Analyzing transcriptomics data for understanding and predicting vaccine response in clinical trials

Rodolphe Thiébaut

Bordeaux, France

Inserm
La science pour la santé
From science to health

Inria

ISPED
SCHOOL OF PUBLIC HEALTH

Université de BORDEAUX

CHU BDX
CENTRE HOSPITALIER UNIVERSITAIRE BORDEAUX

SISTM / Statistics in Systems biology and Translational Medicine

Inserm
La science pour la santé
From science to health

Inria

SISTM / Statistics in Systems biology and Translational Medicine

and its environment…

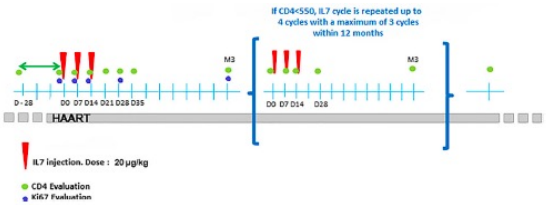
euclid
F-CRIN platform for clinical trials

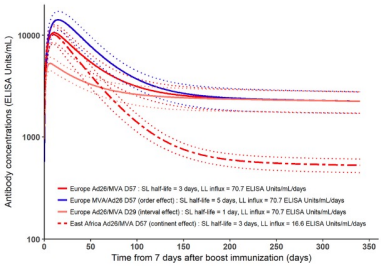
VRI
VACCINE RESEARCH INSTITUTE

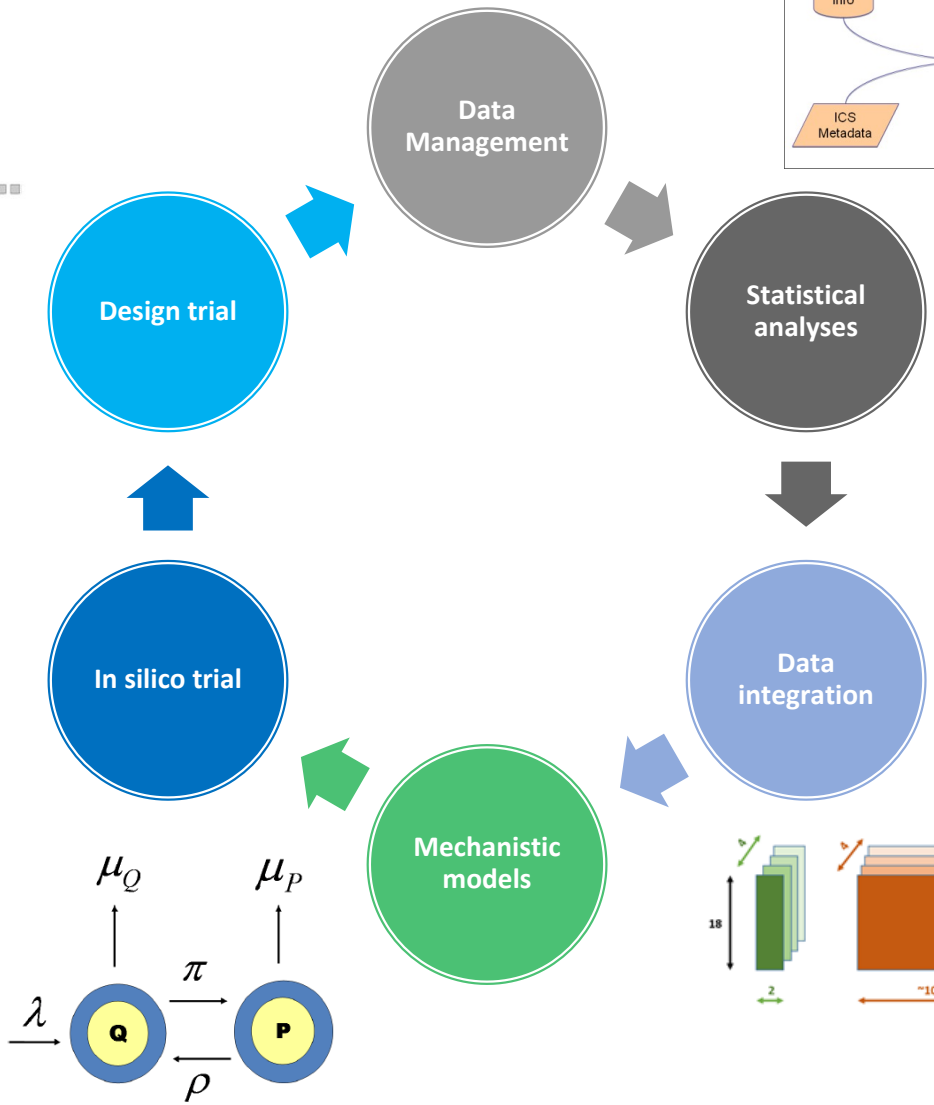
anrs
MALADIES INFECTIEUSES ÉMERGENTES Inserm

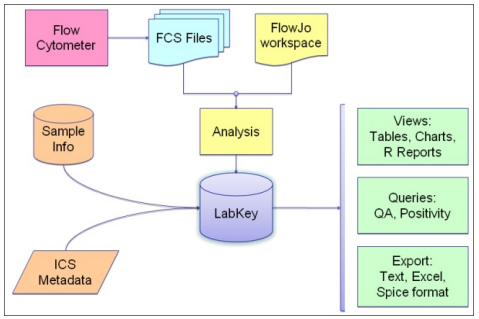
Bordeaux CTU

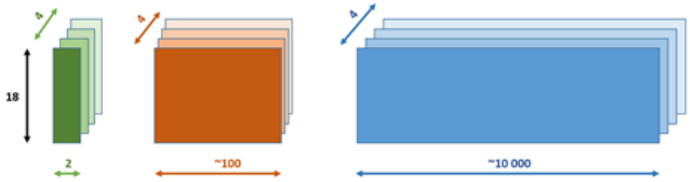
A big picture











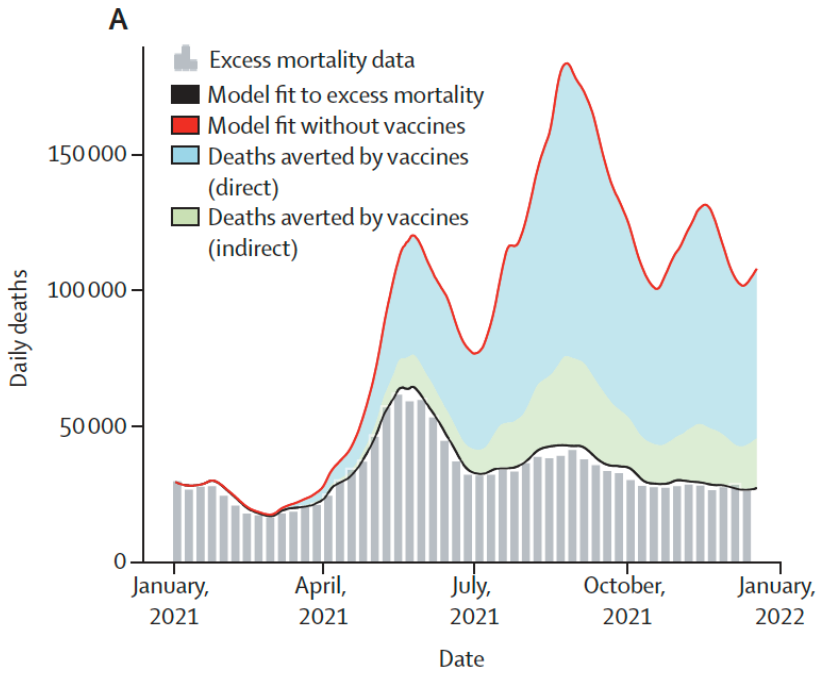
Outline

- Systems vaccinology
- Gene expression differential analysis
- Random forests for longitudinal data

Vaccines

Global impact of the first year of COVID-19 vaccination: a mathematical modelling study

Oliver J Watson, Gregory Barnsley*, Jaspreet Toor, Alexandra B Hogan, Peter Winskill, Azra C Ghani*

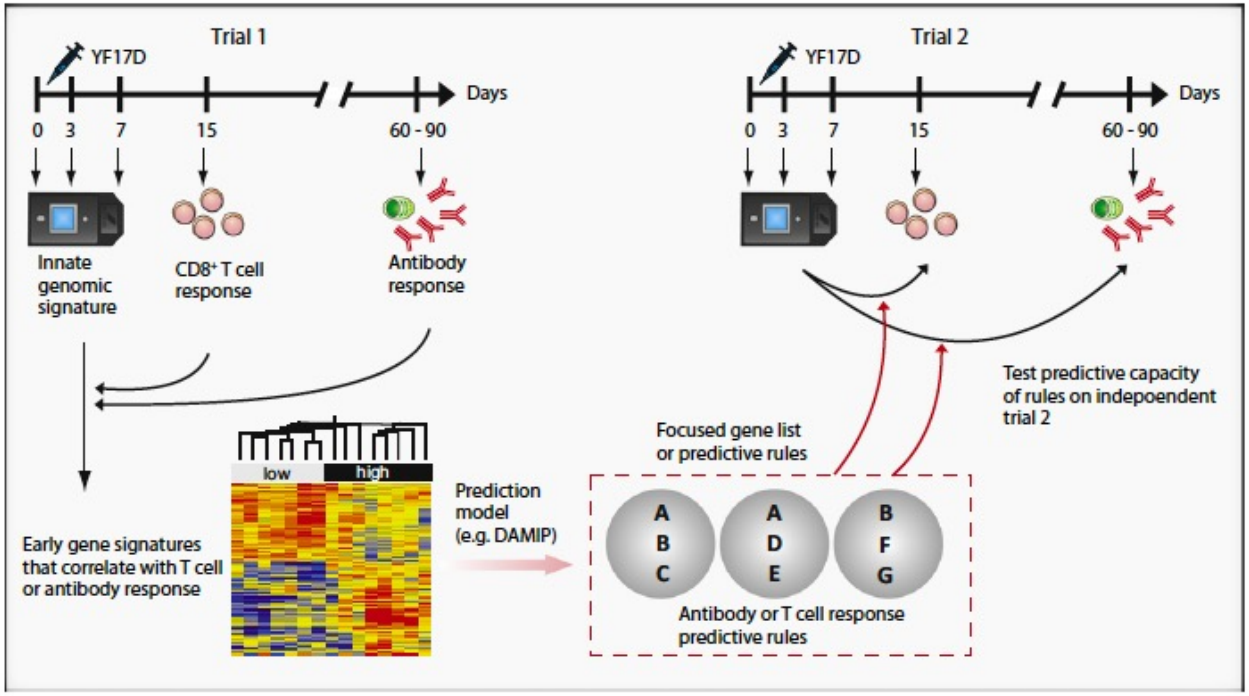


	Total COVID-19 deaths	Vaccination coverage (%)	Estimated deaths averted by vaccinations		
			Total	Per 10 000 people	Per 10 000 vaccines
Worldwide	5 469 000 (5 339 000–5 613 000)	38·30%	14 400 000 (13 650 000–15 900 000)	22·81 (21·63–25·18)	25·99 (24·64–28·69)

Systems vaccinology

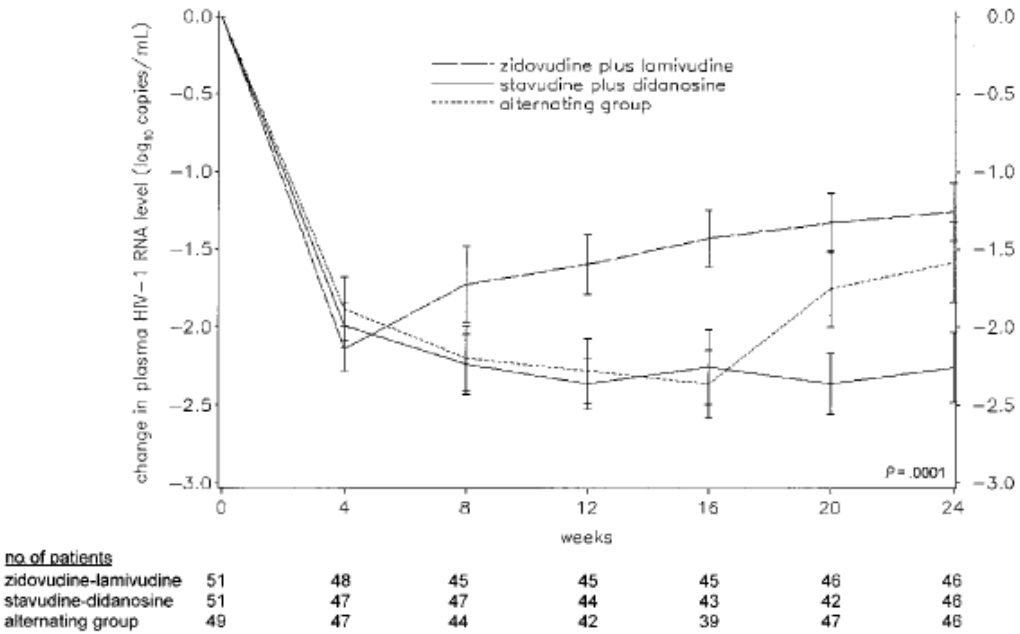
Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans

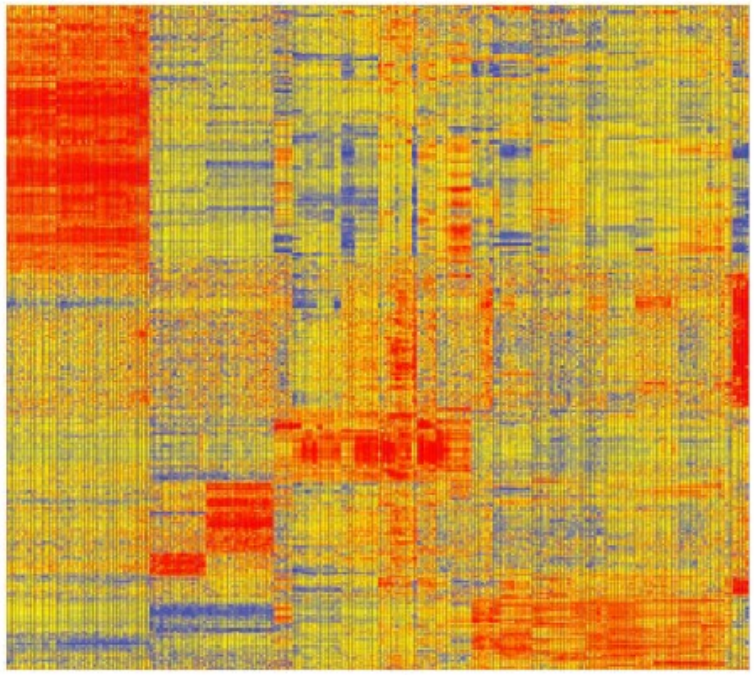
Troy D Querec[1,8], Rama S Akondy[1,8], Eva K Lee[2], Weiping Cao[1], Helder I Nakaya[1], Dirk Teuwen[3], Ali Pirani[4], Kim Gernert[4], Jiusheng Deng[1], Bruz Marzolf[5], Kathleen Kennedy[5], Haiyan Wu[5], Soumaya Bennouna[1], Herold Oluoch[1], Joseph Miller[1], Ricardo Z Vencio[5], Mark Mulligan[1,6], Alan Aderem[5], Rafi Ahmed[1] & Bali Pulendran[1,7]



Gene symbol	Gene ID	1	2	3
BEND4	Hs.120591			
	Hs.139006	X		
PFKFB3	Hs.195471		X	
TNFRSF17	Hs.2556	X	X	X
TPD52	Hs.368433			
	Hs.481166			
KBTBD7	Hs.63841	X	X	X
	Hs.649726			
NAP1L2	Hs.66180			X
		80	80	80
		100	100	100
		97	99	94

Data in vaccinology



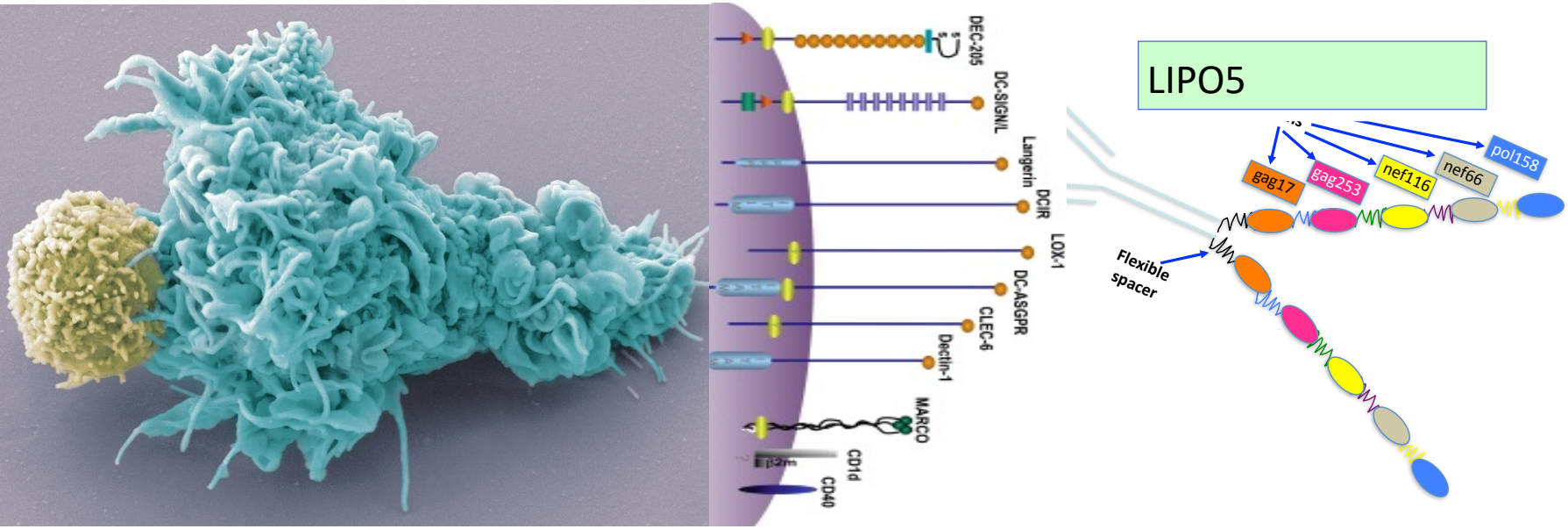


Trial	Year	Sample size	Folder size
ALBI ANRS 070	1999	151	67 Ko
DALIA-1	2014	19	200 Go

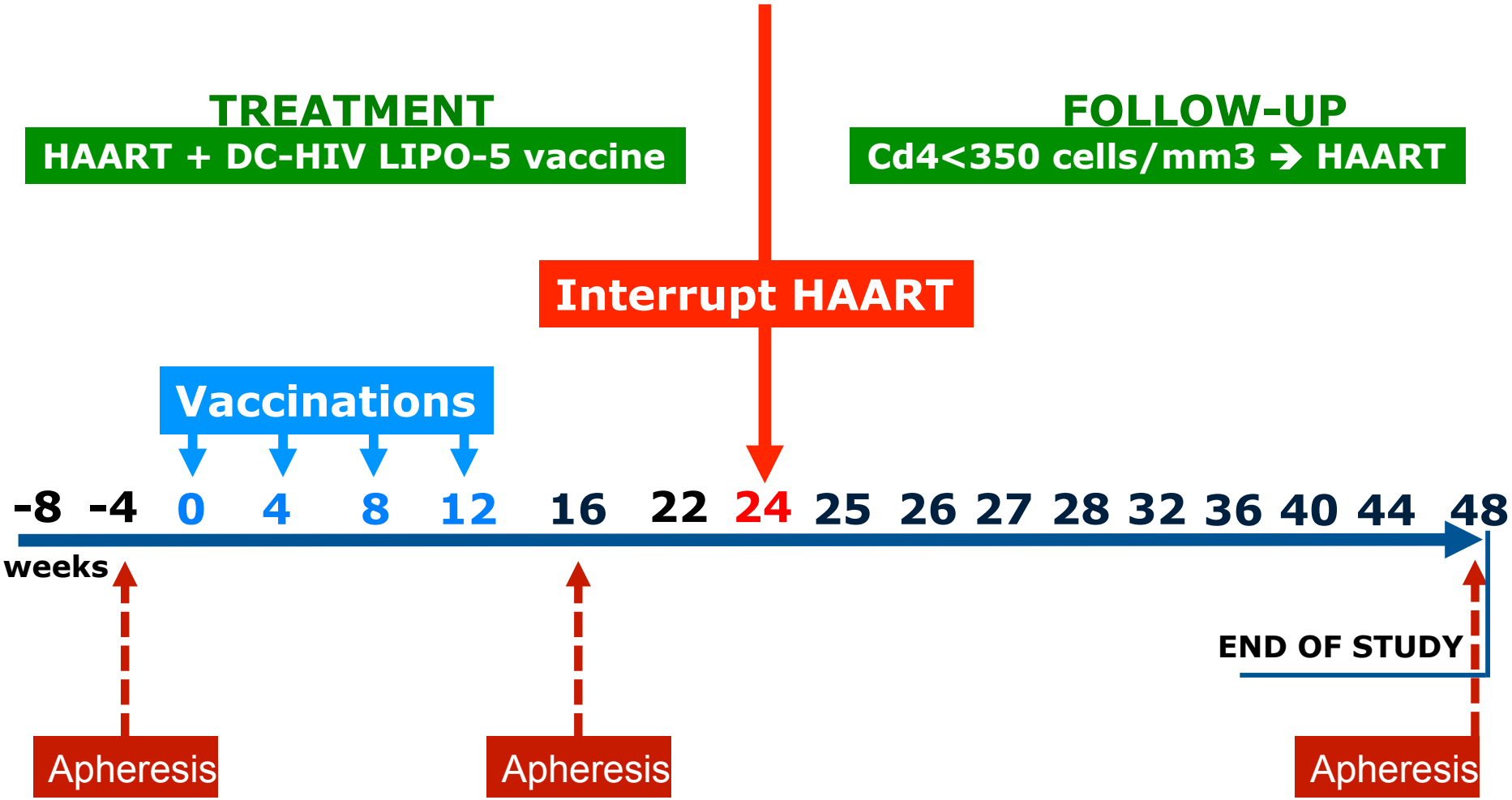
VACCINE RESEARCH INSTITUTE

Dendritic cell based vaccine

- Therapeutic vaccine in HIV-infected patients
- Dendritic Cells are loaded with 5 HIV peptides



DALIA-1 trial design

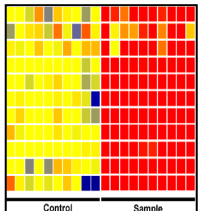


Levy et al. J Immunology 2014

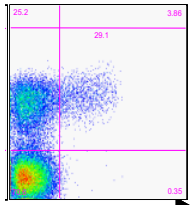
Data in vaccinology

- 846 000 probes (18 temps x 47 000 sondes) 26 Mo
- 18 612 000 beads (22 billes/sonde) 6 Go

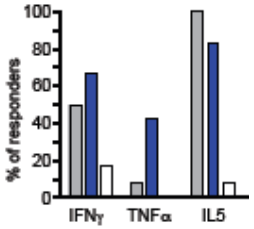
- 30 populations cellulaires 0.05 Mo
- 2160 anticorps (18 temps x 15 tubes x 8 anticorps) pour 2.6 Go



