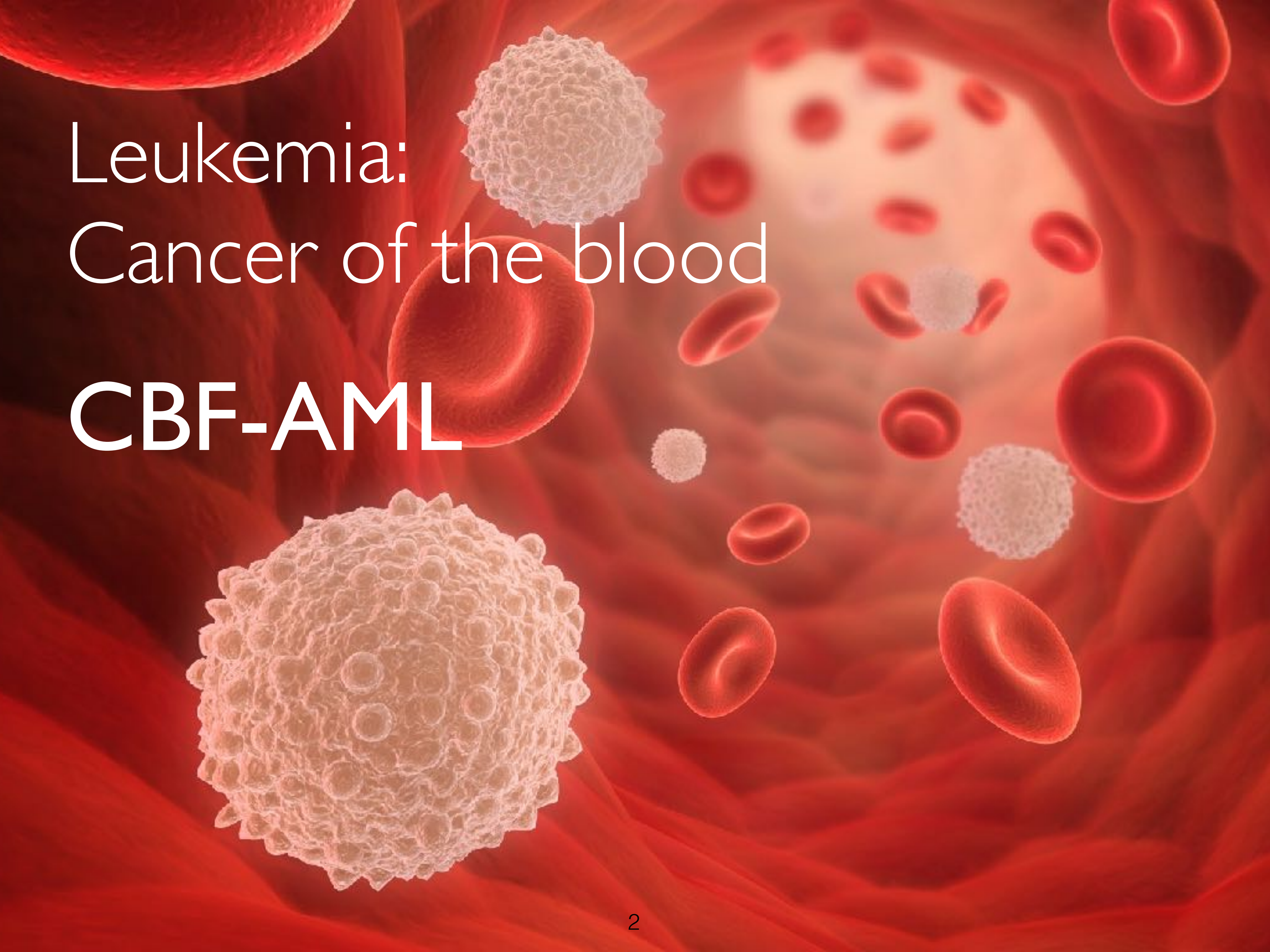
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

