Implementing a Proteomics Data Pipeline and Database on LabKey to promote in-depth analysis, data sharing & integration

Wen Yu, Jonathan Pryke, Gina Dangelo, Raghothama Chaerkady, Sonja Hess, Adolf Brown, Paula Gegwich and David Fenstermacher

Research Bioinformatics, RD&I, Statistical Science and AD&PE

MedImmune LLC, One MedImmune Way, Gaithersburg, Maryland 20878 USA

MedImmune

Enabling Data-driven Discovery

- Systematic collection of quantitative molecular phenotypes to probe what has happened;
- Focused experimental design with specified outcomes;
- Big-data approaches leading to deep, holistic, perhaps unexpected understanding of the system and biology.
- DATA → Understanding → Predictive

engineering	→ LabKey	→ *assay, execution*
analytics	→ algorithms	→ *interpretation*
vis/integration	→ network biology	→ *big picture, discovery.*

Genotypes
(Genomics, Genetics)

Proteins, Peptides,
Lipids & Metabolites

Signals, Causality
Targets, Biomarkers

MedImmune

NextGen Proteomics: ~ Complete Quantitation

FASP trypsin digest
TMT Labelling/mux

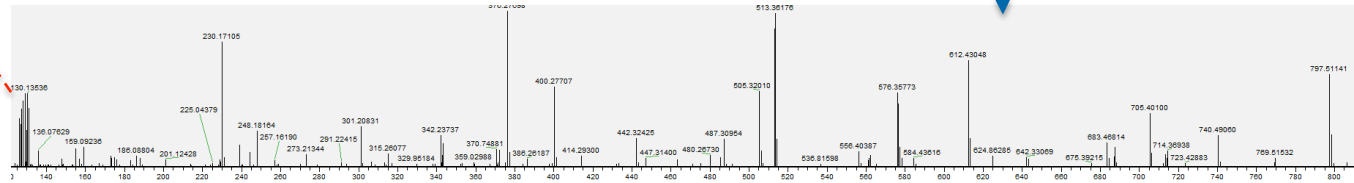
Sample	10 TMT	Sample	20 TMT
1	126	11	126
2	127N	12	127N
…	…	…	…
	131		131

Thermo Scientific Orbitrap Fusion Tribrid LC-MS (I) with Thermo Scientific Dionex UltiMate 3000 Series UHPLC.