Gene profiling

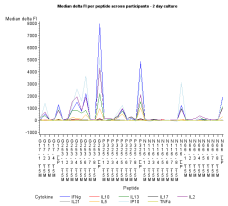


Cell responses

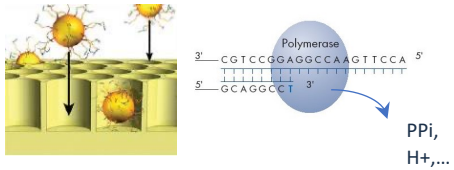


Cytokines

- 800 mesures/temps
- 0.35 Mo



Epitope mapping



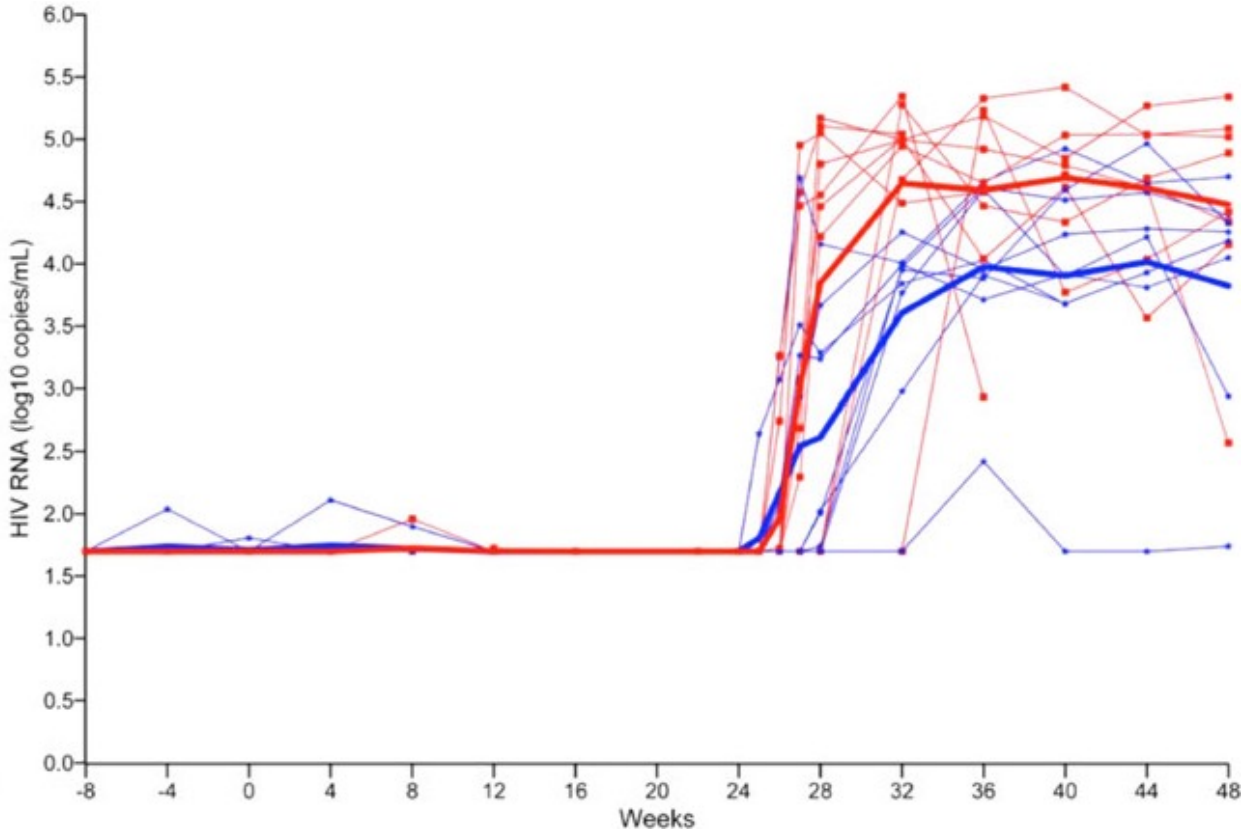
Viral changes/ adaptation

- 200 séquences
- 20 Mo

Levy et al. J Immunology 2014

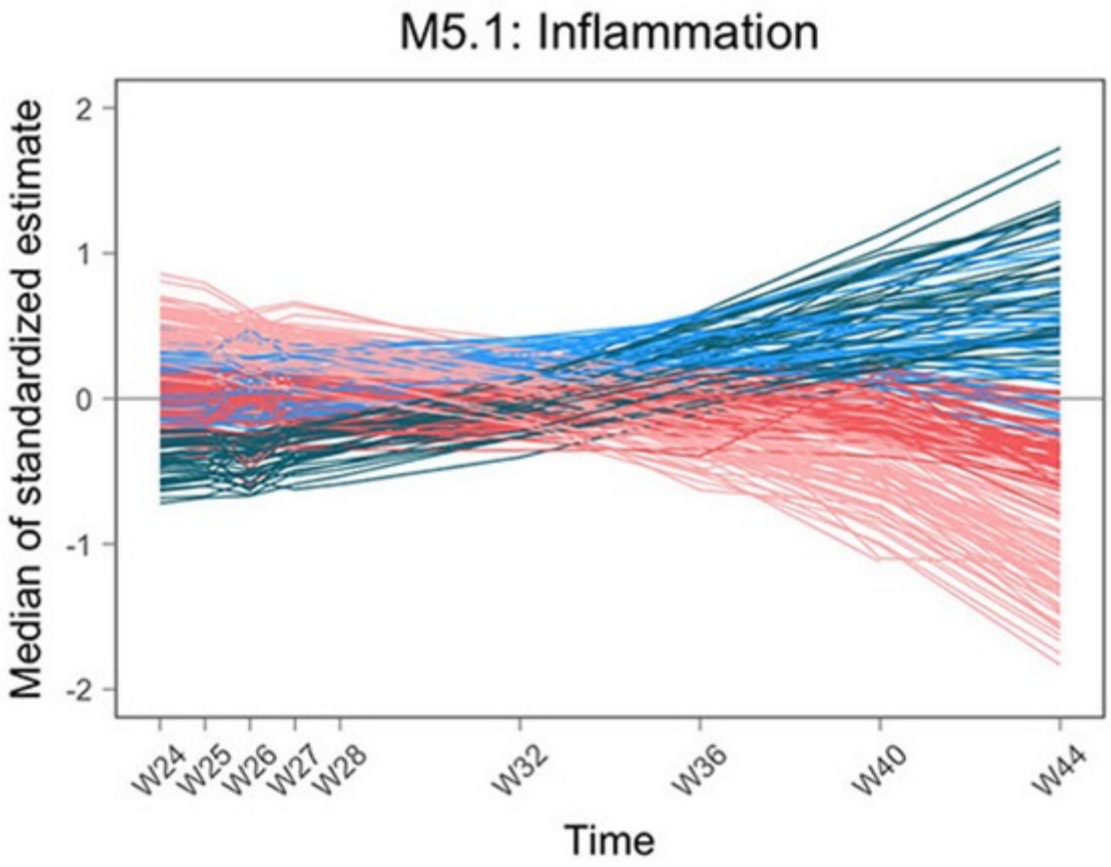
DALIA-1 results: HIV RNA viral load after treatment interruption

VACCINE RESEARCH INSTITUTE



Levy et al. J Immunology 2014

DALIA-1 results: gene expression after treatment interruption

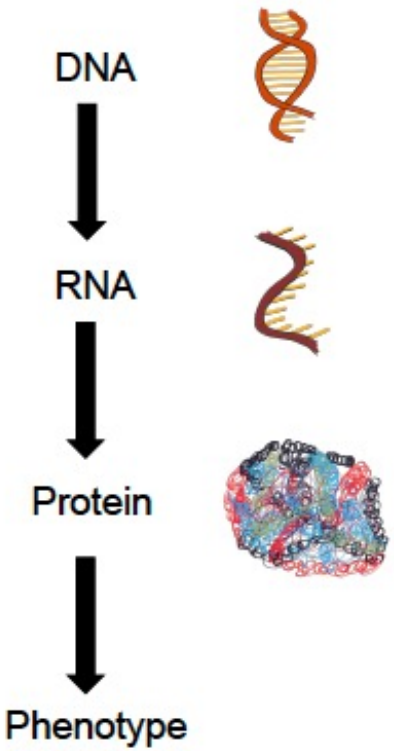


Thiebaut et al. Front Immunol 2019 & Hejblum Plos Comp Biol 2015

Aims

- To understand the mechanism of vaccine response
- To predict the vaccine response

The gene expression (transcriptome)

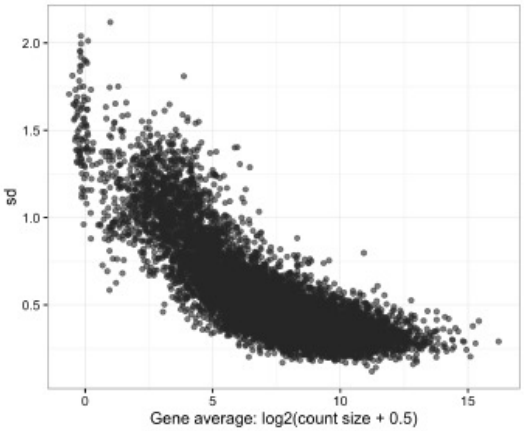


- Messenger RNA
- Measured by
  - Microarray: fluorescence intensity (continuous)
  - RNA-sequencing: count data
- Could be measured:
  - At the single-cell level
  - In the whole blood (bulk): gene abundance

RNA-seq differencial analysis

- Objective: to compare gene expression/abundance between/within groups

- Taking into account:
  - Test multiplicity
    (control of Type-I error and False Discovery rate)
  - RNA-seq data heterosckedasticity



RNA-seq differencial analysis

- The reference approaches
  - EdgeR: Robinson et al. Bioinformatics 2010 (>17K citations)
  - DESeq2: Love et al. Genome Biology 2014 (>24K citations)
  - Voom-lima: Law et al. Genome Biology 2014 (>2K citations)
- But limitations (Rapaport 2013, Rocke 2015, Germain 2016)
  - Strong parametric assumptions
  - Tailored for small studies

Dearseq method

Marine Gauthier

Boris Hejblum

We rely on the following **working model** for each gene g:

Model

$$\boldsymbol{y}_i = \boldsymbol{\alpha}_0 + X_i\boldsymbol{\alpha} + \Phi_i\boldsymbol{\beta} + \Phi_i\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i,$$

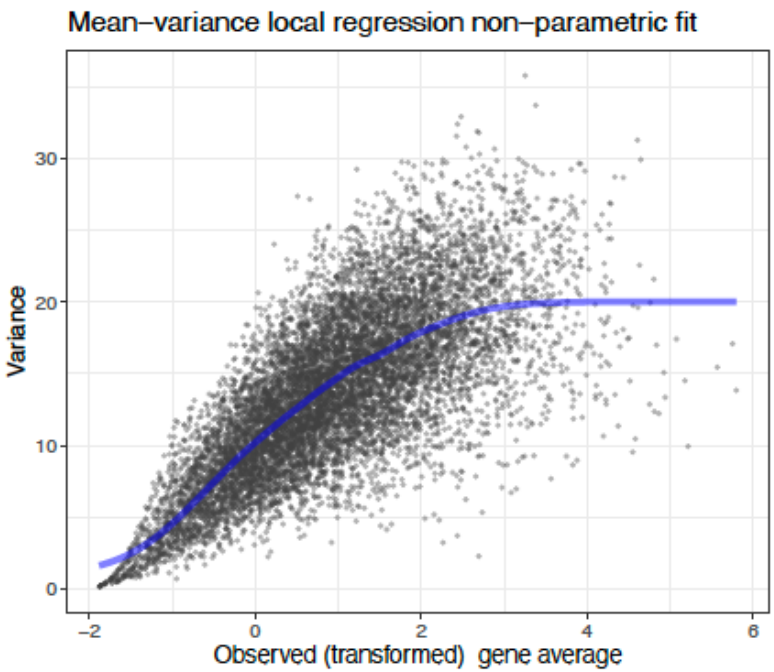
$$\boldsymbol{\xi}_i \sim N(0, \Sigma_\xi), \boldsymbol{\epsilon}_i \sim N(0, \Sigma_i)$$

$$\forall i = 1, ..., n$$

- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ is a $n_i \times 1$ vector of normalized gene expression measurements
- $\boldsymbol{\alpha}_0$ is a $n_i \times 1$ vector of intercepts
- X_i is the $n_i \times p$ matrix of covariates
- $\boldsymbol{\alpha}$ is a $p \times 1$ vector of fixed effects
- Φ_i is the $n_i \times m$ matrix of the variables of interest
- $\boldsymbol{\beta}$ is a $m \times 1$ vector of fixed effects
- $\boldsymbol{\xi}_i \sim N(0, \Sigma_\xi)$ is a $m \times 1$ vector of individual-level random effects of the variables of interest
- $\boldsymbol{\epsilon}_i$ is a $n_i \times 1$ vector of measurement error

Gauthier et al. NAR Genomics and Bioinformatics 2020

Dearseq method



Measurements errors: $\epsilon_i \sim N(0, \Sigma_i)$

$$v_{ij}^g = Var(y_{ij}^g \mid X_{ij}, \boldsymbol{\xi}_i^g) \quad m_{ij}^g = E(y_{ij}^g \mid X_{ij}, \boldsymbol{\xi}_i^g)$$

Following Law *et al.*, we model the mean-variance relationship at the gene level through Σ_i:

$$v^g = \omega(m^g) + e^g$$

for some unknown function $\omega(.)$ and

$$E(e^g) = 0, V(e^g) = \tau^2, \tau > 0.$$

$\Longrightarrow$ **Local linear regression** [Wasserman, 2006] **borrowing information across all genes**

$$diag(\widehat{\Sigma}_i^g) = \widehat{\omega}_n(\widehat{m}_{ij}^{(g)})$$

The parameters are estimated by Ordinary Least Square.

Gauthier et al. NAR Genomics and Bioinformatics 2020

Dearseq method

We derive a variance component score test statistic for the effects of interest. The null hypothesis of no effect of interest is:

$$H_0 : \text{“}\boldsymbol{\beta} = 0 \text{ and } \Sigma_\xi = 0\text{”}$$

Under H_0, $Q = q^T q$

$$\text{with } q^T = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{y}_{\mu_i}^T \Sigma_i^{-1} \Phi_i$$

where $\boldsymbol{y}_{\mu_i} = \boldsymbol{y}_i - \boldsymbol{\mu}_i = \boldsymbol{y}_i - \boldsymbol{\alpha}_{i0} + X_i \boldsymbol{\alpha}$

Asymptotic test

$$Q \underset{+\infty}{\sim} \sum_{l=1}^{n_i} a_l \chi_1^2$$

where the mixing coefficients a_l depend on the covariance of $\boldsymbol{q}$

Permutation test

Permutations $\Rightarrow$ Naive p-values

[Phipson & Smyth (2010)]

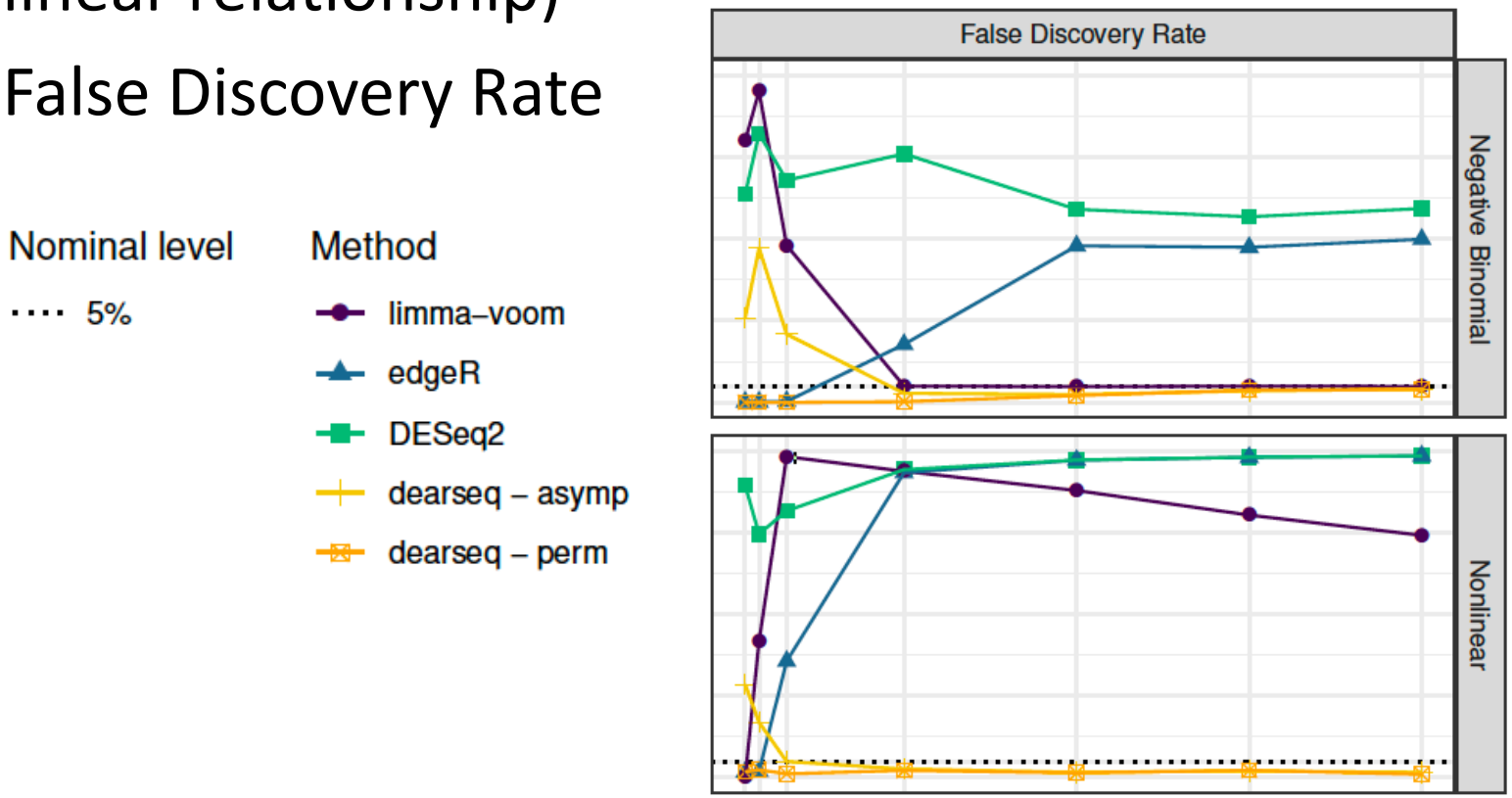
$\Rightarrow$ **exact p-values computations**