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

Anna Quaglieri
PhD Student

RSS 2018 International Conference 2018



Leukemia: Cancer of the blood

CBF-AML

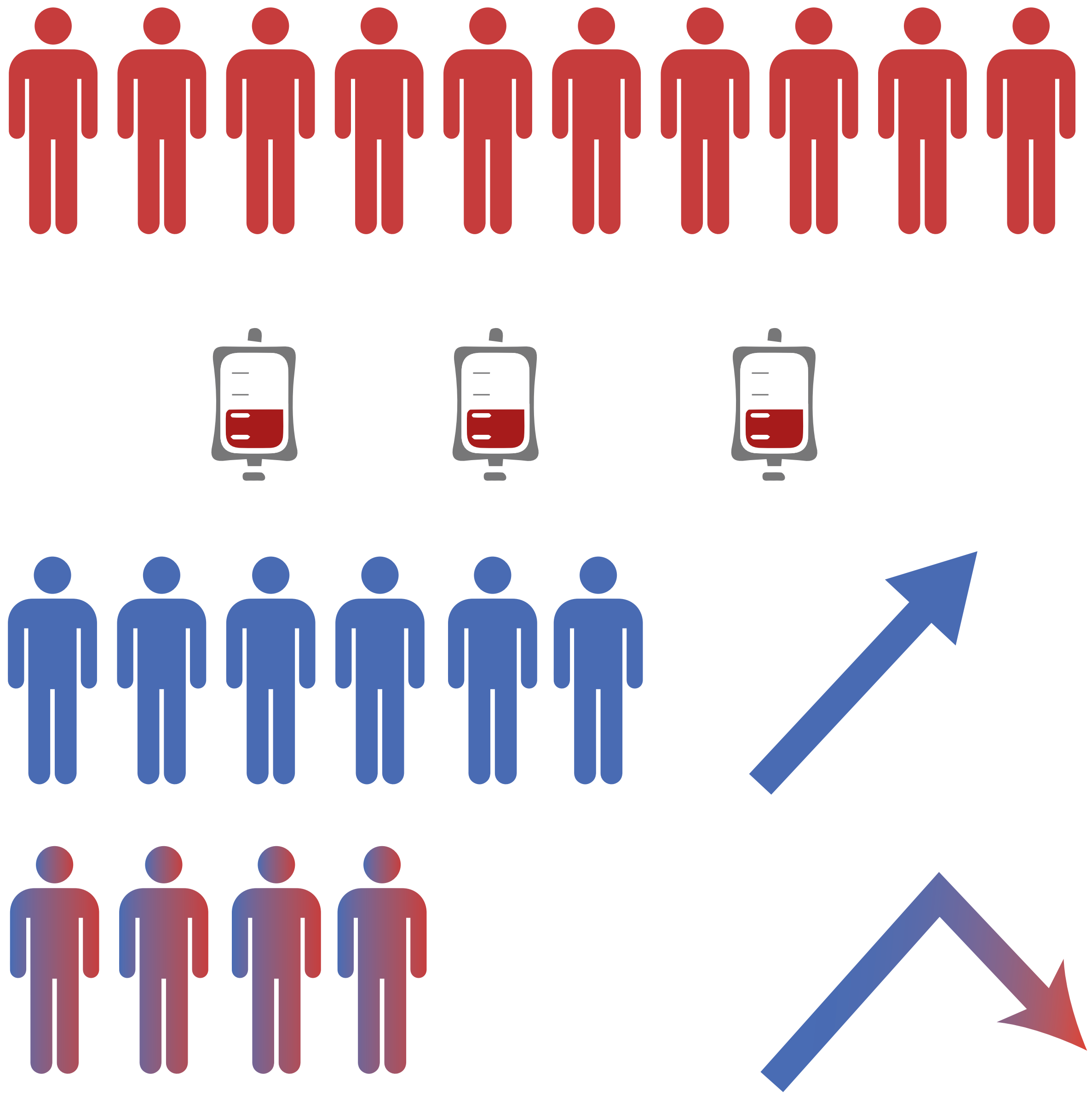
CBF-AML

Diagnosis

Chemotherapy

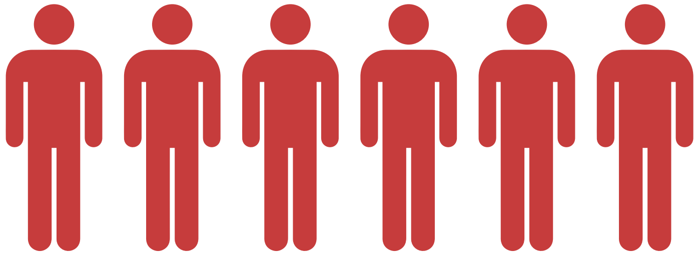
60%
Long term remission

40%
Relapse

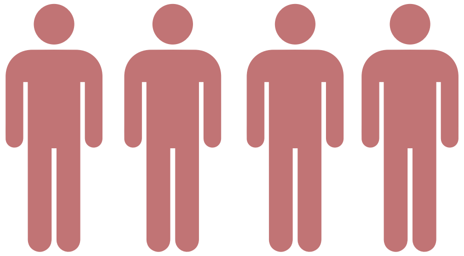


Diagnosis

Long Remission (LR)



