Pathway recognition and augmentation by computational analysis of microarray expression data

(Barbara A. Novak and Ajay N. Jain)

Priit Adler

JournalClub
13. march 2006.

result

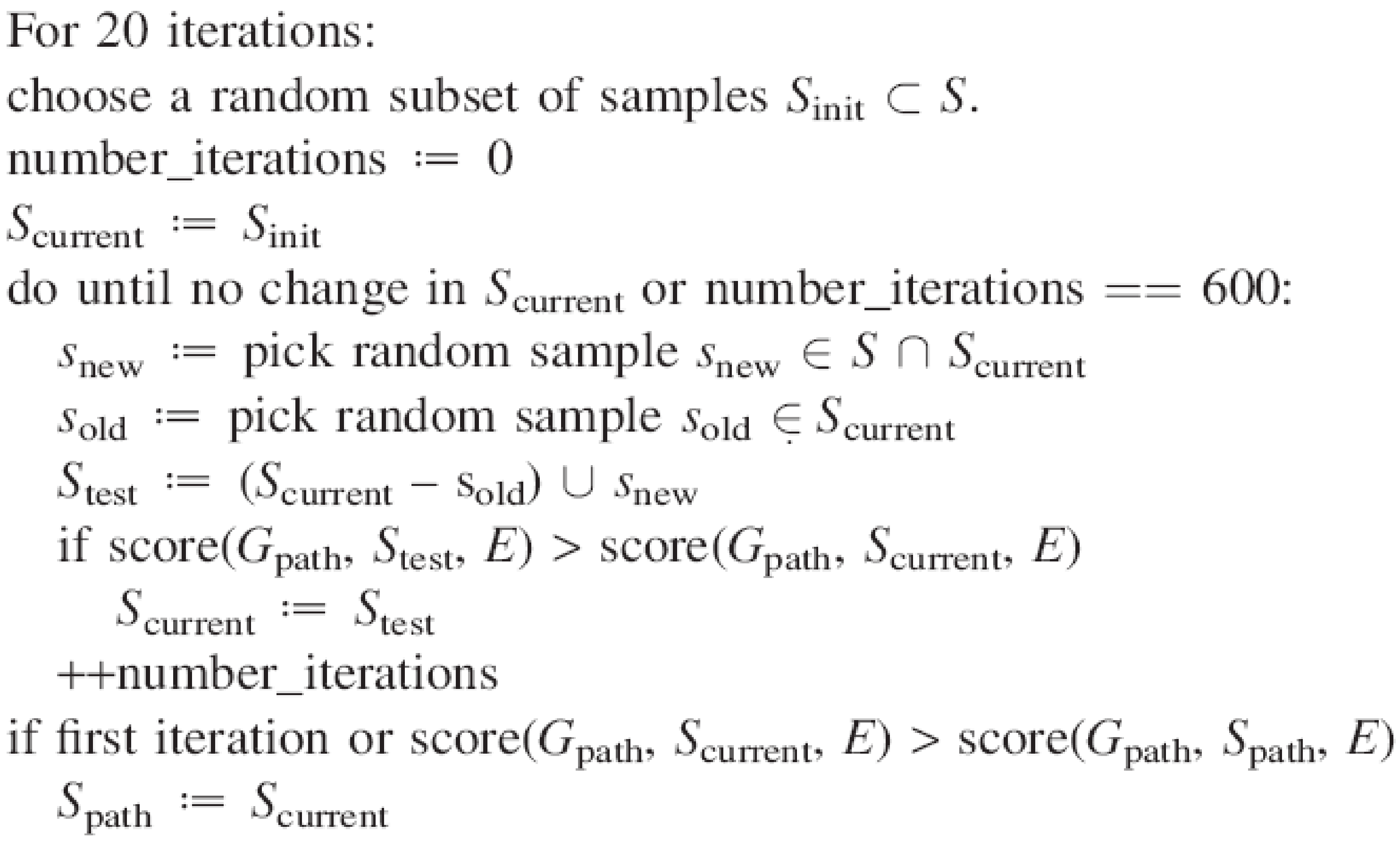
- they present a system, QPACA (Qunatitative Pathway Analysis in Canser)
  - supports data visualization and both fine- and coarse-grained specification
  - **adresses the problems of** *pathway recognition and pathway augmentation*

why ?

