

Correcting unwanted variation in RNA sequencing data derived from a multi-centre study of leukemia

Anna Quaglieri
PhD Student

RSS 2018 International Conference 2018

A detailed illustration of a blood vessel interior. The background is a deep red, wavy surface representing the vessel wall. Numerous red blood cells, depicted as biconcave discs, are scattered throughout the scene. Several white blood cells are also present; they are larger than the red blood cells and have a distinct, textured, spherical appearance. The lighting is soft, creating a sense of depth and highlighting the various cellular components.

Leukemia:
Cancer of the blood

CBF-AML

CBF-AML

Diagnosis

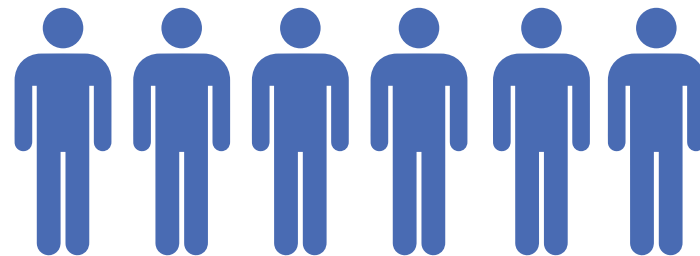


Chemotherapy



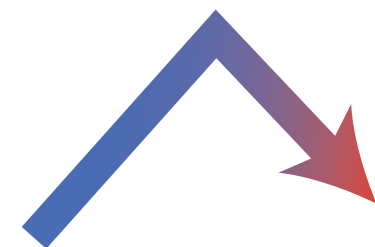
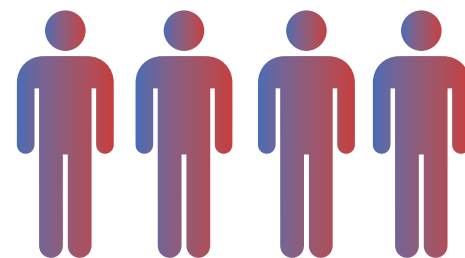
60%

Long term remission



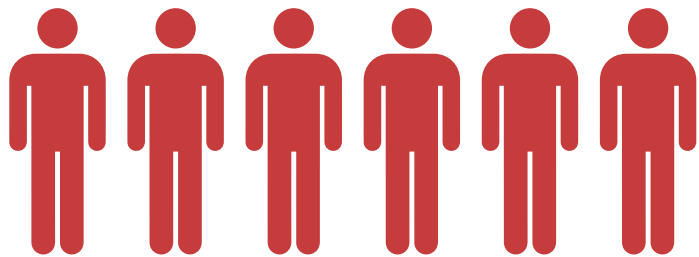
40%

Relapse

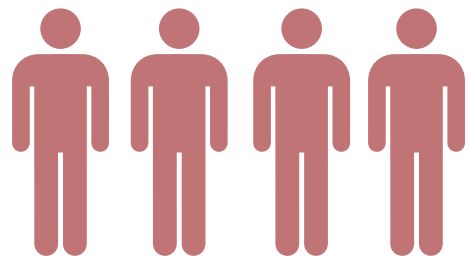


Diagnosis

Long Remission (LR)



Relapse (R)

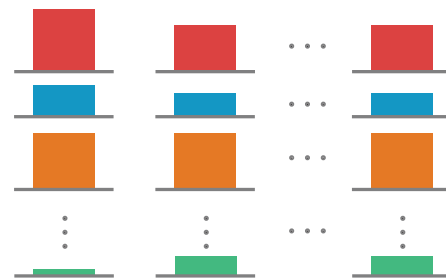


Diagnosis

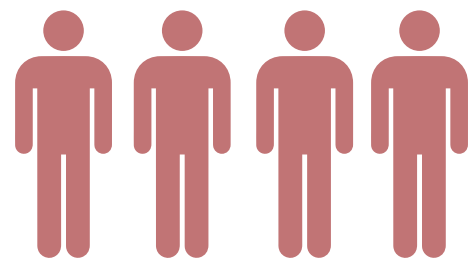
Long Remission (LR)



Gene expression



Relapse (R)



What **genes** are associated with different **outcomes**?

