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

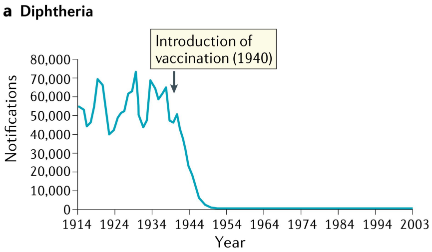
Arthur Hughes[1,2], Denis Agniel[4,5],
Rodolphe Thiébaut[1,2,3], Boris Hejblum[1,2]

[1]University of Bordeaux
[2]Vaccine Research Institute, France
[3]CHU de Bordeaux
[4]Harvard Medical School
[5]RAND Corporation

24th July 2024

Vaccination

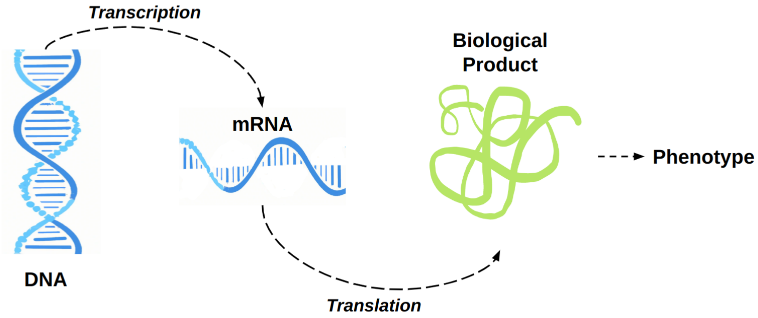
- Most **effective** measure in public health
  - **154 million** deaths prevented in 50 years
- Historically developed empirically



