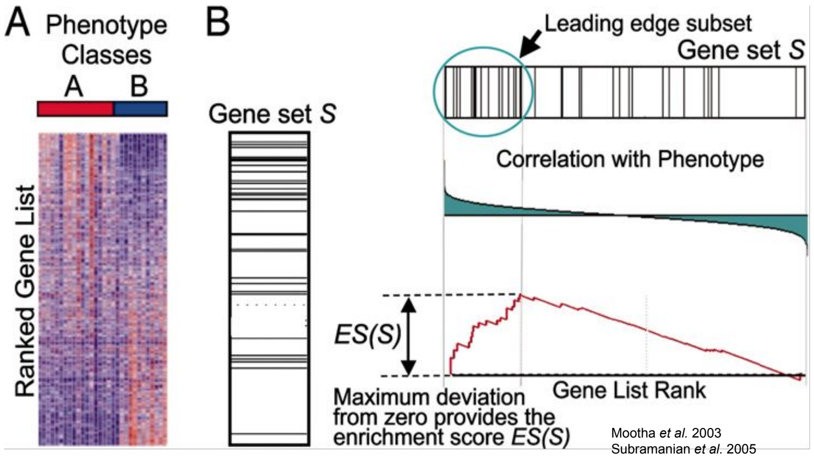
Data driven refinement of gene signatures for enrichment analysis and cell state characterization

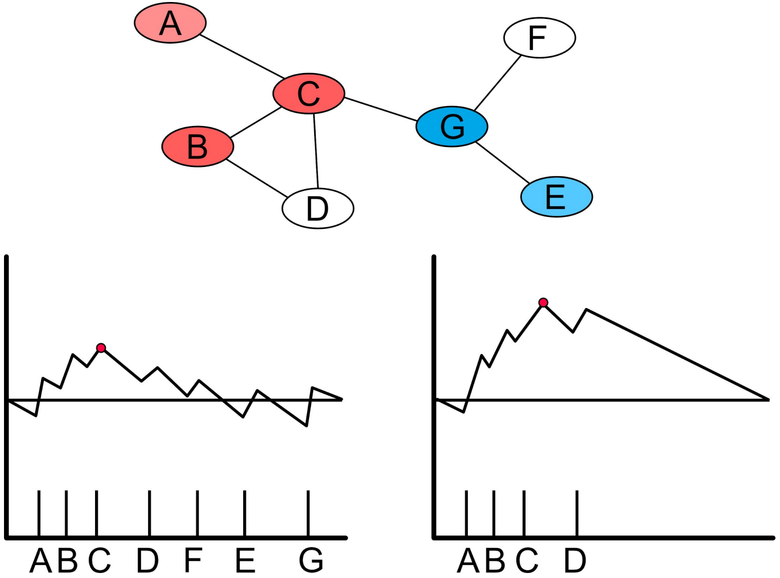
Alexander T. Wenzel
Mesirov Lab, University of California, San Diego

ITCR 2022
9/13/2022

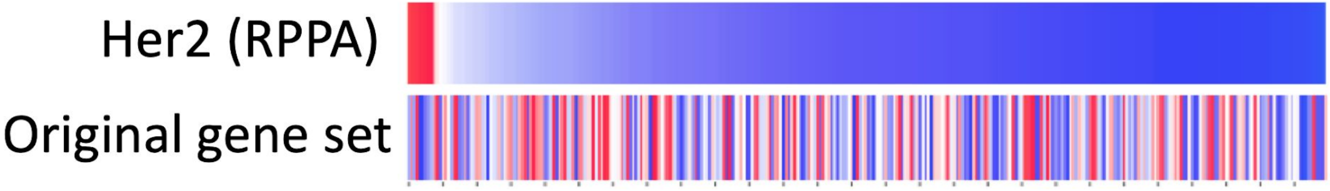
GSEA - Gene sets can lack context-specific coordinate regulation