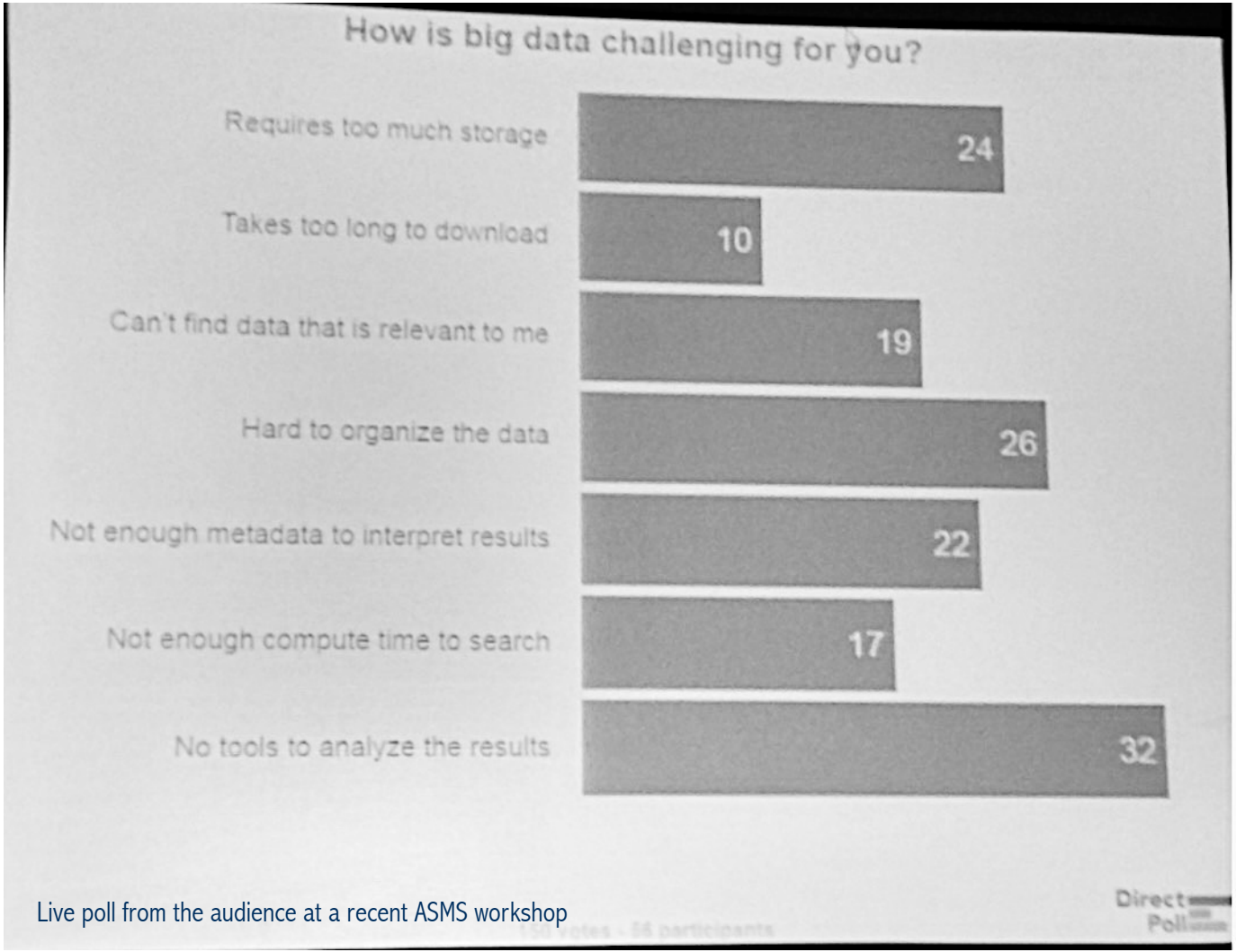
OrbiTrap Fusion Tribrid

~8,000 (cells, tissue) or >1500 (plasma) proteins identified & quantified

480k MS/MS reads / 10-samples



MedImmune



MedImmune

Live poll from the audience at a recent ASMS workshop

Internal & public Proteomics Studies

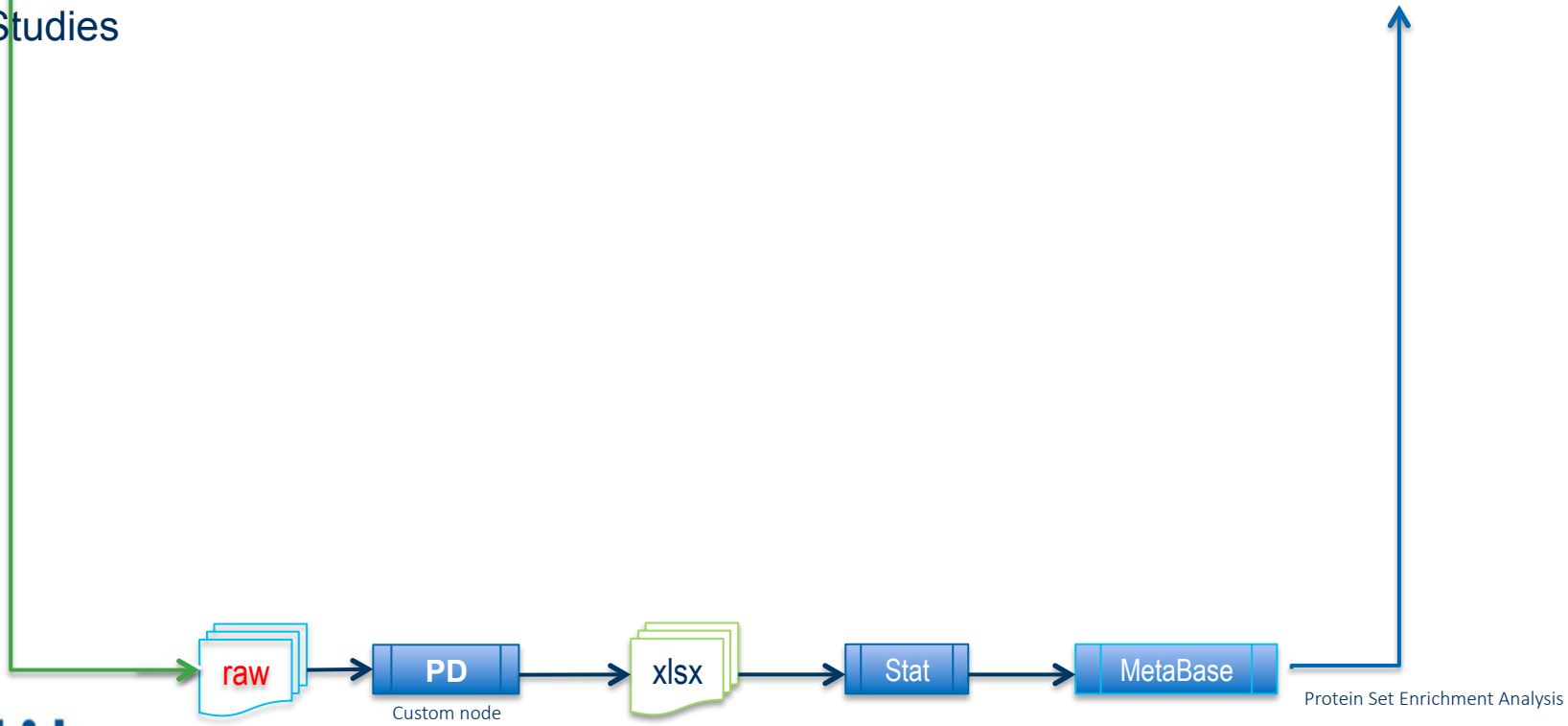
Cell-surface Proteins

Protein Expression Profiles or Signature

Reports

Mechanism of Action

Biomarkers



MedImmune

Phase 3

Internal & public Proteomics Studies

Cell-surface Proteins

Protein Expression Profiles or Signature

Reports

Mechanism of Action

Biomarkers

The Proteomics Data Pipeline

Signature,
Protein Matrix
Proteins, Peptides
Abundances, Spectra

Expt, Sample, Study

RDB

Protein Matrix

Sample	MEM	A	B
Protein 1	*yes*	*expr*	*expr*
Protein 2	*no*	*expr*	*expr*
…			

Visualization

QC Metrics
Supporting Evidences

Stat

Unsupervised classification

FC & pval / fdr

Signature

Pathway analysis

Network of overlapping gene sets

(2) Msf files Import

raw

PD

Custom node

Protein Set Enrichment Analysis

MedImmune

Phase 3

Picking a Solution, aka House Hunting

Fully Furnished Estate

Commercial systems
with built-in analysis pipeline

Custom Home

Proprietary application

MedImmune

The Pipeline Design: the DIY Edition

Component	Function	Solution
Study Design	Sample annotation and experimental factors	Proteome Discoverer
Protein ID and Quant	Peptide/protein ID, reporter ion intensities	
Data Preparation	Protein quants, imputation, data cleansing & formatting	**Custom R-scripts**
Statistical modeling	Ad hoc and formal analysis for the significant changes in protein abundances	R/SAS, **LabKey**
Data Visualization	Delivery to the investigators for access and exploration. Ad hoc experimentation.	R/Shiny, **LabKey**
Pipelining	Streamlining the workflow	**LabKey**
Data management	Repository, project tracking, visualization, R-integration	**LabKey**
Pathway, G/PSEA	Biological contextualization and hypothesis generation	R/Shiny