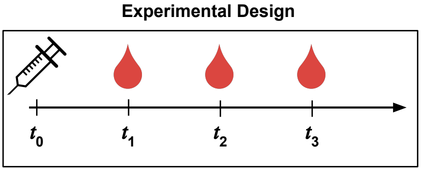
Pollard and Bijker, 2020

Transcriptomics

- **Gene expression**: Genes → Product



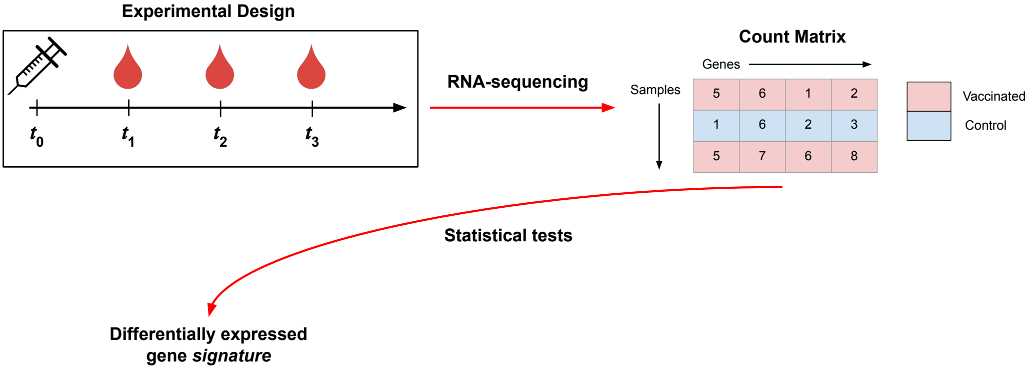
Transcriptomics in vaccinology



Transcriptomics in vaccinology

5	6	1	2
1	6	2	3
5	7	6	8

Transcriptomics in vaccinology



Experimental Design

RNA-sequencing

Count Matrix

5	6	1	2
1	6	2	3
5	7	6	8

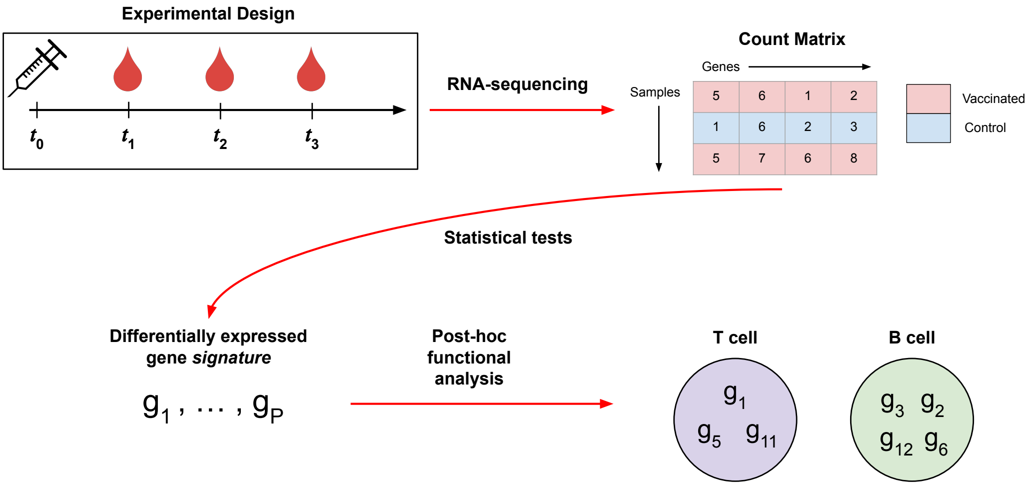
🟥 Vaccinated
🟦 Control

Statistical tests

Differentially expressed gene *signature*

$$g_1, \dots, g_P$$

Transcriptomics in vaccinology



Experimental Design

t_0 t_1 t_2 t_3

RNA-sequencing

Count Matrix

Genes →

Samples ↓

5	6	1	2
1	6	2	3
5	7	6	8

🟥 Vaccinated

🟦 Control

Statistical tests

Differentially expressed gene *signature*

$g_1, \dots, g_P$

Post-hoc functional analysis

T cell

g_1 g_5 g_{11}

B cell

g_3 g_2 g_{12} g_6

Why study gene expression?

Huge potential :

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

But challenges with high-dimensionality...

Why study gene expression?

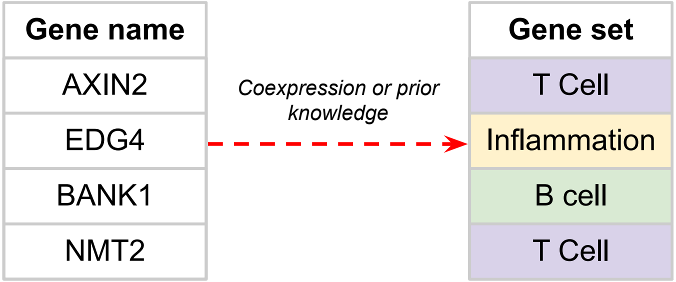
Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

But challenges with high-dimensionality...

- ⚠ Interpretability
- ⚠ Sensitivity to investigator choices
- ⚠ Low signal-noise ratio

