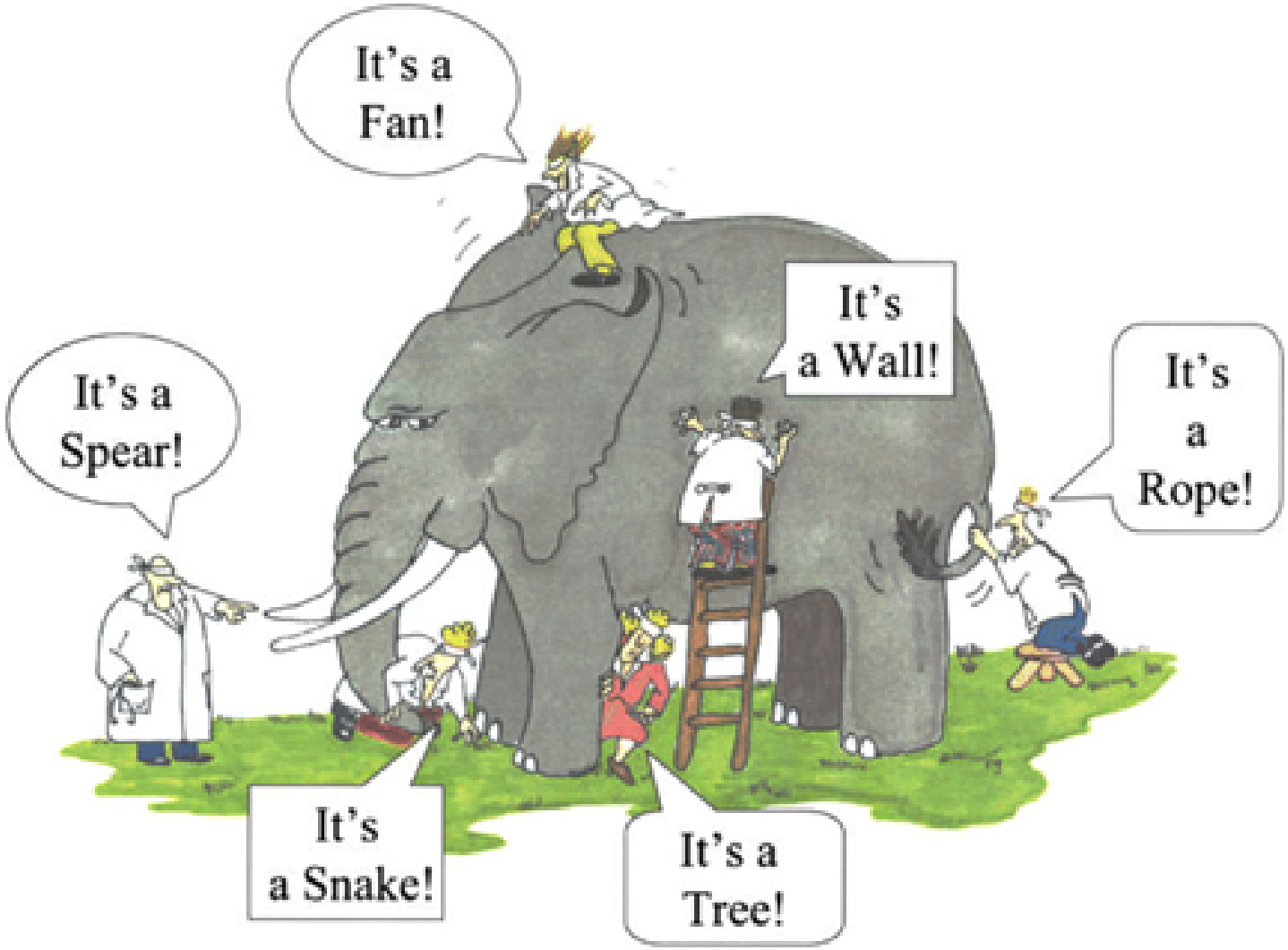
Integrated differential analysis of multi-omics data using a joint mixture model: `idiffomix`

Koyel Majumdar*, Florence Jaffrézic†, Andrea Rau†,
Isobel Claire Gormley*, Thomas Brendan Murphy*

*School of Mathematics and Statistics, University College Dublin, Ireland.
ψINRAE, Université Paris-Saclay, France.

Research Ireland Centre for Research Training in Foundations of Data Science

Why *integrated* differential analysis?



The challenge

- Gene expression and DNA methylation are interconnected biological processes.
- Aim: identification of differentially methylated CpG sites (DMCs) and differentially expressed genes (DEGs) between e.g., healthy and affected samples.
- Typically DMCs and DEGs are identified through independent analyses of methylation and gene expression data; relations between them are subsequently explored.
- Typically DMCs and DEGs detected using t-test/p-valued based approaches e.g., methods such as `limma`[1] state of the art.
- Inherent dependencies and biological structure generally ignored.
- Propose a model-based clustering approach that allows for joint modelling of multiple data sets, incorporation of biological dependencies and simultaneous identification of DMCs and DEGs.

[1] Ritchie et al [2015]

Our proposal: `idiffomix`

- A joint mixture model that integrates information from both data types at the modelling stage, enabling simultaneous identification of DMCs and DEGs.
- Parameter estimation: an expectation-maximisation algorithm.
- Analyse RNA-Seq and DNA methylation array data from matched healthy and breast cancer samples.
- Several non-differential genes, under independent analyses, had high likelihood of being DEGs under the integrated analysis.
- Gene ontology analysis indicated DMCs and DEGs involved in important, cancer related, biological processes and pathways.
- Cross-omics information simultaneously utilised providing comprehensive view.
- An open source R package `idiffomix` is available.

Breast cancer study data

- Analyse RNA-Seq and DNA methylation array data from $N = 5$ matched healthy and breast cancer samples.
- RNA-Seq data: log-fold changes between tumour and benign samples for $G = 15,722$ genes.
- For gene g:

$$\begin{aligned} \boldsymbol{x}_g &= (x_{g1}, \ldots, x_{gn}, \ldots, x_{gN}) \\ \text{where } x_{gn} &= \text{ log-fold change in } g\text{th gene from } n\text{th patient.} \end{aligned}$$

- Methylation data: difference in M-values (= logit transformed beta values) between tumour and benign samples at $C = 94,873$ CpG sites in promoter regions.
- For CpG site c on gene g:

$$\begin{aligned} \boldsymbol{y}_{gc} &= (y_{gc1}, \ldots, y_{gcn}, \ldots, y_{gcN}) \\ \text{where } y_{gcn} &= \text{ difference in } M\text{-values at CpG site } c\text{, on gene } g\text{, patient } n. \end{aligned}$$

Gene expression and methylation data

DEGs...

- Expression levels at gene g assumed to undergo one of $K = 3$ possible state changes between benign and tumour conditions:
  - Downregulated (E-): expression levels decrease (large negative log-fold change) between tumour and benign samples.
  - Upregulated (E+): expression levels increase in tumour sample (large positive log-fold change).
  - Non-differentially expressed (E0): no change (log-fold change ≈ 0).

...and DMCs

- Methylation levels at CpG site c assumed to undergo one of $L = 3$ possible state changes:
  - Hypomethylated (M-): methylation level decreases (large negative differences) between tumour and benign samples.
  - Hypermethylated (M+): methylation increases in tumour sample (large positive differences).
  - Non-differentially methylated (M0): difference in M-values ≈ 0.

A joint mixture model

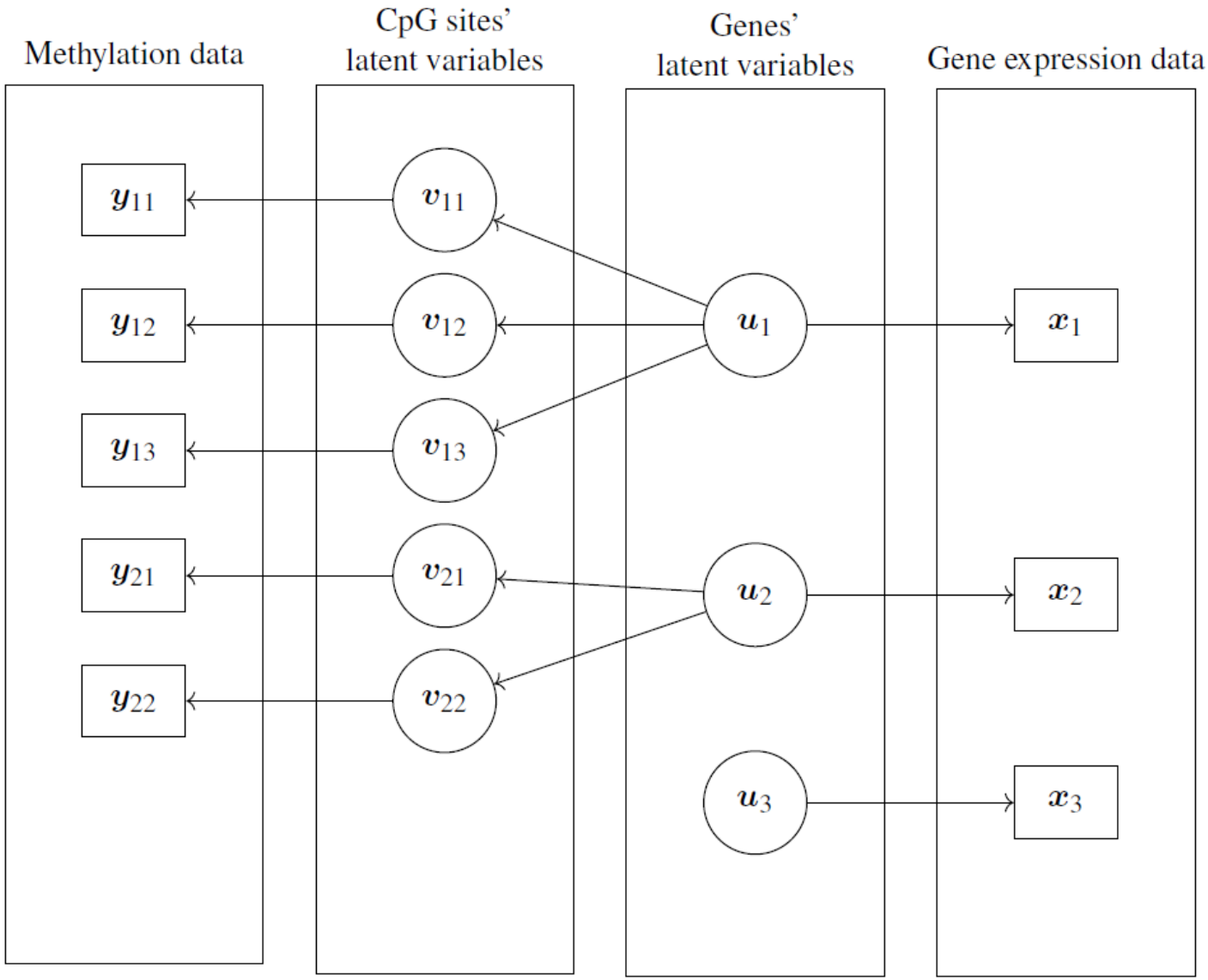
- Mixture model: incomplete data approach employed to facilitate inference.
- Introduce latent variables:

$$u_{gk} = 1 \text{ if gene } g \text{ belongs to cluster } k, \text{ 0 otherwise.}$$

$$v_{gc\ell} = 1 \text{ if CpG site } c, \text{ located in neighbourhood of gene } g, \text{ belongs to cluster } \ell, \text{ 0 otherwise.}$$

- Use these latent variables to account for nested structure, integrating the expression and methylation mixture models together.

The `idiffomix` joint mixture model



The `idiffomix` joint mixture model

- Within each component, log-fold change data assumed to be i.i.d Gaussian:

$$x_{gn}|(u_{gk} = 1) \sim N(\mu_k, \sigma_k^2)$$

- Differences in *M*-values also assumed to be i.i.d. Gaussian within a component:

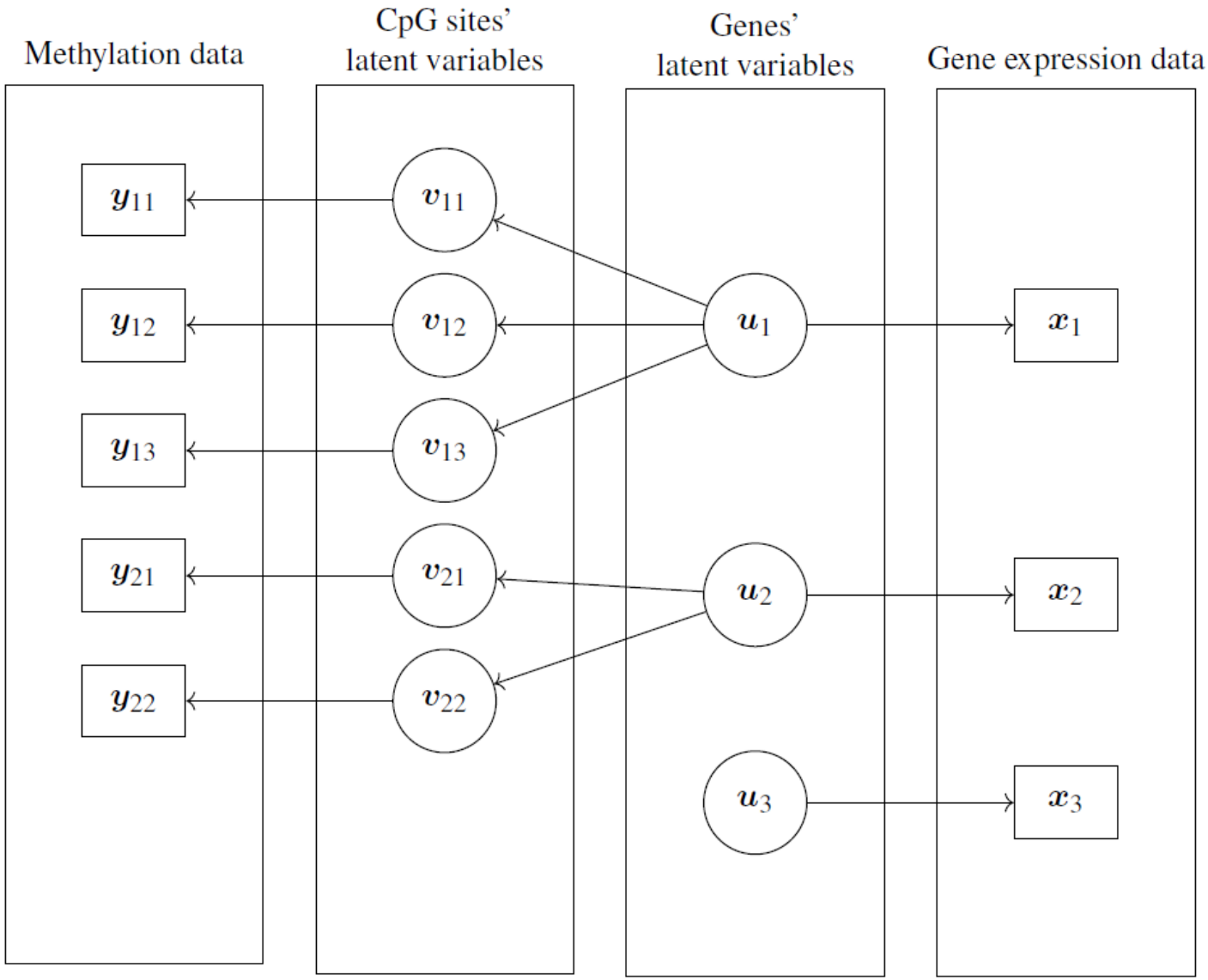
$$y_{gcn}|(v_{gcl} = 1) \sim N(\lambda_l, \rho_l^2)$$

- Proportion of genes in each cluster: $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$.

- Dependencies between genes and CpG sites accounted for through $L \times K$ matrix parameter $\boldsymbol{\pi}$.

$$\pi_{l|k} \quad = \quad \text{probability of a CpG site belonging to cluster } l, \text{ given its associated associated gene} \in k.$$

The idiffomix joint mixture model



The idiffomix joint mixture model

$$P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{V} | \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{g=1}^{G} \left\{ \prod_{k=1}^{K} P(\boldsymbol{x}_g | \boldsymbol{\theta}_k)^{u_{gk}} \prod_{c=1}^{C_g} \prod_{l=1}^{L} P(\boldsymbol{y}_{gc} | \boldsymbol{\phi}_l)^{v_{gcl}} \right\}$$
$$\times \prod_{g=1}^{G} \prod_{k=1}^{K} \left\{ \tau_k \prod_{c=1}^{C_g} \prod_{l=1}^{L} \pi_{l|k}{}^{v_{gcl}} \right\}^{u_{gk}}$$

- If $\pi_{l|k} = \pi_{l|k'}$ for all $k, k' \Rightarrow$ status of CpG sites and genes are independent $\Rightarrow$ model is equivalent to two independent mixture models.
- Inference proceeds via EM algorithm.
- Due to independence of chromosomes and to ease the computational burden, model fitted to each chromosome independently in parallel.
- Initialisation: quantile based approach to specify cluster memberships.
- Convergence: absolute change in all parameter estimates between successive iterations $< 1 \times 10^{-5}$.

idiffomix: inference

- E-step: required expected values of the latent variables are intractable.
- Tractable approximation via computing conditional expected value of latent variable given the others[2] at E-step.
- Iteratively computed until convergence:

$$\begin{aligned}
\mathbb{E}(u_{gk}|\cdots) &\approx u_{gk}^{(S)} = \hat{u}_{gk} \\
\mathbb{E}(v_{gcl}|\cdots) &\approx v_{gcl}^{(S)} = \hat{v}_{gcl} \\
\mathbb{E}(u_{gk}v_{gcl}|\cdots) &\approx u_{gk}^{(S)} v_{gcl}^{(S)} = \widehat{u_{gk}v_{gcl}}.
\end{aligned}$$

- In practice, $S \approx 10$ required to achieve convergence per EM iteration.

[2] Salter-Townshend and Murphy [2013], Chamroukhi and Huynh [2018]

idiffomix: inference

- M-step: the expected complete data log-likelihood function is maximised with respect to the model parameters $\boldsymbol{\tau}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ $\Rightarrow$ closed form solutions.
- On convergence, for each gene and CpG site:
  latent variable estimates = posterior probabilities of cluster membership.
- Cluster assignment performed using the maximum *a posteriori* (MAP) rule:
  - DEGs: genes in clusters E- and E+
  - DMCs: CpGs in clusters M- and M+

Simulation study: set up

- Simulated data that mirrored the breast cancer data settings.
- Considered three settings of π.
- Values represent probabilities of a CpG site belonging to cluster M+, M0 or M-, conditional of their associated gene belonging to cluster E-, E0 or E+.

(a) Case 1: à la breast cancer data

	E-	E0	E+
M+	0.4	0.05	0.1
M0	0.5	0.9	0.5
M-	0.1	0.05	0.4

(b) Case 2: high level of dependency

	E-	E0	E+
M+	0.8	0.1	0.1
M0	0.1	0.8	0.1
M-	0.1	0.1	0.8

(c) Case 3: independence between datasets

	E-	E0	E+
M+	0.2	0.6	0.2
M0	0.2	0.6	0.2
M-	0.2	0.6	0.2

Simulation study: results

- Mean performance metrics for 100 simulated datasets given π under case 1.

(a) DEG identification performance

	FDR	Sensitivity	Specificity	ARI
`idiffomix`	**0.014** (0.011)	**0.976** (0.015)	**0.997** (0.003)	**0.966** (0.017)
`mclust`	0.102 (0.049)	0.873 (0.046)	0.975 (0.015)	0.800 (0.041)
`limma`	0.038 (0.021)	0.764 (0.064)	0.993 (0.005)	0.760 (0.059)

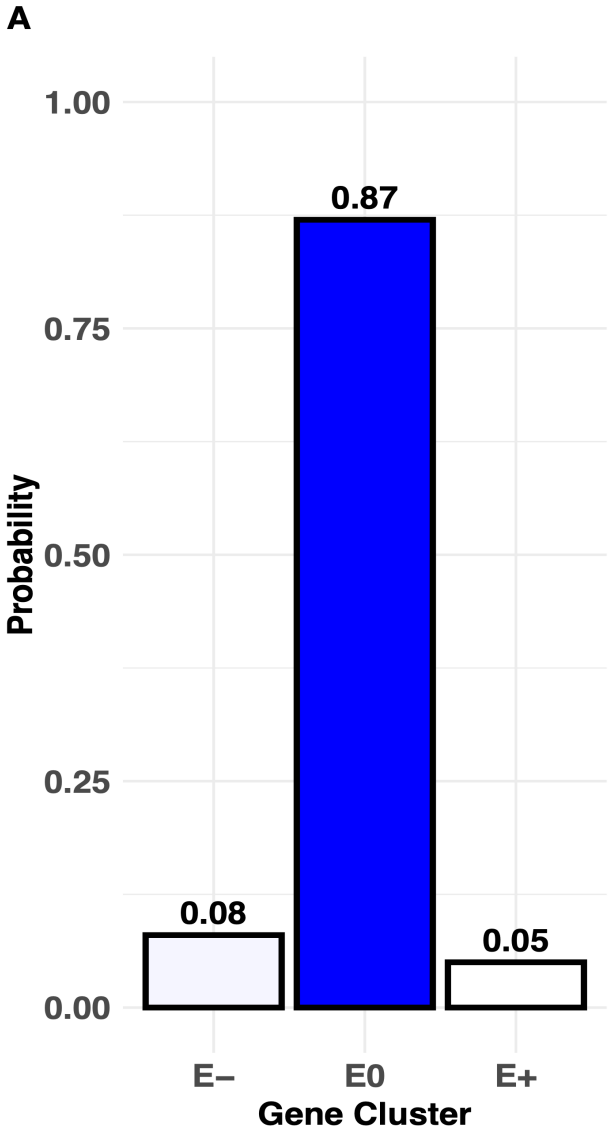
(b) DMC identification performance

	FDR	Sensitivity	Specificity	ARI
`idiffomix`	**0.016** (0.005)	0.999 (0.001)	**0.997** (0.001)	**0.986** (0.004)
`mclust`	0.019 (0.006)	0.999 (0.001)	0.996 (0.001)	0.983 (0.005)
`limma`	0.058 (0.006)	**1.000** ($<$0.001)	0.987 (0.002)	0.948 (0.006)

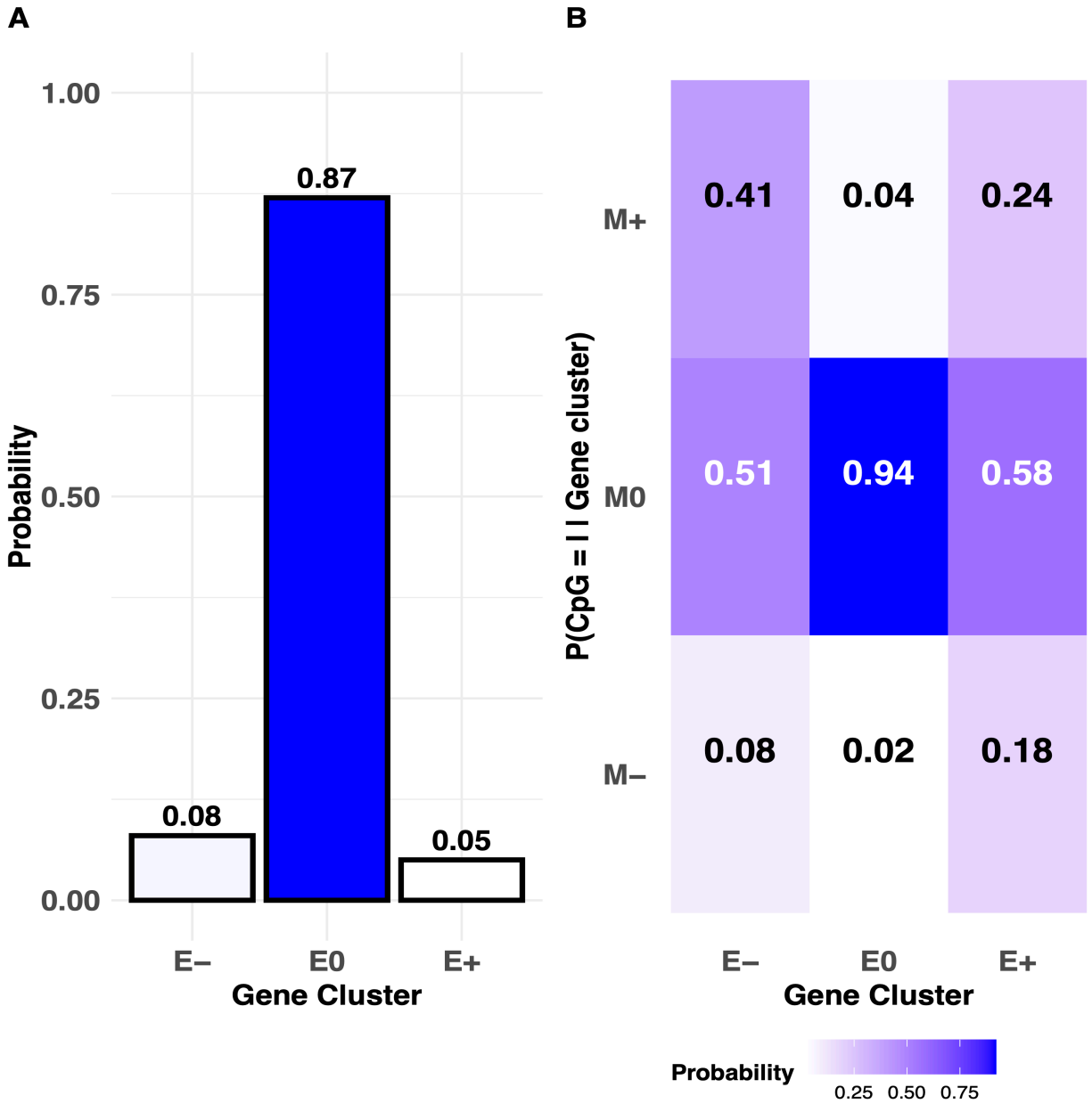
*Standard deviations in parentheses and the top performing method for each metric highlighted in boldface.

Application on breast cancer data

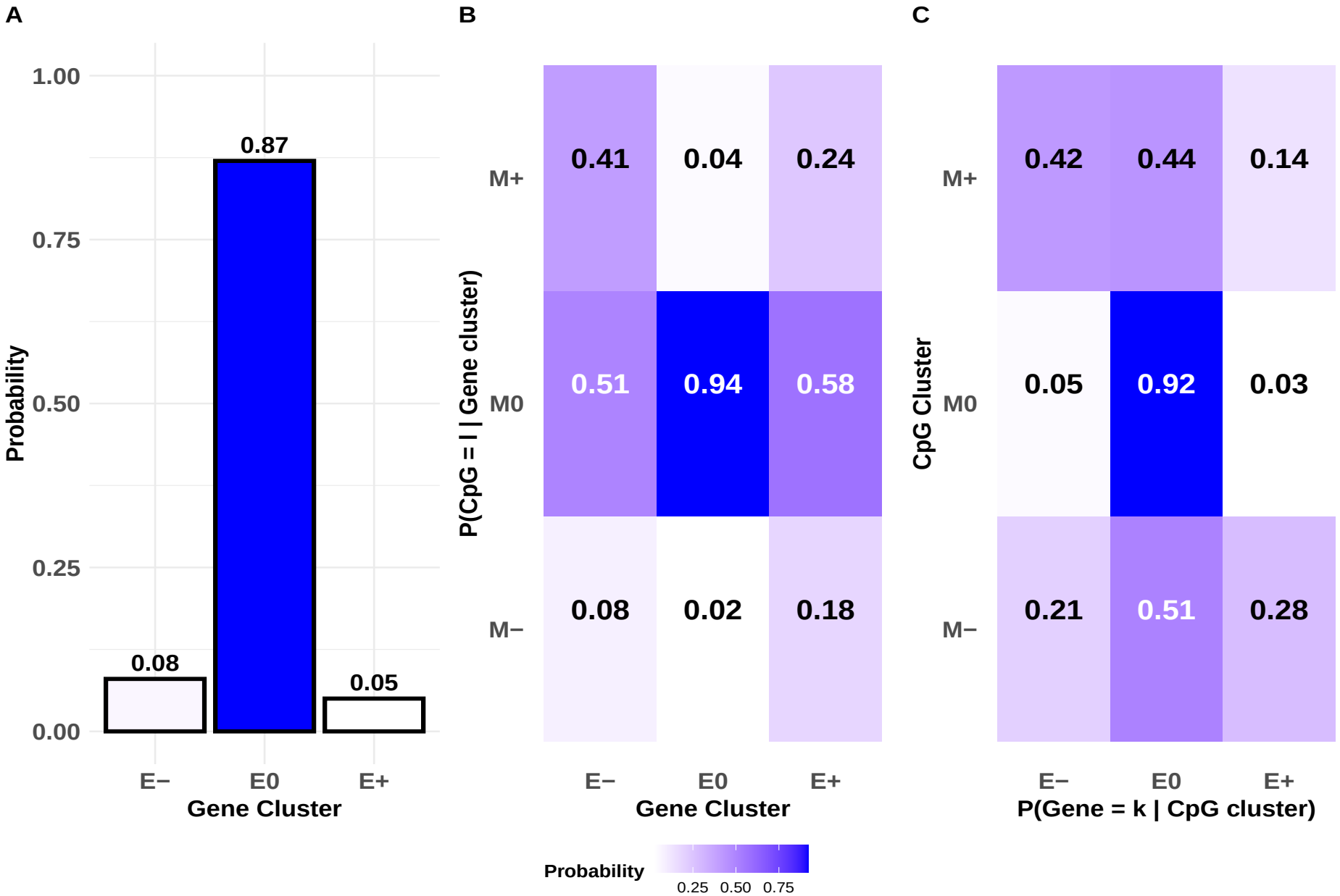
- Matched healthy and tumour tissue from $N = 5$ patients, RNA-seq ($\approx$ 15k genes) + methylation array ($\approx$ 94k CpG sites).



Application on breast cancer data

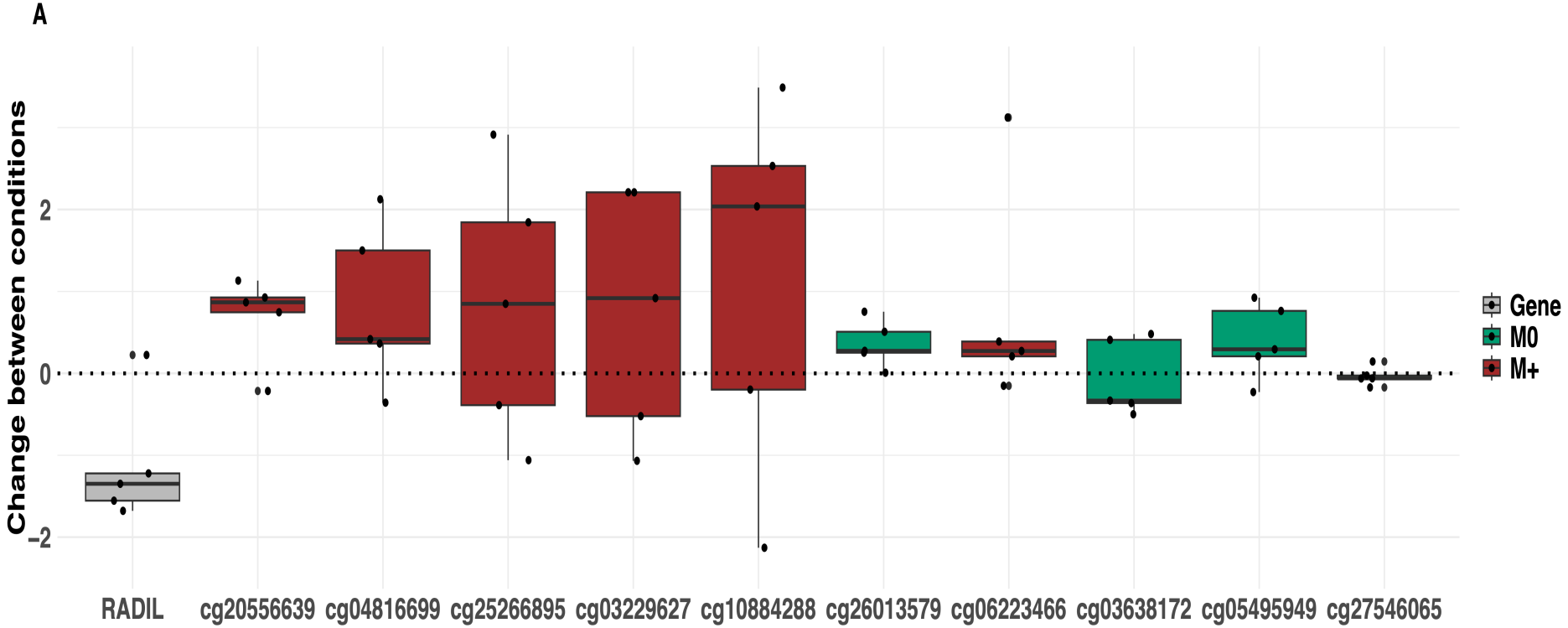


Application on breast cancer data

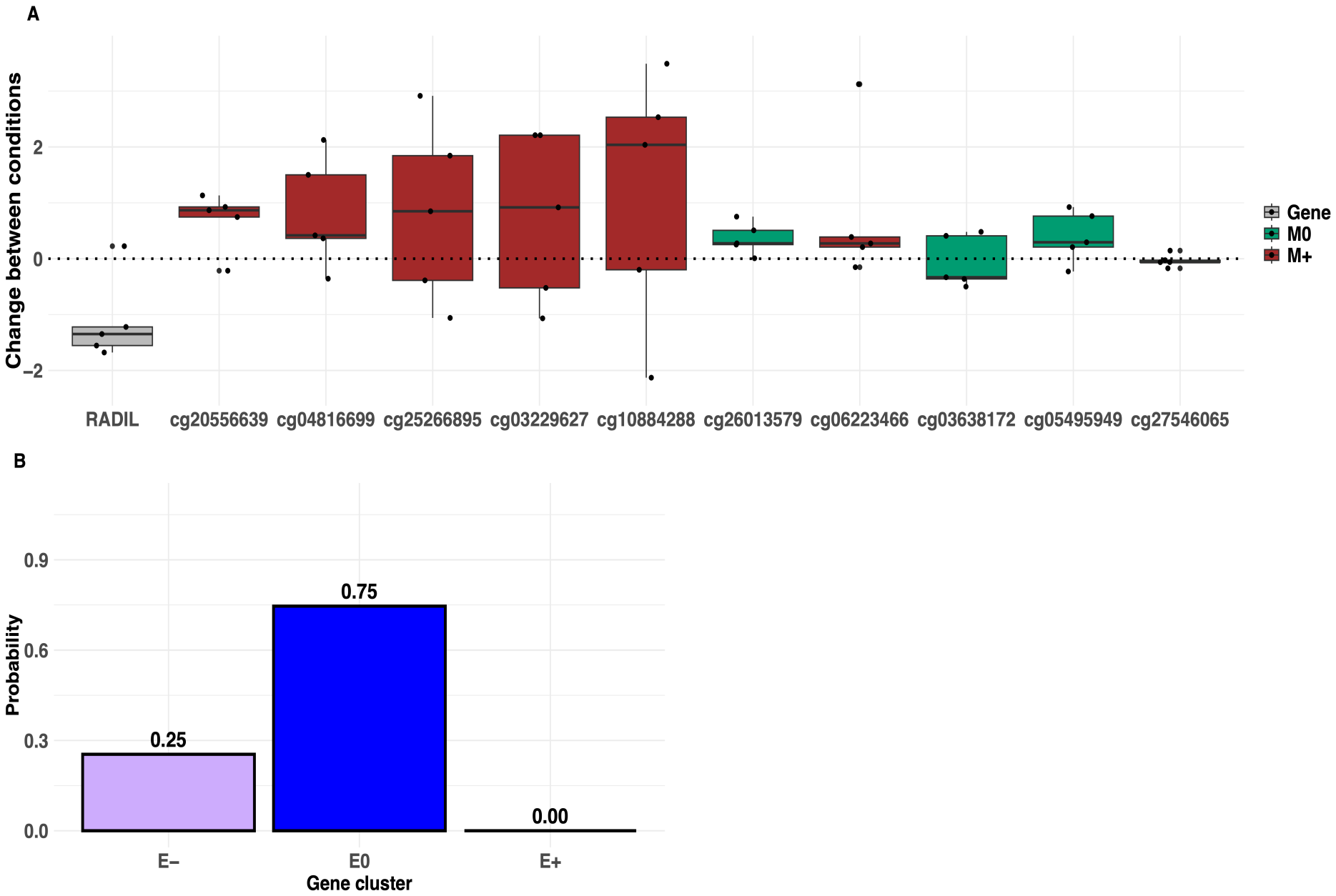


Genes of interest

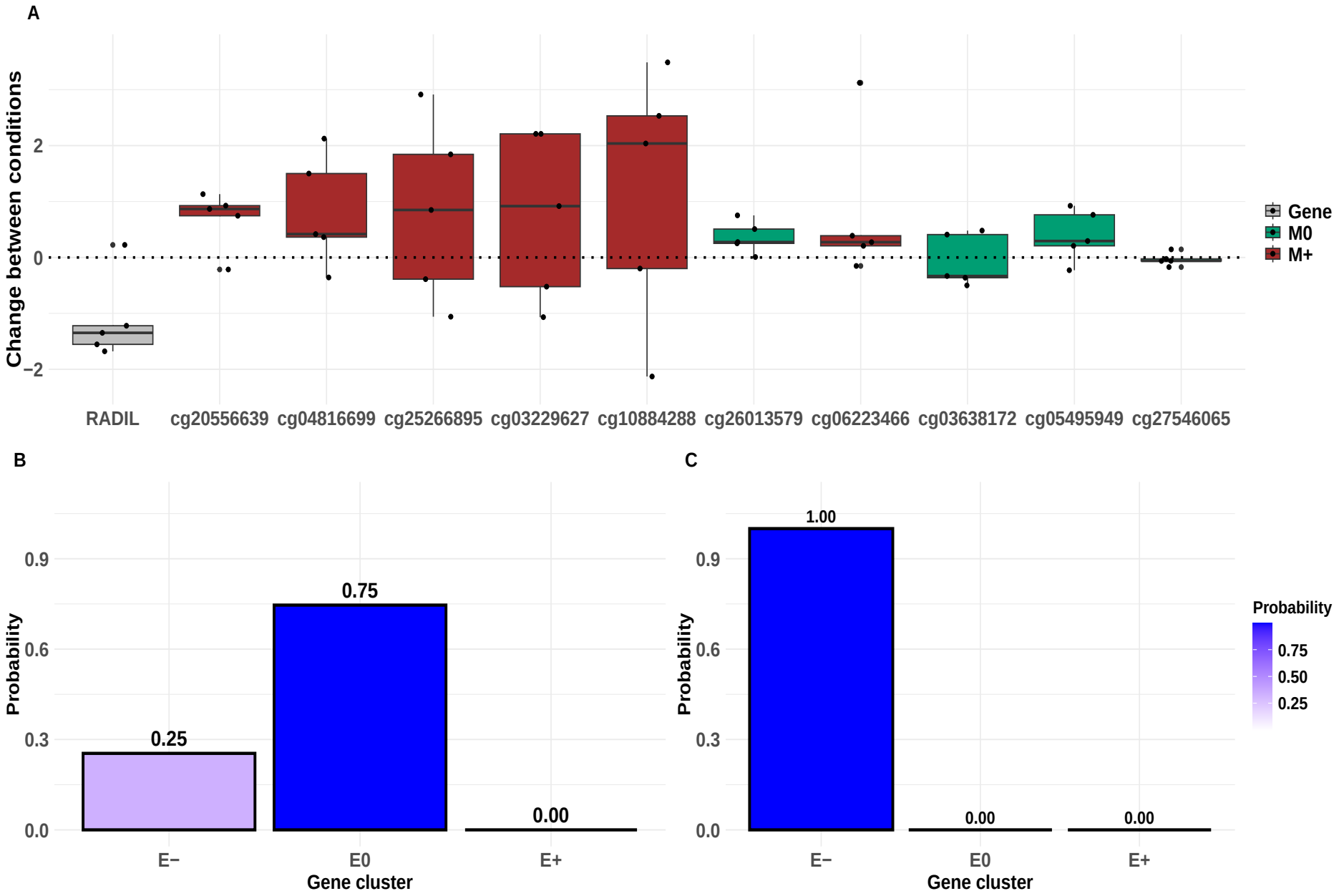
- Genes for which differential status differed between independent and integrated analyses of interest e.g., *RADIL* gene.



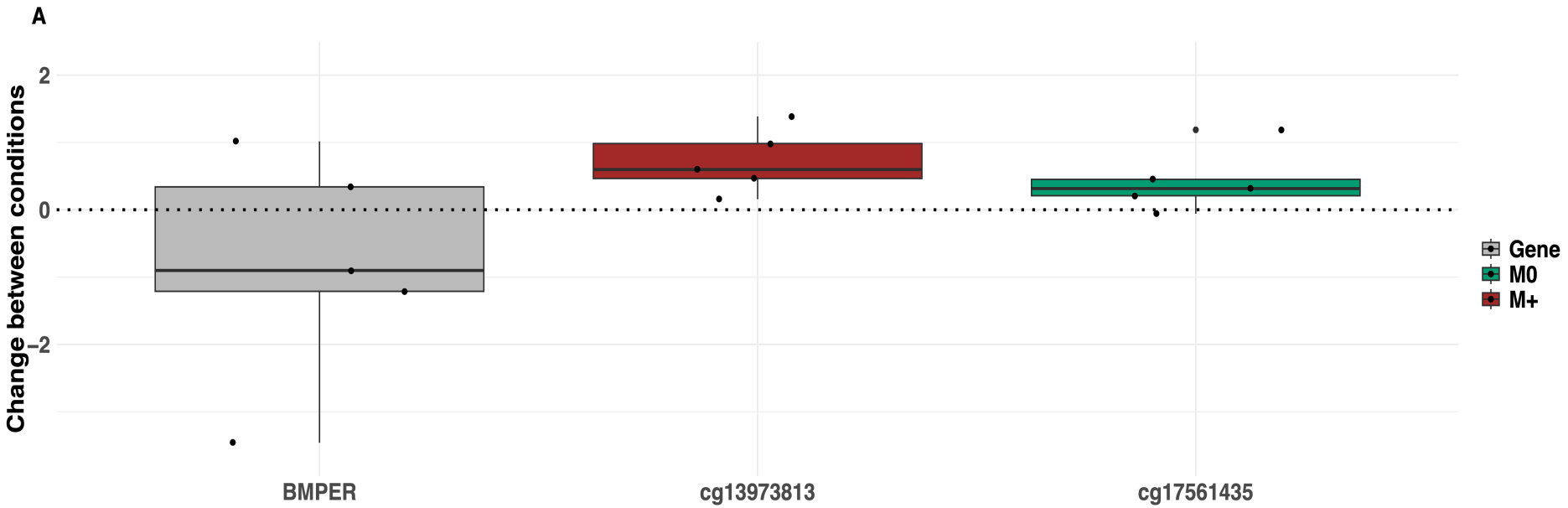
Gene of interest: *RADIL*



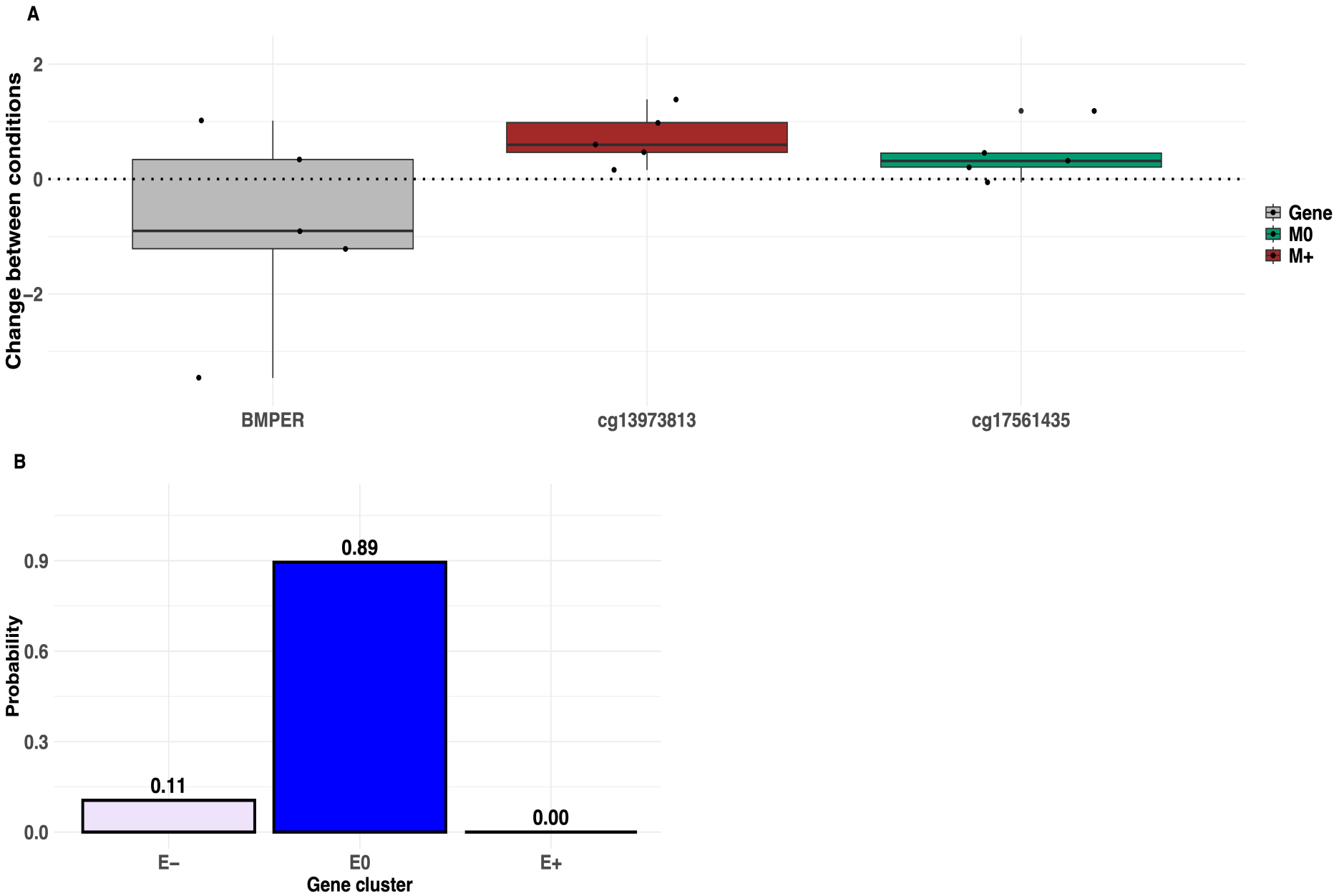
Gene of interest: *RADIL*



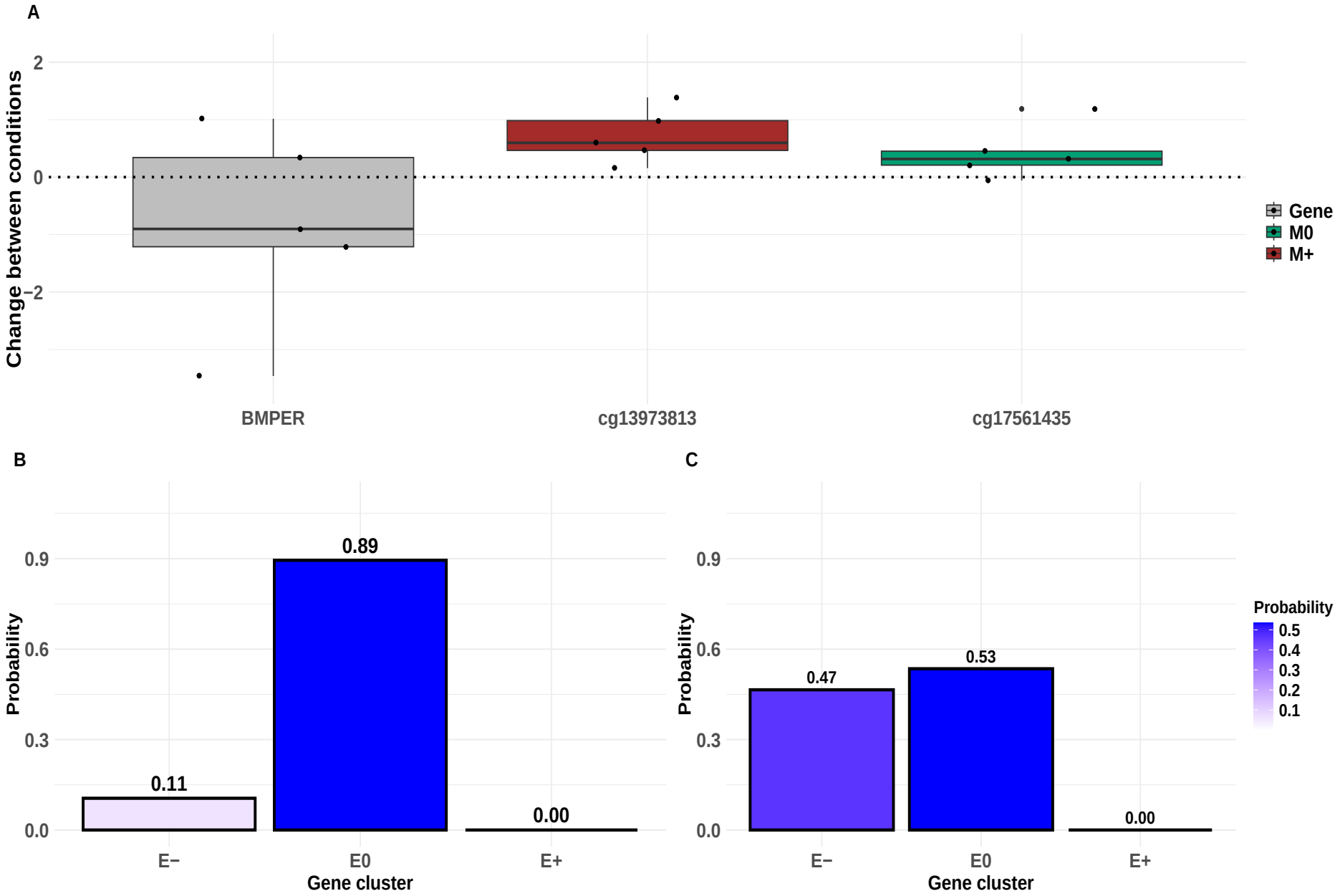
Clustering uncertainty: *BMPER*



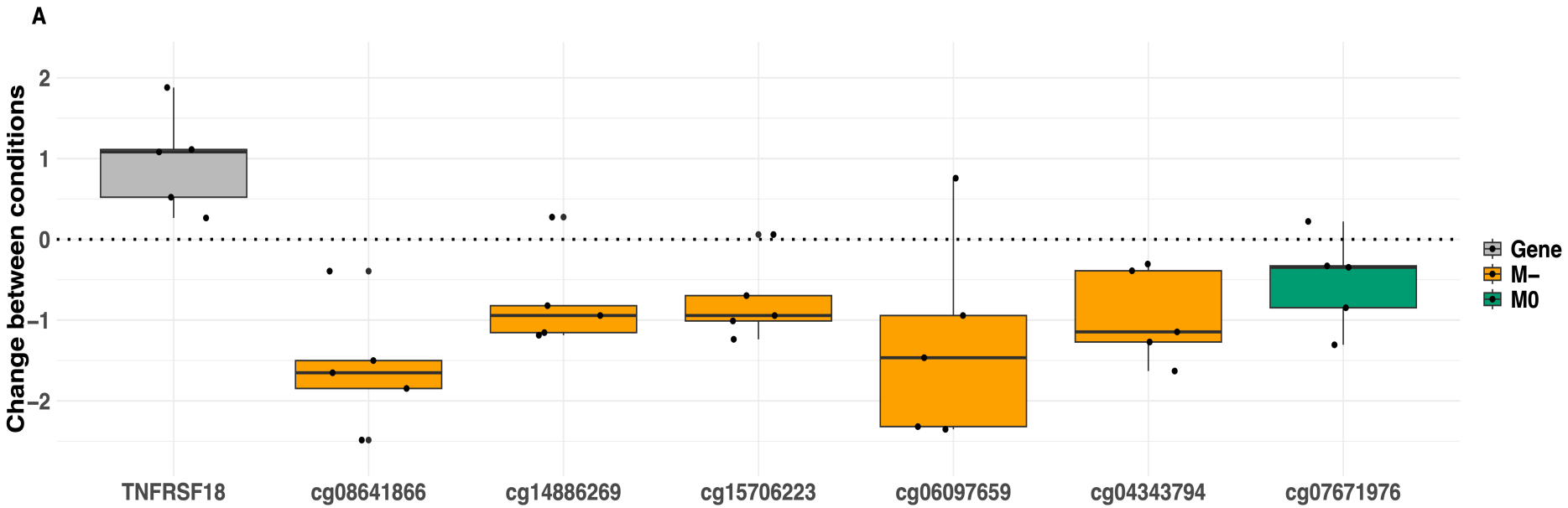
Clustering uncertainty: *BMPER*



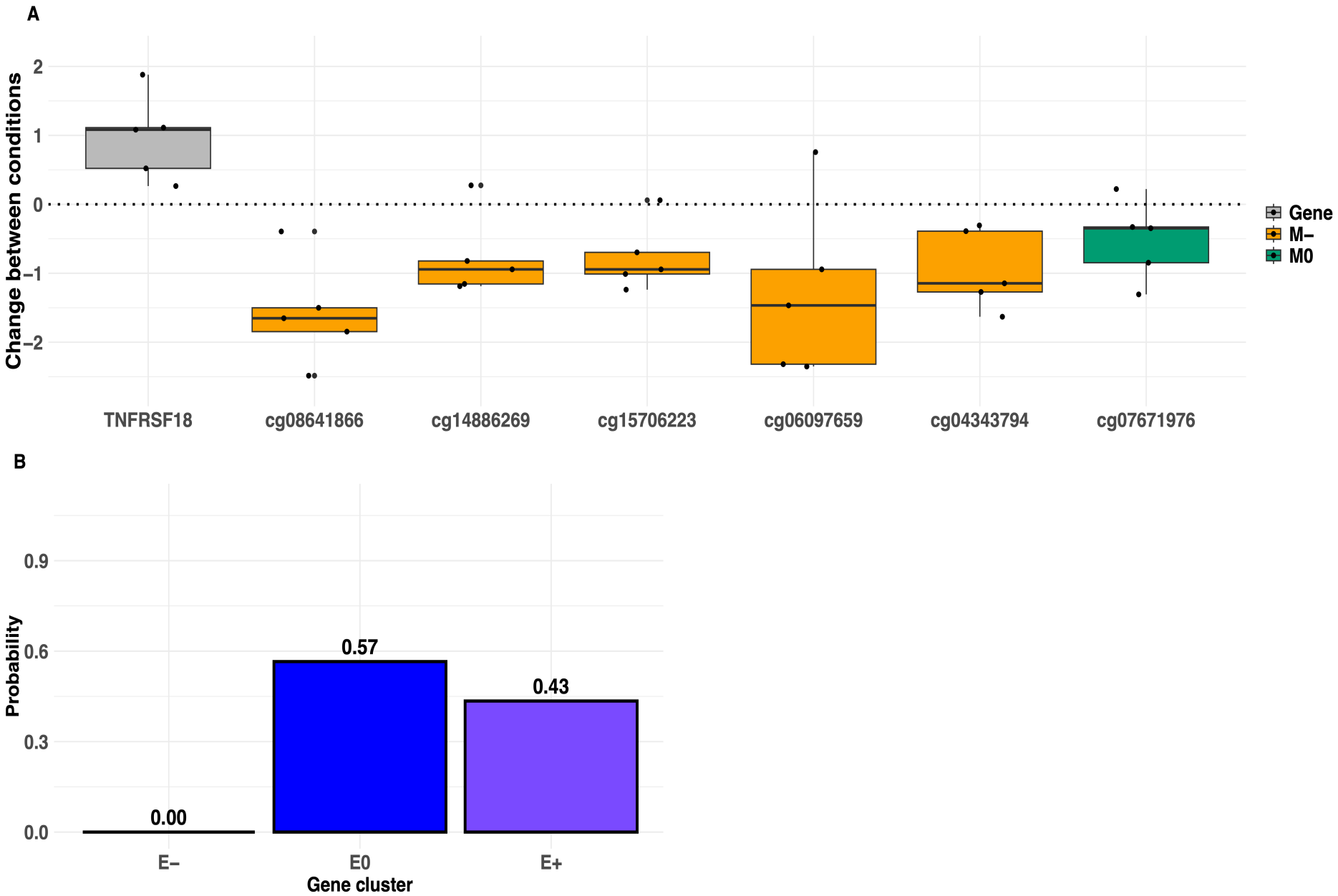
Clustering uncertainty: *BMPER*



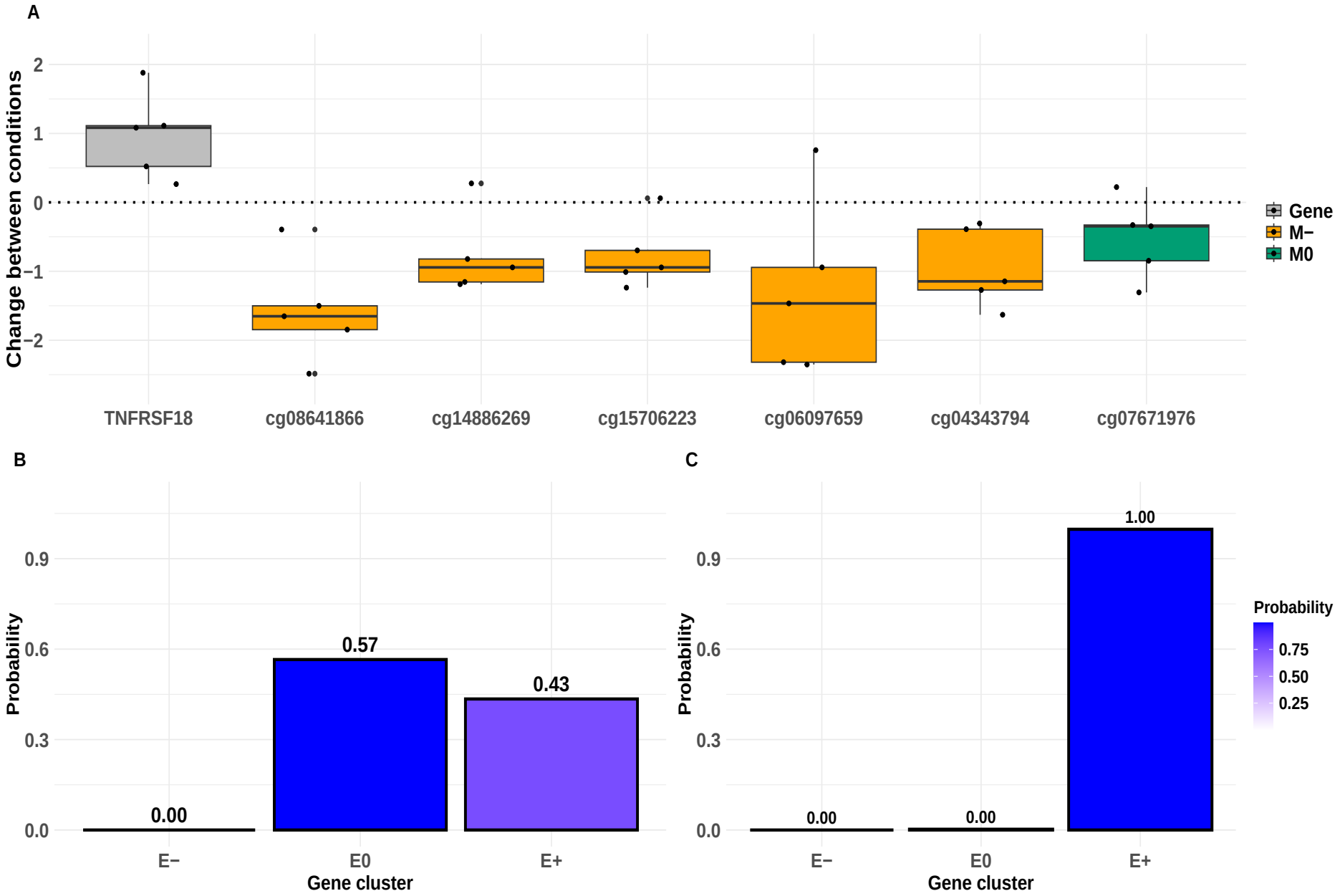
TNFRSF18: role in development & progression of breast cancer



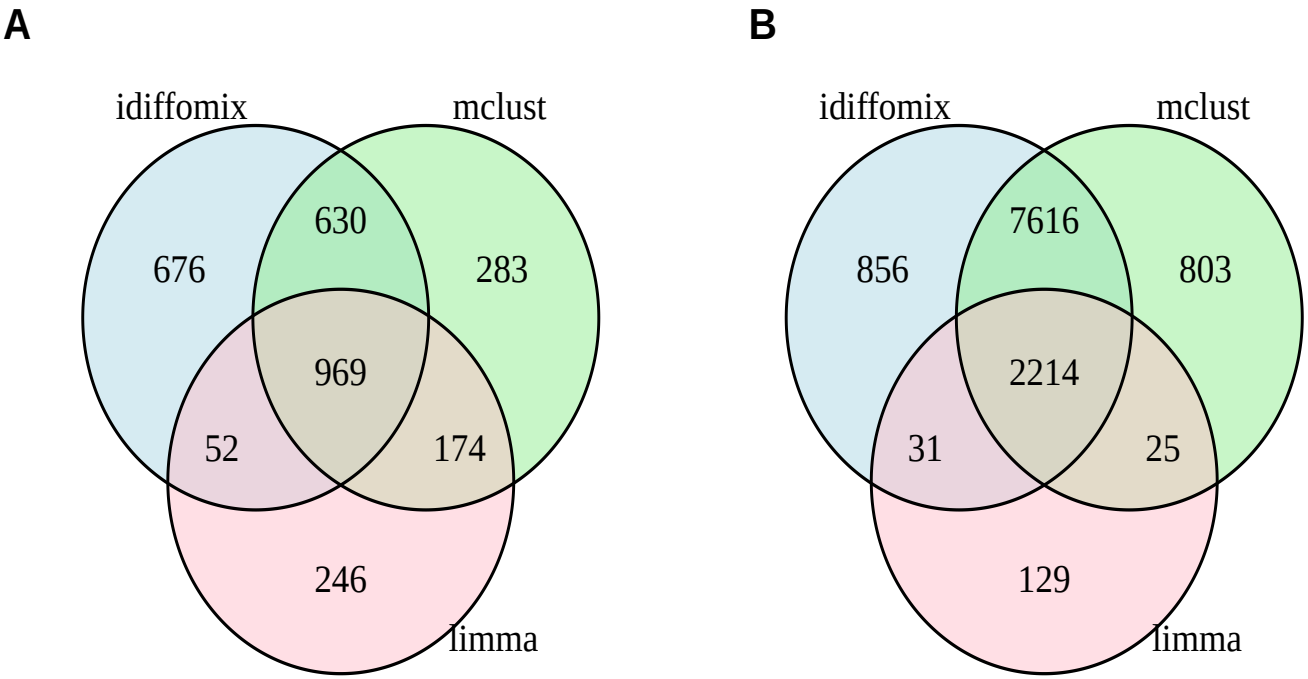
TNFRSF18: role in development & progression of breast cancer



TNFRSF18: role in development & progression of breast cancer



Independent v integrated results



(A) DEGs

(B) DMCs

- Gene enrichment analysis:
  some biological processes and pathways which play essential role in breast cancer development and prognosis identified only under `idiffomix` approach.

`idiffomix`: take-home messages...

- When identifying differential expression and methylation, should account for inherent biological dependencies between gene sequencing and methylation data.
- Take a model-based clustering approach to identify DEGs and DMCs.
- Proposed a joint mixture model that integrates both data types at the modelling stage by directly modelling their nested structure.
- Allows for a genome-wide, cross-omics analysis that simultaneously identifies DMCs and DEGs.
- Simulation studies and application to breast cancer data demonstrated utility.
- General framework: could be generalized to other experimental designs or other omics data.
- `idiffomix` R package available.

idiffomix: ...but!

- Modelling log-fold changes and differences in M-values makes results less biologically interpretable: model the inherent data distributions directly.
- Cases where healthy and diseased tissues do not come from the same subjects, or when sample sizes differ between conditions require model changes?
- Integrate other data? E.g., proteomics + methylation + RNA-Seq?
- Spatial information also available: locations of CpG sites known and could be incorporated (and same for genes).
- Methylation patterns and gene expression regulation also dependent on other factors e.g., environmental stress, food habits: include as covariates.

Bibliography

- Majumdar, K. et al. (2025+)
  Integrated differential analysis of multi-omics data using a joint mixture model: idiffomix.
  Under review
  R package: `idiffomix`

- Majumdar, K. et al. (2024)
  A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer.
  PLOS One
  R package: `betaclust`

- Majumdar, K. et al. (2025+)
  `betaHMM`: a hidden Markov model to identify differentially methylated sites and regions from beta-valued DNA methylation data.
  Under submission
  Bioconductor package: `betaHMM`

Thank you!