Solutions in the Era of Open Source

LabKey, R/Shiny

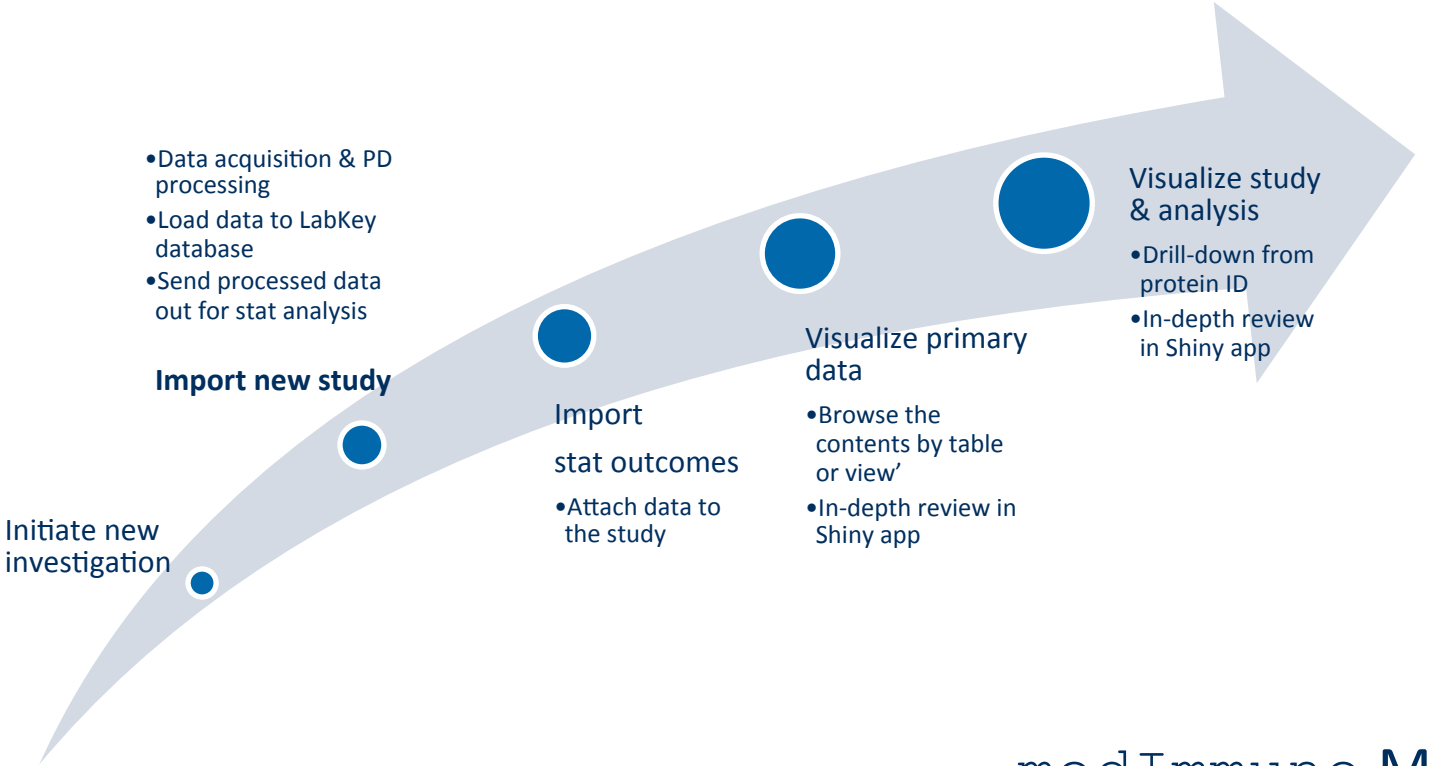
Free Widgets
DIY

Designers
Builders

Free new construction

MedImmune

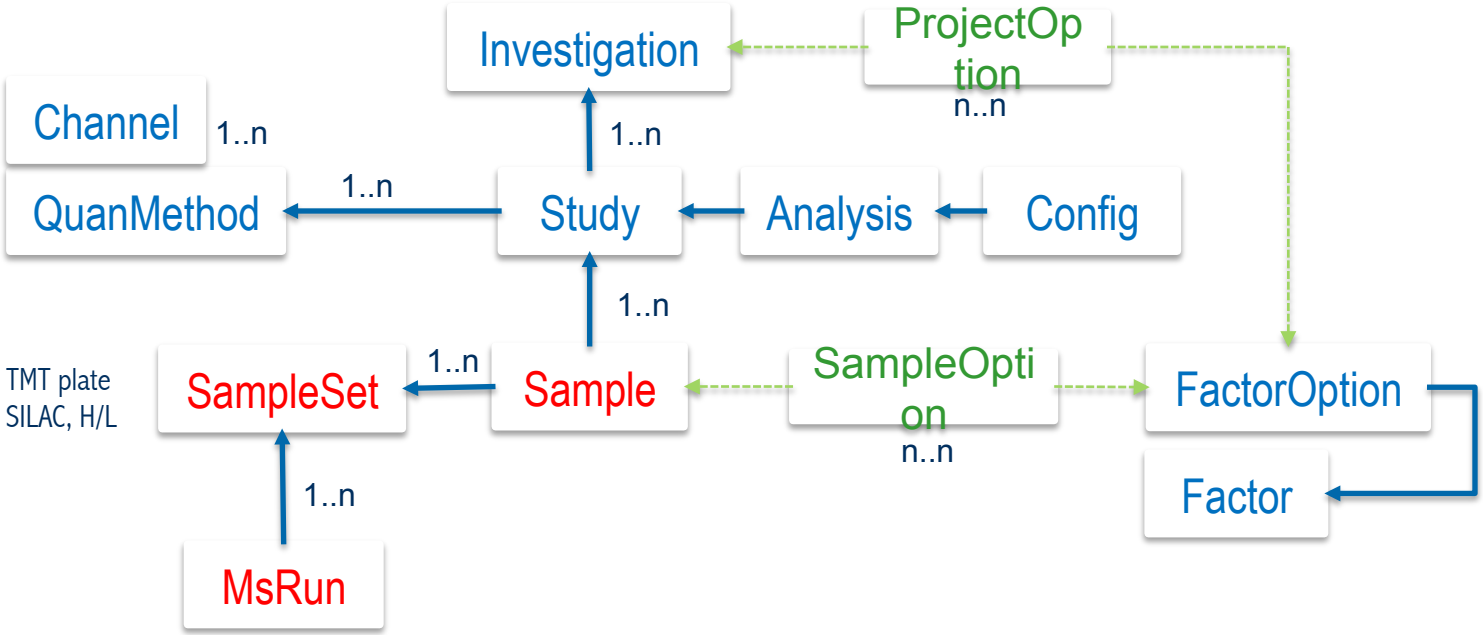
LabKey Implementation: Data Import



MedImmune

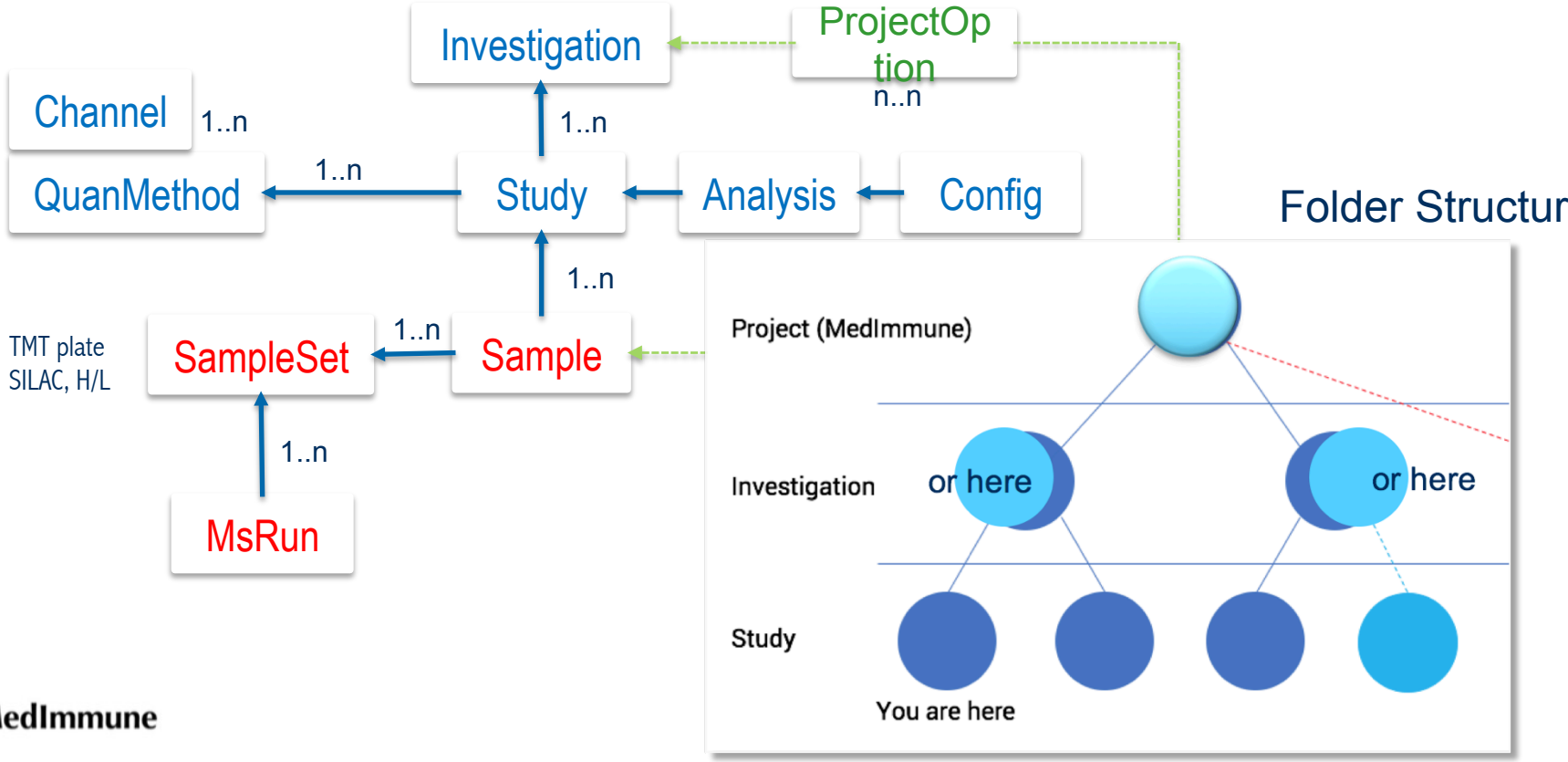
medImmune Module

Data Model: Meta Data



MedImmune

Folder Structure vs Meta Data Hierarchy



MedImmune

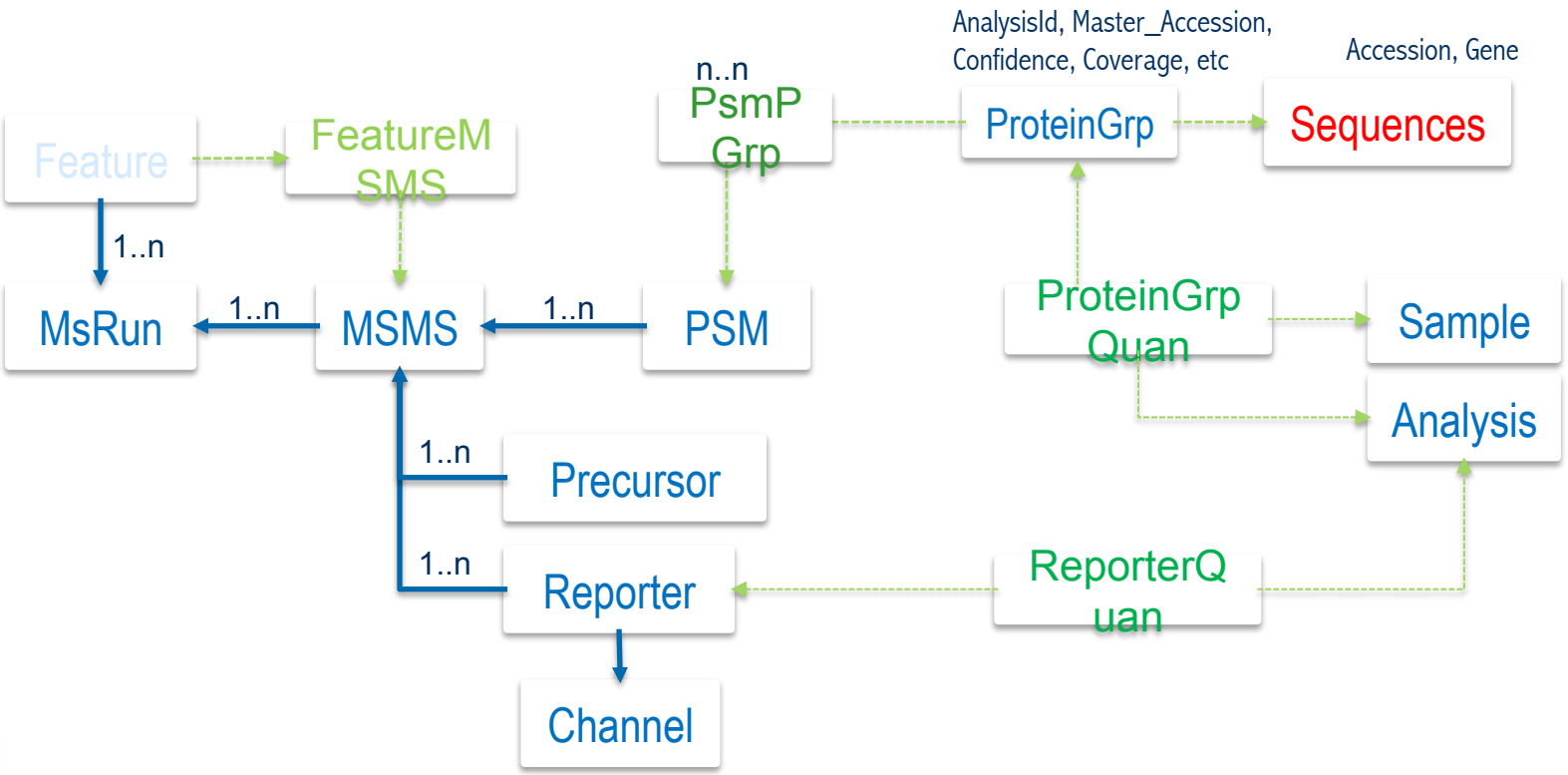
Data Loading Logics: Samples

- Populate the ‘Factor’ and ‘Option’ tables if necessary as described previously
- Create entries in ‘SampleOption’ table that link samples to options
- Cross-reference the ’Channel’ table via ‘Channel’ field and populate the ‘ChannelId’

Channel	PlateName	Plate	Group	Cohort	Replicate	ChannelName
X126	151124_ _TMT1_MS2_F1	1	NS		1	X126
X126	151124_ _TMT2_MS2_F	2	NS		3	X126
X127_C	151124_ _TMT2_MS2_F	2	LPS		3	X127C
X127_C	151124_ _TMT1_MS2_F1	1	LPS		1	X127C
X127_N	151124_ _TMT1_MS2_F1	1	NS		2	X127N
X127_N	151124_ _TMT2_MS2_F	2	NS		4	X127N
X128_C	151124_ _TMT1_MS2_F1	1	LI		1	X128C
X128_C	151124_ _TMT2_MS2_F	2	LI		3	X128C
X128_N	151124_ _TMT2_MS2_F	2	LI		2	X128N
X128_N	151124_ _TMT1_MS2_F1	1	LPS		2	X128N
X129_C	151124 _TMT2_MS2_F	2	LA		2	X129C

MedImmune

Data Model: MS Data



MedImmune

MSMS, PSM, PsmPGrp

- For each row in PSM.all.Rdata
