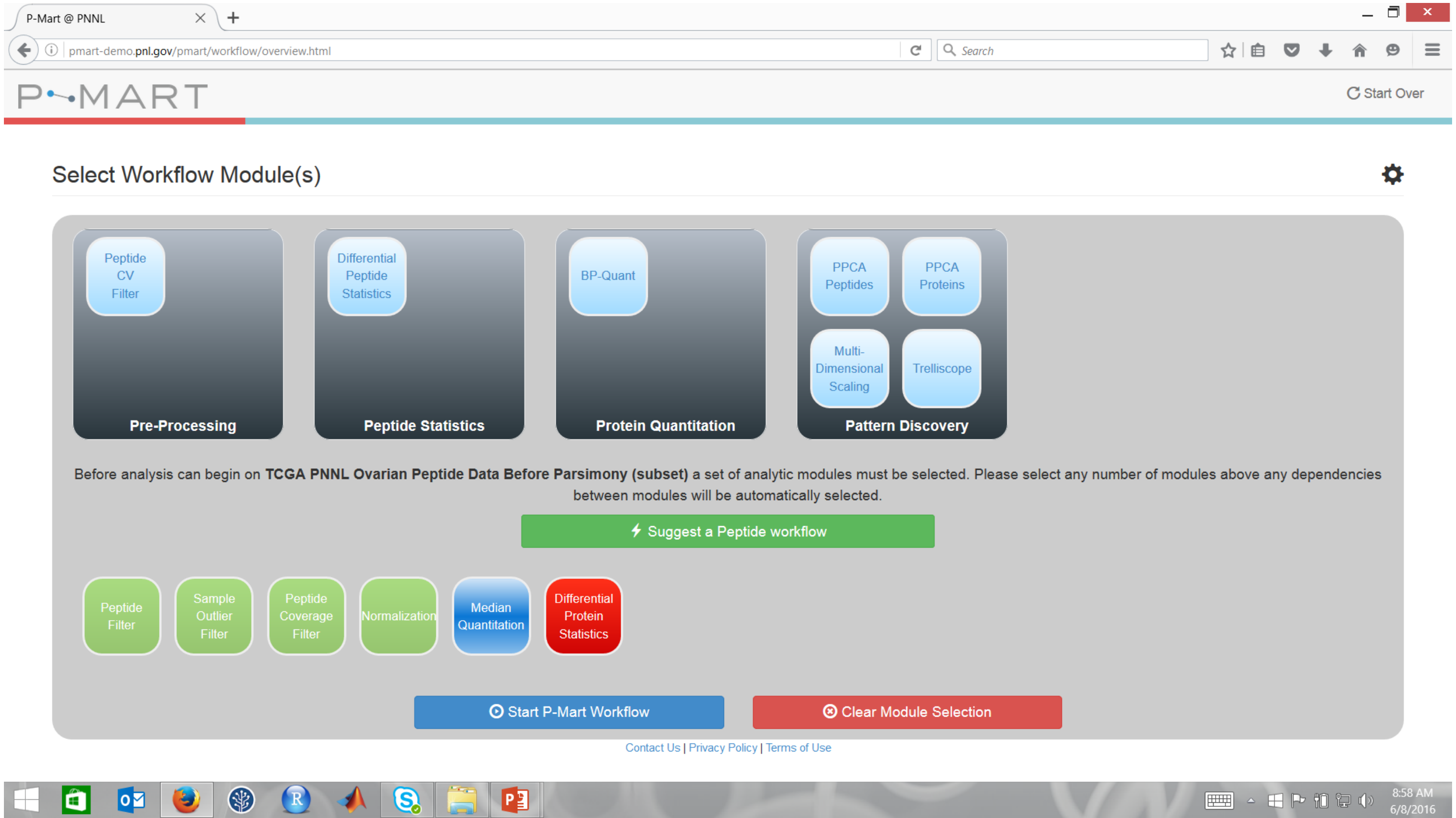
- User may create a workflow (under constraints)
- P-Mart will make a suggestion based on data type



Before analysis can begin on **TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)** a set of analytic modules must be selected. Please select any number of modules above any dependencies between modules will be automatically selected.