Diagnosis

Long Remission (LR)



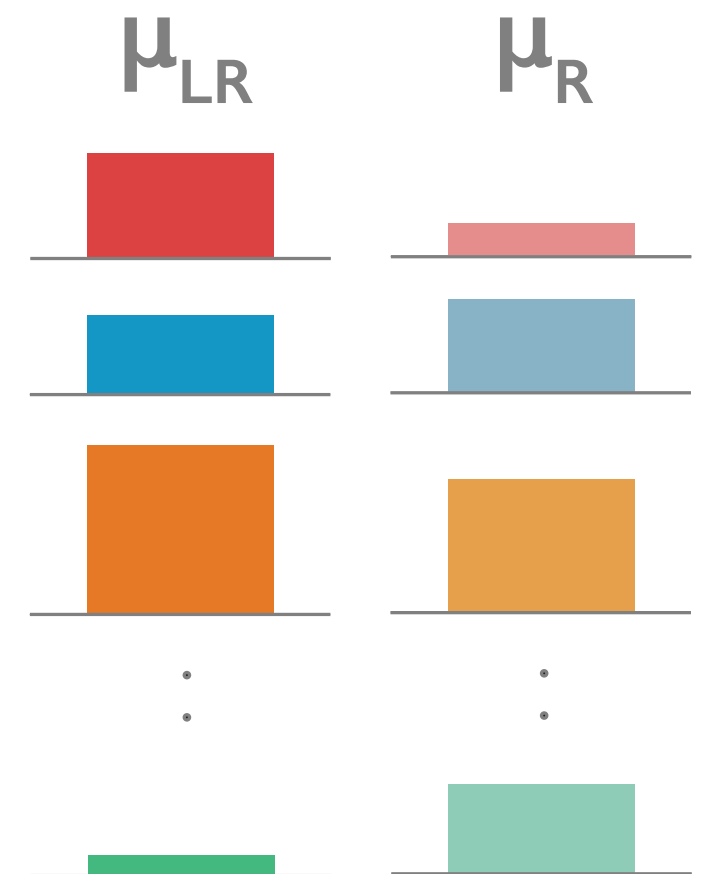
Relapse (R)



