Data Analysis

Kelly Ruggles, Ph.D.
Assistant Professor, Department of Medicine
NYU Langone Medical Center
www.ruggleslab.org

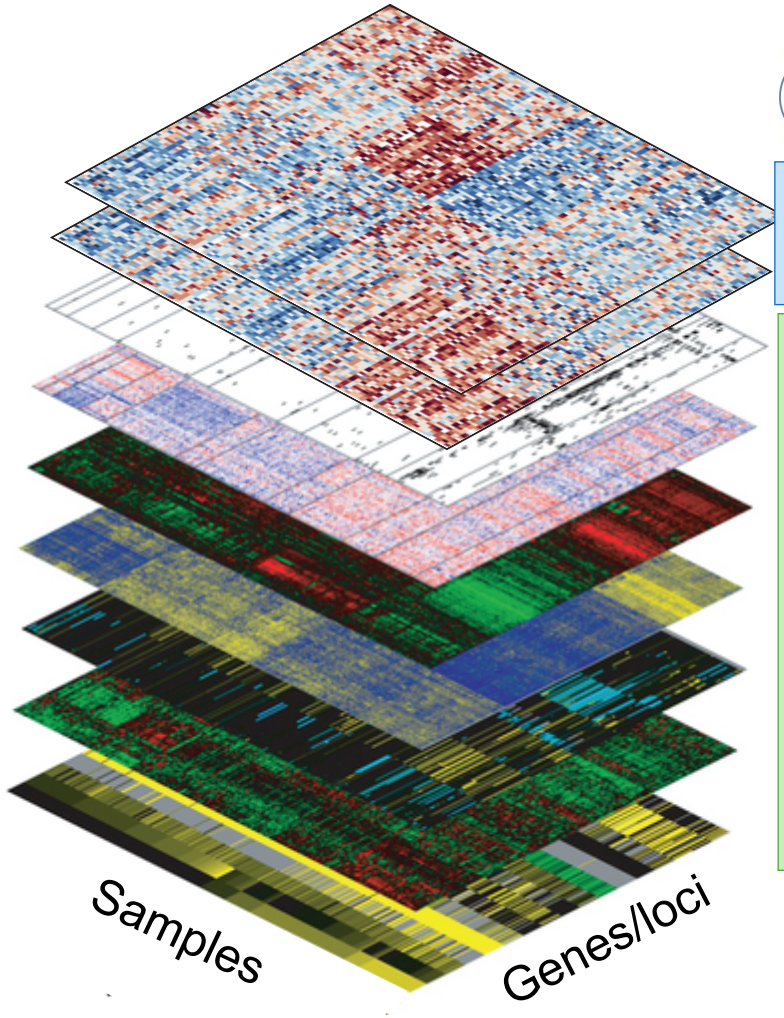
September 18, 2017
Methods in Quantitative Biology

Let’s make it less vague

- How do we explore and analyze matrices of gene/protein expression?

Gene Name	Description	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
plectin isoform 1	NP_958782	1.10	2.61	-0.66	0.20	-0.49	2.77	0.86	1.41	1.19	1.10
plectin isoform 1g	NP_958785	1.11	2.65	-0.65	0.22	-0.50	2.78	0.87	1.41	1.19	1.10
plectin isoform 1a	NP_958786	1.11	2.65	-0.65	0.22	-0.50	2.78	0.87	1.41	1.19	1.10
plectin isoform 1c	NP_000436	1.11	2.65	-0.63	0.21	-0.51	2.80	0.87	1.41	1.19	1.10
plectin isoform 1e	NP_958781	1.12	2.65	-0.64	0.22	-0.50	2.79	0.87	1.41	1.20	1.09
plectin isoform 1f	NP_958780	1.11	2.65	-0.65	0.22	-0.50	2.78	0.87	1.41	1.19	1.10
plectin isoform 1d	NP_958783	1.11	2.65	-0.65	0.22	-0.50	2.78	0.87	1.41	1.19	1.10
plectin isoform 1b	NP_958784	1.11	2.65	-0.65	0.22	-0.50	2.78	0.87	1.41	1.19	1.10
epiplakin	NP_112598	-1.52	3.91	-0.62	-1.04	-1.85	2.21	1.92	3.20	1.05	-2.41
myosin-9	NP_002464	2.04	1.59	-1.27	1.03	0.11	1.25	0.42	0.12	1.15	1.96
myosin-10 isoform 3	NP_001243024	2.10	0.51	-0.67	-0.82	0.23	1.33	0.44	-1.76	2.83	1.91
myosin-10 isoform 1	NP_001242941	2.10	0.51	-0.66	-0.82	0.23	1.29	0.43	-1.76	2.81	1.91
myosin-11 isoform SM1A	NP_002465	-0.23	-2.18	-3.12	0.69	-1.93	-1.67	-0.63	-2.52	2.29	-0.09
myosin-10 isoform 2	NP_005955	2.10	0.51	-0.69	-0.82	0.23	1.35	0.43	-1.75	2.83	1.94
myosin-11 isoform SM2B	NP_001035202	-0.23	-2.14	-3.12	0.67	-1.94	-1.67	-0.62	-2.53	2.29	-0.12
myosin-14 isoform 1	NP_001070654	-0.88	-2.88	-1.97	0.26	-0.05	3.78	-2.42	-3.10	1.56	-0.71
myosin-14 isoform 2	NP_079005	-0.88	-2.90	-1.97	0.27	-0.04	3.80	-2.47	-3.10	1.58	-0.74
unconventional myosin-Va isoform 1	NP_000250	-0.16	0.92	-2.73	0.03	0.45	-0.29	-1.18	1.27	1.08	-0.43
unconventional myosin-Vb	NP_001073936	-0.07	-0.88	-2.28	1.87	-0.98	0.46	-2.78	1.25	0.27	-0.17
unconventional myosin-Vc	NP_061198	-0.35	-1.02	0.02	-0.88	-1.52	2.07	1.44	-1.40	1.73	0.07
unconventional myosin-Ic isoform a	NP_001074248	0.32	-0.44	0.09	0.78	-0.61	-0.39	2.44	-0.89	1.04	-0.01
unconventional myosin-Ic isoform b	NP_001074419	0.32	-0.44	0.09	0.79	-0.62	-0.39	2.44	-0.88	1.05	0.01
unconventional myosin-Id	NP_056009	0.97	1.64	-0.91	0.02	0.85	1.11	1.63	-0.05	3.59	0.60
unconventional myosin-Ib isoform 2	NP_036355	1.53	2.93	-2.38	-0.76	0.56	-0.05	-0.79	1.26	0.14	1.18

Sample Dataset: Breast Cancer Proteogenomics



CPTAC

Proteomics
Phosphoproteomics

77 Human
Breast Tumors

Mutation
Copy Number
Gene Expression
DNA methylation
MicroRNA
RPPA
Clinical Data

825 Human
Breast Tumors

TCGA

TCGA. Nature 490, 61-70 (2012)

Ozenberger KE, et al., Nature Genetics 45, 1113-1120 (2013)

Mertins P*, Mani DR*, Ruggles KV*, Gilette M* et al., Nature 534, 55-62 (2016)

Data Types in Proteogenomics

GENOMICS

WGS, WXS

RNA-Seq

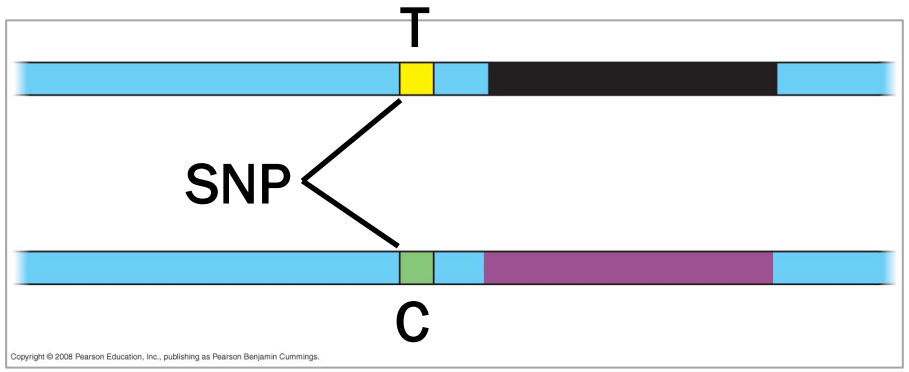
Copy number Alterations (**CNA**)

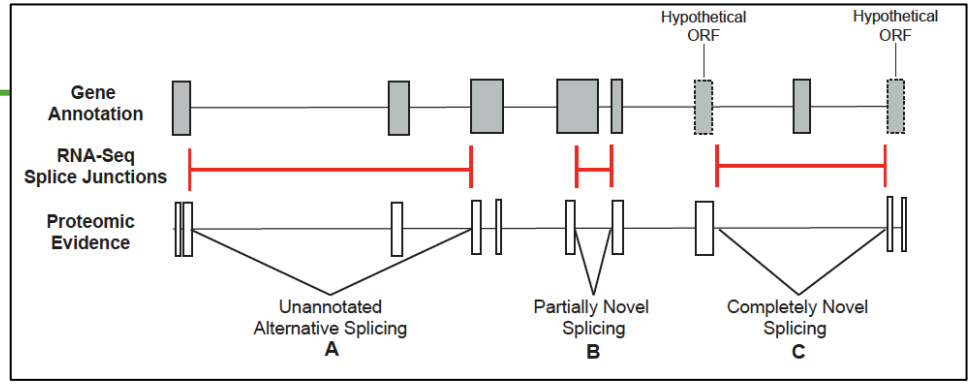
Single Nucleotide Polymorphisms (**SNPs**)

Novel Splice Junctions

Gene Expression

Single base-pair sites that vary in a population





Splicing of exons, creating new protein isoforms

PROTEOMICS

LC-MS/MS

Global Protein Expression

Phosphoprotein Abundance

Targeted Proteomics

Data Types in Proteogenomics

GENOMICS

Amplifications or deletions in the genome

WGS, WXS

Copy number Alterations (**CNA**)

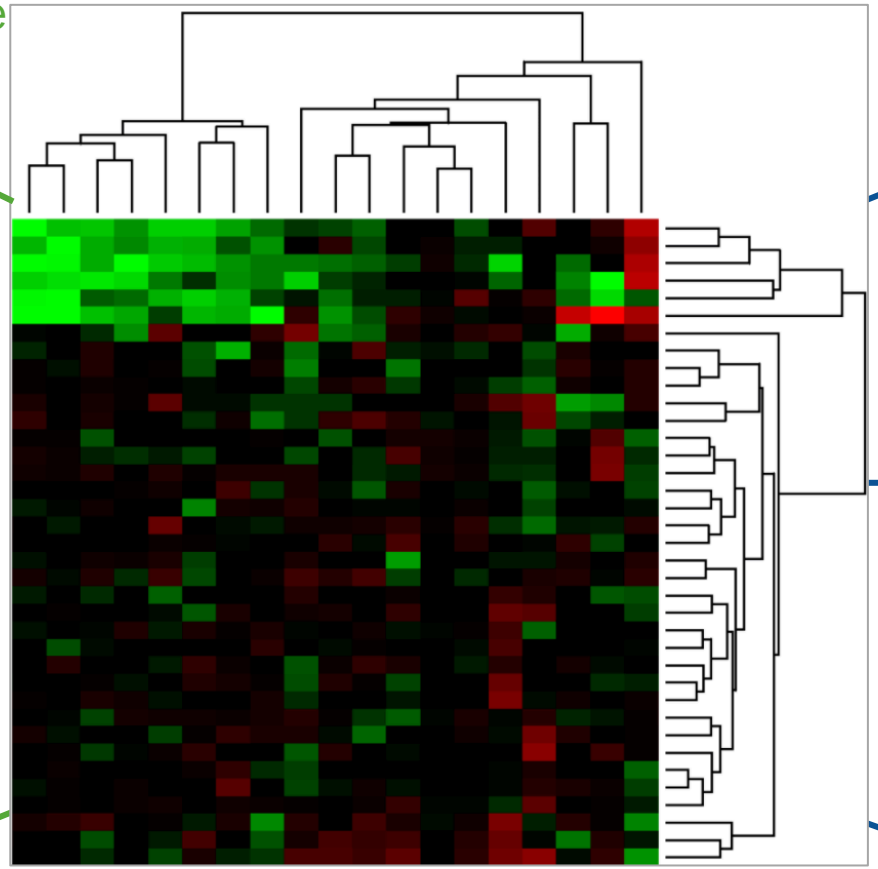
Single Nucleotide Polymorphisms (**SNPs**)

RNA-Seq

Novel Splice Junctions

Gene Expression

Potential protein quantitation



PROTEOMICS

Global Protein Expression

Relative quantitation

Phosphoprotein Abundance

Signaling

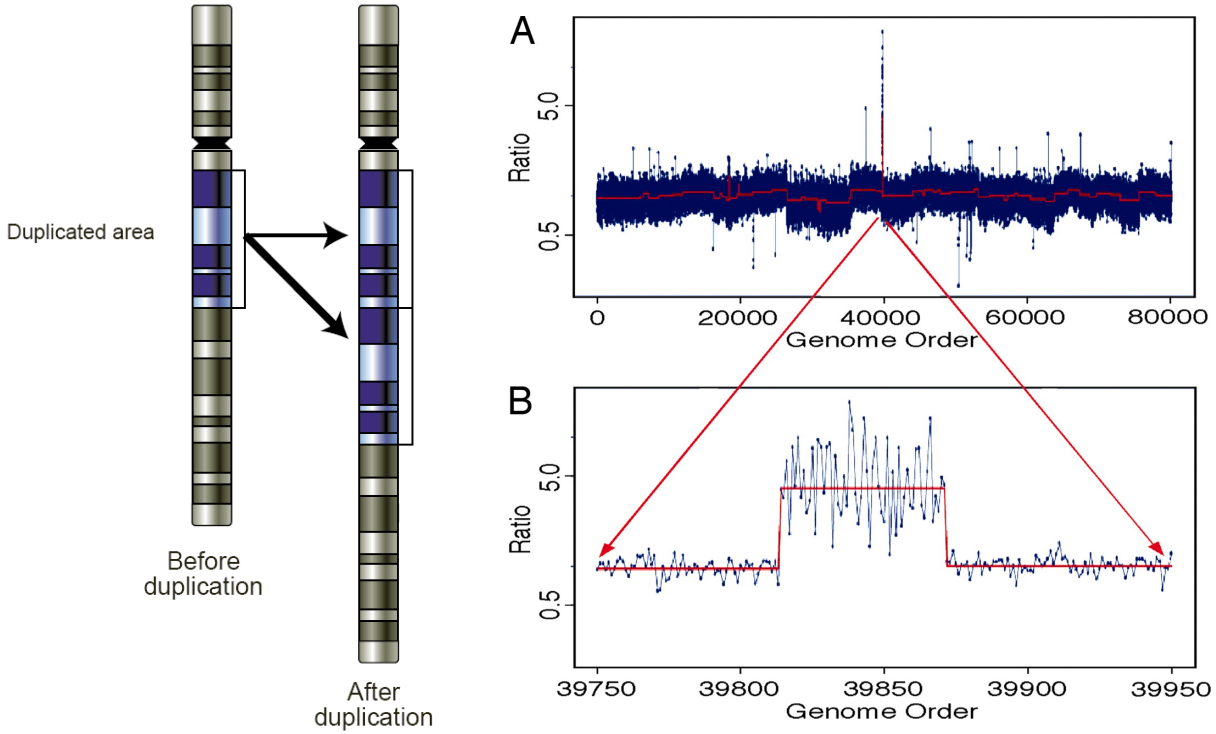
Targeted Proteomics

Absolute quantitation

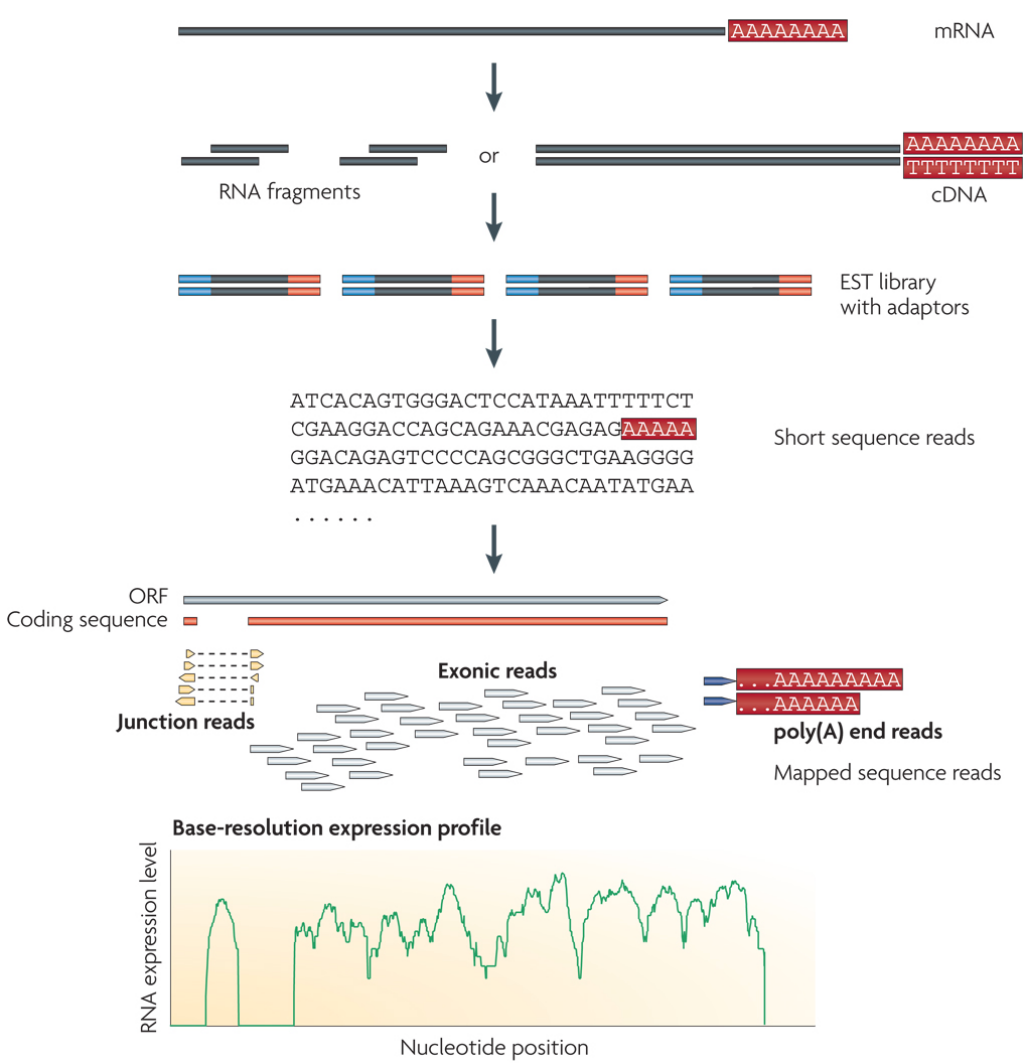
LC-MS/MS

Copy Number Alterations (CNA)

- Changes in the genome due to duplication or deletion of large regions of DNA (>1kb)
- Thought to cover >10% of human genome



Gene Expression using RNA-Seq



RNAs are converted into cDNA fragment library

Sequence adapters (blue) are added to cDNA fragments

Short sequence reads from each cDNA are obtained

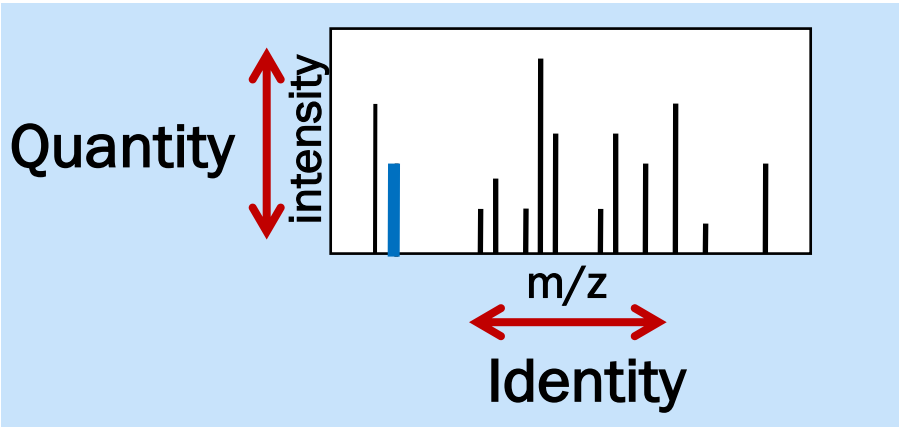
Reads are aligned to reference sequence and classified as exonic reads, junction reads or poly(A) end-reads

