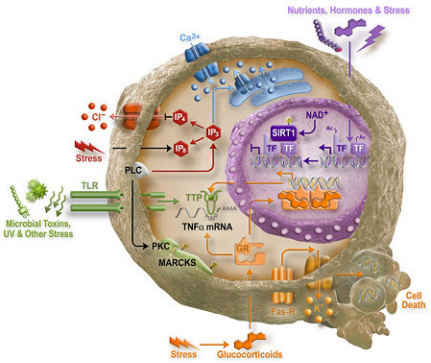
Statistics and Machine Learning Methods for Proteogenomic Data Analysis

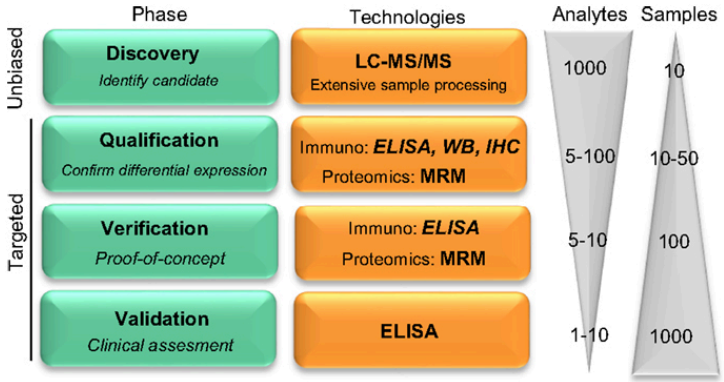
D. R. Mani
Broad Institute

Northeastern University Short Course
Computation and Statistics for Targeted Proteomics
May 5, 2016

Finding protein targets for Targeted Proteomics



- Pathways
  - Disease/phenotype related



- Unbiased Discovery

- Literature

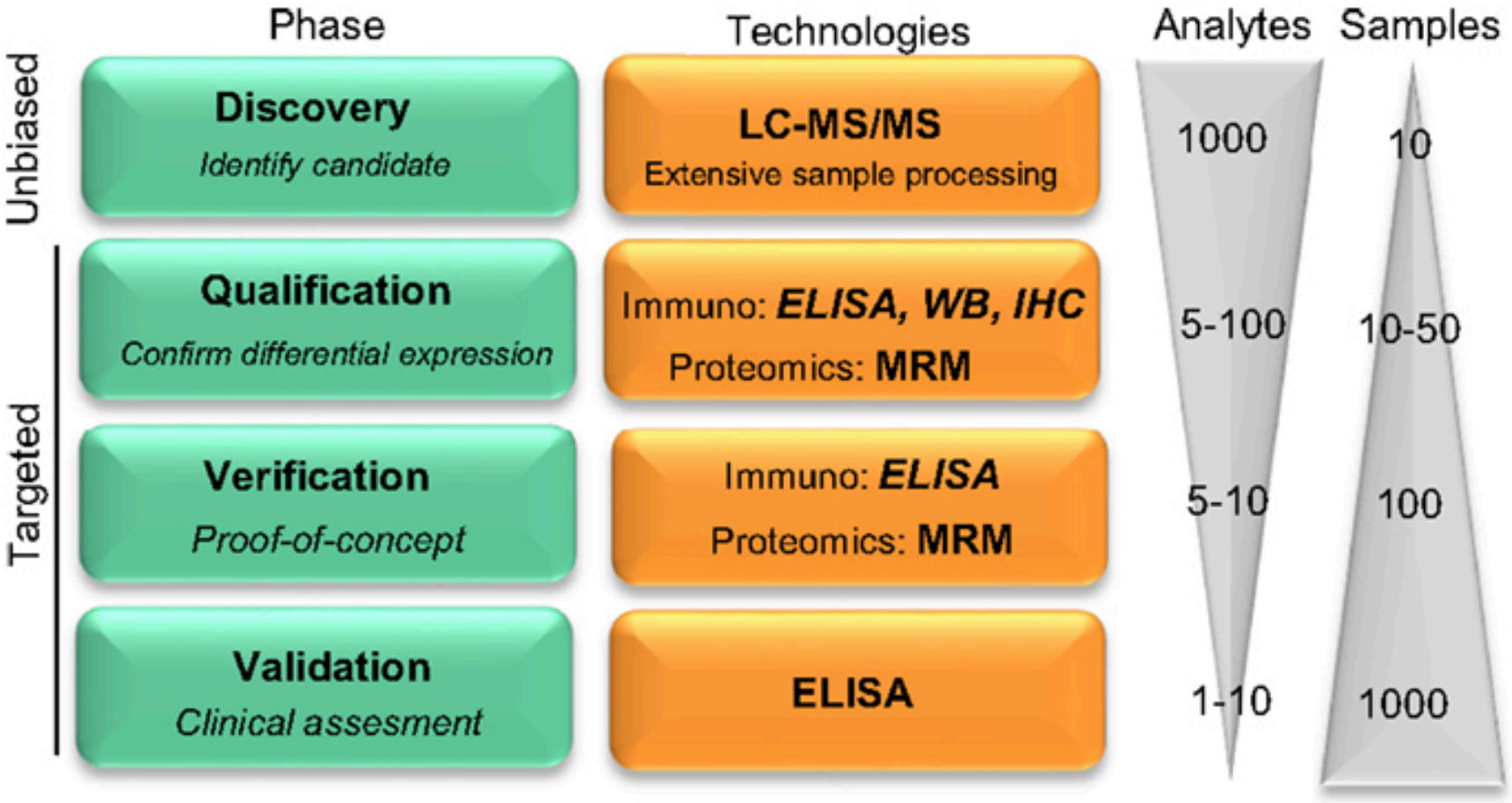
- **Proteins** for targeted proteomics

- Characteristic Subset
  - E.g: P100

gpm ESP PeptideAtlas

- Observed/predicted proteins & peptides

Unbiased biomarker discovery yields targets for Targeted Proteomics



Adapted from Rifai, et. al., *Nature Biotech*., 2006.

Unbiased discovery is increasingly Proteogenomic (or Multi-omic)

- Discovery efforts include multi-omic profiling
  - Omic profiling is getting cheaper
- Proteomic profiles are increasingly common
  - Smaller sample numbers due to higher cost
  - More input material needed

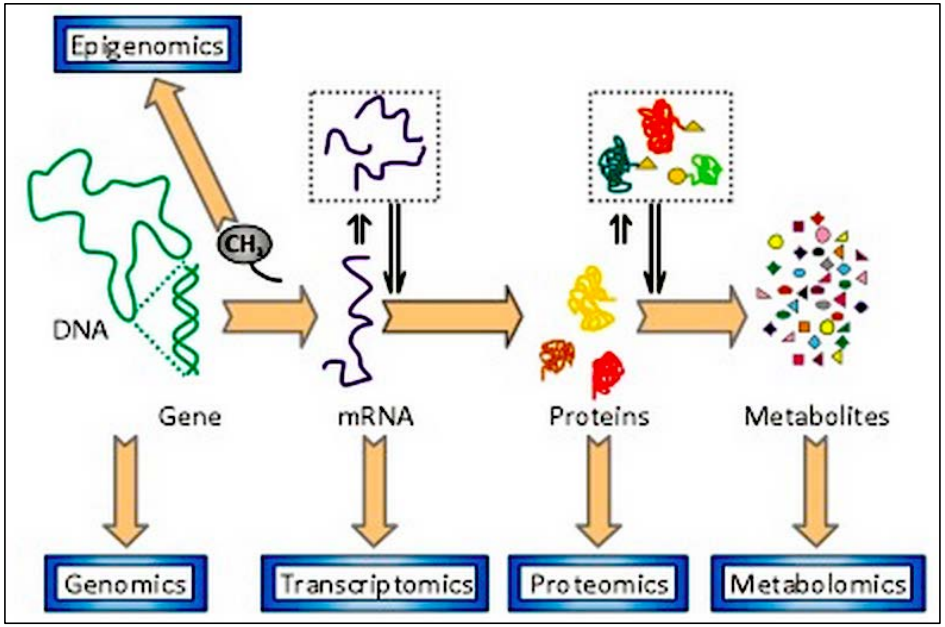


Image Source: Goodacre, J. Exp. Bot 2005.

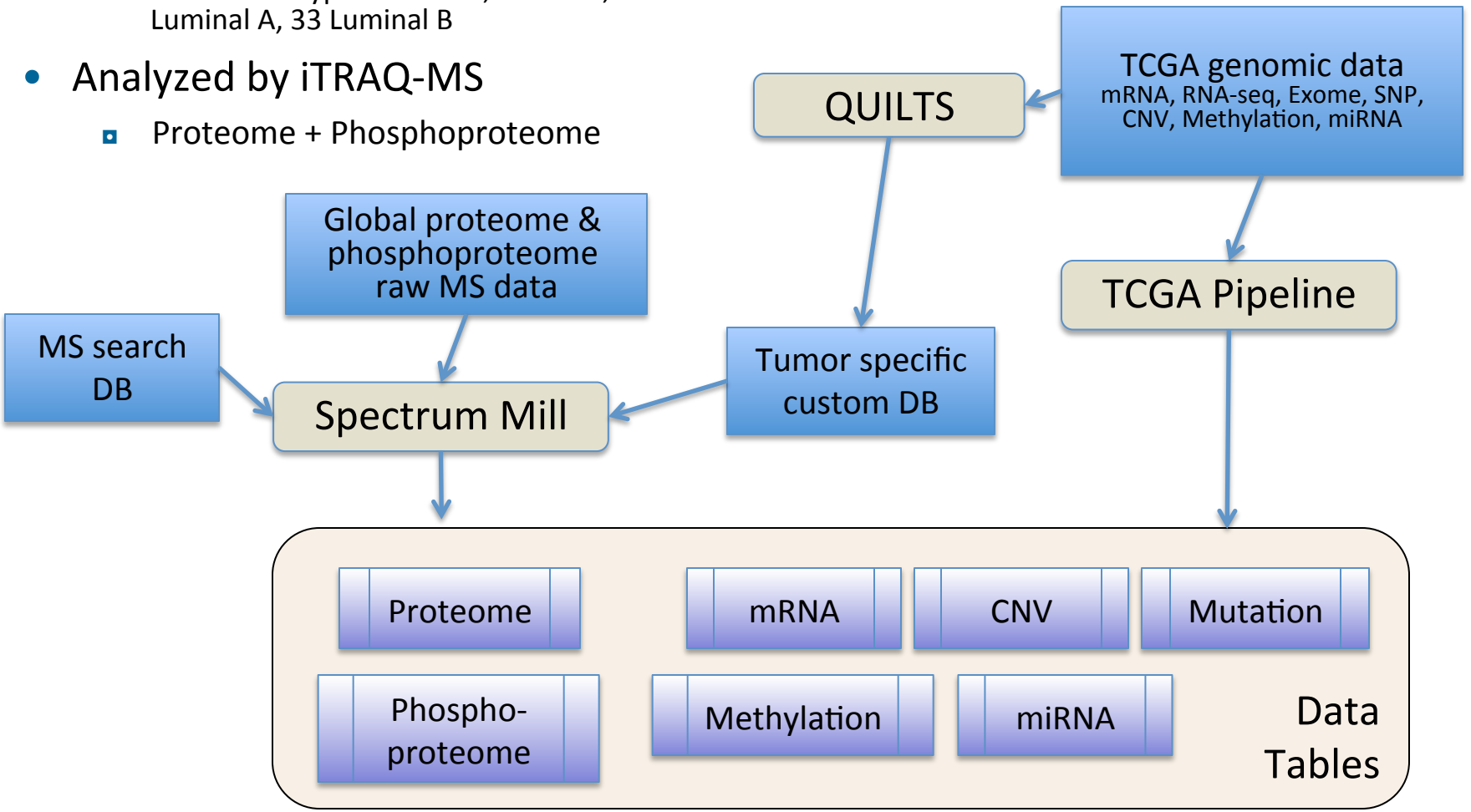
Proteogenomic analysis of Breast Cancer provides generalizable methods