 Gene expression



Compare group means



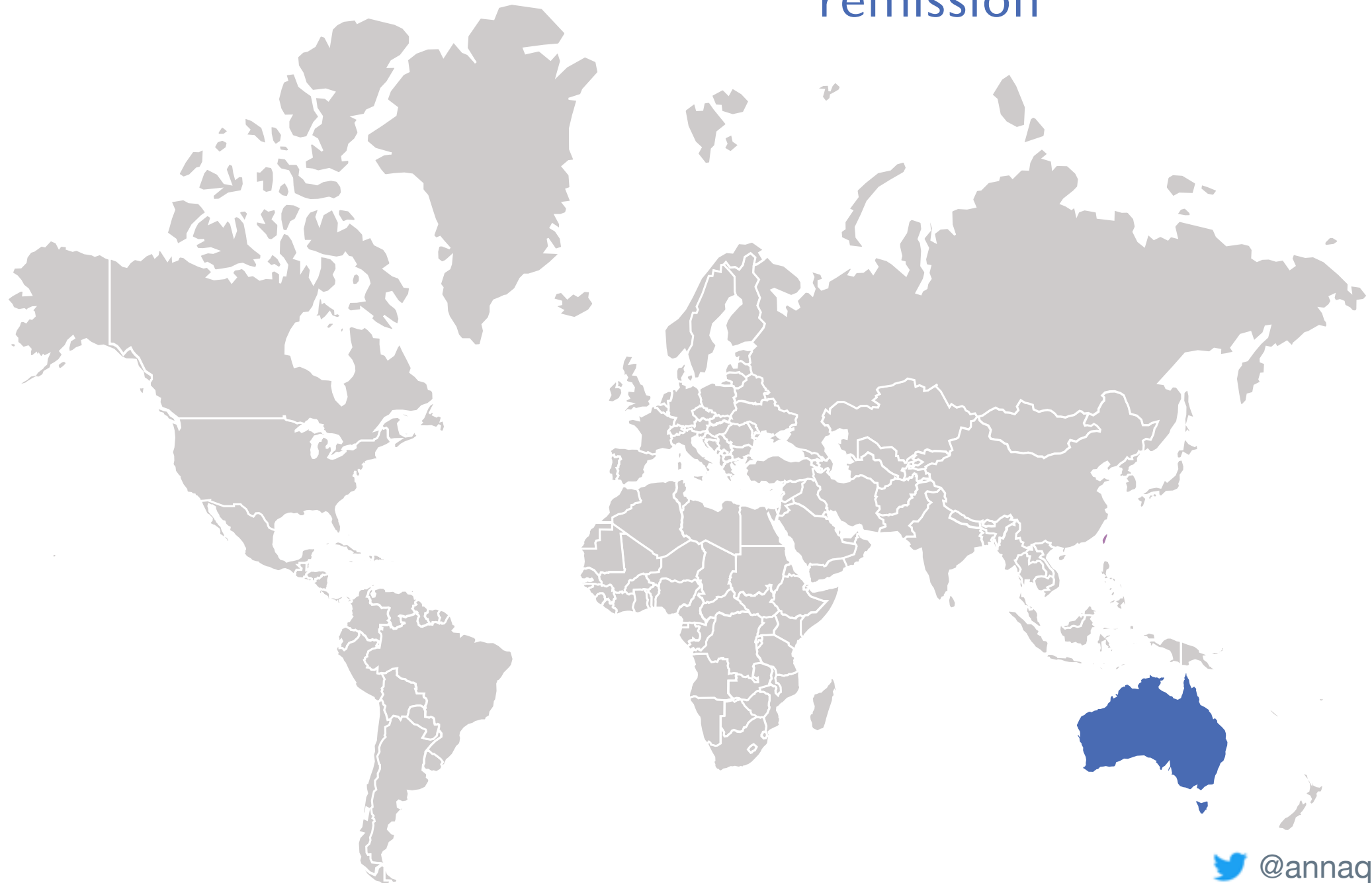
Our initial CBF-AML cohort