  - Grab the FK to 'MsRun' with the 'Run' field
  - Create an entry in 'MSMS' if the Run/Scan combination is not present already
  - Create entry in 'Precursor' table with FK to 'MSMS'
  - Create entries in 'Reporter' table for each channels with FK to 'Channel' rows sharing the same 'ChannelName'
  - Create an entry in 'PSM' table with FK to the 'MSMS' entry
  - Grab the FK to 'ProteinGrp' with the 'Accessions' and 'AnalysisId' fields;
  - Create an entry in 'PsmPGrp' table with FK to PSM and ProteinGrp.

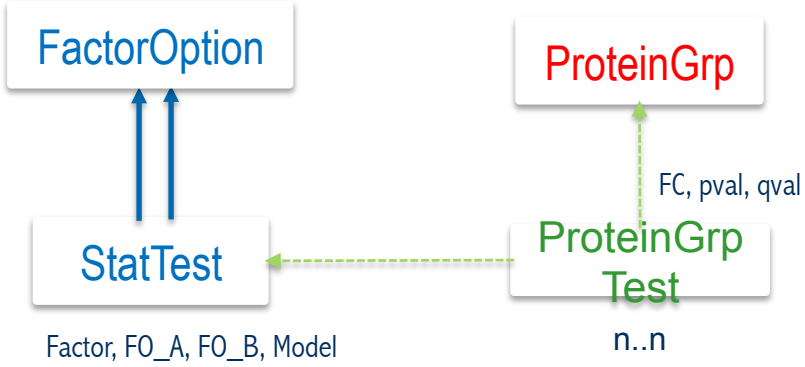
Protein	Confidence	Peptide	Mods	Charge	Charge.obs	Rank
	High	[R].	N-Term(TMT6plex); K8(TMT6plex)	3	3	1
	High	[R].]	N-Term(TMT6plex)	4	4	1
	High	[R].	N-Term(TMT6plex); K10(TMT6plex)	4	4	1
	High	[R].	N-Term(TMT6plex)	2	2	1

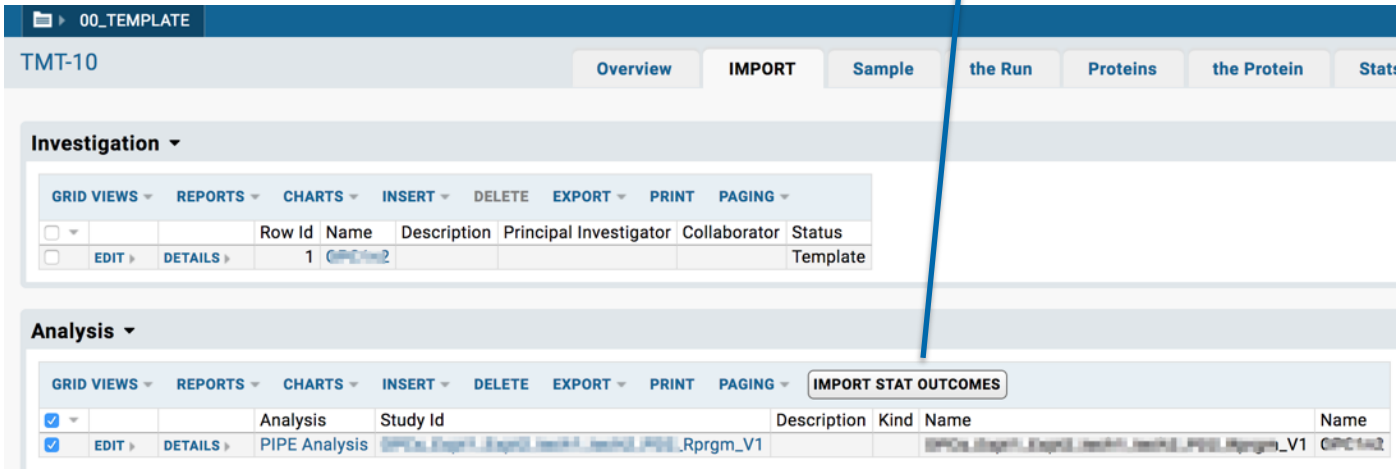
mz	MH	ppm	Intensity	SN.Reporter	Injection	RT	qVal	Usage	Info	Sequence	Accessions	Interference	Scan
481.2494	1441.734	0.8534145	173030112	9.59	1	13.14821	0.001	Use	Unique	D		5.807800	1578
433.2587	1730.013	-0.3717548	3189741	28.87	33	13.24840	0.001	Use	Unique	R		8.217509	1614
443.0070	1769.006	-0.6066390	5019254	287.87	33	13.45662	0.000	Use	Unique	S		7.326426	1690
567.3093	1133.611	0.1079945	18520848	10.33	5	13.47412	0.002	Use	Unique	A		3.778577	1695
581.6573	1742.957	1.3071											
601.0036	1800.996	-0.3560											