- NIH CPTAC initiative to perform large-scale proteogenomic analysis of cancer samples
  - Breast Cancer—Broad Institute
  - Colon Cancer—Vanderbilt
  - Ovarian Cancer—Johns Hopkins/PNNL

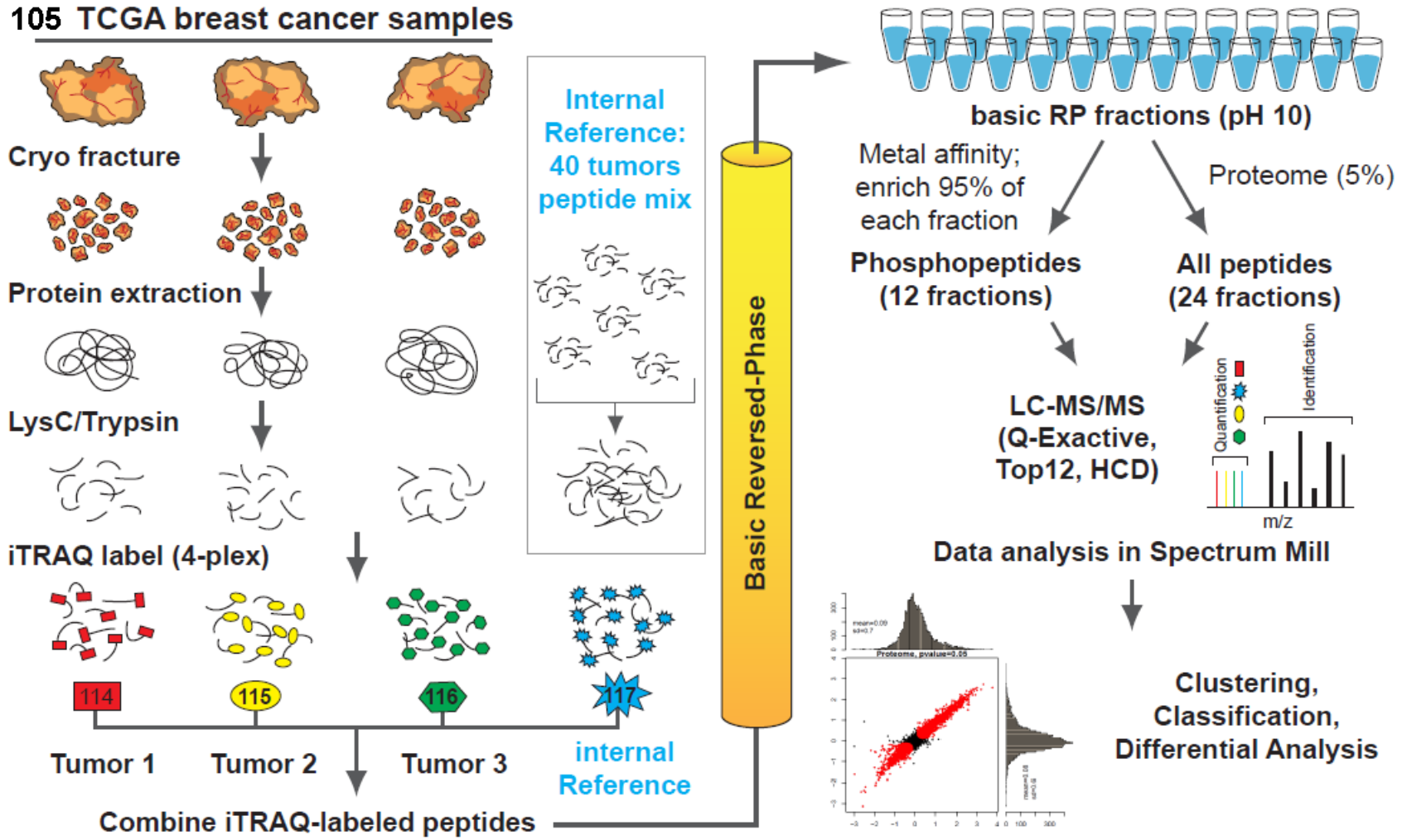
- Presentation Goals:
  - Data analysis algorithms and toolkit for proteogenomics
    - Applied to breast cancer analysis, but generalizable
  - Generalized applicability to wide range of data sets
    - Potential use for targeted data analysis
      - » Some methods applicable, others need to be modified/applied with care

Profiling of 105 TCGA samples produced largest proteomic dataset yet generated at Broad

- 105 BC Tumor Samples
  - PAM50 Subtypes: 18 Her2, 25 Basal, 29 Luminal A, 33 Luminal B
- Analyzed by iTRAQ-MS
  - Proteome + Phosphoproteome



Sample processing



1 mg total protein per tumor
Internal reference: equal representation of basal, Her2 and Luminal A/B subtypes
Tumor-specific databases based on whole exome seq and RNA seq

Sample processing: The basics

3 samples are included in each iTRAQ run. Each run also includes a Common Reference sample.

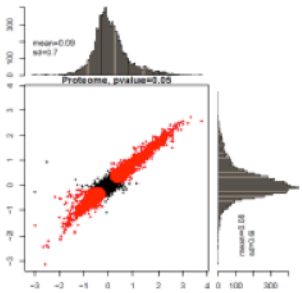
37 iTRAQ Runs
105 samples

Samples are fractionated for increased depth of coverage

[[enrichment]]

Phospho-proteome Proteome

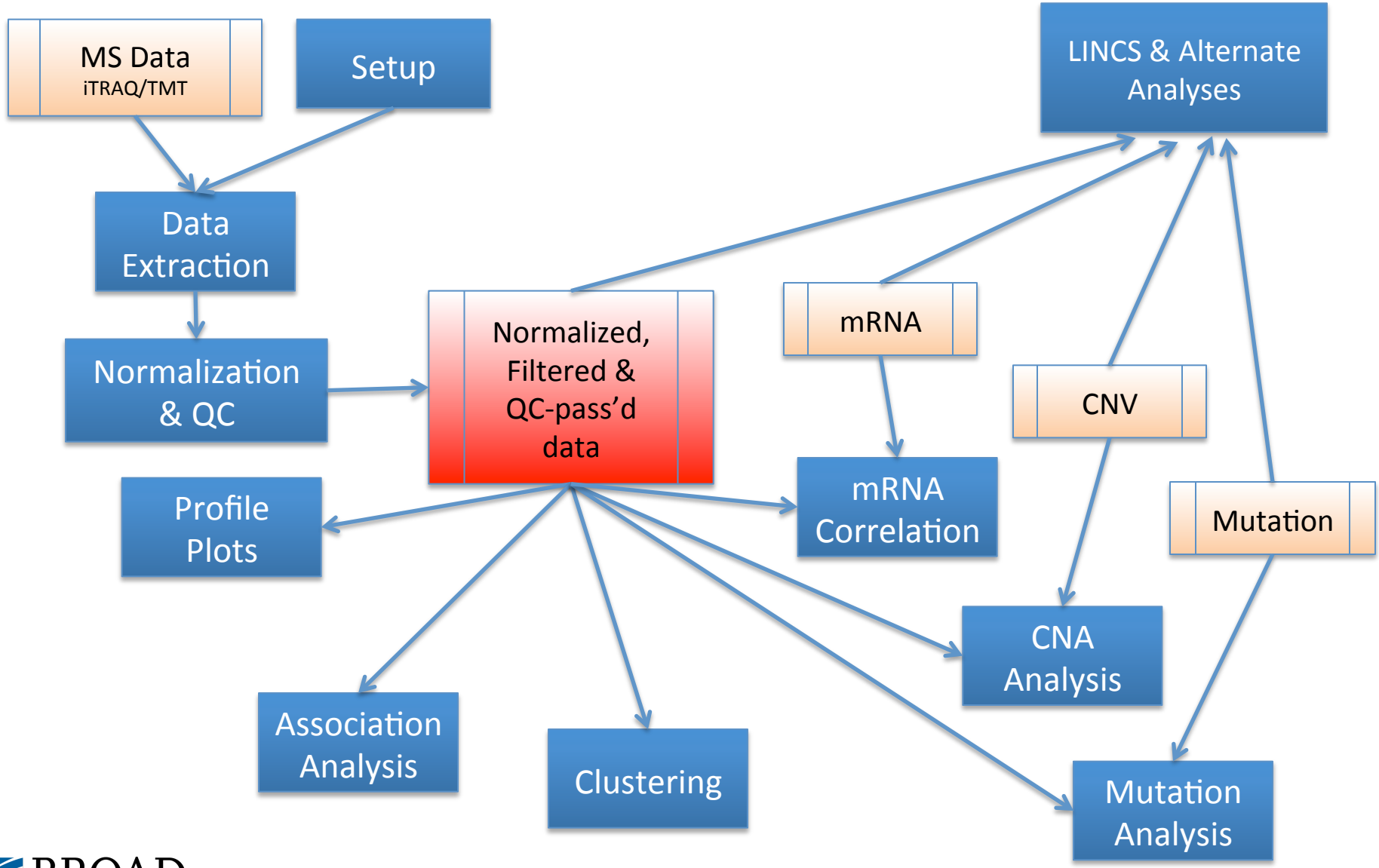
Spectrum Mill DATA output:
Protein/peptide
$\log_2$(**ratio to common reference**)



Data Analysis Pipeline

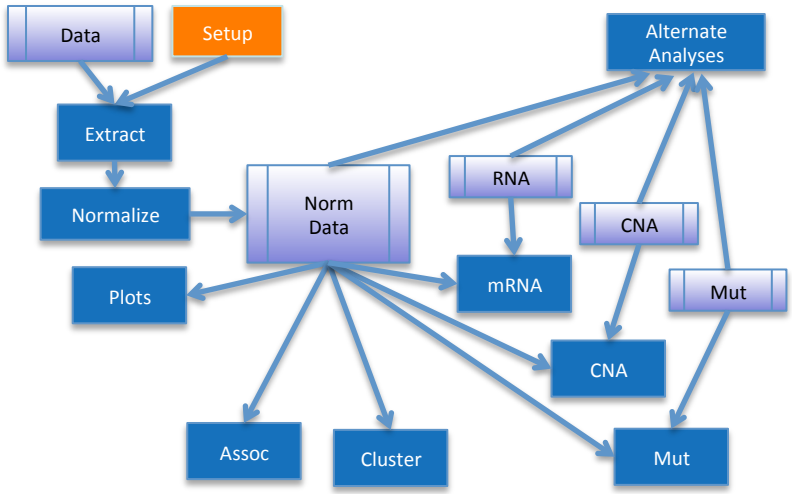
1 mg total protein per tumor
Internal reference: equal representation of basal, Her2 and Luminal A/B subtypes
Tumor-specific databases based on whole exome seq and RNA seq