11

Long term
remission

8

Relapse



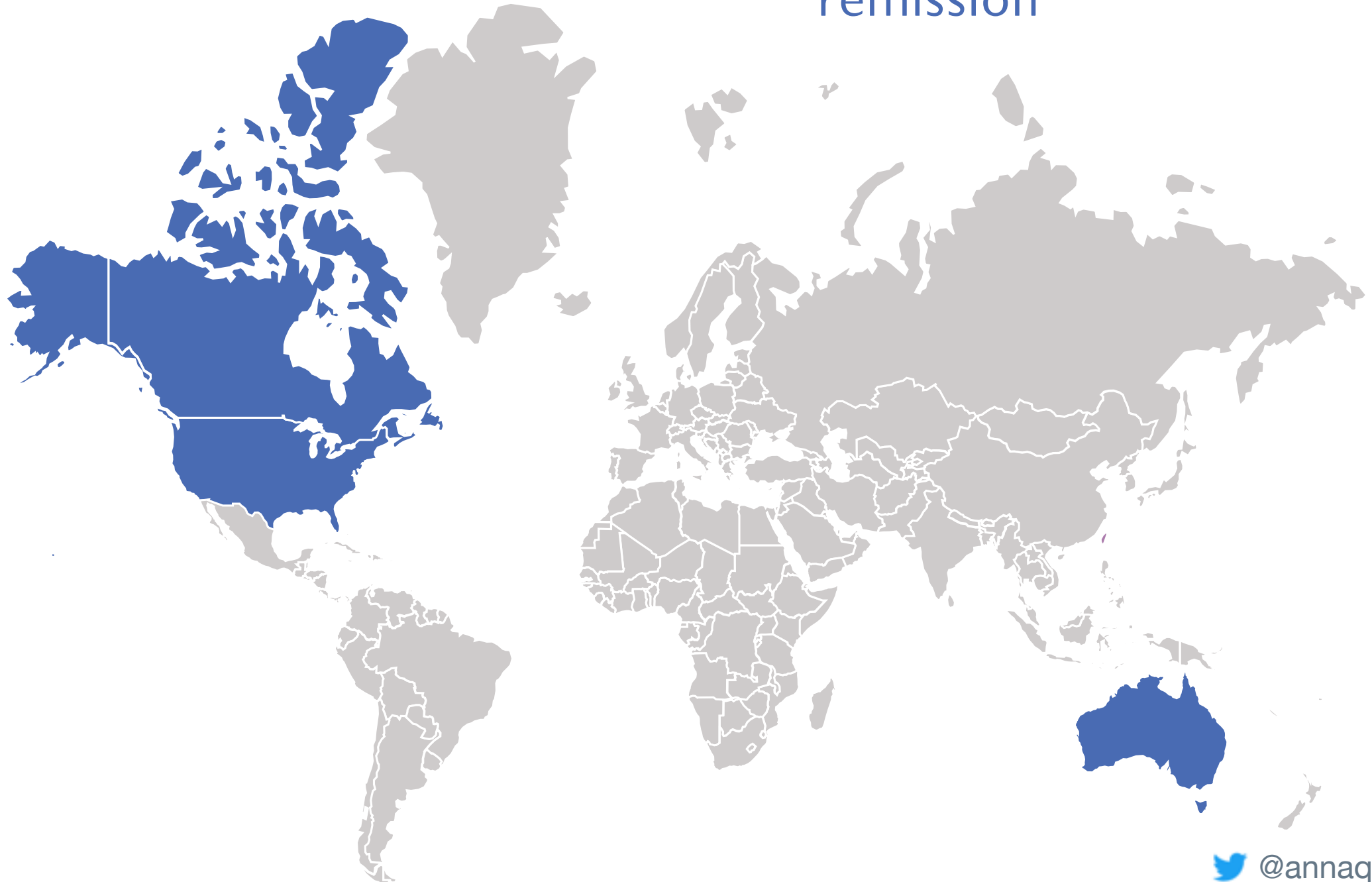
Combined CBF-AML cohort