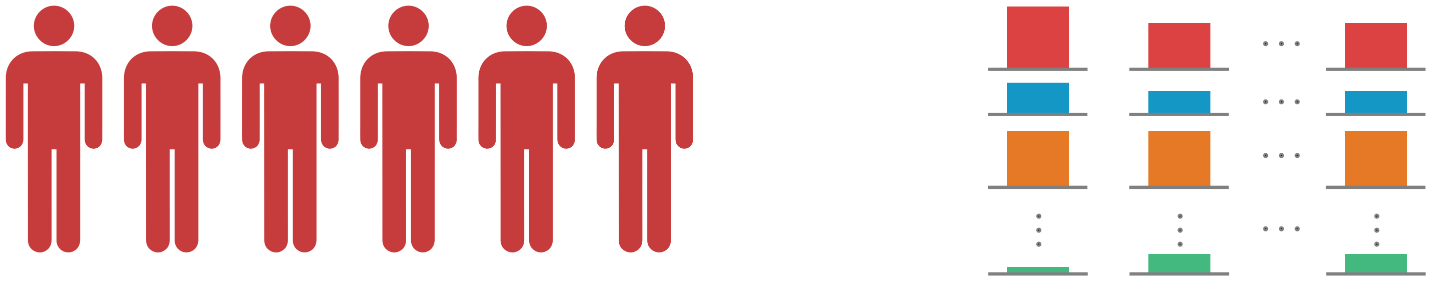
Relapse (R)



Diagnosis

Long Remission (LR)

Gene expression

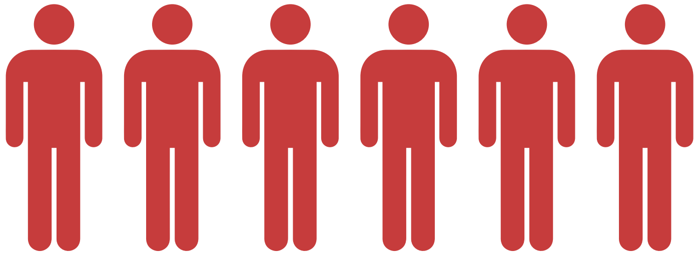


Relapse (R)

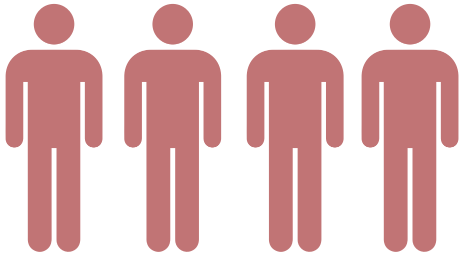
What **genes** are associated with different **outcomes**?

Diagnosis

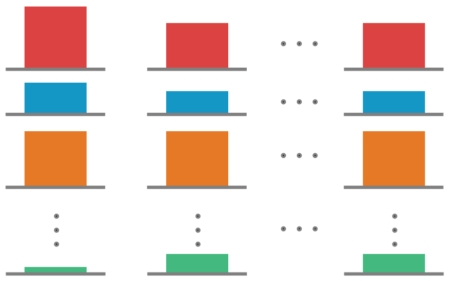
Long Remission (LR)

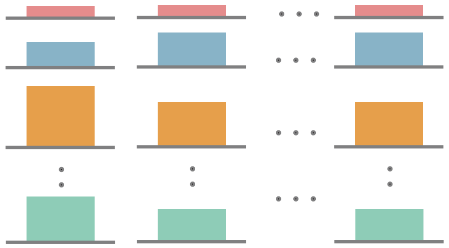


Relapse (R)