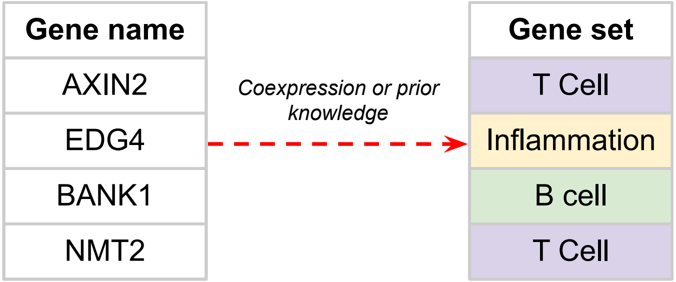
Gene Set Approaches



Gene name	Gene set
AXIN2	T Cell
EDG4	Inflammation
BANK1	B cell
NMT2	T Cell

Coexpression or prior knowledge

Gene Set Approaches

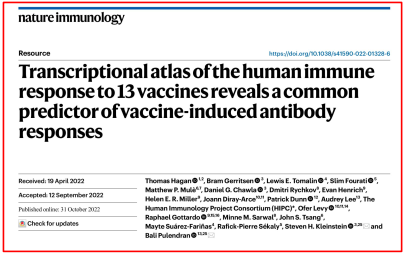


Gene name	Gene set
AXIN2	T Cell
EDG4	Inflammation
BANK1	B cell
NMT2	T Cell

Coexpression or prior knowledge

- ✓ Reduced dimensionality
- ✓ Biological interpretability
- ✓ Boost signal

Comparing vaccine signatures reveals insight



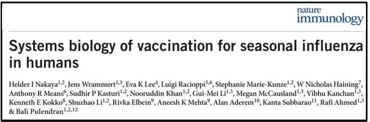
nature immunology

Resource https://doi.org/10.1038/s41590-022-01328-6

Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses

Received: 19 April 2022
Accepted: 12 September 2022

Thomas Hagan[1,2], Bram Gerritsen[3], Lewis E. Tomalin[4], Slim Fourati[5], Matthew P. Mulè[6,7], Daniel G. Chawla[3], Dmitri Rychkov[8], Evan Henrich[9], Helen E. R. Miller[9], Joann Diray-Arce[10,11], Patrick Dunn[12], Audrey Lee[13], The Human Immunology Project Consortium (HIPC)*, Ofer Levy[10,11,14], Raphael Gottardo[9,15,16], Minne M. Sarwal[8], John S. Tsang[6], Mayte Suárez-Fariñas[4], Rafick-Pierre Sékaly[5], Steven H. Kleinstein[3,25] and Bali Pulendran[13,25]



nature immunology

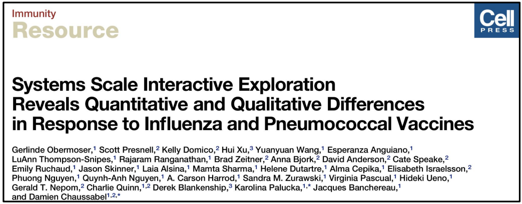
Systems biology of vaccination for seasonal influenza in humans

Helder I Nakaya[1,2], Jens Wrammert[1,3], Eva K Lee[4], Luigi Racioppi[5,6], Stephanie Marie-Kunze[1,2], W Nicholas Haining[7], Anthony R Means[6], Sudhir P Kasturi[1,2], Nooruddin Khan[1,2], Gui-Mei Li[1,3], Megan McCausland[1,3], Vibhu Kanchan[1,3], Kenneth E Kokko[8], Shuzhao Li[1,2], Rivka Elbein[9], Aneesh K Mehta[9], Alan Aderem[10], Kanta Subbarao[11], Rafi Ahmed[1,3] & Bali Pulendran[1,2,12]

nature immunology

Molecular signatures of antibody responses derived from a systems biology study of five human vaccines

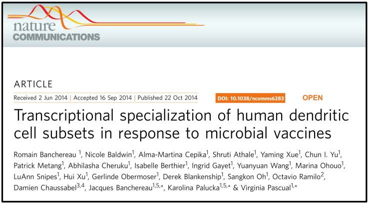
Shuzhao Li[1,2,10], Nadine Rouphael[1,3,10], Sai Duraisingham[1,2,10], Sandra Romero-Steiner[4], Scott Presnell[5,6], Carl Davis[1,7], Daniel S Schmidt[4], Scott E Johnson[4], Andrea Milton[4], Gowrisankar Rajam[4], Sudhir Kasturi[1,2], George M Carlone[4], Charlie Quinn[5,6], Damien Chaussabel[5,6], A Karolina Palucka[6], Mark J Mulligan[1,3,7], Rafi Ahmed[1,8], David S Stephens[1,7], Helder I Nakaya[1,2,9] & Bali Pulendran[1,2,9]



Immunity
Resource
Cell PRESS

Systems Scale Interactive Exploration Reveals Quantitative and Qualitative Differences in Response to Influenza and Pneumococcal Vaccines