Data Analysis Pipeline Overview



BROAD INSTITUTE

Setup initiates automatic pipeline execution

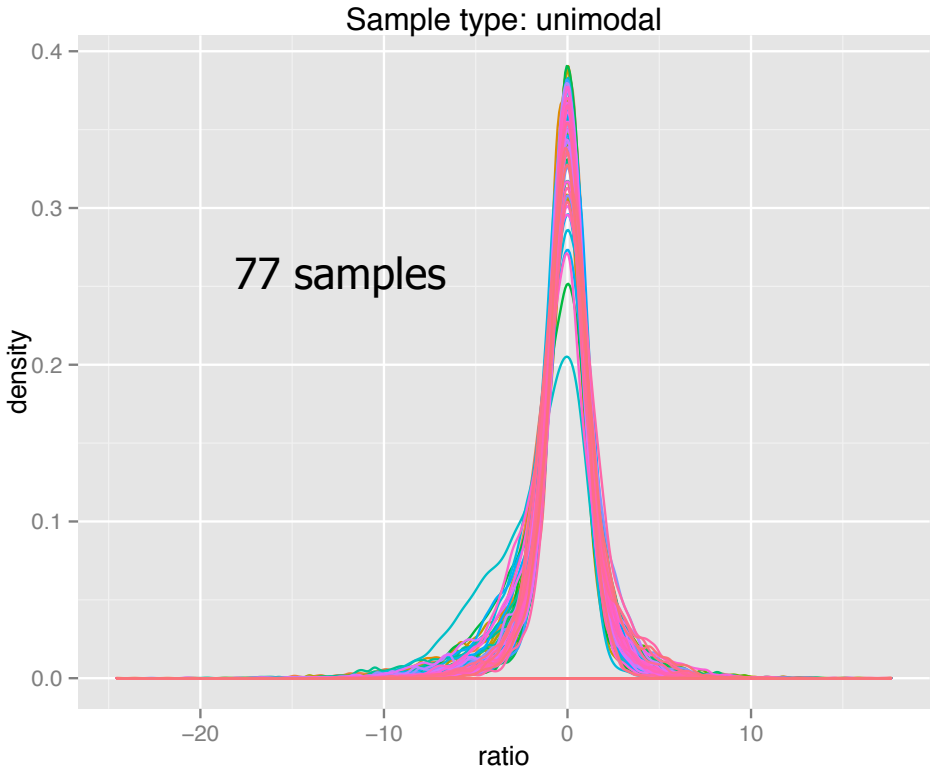


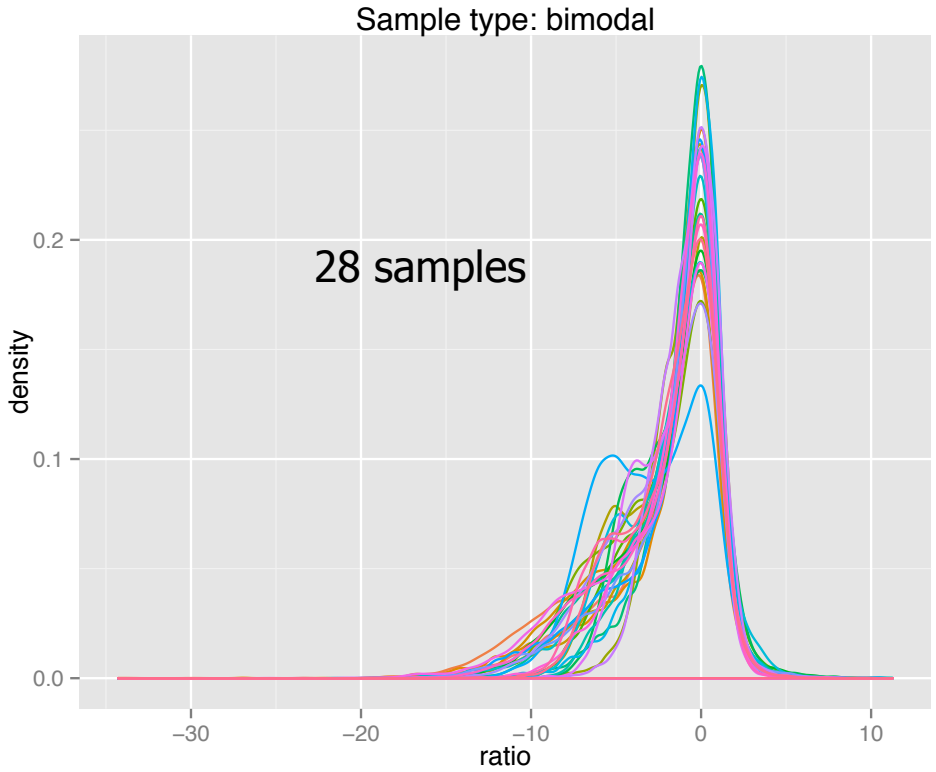
- Tools:
  - bash
  - symbolic links
  - subversion (svn)
  - UGER
    (Univa grid engine for research)

- Unix shell script
- Create directories
- Copy input data files
- Assemble required code and additional data files
  - Code & data are versioned
- Execute all core analysis components
- Use Grid Engine for parallelization at multiple levels
  - Account for data dependencies

Quality Control: Profile plots identify bimodal samples

- Bimodal samples are identified using (mixture) model-based clustering



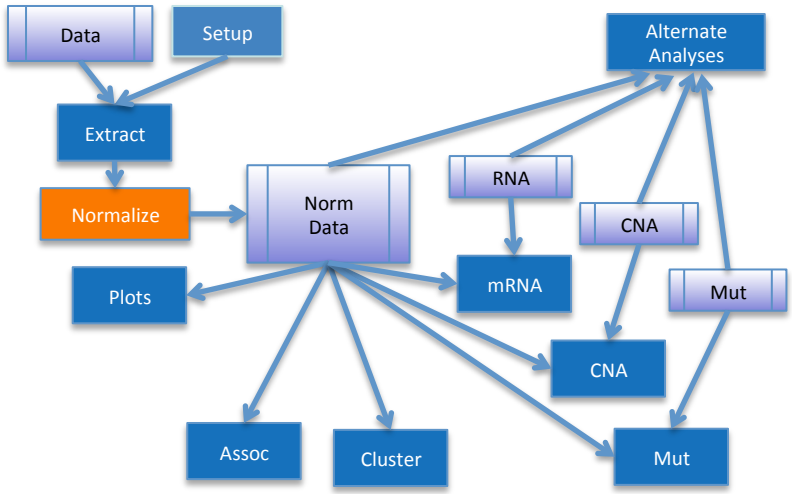


- Tools:
  - Mclust (R)
- Bimodality is most likely due to poor sample quality

Defining bimodal samples: Challenges

- Identify a metric that can separate bimodal/tailing samples from unimodal samples
  - Bimodality coefficient (too conservative—too many bimodals)
  - Dip statistic (too stringent—very few bimodal samples)
  - Measures of dispersion
    - IQR
    - Standard deviation (balanced metric)
- Classify new samples as unimodal/bimodal
  - Train classifier using single-shot (label-free) MS data
    - Use unimodal/bimodal designation as class vector
  - Apply to new samples as a QC check

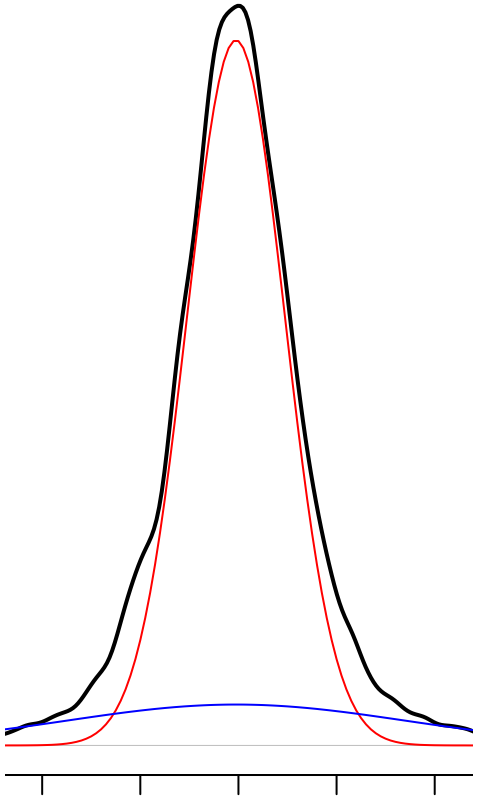
Normalization



- Each sample contains regulated and unregulated proteins.
  - Unregulated: $\log_2$ (ratio) ~ 0
  - Regulated: Extreme (+/-) ratios
- Normalize samples using only unregulated proteins.
- Unified method for both unimodal and bimodal samples

Normalization Algorithm
Using 2-component Gaussian mixture model

- Unimodal samples:
  - Find the mode M using kernel density estimation (Gaussian kernel with Shafer-Jones bandwidth)
  - Fit mixture model with mean for both components constrained to be equal to M
  - Normalize (standardize) samples using mean M and smaller std. dev. from mixture model fit



Normalization Algorithm
Using 2-component Gaussian mixture model

- Bimodal samples:
  - Find the major mode M1 by kernel density estimation (Gaussian kernel with Shafer-Jones bandwidth)
  - Fit mixture model with one component mean constrained to M1
  - Normalize (standardize) samples using mean (M1) and resulting std. dev.

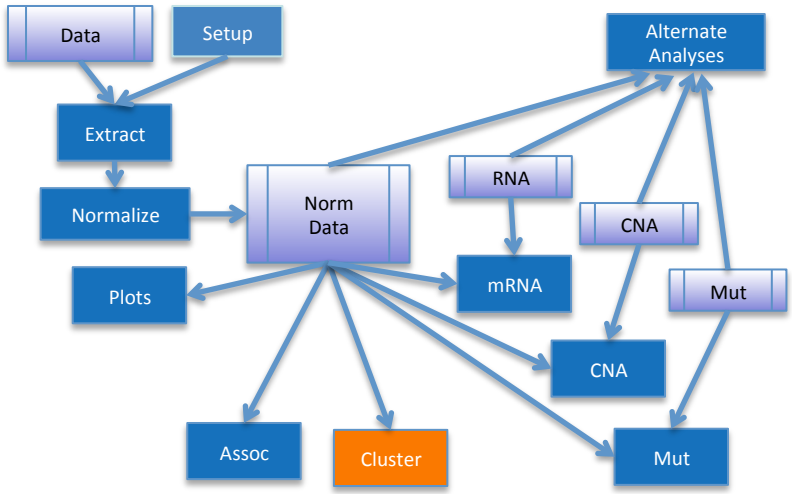
- Tools:
  - mixtools (R)
    - normalmixEM for EM estimation of mixture parameters (μ, σ^2) with constrained mean
  - Mclust (R)

Normalization: Challenges

- Mixtools estimation is not robust, and can produce one-off results
  - Unrealistic mean and variance estimates
  - Large variation is estimates when re-fit