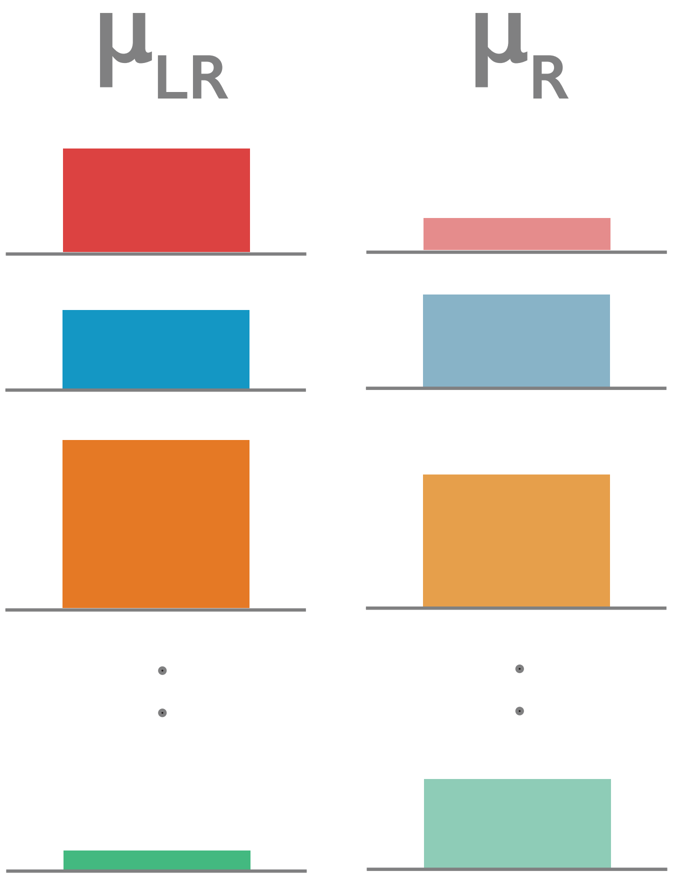


Gene expression





Compare group means



@annaquagli

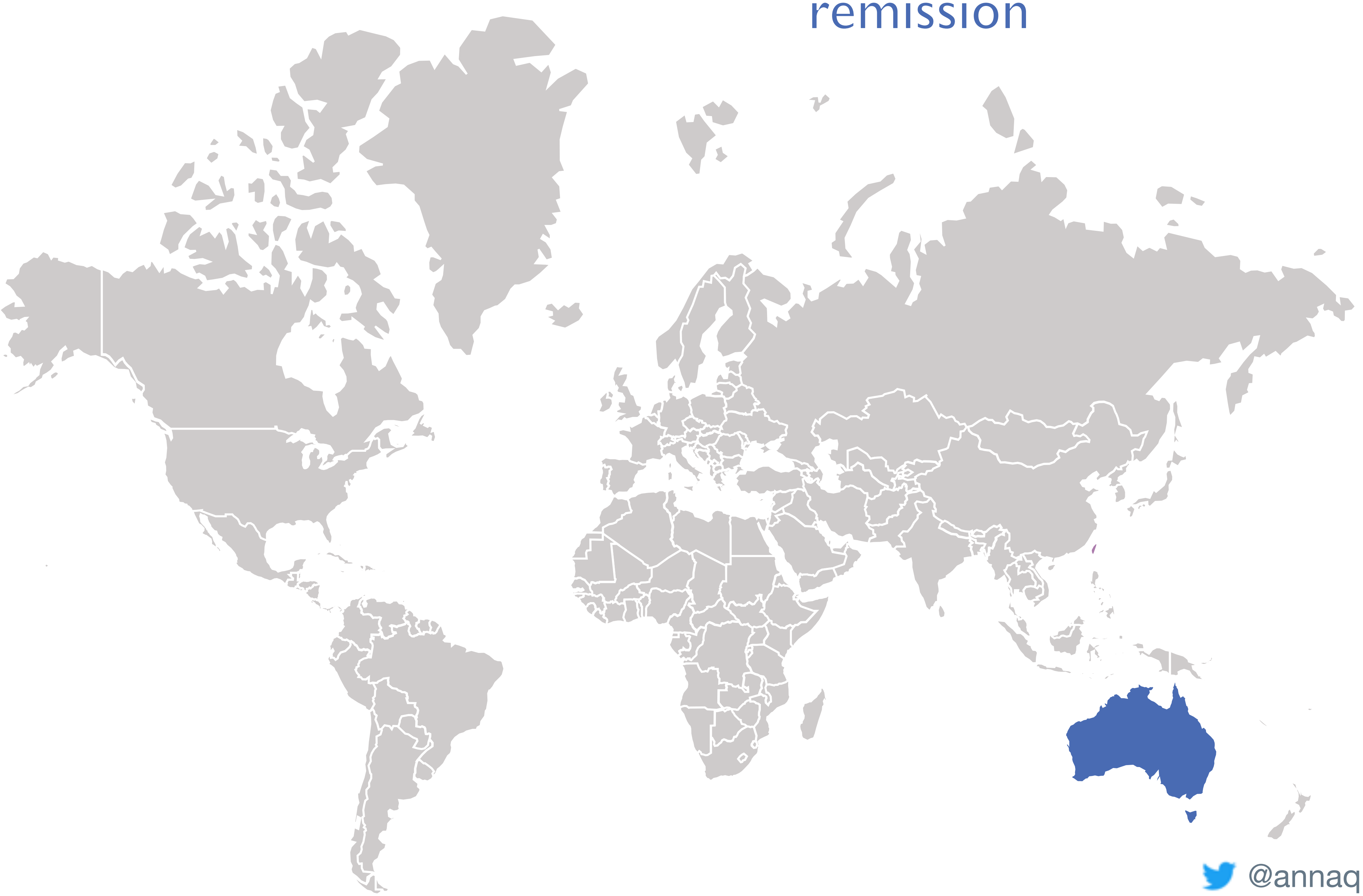