P-Mart – Customized Workflows

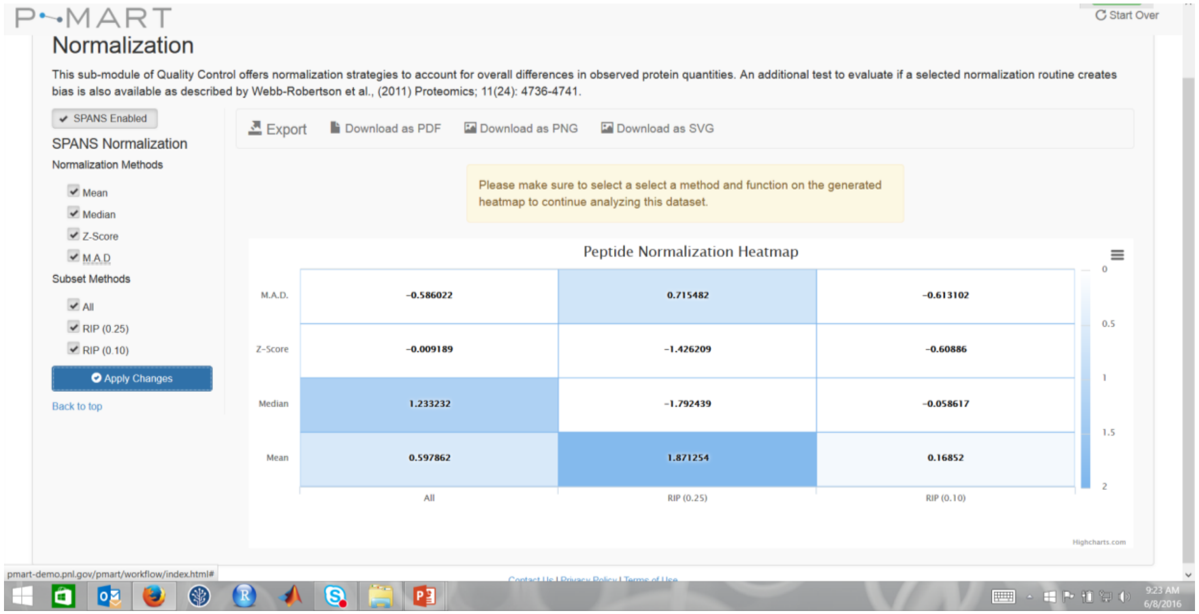
- User may create a workflow (under constraints)
- P-Mart will make a suggestion based on data type



Before analysis can begin on **TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)** a set of analytic modules must be selected. Please select any number of modules above any dependencies between modules will be automatically selected.

P-Mart – Pre-Processing

- Samples
  - Looks at statistical metrics of the distribution of peptide abundances for anomalies
- Peptides/Protein
  - Statistical evaluation of peptide/protein coverage coefficient of variation within peptides
- Samples/Peptides
  - Data normalization approaches evaluated to identify the most robust approaches



Normalization

This sub-module of Quality Control offers normalization strategies to account for overall differences in observed protein quantities. An additional test to evaluate if a selected normalization routine creates bias is also available as described by Webb-Robertson et al., (2011) Proteomics; 11(24): 4736-4741.

☑ SPANS Enabled

SPANS Normalization

Normalization Methods

☑ Mean
☑ Median
☑ Z-Score
☑ M.A.D

Subset Methods

☑ All
☑ RIP (0.25)
☑ RIP (0.10)

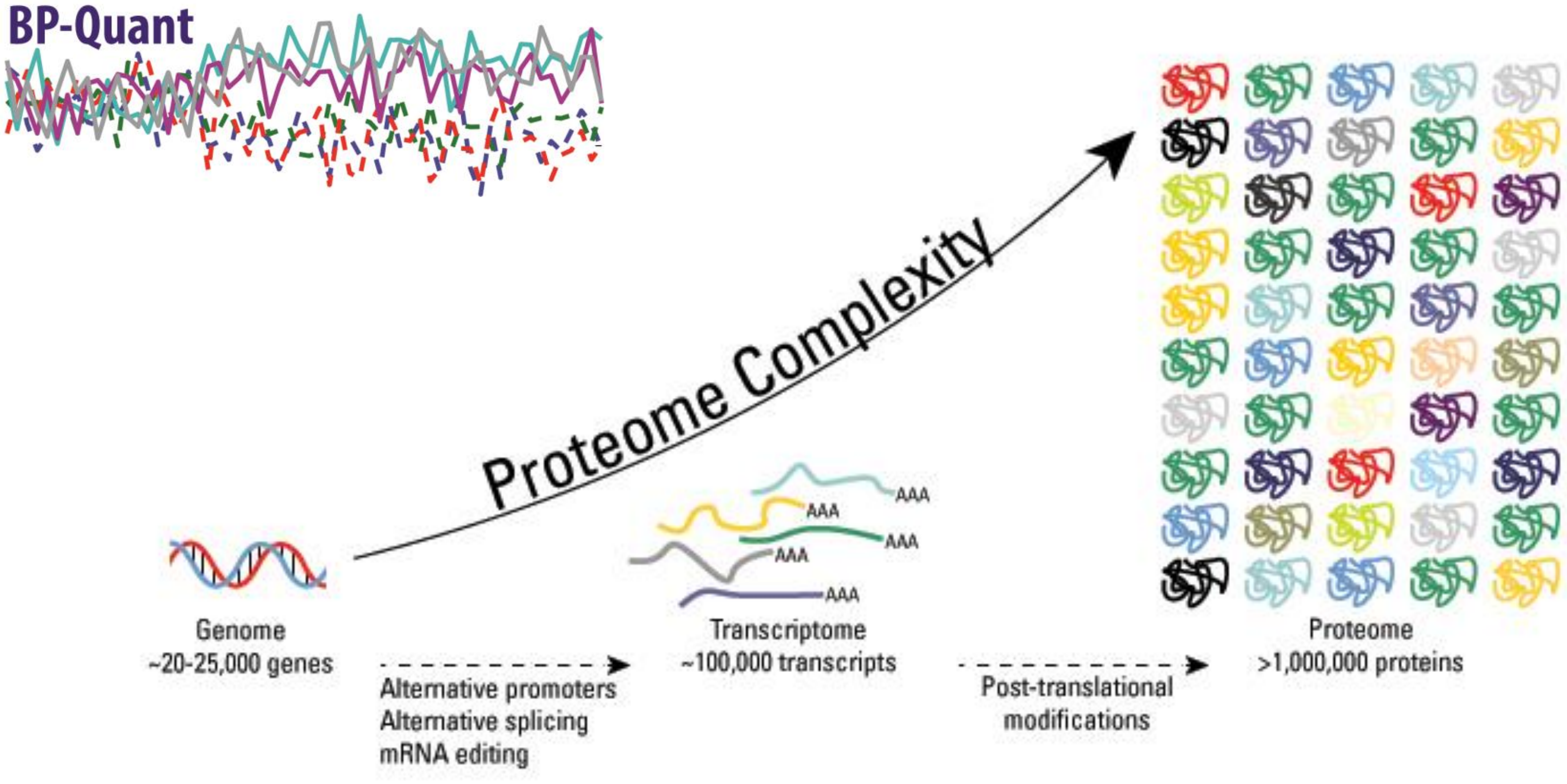
Please make sure to select a select a method and function on the generated heatmap to continue analyzing this dataset.