$\Longrightarrow$ Benjamini-Hochberg (1995) correction for multiple testing

Gauthier et al. NAR Genomics and Bioinformatics 2020

Simulations

- Monte Carlo estimation over 1000 simulations (non linear relationship)
- False Discovery Rate



Gauthier et al. NAR Genomics and Bioinformatics 2020

Valorisation

- Package dearseq

Volume 2, Issue 4
December 2020

dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate

Marine Gauthier, Denis Agniel, Rodolphe Thiébaut, Boris P Hejblum

NAR Genomics and Bioinformatics, Volume 2, Issue 4, December 2020, lqaa093,
https://doi.org/10.1093/nargab/lqaa093
Published: 19 November 2020 **Article history** ▾

Oups, were we wrong?

Li *et al. Genome Biology* (2022) 23:79
https://doi.org/10.1186/s13059-022-02648-4

Genome Biology

SHORT REPORT Open Access

Exaggerated false positives by popular differential expression methods when analyzing human population samples

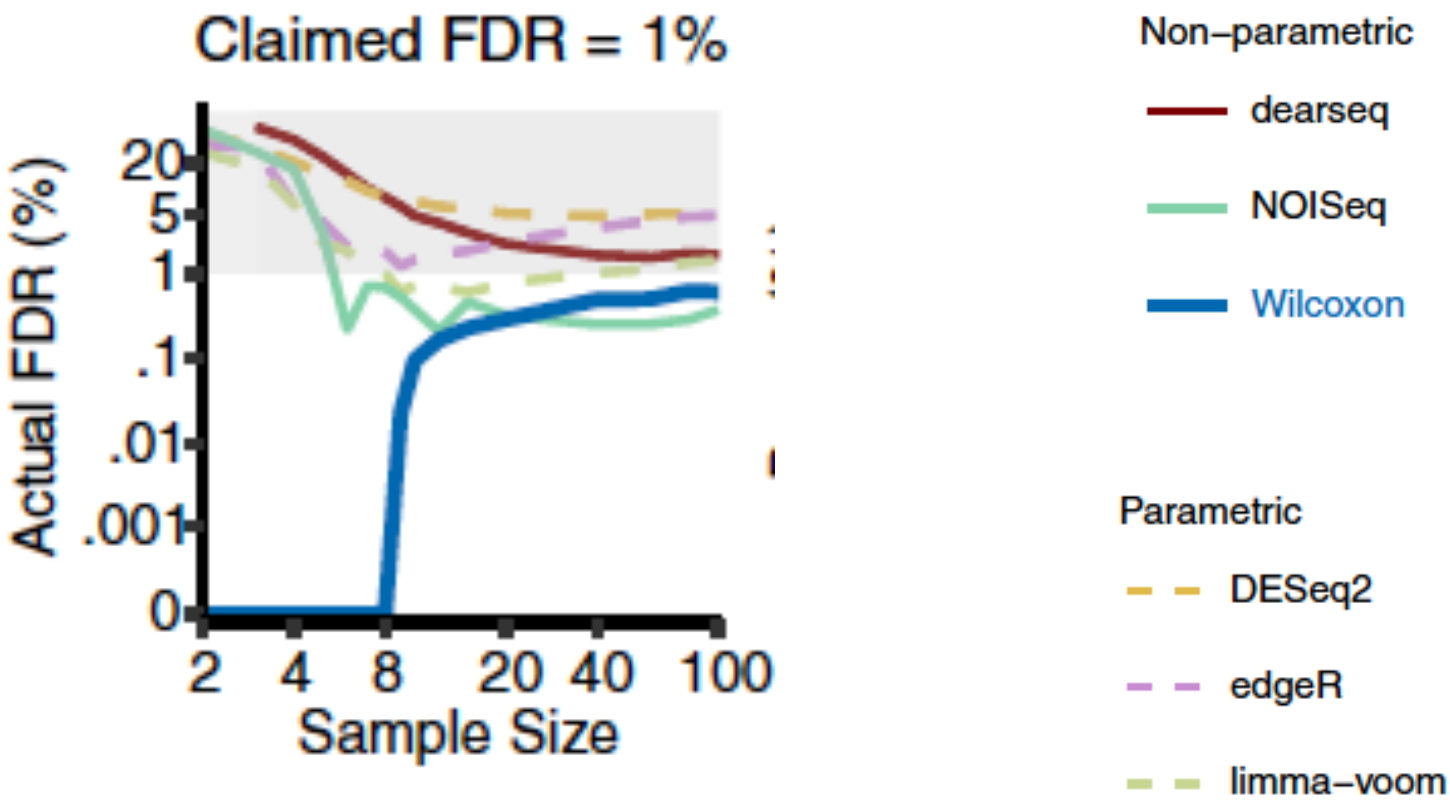
Yumei Li[1†], Xinzhou Ge[2†], Fanglue Peng[3], Wei Li[1*] and Jingyi Jessica Li[2,4,5,6,7*]

*Correspondence: wei.li@uci.edu; lijy03@g.ucla.edu
†Yumei Li and Xinzhou Ge contributed equally to this work.
[1] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[2] Department of Statistics,

Abstract

When identifying differentially expressed genes between two conditions using human population RNA-seq samples, we found a phenomenon by permutation analysis: two popular bioinformatics methods, DESeq2 and edgeR, have unexpectedly high false discovery rates. Expanding the analysis to limma-voom, NOISeq, dearseq, and Wilcoxon rank-sum test, we found that FDR control is often failed except for the Wilcoxon rank-sum test. Particularly, the actual FDRs of DESeq2 and edgeR sometimes exceed 20% when the target FDR is 5%. Based on these results, for population-level RNA-seq studies with large sample sizes, we recommend the Wilcoxon rank-sum test.

Oups, were we wrong?

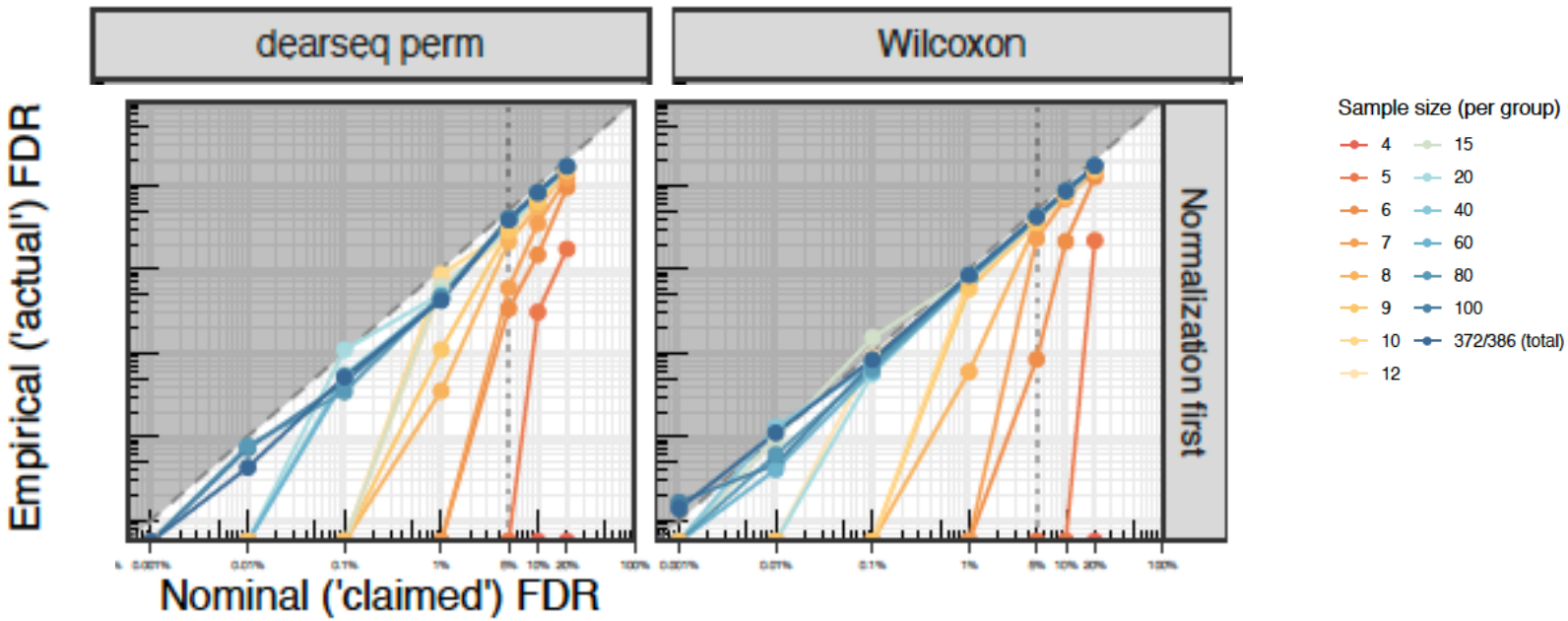


Li et al. Genome Biology 2022

Actually, no. It is ok.

Neglecting normalization impact in semi-synthetic RNA-seq data simulation generates artificial false positives

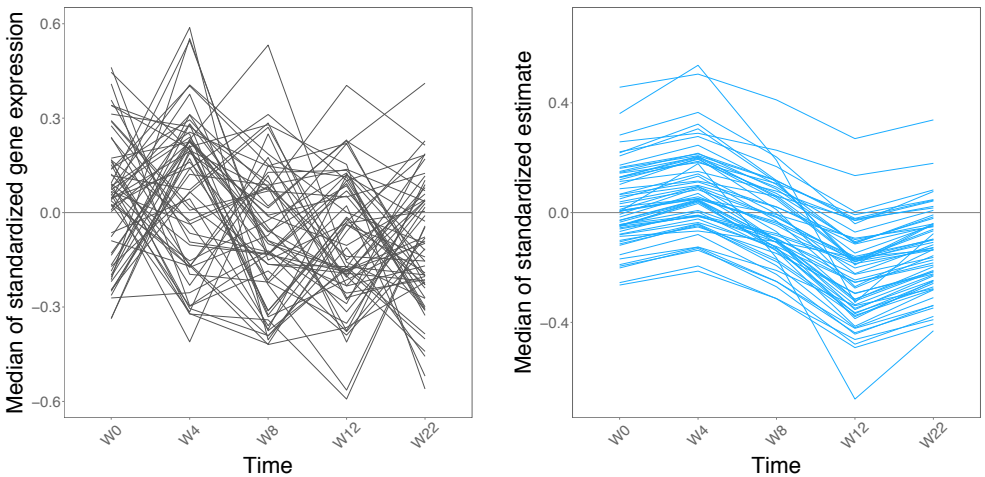
Boris P Hejblum[1,2,*], Kalidou Ba[1,2], Rodolphe Thiébaut[1,2,3], Denis Agniel[4,5]



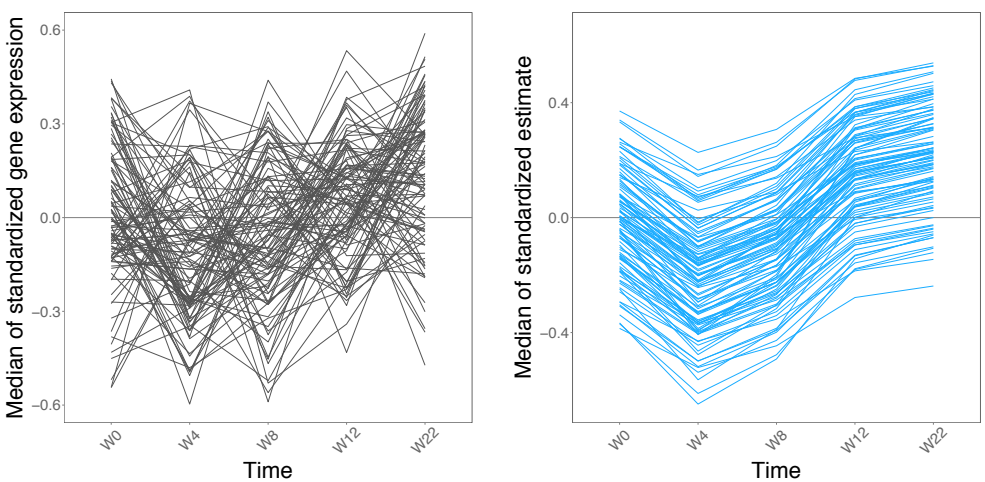
https://www.biorxiv.org/content/10.1101/2022.05.10.490529v1