Our initial CBF-AML cohort

11

Long term remission

8

Relapse

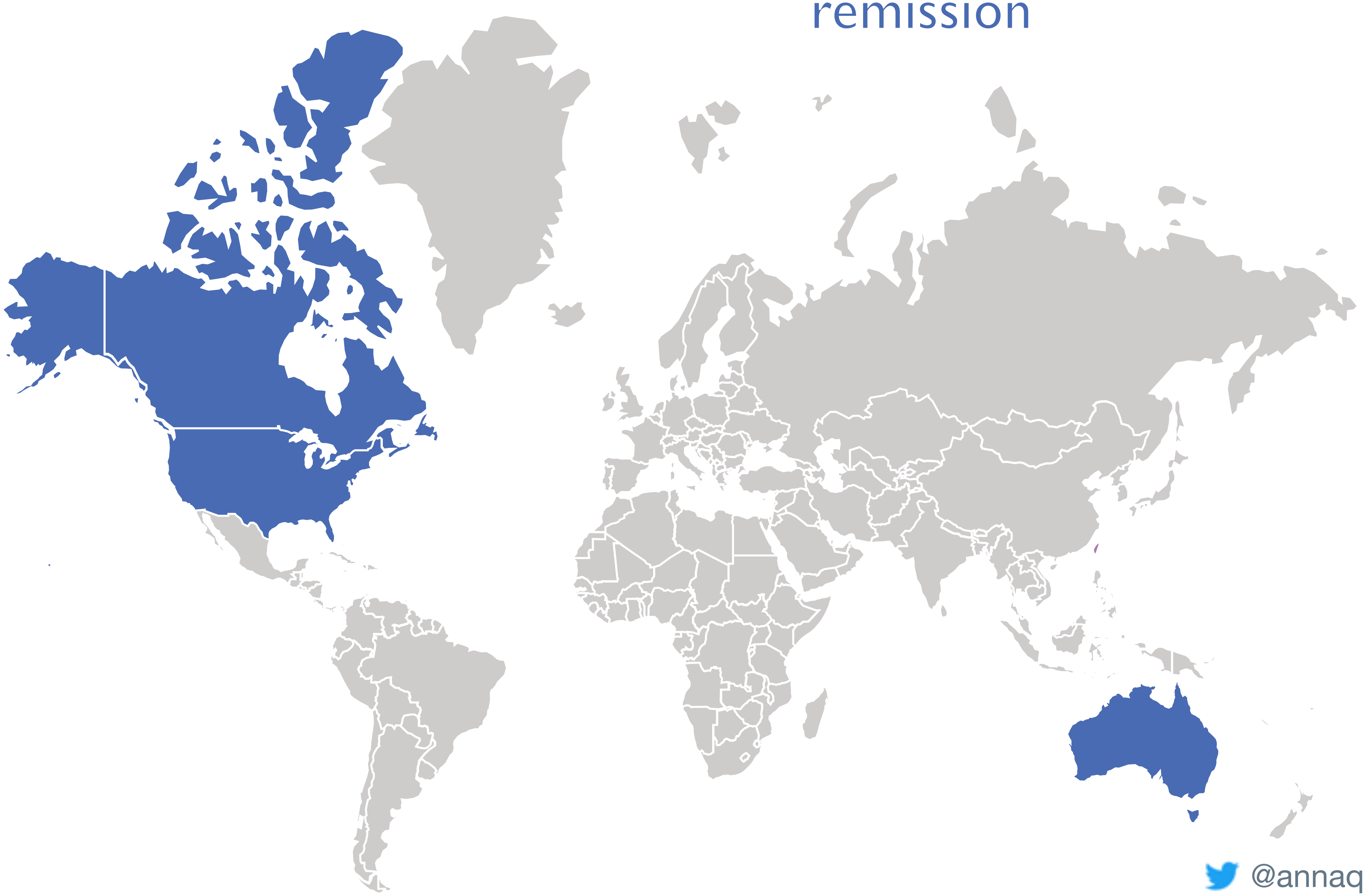


@annaquagli