Gerlinde Obermoser,[1] Scott Presnell,[2] Kelly Domico,[2] Hui Xu,[3] Yuanyuan Wang,[1] Esperanza Anguiano,[1] LuAnn Thompson-Snipes,[1] Rajaram Ranganathan,[1] Brad Zeitner,[2] Anna Bjork,[2] David Anderson,[2] Cate Speake,[2] Emily Ruchaud,[1] Jason Skinner,[1] Laia Alsina,[1] Mamta Sharma,[1] Helene Dutartre,[1] Alma Cepika,[1] Elisabeth Israelsson,[2] Phuong Nguyen,[1] Quynh-Anh Nguyen,[1] A. Carson Harrod,[1] Sandra M. Zurawski,[1] Virginia Pascual,[1] Hideki Ueno,[1] Gerald T. Nepom,[2] Charlie Quinn,[1,2] Derek Blankenship,[3] Karolina Palucka,[1,*] Jacques Banchereau,[1] and Damien Chaussabel[1,2,*]



nature COMMUNICATIONS

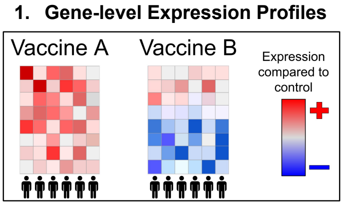
ARTICLE

Received 2 Jun 2014 | Accepted 16 Sep 2014 | Published 22 Oct 2014 DOI: 10.1038/ncomms6283 OPEN

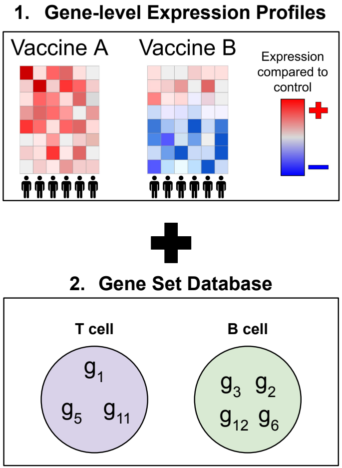
Transcriptional specialization of human dendritic cell subsets in response to microbial vaccines

Romain Banchereau [1], Nicole Baldwin[1], Alma-Martina Cepika[1], Shruti Athale[1], Yaming Xue[1], Chun I. Yu[1], Patrick Metang[1], Abhilasha Cheruku[1], Isabelle Berthier[1], Ingrid Gayet[1], Yuanyuan Wang[1], Marina Ohouo[1], LuAnn Snipes[1], Hui Xu[1], Gerlinde Obermoser[1], Derek Blankenship[1], Sangkon Oh[1], Octavio Ramilo[2], Damien Chaussabel[3,4], Jacques Banchereau[1,5,*], Karolina Palucka[1,5,*] & Virginia Pascual[1,*]

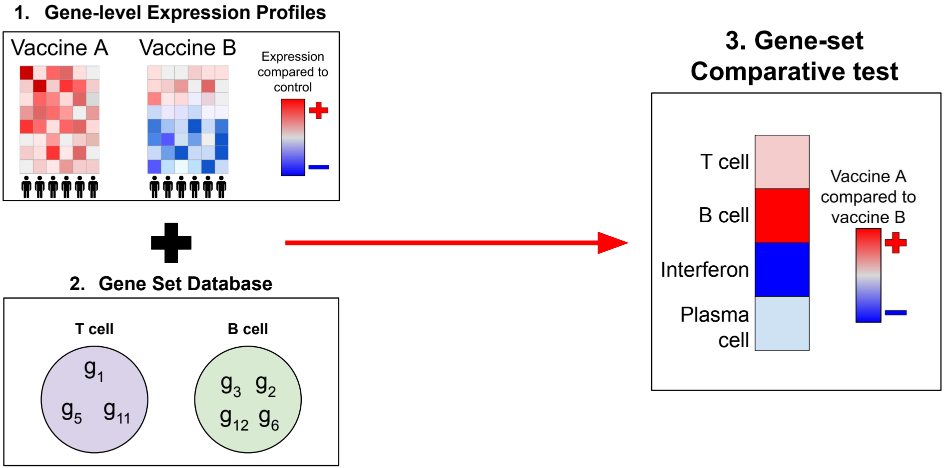
A statistical test for comparing signatures of vaccines