Peptide Normalization Heatmap

	All	RIP (0.25)	RIP (0.10)
M.A.D.	-0.586022	0.715482	-0.613102
Z-Score	-0.009189	-1.426209	-0.60886
Median	1.233232	-1.792439	-0.058617
Mean	0.597862	1.871254	0.16852

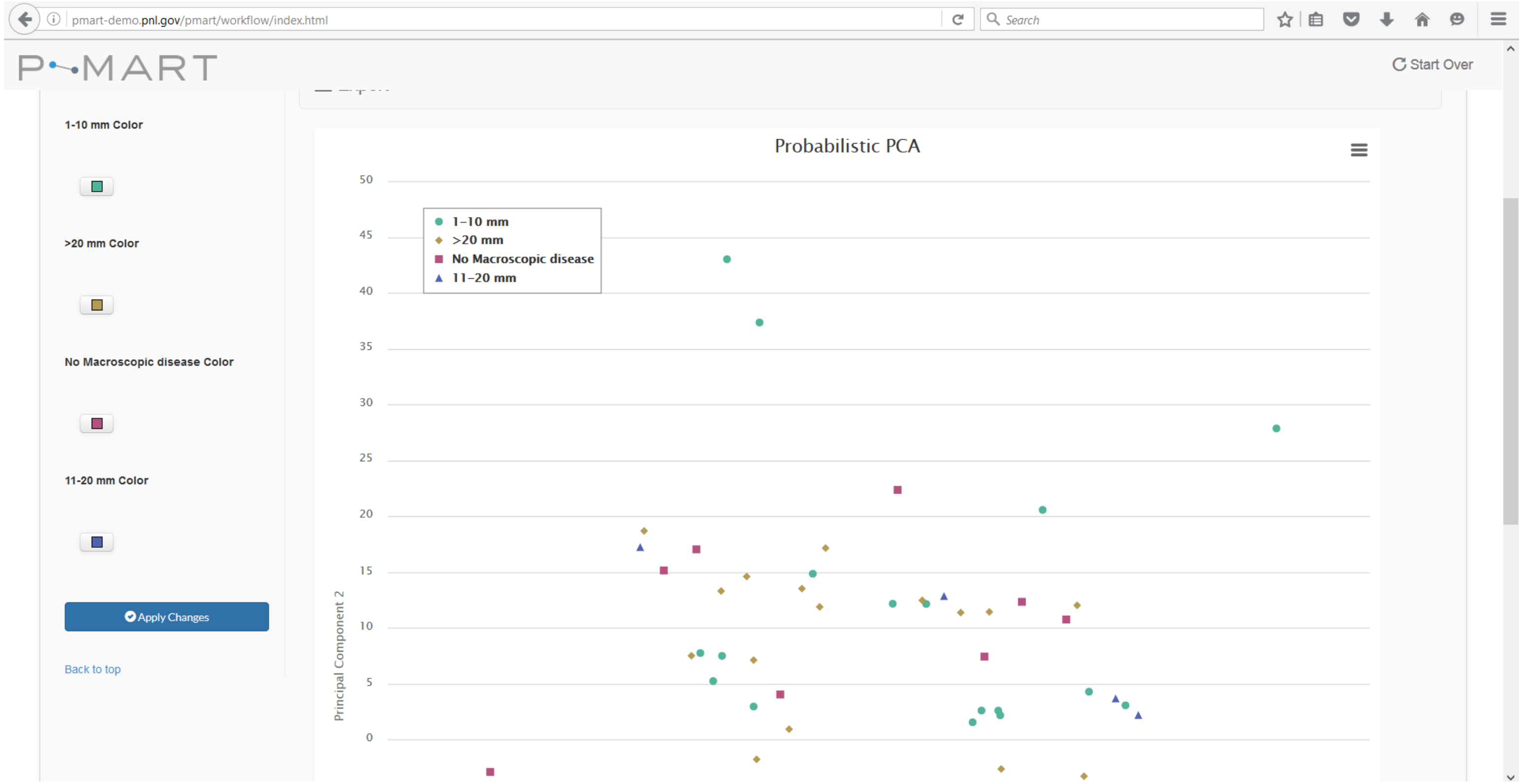
P-Mart – Protein Quantification

- Offers multiple standard protein quantification methods
- Uses unique approaches to identify proteoforms



P-Mart – Exploratory Data Analysis

- Perform PCA in a manner that is robust to missing data
- Allow exploration of potential biomarkers through smart queries defined by the user