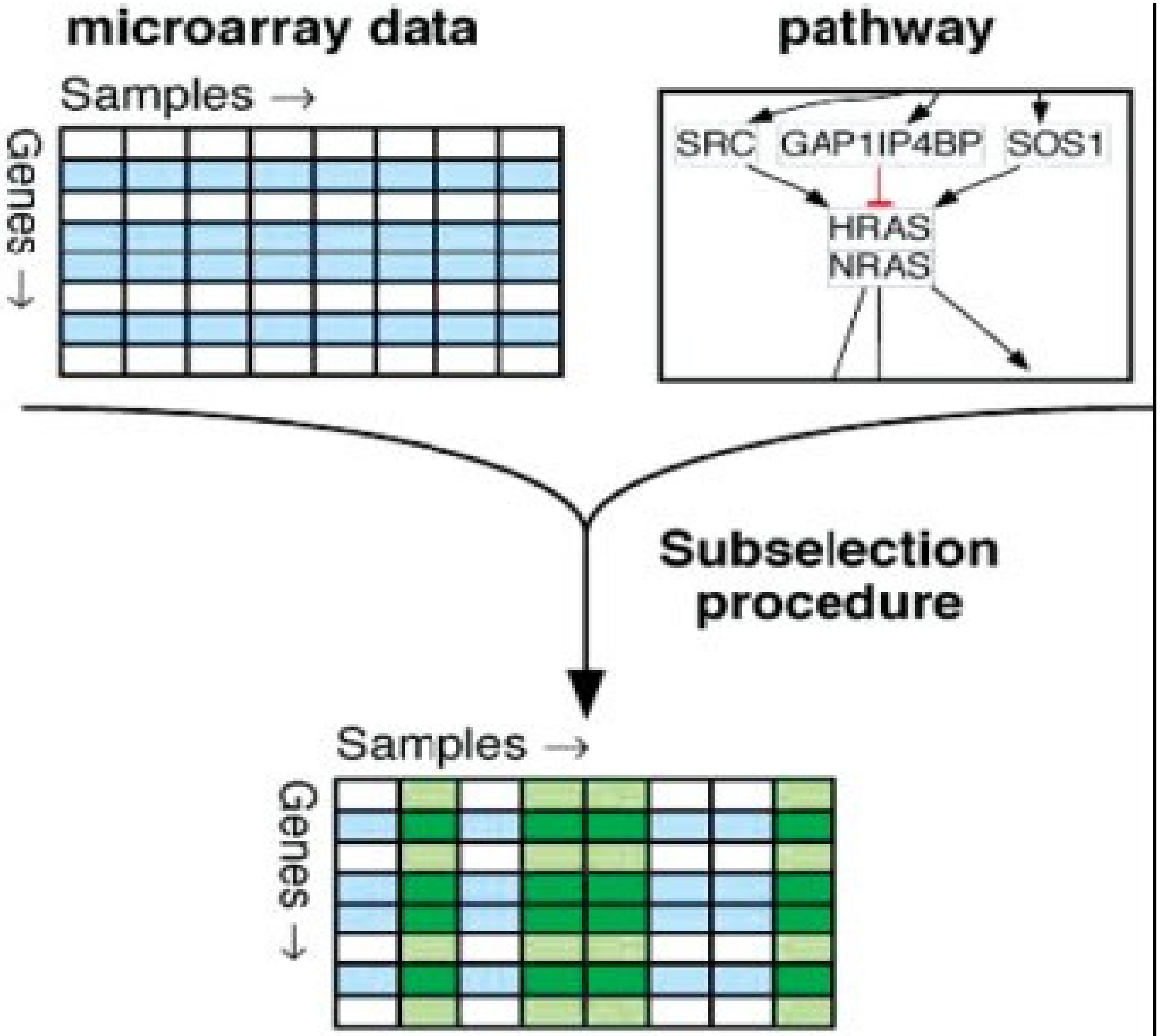
- there's quite a bit of experimental data, but is it always revalent ?
- same condition may be resulted from different causes (like cancer from different tumors).