A statistical test for comparing signatures of vaccines



A statistical test for comparing signatures of vaccines



Working linear mixed effects model

Derive test statistic from

$$\boldsymbol{y_i}^G = \boldsymbol{\alpha_0} + \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{\Phi}_i\boldsymbol{\beta} + \boldsymbol{\Phi}_i\boldsymbol{\xi_i} + \boldsymbol{\epsilon_i}$$

- $\boldsymbol{y_i}^G = ((\boldsymbol{y_i}^1)^T, ..., (\boldsymbol{y_i}^p)^T)^T$ - expression of p genes in set G
- $\boldsymbol{X_i}$ - covariates to control for
- $\boldsymbol{\Phi}_i$ - K test variables

Working linear mixed effects model

Derive test statistic from

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

- $\boldsymbol{y_i}^G = ((\boldsymbol{y_i}^1)^T, ..., (\boldsymbol{y_i}^p)^T)^T$ - expression of p genes in set G
- $\boldsymbol{X_i}$ - covariates to control for
- $\boldsymbol{\Phi}_i$ - K test variables

$\Longrightarrow$ $\boldsymbol{\beta}$ - fixed effects of test variables

$\Longrightarrow$ $\boldsymbol{\xi_i} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\xi)$ - random effects of test variables

$\Longrightarrow$ $\boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_i)$ - residual error

Working linear mixed effects model

Derive test statistic from

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

- $\boldsymbol{y_i}^G = ((\boldsymbol{y_i}^1)^T, ..., (\boldsymbol{y_i}^p)^T)^T$ - expression of p genes in set G
- $\boldsymbol{X_i}$ - covariates to control for
- $\boldsymbol{\Phi}_i$ - K test variables

$\Longrightarrow$ $\boldsymbol{\beta}$ - fixed effects of test variables

$\Longrightarrow$ $\boldsymbol{\xi_i} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\xi)$ - random effects of test variables

$\Longrightarrow$ $\boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_i)$ - residual error

Null Hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{0}, \Sigma_\xi = \boldsymbol{0}$$

Test statistic

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

$$\boldsymbol{q}^T = n^{-1/2} \sum_{i=1}^{n} \underbrace{\left(\boldsymbol{y_i}^G - (\boldsymbol{\alpha_0} + \boldsymbol{X_i}\boldsymbol{\alpha})\right)^T}_{\text{Gene-set expression}} \qquad \overbrace{\Sigma_i^{-1}}^{\text{Gene-set Covariance}} \qquad \underbrace{\Phi_i}_{\text{Vaccine indicator}}$$

$\Longrightarrow Q \underset{+\infty}{\sim} \sum_{k=1}^{pK} a_k \chi_1^2$ where a_k is kth eigenvalue of $\mathrm{cov}(\boldsymbol{q})$.

Application : vaccine comparison

Yellow Fever 17D
(live attenuated)

