Adding value through statistics

(Gene ranking with desirability functions and Bayesian rank aggregation)

Stanley E. Lazic, PhD

5 Nov 2015

Selecting genes

Case 1: Want to prioritise/rank/select genes from a single list.

Case 2: Want to prioritise/rank/select genes from multiple lists.

Desirability approach

1) Choose variables for selection criteria
2) Map values to 0–1 with desirability functions
3) Calculate the overall desirability as a weighted combination of the individual desirabilities

Desirability approach

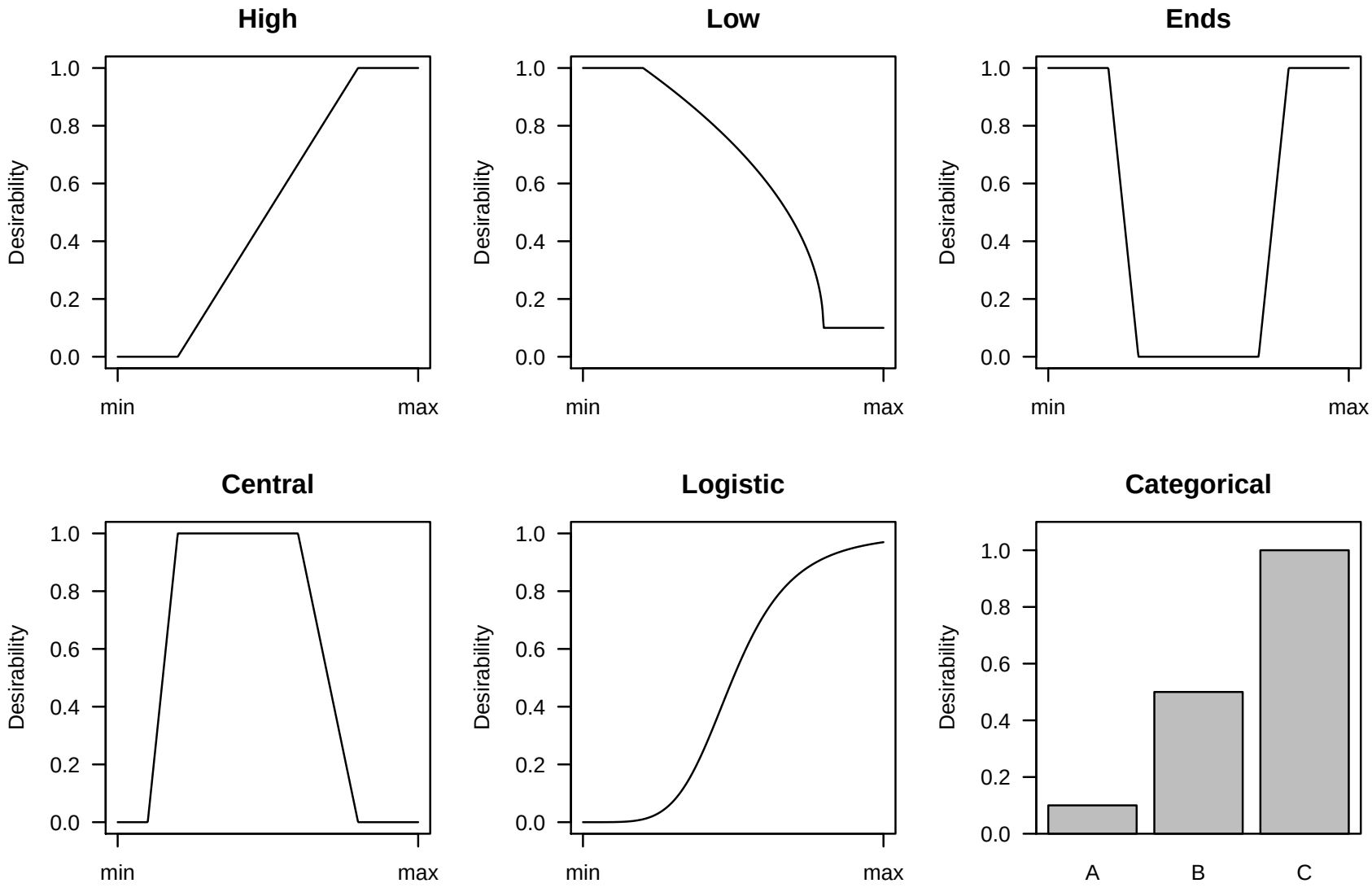
1) Choose variables for selection criteria
2) Map values to 0–1 with desirability functions
3) Calculate the overall desirability as a weighted combination of the individual desirabilities

The aim is to generalise and formalise current methods of selecting genes, and to avoid binary thresholds.