79

Long term
remission

31

Relapse



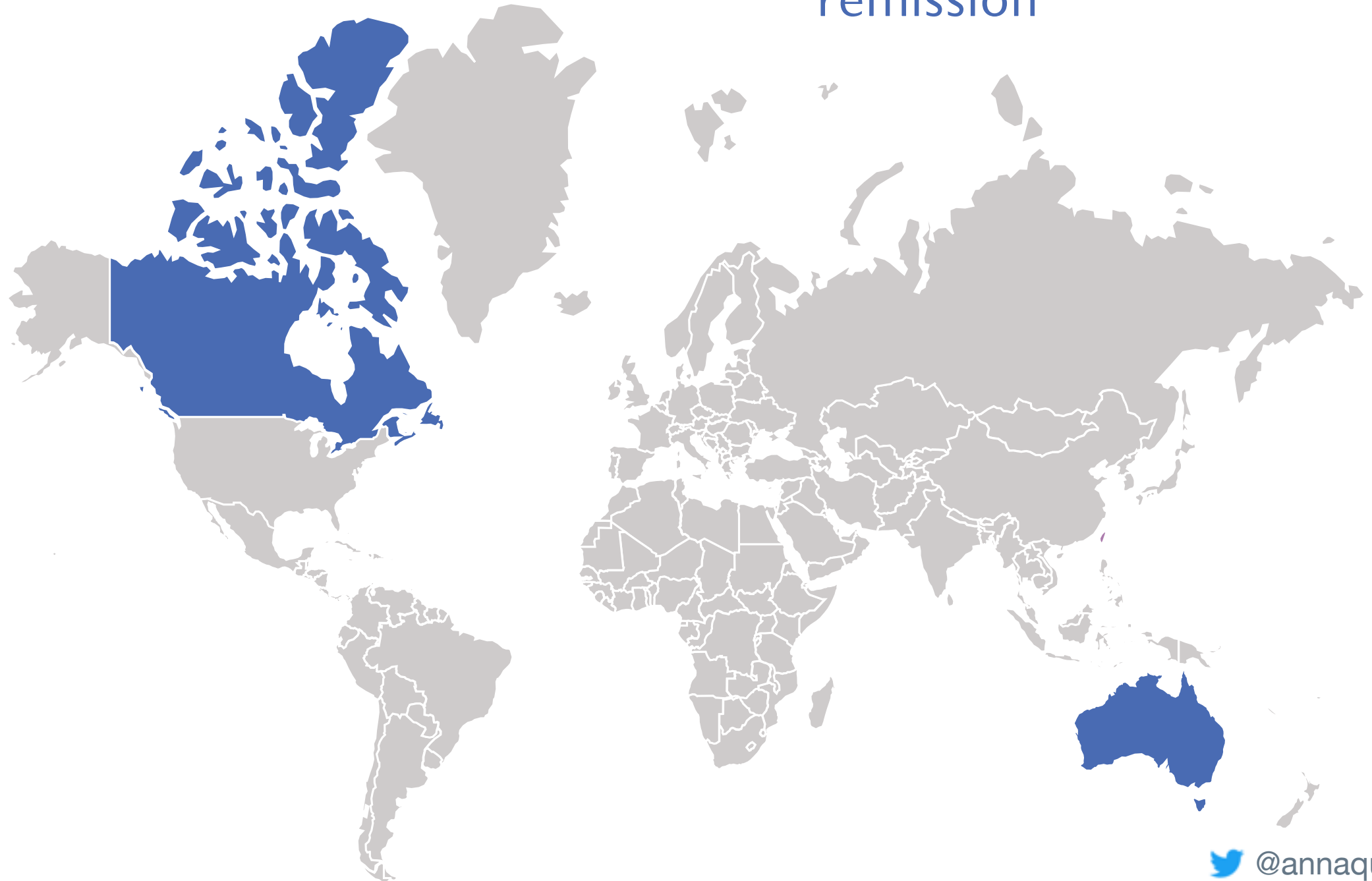
Used so far...