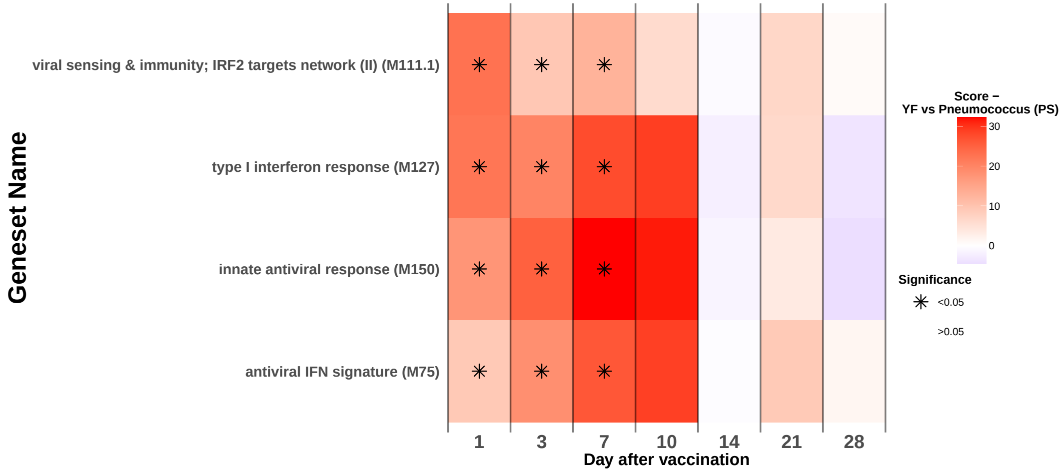
VS

Pneumoccocus
(polysaccharide)

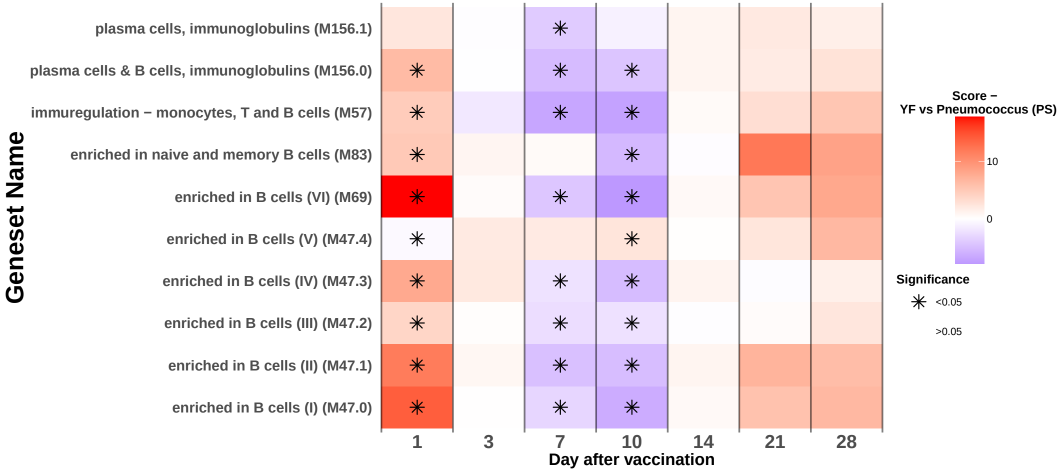
⟹ **What differs between how these vaccines provoke immunity?**

Public Data from *Human Immune Project Consortium* (Diray-Arce et al., 2022)

Strong antiviral signal after Yellow Fever



Distinct dynamics of B-cell response



Lessons learnt : Yellow Fever vs Pneumococcus

- Characterised by **different immune responses**
- Early antiviral response may characterise yellow fever
- Yellow fever induces an early B-cell response

Take-home message

- The **potential and challenges** of gene expression for vaccine development

- **A statistical test for comparing pathways** activated through vaccination

- Illustrated **interpretable differences** between two vaccines

Take-home message

- The **potential and challenges** of gene expression for vaccine development
- **A statistical test for comparing pathways** activated through vaccination
- Illustrated **interpretable differences** between two vaccines

Thank you for listening!

References

- Diray-Arce, Joann et al. (Oct. 2022). "The Immune Signatures data resource, a compendium of systems vaccinology datasets". In: *Scientific Data* 9.1. ISSN: 2052-4463.
- Law, Charity W et al. (2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biology* 15.2, R29. ISSN: 1465-6906.
- Pollard, Andrew J. and Else M. Bijker (Dec. 2020). "A guide to vaccinology: from basic principles to new developments". In: *Nature Reviews Immunology* 21.2, pp. 83–100. ISSN: 1474-1741.