method

- optimization



```
For 20 iterations:
choose a random subset of samples $S_{\text{init}} \subset S$.
number_iterations $:=$ 0
$S_{\text{current}} := S_{\text{init}}$
do until no change in $S_{\text{current}}$ or number_iterations $==$ 600:
  $s_{\text{new}} :=$ pick random sample $s_{\text{new}} \in S \cap S_{\text{current}}$
  $s_{\text{old}} :=$ pick random sample $s_{\text{old}} \underset{\cdot}{\in} S_{\text{current}}$
  $S_{\text{test}} := (S_{\text{current}} - \mathrm{s}_{\text{old}}) \cup s_{\text{new}}$
  if score($G_{\text{path}}, S_{\text{test}}, E$) > score($G_{\text{path}}, S_{\text{current}}, E$)
     $S_{\text{current}} := S_{\text{test}}$
  ++number_iterations
if first iteration or score($G_{\text{path}}, S_{\text{current}}, E$) > score($G_{\text{path}}, S_{\text{path}}, E$)
  $S_{\text{path}} := S_{\text{current}}$
```

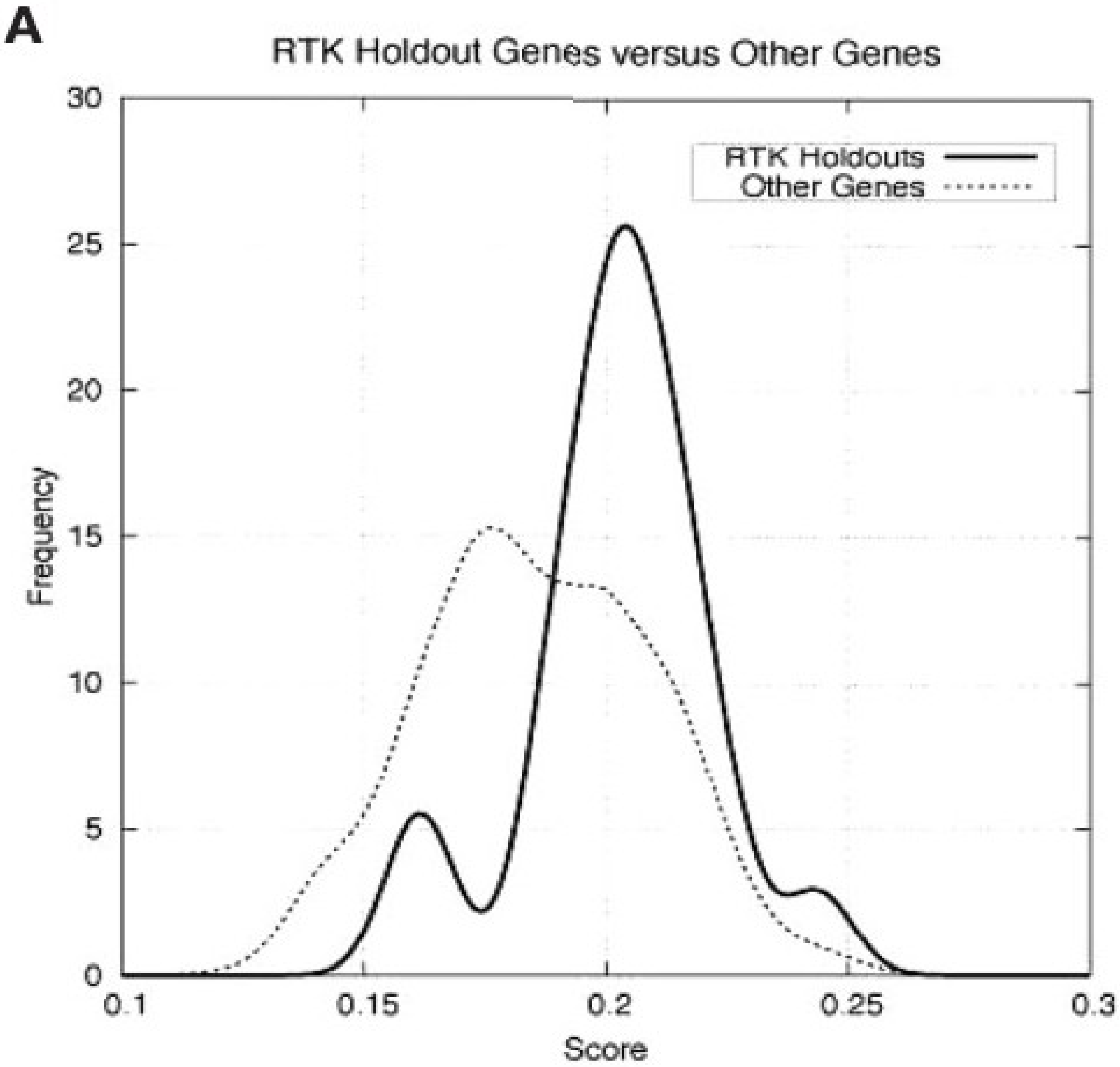


a subset of experiments (green) that is optimized for high score among the pathway genes (blue)

- performance evaluation
  - one possible application is to test whether a gene set is behaving in a coordinated fashion, and whether the coordination is relevant to our notion of a biological pathway
  - also can apply question of biological pathway augmentation
- datasets
  - yeast
  - human (cancer)

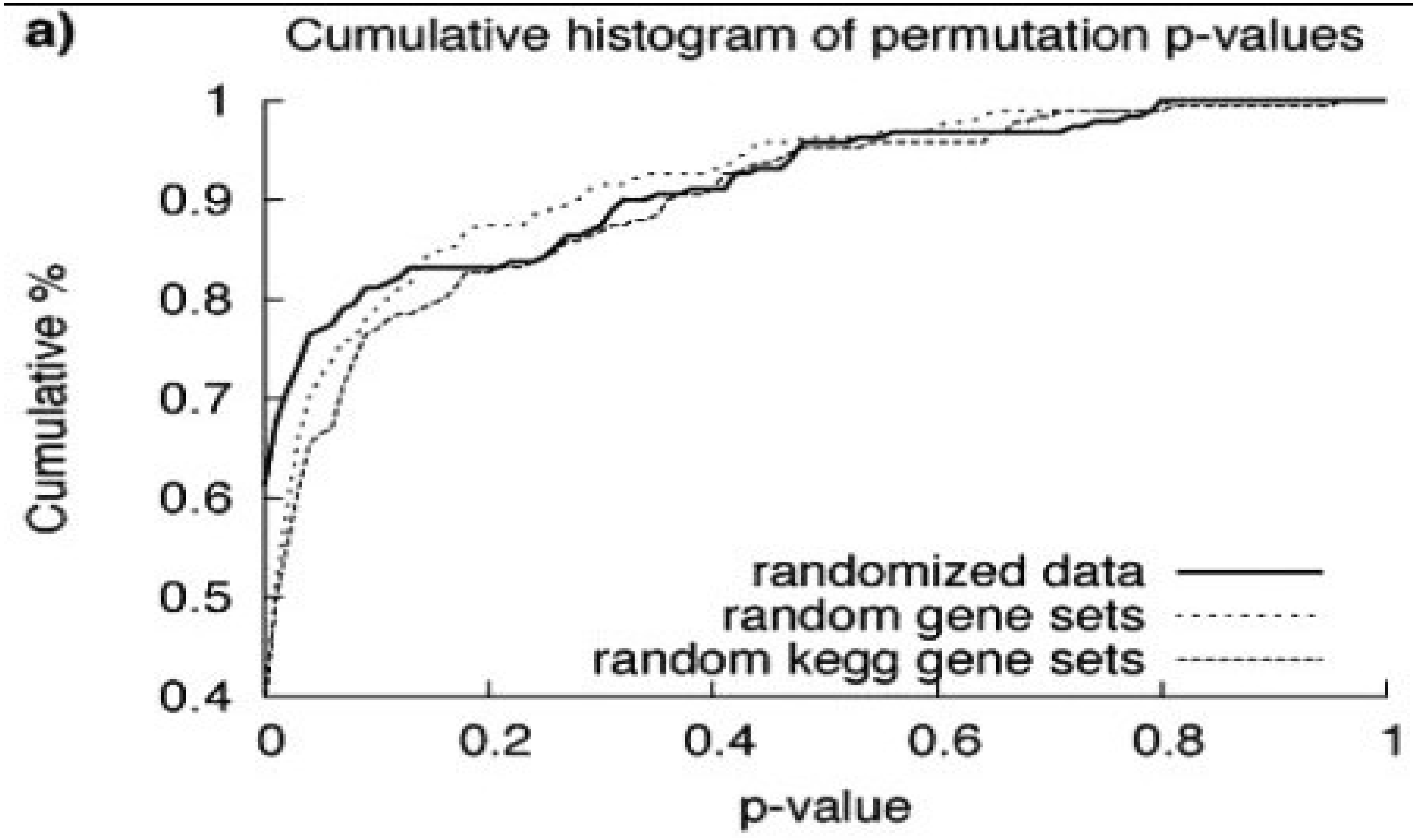
- Permutation testing and *p*-value calculation
  - scores resulting from randomly chosen gene sets of the same size as the pathway set in question
  - scores resulting from randomizing the gene expression data itself
  - scores resulting from randmoly chosen gene sets of the same size as pathway set in question, restricted to genes represented within KEGG