Run	X126	X127N	X127C	X128N	X128C	X129N	X129C	X130N	X130C	X131
151124. TMT1_MS2_F1.raw	7.764347	2.351445	15.84418	12.49247	6.413736	NA	12.13688	19.663802	8.552111	7.775993
151124. TMT1_MS2_F1.raw	18.375119	13.361465	50.76977	61.27787	24.895180	NA	21.25711	52.323013	18.376584	17.905292
151124. TMT1_MS2_F1.raw	169.340902	90.950866	556.83933	553.17669	233.141770	26.5068669	224.81314	562.603054	175.340205	176.183739
151124. TMT1_MS2_F1.raw	10.360915	3.758381	16.10108	10.62974	6.913472	NA	18.67190	9.994884	11.368426	12.230057
151124. TMT1_MS2_F1.raw	27.564634	16.196917	52.95002	51.94636	29.482079	0.2259693	36.25498	60.995901	34.147985	29.769754
151124. TMT1_MS2_F1.raw	32.085319	11.529672	98.70601	95.54057	47.368985	NA	31.44710	109.052791	33.864875	28.996786

MedImmune

Data Model: Quantitative Outcomes





MedImmune

medImmune Module and TMT Template

- All database tables and server-side logics are implemented in a new MedImmune module
- UI layout and interactive reporting are written as a "template" folder, which can then be used to create a new "investigation".

Overview | IMPORT | Sample | MsRun | Proteins | Selected Protein | Stats

MedImmune

Name:

My new project

☑ Use name as title

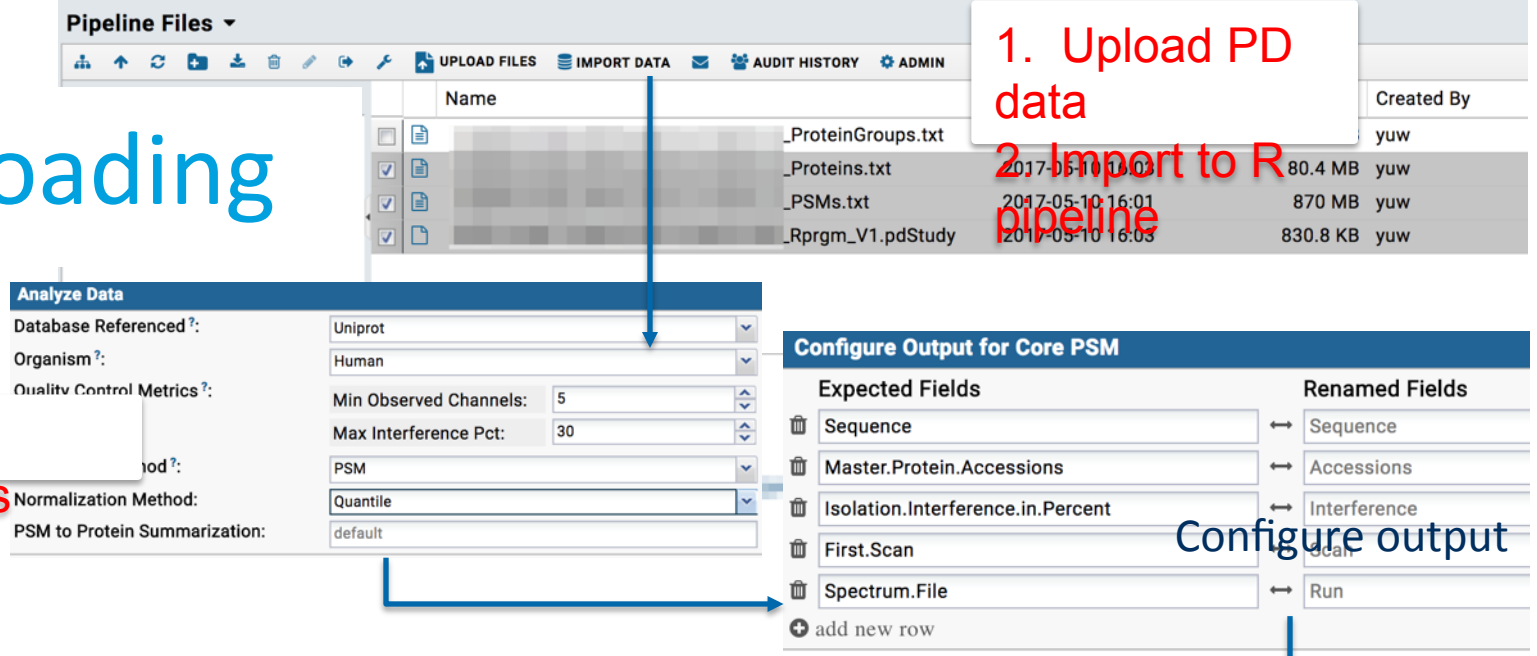
Folder Type:

- Assay
- Collaboration
- Flow
- MS1
- MS2
- MedImmune
- Microarray
- Panorama
- Study
- Custom
- Create From Template Folder

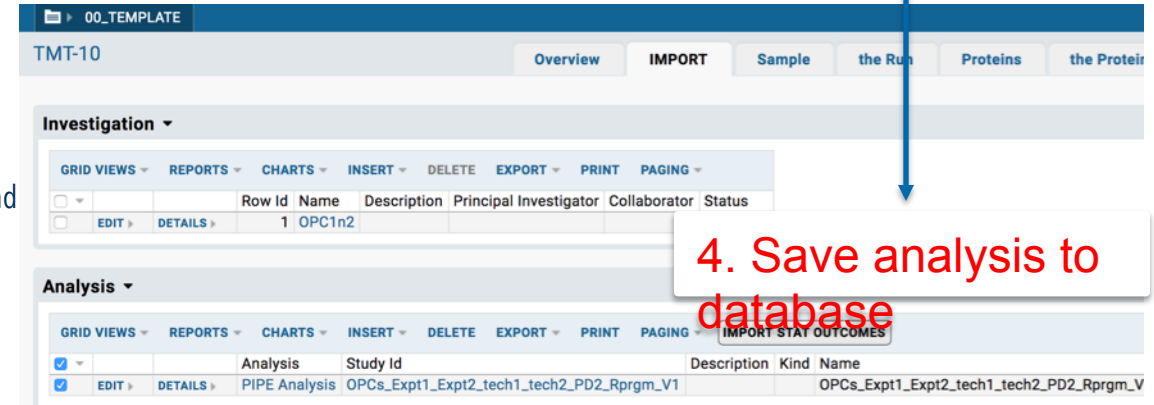
Choose Template Folder:

/00_template
/00_template/TMT
/00_template/TMT with Stats

Data Loading

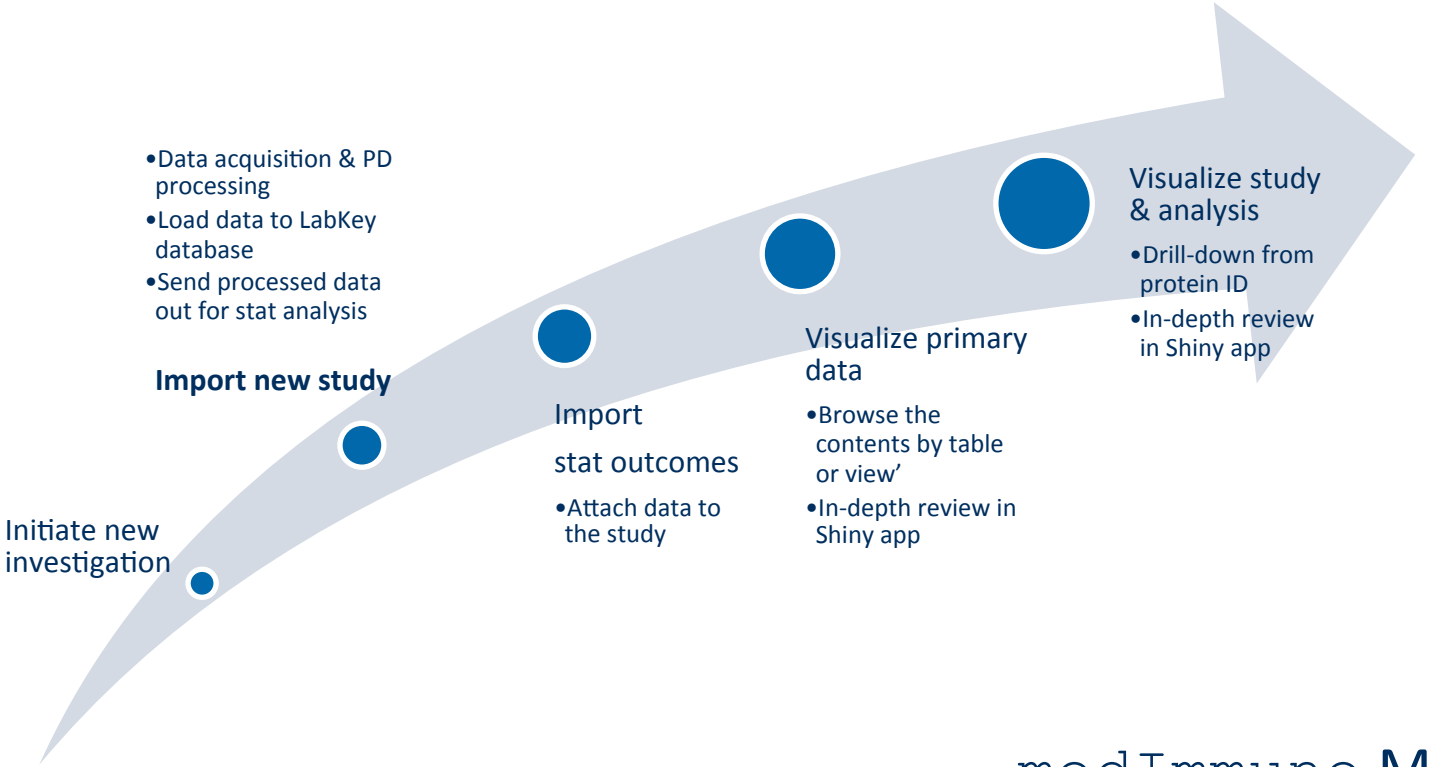


One of the first workflow being implemented is TMT-based multiplex quantitation of the total proteome. Following the data acquisition and processing in ProteomeDiscoverer, the experimental design, peptide and protein identification and quantitation are imported into LabKey where a custom-built data ingestion pipeline written in R will transform the data and prepare them for deposition in a Microsoft SQL database.



MedImmune

LabKey Impl: Results and Visualization



MedImmune

medImmune Module

Study Portal

TMT with Stats

Overview | IMPORT | Sample | MsRun | Proteins | Selected Protein | Stats | About | How Do I...

Investigation - Study - Analysis

GRID VIEWS | REPORTS | CHARTS | EXPORT | PRINT | PAGING

Investigation	Description	Created On	Study	Analysis	Kind	Analyzed On	Proteins
TMT with Stats -- Template		2017-04-11 20:02	[redacted]	PIPE Analysis	Initial Import	2017-04-11 20:13	8845

Data Views

name, category, etc.

Mine

Name	Details	Access
Uncategorized		
Volcano Example Test		public
bwplot_ProteinVsCohort		private
Protein.Abund.Plot		public
Peptide.PSM.Plot		public
Protein.Reporters		public

Protein Search

Protein name *?

Minimum prob?

Search in subfolders?

Exact matches only?

SEARCH

Search

SEARCH

<Select Web Part>

ADD

MedImmune

DIY UI and Report

- Data grid + custom query offers flexible way to review the dataset
- Custom URL enable master-detail view into complex hierarchical data

Customize Query

Web Part Title:	MsRun Summary
Schema:	medimmune
Query and View:	○ Show the list of queries in this schema. ◉ Show the contents of a specific query and view.
Query:	MsRunSummaryView
View:	[default view]
Allow user to choose query?:	No
Allow user to choose view?:	Yes
Button bar position:	Top

```
SELECT MsRunSumPSM.RowId,
   MsRunSumPSM.Investigation,
   MsRunSumPSM.Study,
   MsRunSumPSM.SampleSet,
   MsRunSumPSM.Run,
   MsRunSumPSM.Plate,
   MsRunSumPSM.Container,
   MsRunSumPSM.PSMs,
   MsRunSumProtein.Proteins
FROM MsRunSumPSM
    JOIN MsRunSumProtein on MsRunSumPSM.RowId = MsRu
```

/project/${container}/begin.view?pageId=MsRun%20%28New%29&qwp1.param.RunParam=${RowId}
&qwp2.param.RunParam=${RowId}&qwp3.param.RunParam=${RowId}&qwp4.param.RunParam=${RowId}
&qwp5.param.RunParam=${RowId}

MsRun Summary

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING 1 - 25 of 72 Next › Last »

Investigation	Study	Sample Set	Run	Plate	PSMs	Proteins
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F12.raw	1	4402	2388
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F2.raw	1	2008	1346
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F9.raw	1	8305	2839
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech2_F2.raw	1	2554	1511
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech2_F9.raw	1	8443	2811
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt2_tech1_F1	160524_[redacted]_expt2_tech1_F15.raw	2	4597	2299
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt2_tech1_F1	160524_[redacted]_expt2_tech1_F9.raw	2	7255	2763
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt2_tech1_F1	160524_[redacted]_expt2_tech2_F15.raw	2	4249	2282
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt2_tech1_F1	160524_[redacted]_expt2_tech2_F5.raw	2	1662	950
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F15.raw	1	4665	2520
[redacted]	[redacted]s_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F16.raw	1	3819	2167

MedImmune

Proteins, PSM and MSMS Results from a MS-Run

Proteins by MsRun

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING — 1 - 25 of 945 Next › Last »

Parameters: RunParam = 4

View: default

Accessions	Gene	Description	Unique Peptides Count	Total Number Peptides	Percent Coverage	Protein Probability	Error Rate	Confidence	Em PAI	Row Id
[redacted]	[redacted]	[redacted]	8	40	8.5	58.6	0.00	High	0.7	7561
[redacted]	[redacted]	[redacted]	2	237	26.5	245.0	0.00	High	4.8	1378
[redacted]	[redacted]	[redacted]	25	239	73.5	260.9	0.00	High	420.7	2659

/project/${container}/begin.view?pageId=MsRun%20%28New%29&qwp1.param.RunParam=${RowId}&qwp2.param.RunParam=${RowId}&qwp3.param.RunParam=${RowId}&qwp4.param.RunParam=${RowId}&qwp5.param.RunParam=${RowId}

MsMs Spectra

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING

Parameters: RunParam = 4

View: default

Ms Run Id	Scan	Ms Level	Precursor Charge	Retention Time	Snr Reporter	Interference Pct	Inj Tin
160524_[redacted]_expt1_tech1_F12.raw	1446	2	3	13.94	3.8	3.1	
160524_[redacted]_expt1_tech1_F12.raw	1486	2	3	14.21	12.8	16.1	
160524_[redacted]_expt1_tech1_F12.raw	1492	2	3	14.25	14.2	0.7	
160524_[redacted]_expt1_tech1_F12.raw	1494	2	3	14.27	2.3	4.6	
160524_[redacted]_expt1_tech1_F12.raw	1541	2	3	14.53	5.4	5.1	

MsRun Summary

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING — 1 - 25 of 72 Next › Last »

Investigation	Study	Sample Set	Run	Plate	PSMs	Proteins
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F12.raw	1	4402	2388
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F2.raw	1	2008	1346
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech1_F9.raw	1	8305	2839
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech2_F2.raw	1	2554	1511
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt1_tech1_F1	160524_[redacted]_expt1_tech2_F9.raw	1	8443	2811
[redacted]	[redacted]_Expt1_Expt2_tech1_tech2_PD2_Rprgm_V1	160524_[redacted]_expt2_tech1_F1	160524_[redacted]_expt2_tech1_F15.raw	2	4597	2299