Selection criteria

- P-values, fold-change (primary criteria)
- Mean expression, variability in expression (nonspecific filters)
- Sequence similarity with human genes
- Expressed in key tissues
- In a relevant pathway, protein complex, cellular compartment
- Target of known drugs
- In a list from previous publications
- Differentially expressed in multiple conditions
- Disease specificity (XOR)
- Consistency of expression
- Based on parameters of linear models

Examples of desirability functions



Calculating the overall desirability

Geometric mean:

$$D = \left(\prod_{i=1}^{n} d_i \right)^{1/n}$$

Weighted geometric mean:

$$D = \left(\prod_{i=1}^{n} d_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i}$$

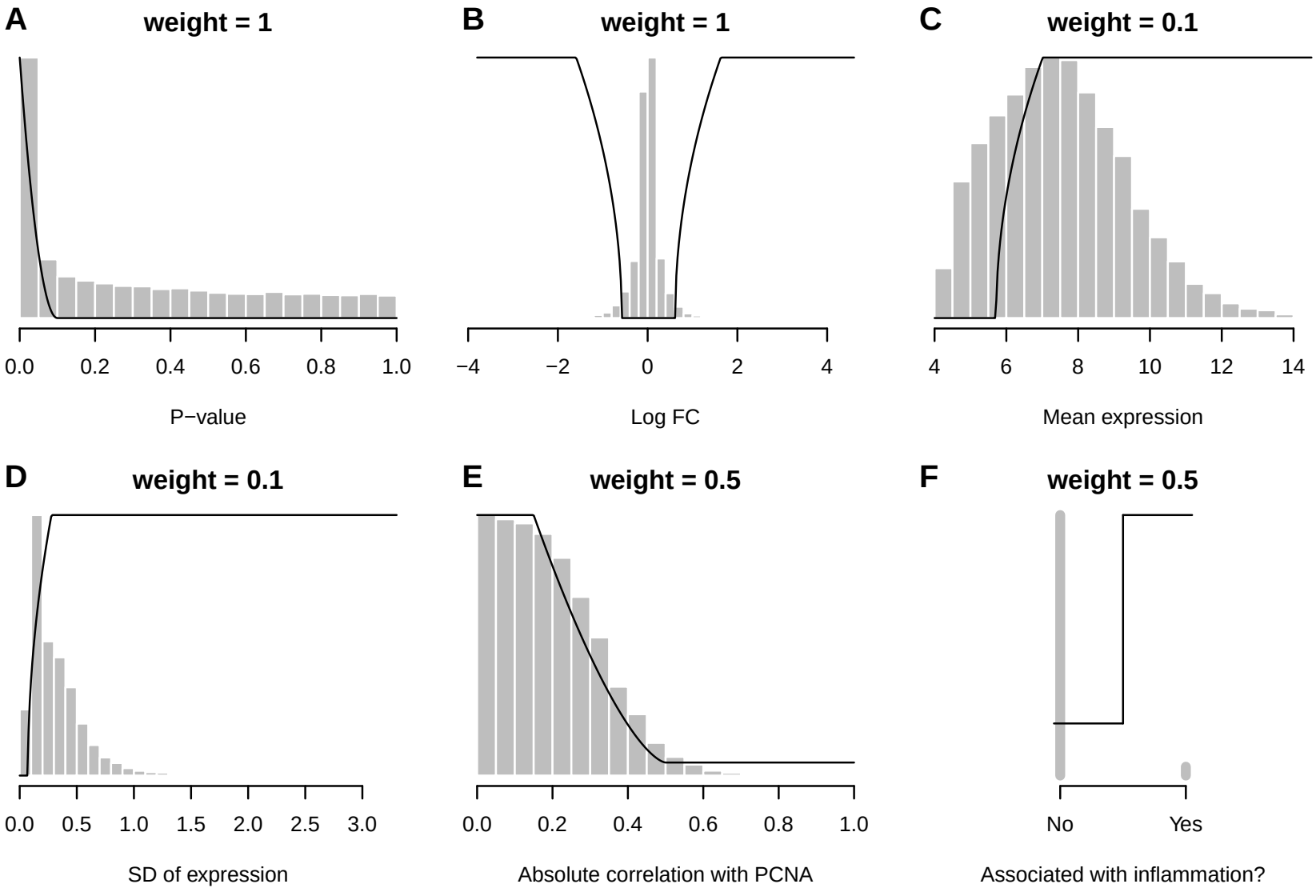
Weighted geometric mean (log-form):

$$D = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln d_i}{\sum_{i=1}^{n} w_i} \right)$$