- Cross-validation in biological pathway augmentation
  - pathways with all tree permutation-based *p*-valuse < 0.1 where excluded. (don't add anything if can't recognise pathway)
  - from rest 10 % of genes were heldout,
  - then scores calculated to those that left
  - added heldouts and backround
  - ranked by score.

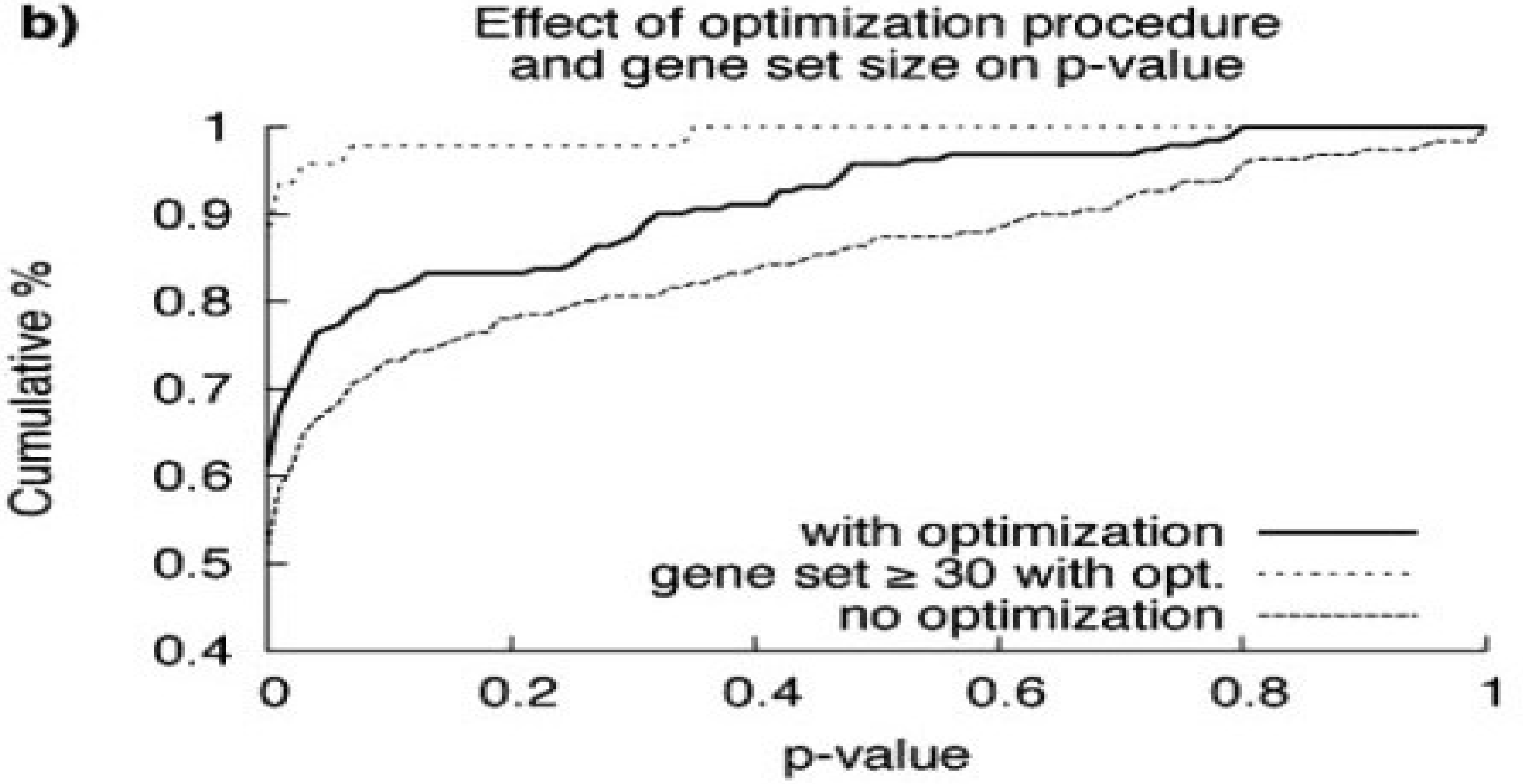

A
RTK Holdout Genes versus Other Genes
RTK Holdouts
Other Genes
Frequency
30
25
20
15
10
5
0
0.1
0.15
0.2
0.25
0.3
Score

results

- *p*-value *x* can be interpreted as the probability that a random set of genes will yield a *p*-value of less than or equal to *x*