VACCINE RESEARCH INSTITUTE

Time course geneset analysis



M 4.1: T cell – 86th percentile



M 4.6: Inflammation – 96th percentile

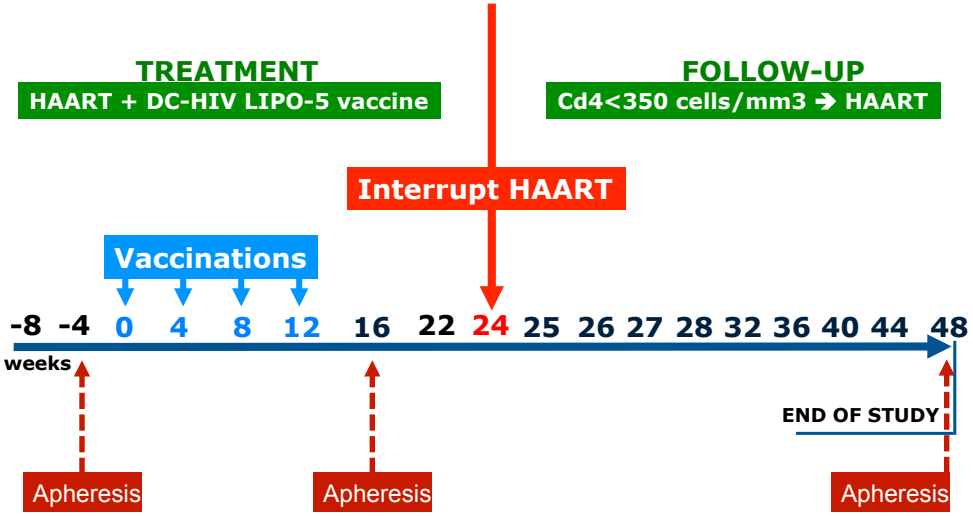
M3.2[99th pctile]: Inflammation 1/1
M1.1[99th pctile]: Platelets 1/1
M4.13[98th pctile]: Inflammation 1/1
M5.10[97th pctile]: Mitochondrial Respiration 1/1
M4.6[96th pctile]: Inflammation 1/1
M5.6[96th pctile]: Mitochondrial Stress / Proteasome 1/1
M3.5[95th pctile]: Cell Cycle 1/1
M5.7[94th pctile]: Inflammation 1/1
M3.1[94th pctile]: Erythrocytes 1/1
M2.3[94th pctile]: Erythrocytes 1/1
M7.1[93th pctile]: Inflammation 1/1
M6.2[93th pctile]: Mitochondrial Respiration 1/1
M4.3[91th pctile]: Protein Synthesis 1/1
M4.11[91th pctile]: Plasma Cells 1/3
M4.11[91th pctile]: Plasma Cells 2/3
M4.11[91th pctile]: Plasma Cells 3/3
M6.13[90th pctile]: Cell Death 1/1
M4.5[89th pctile]: Protein Synthesis 1/1
M4.14[88th pctile]: Monocytes 1/1
M4.2[88th pctile]: Inflammation 1/1
M4.7[87th pctile]: Cell Cycle 1/1
M4.1[86th pctile]: T cell 1/1
M3.6[85th pctile]: Cytotoxic/NK Cell 1/1
M5.9[84th pctile]: Protein Synthesis 1/1
M5.15[82th pctile]: Neutrophils 1/1
M4.15[81th pctile]: T cells 1/1
M6.6[80th pctile]: Apoptosis / Survival 1/1
M5.1[74th pctile]: Inflammation 1/1

Hejblum Plos Comp Biol 2015

Question

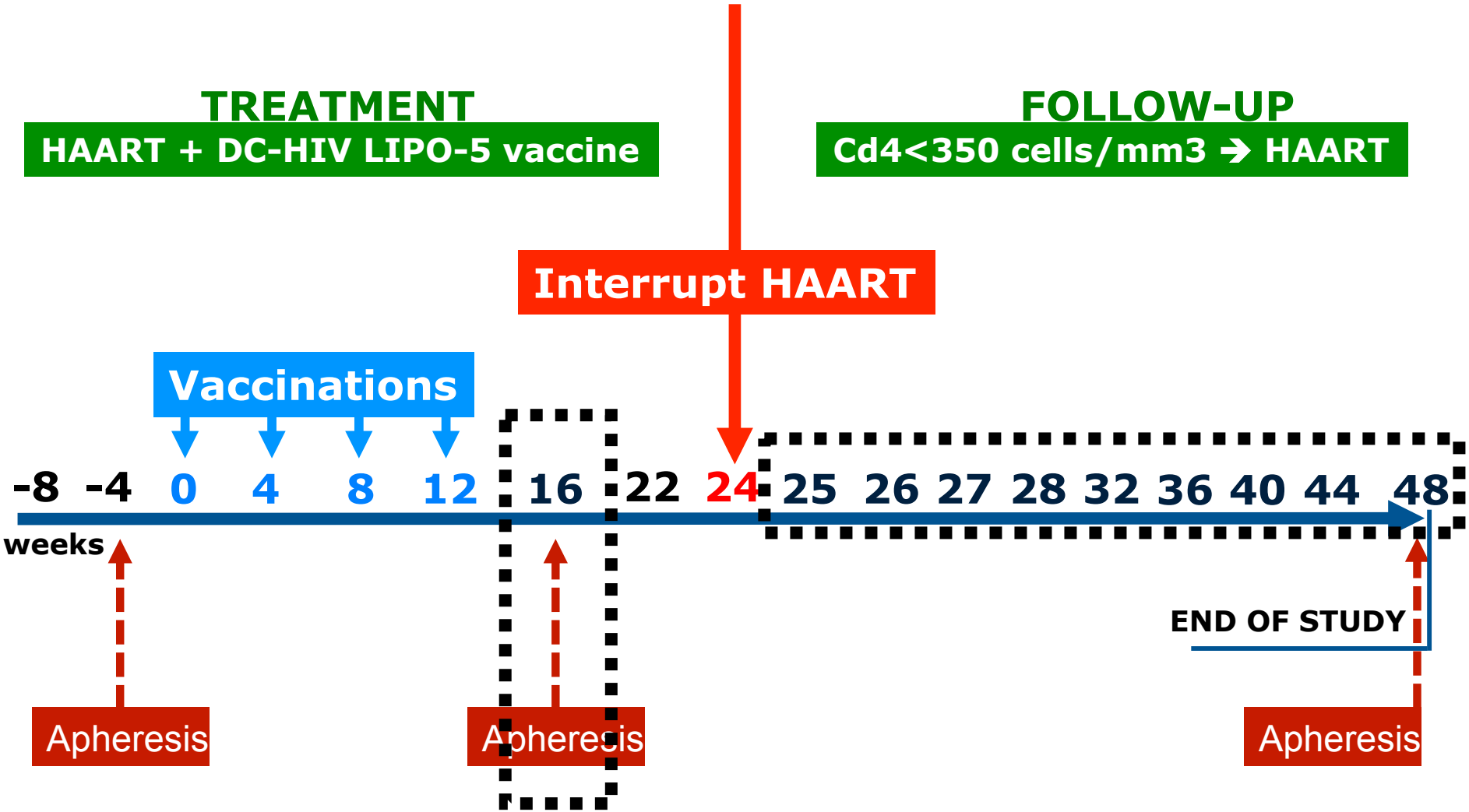
- Knowing which genes / genesets are moving over time,

- which ones are associated to the viral load dynamics?



VACCINE RESEARCH INSTITUTE

DALIA-1 trial design
Integrative Analysis



Sparse group Partial Least Squares method

Benoit Liquet

Number of predictors (genes) > number of patients:

PLS + LASSO

Regression method

- Independent linear combinations of predictors, ie **genes**
- Independent linear combinations of explaining variables, ie **immunological measures**
- Components **maximize their covariance**

Penalization method

- Select only a few predictors, i.e. **genes**
- **Sparse**: all non selected predictor Coefficients are estimated **as exactly 0**

- *Each **sPLS** component is **sparse:** only a few variables contributes to each components*
- *Selection could be done by **group** of genes*

Liquet et al. Bioinformatics 2015; Sutton et al. Statist Med 2018;37:3338

Group PLS

Aim: Select group variables taking into account the data structures

- PLS components
  $C^k = u_1 \times X_1 + u_2 \times X_2 + u_3 \times X_3 + \ldots + u_p \times X_p$
- sparse PLS components (sPLS)
  $C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$
- group PLS components (gPLS)

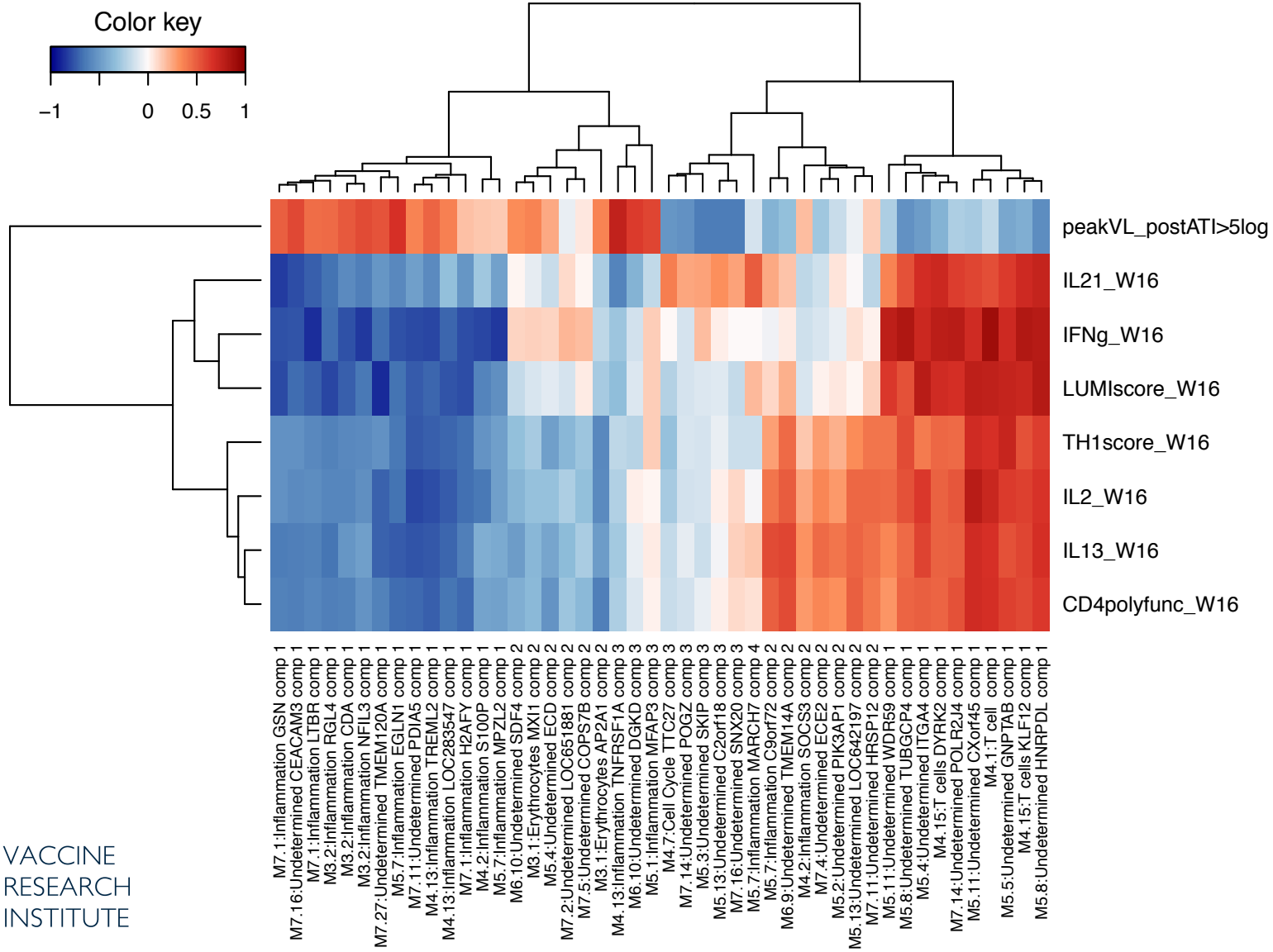
$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

$\hookrightarrow$ select group of variables; all the variables within a group are selected otherwise none of them are selected

sgPLS

Liquet et al. Bioinformatics 2015; Sutton et al. Statist Med 2018;37:3338

Correlations with immune response (W16) and peak of viral load (post ATI)



VACCINE
RESEARCH
INSTITUTE

Need more

- To take into account longitudinal setting
- Repeated measures of HIV RNA viral load
- Repeated measures of gene abundance

Random forest for longitudinal data

Article

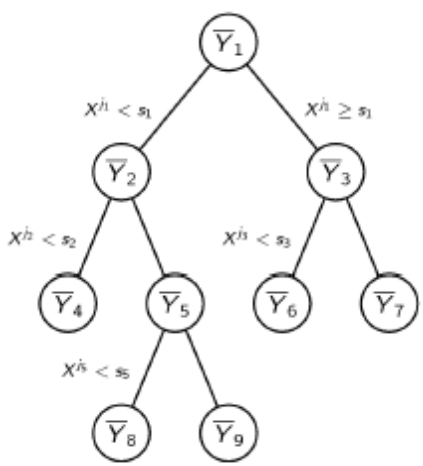
SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

