An evaluation of machine learning and traditional statistical methods for discovery in large-scale data

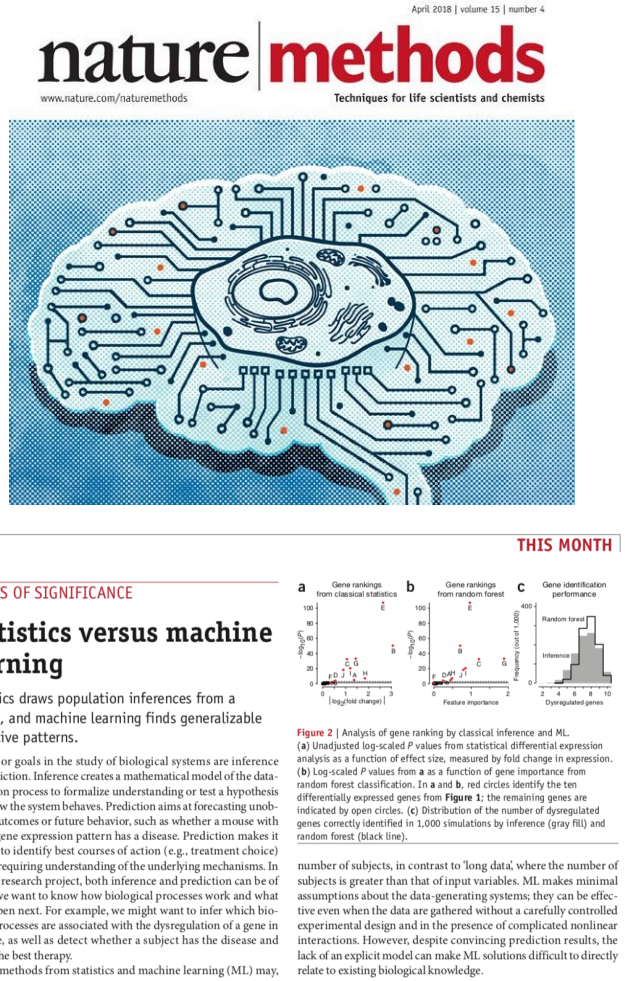
Megan Hollister

Vanderbilt University Department of Biostatistics

March 25^{th}, 2019

Table of Contents

Background



- Recent paper in *Nature Methods* on statistical discovery in large-scale data
- Concluded random forests outperformed Benjamini-Hochberg p-value based approaches
- Based on simulations of dysregulated genes in expression data
- Not all approaches were given the same a priori information

Goals

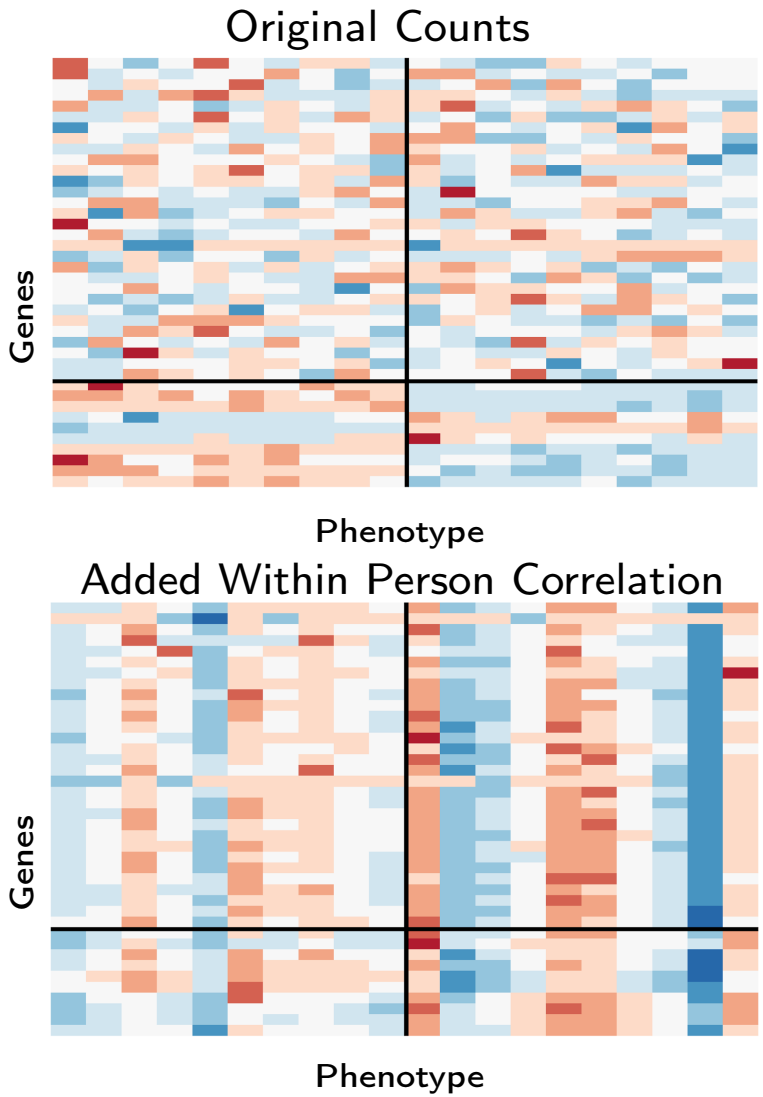
$\rightarrow$ Paper received much press and substantial twitter discussion