Used to generate a base-resolution expression profile for each gene

Protein Identification and Quantitation by Mass Spectrometry

Tumor Sample → Lysis → Fractionation → Digestion → Peptides →

Tandem Mass Spectrometry



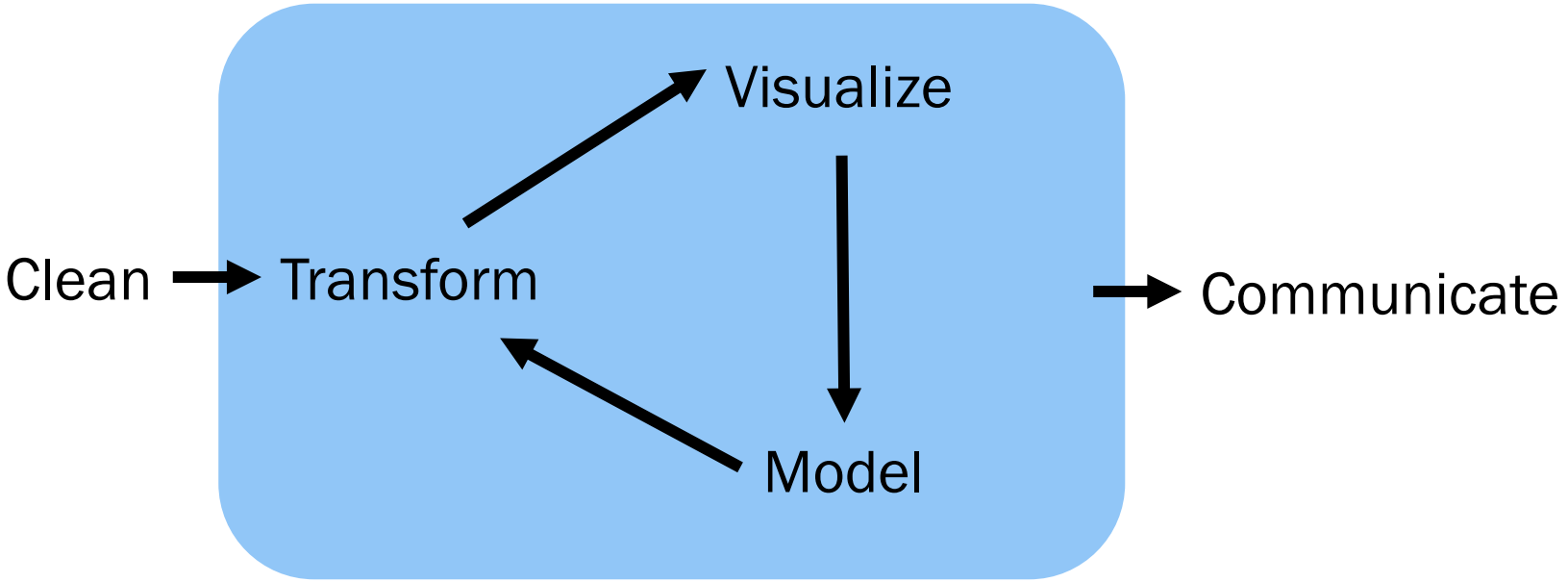
Discovery Proteomics:

- Used to measure global protein expression (whole cell proteome)
- Can enrich for phosphopeptides to measure phosphorylation status

Targeted Proteomics:

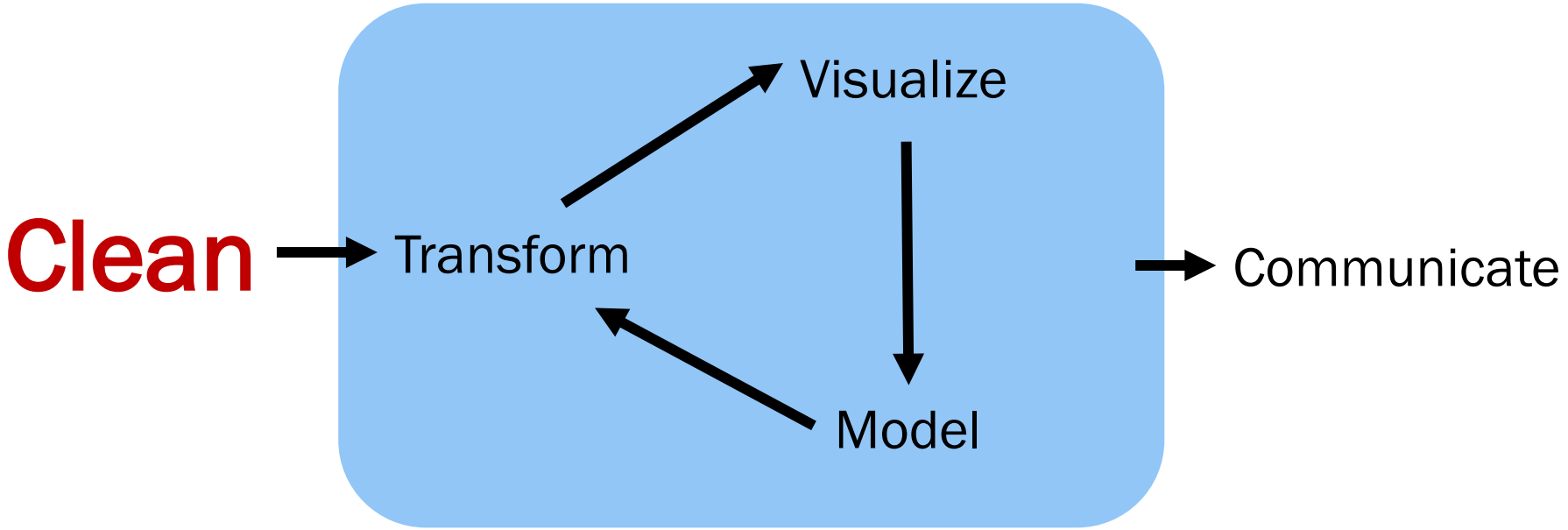
- Hypothesis driven analysis
- Select proteins and representative peptides of these proteins to measure prior to run

Data Exploration



Modified from R for Data Science, Wickham & Grolemund

Data Exploration



Modified from R for Data Science, Wickham & Grolemund

Data Cleaning

- Often gene and sample names are not formatted exactly as needed for downstream analysis

	TCGA-A2-A0CM-01A-31R-A034-07	TCGA-A2-A0D0-01A-11R-A00Z-07	TCGA-A2-A0D1-01A-11R-A034-07
UBC|7316	0.052	0.360	-0.476
GUCY2D|3000		-2.085	3.337
C11orf95|65998	0.405	0.446	1.011
C17orf81|23587	-0.129	0.273	-0.024
ANKMY2|57037	-0.890	-1.851	-1.510
TTC36|143941		-6.382	

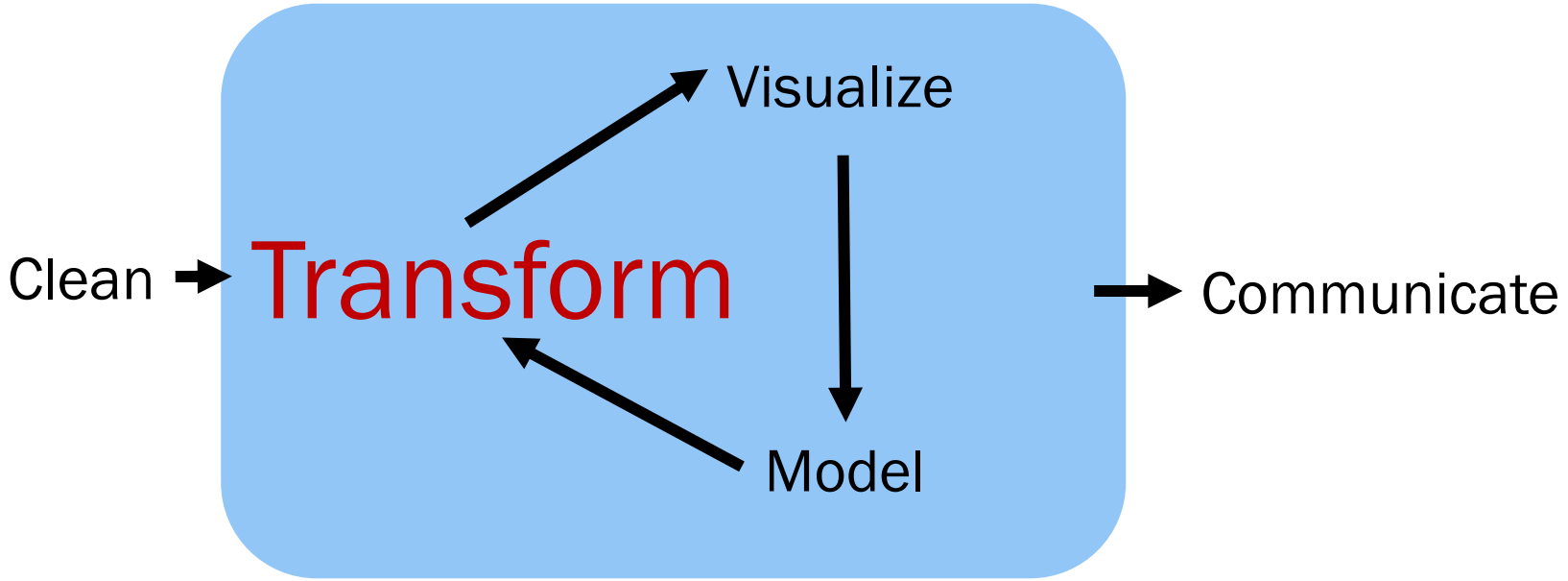
- Or a different reference database was used and the accessions don't match (ex: Ensembl vs. RefSeq)

	AO-A12D.01TCGA	C8-A131.01TCGA	AO-A12B.01TCGA
NP_958782	1.10	2.61	-0.66
NP_958785	1.11	2.65	-0.65
NP_958786	1.11	2.65	-0.65
NP_000436	1.11	2.65	-0.63
NP_958781	1.12	2.65	-0.64
NP_958780	1.11	2.65	-0.65

Data Cleaning

- Missing data:
  - Are missing values in the dataset coded as ‘0’, ’NA’, ‘NaN’, Blanks?
  - Should genes (rows) be removed if they have more than a certain number of missing values?
- Are there repeat samples in the matrix?
  - Technical or experimental replicates?
- Are there repeat genes or proteins in the matrix?

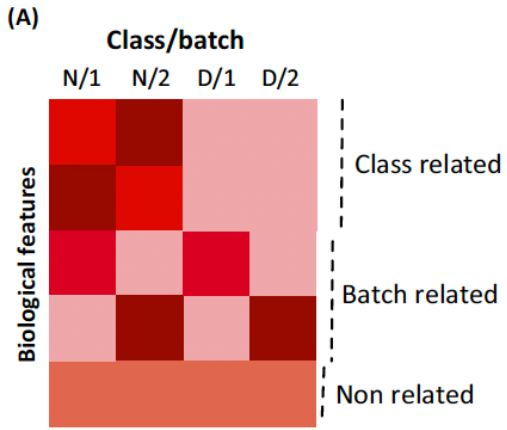
Data Exploration



Modified from R for Data Science, Wickham & Grolemund

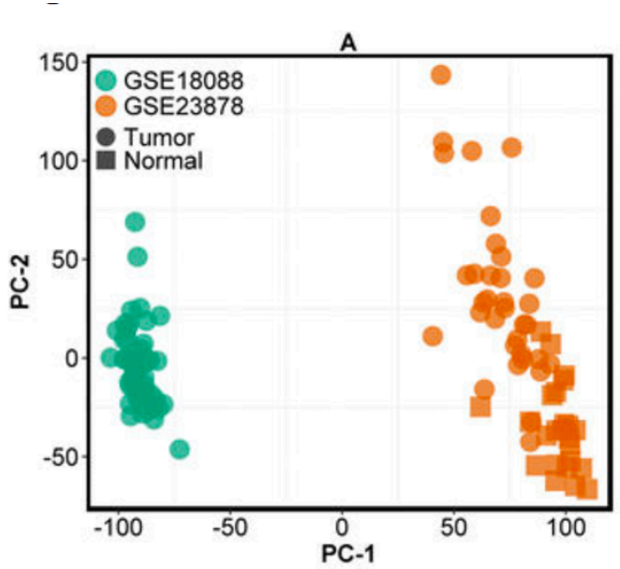
Data Transformation

- Bias in omics can be defined as non-biological signal or features of the data that can be explained by experimental or technical reasons
  - ”Batch Effect”
- Normalization can be used to remove these biases



Class related:
e.g. Normal vs. disease

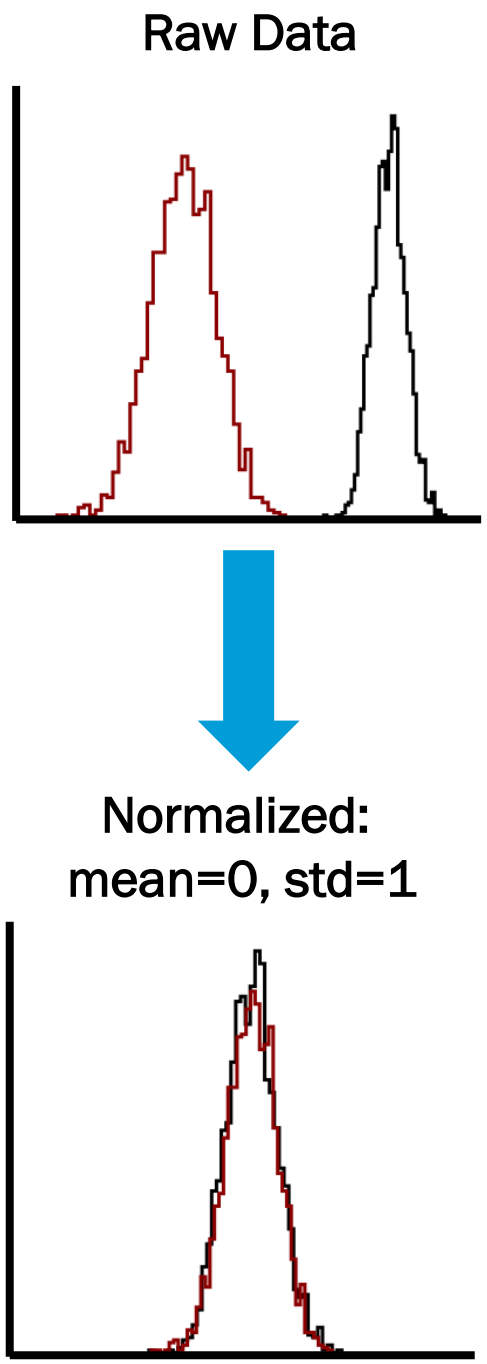
Goh, 2017



Nyamundanda, 2017

Data Normalization

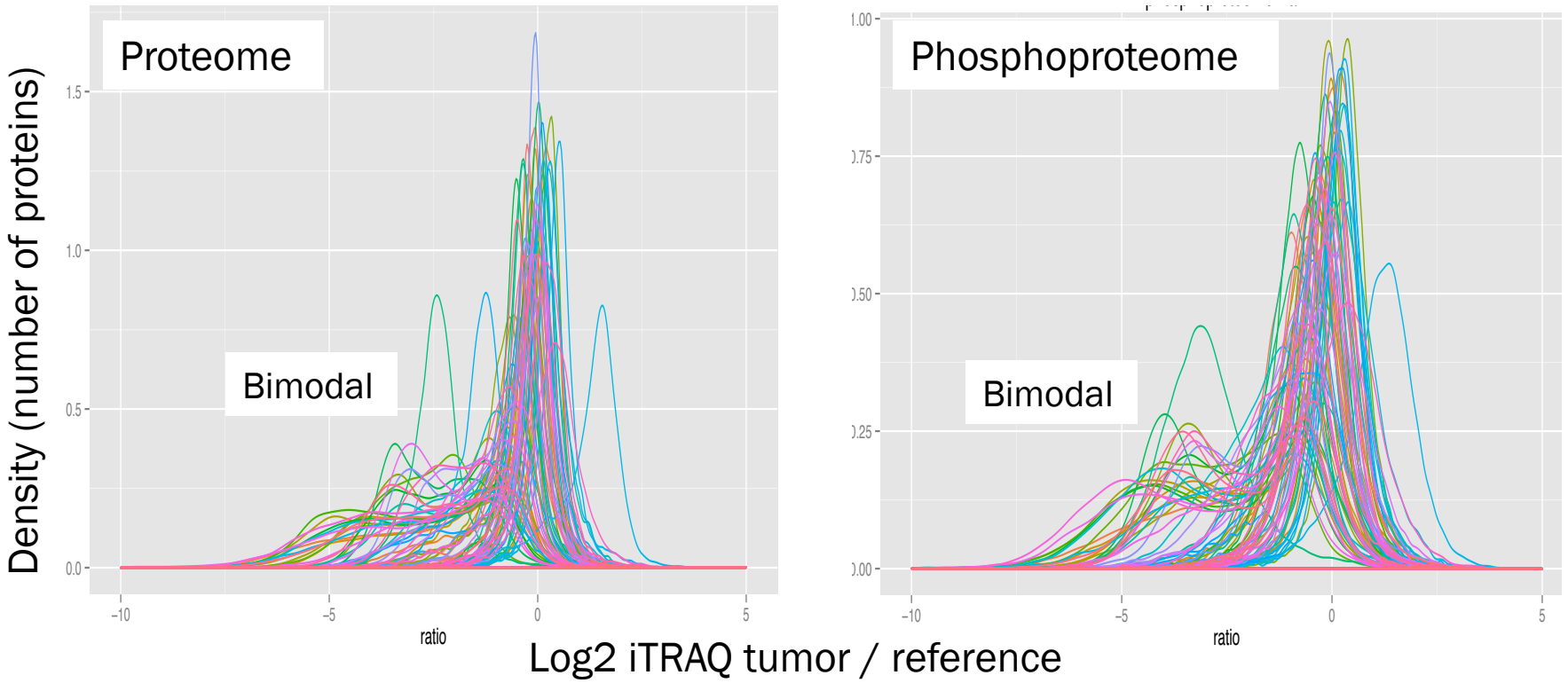
- Simple cases: adjusting values measured on different scales to a common scale
  - Allow the comparison of values from different data sets or with different protein concentrations
- Complicated cases: intention is to bring the entire probability distribution of adjusted values into alignment
  - Align all data to a normal distribution
  - Align quantiles of different measurements



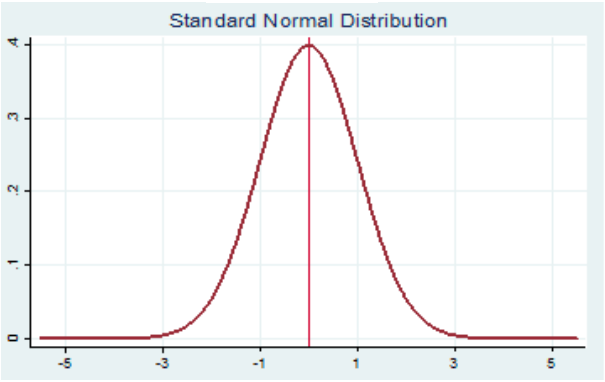
Normalization Methods

- Global Adjustment
  - Used to force the distribution of the log intensity values to center around the mean or median for each sample
  - Assumptions:
    - Most gene abundances do not change, so distribution of intensities across samples should be similar
  - LOG2 normalization
    - Simplifies statistics
    - LOG2 used because we can easily translate into fold change
- Lowess regression: used in microarrays
- Quantile Normalization
- Two component Gaussian
- Z-score Normalization

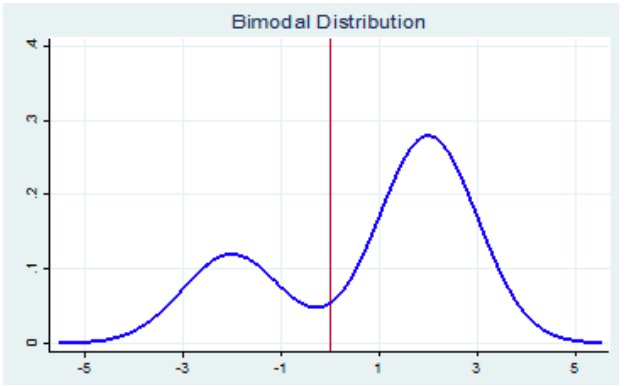
Remove “Wonky” samples



Normal



Bimodal



- Some tumors have bimodal distribution of both proteins and phosphopeptides with lower overall abundance
- Not a processing or technical artifact
- Not specific to subtype, PAM50 status or histology

Normal: 54 (total 75)
Bimodal: 26 (total 30)

Data Imputation

- Replacing missing data with substituted values
- Problems caused by missing data:
  - Introduces bias if the missingness is not random
  - Makes analysis of data more difficult
- Imputing data can also introduce new bias
- In many statistical packages, if one or more missing values are present that case is discarded
  - Does not add any bias but reduces sample size/power

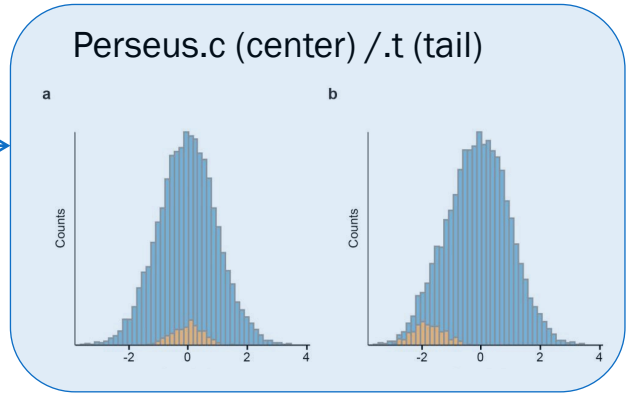
Data Imputation Tools

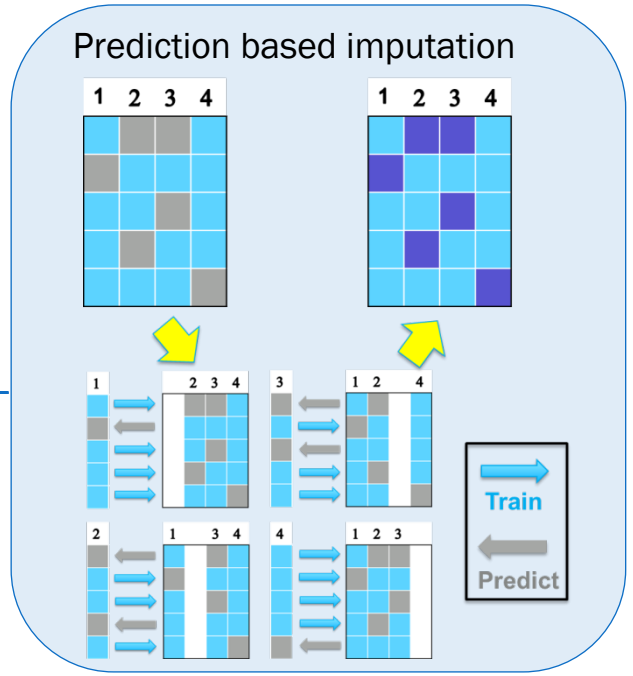
1. Non-informative Imputation
   - Fixed-value imputation: median or minimum
   - Perseus (*S. Tyanova, et al. 2016*): sampling from a non-informative distribution.