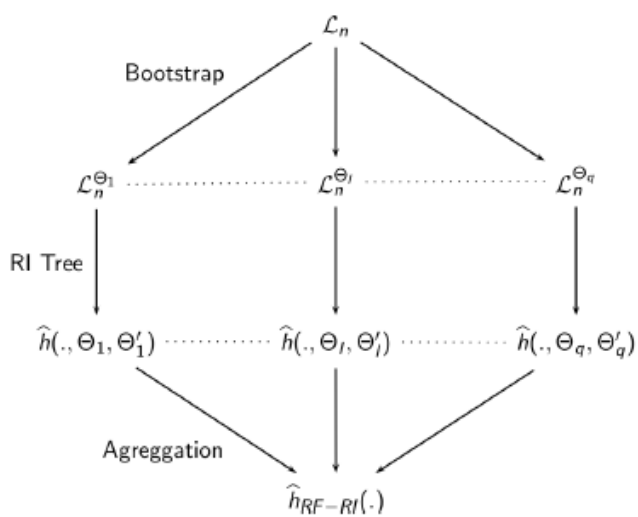
Statistical Methods in Medical Research
0(0) 1–19

DOI: 10.1177/0962280220946080
journals.sagepub.com/home/smm
SAGE

Random forests for high-dimensional longitudinal data

Louis Capitaine, Robin Genuer and Rodolphe Thiébaut





Random forest for longitudinal data

Model

Suppose Y_{ij}, the viral load of the ith individual at time t_{ij}, satisfies

$$Y_{ij} = f(X_{ij}) + Z_{ij} b_i + \omega_i(t_{ij}) + \varepsilon_i \quad \forall i = 1, \ldots, n; j = 1, \ldots, n_i$$

- $f : \mathbb{R}^p \longrightarrow \mathbb{R}$: unknown regression function
  X_{ij} the $p \times 1$ vector of covariates : fixed effects matrix *(gene expressions)*
- $b_i \underset{i.i.d}{\sim} \mathcal{N}(0, B)$: random effects, Z_{ij} the $1 \times q$ random effects covariates
- $(\omega_i(t))_{t \geq 0}$: centered Gaussian process with covariance function $\gamma^2 K_i(s,t)$
- $\varepsilon_{ij} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$: residual error

Capitaine et al. SMMR 2020

Random forest for longitudinal data

Algorithm 1: General estimation procedure for model (1)

Initialization: Let $r = 0$, $\hat{b}_{i,(0)} = 0_q$, $\hat{\omega}_{i,(0)} = 0_{n_i}$, $\hat{B}_{(0)} = I_q$, $\hat{\gamma}^2_{(0)} = 1$ and $\hat{\sigma}^2_{(0)} = 1$.

Repeat

1. Set $r = r + 1$, compute $\tilde{Y}_{ij,(r-1)} = Y_{ij} - Z_{ij}\hat{b}_{i,(r-1)} - \hat{\omega}_{ij,(r-1)}$ estimate f in the standard regression framework (with all N observations):

$$\tilde{Y}_{ij,(r-1)} = f(X_{ij}) + \varepsilon_{ij}$$

   to get $\hat{f}_{i,(r)}$.
   Then predict $\hat{b}_{i,(r)}$ and $\hat{\omega}_{i,(r)}$ using $\hat{B}_{(r-1)}, \hat{\gamma}^2_{(r-1)}, \hat{\sigma}^2_{(r-1)}$ and $\hat{f}_{i,(r)}$.

2. Update $\hat{B}_{(r)}$, $\hat{\gamma}^2_{(r)}$ and $\hat{\sigma}^2_{(r)}$ using $\hat{f}_{i,(r)}$, $\hat{b}_{i,(r)}$ and $\hat{\omega}_{i,(r)}$,

until *convergence*;

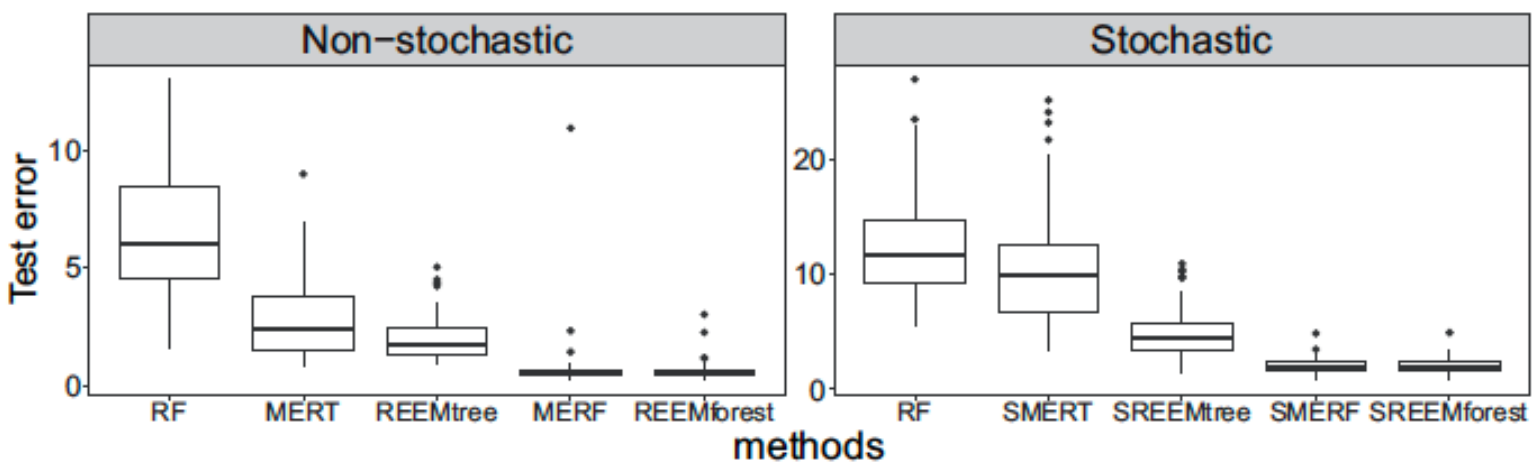
Capitaine et al. SMMR 2020

LongituRF

Random forest for longitudinal data

Table 3. Squared bias of the estimated parameters, averaged on 100 datasets respectively simulated under model (4) and (5) in the high-dimensional case.

	f	B	γ^2	σ^2
Non-stochastic model				
MERT	1.902	0.603	*	0.112
REEMtree	1.543	0.499	*	0.070
MERF	0.750	0.504	*	0.005
REEMforest	0.729	0.493	*	0.005
Stochastic model				
SMERT	5.229	0.926	0.113	0.590
SREEMtree	3.519	0.738	0.071	0.065
SMERF	1.378	0.511	0.024	0.010
SREEMforest	1.367	0.496	0.023	0.011



Application to DALIA-1

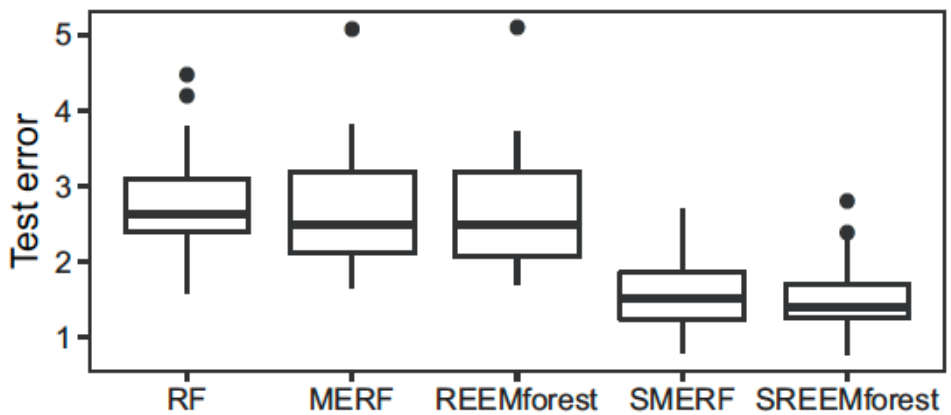
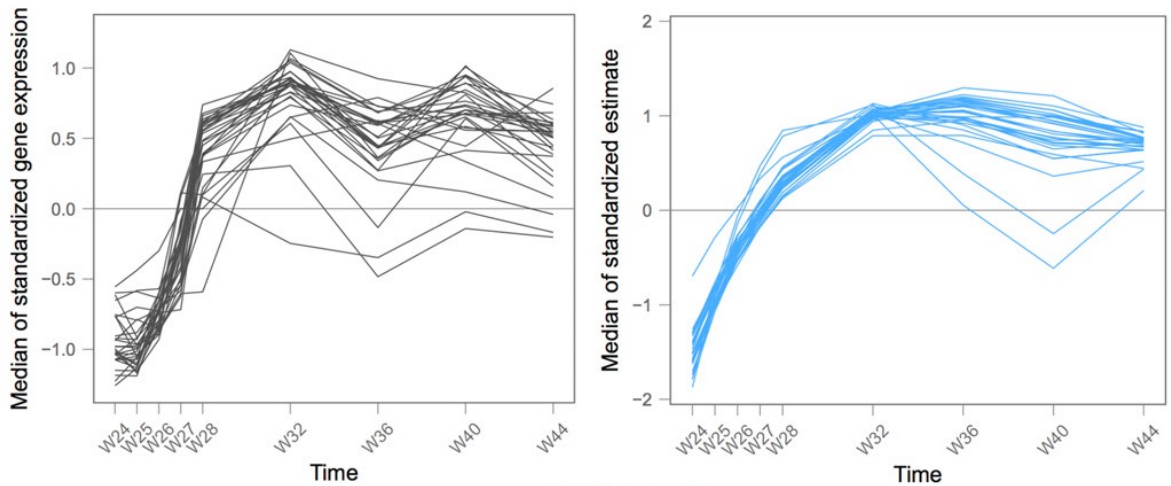


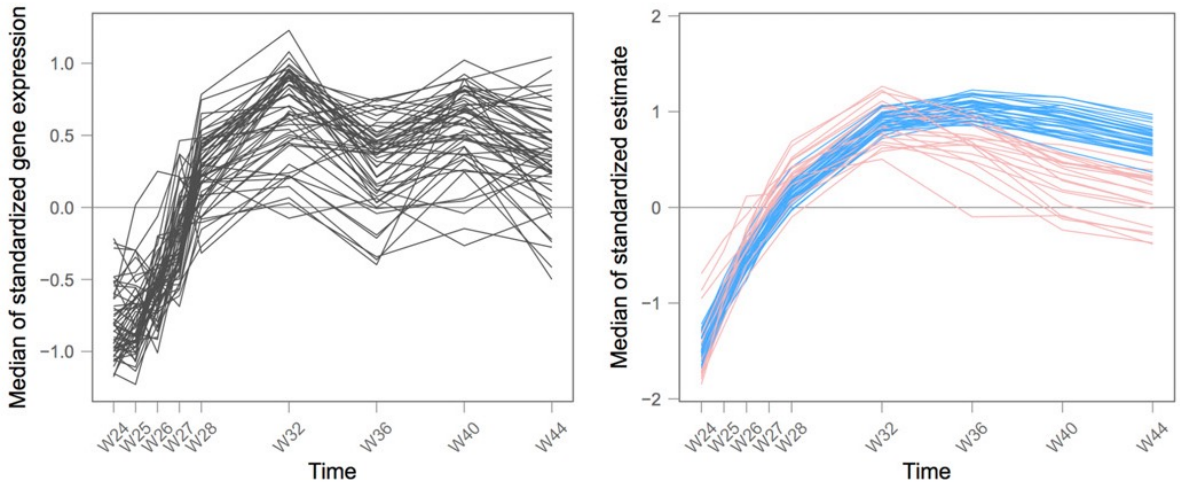
Figure 10. Boxplots of test errors computed using 25 training/test sets random splits, for Breiman's RF, **MERF**, **REEMforest**, **SMERF**, and **SREEMforest**, DALIA trial.