Combined CBF-AML cohort

79
Long term remission

31
Relapse

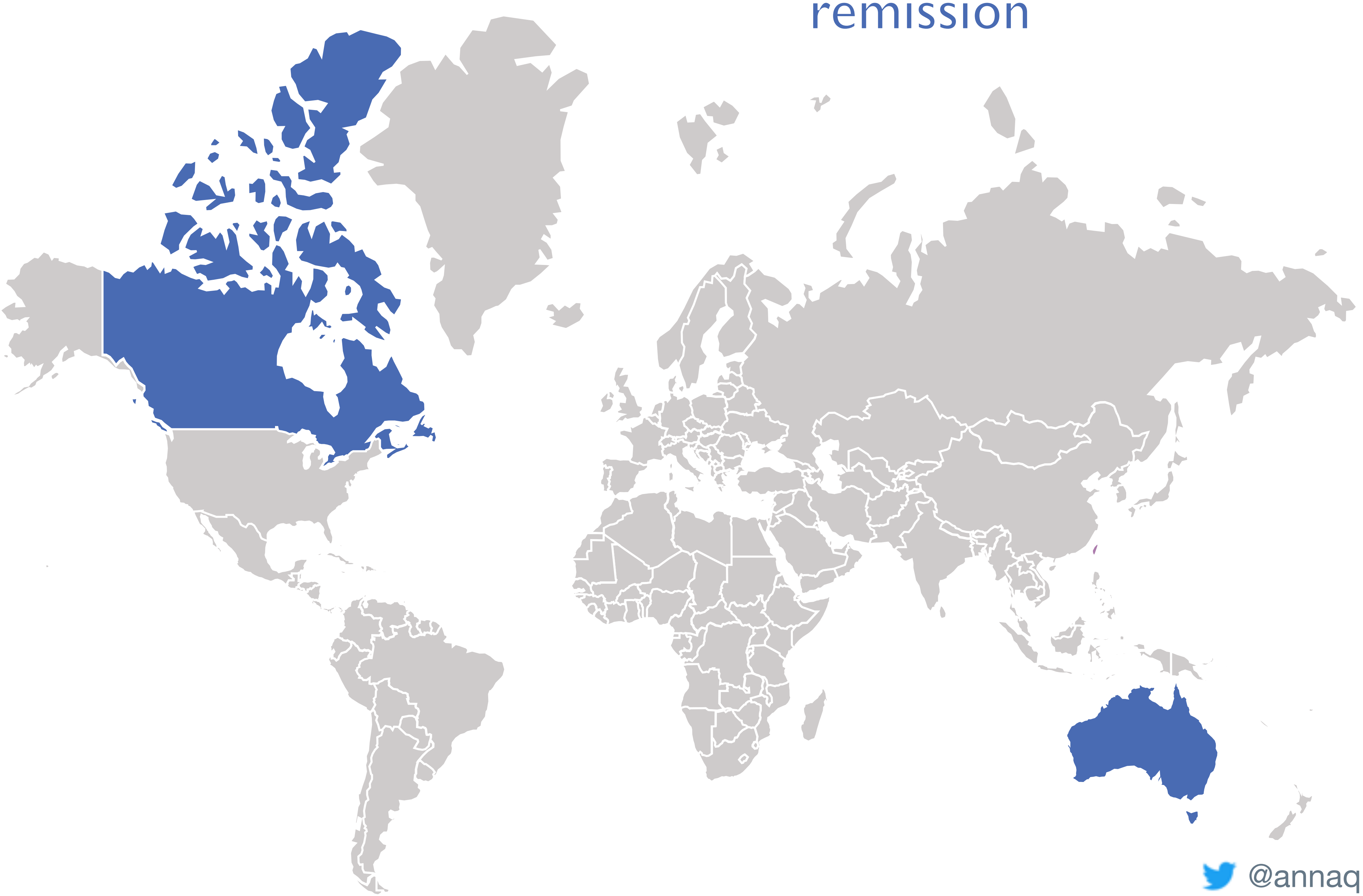


@annaquagli

Used so far...