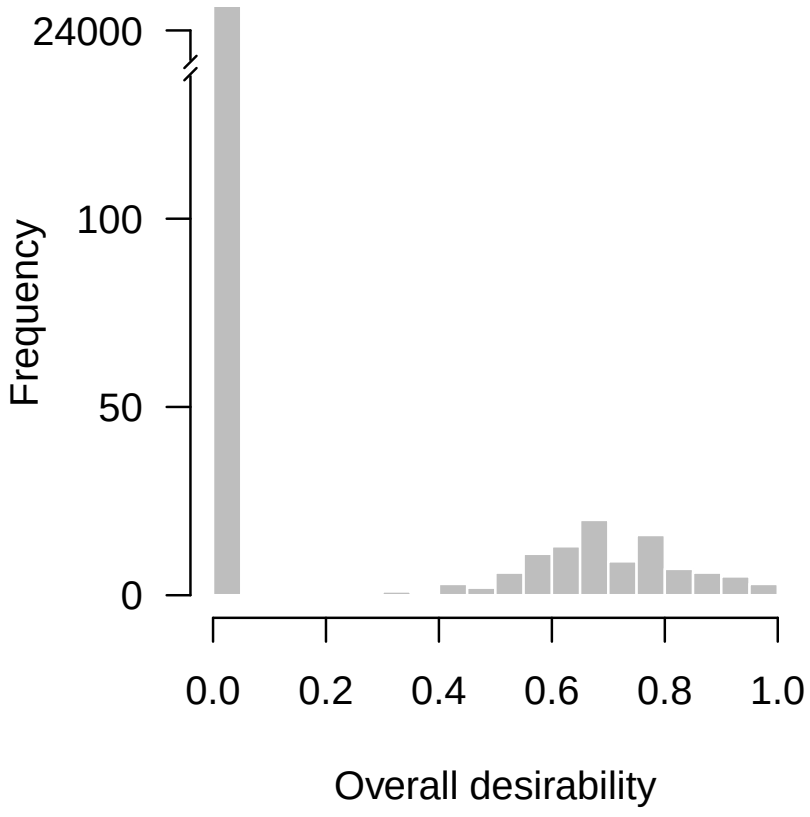
Example: Breast cancer microarray data set

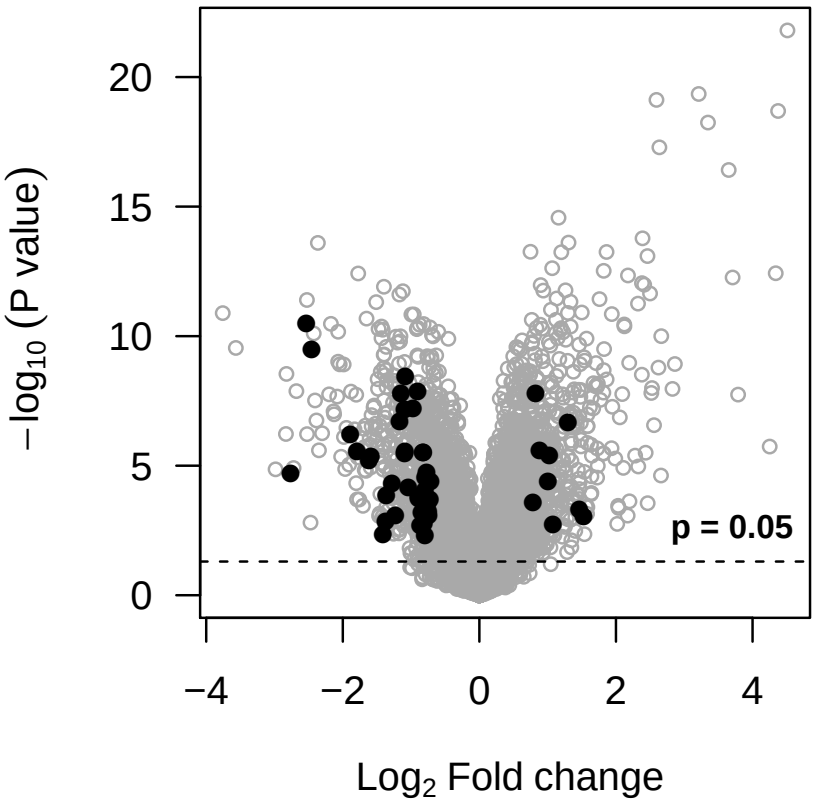
- Data from Farmer et al. *Oncogene* 2005; GEO: GDS1329
- Comparison between basal ($n = 16$) and luminal ($n = 27$) samples
- 4830 probe sets differentially expressed (FDR < 0.05)