Objectives:

1. Examine claims using unbiased and fair comparisons
2. Estimate accuracy of machine learning and "traditional" methods
3. Identify methods with the best performance characteristics

Methods

Simulated Gene Expression Data



- 40 genes ; 20 people
- 10 phenotype positive ; 10 negative
- 25% (10) of genes are "dysregulated" across phenotype
- Computed pseudo-counts = normalized counts (Robinson and Smyth, 2008)

- *Allowed within person correlation across genes (new)*

Methods

Discovery Methods

Traditional	**Machine Learning**
Nominal *p*-values	Random Forest importance levels
Bonferroni adjusted *p*-values	Neural Net prediction weights
Benjamini-Hochberg Emp FDRs	Penalized Regression (forthcoming)
Second-generation *p*-values	

1. 5% significance level / FWER / FDR
2. Top 10 ranked genes by ML criteria
3. *Top 10 ranked genes by Traditional criteria (new)*

Second Generation p-values

Overview

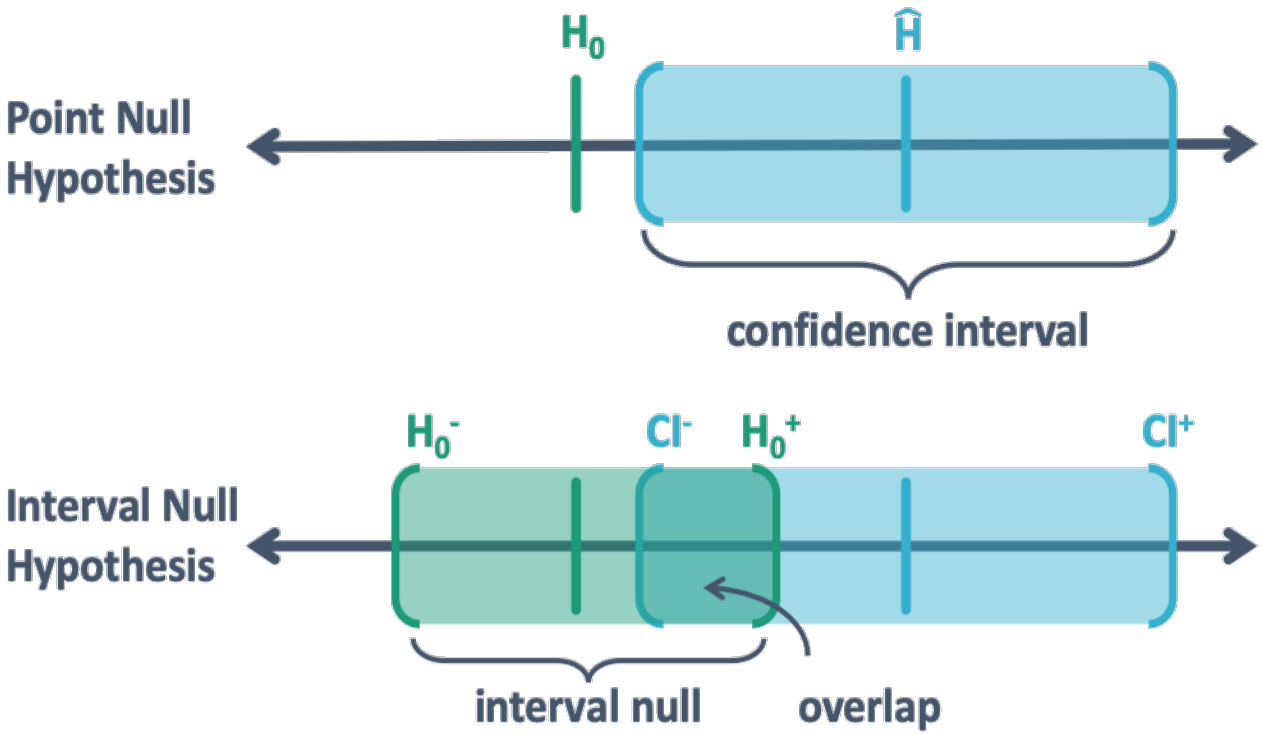
- SGPV is in $[0, 1]$ and denoted by p_δ
- δ indicates dependence on (pre-specified) interval null hypothesis