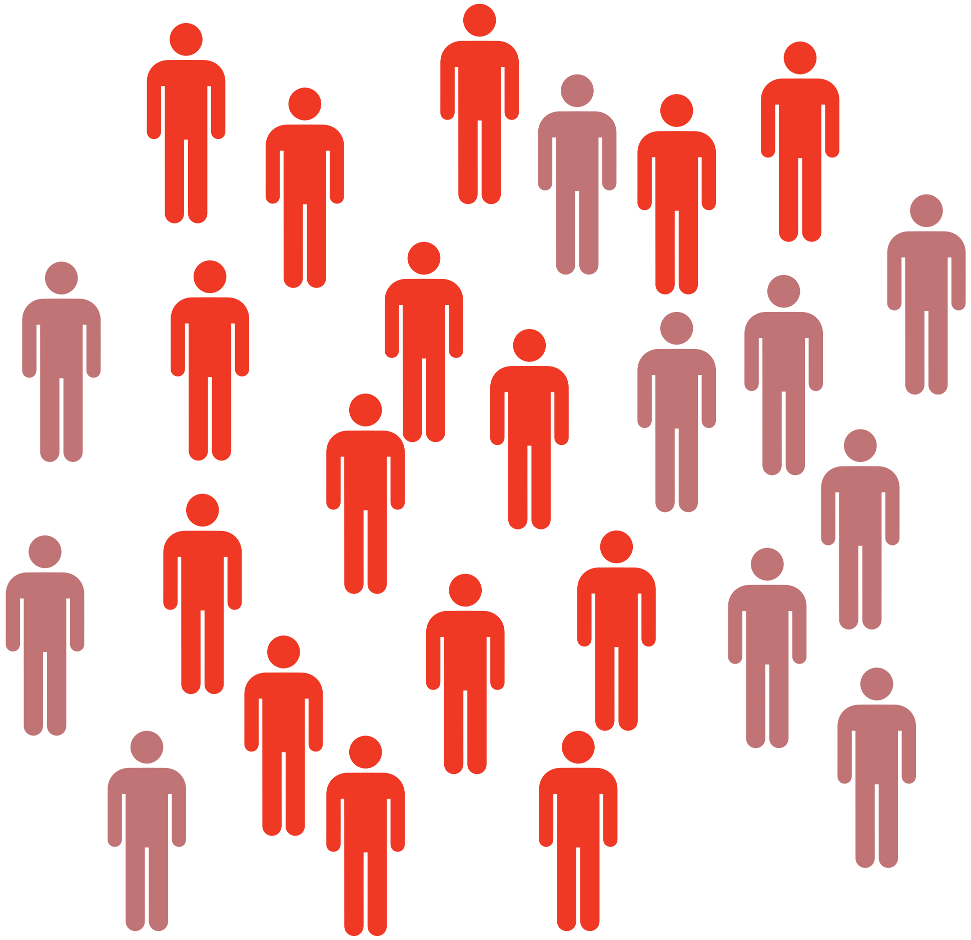
45
Long term remission

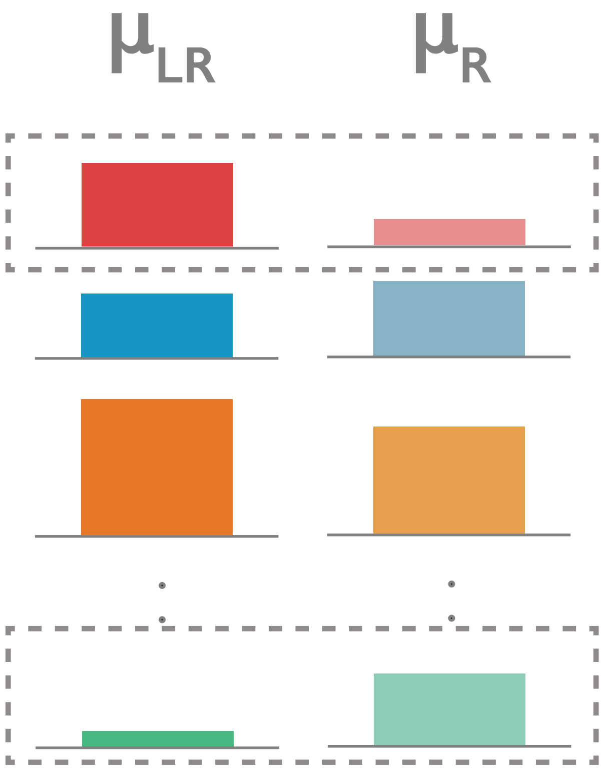
20
Relapse



@annaquagli

Larger sample can give more power to detect signal

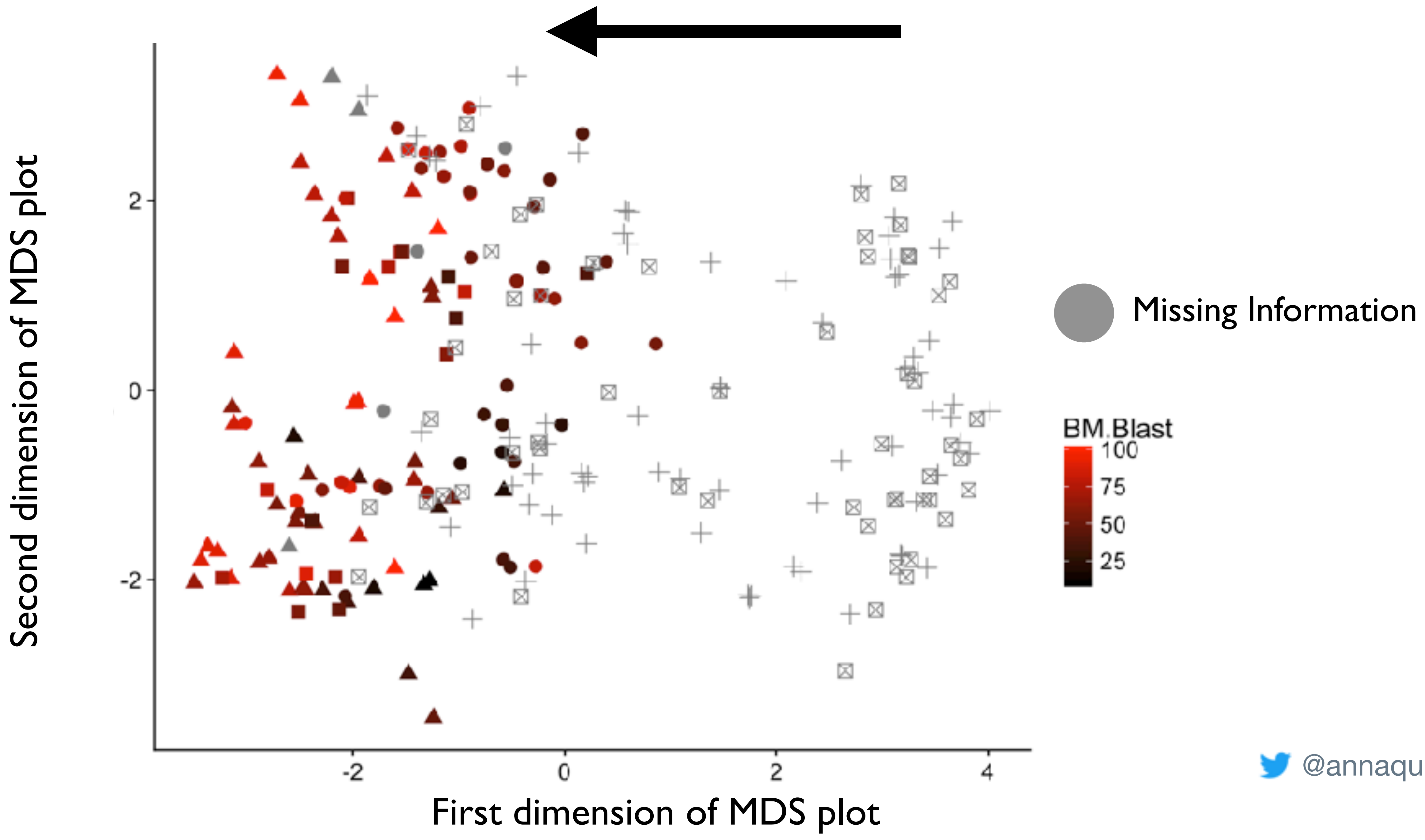




@annaquagli

Larger sample also increases unwanted heterogeneity

Higher burden of Leukemia



@annaquagli

That's why we needed RUV!

Removing Unwanted Variation

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

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

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

Log2 gene expression matrix

m samples

n genes

Model estimated with RUV-4

