# True discovery guarantees in differential gene expression analysis

SMPGD 2026 – Grenoble, 30 January 2026

A. Vesely, L. Finos, J.J. Goeman

University of Bologna
anna.vesely2@unibo.it

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

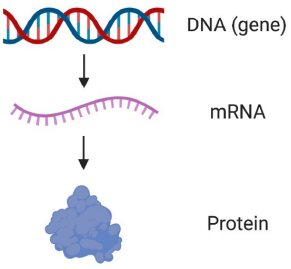
# Overview

# Differential gene expression analysis

Studies gene expression
(quantity of gene product, often a protein)
to infer differences between two populations

**Genes $i \in M = \{1, \ldots, m\}$**
About 20,000 highly correlated units
$H_i$ : *no difference in gene $i$*

**Pathways $S \subseteq M$**
Collections of genes associated with a process
$H_S = \bigcap_{i \in S} H_i$ : *no differences in subset $S$*



# Breast invasive carcinoma dataset

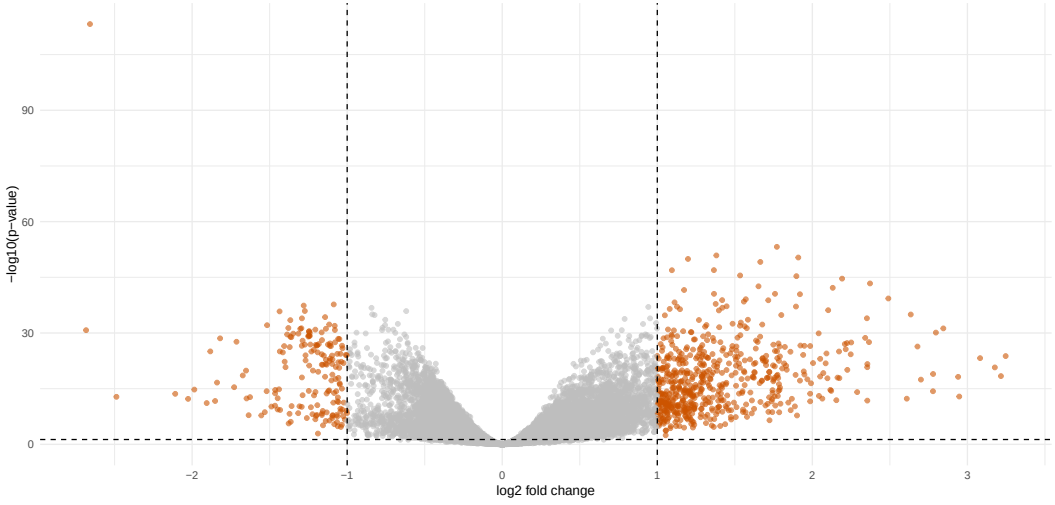
Assess differences between two histological types of primary tumor
According to the origin: ductal vs lobular[1]

$H_i$ : gene $i$ is not DE $\qquad$ $p_i$ = p-value from two-sample t-test

- 985 subjects
- 15,678 genes
- 352 KEGG pathways[2]

[1] The Cancer Genome Atlas (TCGA) research network, `https://www.cancer.gov/tcga`
[2] Kanehisa and Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 2000

# Breast invasive carcinoma dataset



# How to control for multiple comparisons?

**Gene-level analysis**

- FWER: very stringent $\rightarrow$ zero or very discoveries
- FDR: controls the expected value of the FDP, but it can be highly variable

**Group-level analysis**

- aggregates gene-level signal to pathways/gene sets
- groups are often more of interest
- can detect differences even if no single gene is individually significant
- more robust to noisy measurements

# A global test for a pre-fixed group?

**Spatial specificity paradox**

$H_S$ is rejected $\Longrightarrow$ $S$ contains at least one DE gene

No information on:

- the proportion of DE genes (TDP)
- their location in $S$

$\rightarrow$ The larger the pathway, the weaker the finding

**Another multiple testing problem**

- Multiple sets
- Possibly chosen post-hoc (e.g., top genes)

# Objectives

- Adapt to the unknown correlation structure of the data
- Make inference on the TDP within subsets
- Allow for post-hoc selection and follow-up inference

**Focus on sum tests**

Allow meaningful statements at the level of subsets $S$, combining features' signals through sums

**Methods**

- Permutation testing for unknown correlation
- Closed testing for post-hoc validity

# Permutation tests

Use a group of transformations of the data
(here, permutations of the group labels)

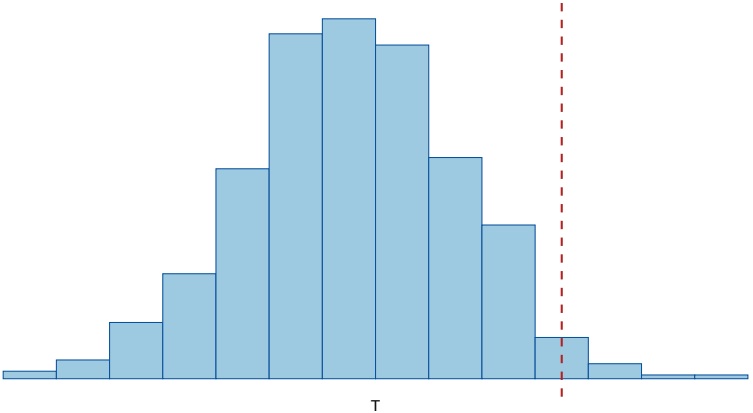
**Assumption: exchangeability**
The joint distribution of p-values of true null hypotheses is invariant under all transformations

**Random permutations**
Use only $B$ elements drawn randomly from the group

- $B \geq 1/\alpha$
- include the identity

# Permutation tests

Reject $H_S$ when $T_S^{\mathrm{id}} > T_S^{(\lceil (1-\alpha)B \rceil)}$, where $T_S^{(1)} \leq \ldots \leq T_S^{(B)}$



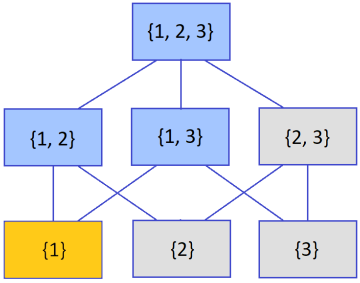
# Closed testing

Closed testing[1] is the optimal way to construct FWER, TDP and related multiple testing procedures[2]

Closed testing rejects $H_S$

$\Longleftrightarrow$

an $\alpha$-level test rejects all $H_V$ with $S \subseteq V \subseteq M$




[1] Marcus et al. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 1976

[2] Goeman et al. Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.*, 2021

# True discovery guarantee

Combining permutation testing and closed testing[1][2] gives simultaneous lower $(1-\alpha)$-confidence bounds for the number of true discoveries $\delta(S)$:

$$P\left(\delta(S) \geq d(S) \text{ for each } S \subseteq M\right) \geq 1-\alpha$$

In particular

$$d(S) = |S| - \max\{|V \cap S| \,:\, V \subseteq M,\ V \text{ not rejected}\}$$

Complexity: $2^m$ operations

[1] Genovese and Wasserman. Exceedance control of the false discovery proportion *JASA*, 2006

[2] Goeman and Solari. Multiple testing for exploratory research *Stat. Sci.*, 2011

# True discovery guarantee by sum tests

# Sum tests

$T_i$ = generic statistic for $H_i$

$$T_S = \sum_{i \in S} T_i \qquad \text{or more generally} \qquad T_S = g\left(\sum_{i \in S} f_i(T_i)\right)$$

**p-value combinations**

- Fisher: $T_i = -2\log(p_i)$
- Cauchy: $T_i = \tan\{(0.5 - p_i)\pi\}$
- generalized means: $T_i = p_i^r$
- ...

**other statistics**

- sum of t-statistics
- sequence kernel association test
- Goeman's global test
- ...

## Single-step shortcut

For any $z \in \{0, \ldots, s-1\}$, it evaluates whether $d(S) > z$:

- yes
- no
- unsure

It studies the collection

$$\mathcal{V}_{|S|-z} = \{V \subseteq M : |V \cap S| \geq |S| - z\}$$

Worst-case complexity: $m \log(m)$ operations

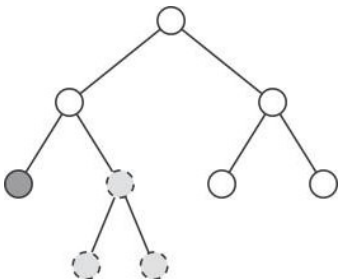
Within a binary search, it approximates closed testing with $d^{(0)}(S) \leq d(S)$

# Iterative shortcut

The single-step shortcut is embedded into a depth-first Branch and Bound search

Improved approximation with
$d^{(0)} \leq d^{(n)}(S) \leq d(S)$

It converges after at most $2^m$ iterations, but may be stopped at any time with valid confidence bounds



# Truncation statistics

$$T_S = \sum_{i \in S} f_i(T_i), \qquad f_i(T_i) = \begin{cases} T_i \text{ if } T_i \geq \tau \\ \gamma \text{ otherwise} \end{cases}$$

$i \notin S$ is such that $f(T_i^\pi) = \gamma$ for all $\pi \neq \mathrm{id}$
$\implies$ $i$ is removed from $M$

$j, k \notin S$ are such that $f(T_j) = f(T_k) = \gamma$
$\implies$ $j$ and $k$ are collapsed into a new index $h$, so that $H_h = H_{\{j,k\}}$

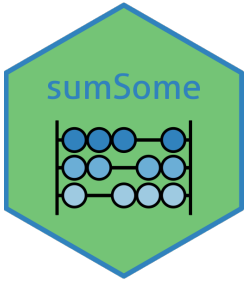
# Results

# sumSome package

The method is implemented in the R package sumSome[1], with underlying code in C++

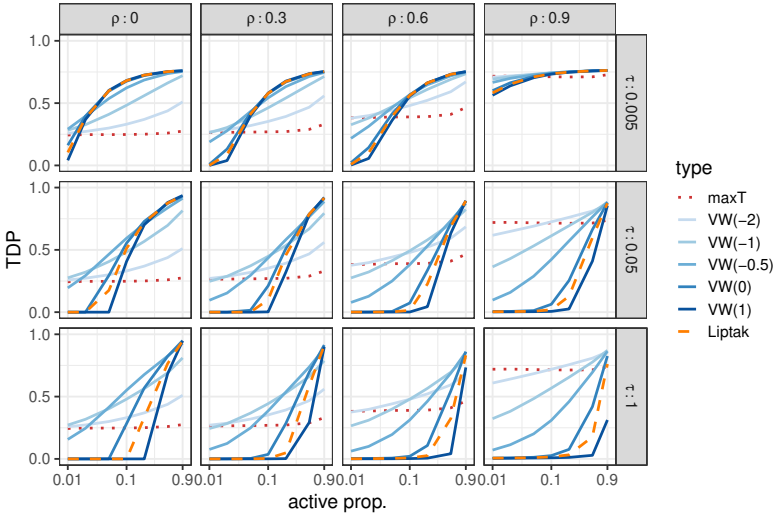
```
permT <- genePvals(exprs, labels,
          type = "vovk.wang", r = -1)

geneAnalysis(permT, pathways)
```

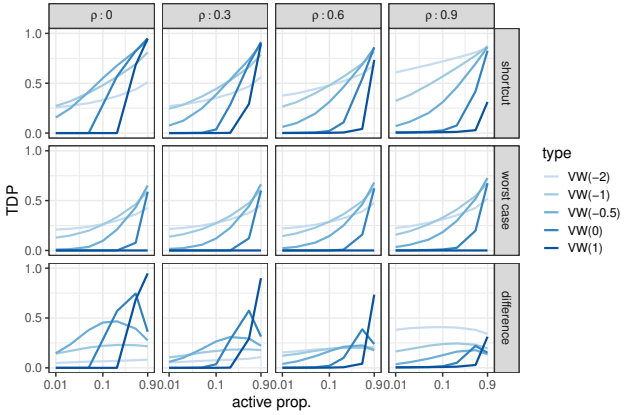



[1] https://CRAN.R-project.org/package=sumSome

# Simulations for p-value combinations



# Comparisons with worst-case distributions[1]




[1]Vovk and Wang. Combining p-values via averaging. *Biometrika*, 2020

# Breast invasive carcinoma dataset

We expect dense signal and medium-high correlation
$\rightarrow$ harmonic mean p-value with truncation ($\tau = \alpha$, $\gamma = 0.5$)

$m = 15{,}678$, $B = 200 \rightarrow \approx 10$ min on a standard PC for all pathways

Among 352 pathways:

- 344 with non-zero TDP
- 56 with TDP higher than the whole gene set

# Breast invasive carcinoma dataset: top 10 pathways

| pathway $S$ | function | size $s$ | TDP (%) $d(S)/s$ |
|---|---|---|---|
| all genes | | 15,678 | 33.08 |
| hsa03450 | non-homologous end-joining | 11 | 72.73 |
| hsa03050 | proteasome | 44 | 70.45 |
| hsa04110 | cell cycle | 120 | 55.83 |
| hsa03030 | DNA replication | 36 | 55.56 |
| hsa03013 | nucleocytoplasmic transport | 100 | 52.00 |
| hsa00900 | terpenoid backbone biosynthesis | 22 | 50.00 |
| hsa03267 | virion | 4 | 50.00 |
| hsa03008 | ribosome biogenesis in Eukaryotes | 69 | 49.28 |
| hsa01210 | oxocarboxylic acid metabolism | 17 | 47.06 |
| hsa00450 | selenocompound metabolism | 13 | 46.15 |

# Other permutation-based proposals

# Methods based on critical vectors/reference families

**Critical vector**

$\ell = (\ell_1, \ldots, \ell_m)$ such that

$$P\left(\ell_1 \leq q_{(1)}, \ldots, \ell_n \leq q_{(n)}\right) \geq 1 - \alpha$$

where $q_{(1)} \leq \ldots \leq q_{(n)}$ are the sorted p-values of true null hypotheses

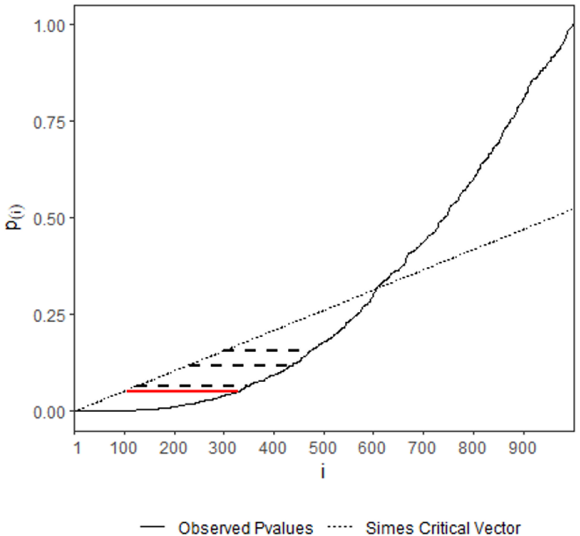
Pre-specified family:

- sansSouci - Blanchard et al., 2020
- pARI - Andreella et al., 2023

Data-dependent:

- Notip - Blain et al., 2022

# Computation of the confidence bound



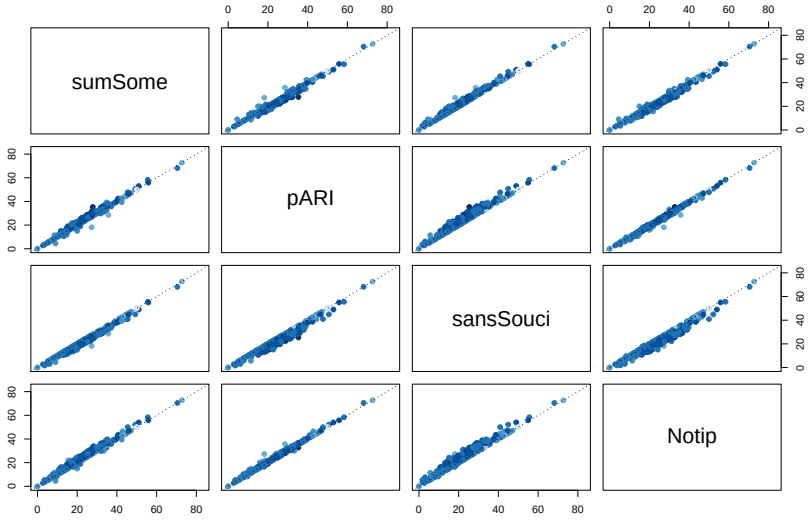

$$d(S) = \max_{u=1,\dots,d} h_S(u)$$

$$h_S(u) = 1 - u + |\{i \in S \,:\, p_i \leq \ell_u\}|$$

$\longrightarrow$ horizontal distance

- sorted p-values
- critical vector

# Breast invasive carcinoma dataset



# Sum tests: advantages

- Sum-based global tests $\rightarrow$ natural, intuitive and popular
- The sum structure can be exploited for more complex problems

**Example: netTDP**[1]

Differential gene co-expression analysis

Quantify difference at the level of network modules

- edges
- nodes (by aggregating)

[1] Cai et al. NetTDP: permutation-based true discovery proportions for differential co-expression network analysis. *Briefings in Bioinformatics*, 2022

# Conclusion

Closed testing procedure for global tests based on sums $\rightarrow$ simultaneous lower confidence bounds for the TDP of all subsets

It is an iterative shortcut that converges to closed testing results, but gives valid bounds even if stopped early

- Reference: Vesely et al. Permutation-based true discovery guarantee by sum tests. *JRSSB*, 2023
- Software: R package `sumSome` (CRAN)

# Back-up slides

# Permutation test

An $\alpha$-level test can be defined using $B$ random transformations from a group $\mathcal{P}$ (where the first is the identity)[1].

**Assumption**

If $N \subseteq M$ contains the indices of true hypotheses, then $(T_i)_{i \in N} \overset{\mathrm{d}}{=} (T_i^{\pi})_{i \in N}$ for each $\pi \in \mathcal{P}$.

**Test**

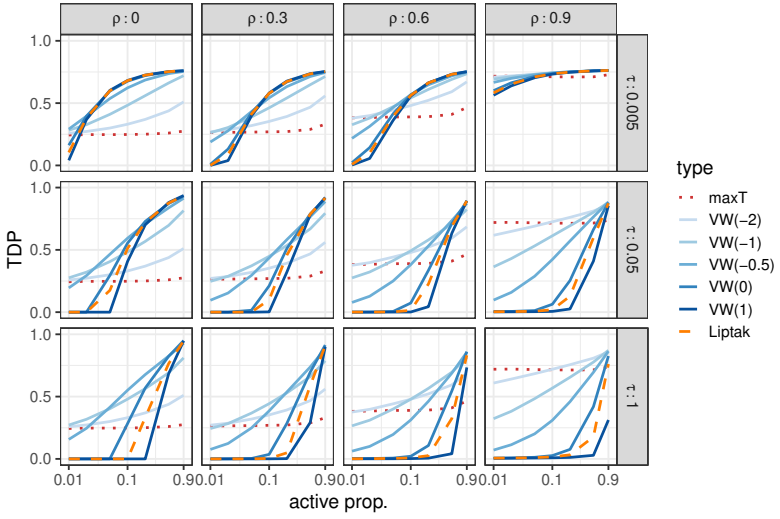
The test rejects $H_S$ when $T_S^{\mathrm{id}} > T_S^{(\lceil (1-\alpha)B \rceil)}$, where $T_S^{(1)} \leq \ldots \leq T_S^{(B)}$.

[1] Hemerik and Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *JRSSB*, 2018

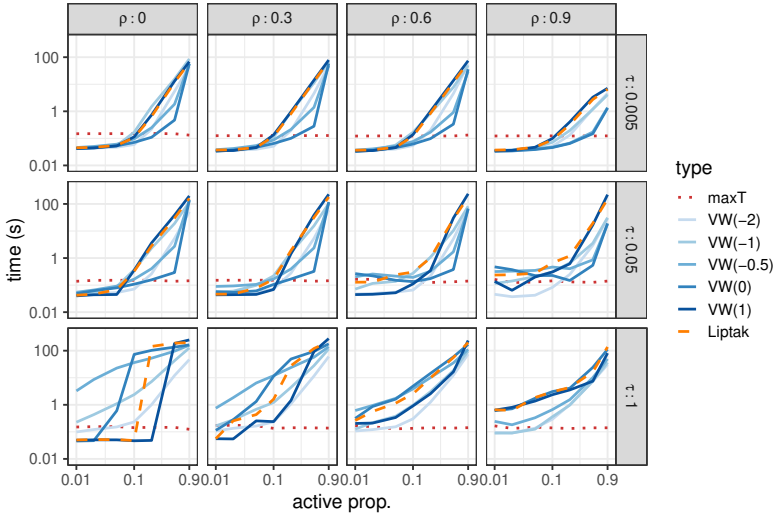
# Simulations

- $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \in \mathbb{R}^{1000}$ with $\mathcal{N}_{1000}(0, \Sigma_\rho)$
- a proportion $a$ of entries in $\boldsymbol{\mu}$ is non-null, with value given by a parameter $\beta$
- 50 observations
- p-values are computed from one-sample t-tests, for 200 permutations
- p-values are then truncated from a threshold $\tau$ to 0.5

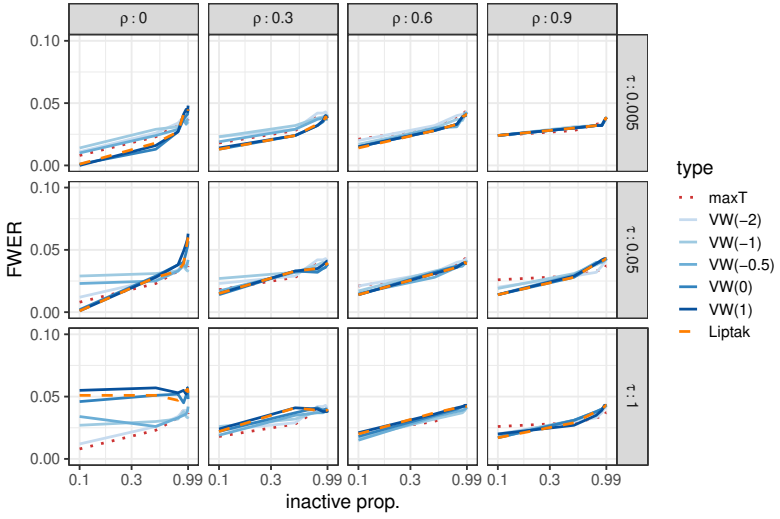
# Results for the TDP



# Computation time



# FWER control



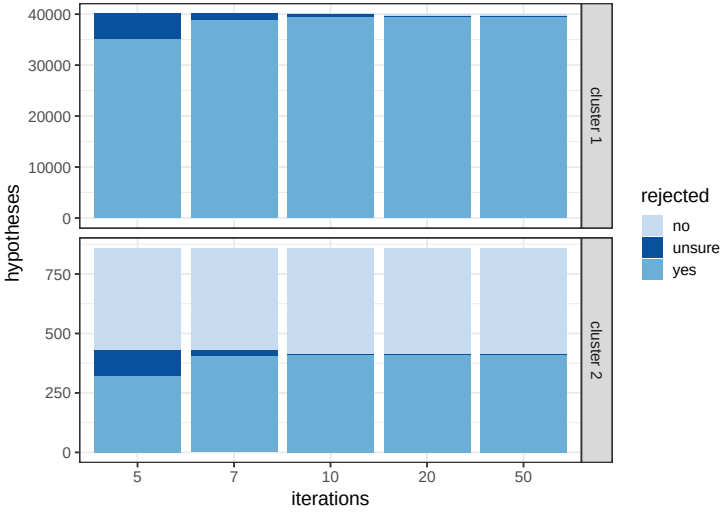
# Arrow data

26 subjects performing an Eriksen Flanker task[1]

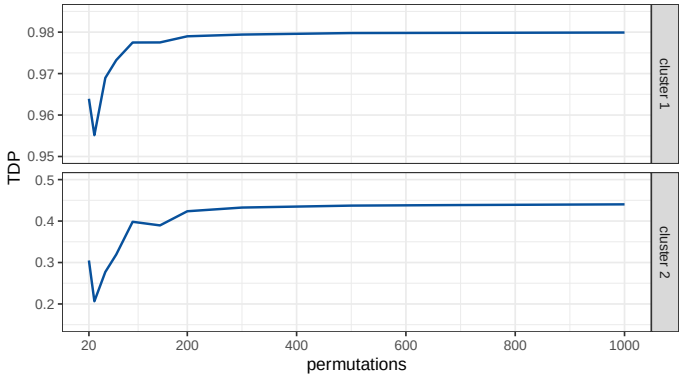
| cluster $S$ | threshold $thr$ | size $s$ | TDP (%) lower conf. bound |
|---|---|---|---|
| PoG/PrG/SPL/LOC | 3.2 | 9,752 | 35.77 |
| LOC/SPL/OFG/OFG | 3.2 | 1,493 | 19.26 |

[1] Kelly et al. Flanker task (event-related). *OpenNeuro dataset*, 2018

# Effect of the number of iterations



# Effect of the number permutations




Note: when $B$ is a multiple of $1/\alpha$ the power peaks, as the test is exact[1].

[1] Hemerik and Goeman. Exact testing with random permutations. *TEST*, 2018