- SGPV reports the fraction of data-supported effect sizes that are null or trivial
- Adjustment for multiple comparisons is automatic

- **Cases:**
  1. $p_\delta = 0$ when data incompatible with null region
  2. $p_\delta = 1$ when data compatible with null region
  3. $0 < p_\delta < 1$ when data are inconclusive

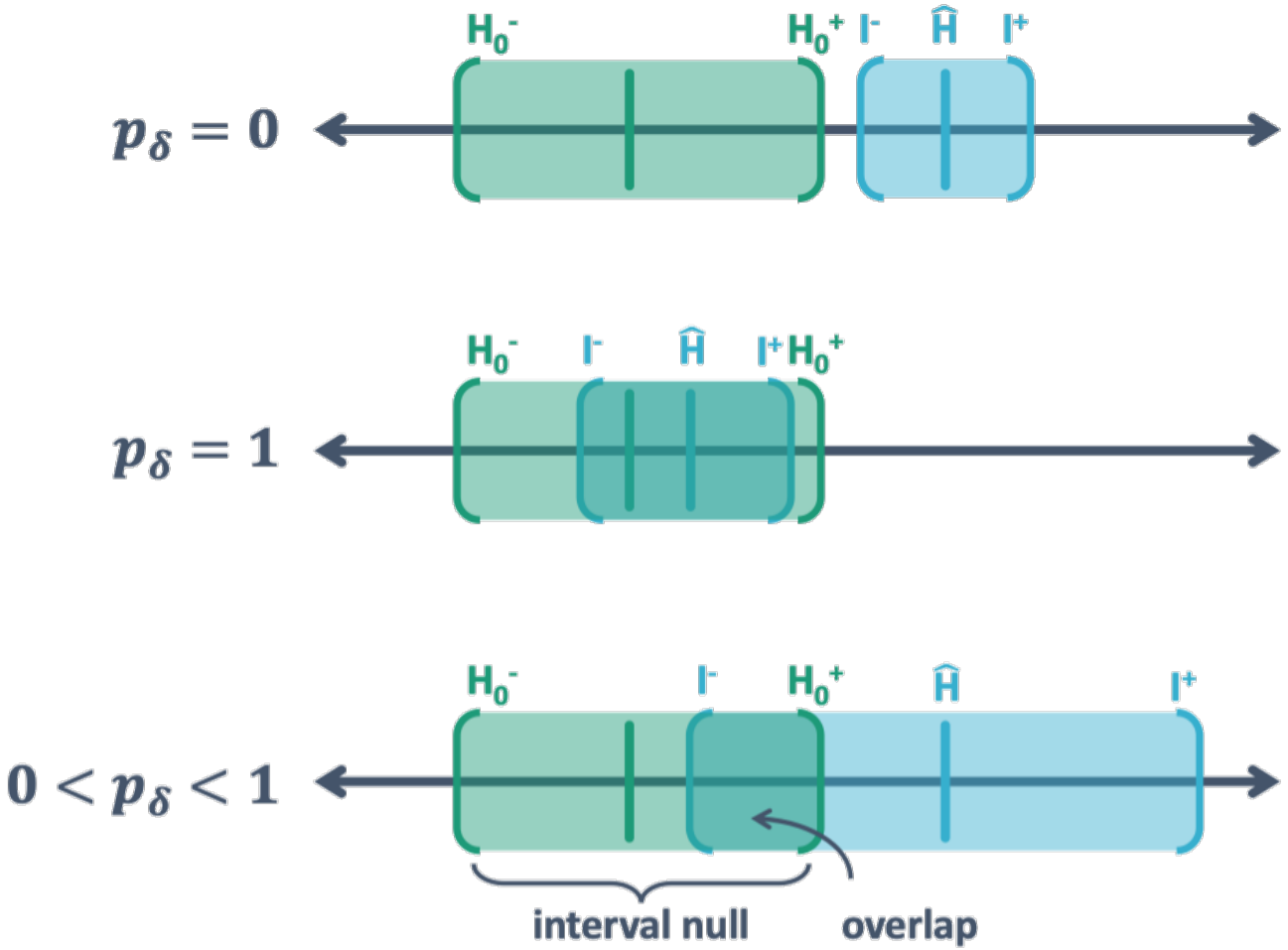
Second Generation p-values

Illustration 1



Second Generation p-values

Illustration 2

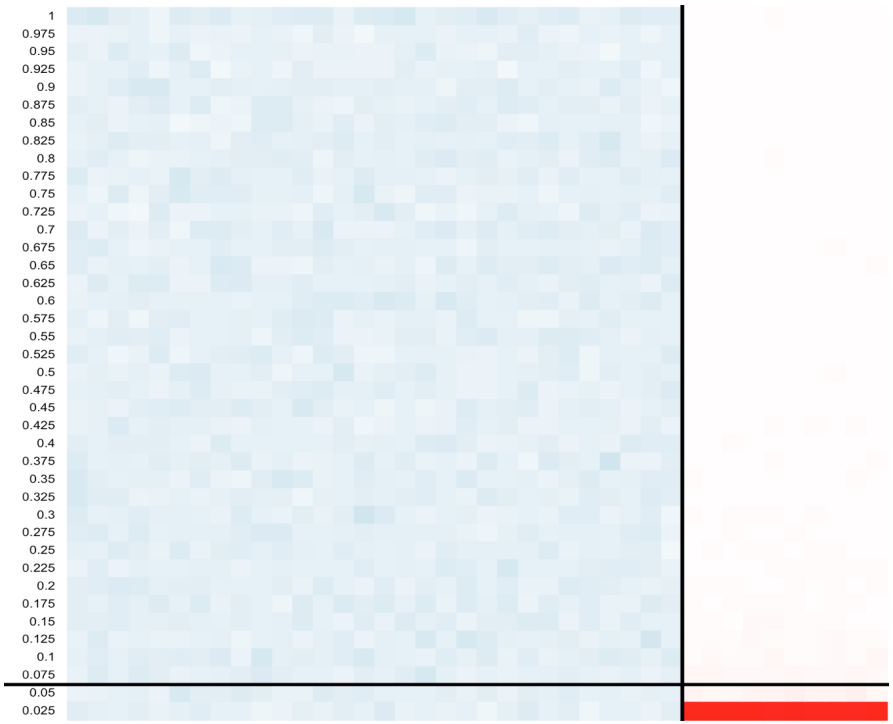


Results

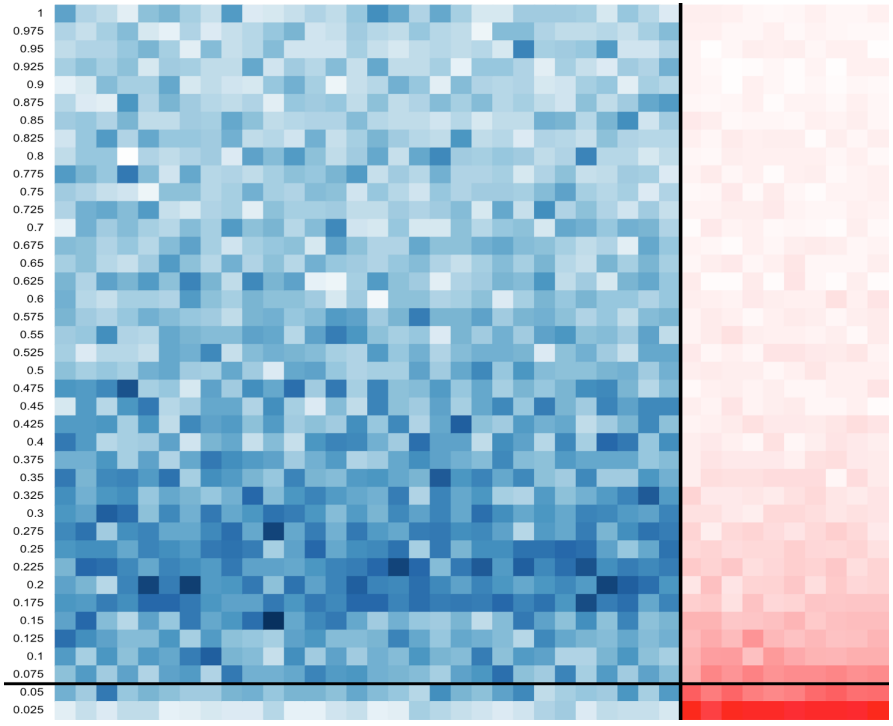
Heatmaps of *p*-values

- Heatmap of discovery *p*-values by nominal *p*-values
- Values below horizontal line less than 0.05

Nominal *p*-values



Nominal *p*-values with
Within Person Correlation

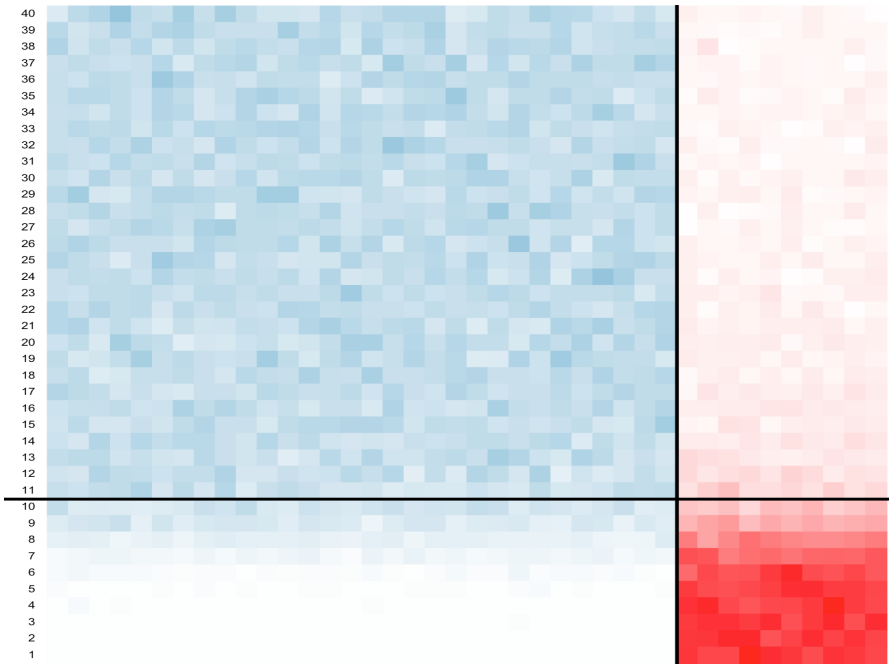


Results

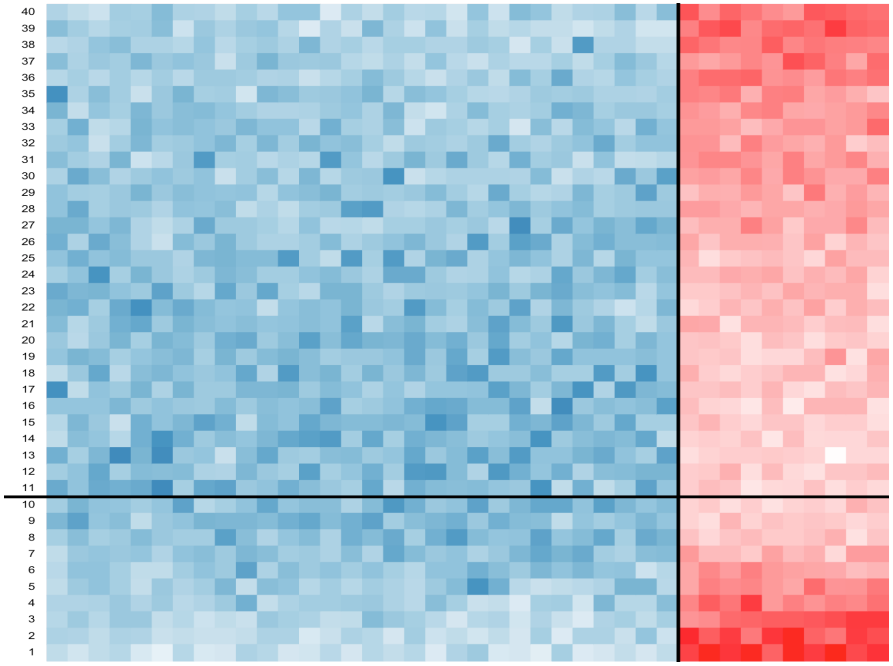
Heatmaps of Rankings

- Heatmap of gene rankings by FDR (Benjamini-Hochberg)
- Top 10 rankings below horizontal line

Rankings of Original Counts



Rankings of Counts with Within Person Correlation

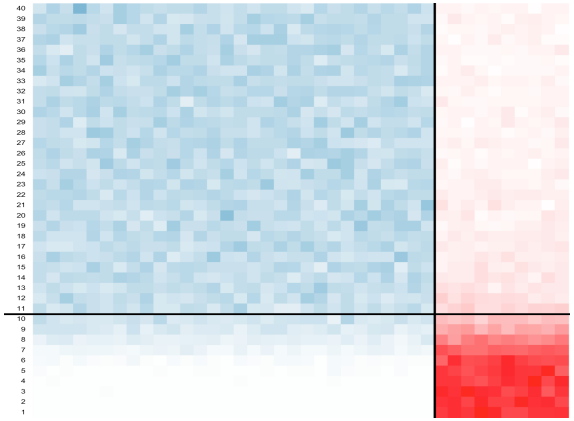


Results

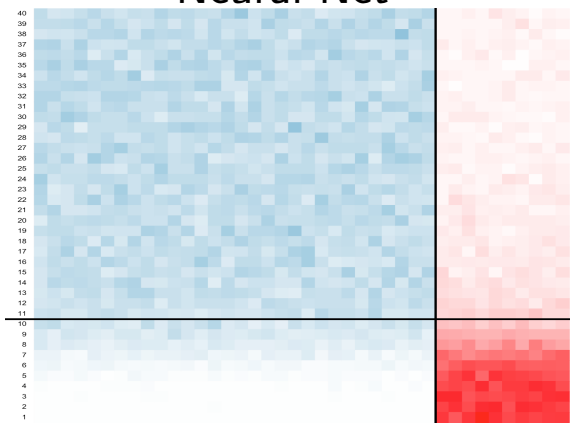
Heatmaps of Rankings

- Heatmaps of rankings of the original gene expression counts

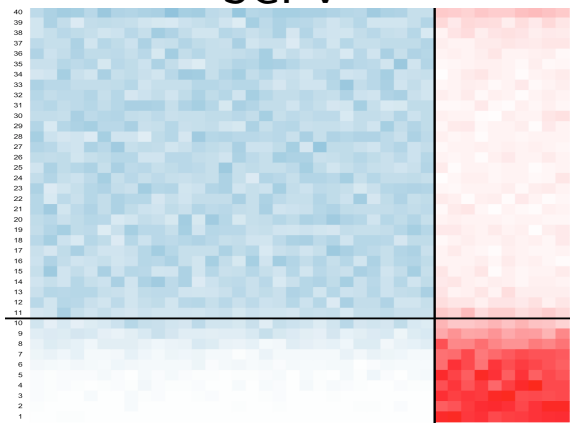
Nominal p-values



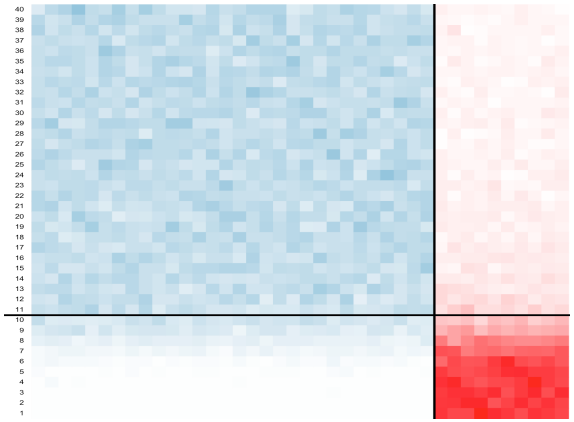
Neural Net



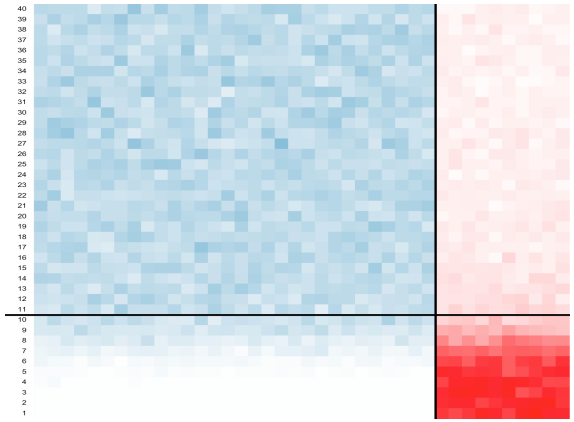