Map criteria to 0–1 scale



Overall desirability





Top ten probe sets sorted by overall desirability

Probeset	Gene	logFC	AveExpr	P-value	P-rank	Overall D
202917_s_at	*S100A8*	2.76	9.42	1.1e-05	935	1.00
204470_at	*CXCL1*	1.59	6.49	1.5e-06	668	0.95
214038_at	*CCL8*	1.36	8.80	1.0e-04	1521	0.95
203535_at	*S100A9*	1.41	7.64	5.6e-03	3168	0.93
210029_at	*IDO1*	1.23	8.90	4.5e-04	2249	0.92
209924_at	*CCL18*	-1.79	9.49	2.7e-06	594	0.92
32128_at	*CCL18*	-1.89	9.43	5.9e-07	440	0.91
206214_at	*PLA2G7*	-1.28	7.55	4.6e-05	1147	0.90
221698_s_at	*CLEC7A*	-1.09	8.15	3.4e-06	625	0.88
216598_s_at	*CCL2*	-1.17	8.81	2.0e-07	347	0.87

Extension to multiple experiments (data integration)

Luo	Welsh	Dhana	True	Singh
HPN	HPN	OGT	AMACR	HPN
AMACR	AMACR	AMACR	HPN	SLC25A6
CYP1B1	0ACT2	FASN	NME2	EEF2
ATF5	GDF15	HPN	CBX3	SAT
BRCA1	FASN	UAP1	GDF15	NME2
LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA
MYC	KRT18	0ACT2	MRPL3	CANX
PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA
WT1	GRP58	KRT18	NME1	FASN
TFF3	PPIB	EEF2	COX6C	SND1

Data from DeConde et al. *Stat App Gene Mol Bio* 5(1), 2006.

Current methods

1) Venn diagram
2) Rank aggregation
3) Meta-analysis
4) Desirability functions

Stealing from psychologists

Behav Res (2013) 45:857–872
DOI 10.3758/s13428-012-0300-3

Bayesian Thurstonian models for ranking data using JAGS

Timothy R. Johnson · Kristine M. Kuhn

OPEN ACCESS Freely available online

PLOS | ONE

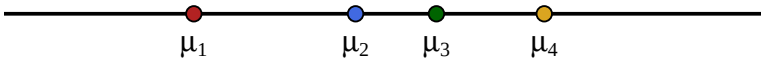
A Cognitive Model for Aggregating People's Rankings

Michael D. Lee*, Mark Steyvers, Brent Miller

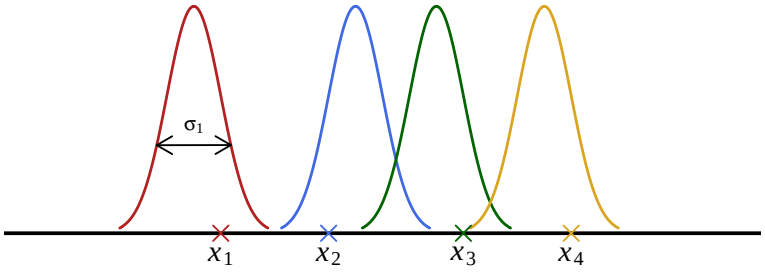
Department of Cognitive Sciences, University of California Irvine, Irvine, California, United States of America

Bayesian rank aggregation

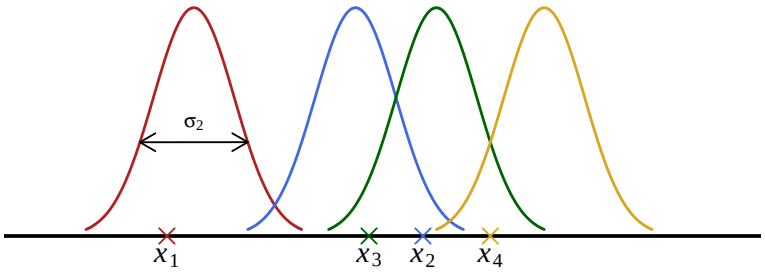
True gene ranking



Experiment 1 $y_1 = (1, 2, 3, 4)$

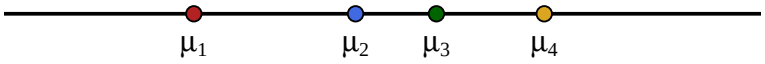


Experiment 2 $y_2 = (1, 3, 2, 4)$



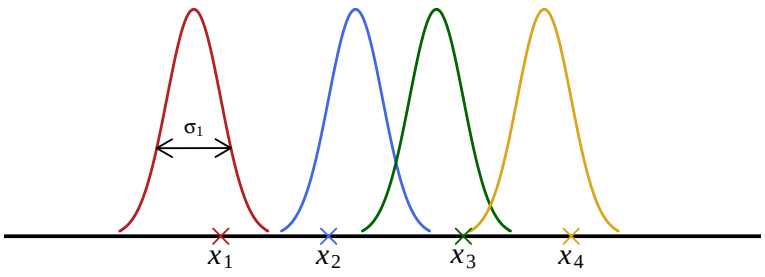
Bayesian rank aggregation

True gene ranking



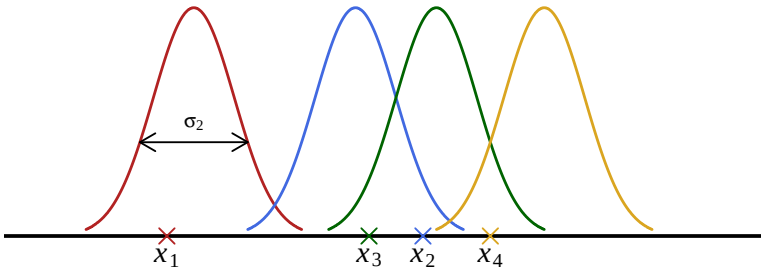
Experiment 1

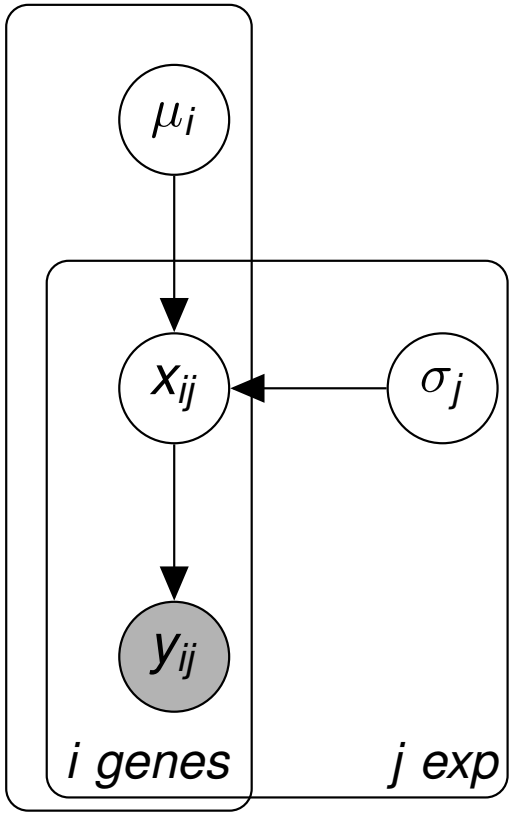
$y_1 = (1, 2, 3, 4)$



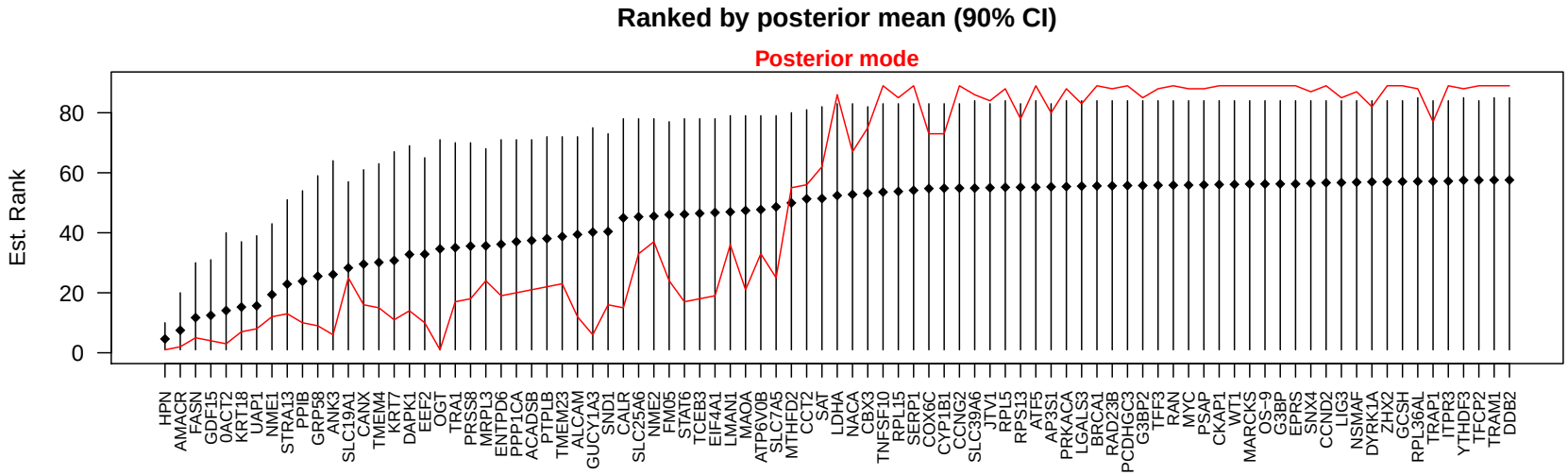
Experiment 2

$y_2 = (1, 3, 2, 4)$

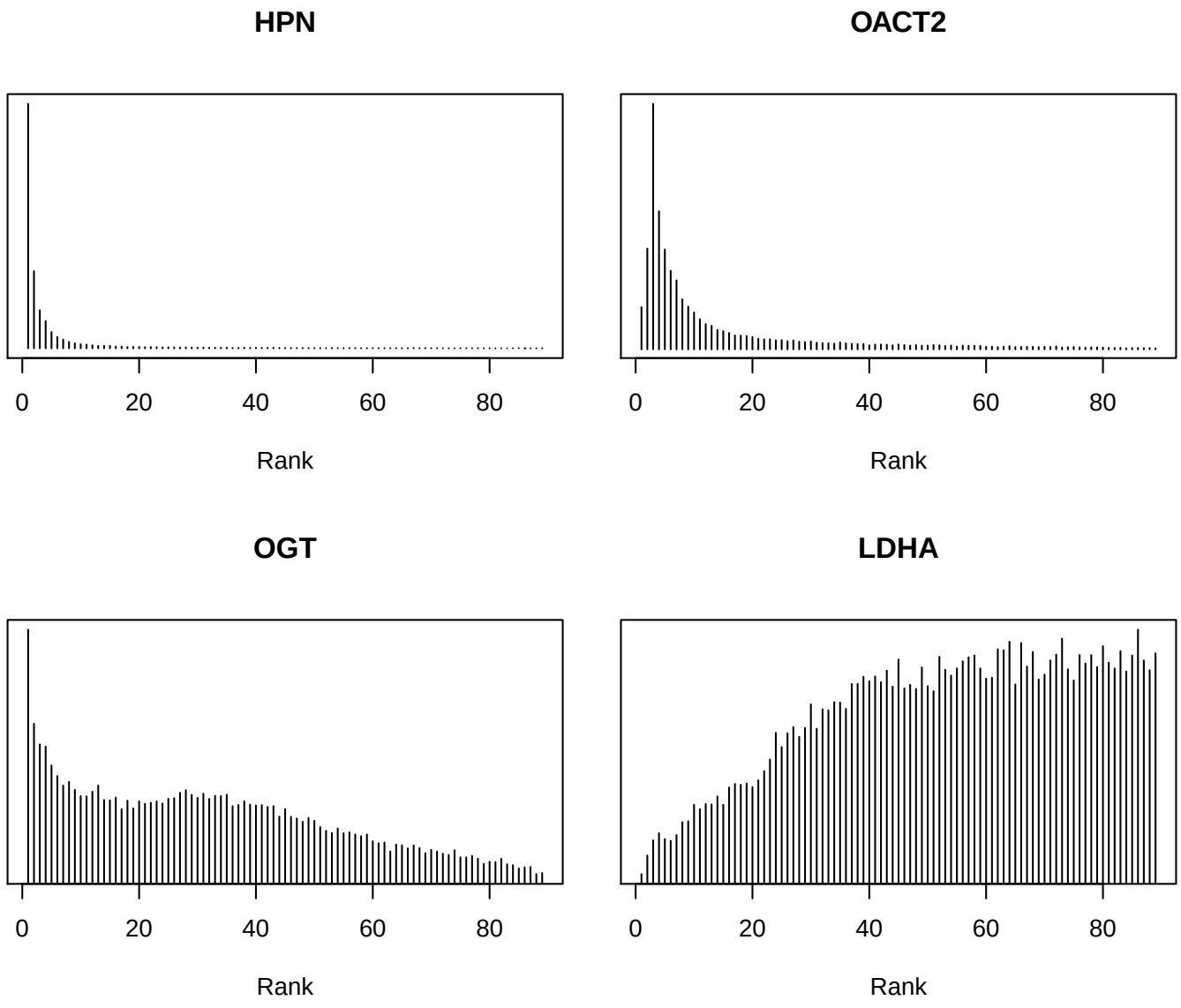




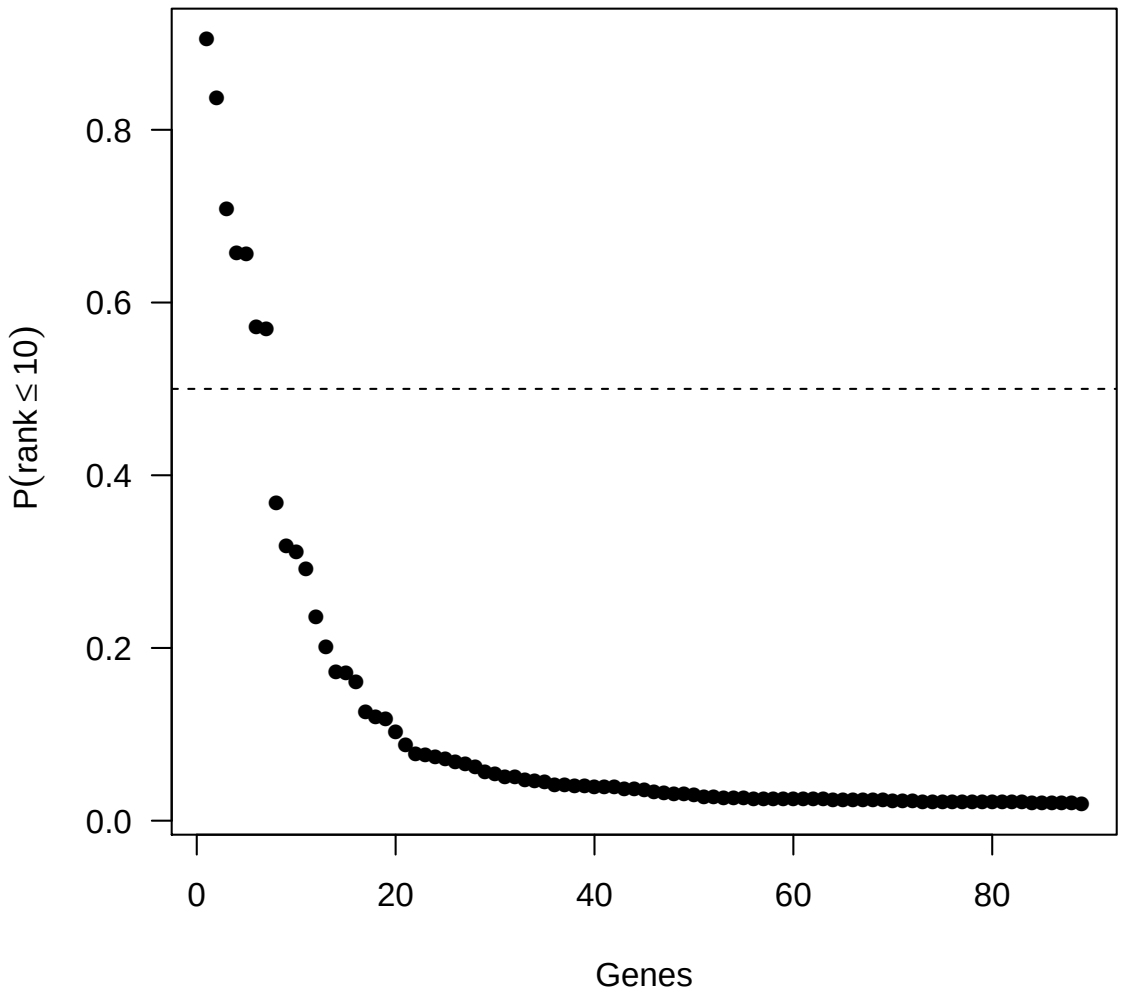
Genes ordered by estimated mean rank



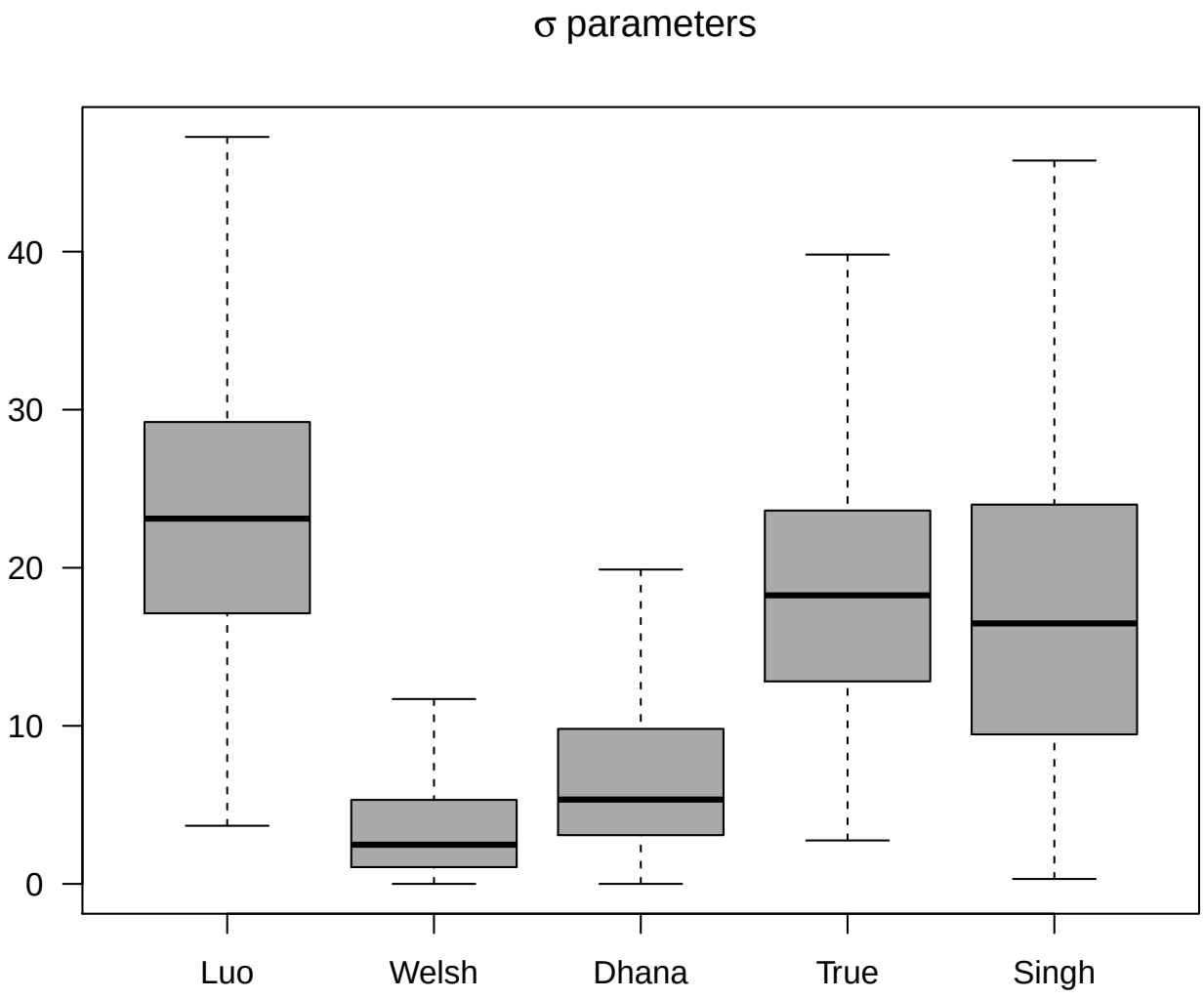
Posterior distribution over ranks



Probability that a gene is in the top 10



Quality/relevance of studies



Conclusions

1) Desirability functions are a fast and intuitive way of selecting genes.

2) Bayesian rank aggregation provides several useful summary statistics for interpreting gene lists.

Acknowledgements

Steffen Renner & Ansgar Schuffenhauer, Novartis
Prof. Michael Lee, UC Irvine

References

1) **Lazic SE** (2015). Ranking, selecting, and prioritising genes with desirability functions. *PeerJ* (in press).
`https://cran.r-project.org/web/packages/desiR/` (stable release)
`https://github.com/stanlazic/desiR` (dev version)