PSM by MsRun

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING — 1 - 25 of 4,402 Next › Last »

Parameters: RunParam = 4

View: default

Scan	RT	Mz	Z	Ppm	Coisolation	Injection	Intensity	Peptide	Qvalue	Mods
10002	47.52	731.1195	3	1.0	0.0	47.0	3105262	[redacted]	0.00	
10004	47.53	554.6912	3	0.0	9.1	20.1	2495223	[redacted]	0.00	
10006	47.53	667.6601	4	0.8	39.0	47.3	1816494	[redacted]	0.00	
10008	47.54	481.9740	3	-1.0	67.3	140.5	1374467	[redacted]	0.00	
10020	47.58	418.5945	3	-0.3	40.8	269.9	524959	[redacted]	0.01	M1(Oxidation)
10036	47.64	577.3499	3	-0.9	36.2	21.9	1813384	[redacted]	0.00	
10037	47.64	660.3864	4	0.2	26.6	32.3	758445	[redacted]	0.00	
10040	47.65	572.3969	2	-0.0	7.3	22.3	8923099	[redacted]	0.00	

- A key strengths of LabKey is the flexibility of custom query, visualization and report with SQL/R or point-n-click interface.
- Once a study is imported, its experimental design, LcMsMs runs, protein identification and quantitation can be inspected via the web-interface as data grids or plots.

MedImmune

Protein Abundances - SQL/R Reporting

Protein Identifications

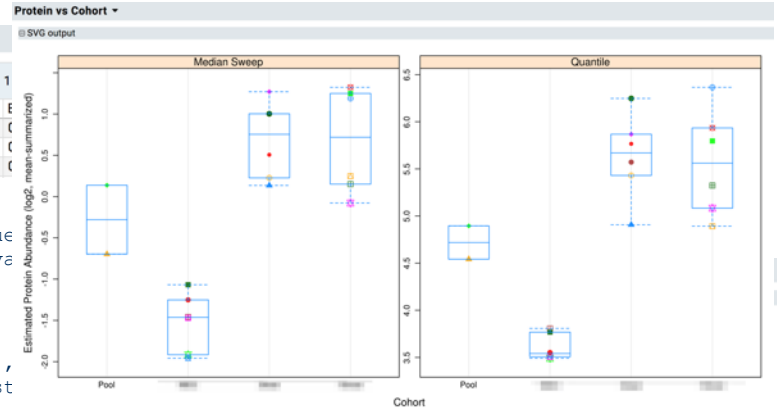
GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING

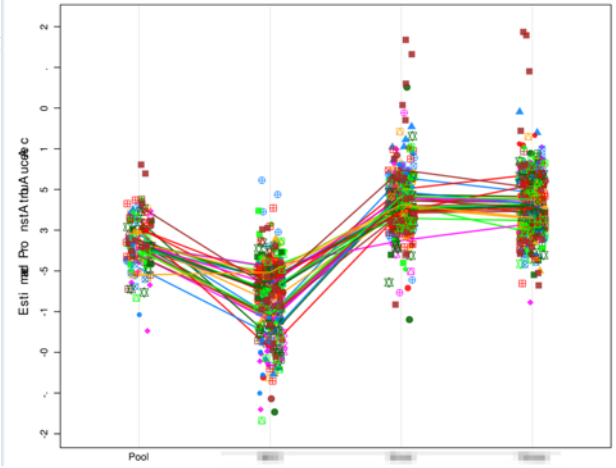
Sequence Id	Gene	Description	Accessions	Peptides	PSMs	Coverage
[illegible]	[illegible]	[illegible]	[illegible]	11	70	22.78
[illegible]	[illegible]	[illegible]	[illegible]	4	28	15.67
[illegible]	[illegible]	[illegible]	[illegible]	13	125	31.73

```
plt01 = dcast(data=labkey.data[,c("name","value","abundance")], abundance~name, value
plt02 = dcast(data=labkey.data[,c("name","value","mswept")],    mswept~name, value.va

plt = rbind(plt01, plt02);

bwplot(abundance~Cohort|Norm, groups=SampleID, data=plt, type=c("p"), layout=c(2, 1),
       par.settings=simpleTheme(pch=c(10:20),cex=1.25,lwd=2),xlab="Cohort", ylab="Est
       scale=list(relation="free",alternating=1, y=list(tick.number=10, log=F, equispaced.log=F),
       panel = function(x, y, ...) {
           panel.dotplot(x, y, par.settings=simpleTheme(pch=c(10:20),cex=0.75,lwd=1), cex=1, alp
           panel.bwplot(x, y, pch = "|", ...)          }
  );
```





For simple visualization, boxplot, volcano plot can be readily generated in LabKey and shared with other researchers.

MedImmune

PSM → MS/MS Spectrum via OpenSlice

Protein Identifications

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING 1 - 25 of 3,005 Next › Last »

Sequence Id	Gene	Description	Accessions	Peptides	PSMs	Coverage	Err%	Confidence	Em PAI
[illegible]	[illegible]	[illegible]	[illegible]	11	70	22.78	0.00	High	4.6
[illegible]	[illegible]	[illegible]	[illegible]	4	28	15.67	0.00	High	2.4