SGPV



BH Emp FDRs

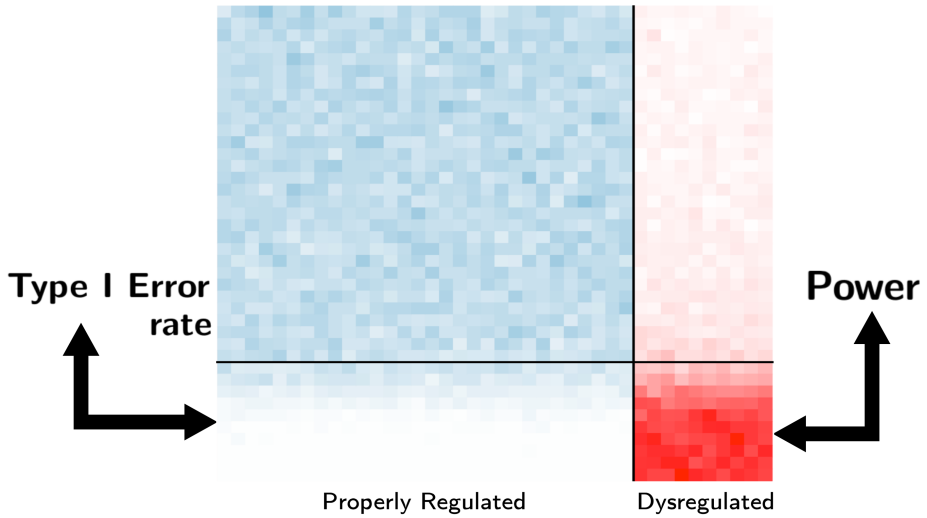


Random Forest



Results

Comparisons

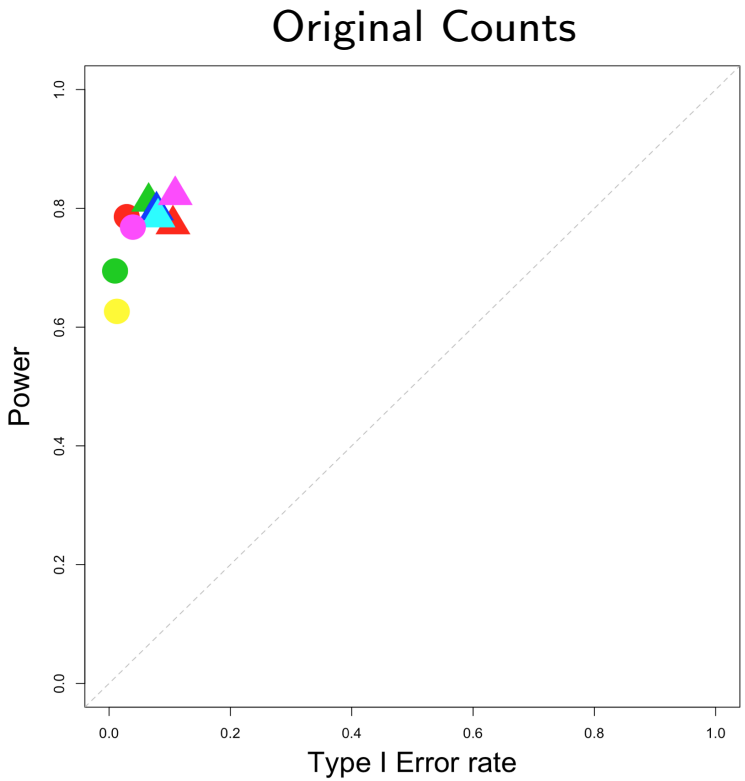


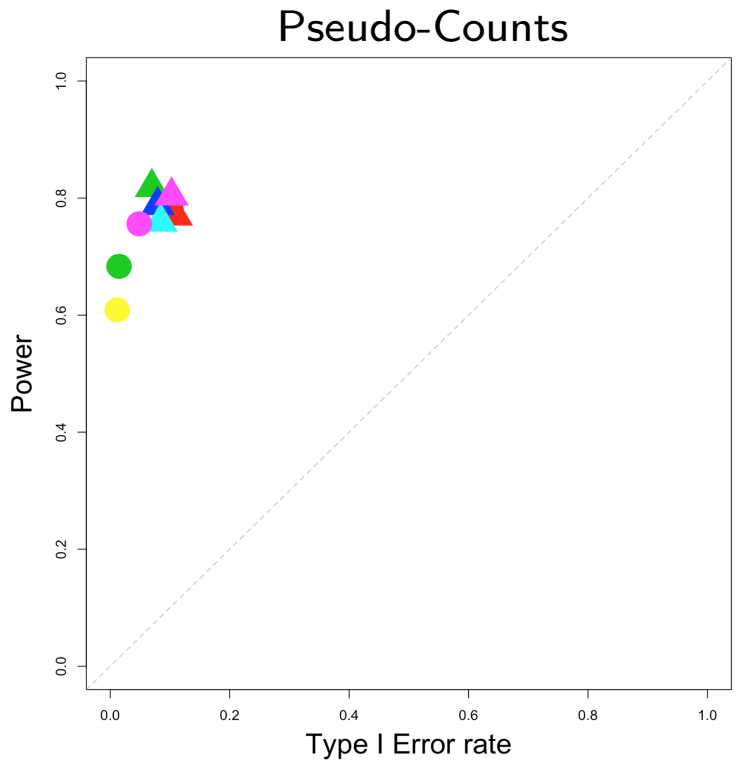
Accuracy statistics:

- **Power**
  - → Proportion of "dysregulated" genes identified as "dysregulated"
- **Type I Error rate**
  - → Proportion of "properly regulated" genes identified as "dysregulated"

Results

Comparisons

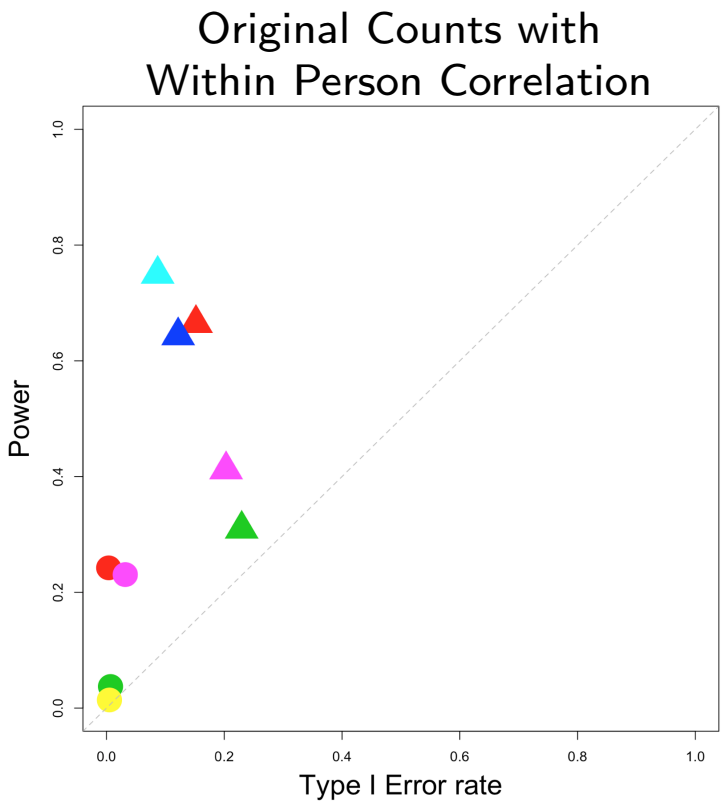


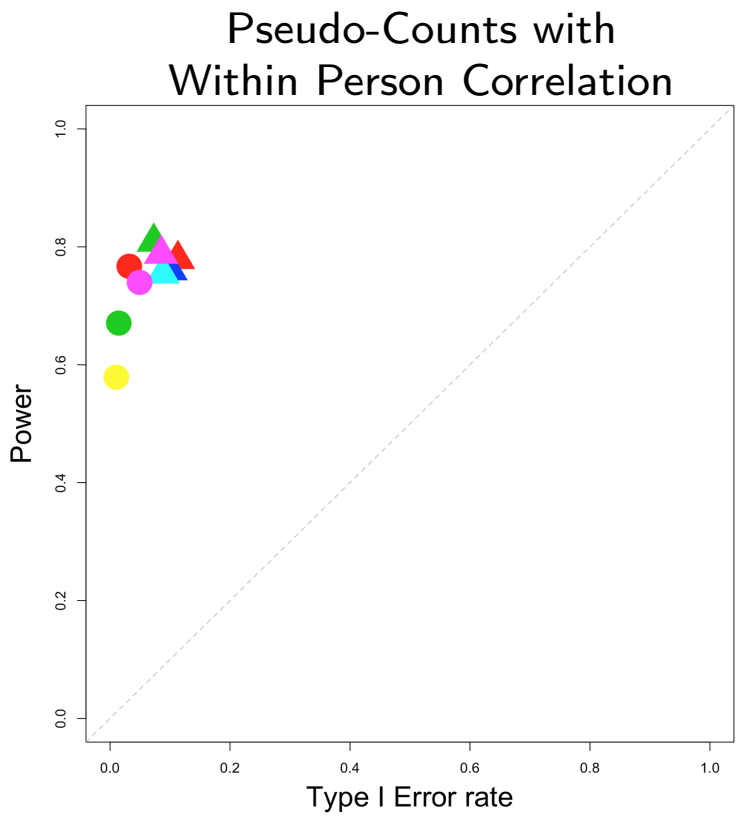


Nominal P-value
Nominal P-value(rank)
Bonferroni
Benjamini-Hochberg
Benjamini-Hochberg(rank)
Random Forest(rank)
Second Generation P-value
Second Generation P-value(rank)
Neural Net(rank)

Results

Comparisons





Nominal P-value
Nominal P-value(rank)
Bonferroni
Benjamini-Hochberg
Benjamini-Hochberg(rank)
Random Forest(rank)
Second Generation P-value
Second Generation P-value(rank)
Neural Net(rank)

Conclusions

- Normalizing step is critical for some methods
- Methods perform identically when properly compared (by rankings)
- Comparing ranking vs threshold discovery gives *false* impression of differential statistical accuracy (ie, *Nature Methods*)

	Traditional Methods	**Machine Learning**
Pros	• Significance level criterion • Can be ranked • Interpretable coefficients	• Handles complexity with ease • Variety of flexible algorithms
Cons	• Complexity poses challenges • Significance criterion not universal • Models can be simplistic	• Must pre-specify number of findings • No threshold criterion • Coefficients hard to interpret

Acknowledgments

NIH Clinical and Translational Science Awards (CTSA) TL1 Training Grant
Statistical Evidence in Data Science (SEDS) Lab:

- Dr. Jeffrey D. Blume (PI) www.statisticalevidence.com
- Dr. Thomas Stewart
- Valerie Welty

References:

1. Bzdok, Danilo and Altman, Naomi and Krzywinski, and Martin (2018). Statistics versus machine learning. *Nature Methods* 15, 233-234. https://doi.org/10.1038/nmeth.4642
2. Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140
3. Blume JD, Greevy RA Jr., Welty VF, Smith JR, Dupont WD (2019). An Introduction to Second-generation *p*-values. *The American Statistician*. https://doi.org/10.1080/00031305.2018.1537893