Capitaine et al. SMMR 2020

Application to DALIA-1



M1.2: Interferon

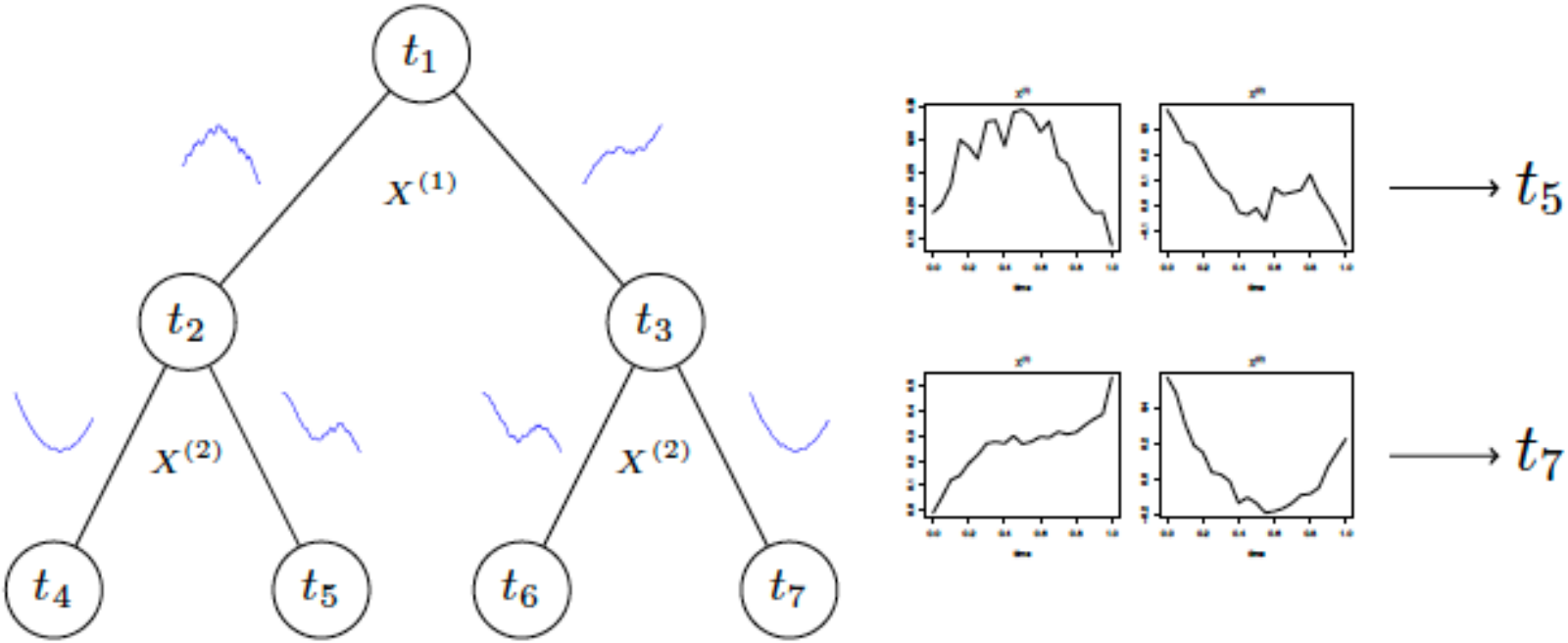


M3.4: Interferon

Capitaine et al. SMMR 2020

Frechet random forest

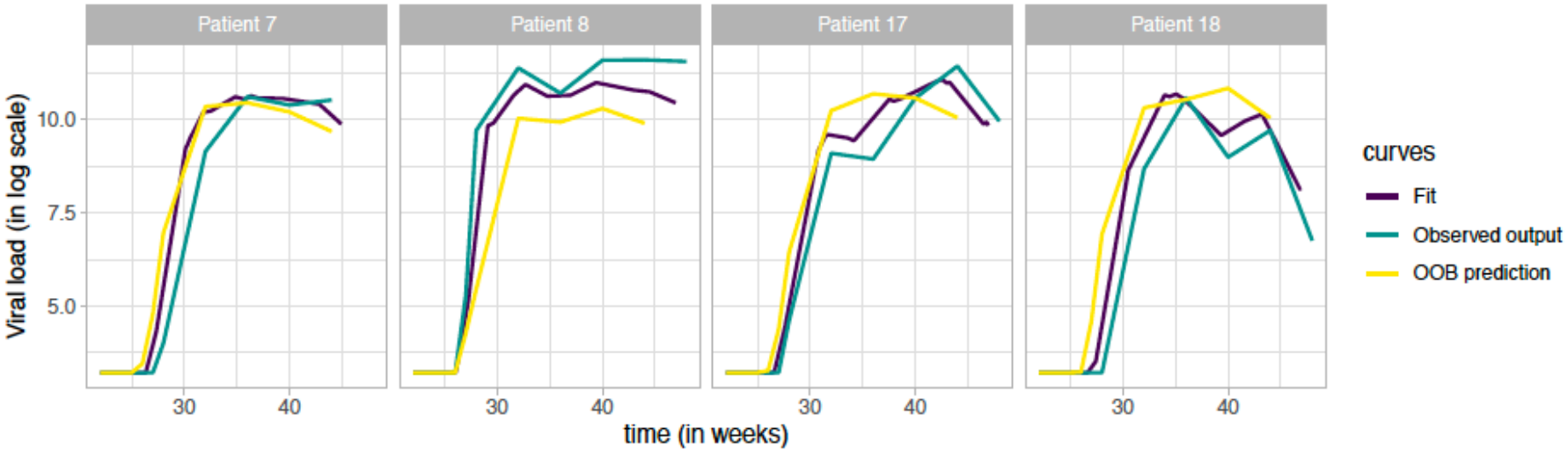
- We use a shape-based distance between curves: the Fréchet distance.
- The split function is the 2-means algorithm adapted to curves (Genolini et.al. 2016).

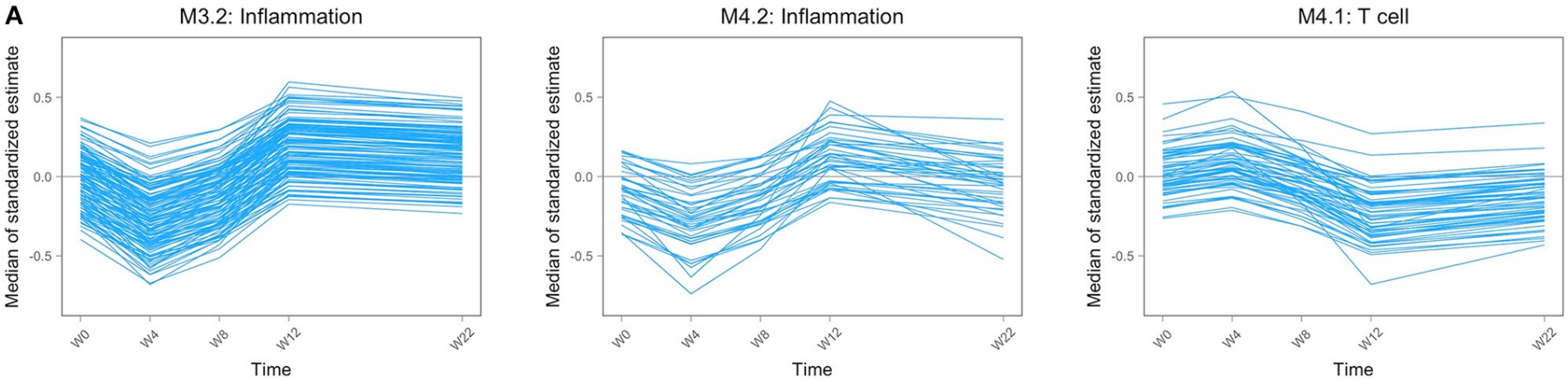


Capitaine et al. https://arxiv.org/abs/1906.01741

Frechforest

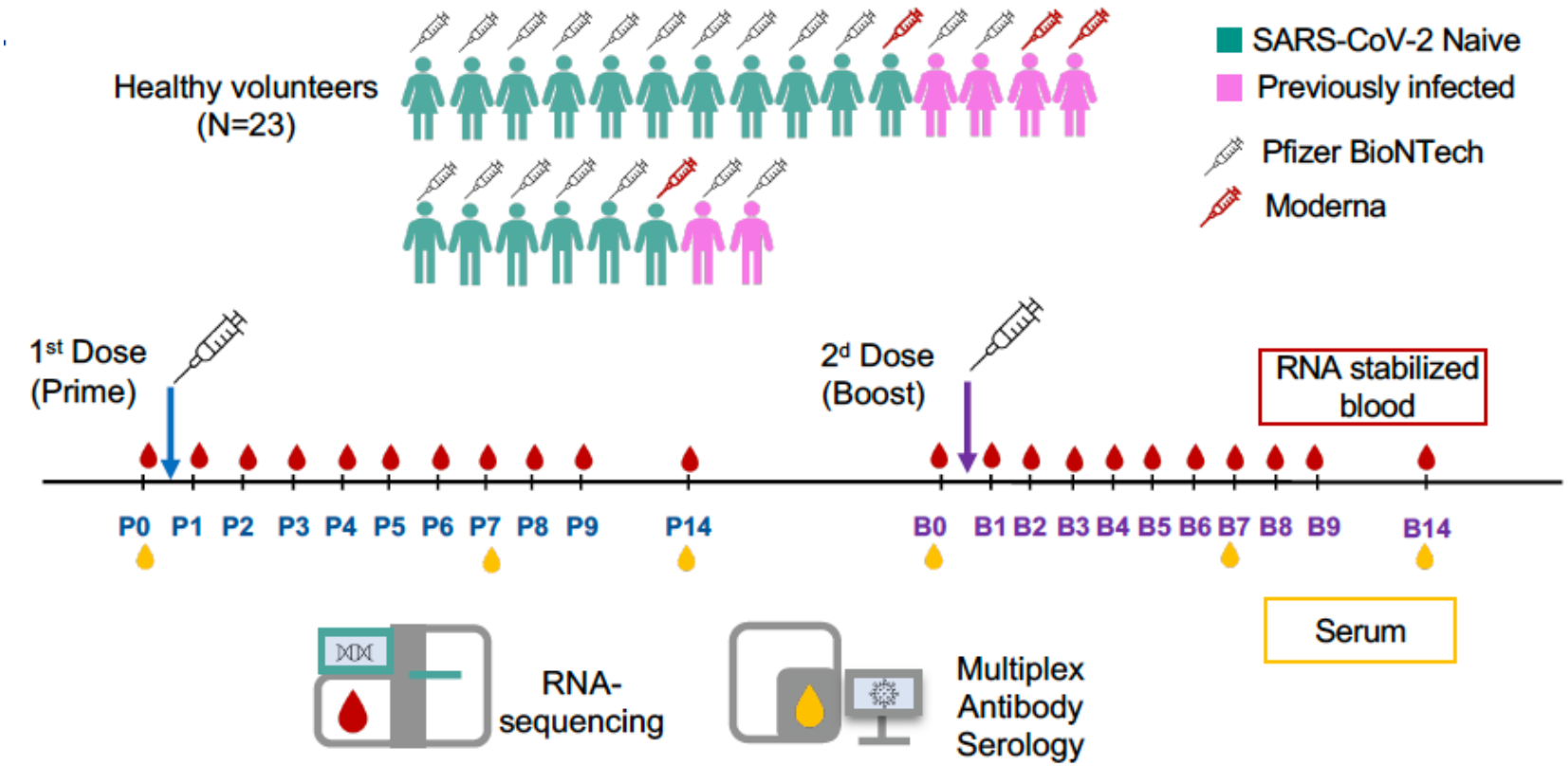
Frechet random forest





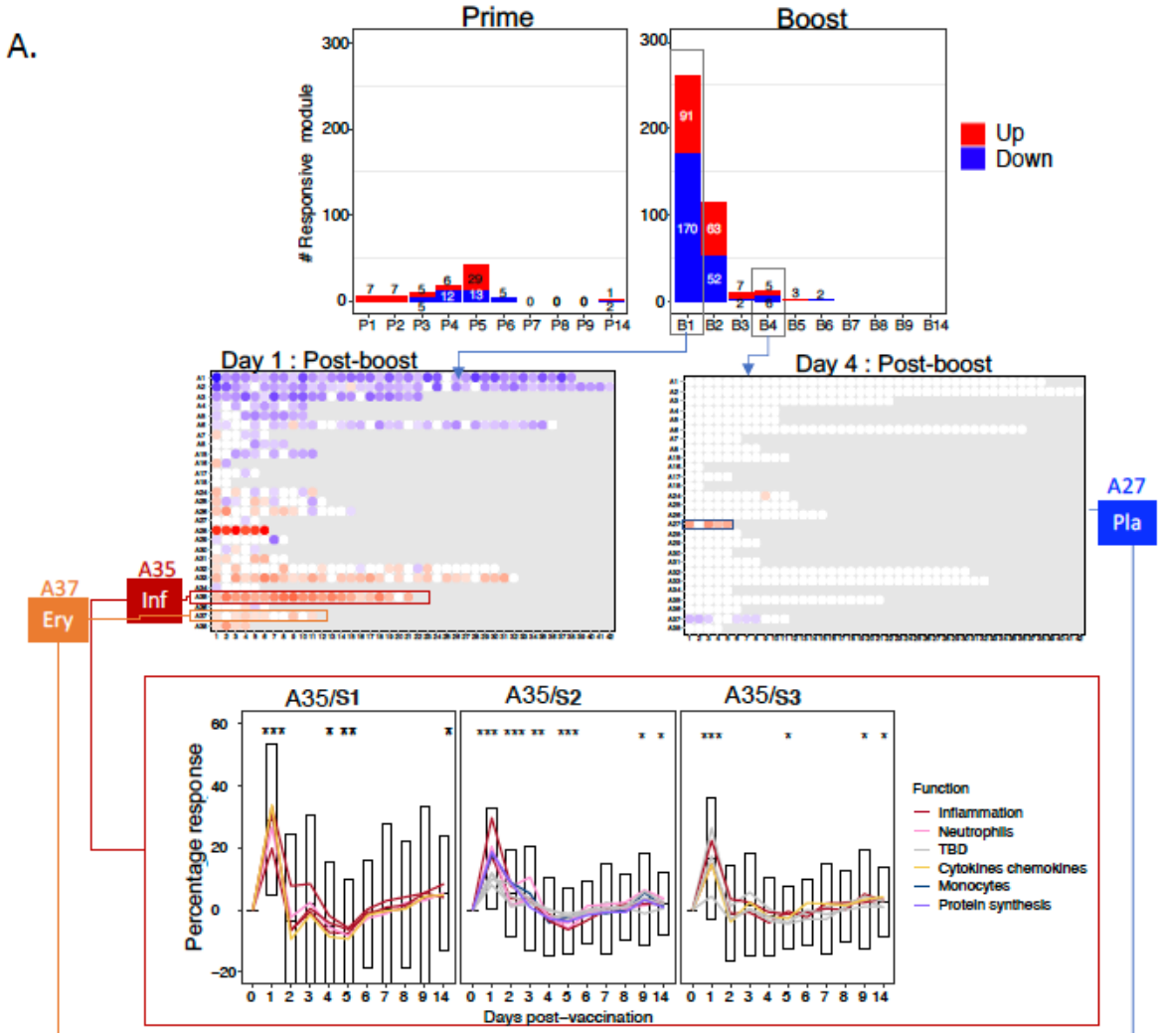
Capitaine et al. https://arxiv.org/abs/1906.01741