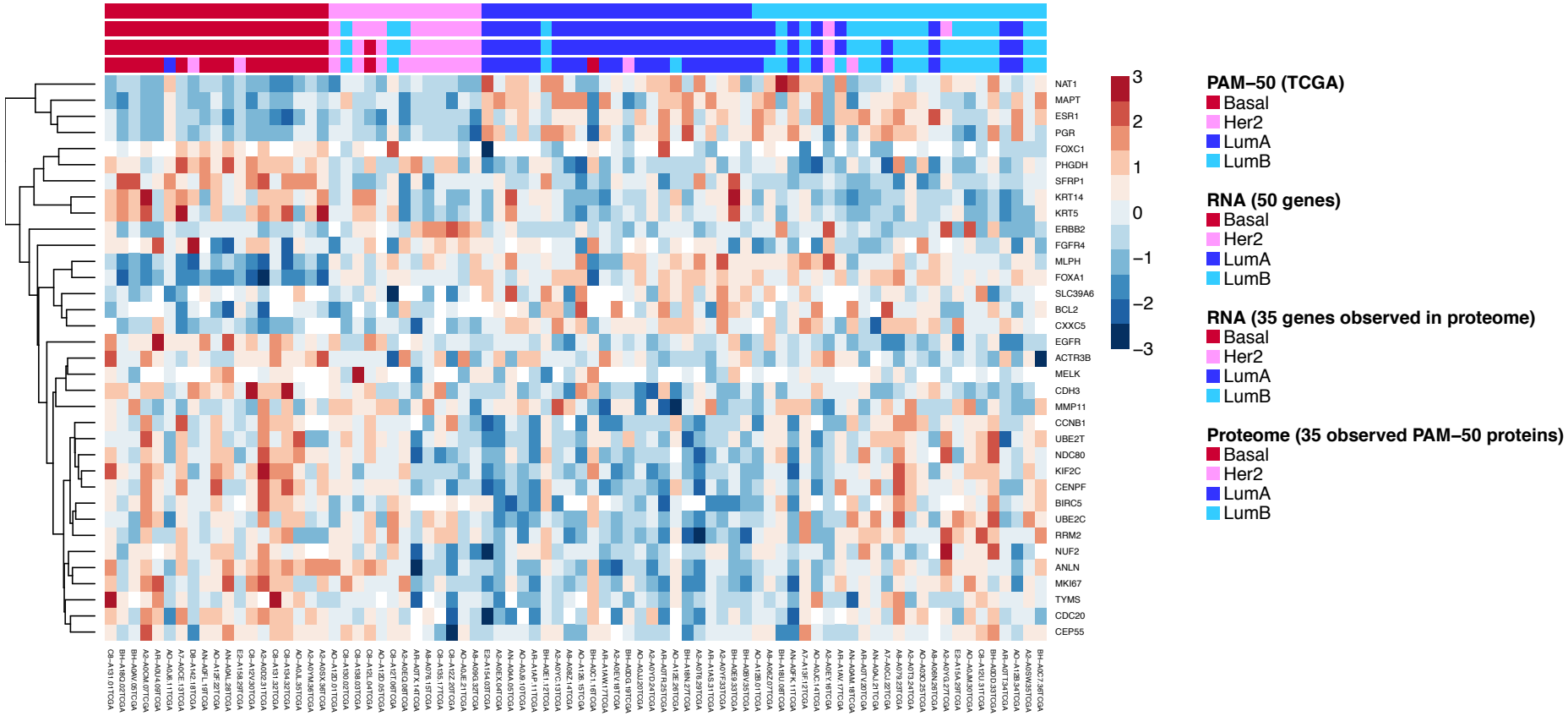
- Use mclust to assess parameter estimates from mixtools
  - Obtain approximate (unconstrained) estimate using mclust
  - Re-fit mixtools model multiple times to ensure repeatable parameter estimates
    - Must be close to mclust estimates

Clustering for proteogenomic analysis



- Does the proteome capture intrinsic RNA-based classes?
- Does tumor heterogeneity invalidate genome-proteome comparisons?
- Define intrinsic proteome and phosphoproteome clusters
- How does phosphoproteome data cluster in pathway space?
  - Based on single-sample Gene Set Enrichment Analysis (ssGSEA)

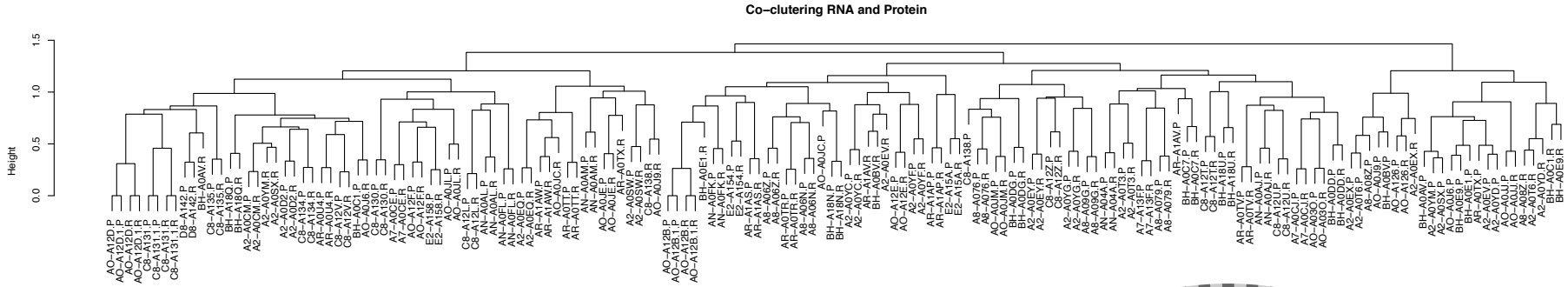
RNA-based PAM-50 clusters are captured in the proteome

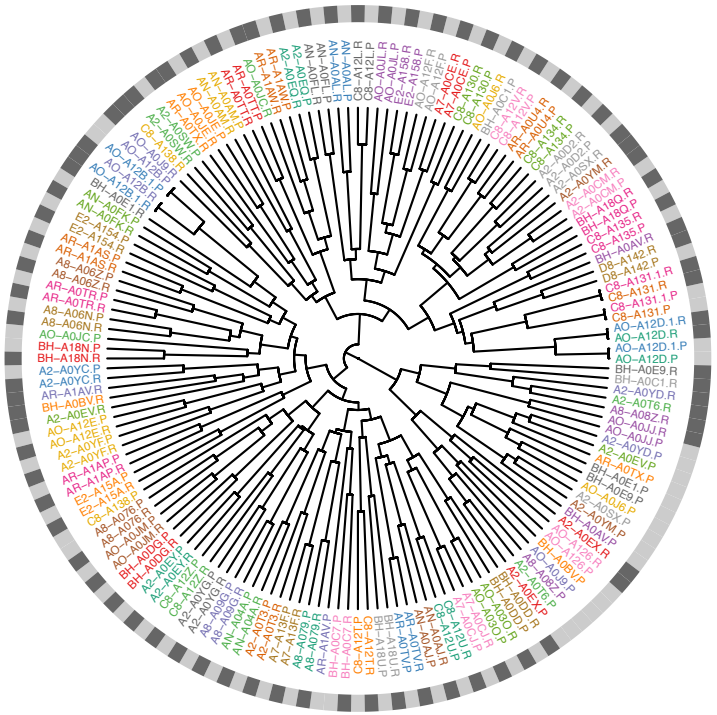


- Tools: FANNY clustering (Kaufman & Rousseeuw, 1990)
  - cluster (R)

FANNY

Proteome and RNA samples co-cluster in the space of correlated genes



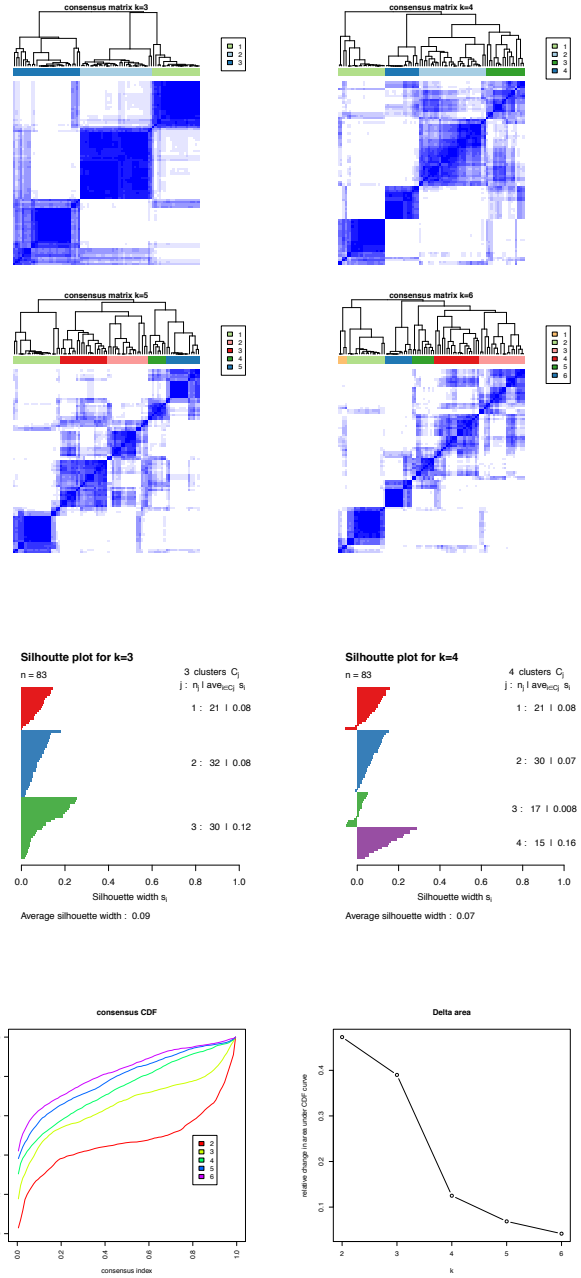


- Dataset: Combined RNA + proteome for 77 samples.
  - 4,291 proteins/genes with moderate to high correlation (R > 0.4)
- Spearman correlation to measure sample similarity
- AGNES hierarchical clustering
- "Fanplot" to show co-clustering
  - 62/77 samples co-cluster

AGNES

(Intrinsic) Proteome Clusters

- 1,521 proteins with
  - No missing values
  - Standard deviation > 1.5
- Consensus *k*-means clustering
  - 1000 bootstrap samples
  - k=3,4,5,6
- Assess cluster coherence
  - Visualization of consensus matrix
  - Consensus CDF/Delta-area plot
  - Silhouette distance