1-10 mm Color 🟩

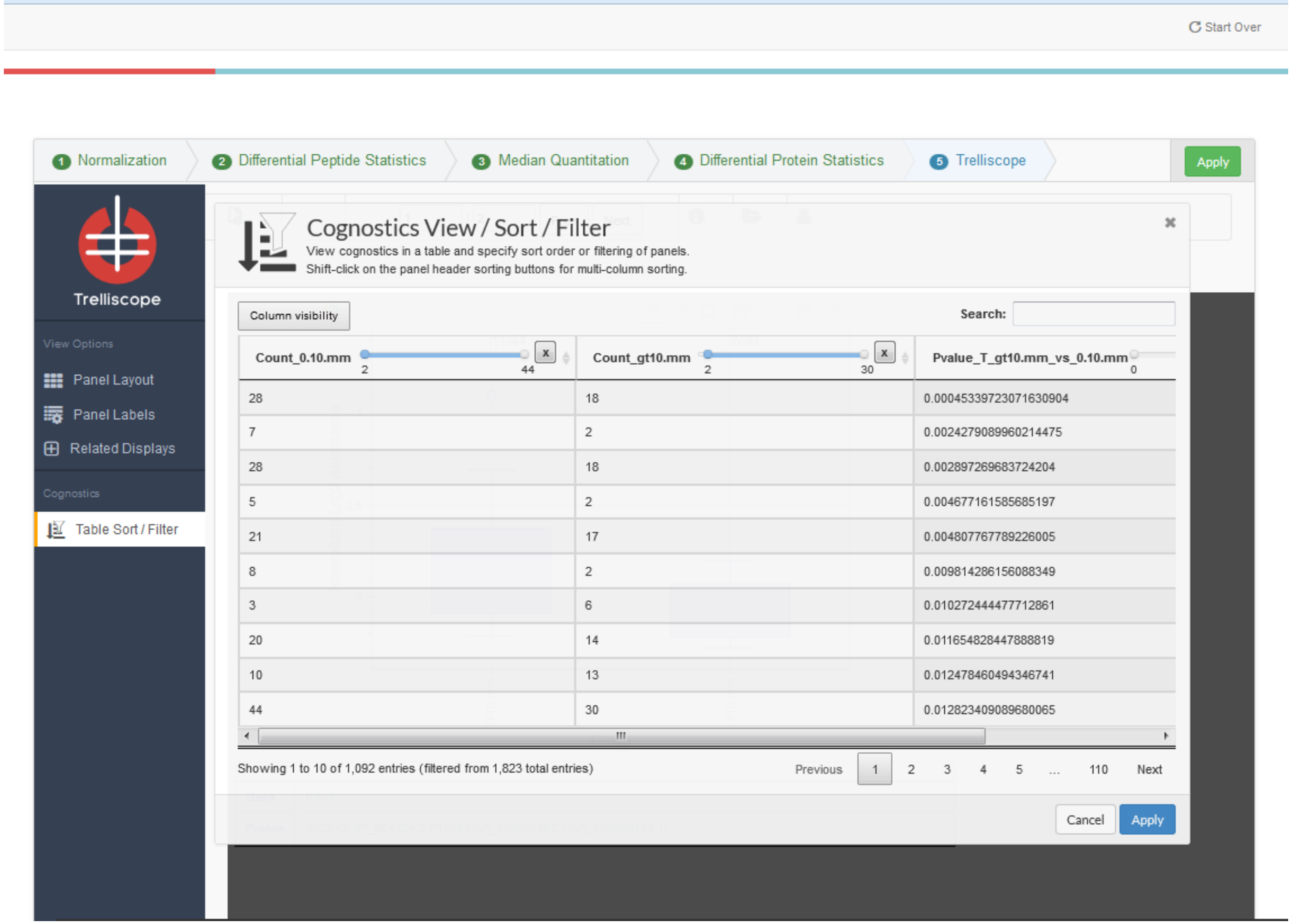
>20 mm Color 🟨

No Macroscopic disease Color 🟥

11-20 mm Color 🟦

P-Mart – Exploratory Data Analysis

- Perform PCA in a manner that is robust to missing data
- Allow exploration of potential biomarkers through smart queries defined by the user



Cognostics View / Sort / Filter

View cognostics in a table and specify sort order or filtering of panels.
Shift-click on the panel header sorting buttons for multi-column sorting.