Application for COVID19 vaccine



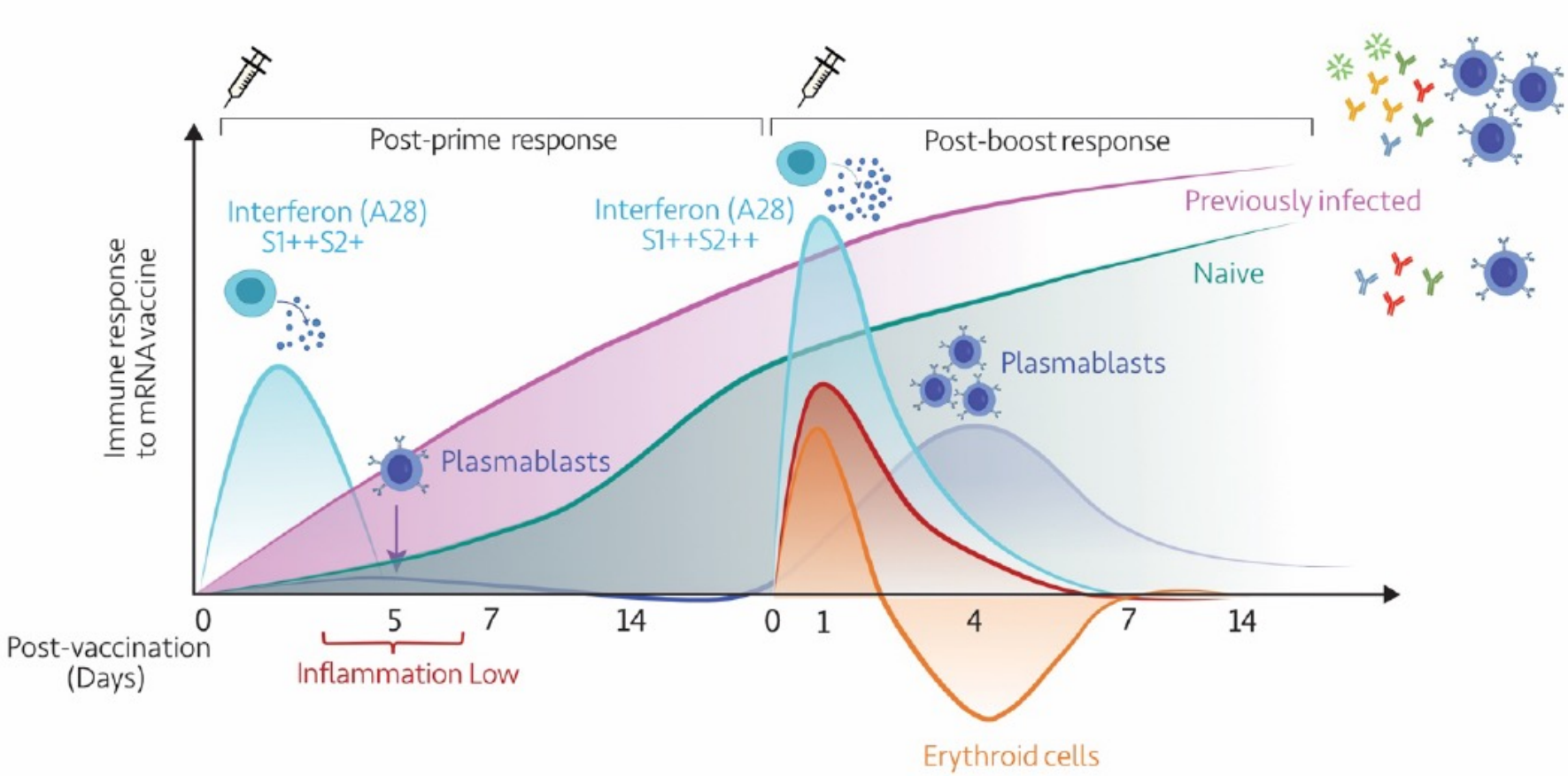
Chaussabel et al. https://www.biorxiv.org/content/10.1101/2021.12.12.472257v2

Application for COVID19 vaccine



Chaussabel et al. https://www.biorxiv.org/content/10.1101/2021.12.12.472257v2

Application for COVID19 vaccine



Chaussabel et al. https://www.biorxiv.org/content/10.1101/2021.12.12.472257v2

Acknowledgments

SISTM team and other collaborators

VACCINE RESEARCH INSTITUTE

EHVA EUROPEAN ALLIANCE HIV VACCINE

eboVAC1

eboVAC2

efpia

imi innovative medicines initiative

http://www.snb2022.paris/

SnB2022
PARIS, 19-21 September

powered by SFdS Biopharmacie & Santé

Home | Program | Committees | Information | Registration | Submission | Sponsors

9th
STATISTICS & BIOPHARMACY CONFERENCE

When Machine Learning meets Statistics
For Drug Development and Evaluation

PARIS
19 | 21
SEPT. 2022

Confirmed invited speakers :

- Gary Collins
- Harald Binder
- Marc Buyse
- Chris Holmes
- Emilie Kaufmann
- Stephen Senn
- Ewout Steyerberg

The sparse Partial Least Squares method

Number of predictors (genes) > number of patients:

PLS + LASSO

Regression method

- Independent linear combinations of predictors, ie **genes**
- Independent linear combinations of explaining variables, ie **immunological measures**
- Components **maximize their covariance**

Penalization method

- Select only a few predictors, i.e. **genes**
- **Sparse**: all non selected predictor Coefficients are estimated **as exactly 0**

*Each **sPLS** component is **sparse:** only a few variables contributes to each components*

Refs: Esposito Vinzi *et al.* (2010), *Handbook of Partial Least Squares* ;
Tibshirani (1996), Regression shrinkage and selection via the lasso. *JRSS-B* ;
Lê Cao,K.-A. *et al.* (2008), A sparse PLS for variable selection when integrating omics data. *Stat. App. In Gen. & Mol. Biol.*

- Aims:
  1. Symmetric situation. Analysis the associations between two blocks of information, analysis focuses on shared information.
  2. Asymmetric situation. **X** matrix= predictors and **Y** matrix= responses variables, analysis focuses on prediction.

- Partial Least Square Family: dimension reduction approaches
  - PLS find pairs of latent vectors $\mathbf{C_X} = \mathbf{Xu}$, $\mathbf{C_Y} = \mathbf{Yv}$ with maximal covariance.

$$e.g., \quad \mathbf{C_X} = u_1 \times SNP_1 + u_2 \times SNP_2 + \ldots + u_p \times SNP_p$$

  - Symmetric situation and Asymmetric situation.
  - Successive matrix decomposition of **X** and **Y** into new latent variables.

PLS

- Output of PLS: K pairs of latent variables $(\mathbf{C_X}^k, \mathbf{C_Y}^k)$, $k = 1, \ldots, K$ with $K << min(p, q)$.
- Reduction method but no variable selection for extracting the most relevant variables from each latent variables.

sparse PLS

- sparse PLS select the relevant SNPs
- Some coefficients u_l are equal to 0
  $C^k = u_1 \times SNP_1 + \underbrace{u_2}_{=0} \times SNP_2 + \underbrace{u_3}_{=0} \times SNP_3 + \ldots + u_p \times SNP_p$
- The sPLS components are linear combinations of the selected variables

Group PLS

Aim: Select group variables taking into account the data structures

- PLS components
  $C^k = u_1 \times X_1 + u_2 \times X_2 + u_3 \times X_3 + \ldots + u_p \times X_p$
- sparse PLS components (sPLS)
  $C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$
- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

$\hookrightarrow$ select group of variables; all the variables within a group are selected otherwise none of them are selected

Sparse Group PLS

Aim: combine both sparsity of groups and within each group.
Example, $\boldsymbol{X}$ matrix= genes, we might be interested in identifying particularly important genes in pathways of interest.

- sparse PLS components (sPLS)

$$C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$$

- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

- sparse group PLS components (sgPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{=0} X_4 + \underbrace{u_5}_{=0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

Optimisation problem

- PLS

$$\min_{||\tilde{\boldsymbol{u}}||_2=1,\tilde{\boldsymbol{v}}} ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^{\mathrm{T}}||_F^2 \qquad (\text{respectively}, \min_{\tilde{\boldsymbol{u}},||\tilde{\boldsymbol{v}}||_2=1} ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^{\mathrm{T}}||_F^2)$$

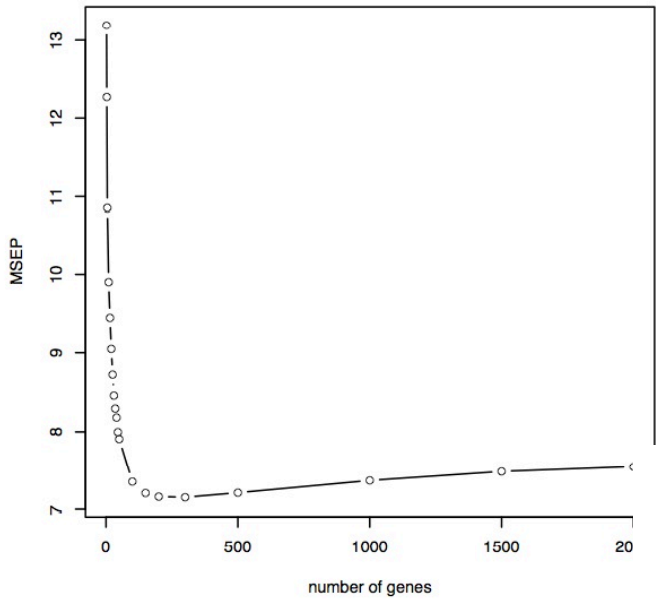
- sPLS

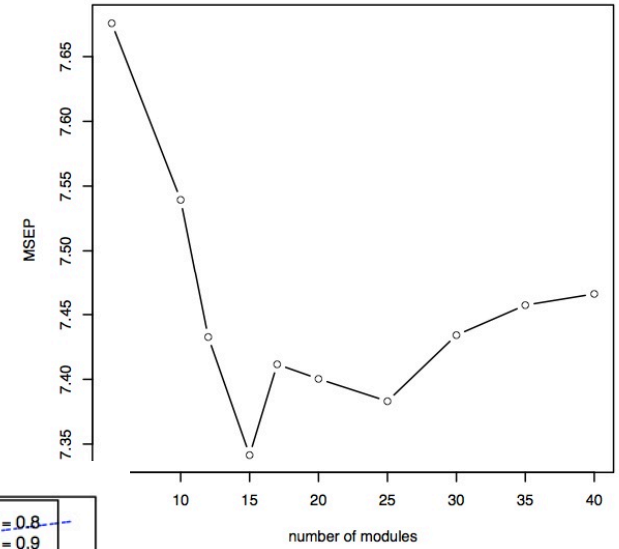
$$\min_{\boldsymbol{u}_h,\boldsymbol{v}_h} ||\boldsymbol{M}_h - \boldsymbol{u}_h\boldsymbol{v}_h^{\mathrm{T}}||_F^2 + P_{\lambda_{1,h}}(\boldsymbol{u}_h) + P_{\lambda_{2,h}}(\boldsymbol{v}_h)$$

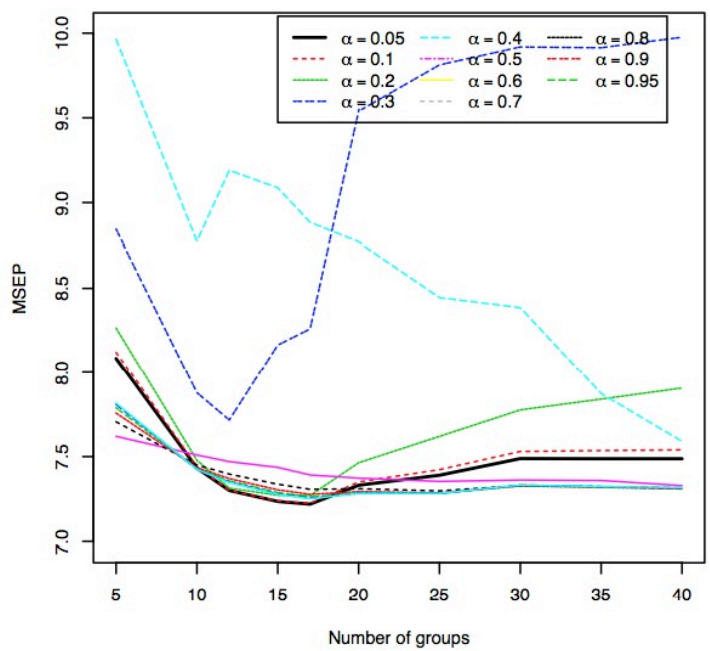
$$P_{\lambda_{1,h}}(\boldsymbol{u}_h) = \sum_{i=1}^{p} 2\lambda_1^h |u_{i,h}|$$

$$P_{\lambda_{2,h}}(\boldsymbol{v}_h) = \sum_{j=1}^{q} 2\lambda_2^h |v_{j,h}|$$

Choice of penalty by 5-fold cross validation of MSEP*







* MSEP: Mean Square Error of Prediction