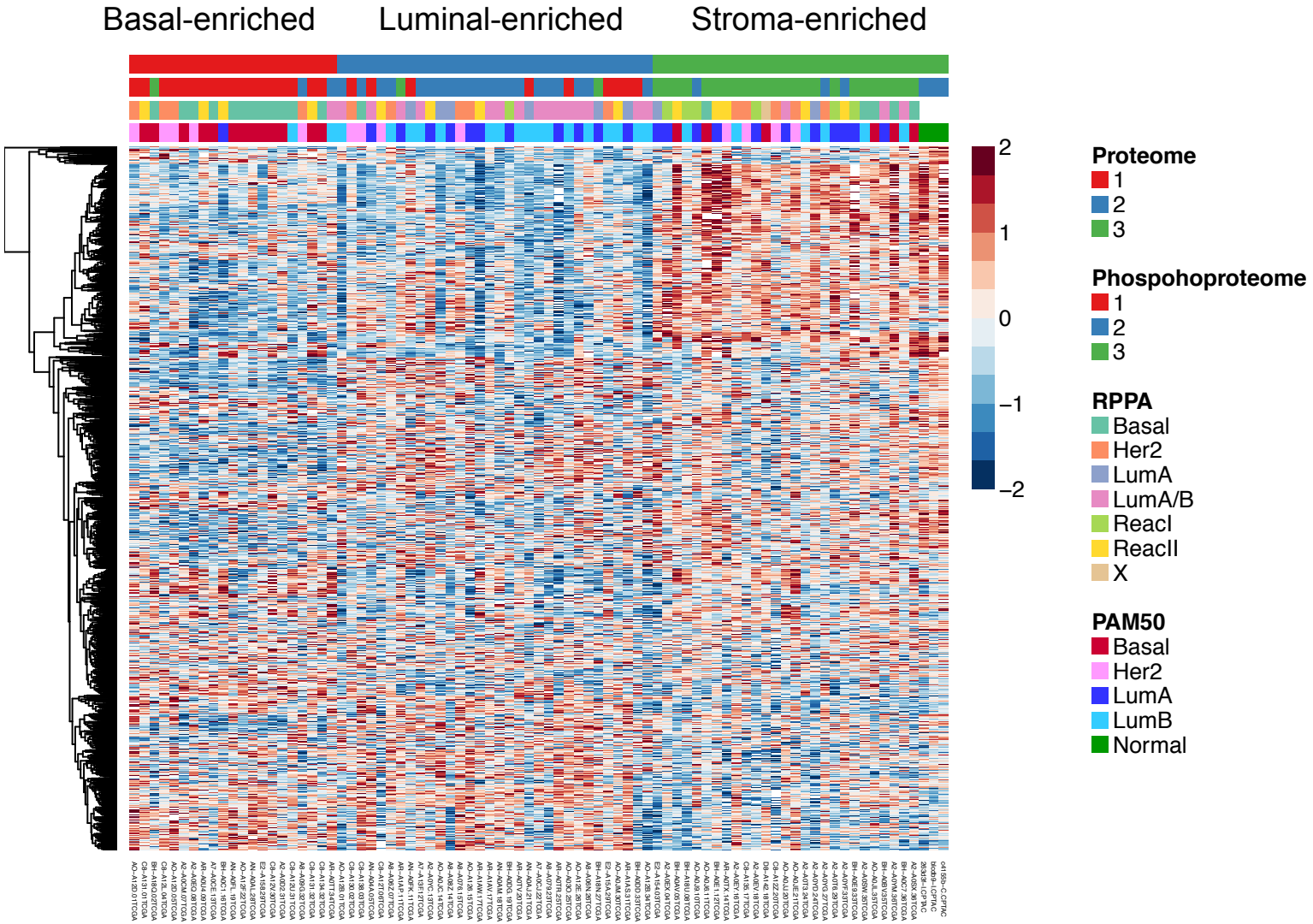


Assessing cluster coherence

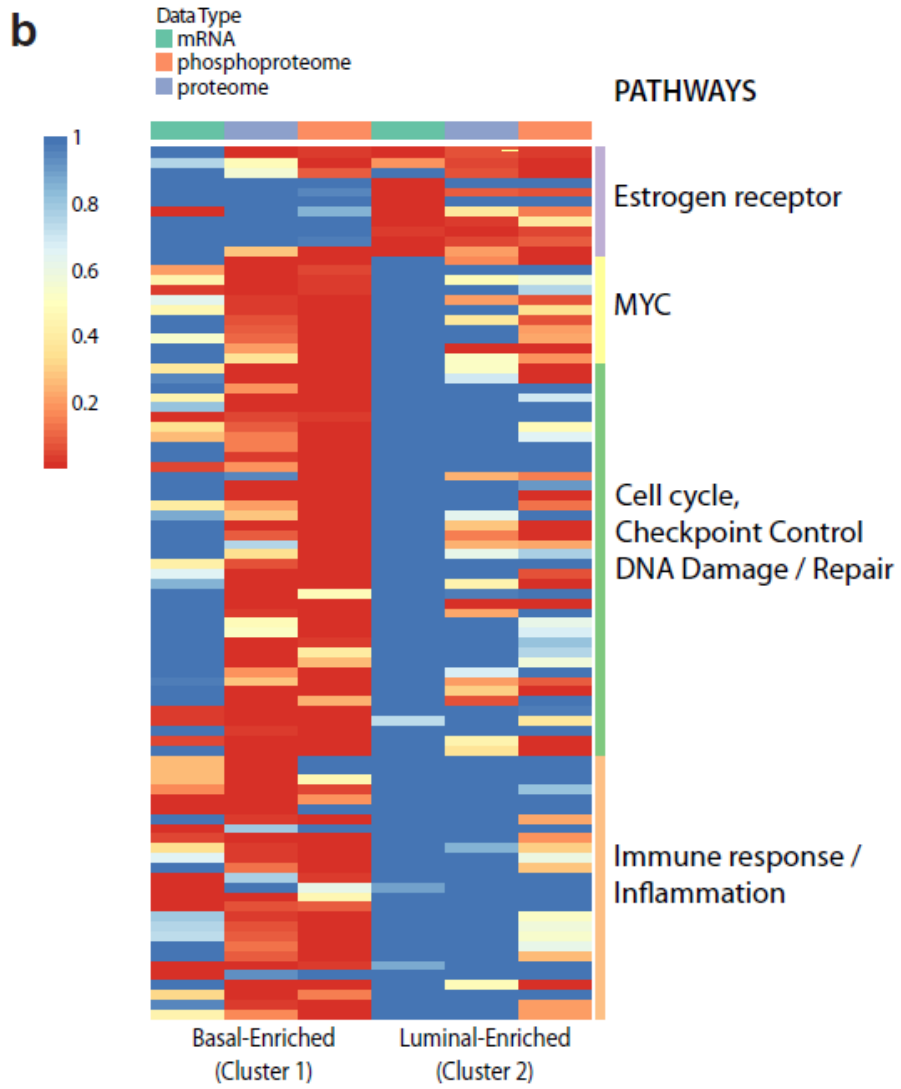
- Silhouette distance
- Consensus CDF
- Delta-area plot

- Tools:
  - cluster (R)
  - consensusClusterPlus (R)

Proteome and Phosphoproteome Clusters



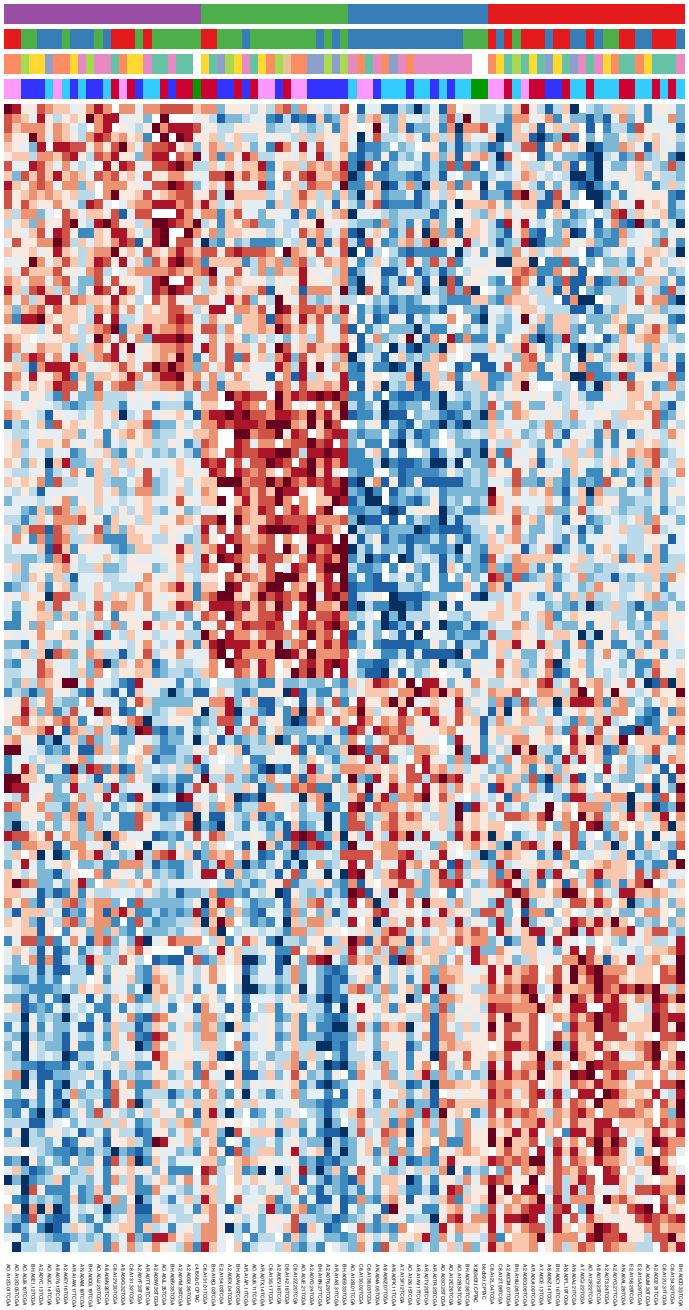
Cell cycle, DNA-damage and immuno-regulatory gene sets are enriched in Basal-like tumors



Pathway enrichment analysis

Phospho-pathway clustering

- Dataset: 5,914 phosphoproteins
  - Filtered Phosphoproteome data
    - Phosphosites with <81 missing values
    - Standard deviation > 0.5 across all samples
  - Phosphosite rolled-up to proteins using median ratio
  - Map phosphoproteins to genes
- Map samples to MSigDB pathways using ssGSEA
  - 908 curated pathways
- Consensus *k*-means clustering in pathway space
- Assess cluster coherence



G protein

G protein-coupled receptor

Inositol phosphate metabolism

ErbB / HER signaling

VEGF signaling

Integrin signaling

Estrogen receptor signaling

Transcription Regulation

Cell cycle

Checkpoint

2
1
0
−1
−2

PhosphoPathway
1
2
3
4

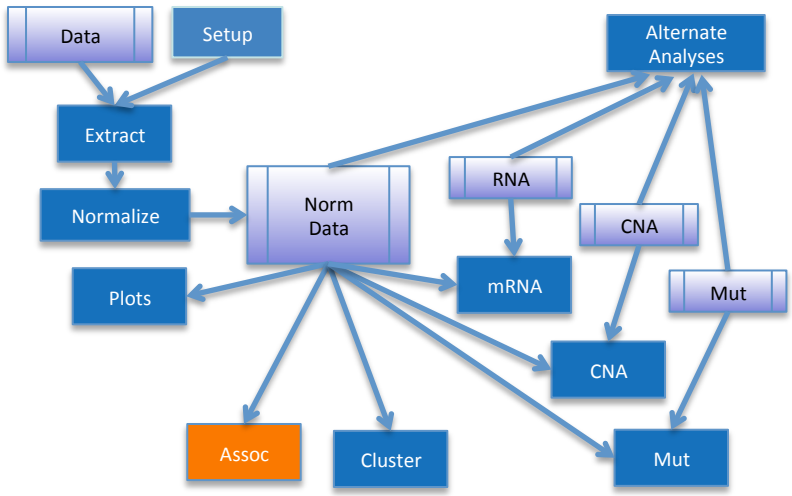
Proteome
1
2
3

RPPA
Basal
Her2
LumA
LumA/B
ReacI
ReacII
X

PAM50
Basal
Her2
LumA
LumB
Normal

Phospho-pathway clustering identifies unique clusters

Association analysis via marker selection and GSEA



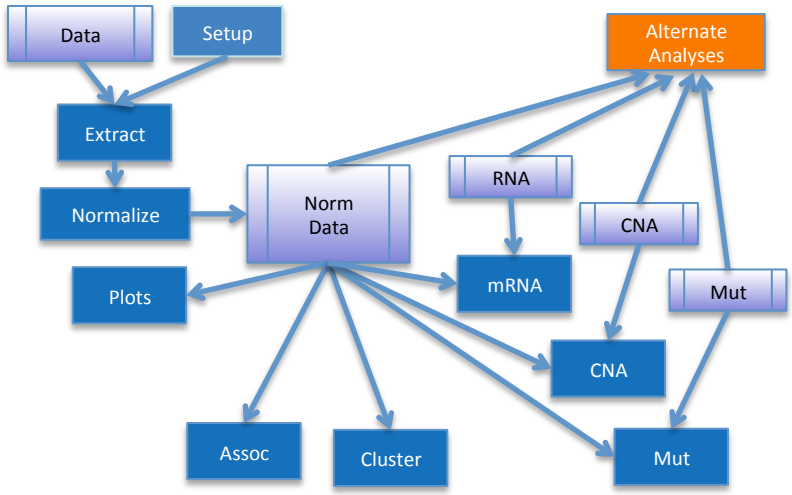
Association Analysis and Marker Selection