2. Low rank matrix completion
   - softImpute (*R. Mazumder, et al. 2010*): imagine processing; a regularized SVD decomposition. R-package: 'softImpute'.

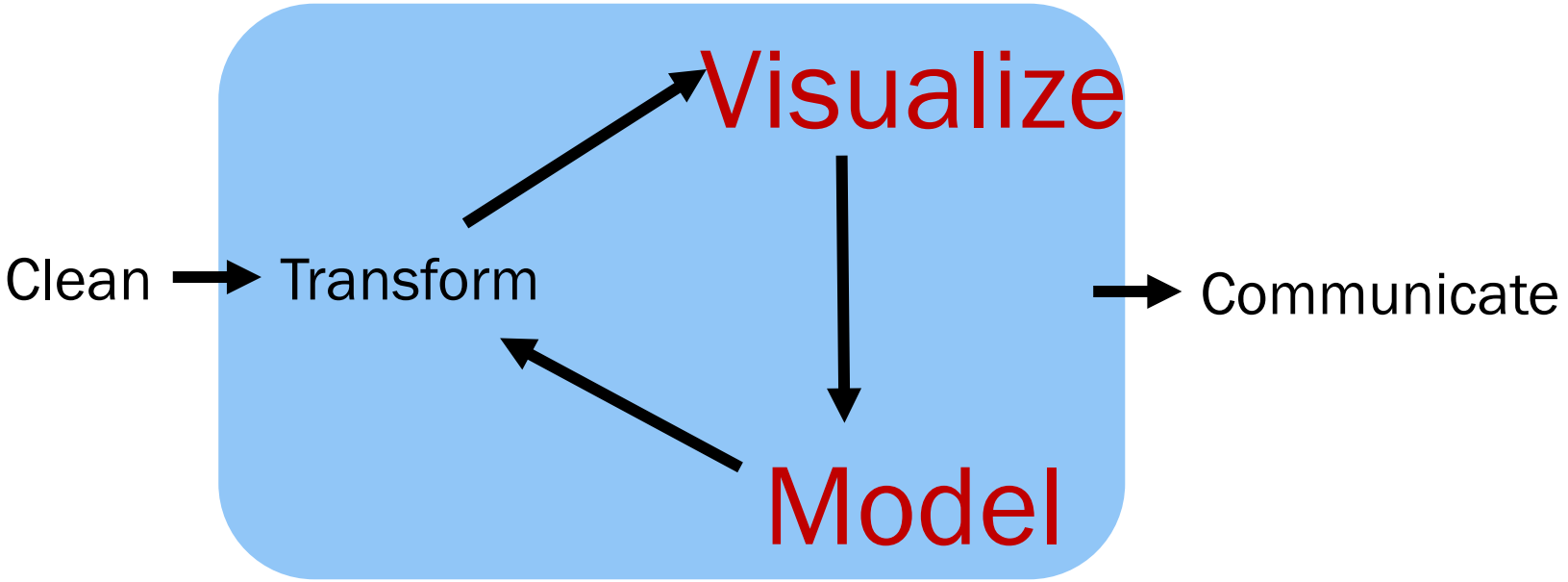
3. Prediction based imputation
   - KNN: R-package: 'pamr'.
   - Lasso: R-package: 'glmnet'.
   - Xgboost (*T. Chen, et al. 2016*): R-package: 'xgboost'.

4. Machine-learning based imputation
   - missForest (*D. J. Stekhoven, et al. 2012*): R-package: 'missForest'.
   - ADMIN: A multi-layer prediction model learned through an iterative procedure.



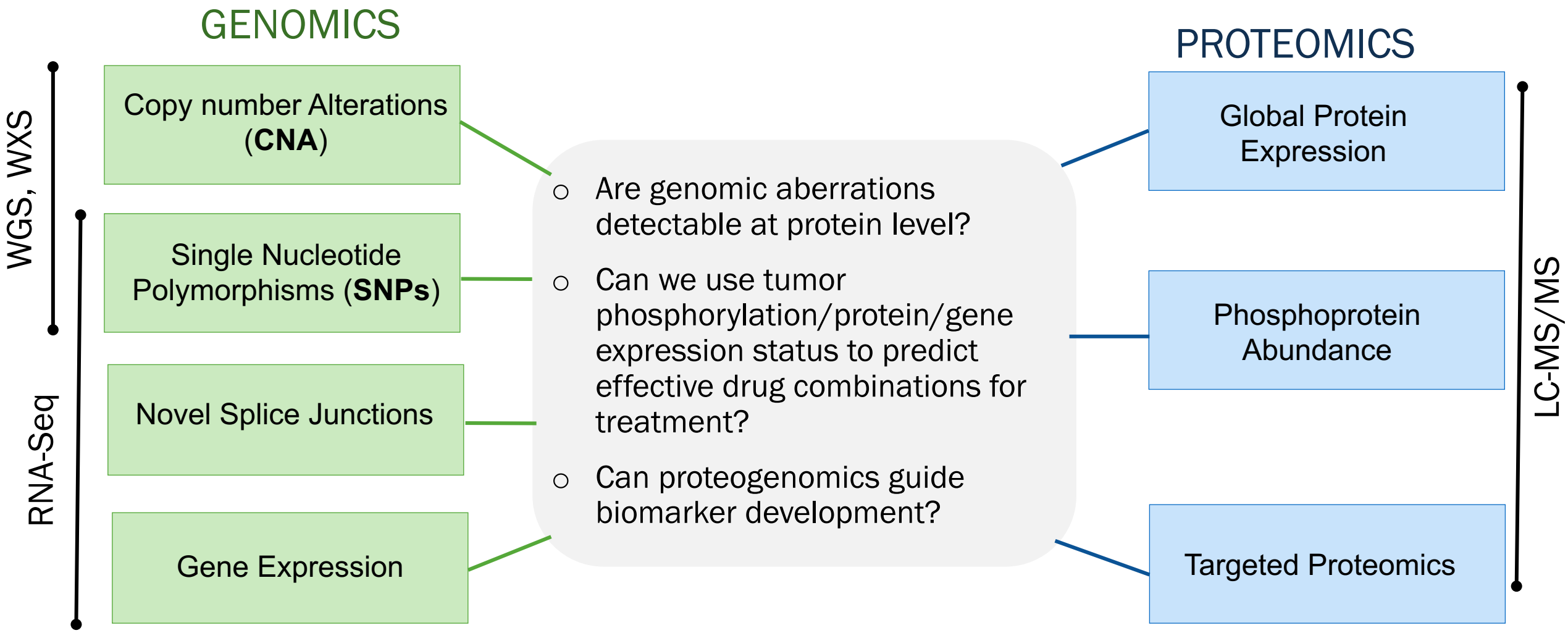


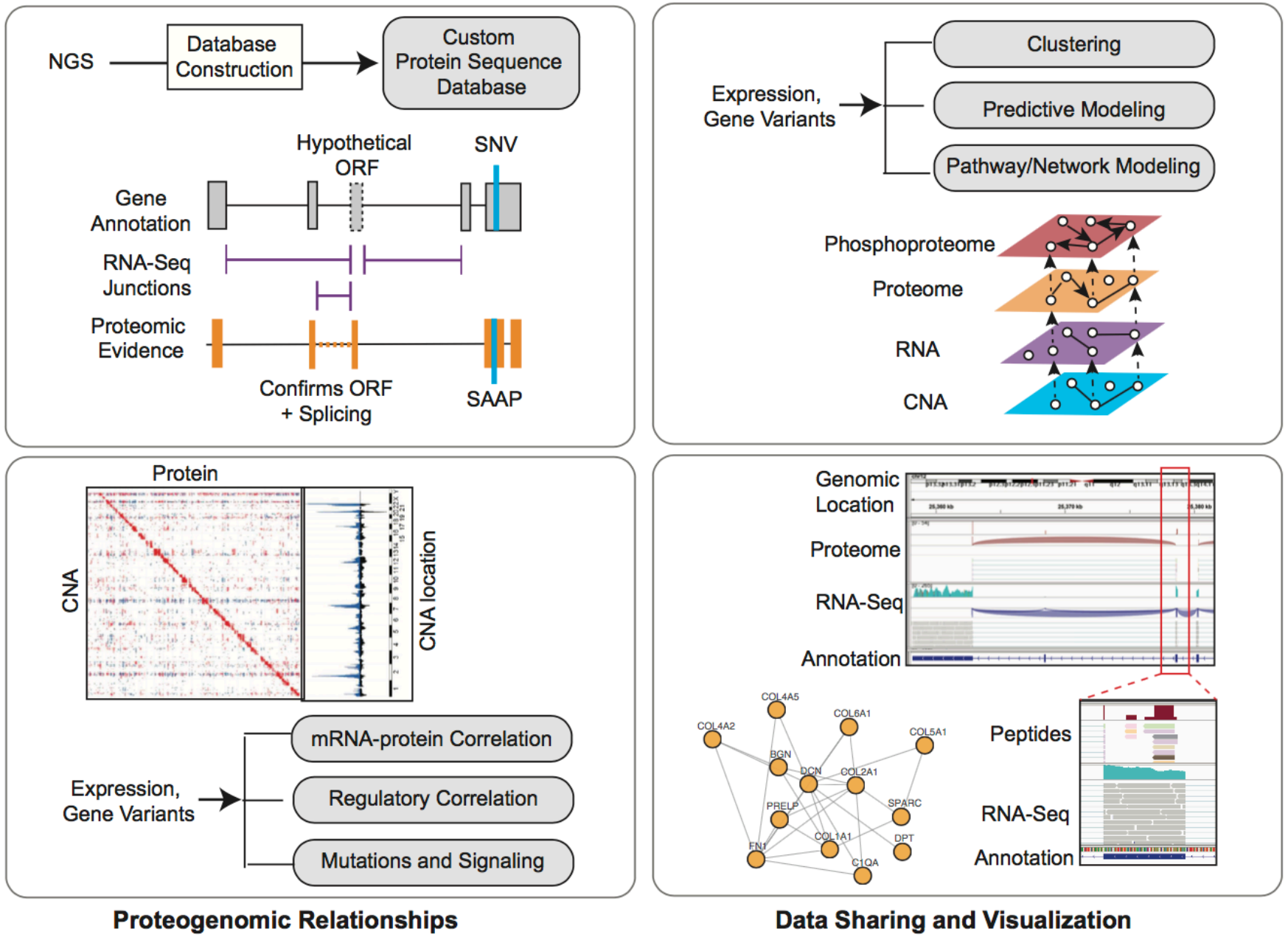
Data Exploration



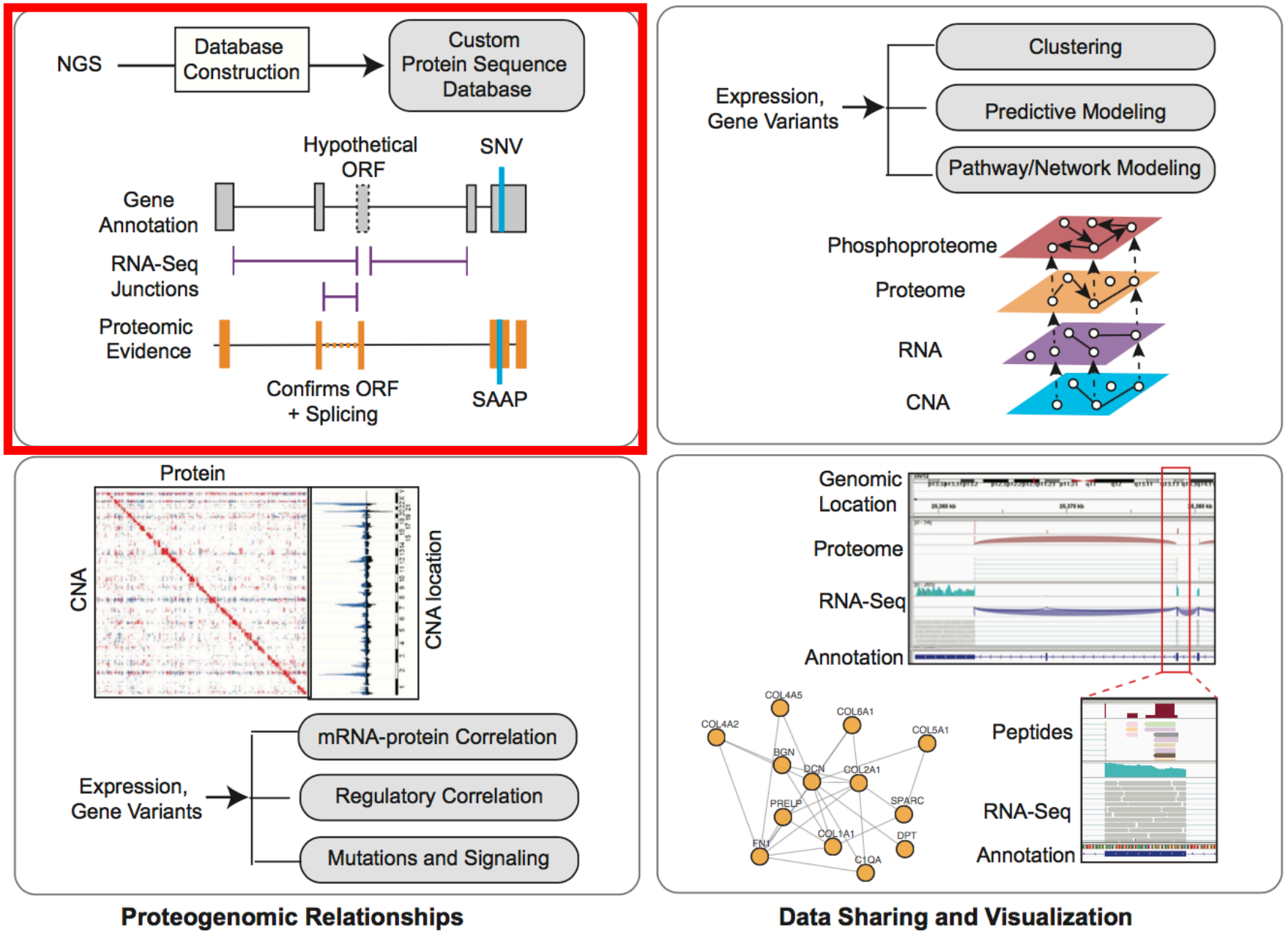
Modified from R for Data Science, Wickham & Grolemund

Goals of Proteogenomic Integration





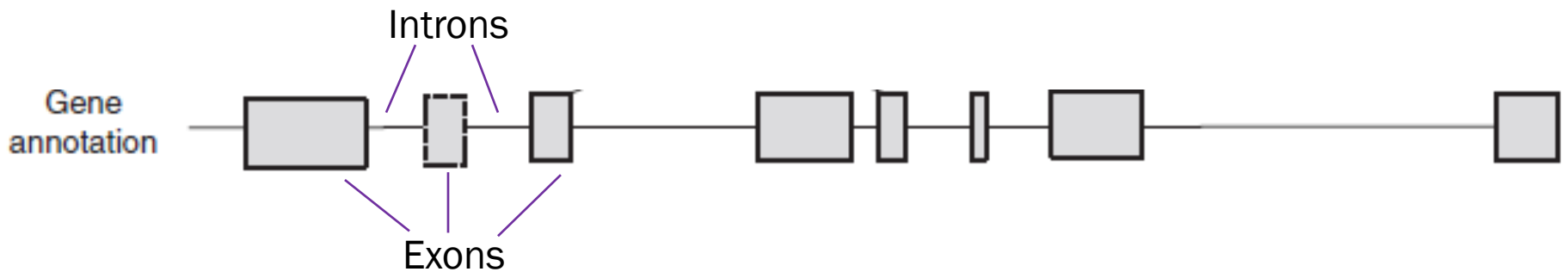
Ruggles et al., MCP 16(6), 959-981 (2017



Ruggles et al., MCP 16(6), 959-981 (2017

Genome Annotation

- To be useful, genomes must be annotated
- Genome annotation:
  - identifying the location and function of protein coding genes
  - Understand cis-regulatory sequences
  - Alternative splicing



Reference Genome

- Serves as a “representative example” of a species’ set of genes
- Created by sequencing a number of donors

Human Reference

Release name	Date of release	Equivalent UCSC version
GRCh38	Dec 2013	hg38
GRCh37	Feb 2009	hg19
NCBI Build 36.1	Mar 2006	hg18
NCBI Build 35	May 2004	hg17
NCBI Build 34	Jul 2003	hg16

Mouse Reference

Release name	Date of release	Equivalent UCSC version
GRCm38	Dec 2011	mm10
NCBI Build 37	Jul 2007	mm9
NCBI Build 36	Feb 2006	mm8
NCBI Build 35	Aug 2005	mm7
NCBI Build 34	Mar 2005	mm6

https://genome.ucsc.edu/FAQ/FAQreleases.html#release1

Reference Sequence Database

- Annotated and curated genes, transcripts and proteins

Curated Protein Coding

Swiss-Prot
UniProt
RefSeq NP

Translated Genes

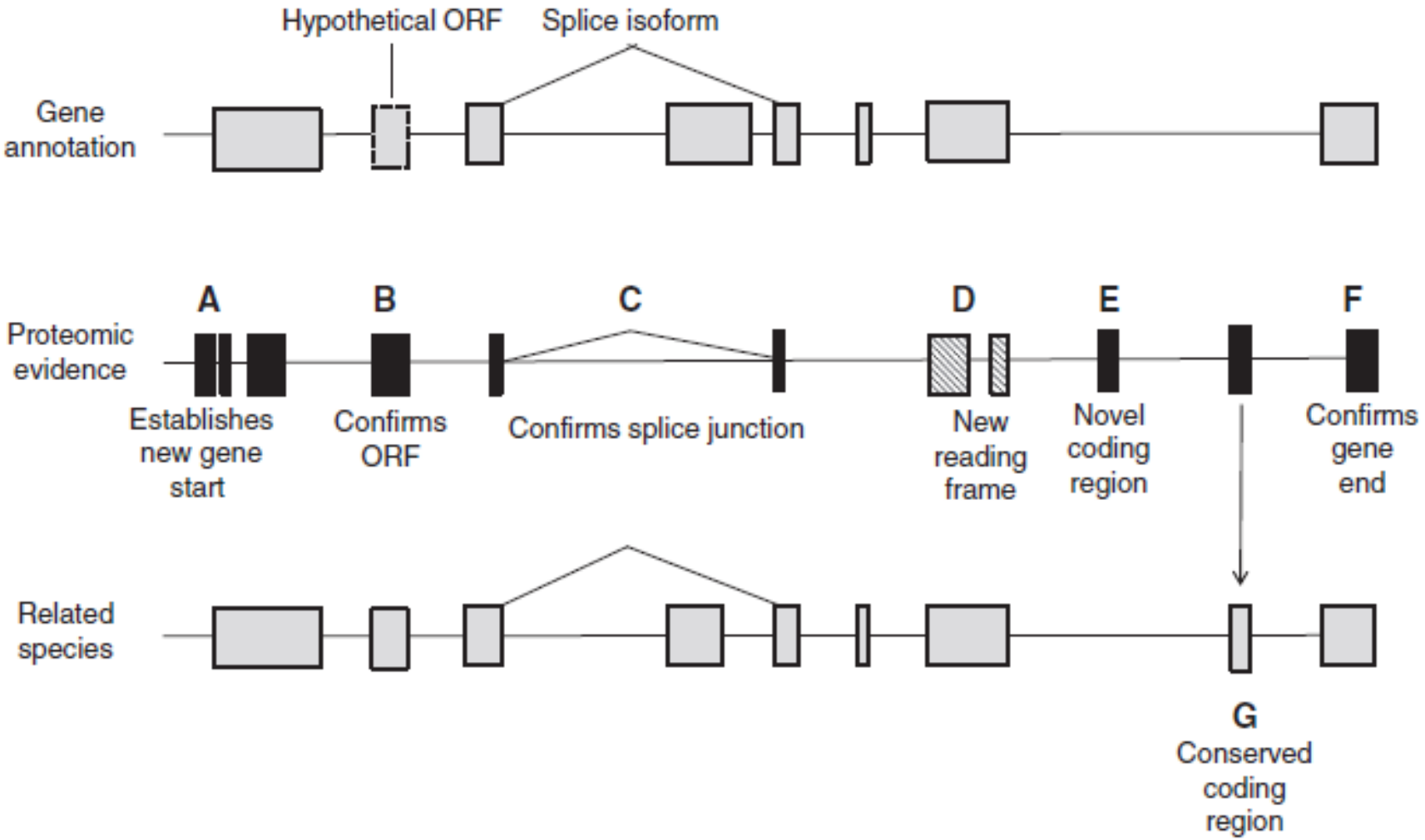
TrEMBL
RefSeq XP, ZP

Annotated Genomes*

Ensembl
UCSC

*Automated annotation through pattern matching of protein to DNA + known proteoin coding genes

Genome Annotation



Ruggles & Fenyo, 2015

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

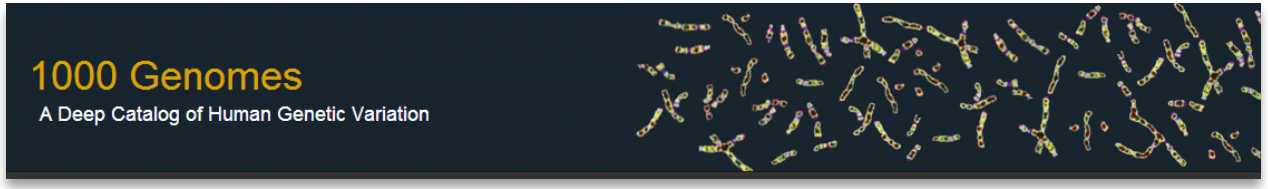
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr17:39,699,560-39,728,662 29,103 bp. enter position, gene symbol, HGVS or search terms go