Mootha *et al.* 2003
Subramanian *et al.* 2005



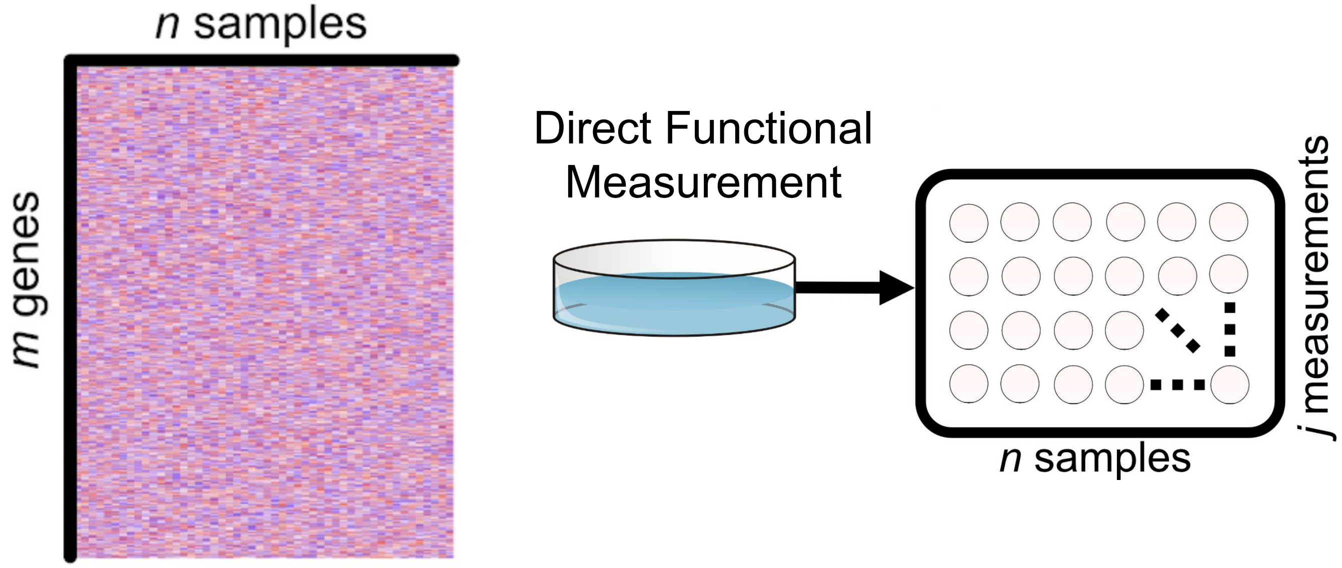
Example: REACTOME_SIGNALING_BY_ERBB2



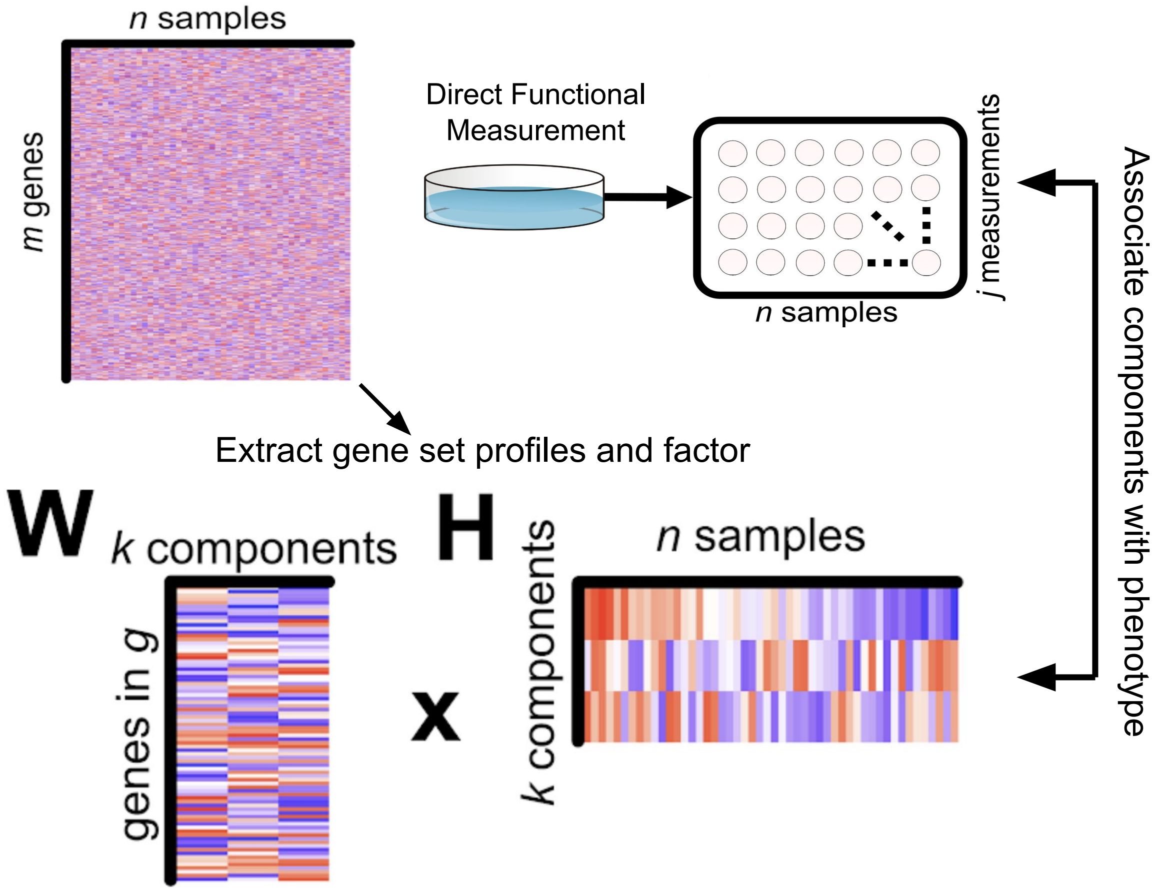
IC	p-value
0.30	0.04

Finding patterns in gene expression data

- Now we have large compendia of expression data *and* functional measurements of the same samples (RPPA, CRISPR essentiality, etc.)
- **Proposal: A data-driven method that uses these compendia to refine a gene set in order to increase its sensitivity and specificity**

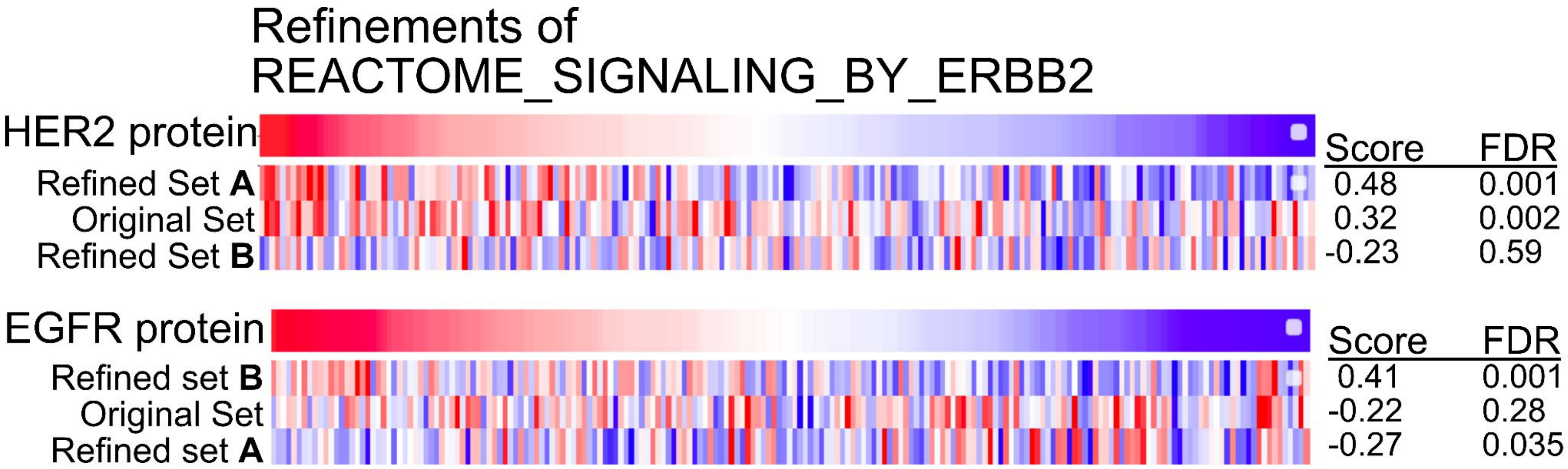


Finding patterns in gene expression data



- Non-negative matrix factorization finds lower-dimensional **patterns** in gene expression data
- Goal: Use patterns found within expression of existing gene sets to define new refined gene sets
- Decomposing gene set-defined subsets of transcriptional profiles from large compendia to identify underlying functional components.

Multi-phenotype refinement of ERBB2 signature



Refinements of REACTOME_SIGNALING_BY_ERBB2

HER2 protein	Score	FDR
Refined Set **A**	0.48	0.001
Original Set	0.32	0.002
Refined Set **B**	-0.23	0.59

EGFR protein	Score	FDR
Refined set **B**	0.41	0.001
Original Set	-0.22	0.28
Refined set **A**	-0.27	0.035

- Gene set refinement identifies two components associated with Her2 and EGFR protein abundance, respectively.
- Her2 homo-dimerizes or heterodimerizes with EGFR (Hsu & Hung 2016)
- Gene sets may represent two signaling axes

Acknowledgements

Jill Mesirov
Pablo Tamayo
Michael Reich
Ted Liefeld
Edwin Juarez
Owen Chapman
Huwate Yeerna

Thesis Committee

Jill Mesirov
Bing Ren
Pablo Tamayo
Hannah Carter
J. Silvio Gutkind

Funding

NIH T15LM011271
NIH F31CA257344