- Collection of algorithms for
  - Identification of statistically significant differential markers
    - Multiclass
    - One-vs-all
  - Training of multiple classifiers
    - Partial Least Squares, Shrunken Centroids, Random Forests, Elastic Nets
    - Other algorithms can be easily added
  - Variable importance from classifiers for further prioritization of differential markers
    - Marker rank aggregation for final marker ranking
  - Class prediction for unknown/new samples
  - Visualization (heatmaps)
  - GSEA for pathway enrichment
  - EnrichmentMap for visualizing enriched pathways

Rank Aggregation for Marker Ranking

- Perform Marker selection:
  - Identify statistically significant differential markers (SAM)
    - Multiclass
    - One-vs-all
  - Train multiple classifiers
    - Partial Least Squares, Shrunken Centroids, Random Forests, Elastic Nets
    - Other algorithms can be easily added
  - Rank markers using variable importance from classifiers
- Combine multiple rankings to a final rank
  - Robust rank aggregation (R. Kolde et. al., *Bioinformatics*, 2012)
    - Calculate final rank based on order statistics
    - Accommodates significant proportion of "noise" markers and occasional "low" ranks

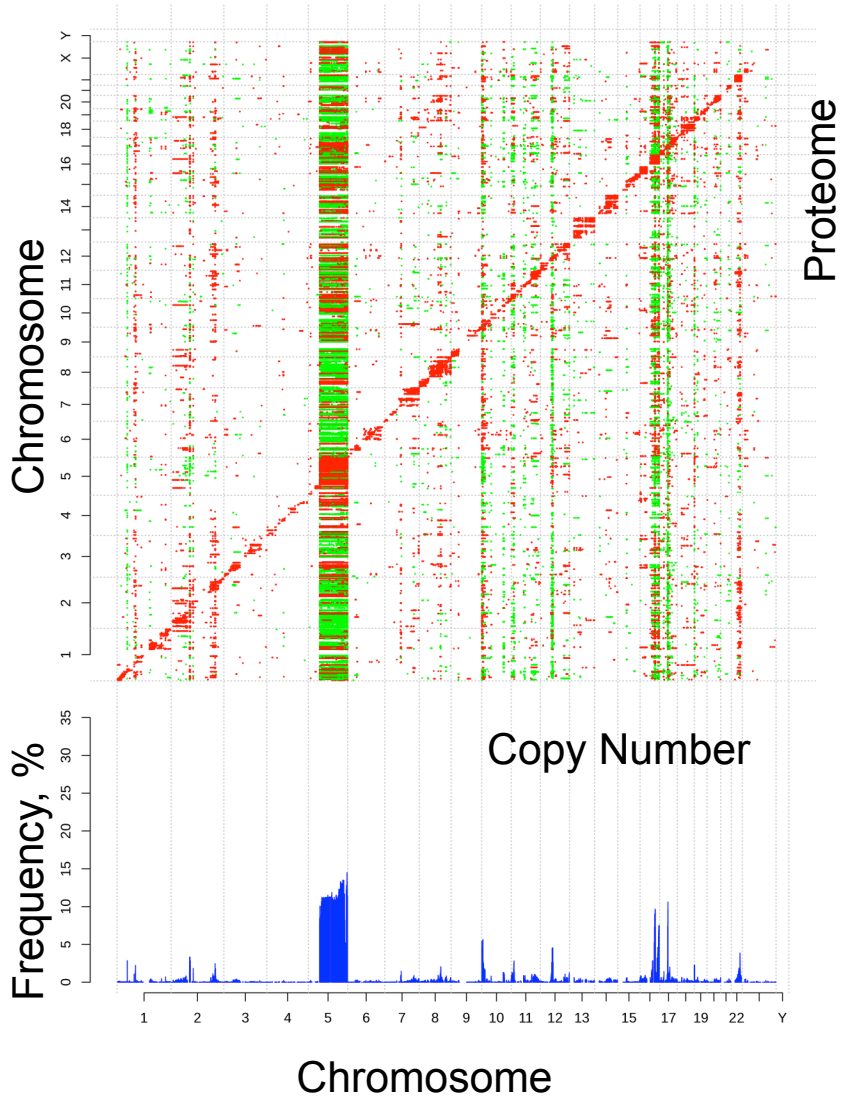
Linking copy number alteration and protein expression using LINCS
(Library of Integrated Cellular Signatures)



Approach

- Compare proteome profiles filtered by CNA TRANS correlations with LINCS functional knock-down data.
  - Genes with LINCS-enriched CIS effects are considered candidate driver genes
  - FDR for candidate driver genes is estimated using a permutation test

CNA-protein correlations show CIS and many TRANS effects



- Correlate copy number (CN) data with proteome for all 60 million gene-protein pairs
- Plot statistically significant correlations (FDR < 0.05)
  - positive correlation
  - negative correlation
- Histogram shows percent of significant correlations at a CN locus
- Highlights "hot-spots" of TRANS-activity

TRANS effect "hot spots" at chromosomes 5q, 10p, 12, 16q, 17q, and 22q

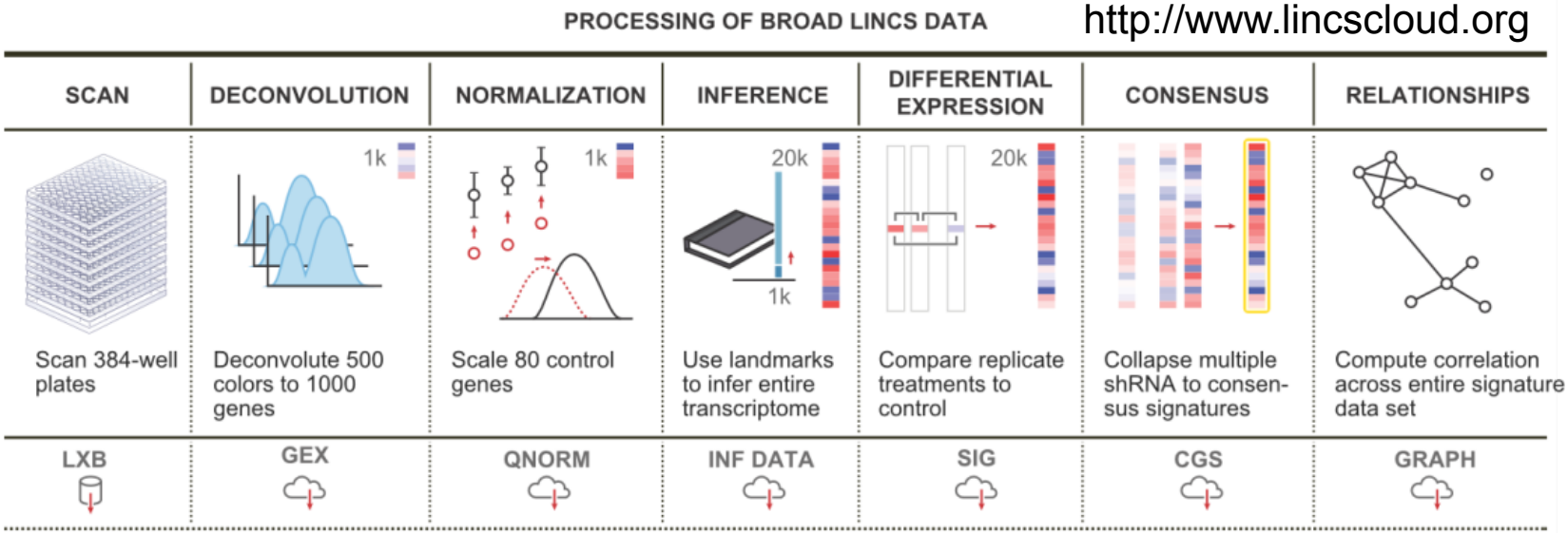
Can proteome profiles identify candidate genes driving response in copy number altered regions?

- A small number of key genes drive observed TRANS-effects
- To identify candidate genes:
  Correlate proteome profiles of CN altered samples with gene knock-down mRNA profiles
- CN amplification negatively correlated to knock down profile and/or CN deletion positively correlated to knock down profile

  ➔ candidate causal gene

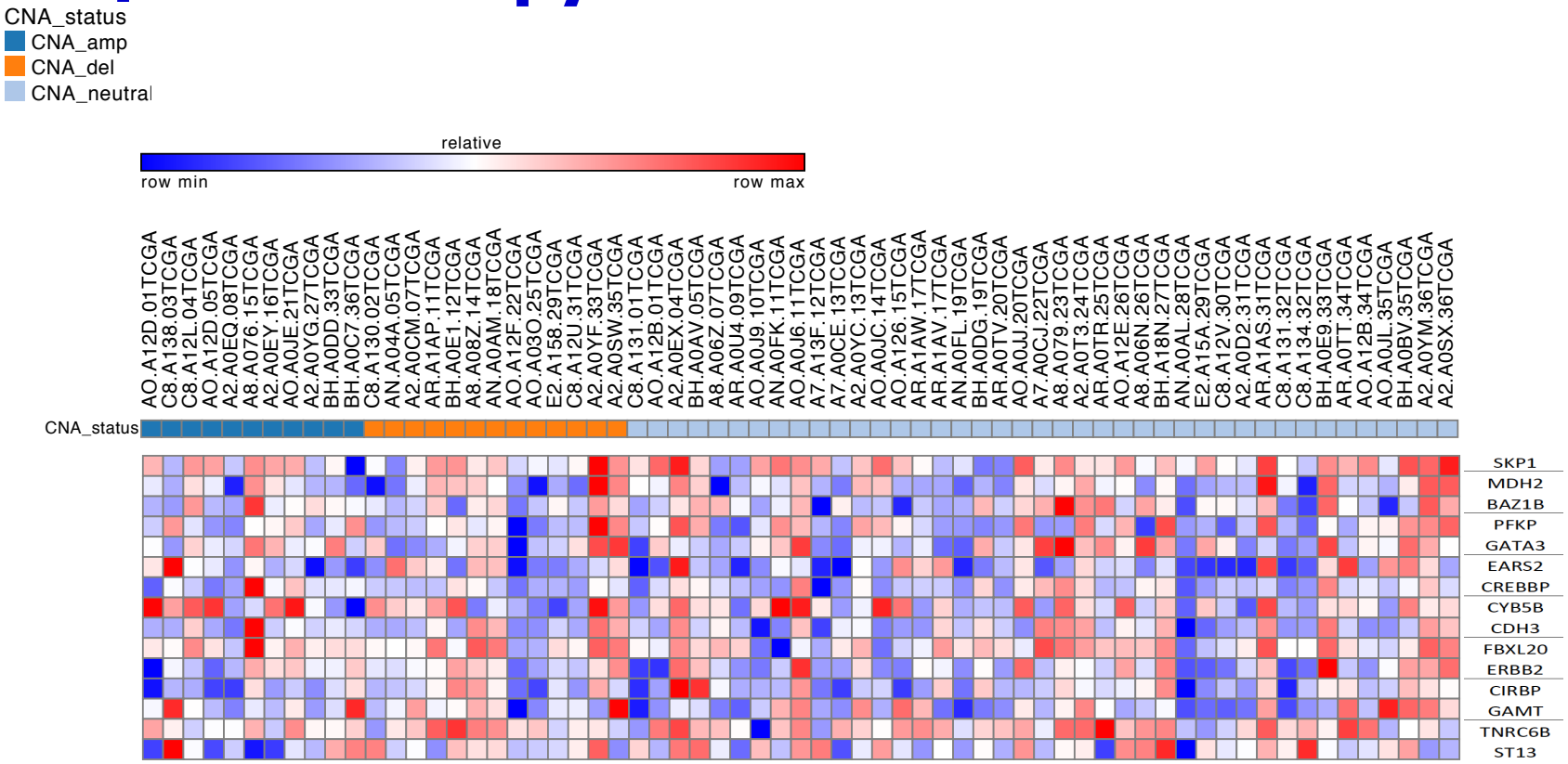
Leveraging large scale perturbation datasets to identify candidate causal genes in CNA regions

Library of Integrated Cellular Signatures (LINCS) aka The Connectivity Map (CMAP)



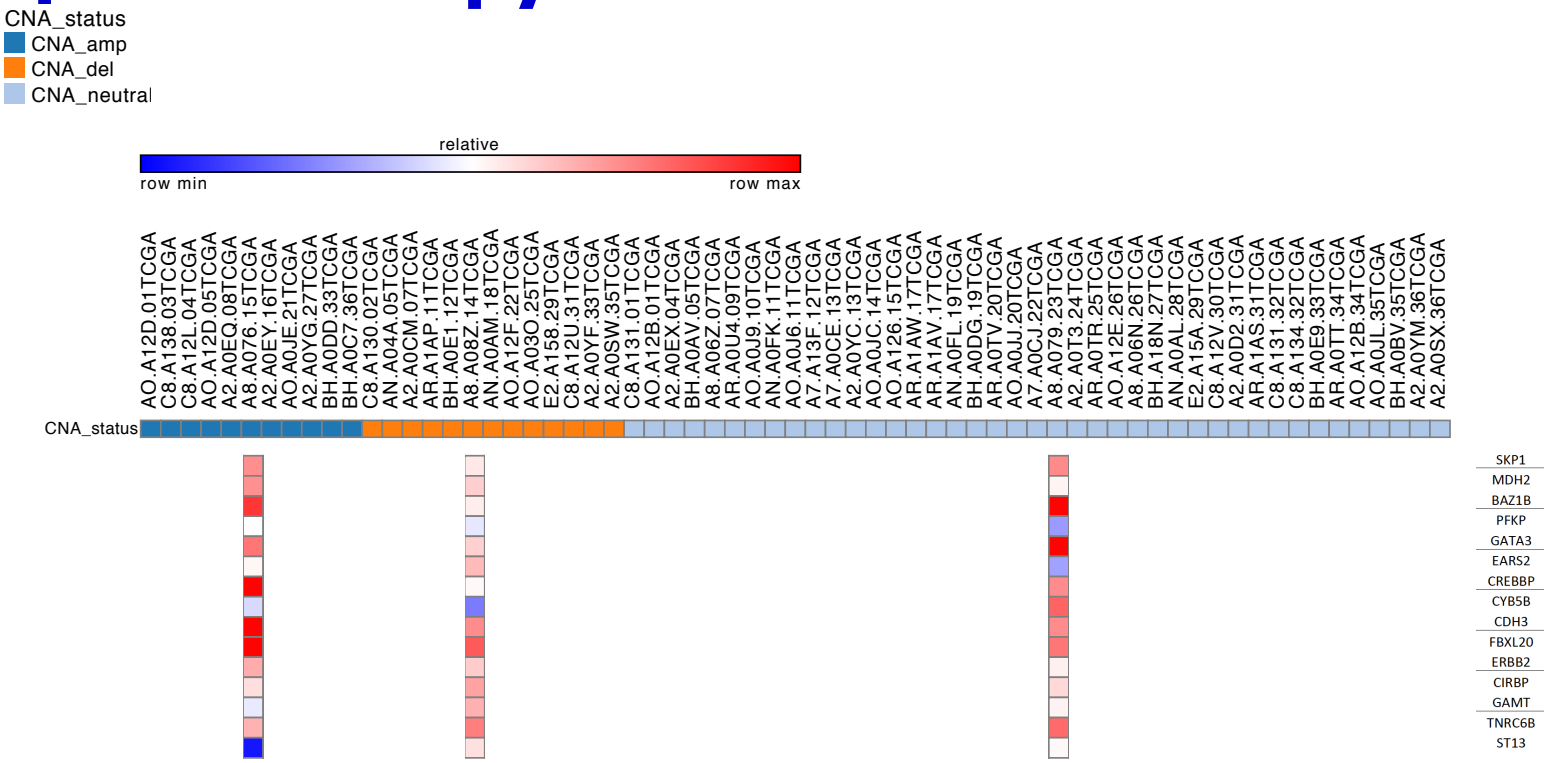
- LINCS Functional knock-down profiles on ~3,800 genes:
  - Multiple hairpins per gene knock-down
  - 1000 landmark genes measured on Luminex assay
  - Complete profile (~22,000 genes) calculated by inference
  - Includes ~20,000 drug perturbagens. Total ~476,000 mRNA profiles

Use LINCS to identify key genes driving response to copy number alterations: STEP 1



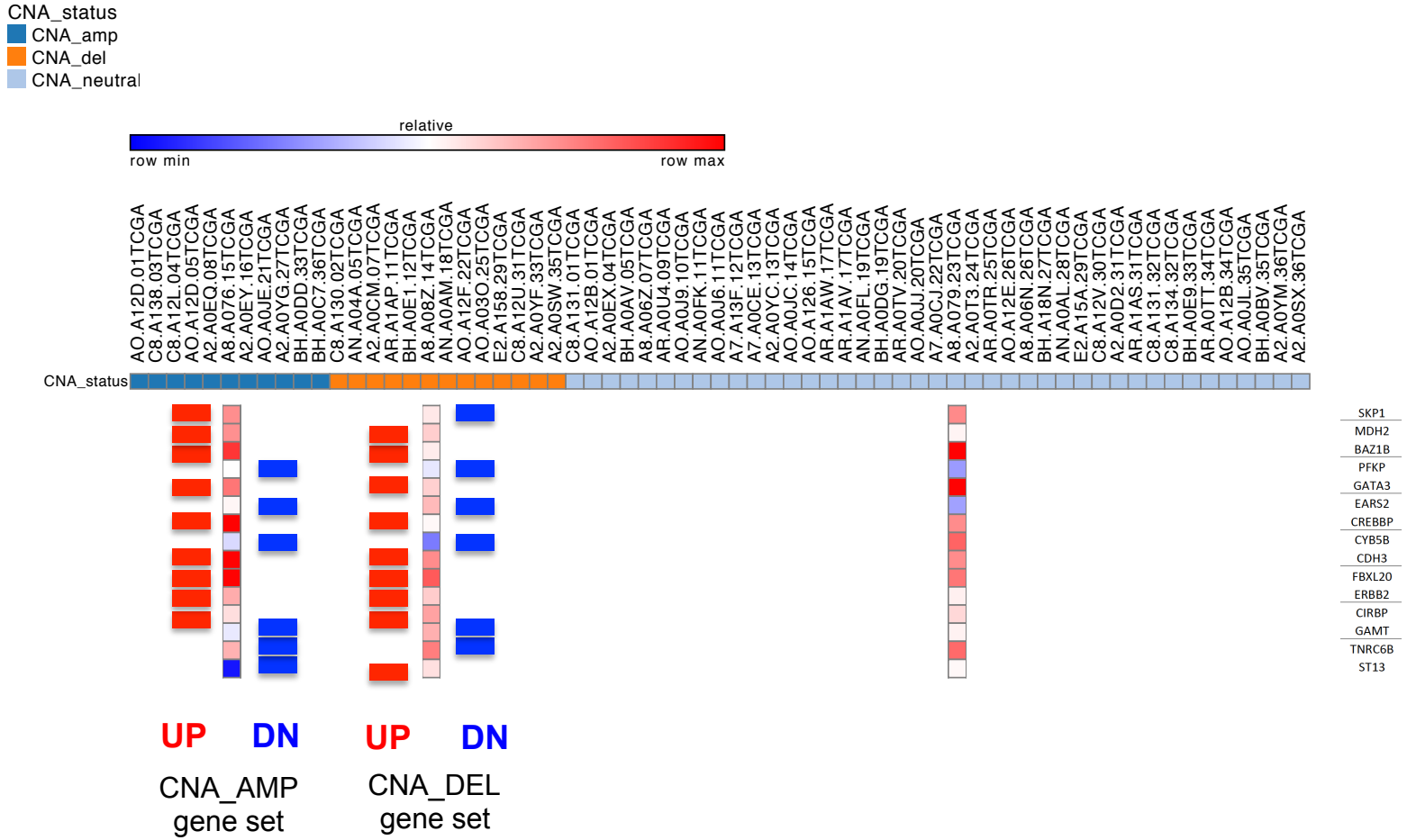
- Identify samples with deletion [log(CN) < −0.3], neutral and amplification [log(CN)>0.3] CNA for a given gene
- Extract protein expression for genes with significant TRANS-effects (FDR < 0.05).

Use LINCS to identify key genes driving response to copy number alterations: STEP 2



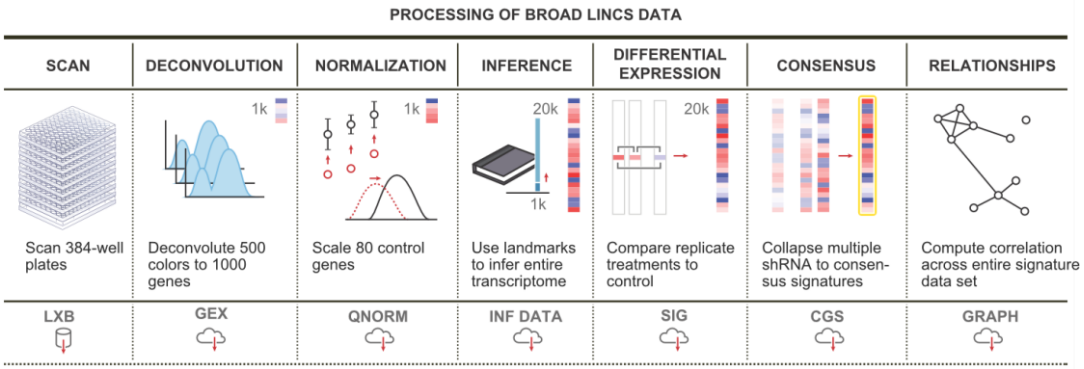
- Summarize expression in CNA_DEL, CNA_AMP and CNA_NEUTRAL groups
  - Median expression for each trans-gene

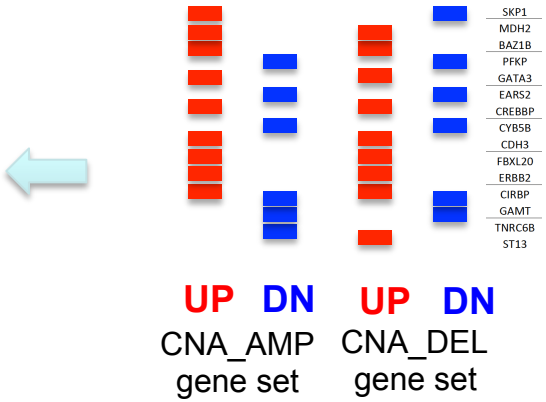
Use LINCS to identify key genes driving response to copy number alterations: STEP 3



- Determine up and down regulated genes in CNA_DEL and CNA_AMP (in comparison to CNA_NEUTRAL expression)

Use LINCS to identify key genes driving response to copy number alterations: STEP 4





- Query LINCS using CNA_AMP/DEL gene sets
  - Convert CNA_AMP/DEL gene sets to Affymetix IDs
  - Run LINCS enrichment test on ~240,000 "gold" consensus signatures (CGS).
  - Extract "CIS-enriched" gene knock downs:
    - Enriched gene knock downs include CIS gene
      - Correct direction of correlation (+ve for CNA_DEL, –ve for CNA_AMP)
    - |mean_rankpt4| > 90
      - Mean percentile in 4 cell lines > 90
  - Extract and analyze z-scores for CIS-enriched genes