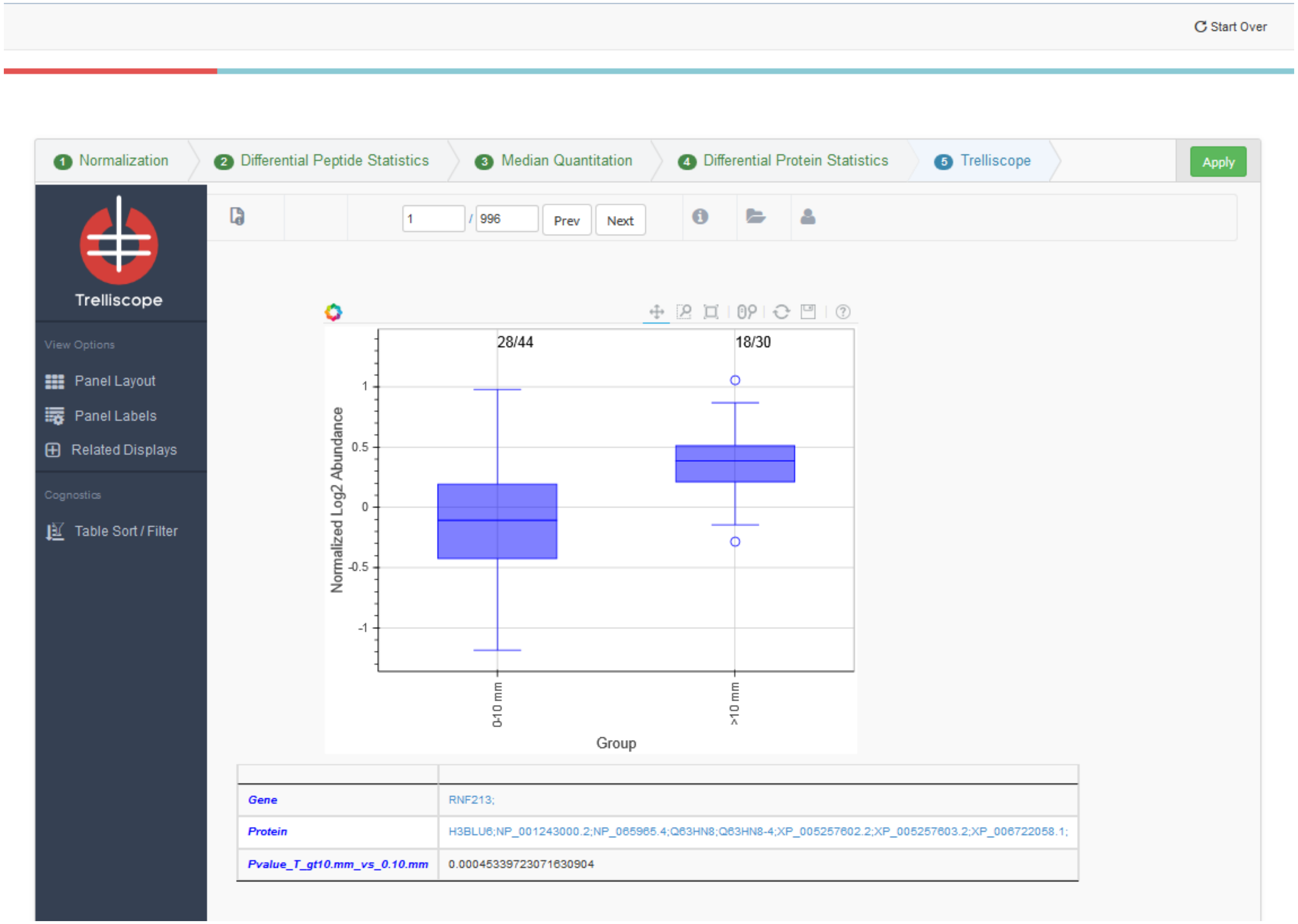
Count_0.10.mm (2 – 44)	Count_gt10.mm (2 – 30)	Pvalue_T_gt10.mm_vs_0.10.mm (0)
28	18	0.00045339723071630904
7	2	0.0024279089960214475
28	18	0.002897269683724204
5	2	0.004677161585685197
21	17	0.004807767789226005
8	2	0.009814286156088349
3	6	0.010272444477712861
20	14	0.011654828447888819
10	13	0.012478460494346741
44	30	0.012823409089680065

Showing 1 to 10 of 1,092 entries (filtered from 1,823 total entries)