Estimating the mean-variance relationship

⚠ RNA-seq is heteroskedastic $\Longrightarrow$ **Mean-variance relationship non-linear**

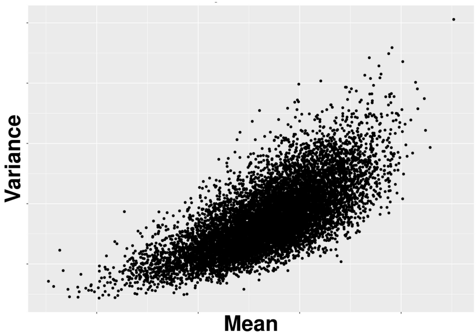


Figure inspired by Law et al., 2014 (simulated data)

Estimating the mean-variance relationship

⚠ RNA-seq is heteroskedastic $\Longrightarrow$ **Mean-variance relationship non-linear**

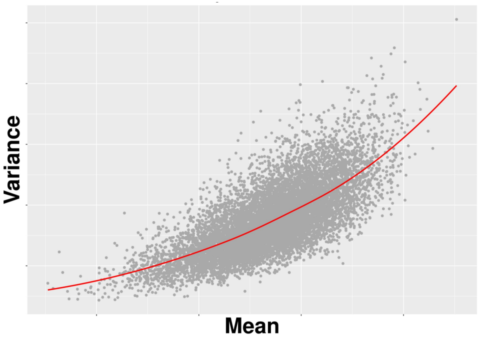


Figure inspired by Law et al., 2014 (simulated data)

Estimating the mean-variance relationship

⚠ RNA-seq is heteroskedastic $\Longrightarrow$ **Mean-variance relationship non-linear**

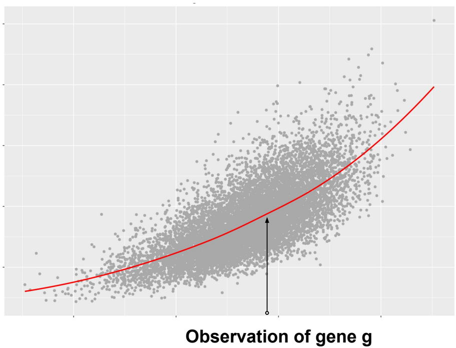


Figure inspired by Law et al., 2014 (simulated data)

Estimating the mean-variance relationship

⚠ RNA-seq is heteroskedastic $\Longrightarrow$ **Mean-variance relationship non-linear**

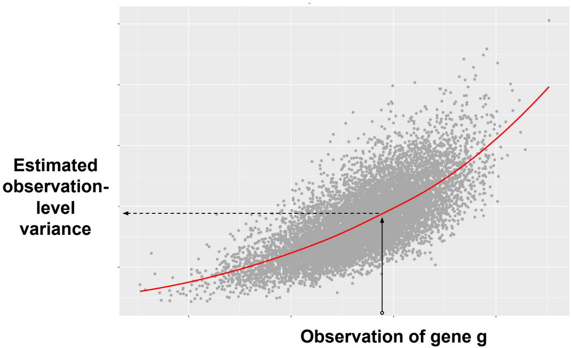


Figure inspired by Law et al., 2014 (simulated data)

Visualisation score

Define gene-level scores $q_i^g = (\boldsymbol{y_i}^g - (\boldsymbol{\alpha_0} + \boldsymbol{x_i}\boldsymbol{\alpha}))\sigma_i^{-1}\phi_i$
where σ_i is the std. deviation of g

$i \in T$ if treated
$i \in C$ if control
$g = 1, ..., p$ genes in a set G
Define the *visualisation score* :

$$d = \frac{1}{p}\sum_{g=1}^{p}\{\frac{1}{|T|}\sum_{i\in T} q_i^g - \frac{1}{|C|}\sum_{i\in C} q_i^g\}$$

i.e. the **average difference in scores** between the two groups.

Human Immune Project Consortium

- **Public resources** to characterise human immune system
- 30 studies comprising 24 vaccines developed against 11 pathogens
  - **13 distinct vaccine types** (pathogen + vaccine platform)
- **4975 transcriptomic samples from 1405 individuals**
  - 10086 common genes across all studies

Diray-Arce et al., 2022

Limitations

- Typically choose Σ_i diagonal i.e. do not model within-gene set correlations
- Visualisation shows mean differences but there may be multiple trends within a gene set
- Test underpowered when model grossly misspecified