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

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

Model estimated with RUV-4

Matrix with Unwanted Variation

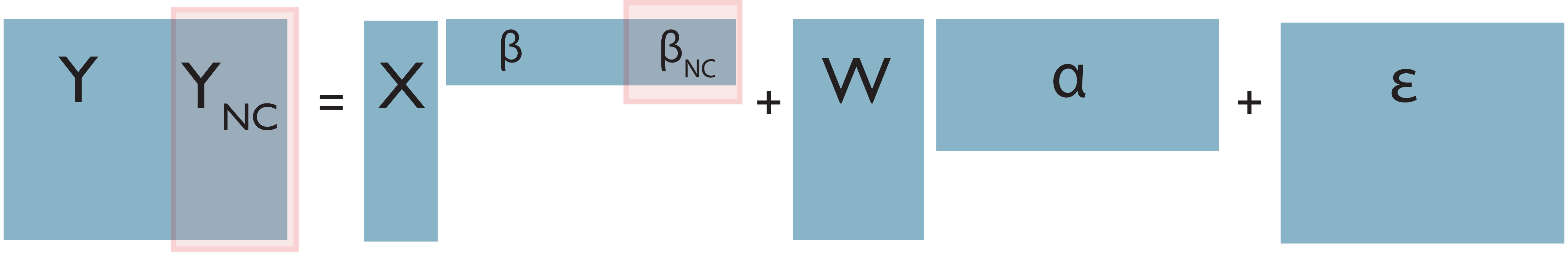
$$Y_{mxn} = X_{mxp}\beta_{pxn} + \boxed{W_{mxk}}\alpha_{kxn} + \epsilon_{mxn}$$

@annaquagli

Model estimated with RUV-4

Estimated using Negative Control genes

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



$$\beta_{NC} = 0$$

@annaquagli

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

FIAT LUX
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

AUSTRALIA
Australian Government

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE