A test for the comparison of gene-set transcriptomic profiles of vaccines

Arthur Hughes [1,2] Denis Agniel [4,5] Rodolphe Thiébaut [1,2,3] Boris Hejblum [1,2]

[1]University of Bordeaux, BPH INSERM U1219, INRIA SISTM [2]Vaccine Research Institute [3]CHU de Bordeaux

[4]Harvard Medical School, Boston, USA [5]RAND Corporation, Santa Monica, USA

Introduction

Comparing gene expression profiles between vaccines has high potential

- Understanding vaccine mechanisms
- Identifying biomarkers

But there are many challenges with high-dimensional data...

⚠ Interpretation of results
⚠ Sensitivity to investigator choices
⚠ Low signal-noise ratio

Gather genes in *gene sets* defined by biological function

✓ Reduce dimension
✓ Ease interpretation
✓ Maximise signal

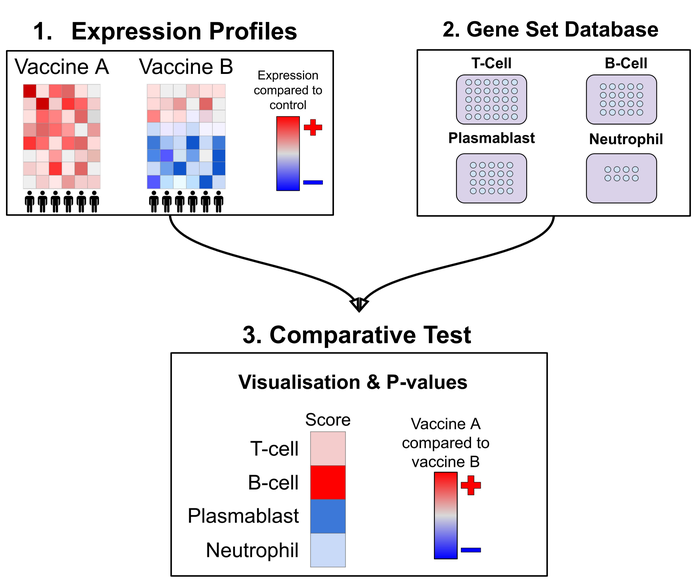

Figure 1. Illustration of the gene set comparative test.

Working Linear Mixed Effects Model

Test statistic derived from a working linear-mixed effects model:

$$\boldsymbol{y_i}^G = \boldsymbol{\alpha_0} + \boldsymbol{X_i}\boldsymbol{\alpha} + \boldsymbol{\Phi_i}\boldsymbol{\beta} + \boldsymbol{\Phi_i}\boldsymbol{\xi_i} + \boldsymbol{\epsilon_i}$$

- $\boldsymbol{y_i}^G = ((\boldsymbol{y_i}^1)^T, ..., (\boldsymbol{y_i}^p)^T)^T$ - expression of p genes in set G for individuals $i = 1, ..., n$
- $\boldsymbol{X_i}$ - matrix of baseline covariates to control for
- $\boldsymbol{\Phi}_i$ - K time-dependent variables whose association with $\boldsymbol{y_i}^G$ is to be tested

$\implies$ $\boldsymbol{\beta}$ - fixed effects of testing variables
$\implies$ $\boldsymbol{\xi_i} \sim \mathcal{N}(\mathbf{0}, \Sigma_\xi)$ - individual-level random effects of testing variables
$\implies$ $\boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$ - random error terms

Null Hypothesis

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \boldsymbol{\xi_i} = \mathbf{0}$$

Variance-Component Score Test

The derived **variance-component score test statistic** [1] is $Q = \boldsymbol{q}^T\boldsymbol{q}$ with

$$\boldsymbol{q}^T = n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{y}_i^G - (\boldsymbol{\alpha}_0 + X_i\boldsymbol{\alpha}))^T\Sigma_i^{-1}\Phi_i$$

- Central limit theorem $\implies Q \underset{+\infty}{\sim} \sum_{k=1}^{pK} a_k\chi_1^2$
  where a_k is the kth eigenvalue of $\mathrm{cov}(\boldsymbol{q})$

✓ Only requires estimation of model under H_0 (i.e. linear model)!
✓ Type 1 error control relies only on the central limit theorem!

Gene-Set Correlation

⚠ Dependence between genes in same set

Take into account in residual variance Σ_i

1. Fit linear model for each gene $g \in G$:
$$y_{ij}^g = \alpha_0^g + \boldsymbol{x_i}^T\boldsymbol{\alpha}^g + \boldsymbol{\phi_{ij}}^T\boldsymbol{\beta}^g + e_{ij}^g$$
2. OLS estimates residuals $r_{ij}^g = y_{ij}^g - \widehat{y_{ij}^g} \implies \boldsymbol{r}^g = (r_{11}^g, ..., r_{1t}^g, ..., r_{n1}^g, ..., r_{nt}^g)^T$
3. Entry $(g1, g2)$ of Σ_i $\implies$ **covariance between gene-wise residuals** $\forall g1, g2 \in G$

$$[\Sigma_i]_{g1,g2} = Cov(\boldsymbol{r}^{g1}, \boldsymbol{r}^{g2})$$

Application - Yellow Fever vs Flu Vaccines

- Compare early gene expression signatures of yellow fever and influenza vaccines*
- Gene sets: BloodGen3 Modules [2] - immunology focused sets

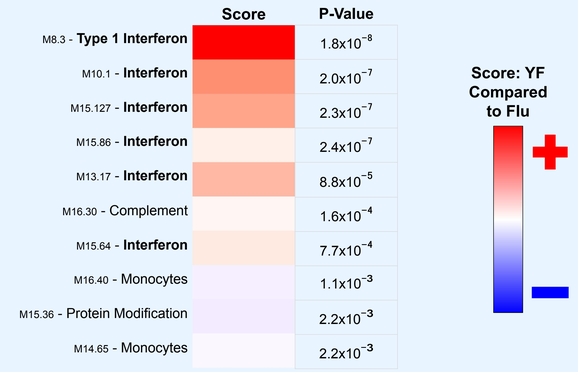

Figure 2. Scores and p-values for the top 10 most significant sets.

$\implies$ 300/382 gene sets significant**
$\implies$ Interferon modules strongly up-regulated early in YF compared to flu vaccine

*Public data from Immune Signatures Data Resource [3] - gene expression prior to 7 days post-vaccination of 1090 samples from 709 individuals vaccinated with either YF-17D or inactivated seasonal influenza vaccines. Analysis controls for age and sex, but not correlation structures.
**P-values corrected for multiple testing with Benjamini-Hochberg procedure

Summary

- Gene sets resolve problems with transcriptomic data
- Variance-component score test is a flexible, powerful method
- Biologically interpretable differences found between two vaccines
- Future work : explore properties of test under correlation structure estimation

References

[1] Marine Gauthier, Denis Agniel, Rodolphe Thiébaut, and Boris P Hejblum. Dearseq: A variance component score test for rna-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics*, 2(4), 2020.

[2] Matthew C. Altman and et al. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nature Communications*, 12(1), July 2021.

[3] Joann Diray-Arce and et al. The immune signatures data resource, a compendium of systems vaccinology datasets. *Scientific Data*, 9(1), 2022.

Arthur Hughes is supported by the Digital Public Health Graduate school, funded by the PIA 3 (Investments for the Future – Project reference: 17-EURE-0019)

arthur.hughes@u-bordeaux.fr