Previous 1 2 3 4 5 ... 110 Next

P-Mart – Exploratory Data Analysis

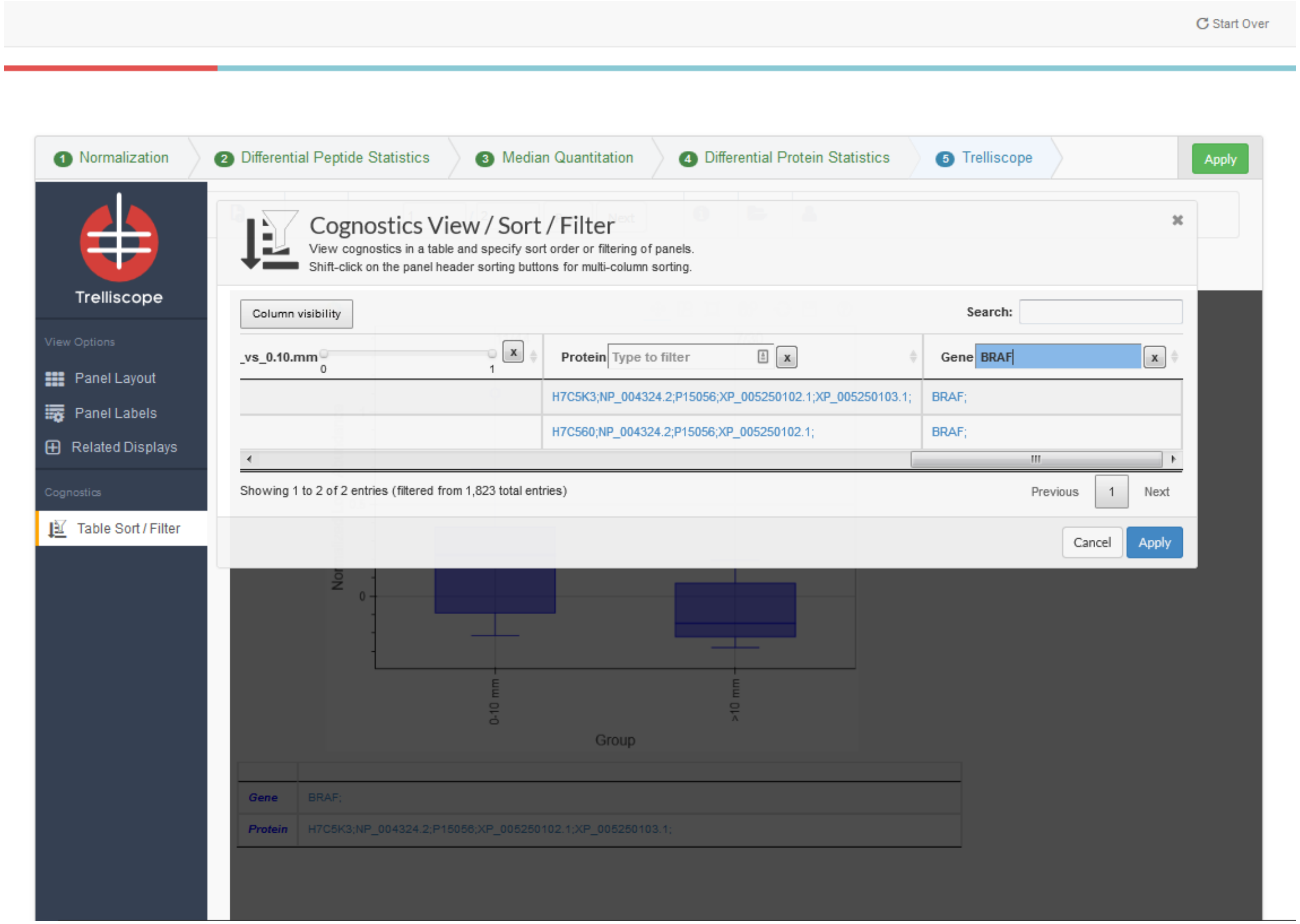
- Perform PCA in a manner that is robust to missing data
- Allow exploration of potential biomarkers through smart queries defined by the user



Gene	RNF213;
Protein	H3BLU6;NP_001243000.2;NP_065965.4;Q63HN8;Q63HN8-4;XP_005257602.2;XP_005257603.2;XP_006722058.1;
Pvalue_T_gt10.mm_vs_0.10.mm	0.00045339723071630904

P-Mart – Exploratory Data Analysis

- Perform PCA in a manner that is robust to missing data
- Allow exploration of potential biomarkers through smart queries defined by the user



Cognostics View / Sort / Filter

View cognostics in a table and specify sort order or filtering of panels.
Shift-click on the panel header sorting buttons for multi-column sorting.