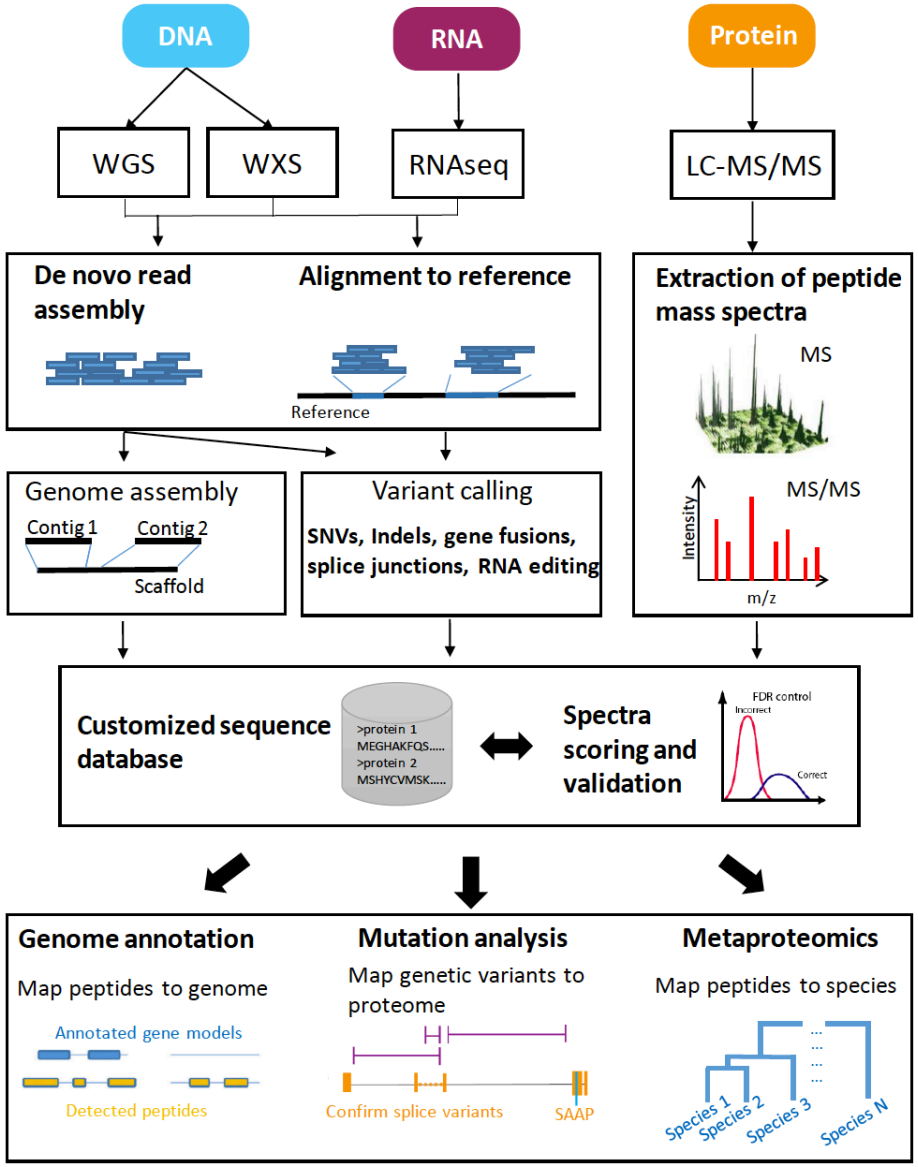
Genetic Variation

- Because the human species is so large, many spontaneous, nonlethal mutations have arisen in all human genes
- With NGS, we can now identify these mutations and study their evolution and inheritance across thousands of humans
- Comparing human genomes, two individuals differ in roughly 1 nucleotide per 1000
- When two sequence variants exist and are both common (~1%) they are called polymorphisms
  - single nucleotide polymorphisms (if substitution in 1 nucleotide)
  - Indels (small insertion or deletion)
  - Copy number variation (CNV), larger insertion/deletion

Genomic Variant Databases



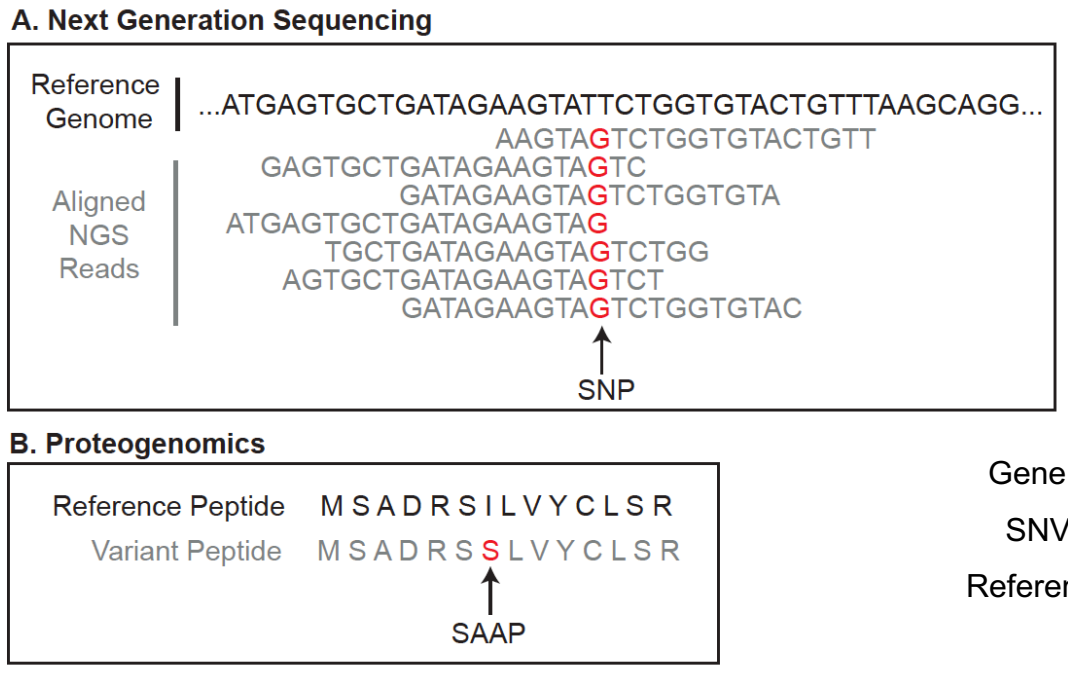
Sequence Focused Proteogenomics



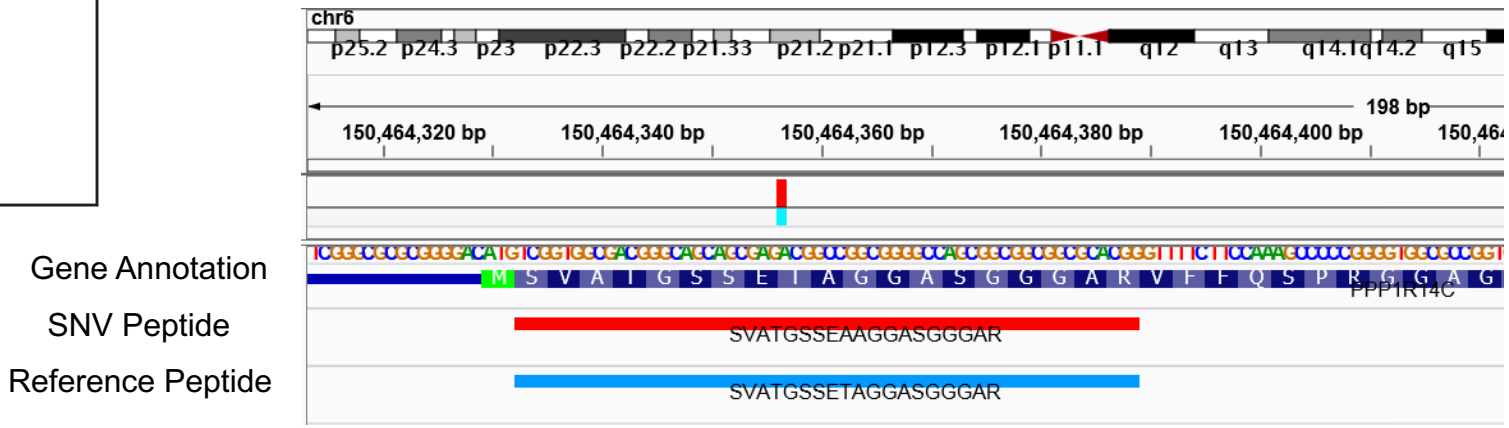
Ruggles et al., 2017

Proteogenomics and SAAP discovery

Modeling

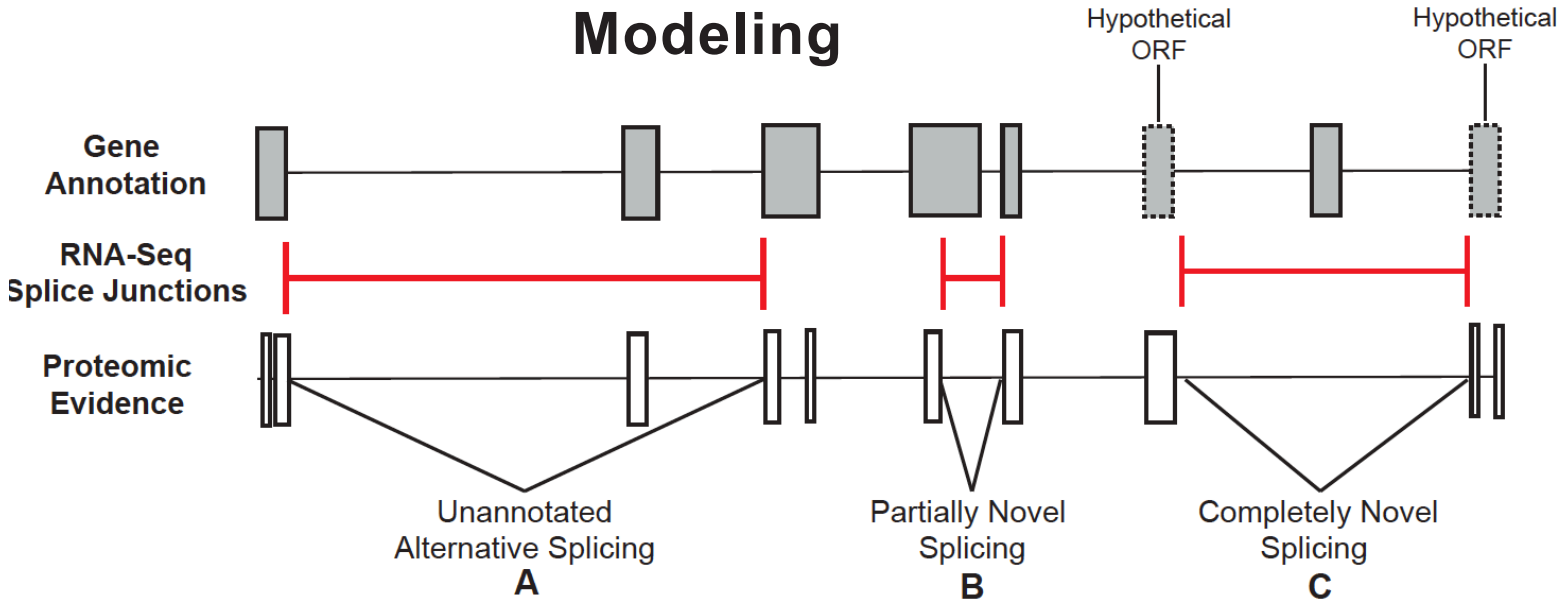


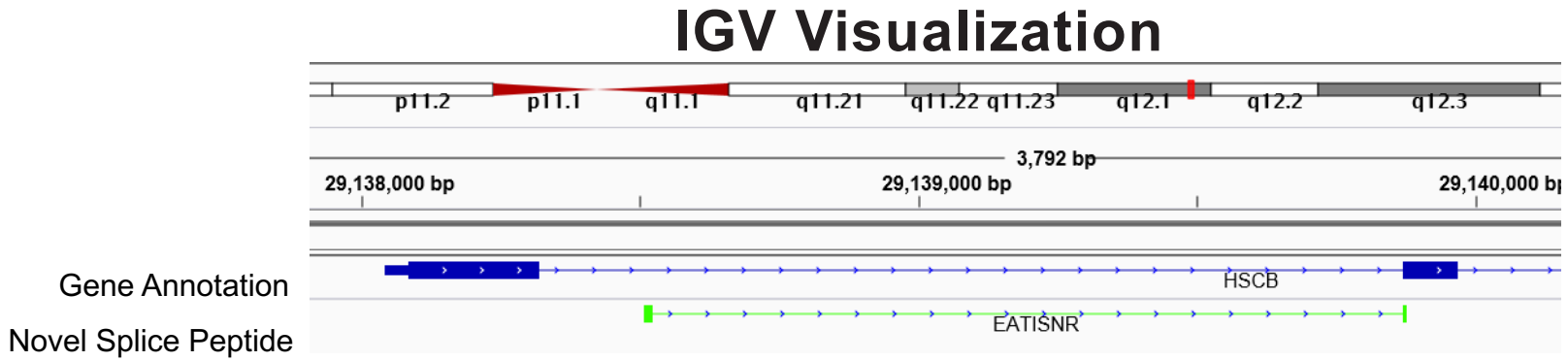
IGV Visualization



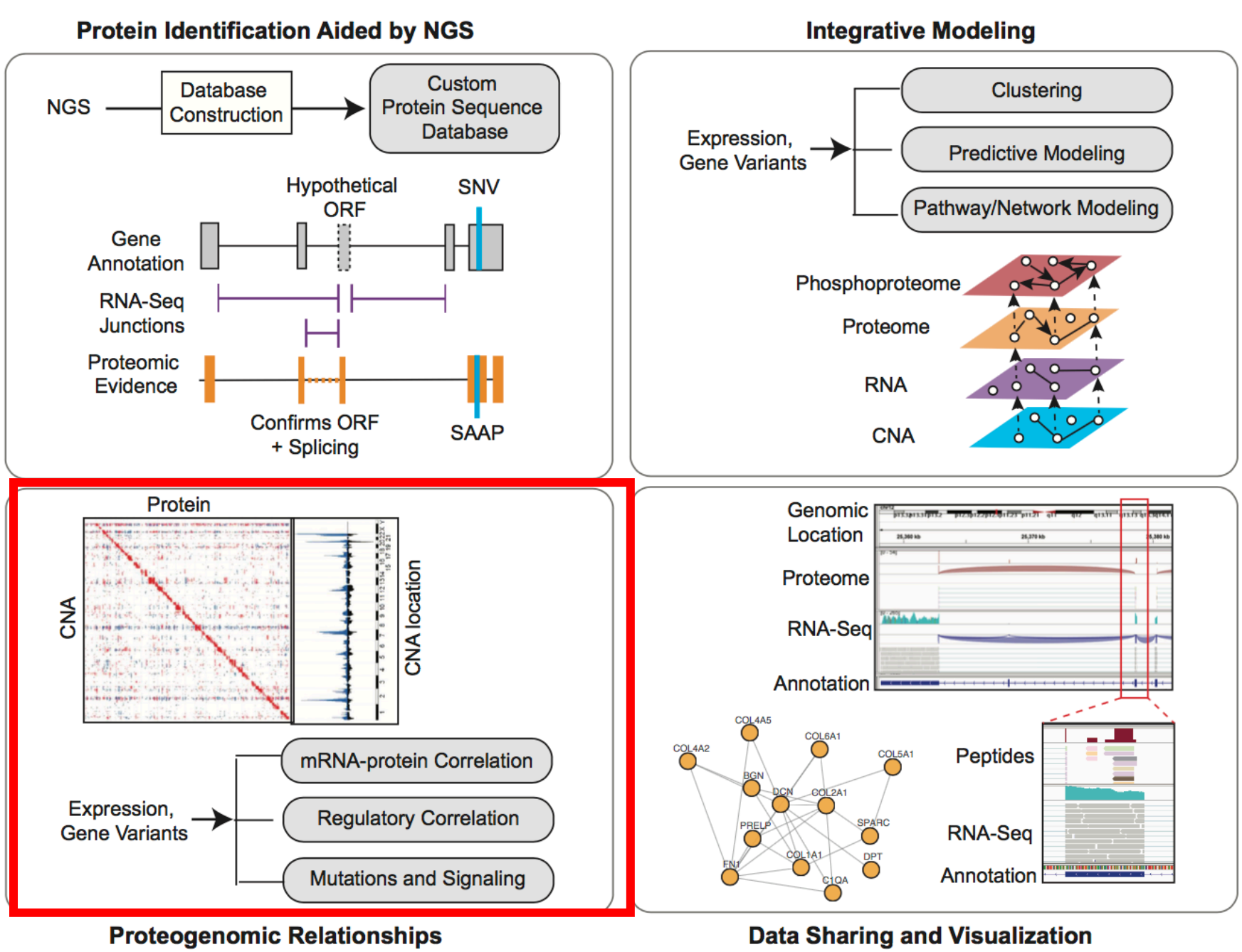
Ruggles & Fenyo, 2016
Ruggles & Fenyo, 2015

Proteogenomics and Novel Junction Discovery



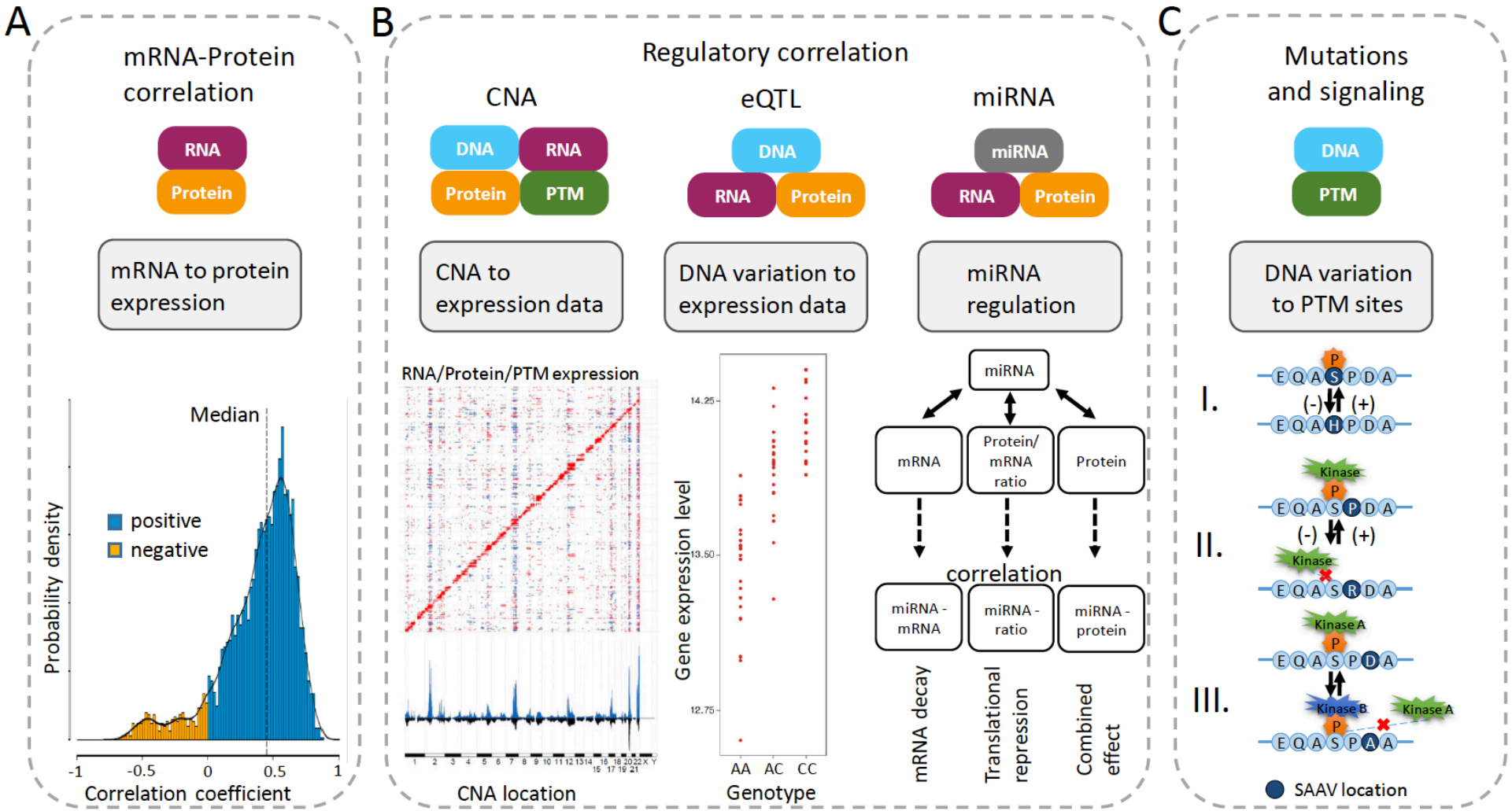


Ruggles & Fenyo, 2016
Ruggles & Fenyo, 2015



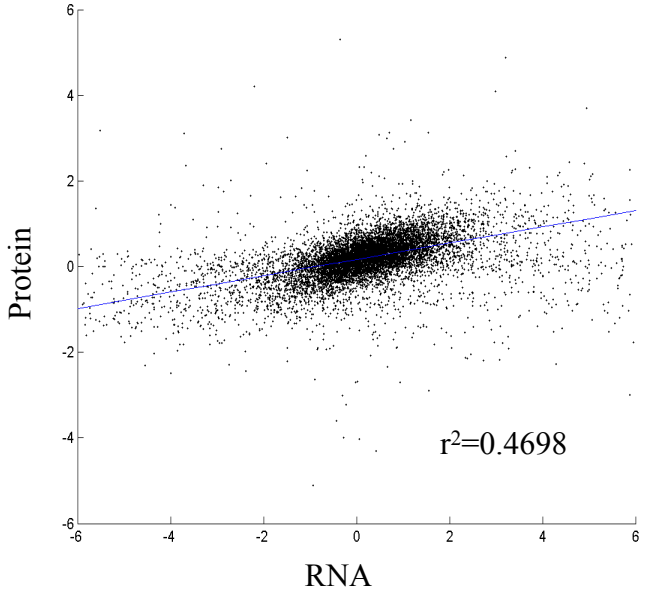
Ruggles et al., MCP 16(6), 959-981 (2017