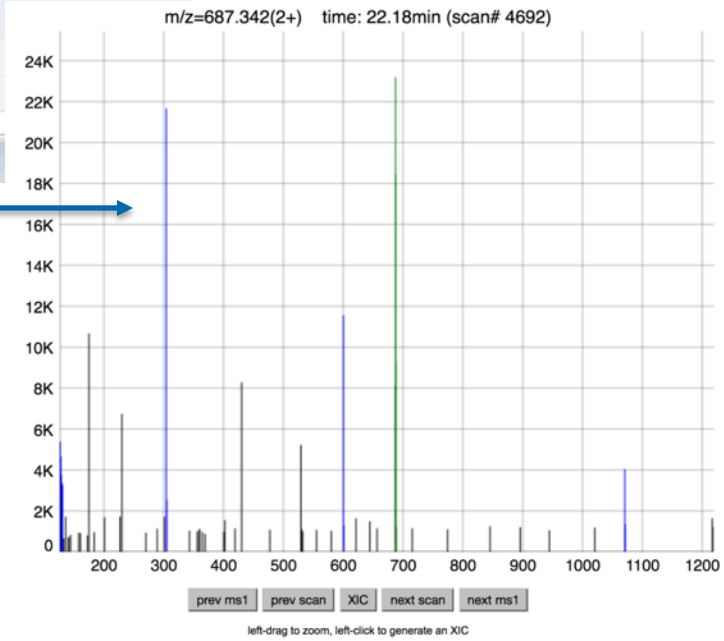
PSMs for the Selected Protein

GRID VIEWS REPORTS CHARTS EXPORT PRINT PAGING

Parameters: Protein = 149733

View: default

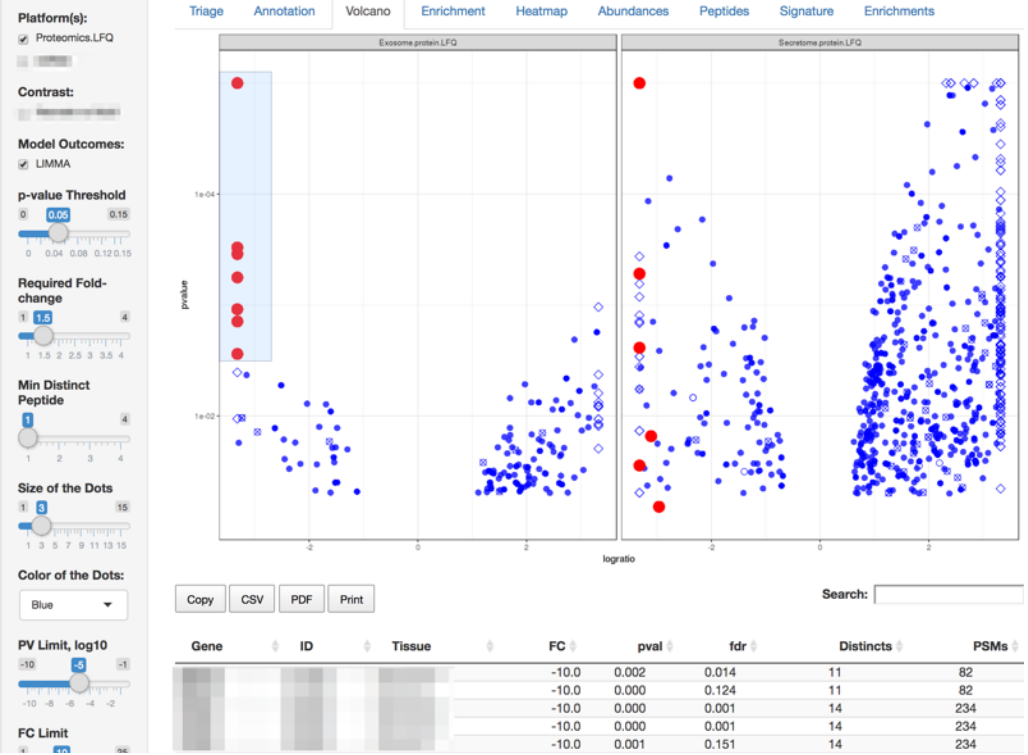
Scan	RT	Mz	Z	Interferene	Injection	Intensity	PPM	Peptide	Qvalue	Modifications	Run
10389	41.95	925.4932	2	0.0	130.0	3651562	1.17	[illegible]	0.001085	M5(Oxidation)	160524_
11428	41.08	917.4944	2	21.3	64.4	6936547	-0.26	[illegible]	0.0		160524_



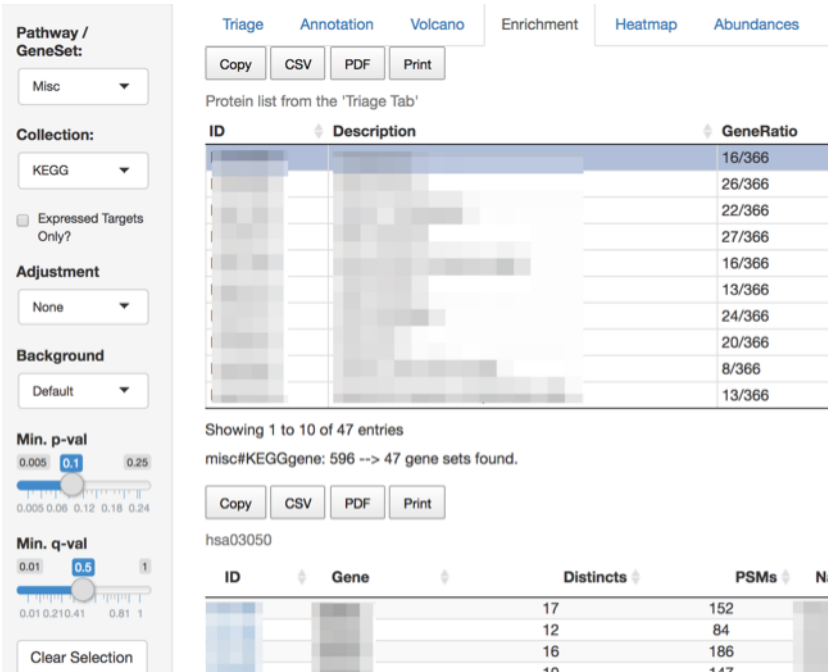
- To visualize the raw MS and MS/MS data, another open-source program, OpenSlice, was adopted. It pre-processes the raw files to allow instantaneous review of spectrum and XIC trace.
- Custom URL in LabKey enables drill-down of the experimental evidences from summary levels downward with OpenSlice.

MedImmune

In-depth Analysis in a R/Shiny App



- Expose LabKey data to Shiny app for in-depth analysis
- Live data tables, linked volcano plots, enrichment analysis, heatmap, and integration with RNASeq, etc.



MedImmune

Lessons Learned

- LabKey, due to its open-source architect and abundance of widgets and customizability, is an ideal environment to manage complex omics data such proteomics
- By externalizing the platform-specific processes, different data types can be readily managed in LabKey.
- Template folder provides a good compromise between UI flexibility and usability.
- Better grasp of the 'folder' concept and the scoping rule is crucial
- Proper division of labors is critical
  - Server-side data management
  - Client-side UI customization and reporting
- Better out-of-box features will reduce the upfront works in a commercial settings.

MedImmune

Future Plan

- Additional workflows for
  - Label-free quantitation by Maxquant
  - Targeted proteomics using the “Panorama” module
- UI refinements
  - to accommodate multiple workflows and to clarify user inputs
  - Factors, factor options and sample attribute editor
  - Request for Stat Analysis

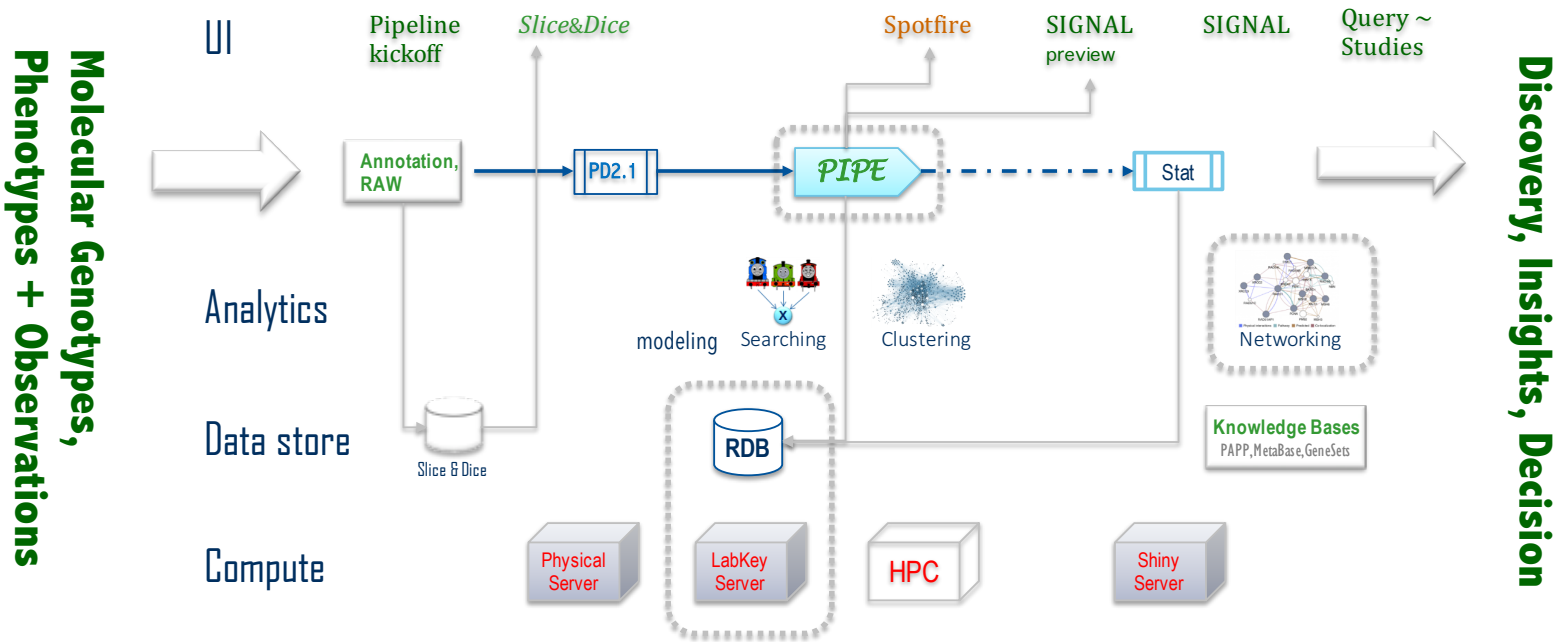
MedImmune

Acknowledgements

- Research Bioinformatics,
- Proteomics, AD&PE
- RD&I
- Statistical Science
- MedImmune, AZ
- Cory Nathe , Frank Lee
- Josh Eckels
- Steve Hanson, Avital Sadot
- LabKey

MedImmune

Data Pipeline for *Proteo/Metabolo/Lipidomics*



MedImmune