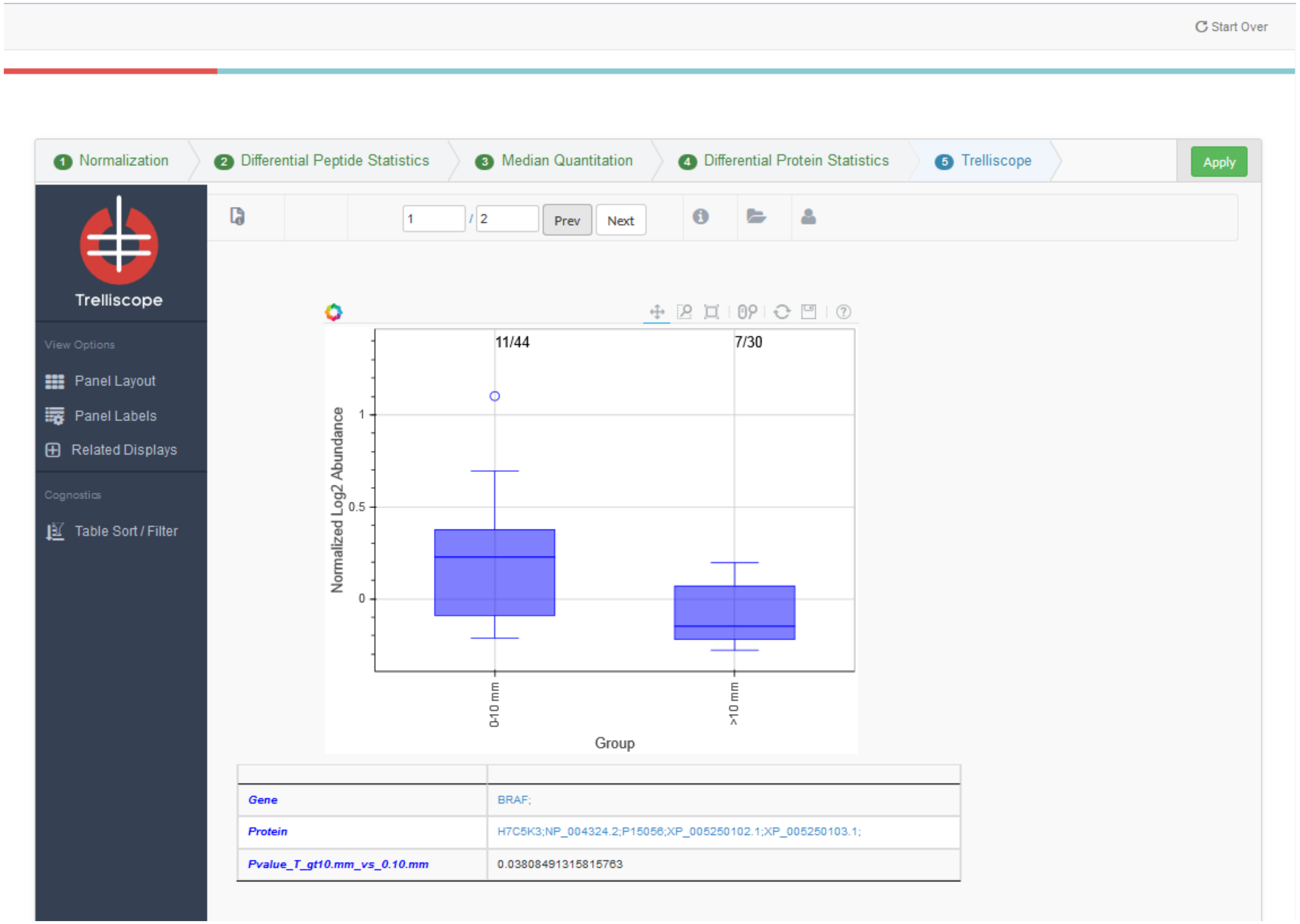
_vs_0.10.mm (0 – 1)	Protein (Type to filter)	Gene (BRAF)
	H7C5K3;NP_004324.2;P15056;XP_005250102.1;XP_005250103.1;	BRAF;
	H7C560;NP_004324.2;P15056;XP_005250102.1;	BRAF;

Showing 1 to 2 of 2 entries (filtered from 1,823 total entries)

Gene	BRAF;
Protein	H7C5K3;NP_004324.2;P15056;XP_005250102.1;XP_005250103.1;