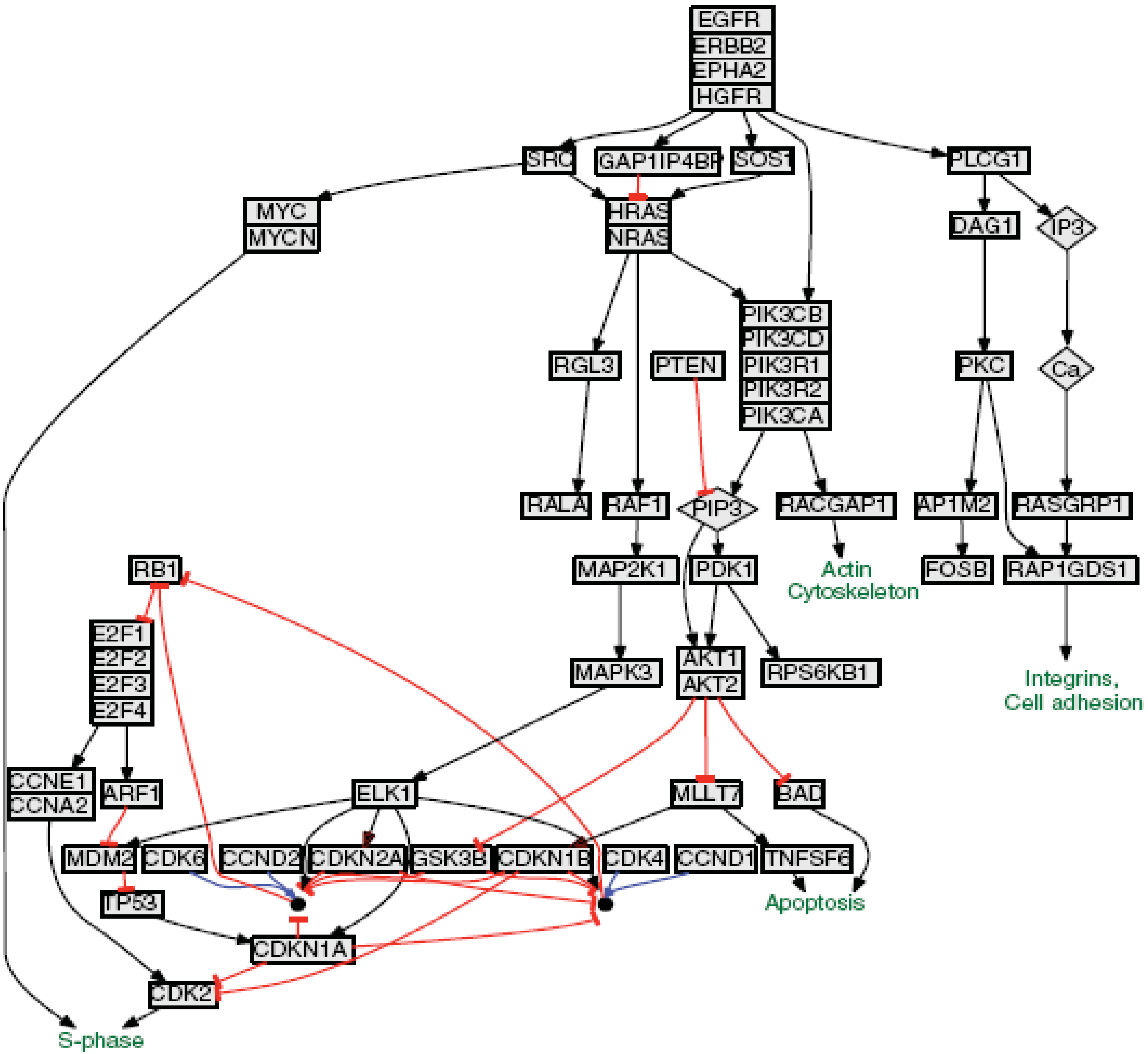


Cumulative histograms of permutation-based *p*-values for all analyzed pathways. Most pathways (117/191 or 61%) had significant scores (p 0.05) under the most stringent of these methods, indicating that this method is reproducible over a large and varied set of pathways.