Calculate Permutation-based FDR

1. For each of the genes input to the LINCS enrichment test, generate a random permutation as follows:
   - Let gene G have N_g TRANS genes
   - From the list of all genes, randomly select N_g genes (without replacement)
2. Run LINCS enrichment for this permuted dataset
3. Determine FP_i, the number of "candidate driver genes" from the random dataset.
4. Repeat Steps 1-3 R times.
5. Calculate FDR as mid point of 95% Score CI assuming Poisson distribution with small rate ($\lambda \approx 0$) and small R (R=6).

$$FDR = E\left(\frac{\#FP}{\#P}\right) = \frac{E(\#FP)}{\#P} = \frac{\overline{FP} + 1.96^2/(2R)}{\#P} \quad \text{where } \overline{FP} = \frac{1}{R}\sum_{i=1}^{R} FP_i$$

$$95\%\ \text{Score CI for } E(\#FP) = \overline{FP} + 1.96^2/(2R) \pm 1.96\frac{\sqrt{4\overline{FP} + 1.96^2/R}}{\sqrt{4R}}$$

Results

- Input gene sets:
  - ≥15 tumors with |CNA| > 0.3
  - Gene must be on CMAP KD list
  - # TRANS genes ≥ 20
  - Total genes tested: 502

- 20 CIS-enriched candidate genes
  - Level I: 10 genes
    - Enriched in both CNA_AMP and CNA_DEL
  - Level II: 10 genes
    - Enriched in either CNA_AMP or CNA_DEL

✧ Level I

CETN3
SKP1
SNX2
FBXO7
TERF2IP
WIPF2
YAP1
RB1
USP7
DHPS

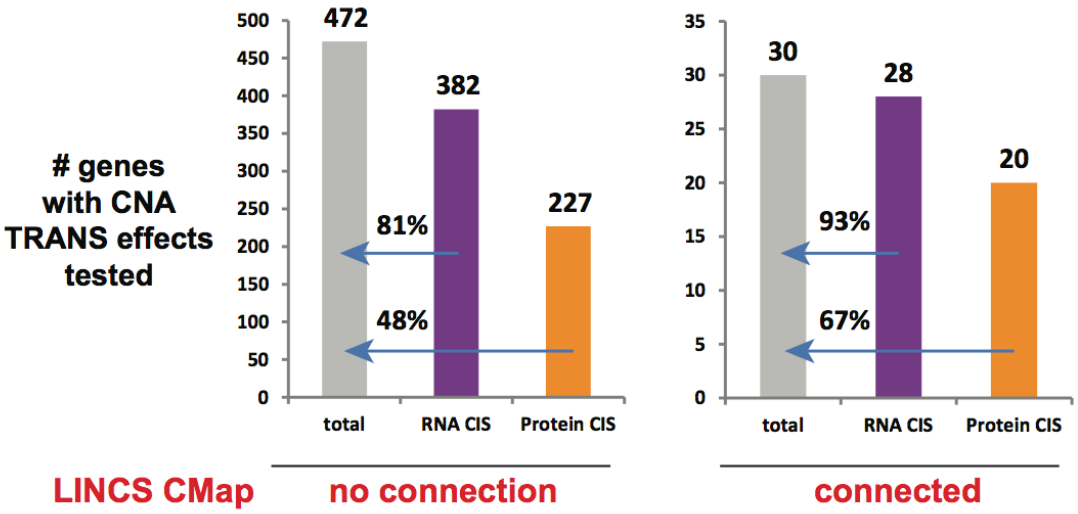
FDR=0.049
[0.003, 0.094]

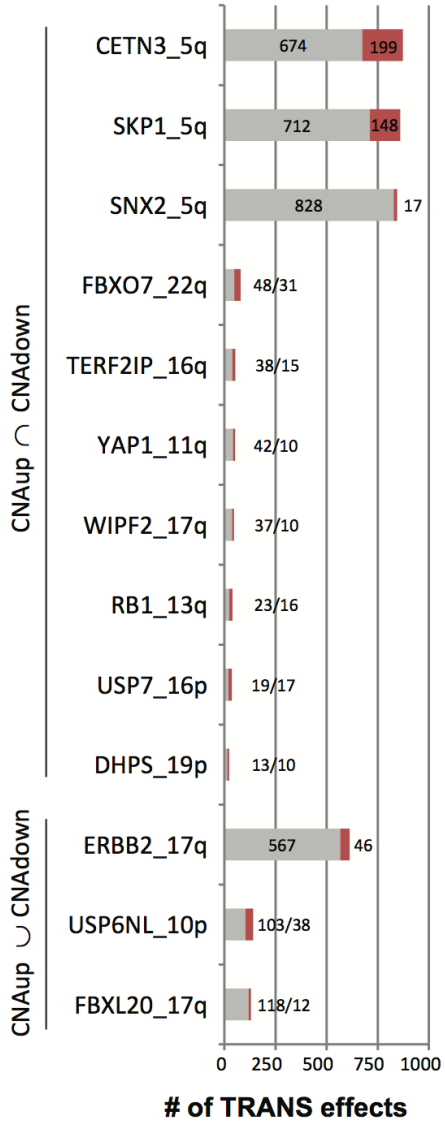
✧ Level II

FBXL20
ERBB2
USP6NL
ARHGEF12
MRPL12
RAB21
EP300
CPNE3
PLCB3
UBE3C

FDR=0.305
[0.225, 0.385]

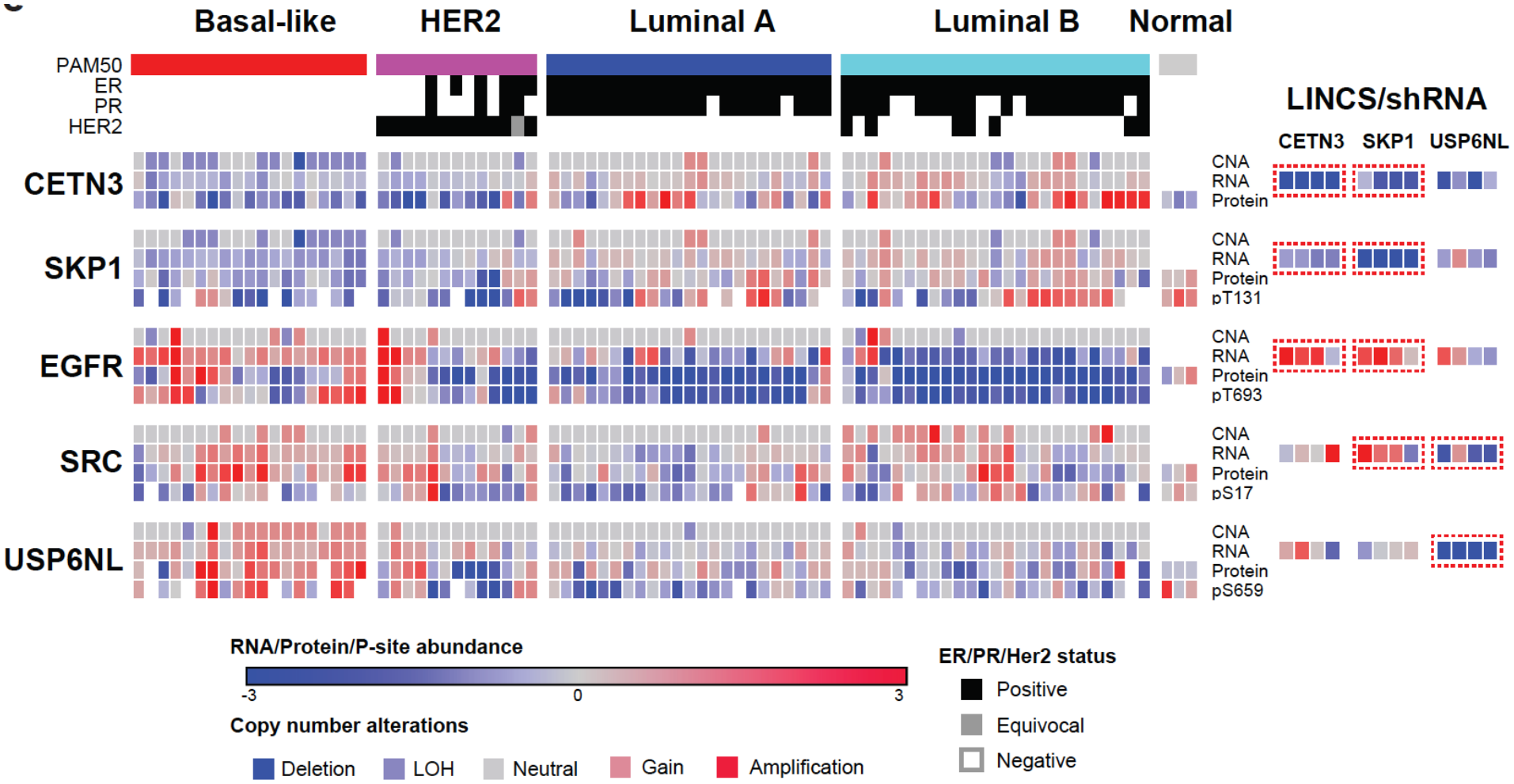
SKP1 and CETN3 are new candidate causal genes for 5q CNA TRANS effects





- 20 candidate causal genes were identified
- ERBB2 serves as a positive control
- CNA/Protein CIS effects are more indicative for a CMAP connection to TRANS regulated genes than CNA/RNA CIS effects

LINCS data: Knock-down of SKP1 and CETN3 increases EGFR, YES1 and DAPK3 expression



CMAP programming interface for large-scale queries

- LINCS Cloud Compute server with command line interface
  - Run large-scale batch queries using a Grid Engine
- Programming interface
  - R, Python, Matlab
  - For accessing and manipulating LINCS data and results
- Web-based API
  - For accessing and querying metadata
    - Perturbations and perturbagens
    - Signatures
    - Measured and inferred genes

- HDF5 (hierarchical data format) for storing data
- Mongo DB for metadata

Challenges and Implementation

Summary

- Automated pipeline enables high throughput analysis of CPTAC data
  - Reproducible and documented process
  - Version controlled
  - Generalizable to other projects
  - Easy comparison of alternatives
  - Effective use of parallelism

- Marker selection and classification can be used for any analysis
  - Automated
  - Multiple ML methods

Acknowledgments

BROAD INSTITUTE

- **Steve Carr**
- **Karl Clauser**
- **Michael Gillette**
- **Jana Qiao**
- Lauren Tang
- **Philipp Mertins**
- **D R Mani**
- Karsten Krug
- Eric Kuhn
- Filip Mundt
- Corey Flynn
- Jacob Asiedu
- Aravind Subramaniyan

FHCRC

- Amanda Paulovich
- Jeffrey Whiteaker
- Pei Wang
- Sean Wang
- Chenwei Lin
- Ping Yan
- Yuzheng Zhang

NCI STAFF

- Emily Boja
- Mehdi Mesri
- Rob Rivers
- Chris Kinsinger
- **Henry Rodriguez**

WASHINGTON U./ NYU/UNC/ VANDERBILT

- Sherri Davies
- **Matthew Ellis**
- **Reid Townsend**
- **Li Ding**
- Song Cao
- Michael McLellan
- Kuan-lin Huang
- Venkata Yellapa
- **David Fenyo**
- **Kelly Ruggles**
- **Chuck Perou**
- Michael Gatza
- **Bing Zhang**
- Jing Wang

FUNDING: National Cancer Institute