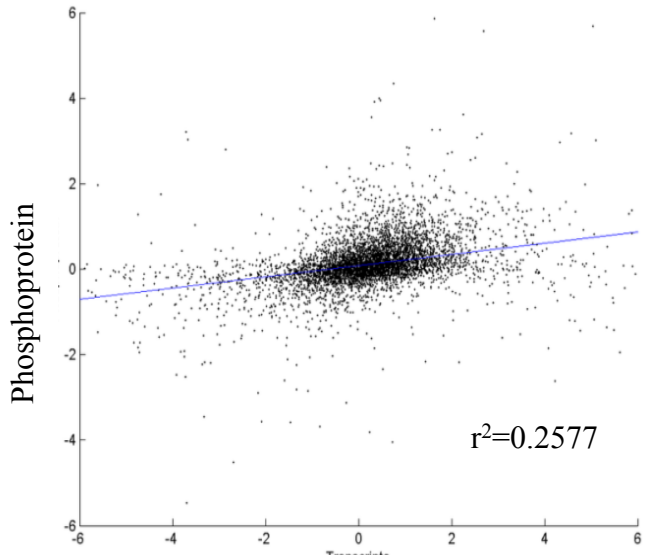
Proteogenomic Relationships

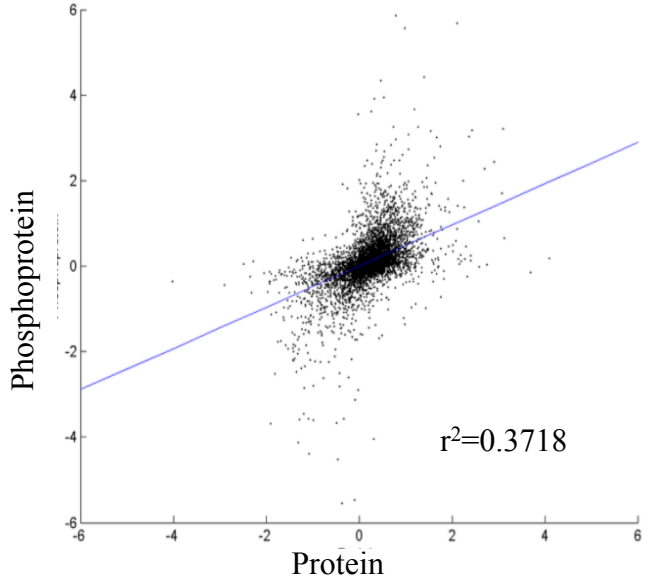


Ruggles et al., 2017

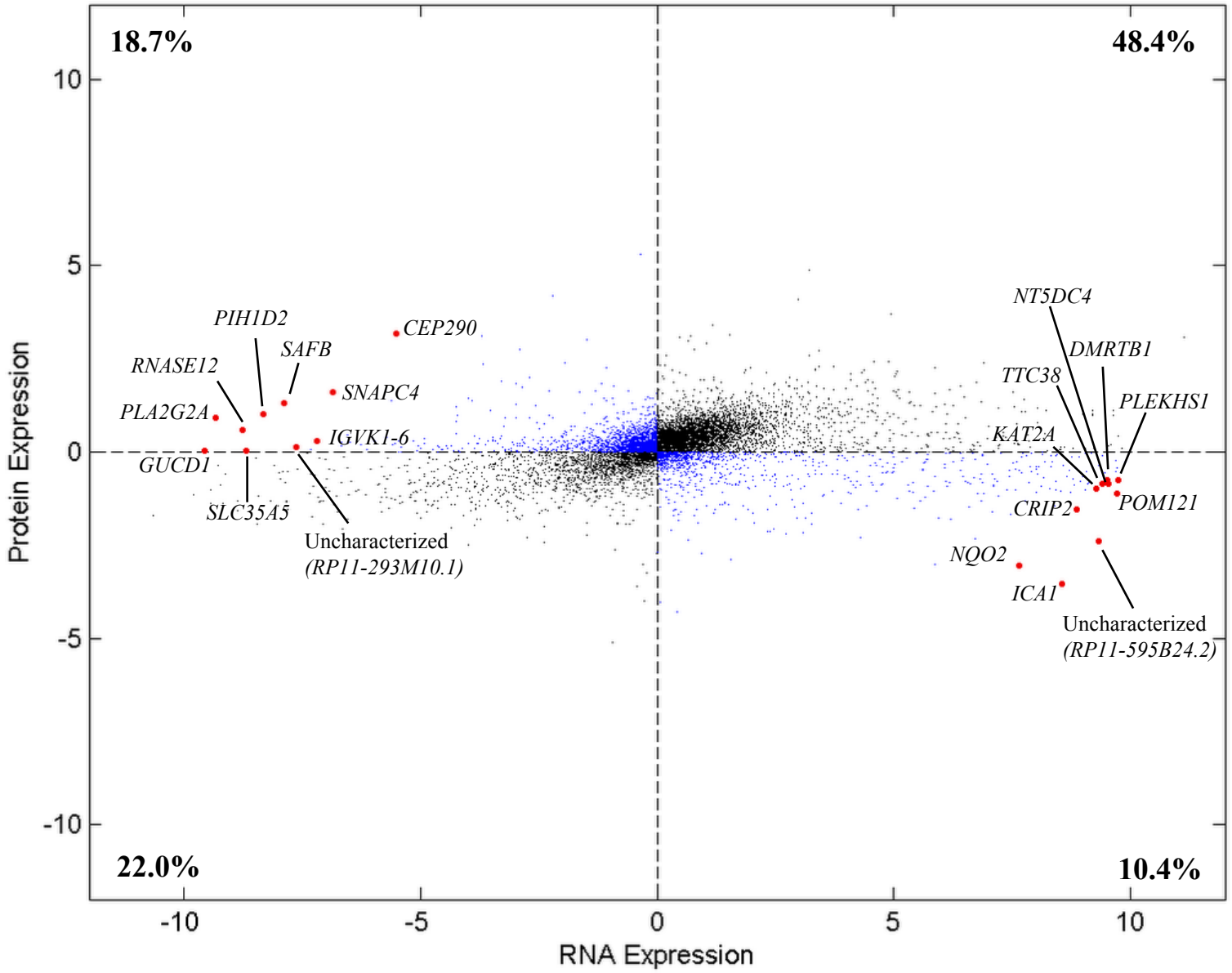
Association Tests Comparing Data Sets



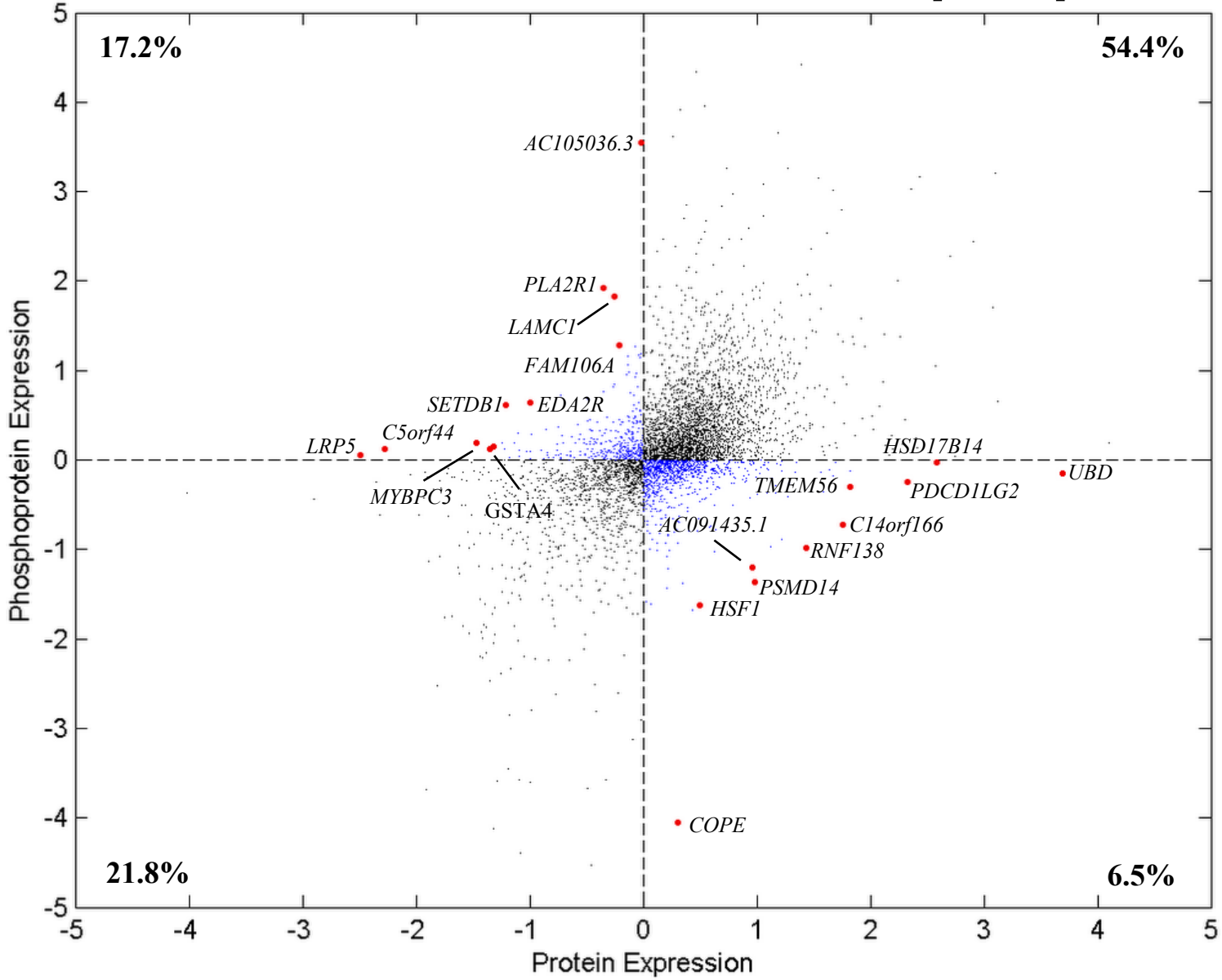




Genes with Differential RNA and Protein expression

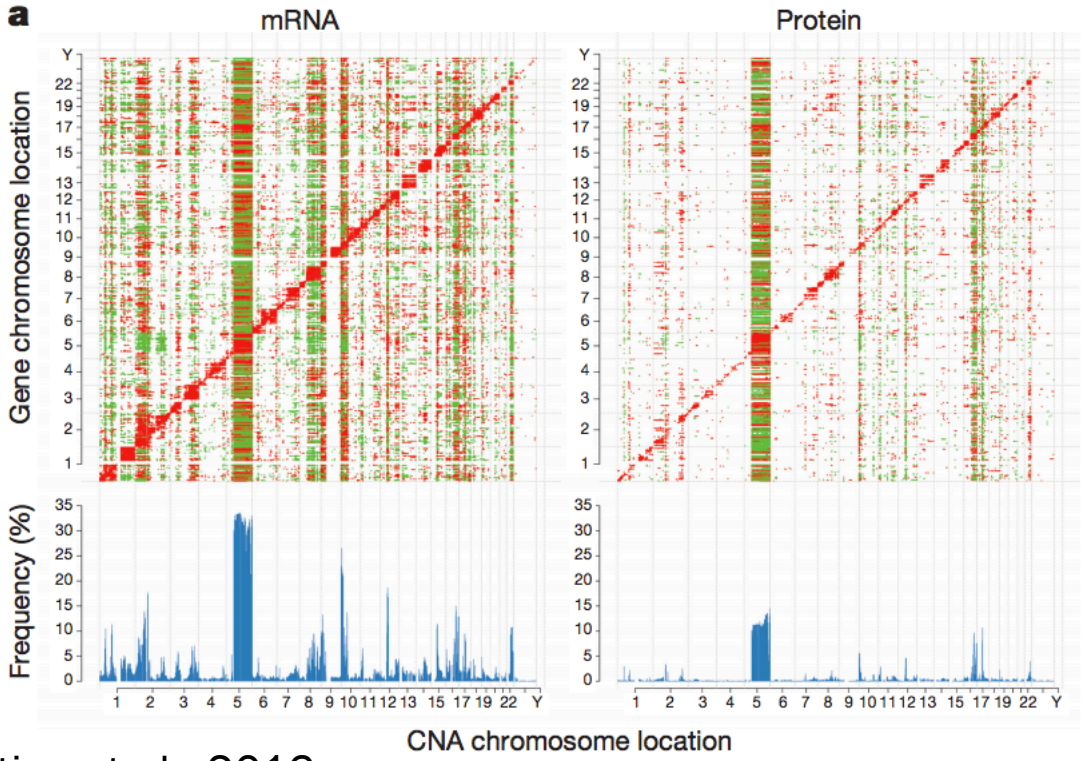


Genes with Differential Protein and Phosphoprotein Expression



Effect of CNA on protein abundance

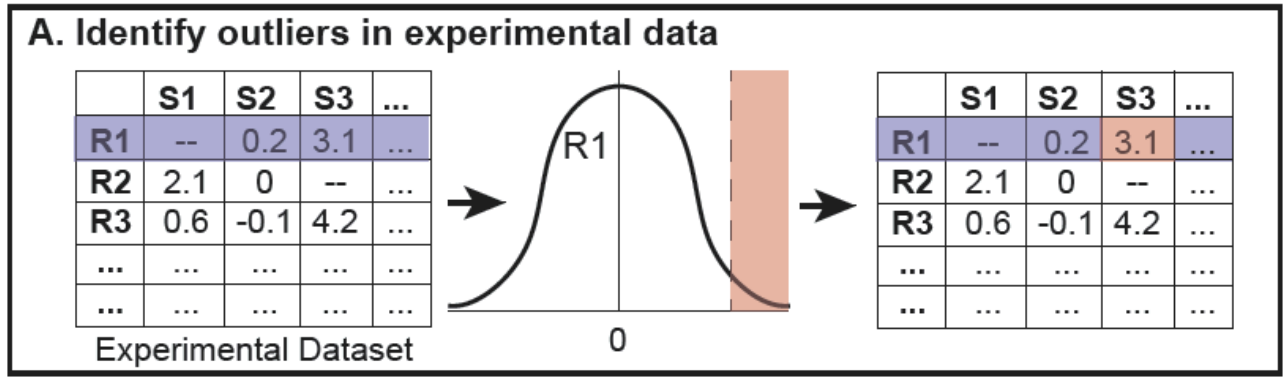
- Determine consequence of CNAs on mRNA and protein abundance both in ‘cis’ and ‘trans’ genes
- Used all genes with CNA, mRNA and protein measurements

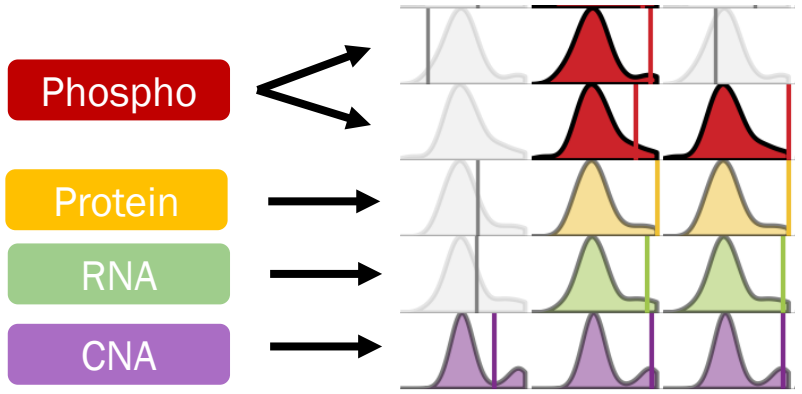


- Multiple test adjusted, Pearson correlation coefficient

Mertins et al., 2016

Identifying Aberrant Proteogenomic Events Using Outlier Analysis

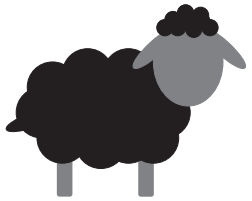




Outlier Status

1. Used log2 normalized data for 668 kinases from all 77 TCGA breast samples
2. Found distribution for each phosphosite across samples
3. Flag samples with normalized phosphosite expression above 1.5 interquartile ranges (IQR) from the median.
4. Repeat for CNA, RNA and protein expression

Black Sheep

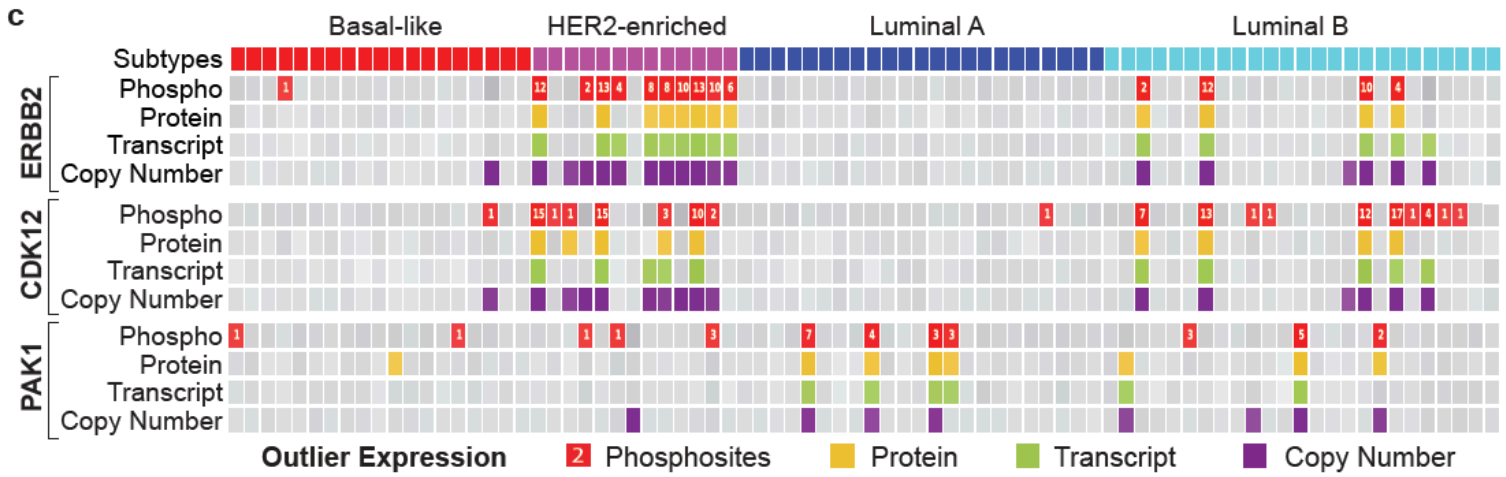


Kinase Outliers

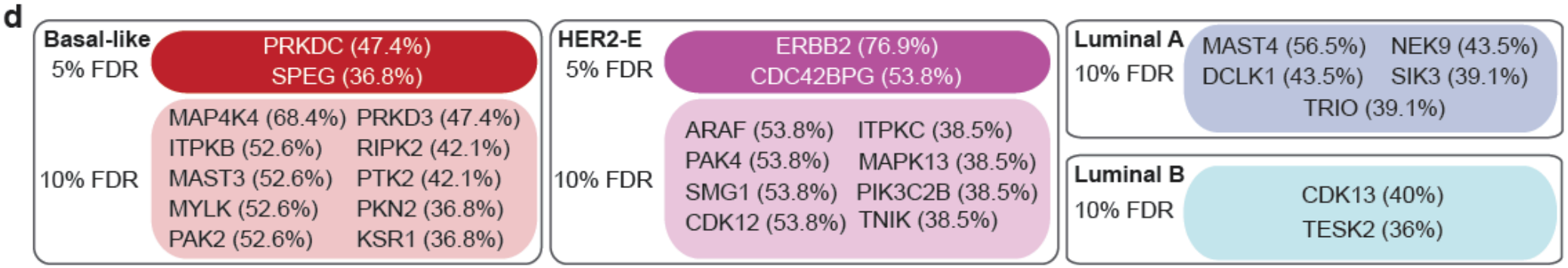
Subtype enrichment
Druggable Drivers

Phosphosite Outlier Enrichment in Breast Cancer Subtypes

181 phosphosite outlier kinases identified

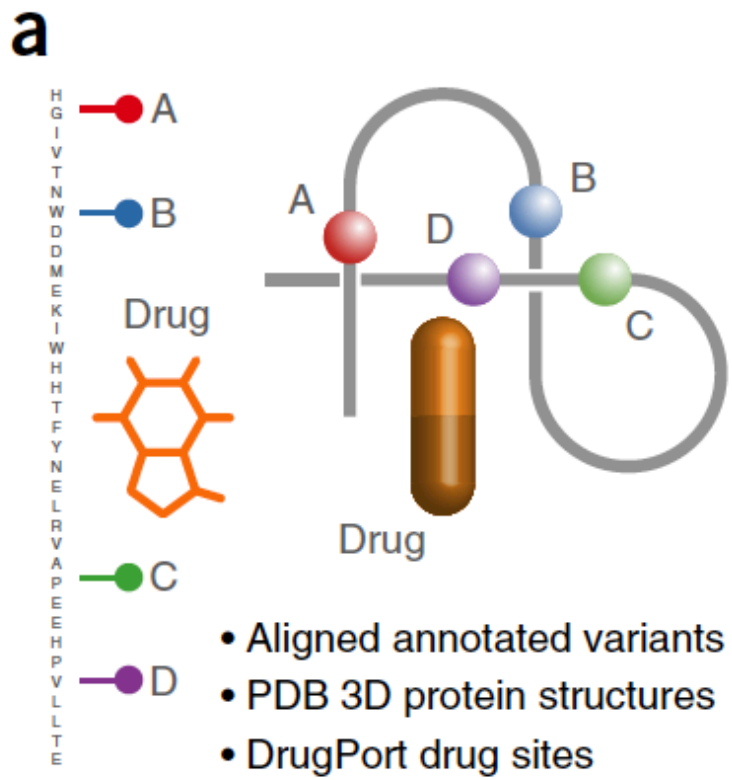


Which phosphosite outlier kinases are enriched in the 4 represented subtypes?

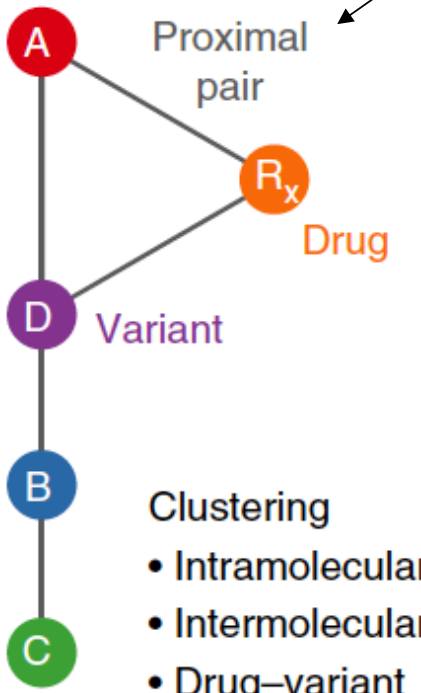


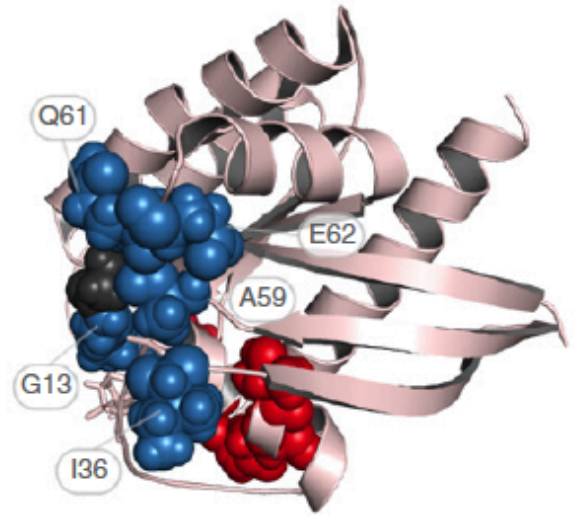
Mertins P*, Mani DR*, Ruggles KV*, Gilette M* et al., Nature 534, 55-62 (2016)

HotSpot3D



Things that are in close proximity in protein structure

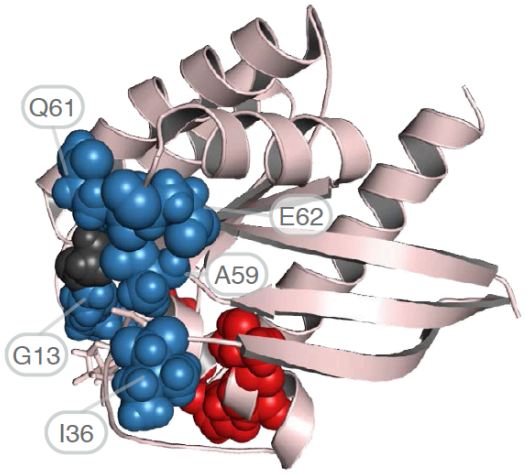




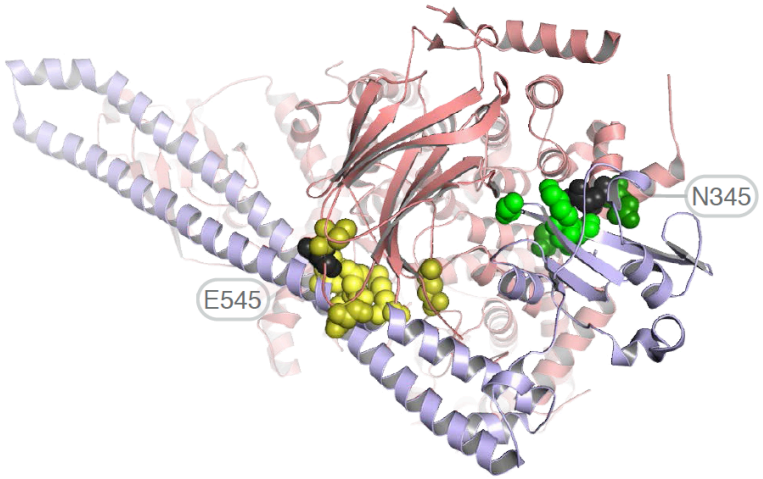
Sequence variants and drug binding are mapped to protein structure	Pairwise correlations used to determine the impact of variants on drug response	Validate the impact of these variants in disease models

Niu, Scott*, Sengupta* et al., Nature Genetics (2016)*

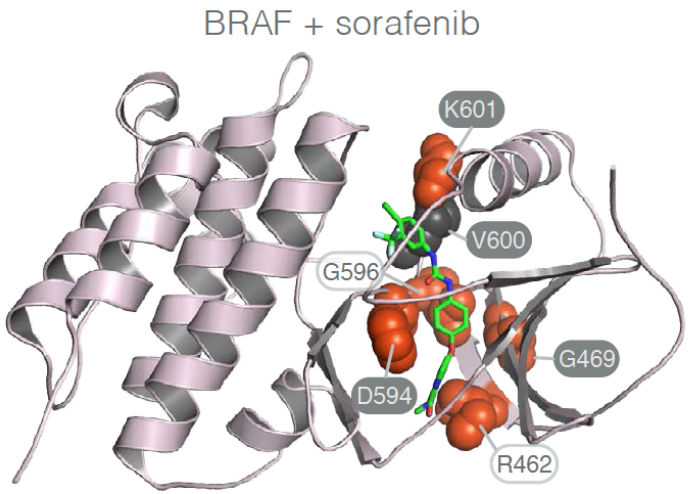
HotSpot3D



Intra-molecular Clusters



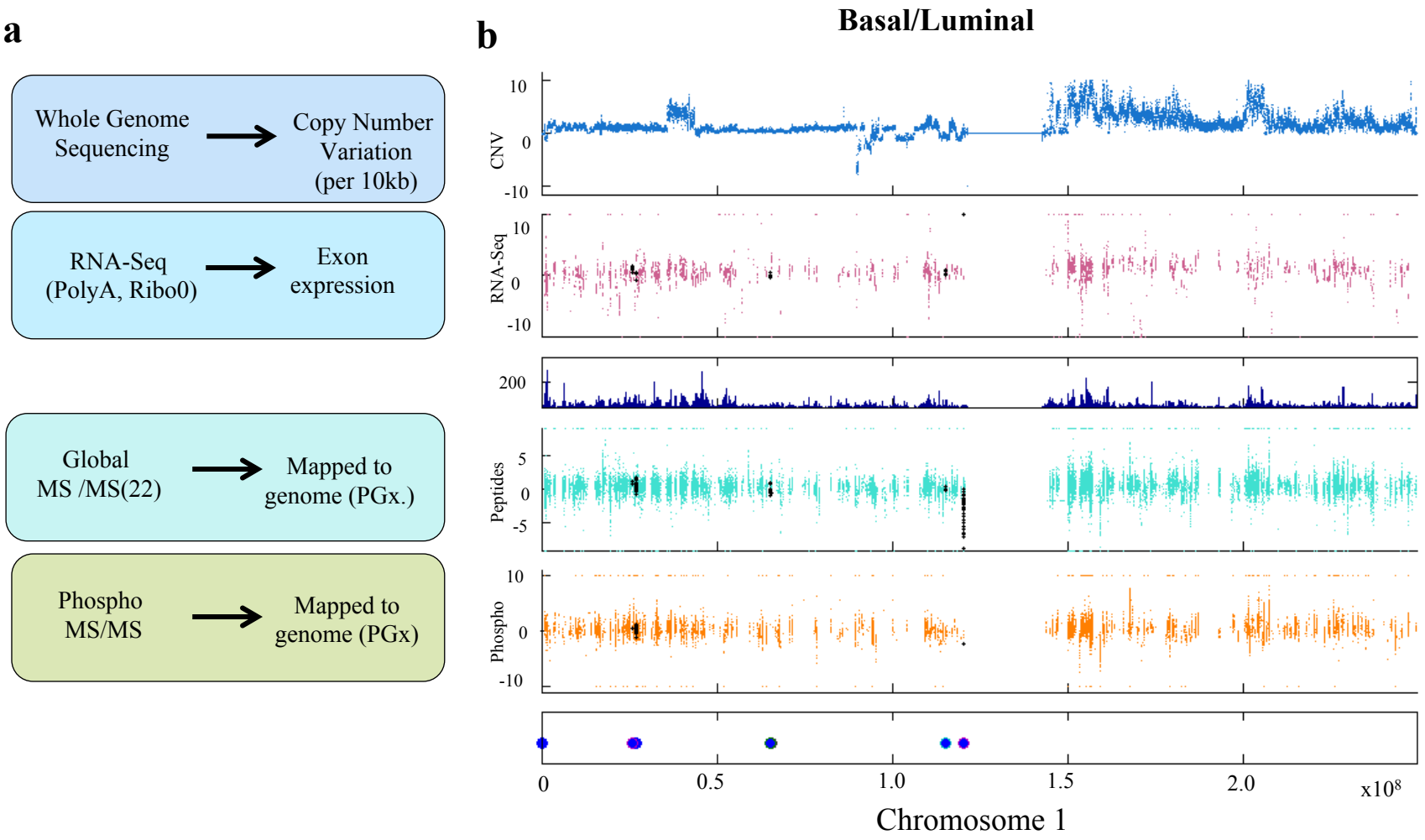
Inter-molecular Clusters



Mutations clustering around Drug binding pockets

Niu, Scott*, Sengupta* et al., Nature Genetics (2016)*

Proteogenomic Mapping



Proteogenomic Mapping

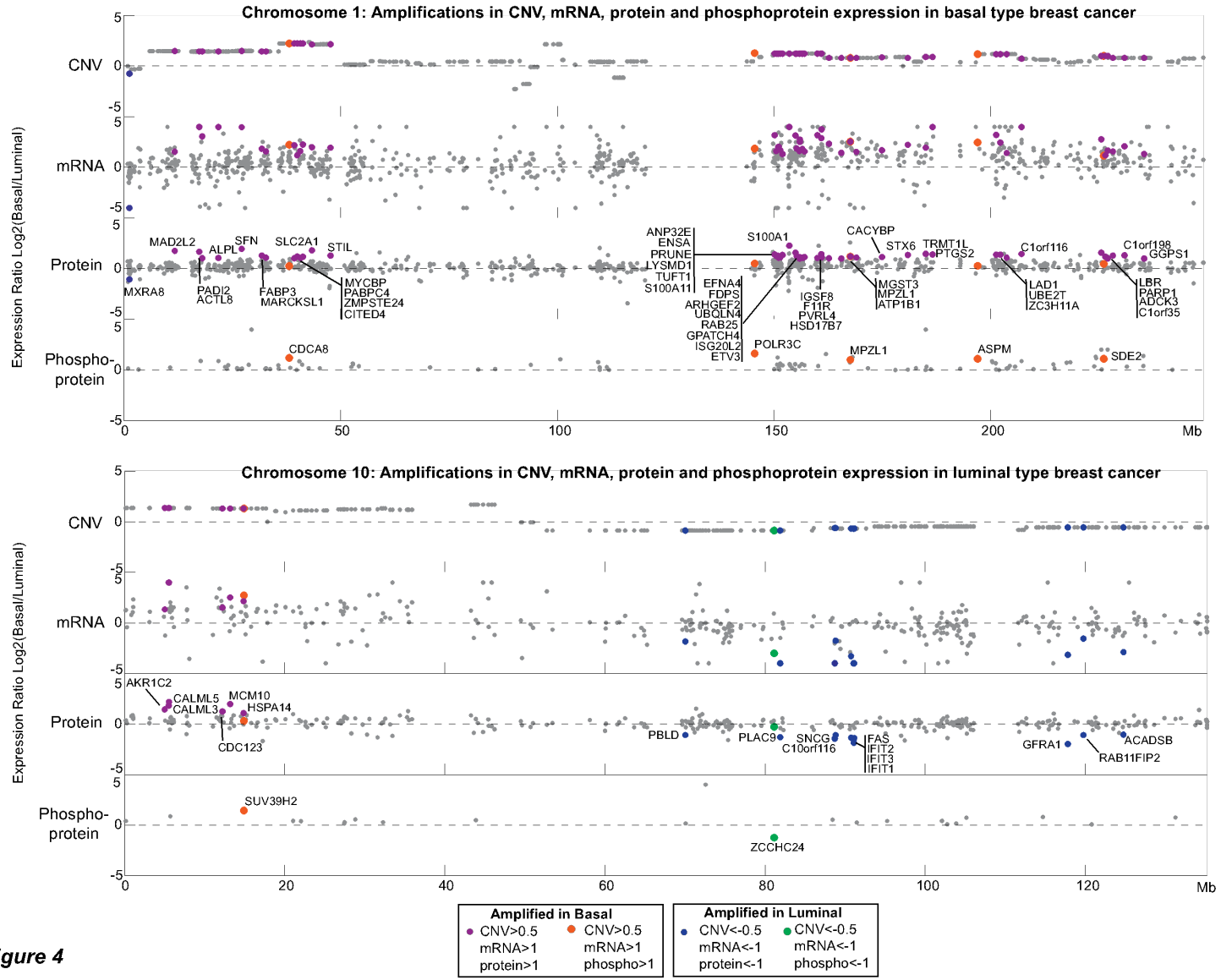
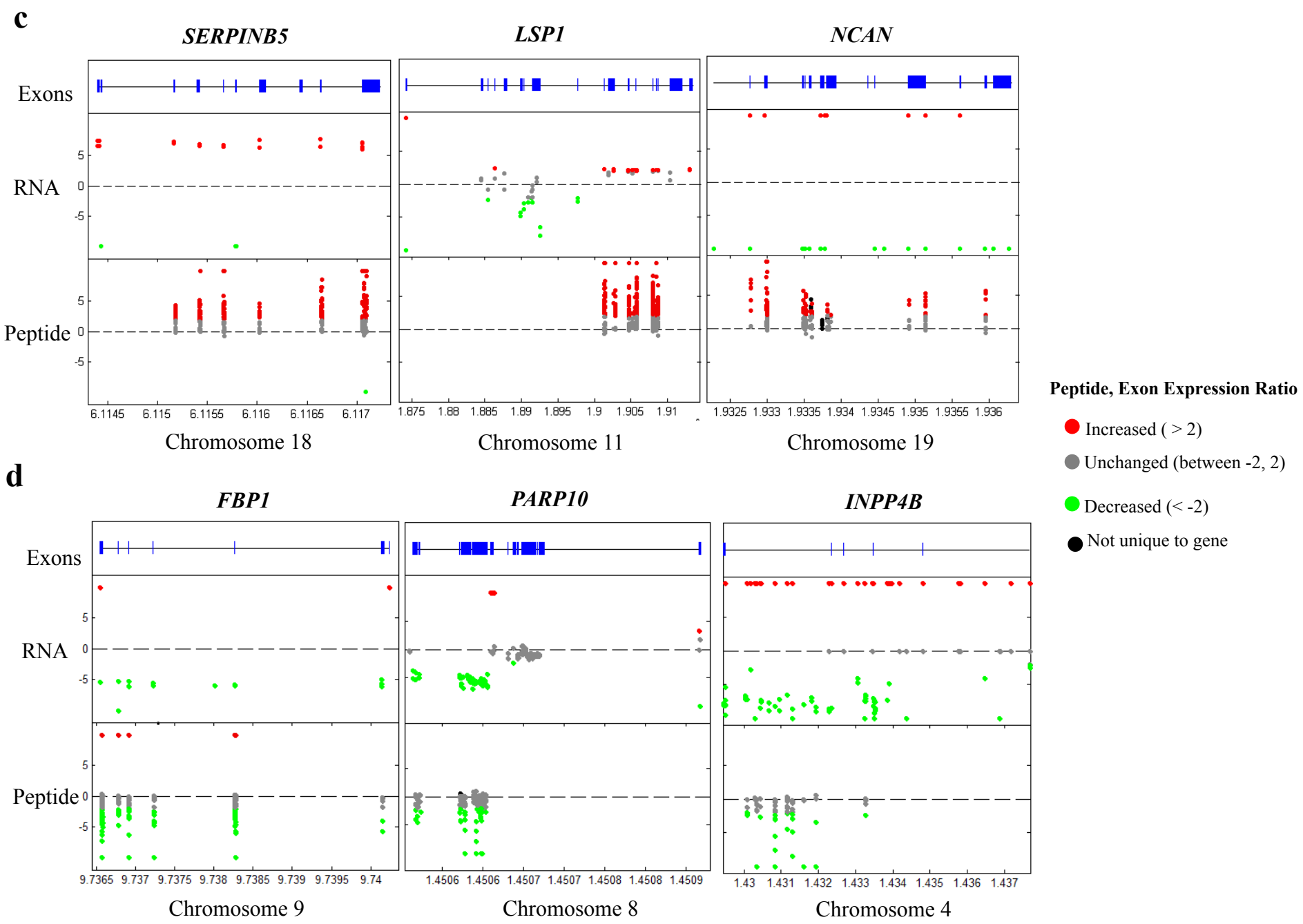
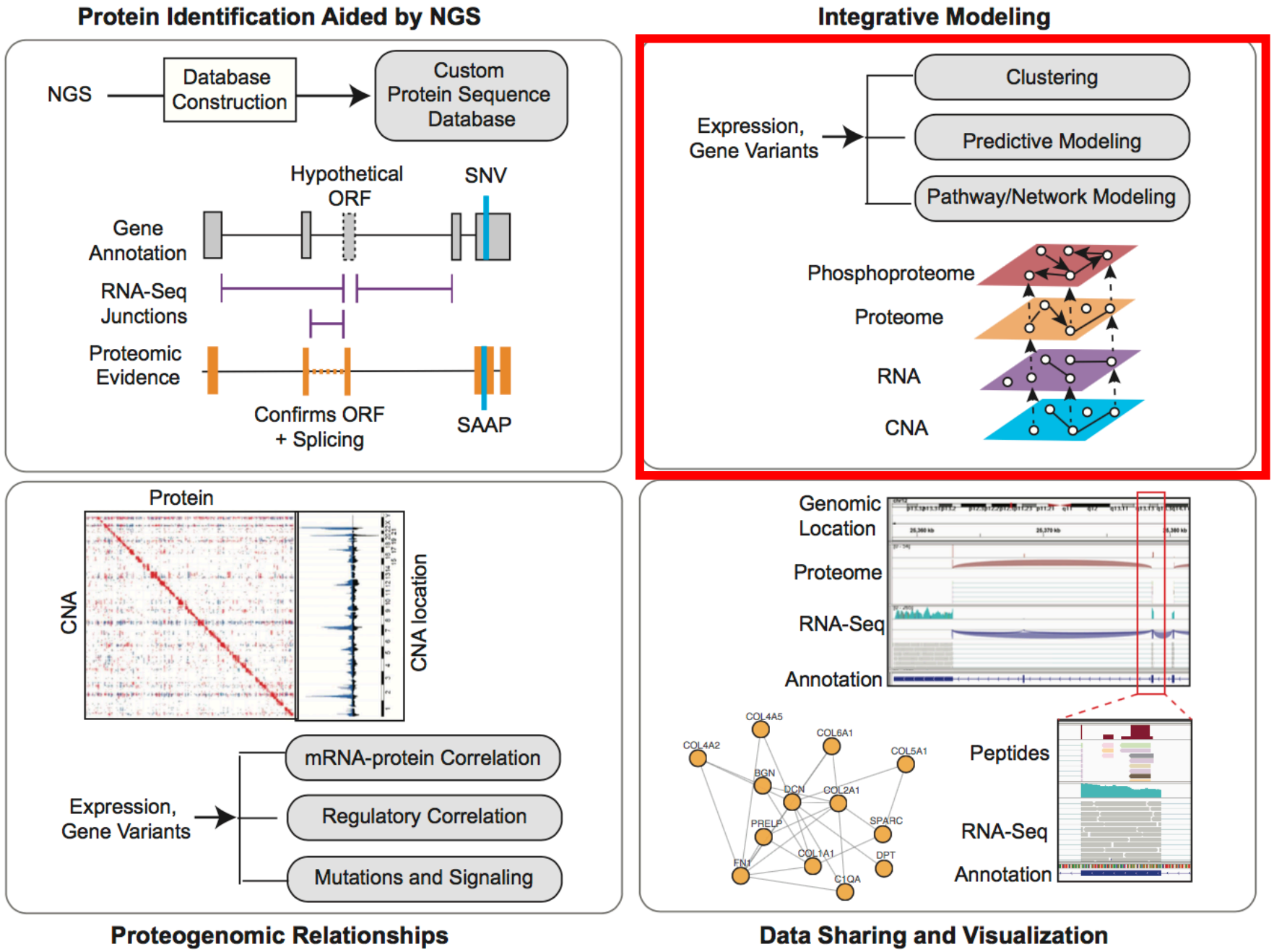
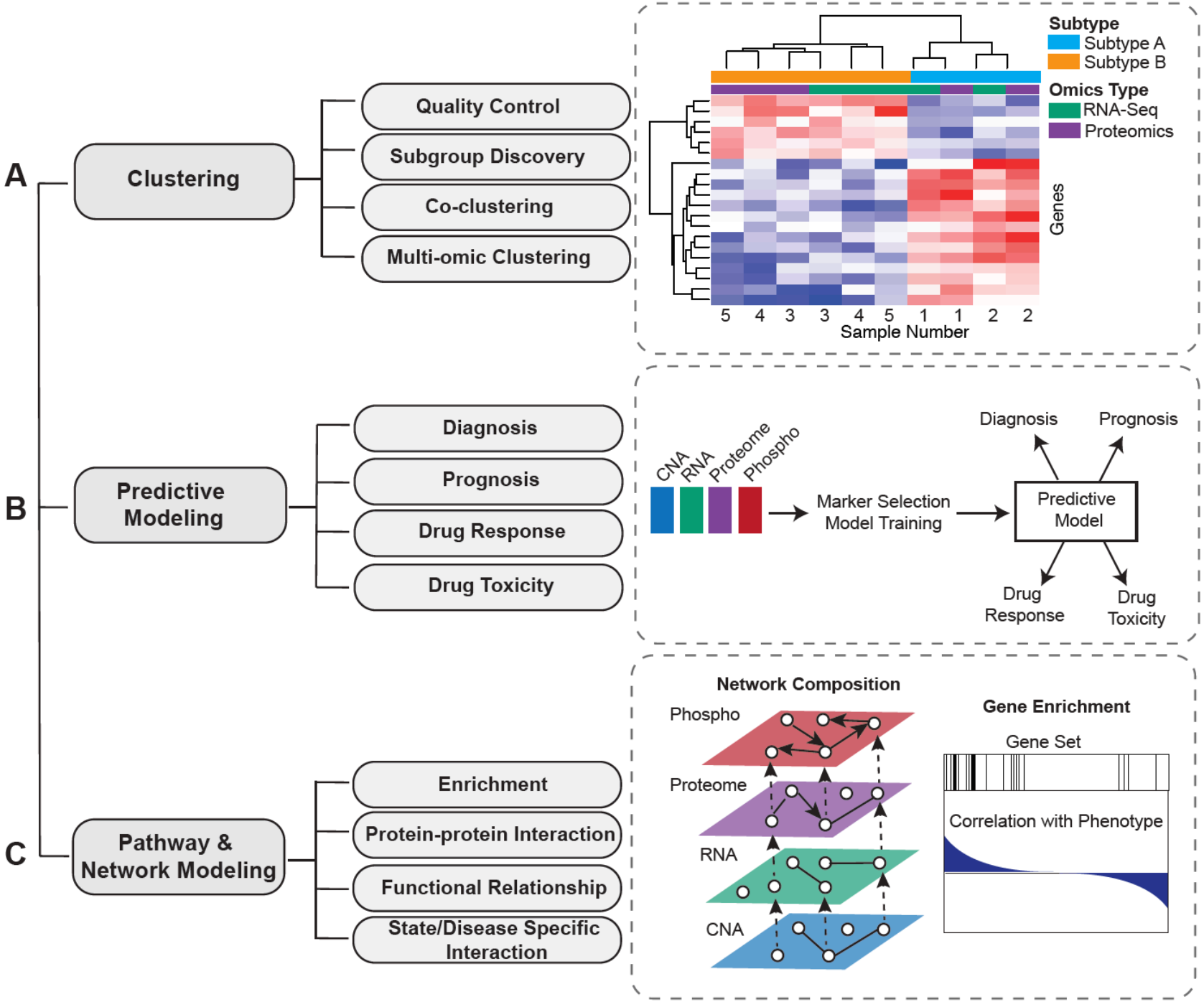


Figure 4


c
SERPINB5
LSP1
NCAN
Exons
RNA
Peptide
5
0
-5
6.1145 6.115 6.1155 6.116 6.1165 6.117
1.875 1.88 1.885 1.89 1.895 1.9 1.905 1.91
1.9325 1.933 1.9335 1.934 1.9345 1.935 1.9355 1.936
Chromosome 18
Chromosome 11
Chromosome 19
d
FBP1
PARP10
INPP4B
9.7365 9.737 9.7375 9.738 9.7385 9.739 9.7395 9.74
1.4506 1.4506 1.4507 1.4507 1.4508 1.4508 1.4509
1.43 1.431 1.432 1.433 1.434 1.435 1.436 1.437
Chromosome 9
Chromosome 8
Chromosome 4
Peptide, Exon Expression Ratio
Increased (> 2)
Unchanged (between -2, 2)
Decreased (< -2)
Not unique to gene



Ruggles et al., MCP 16(6), 959-981 (2017

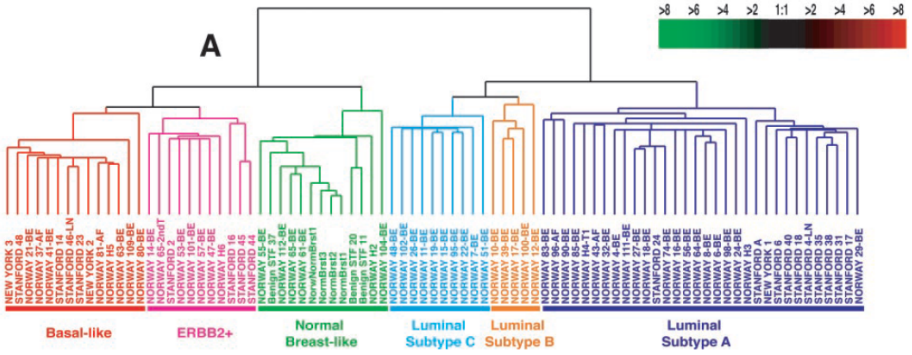


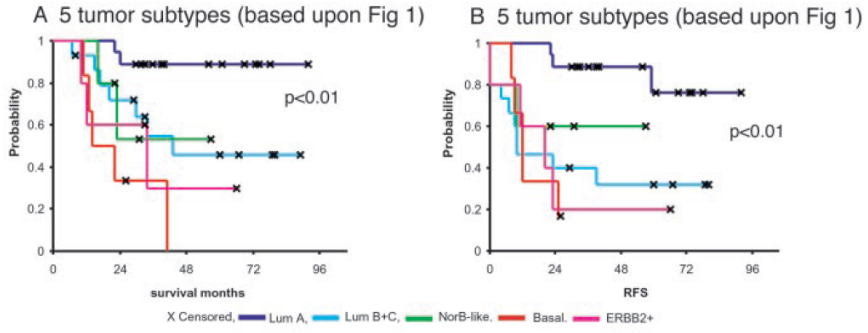
Ruggles et al., 2017

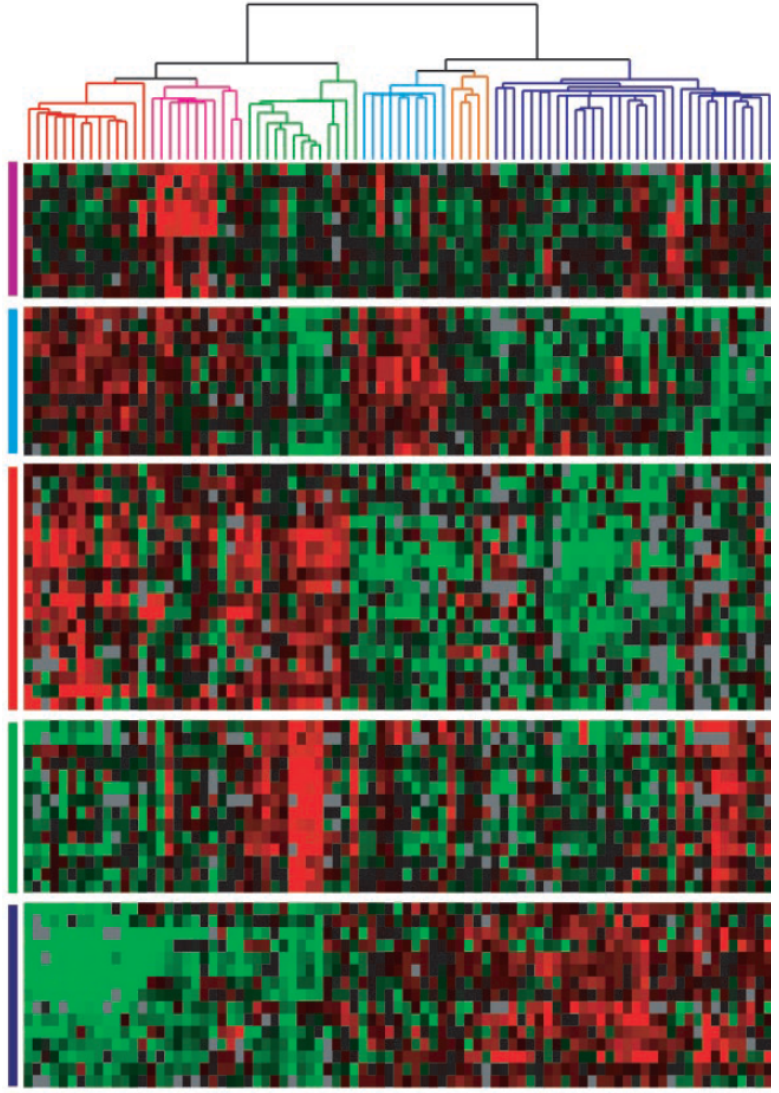
Unsupervised Learning: Unlabeled Data

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

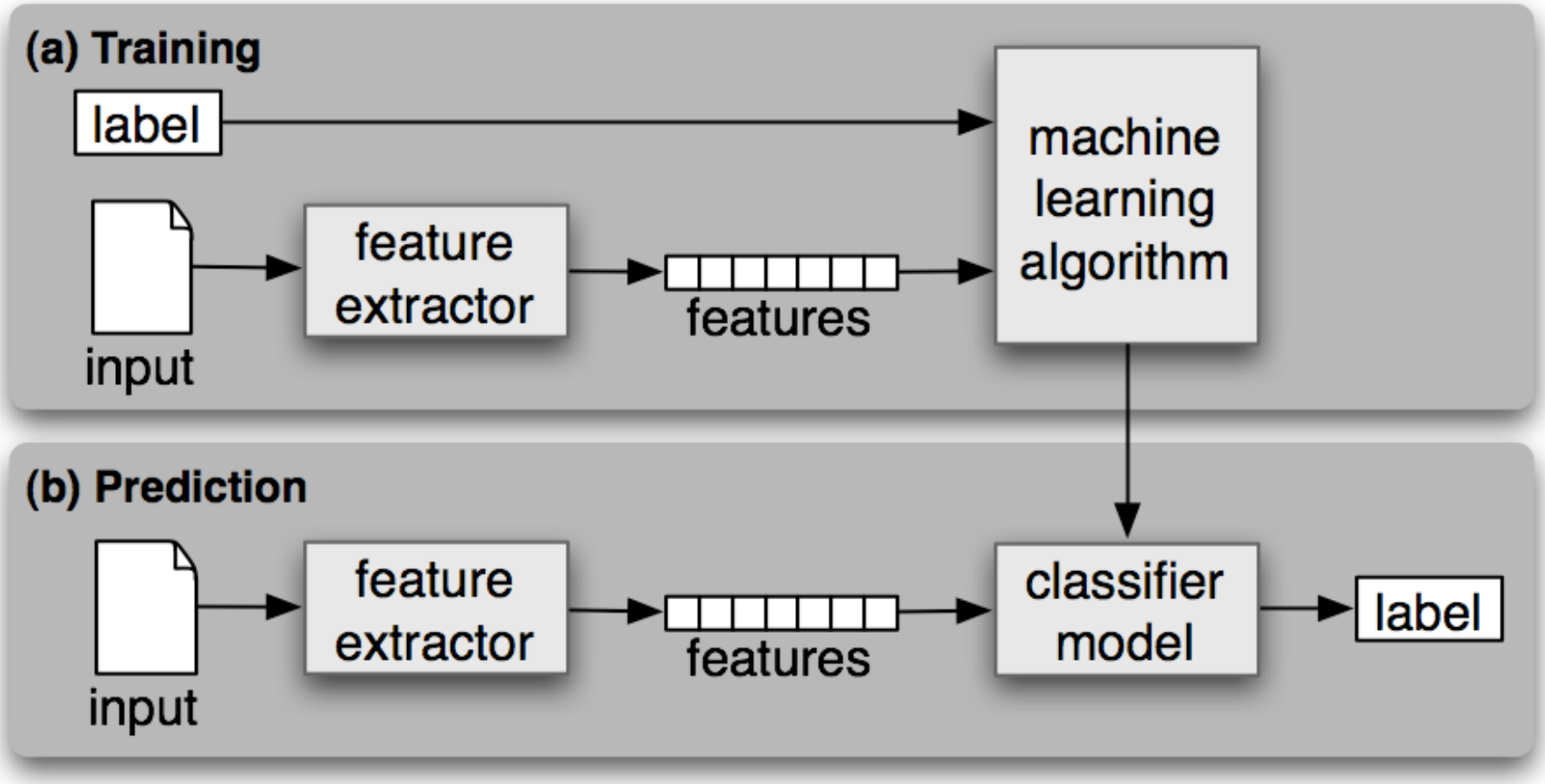
Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]





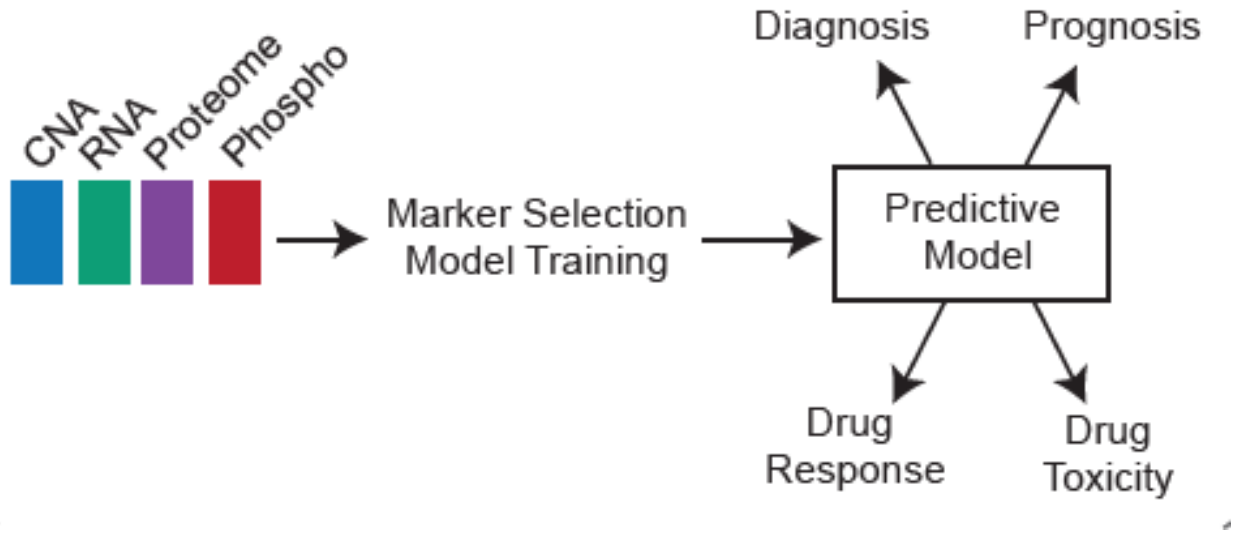


Supervised Learning: Labeled Data



Machine Learning and Disease Phenotypes

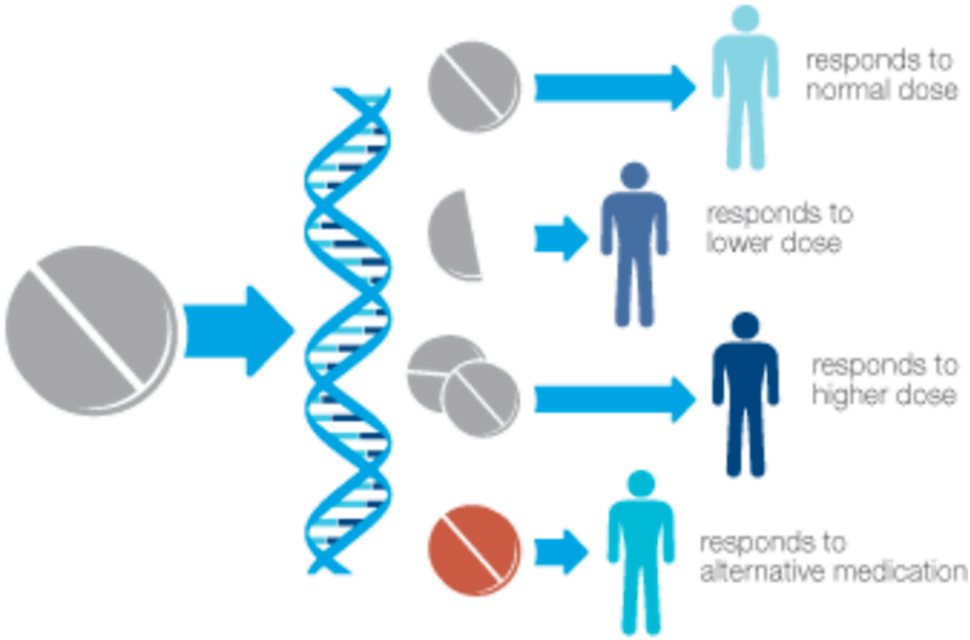
- Input can also be expression matrices
  - RNA-seq
  - DNAse-seq
  - ChIP-seq
  - Microarray
  - Proteomics etc.
- Can be used to distinguish between disease phenotypes and/or to identify potentially valuable disease biomarkers



Ruggles et al. (2017) Methods, tools and perspectives in proteogenomics. MCP.

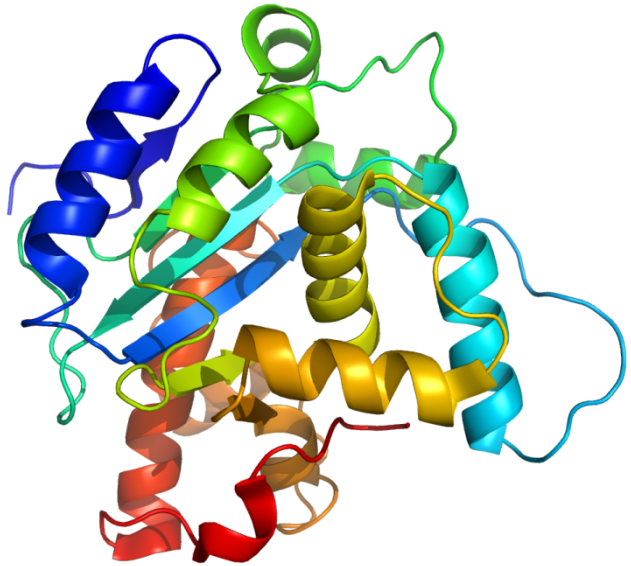
Personalized Medicine

- Personalized medicine: algorithm that optimizes treatment to maximize efficacy and minimize risk based on genetic make-up
- Patient populations show high inter-individual variability in drug response and toxicity.
- Gene factors account for 15-30% of drug metabolism differences
- Ability to identify gene biomarkers corresponding to a therapeutic effect



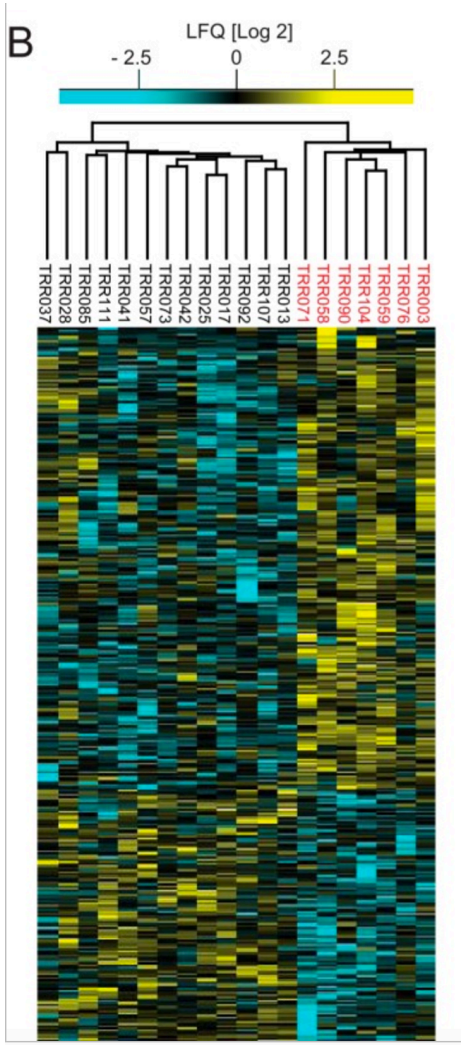
Machine Learning in Multiomics

- One would expect the predictive analysis of proteome and phosphoproteome data to be more informative regarding clinical outcomes compared to NGS data, as these data modalities are more proximal to the disease.
- These techniques have been applied to proteomics data to
    1. Classify clinically-relevant disease subtypes in cancer
    2. Define prognosis
    3. Identify biomarkers predicting drug sensitivity



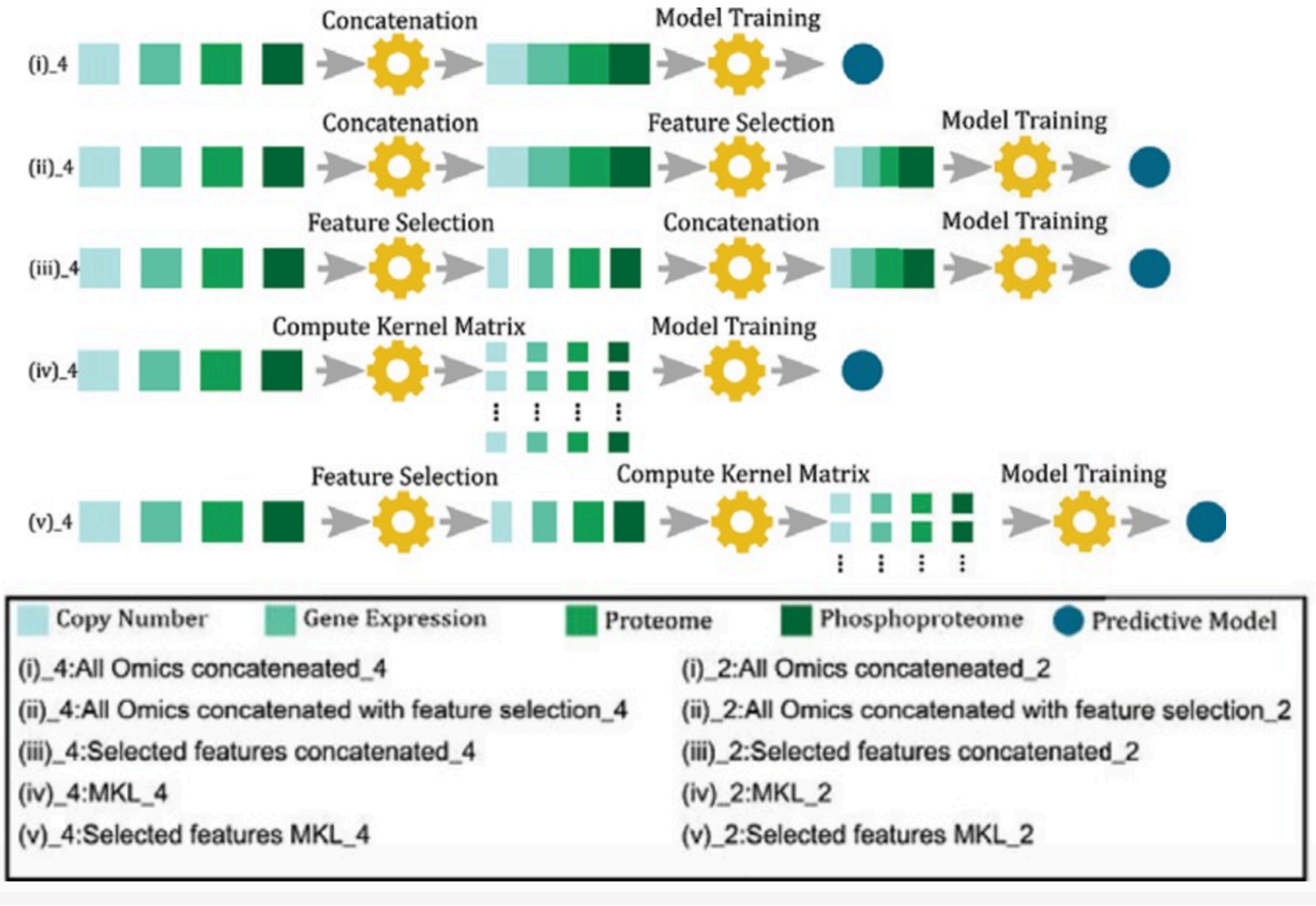
Can we accurately classify patients using protein expression?