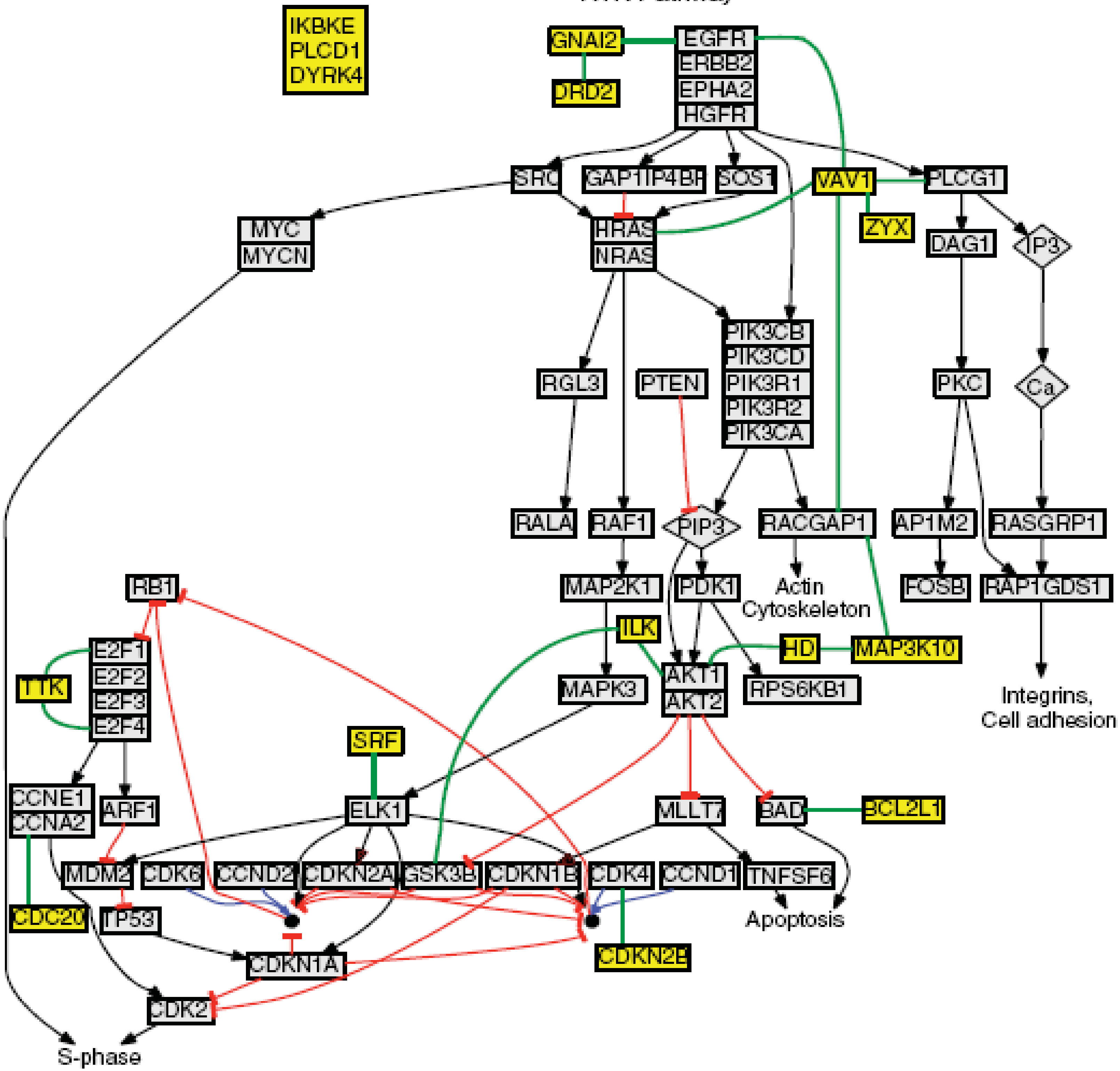


Cumulative histograms of the *p*-values with optimization for all pathways (solid line), all pathways without optimization (dotted line) and for those pathways with >= 30 genes using optimization (dashed line). For each case, the randomization permutation method was used. The optimization method clearly improves the *p*-value distribution over no optimization (p << 0.001 by Mann–Whitney), and restriction to larger gene set sizes also improves the p-value distribution (p << 0.001 by Mann–Whitney).

Thank you !


RTK Pathway
EGFR
ERBB2
EPHA2
HGFR
SRC
GAP1IP4BP
SOS1
PLCG1
MYC
MYCN
HRAS
NRAS
DAG1
IP3
PIK3CB
PIK3CD
PIK3R1
PIK3R2
PIK3CA
RGL3
PTEN
PKC
Ca
RALA
RAF1
PIP3
RACGAP1
AP1M2
RASGRP1