45

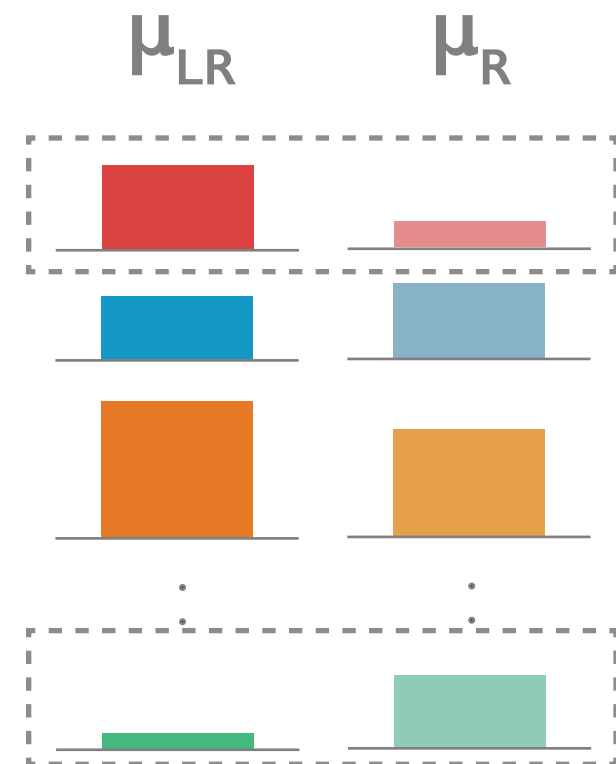
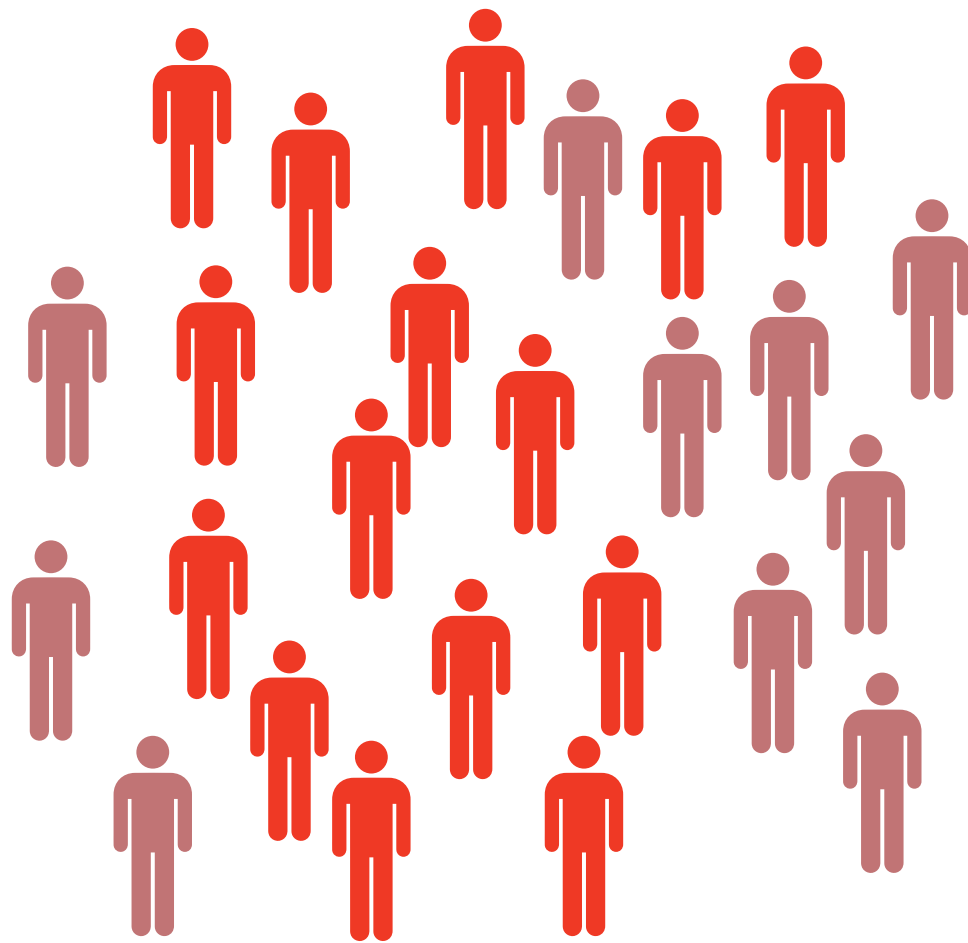
Long term
remission