- Deeb et al. used global expression patterns from shotgun proteomics
  - ~9000 tumor proteins
  - 20 Large B-Cell lymphoma patients
- Used SVMs to extract candidate proteins with highest segregating power
- Identified four proteins (PALD1, MME, TNFAIP8 and TBC1D4) to accurately classify Large B-Cell lymphoma patients, which are usually morphologically indistinguishable



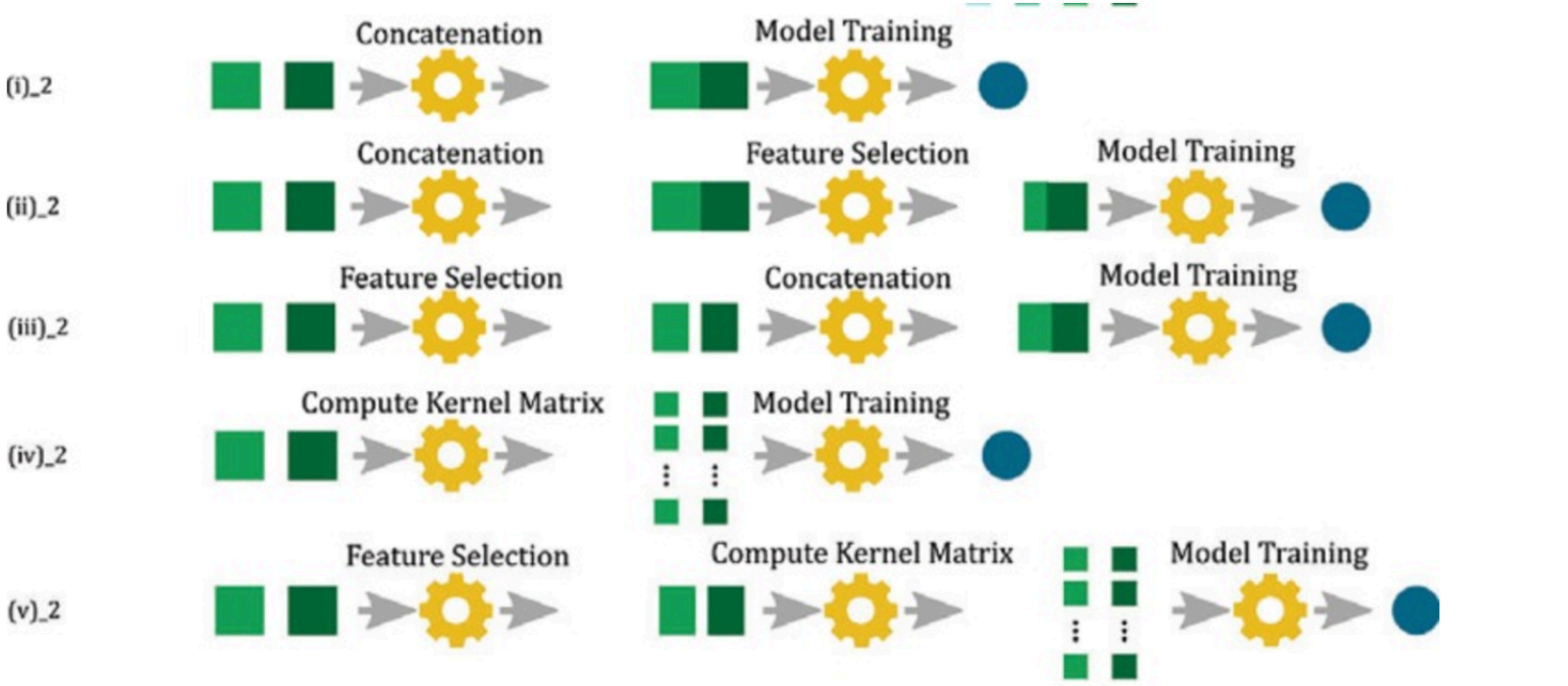
Deeb, et al. (2015) *Mol. Cell. Proteomics MCP*. **14**, 2947–2960

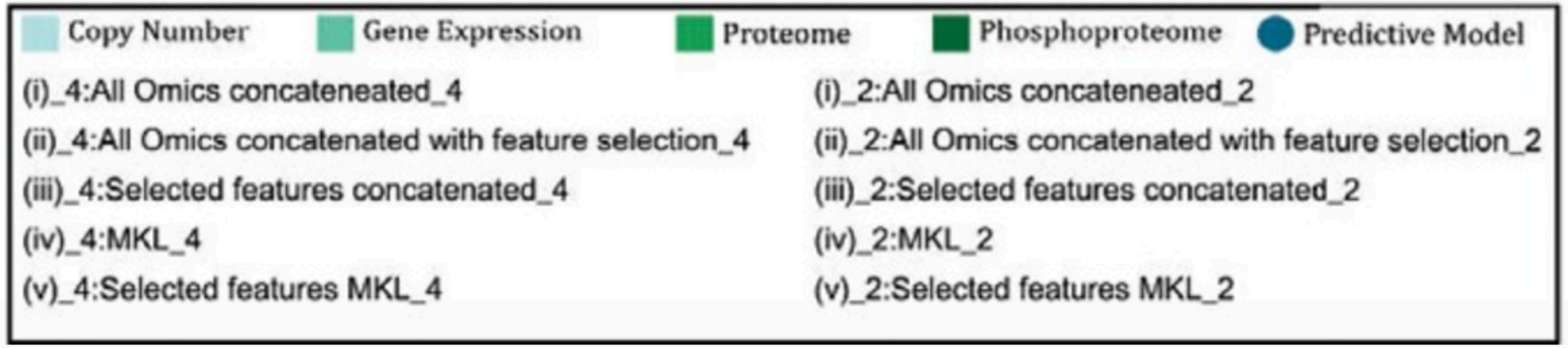
Data Integration Strategies



Ma, S et al. (2016) *AMIA Summits Transl. Sci. Proc.* **2016**,

Data Integration Strategies continued

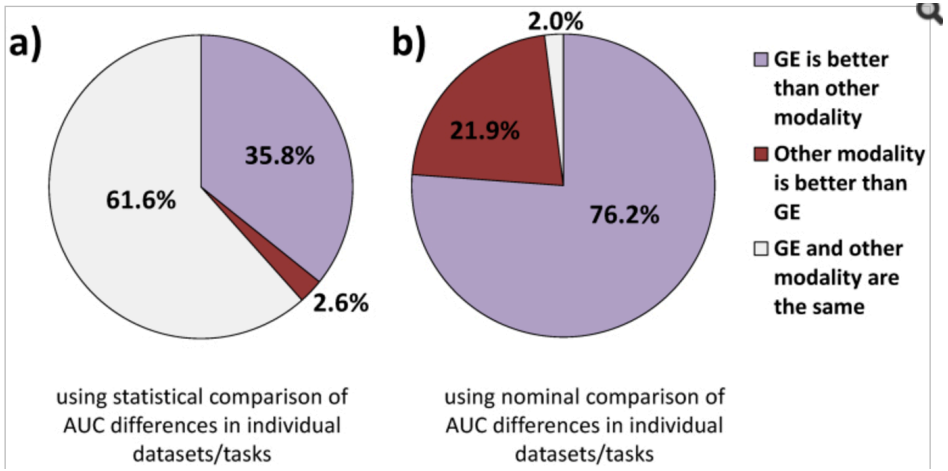




Ma, S et al. (2016) *AMIA Summits Transl. Sci. Proc.* **2016**,

Does multimodal analysis increase predictive power?

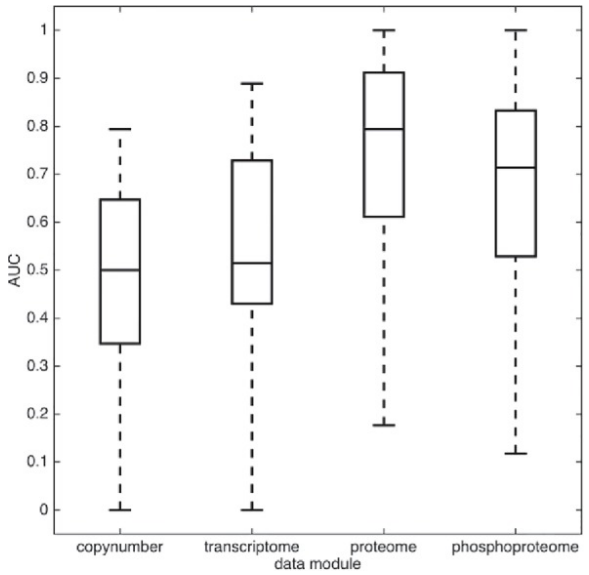
- Ray et al. used unimodal and “multi-modal” approaches to predict clinical phenotypes using
  - RNA-Seq, gene expression, and Reverse Phase Protein Array (RPPA)
- Found no advantage to combining data modalities compared to individual platform analysis
- Gene expression data was consistently more predictive than RPPA-based proteomics



Ray, B., et al. (2014) *Sci. Rep.* **4**, 4411

Does multimodal analysis increase predictive power? Take 2

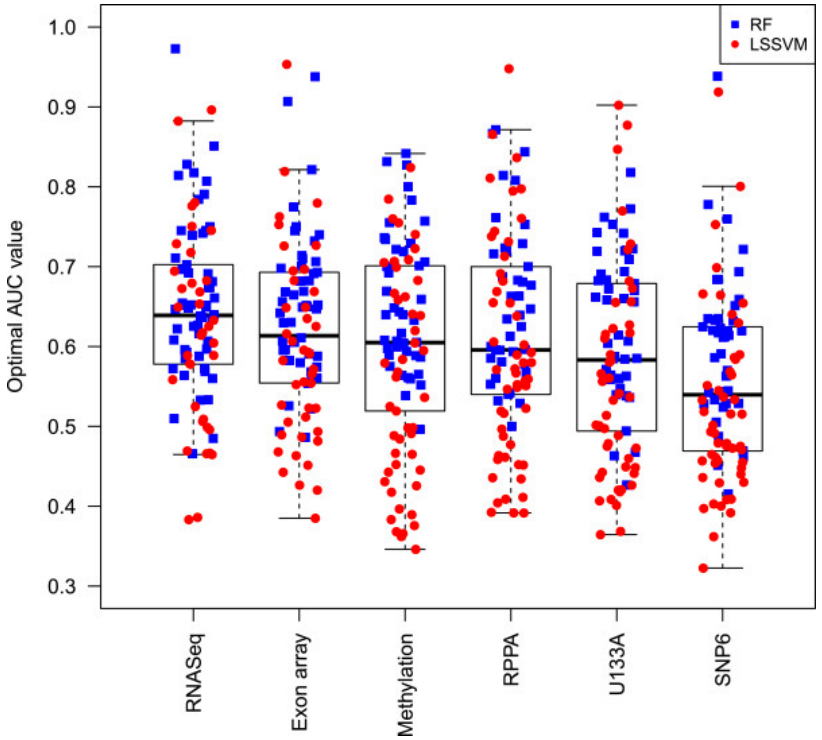
- Ma et al used proteogenomics data from 77 breast tumors to predict 10 year survival in breast cancer
- Found that fusion of 4 data types did not improve model performance
- Proteomics outperformed genomics and transcriptomics



Ma, S et al. (2016) *AMIA Summits Transl. Sci. Proc.* **2016**, 52–59

Can we identify markers of drug response in cancer?

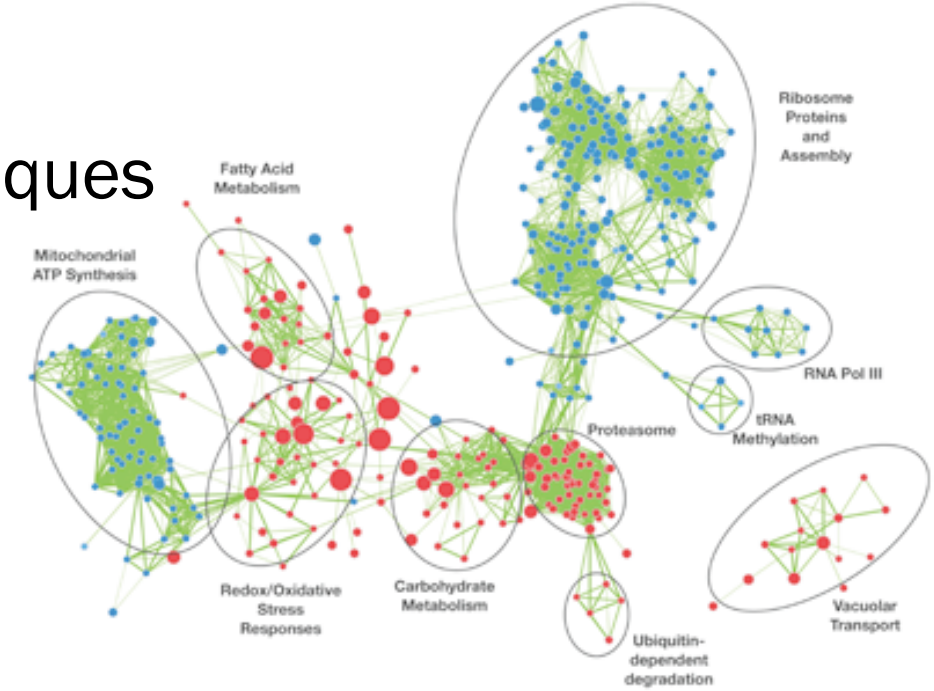
- Daemen et al, used an SVM and Random Forest approach to identify molecular features associated with drug response of 90 drugs in 70 breast cancer cell lines.
- Input data was CNA, mutations, gene expression, promoter methylation and protein expression
- Found that RNA-expression had the best prediction but other data types improved the prediction in a subset of cases

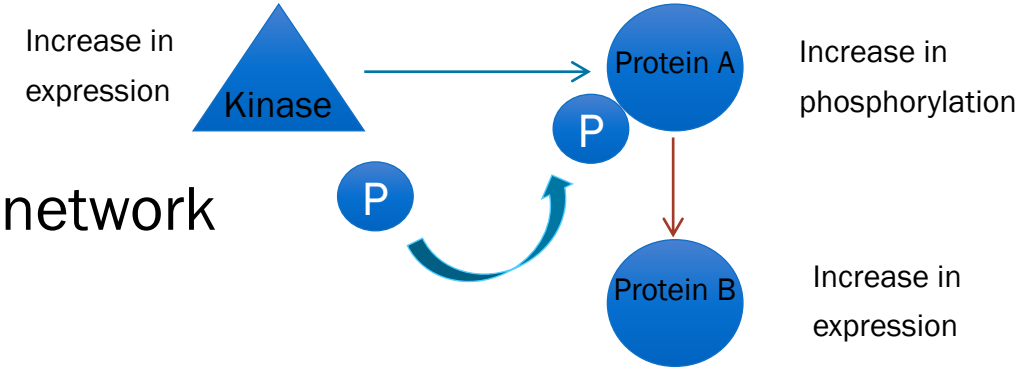


Daemen et al. (2013) *Genome Biol.* **14**, R110

Pathway and Network Analysis

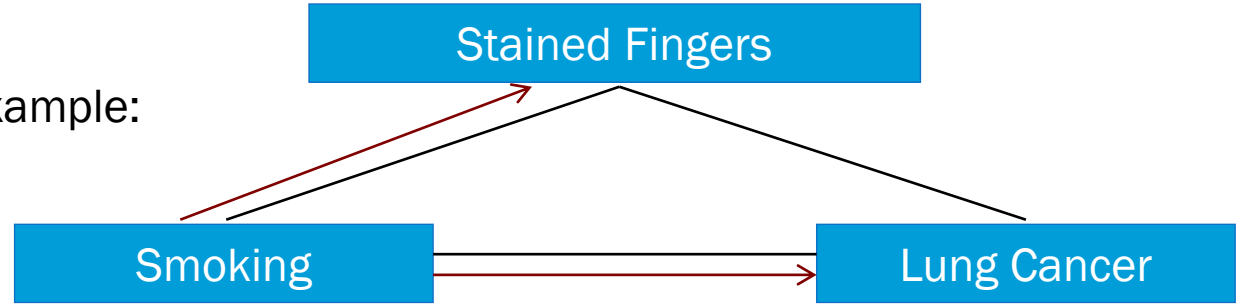
- Classical pathway analysis techniques
  - KEGG
  - Pathway Studio
  - IPA
- Network analysis
  - Cytoscape
  - GSEA
- **Causal Discovery**
  - PC algorithm
  - Markov Blanket/Bayesian network





Causal Discovery and Cancer Signaling

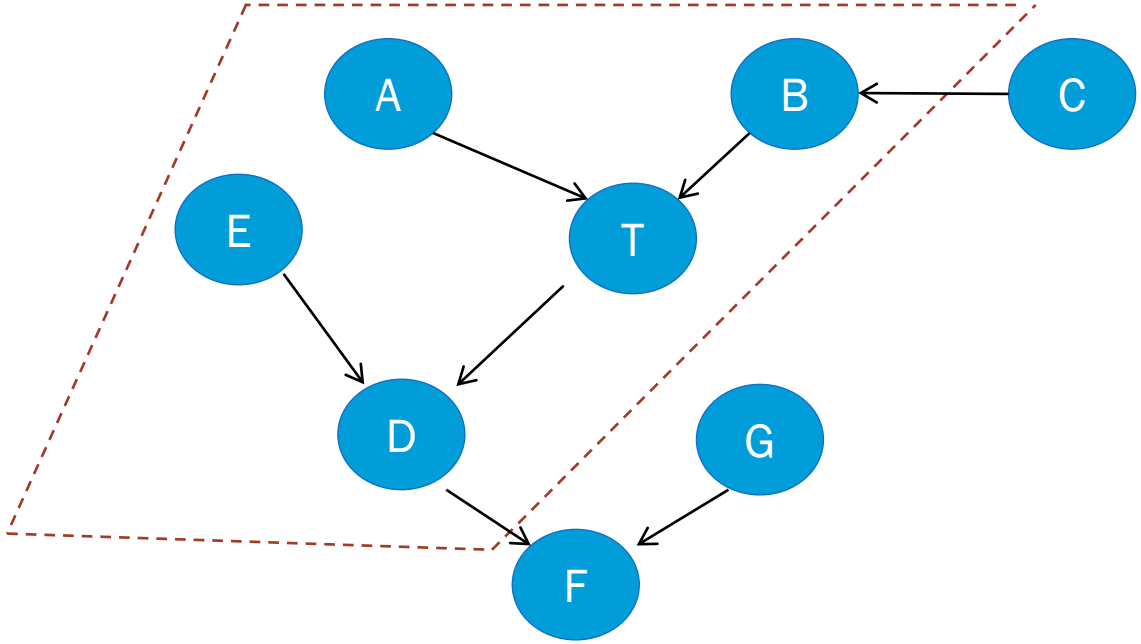
Classic Causal Discovery Example:



- Goal: To use causal discovery algorithms along side phosphoproteomic data to better understand cancer signaling, discover novel drug targets and subtype based on pathway activity.
- Use data from phosphorylation measurement

Markov Blanket

- A method that looks only at a single variable and its immediate surroundings
- Determines direct, close proximity causes and effects of known aberrant proteins
- This allows us to focus on possible clinically useful targets without the complication of distant causes and effects



Open Questions

- What is the best method for
  - Integrating different data modalities?
  - Visualizing our findings?
- Where should the investment be in the future in terms of data collection?
- Are we missing integral data types in our analysis?
  - Metabolomics
  - Other protein modifications
- Data sharing
- Tool sharing

Paper Presentations

- Anna Yeaton: *Mertins et al., Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534 (2016) 55-62.*
- Runyu Hong: *Bermudez-Hernandez et al., A Method for Quantifying Molecular Interactions Using Stochastic Modelling and Super-Resolution Microscopy, bioRxiv (2017)*
- Alexi Archambault: *Rotmensch et al., Learning a Health Knowledge Graph from Electronic Medical Records. Sci Rep. 7 (2017) 5994*