RB1
MAP2K1
PDK1
Actin
Cytoskeleton
FOSB
RAP1GDS1
E2F1
E2F2
E2F3
E2F4
MAPK3
AKT1
AKT2
RPS6KB1
Integrins,
Cell adhesion
CCNE1
CCNA2
ARF1
ELK1
MLLT7
BAD
MDM2
CDK6
CCND2
CDKN2A
GSK3B
CDKN1B
CDK4
CCND1
TNFSF6
TP53
Apoptosis
CDKN1A
CDK2
S-phase

RTK Pathway


IKBKE
PLCD1
DYRK4
GNAI2
DRD2
EGFR
ERBB2
EPHA2
HGFR
SRC
GAP1IP4BP
SOS1
VAV1
PLCG1
ZYX
MYC
MYCN
HRAS
NRAS
DAG1
IP3
PIK3CB
PIK3CD
PIK3R1
PIK3R2
PIK3CA
RGL3
PTEN
PKC
Ca
RALA
RAF1
PIP3
RACGAP1
AP1M2
RASGRP1
RB1
MAP2K1
PDK1
Actin
Cytoskeleton
FOSB
RAP1GDS1
ILK
HD
MAP3K10
E2F1
E2F2
E2F3
E2F4
TTK
AKT1
AKT2
MAPK3
RPS6KB1
Integrins,
Cell adhesion
SRF
CCNE1
CCNA2
ARF1
ELK1
MLLT7
BAD
BCL2L1
MDM2
CDK6
CCND2
CDKN2A
GSK3B
CDKN1B
CDK4
CCND1
TNFSF6
CDC20
TP53
Apoptosis
CDKN2B
CDKN1A
CDK2
S-phase