P-Mart – Exploratory Data Analysis

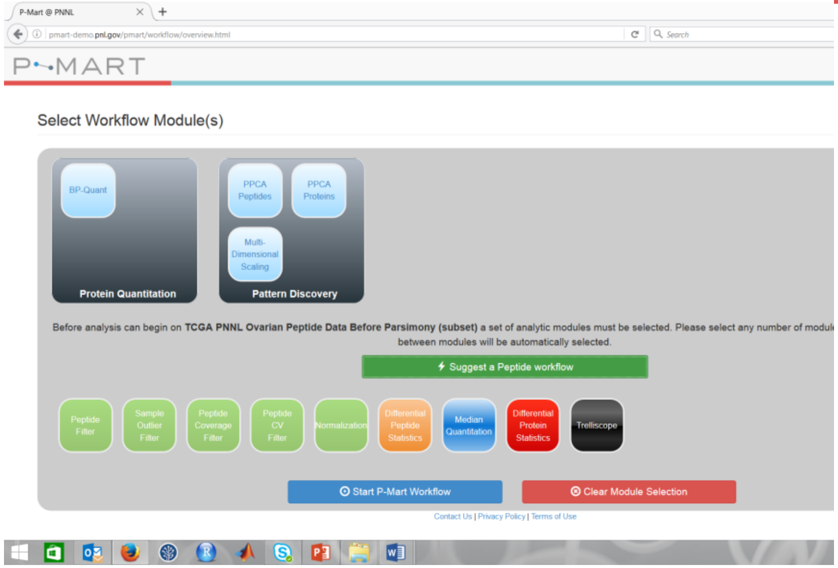
- Perform PCA in a manner that is robust to missing data
- Allow exploration of potential biomarkers through smart queries defined by the user



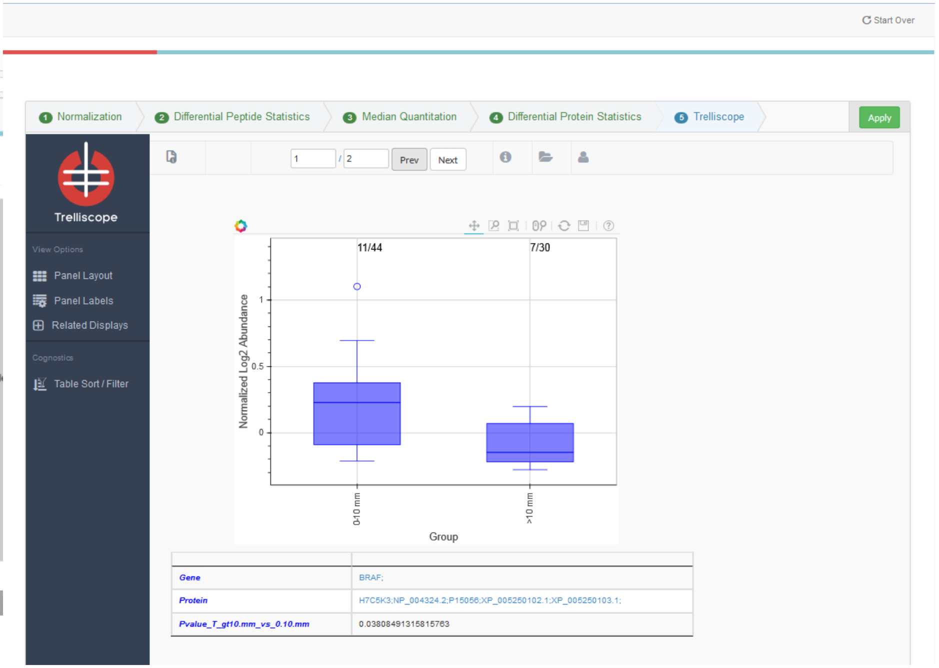
Gene	BRAF;
Protein	H7C5K3;NP_004324.2;P15056;XP_005250102.1;XP_005250103.1;
Pvalue_T_gt10.mm_vs_0.10.mm	0.03808491315815763

P-Mart Outcomes

- Easy to use online software driven by visual interactions
- Customizable workflows documented for reproducibility
- Open-source code of statistical methods (R)
- Easy exports of data at multiple levels of processing



Before analysis can begin on **TCGA PNNL Ovarian Peptide Data Before Parsimony (subset)** a set of analytic modules must be selected. Please select any number of modul
between modules will be automatically selected.



Gene	BRAF;
Protein	H7C5K3;NP_004324.2;P15056;XP_005250102.1;XP_005250103.1;
Pvalue_T_gt10.mm_vs_0.10.mm	0.03808491315815763

P-Mart Status

- Completed
  - Website deployed with Azure cloud technology
    http://pmart.azurewebsites.net/pmart/index.html
  - Statistical methods robust to missing data released on GitHub
    https://github.com/MSomics-StatTools/MSomicsQC
  - Peptide and Protein level statistics and associated meta-data can be queried to rapidly identify biomarkers of interest
- In Progress
  - Documentation allows for reproducible analyses
  - Exports allow the user access to data at various levels of quality control and processing to enable downstream analyses (e.g., pathway analysis)
  - Machine learning aides in validation and feature selection tasks to select candidates for validation

Acknowledgements 1U01CA184783-01

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

P-MART Team

Bobbie-Jo Webb-Robertson

bj@pnnl.gov

Lisa Bramer

Kelly Stratton

Markus Kobold

Jeffrey Jensen

Mathew Thomas

Amanda White

Clinical Proteomic Tumor Analysis Consortium

VANDERBILT UNIVERSITY

BROAD INSTITUTE

ESAC
Enterprise Science And Computing

Pacific Northwest
NATIONAL LABORATORY

Early Detection Research Network
Biomarkers: the key to early detection

NASA
Jet Propulsion Laboratory
California Institute of Technology