20

Relapse

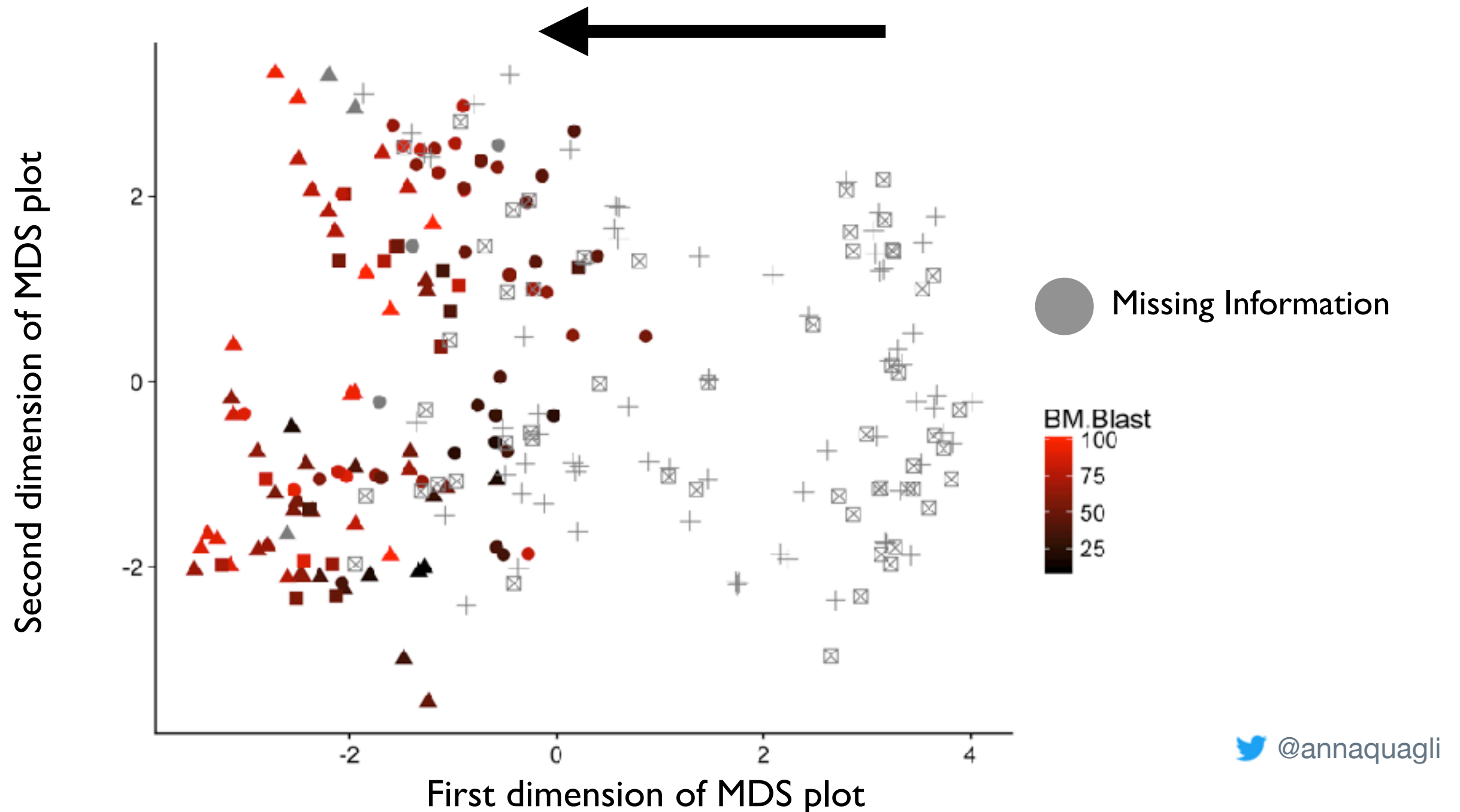


Larger sample can give more power to detect signal



Larger sample also increases unwanted heterogeneity

Higher burden of Leukemia



That's why we needed RUV!

Removing Unwanted Variation

JA. Gagnon-Bartsch, L. Jacob, T. Speed



```
library(ruv)  
ruv::RUV4
```

$$Y_{m \times n} = X_{m \times p} \beta_{p \times n} + W_{m \times k} \alpha_{k \times n} + \epsilon_{m \times n}$$

↙
Log2 gene expression matrix
m samples
n genes

Model estimated with RUV-4

Gene-wise comparison of *Relapse* vs *Long Remission* patients

$$Y_{mxn} = X_{mxp}\beta_{pxn} + W_{mxk}\alpha_{kxn} + \epsilon_{mxn}$$

Model estimated with RUV-4

Matrix with Unwanted Variation

$$Y_{mxn} = X_{mxp}\beta_{pxn} + W_{mxk}\alpha_{kxn} + \epsilon_{mxn}$$

Model estimated with RUV-4

Estimated using Negative Control genes

$$Y_{mxn} = X_{mxp}\beta_{pxn} + W_{mxk}\alpha_{kxn} + \epsilon_{mxn}$$

The diagram illustrates the RUV-4 model structure using matrix blocks. It shows the equation $Y_{mxn} = X_{mxp}\beta_{pxn} + W_{mxk}\alpha_{kxn} + \epsilon_{mxn}$ represented by blue rectangular blocks. The first block is a large rectangle labeled 'Y'. To its right is a smaller rectangle labeled 'Y_{NC}', which is highlighted with a red border. This is followed by an equals sign. Then is a block labeled 'X', followed by a block labeled 'β'. To the right of 'β' is a smaller rectangle labeled 'β_{NC}', which is also highlighted with a red border. This is followed by a plus sign, then a block labeled 'W', followed by a block labeled 'α', followed by another plus sign, and finally a block labeled 'ε'.

$$\beta_{NC} = 0$$

Consistent with published results!

- We could remove unwanted variations even where information was not available

Found consistency with recently published results

- High expression of **DOCK1** confers poor prognosis in AML (Lee Sh (2017))
- “...This suggests that **PID1** may contribute to responsiveness to chemotherapy.” Xu J (2017), Scientific Reports



Looking to clean your data?

Check the RUV tutorial that we gave at useR! 2018

[Looking to clean your data? Learn how to Remove Unwanted Variation with R - Part 1](#)

[Looking to clean your data? Learn how to Remove Unwanted Variation with R - Part 2](#)

Thanks to...

Terry Speed

Ian Majewski

Edward Chew



Australian Government



THE UNIVERSITY OF
MELBOURNE