Correcting unwanted variation in RNA sequencing data derived from a multicentre study of leukemia

Anna Quaglieri[1,2], Edward Chew[1,2], Terry Speed[1,3], Ian Majewski[1]

[1]Walter and Eliza Hall Institute of Medical Research
[2]The University of Melbourne, Faculty of Medicine, Dentistry and Health Sciences
[3]The University of Melbourne, Department of Mathematics and Statistics

Walter+Eliza Hall Institute of Medical Research
DISCOVERIES FOR HUMANITY

Background and data

Core-Binding Factor Acute Myloid Leukemia (CBF-AML)

AML is characterised by a rapid abnormal growth of undifferentiated white blood cells called **blasts**.
CBF-AML constitutes about 15% of AML patients and it is characterised by two recurrent gene fusions: a translocation between chromosomes 8 and 21 (t(8;21)) or an inversion within chromosome 16 (inv(16)).

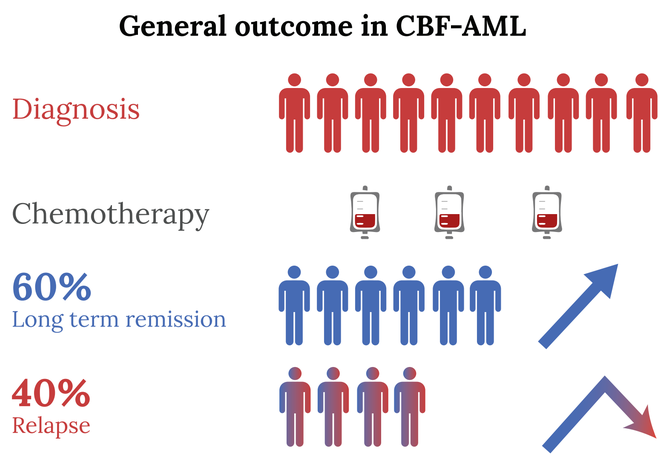


Figure 1. Outcome in CBF-AML patients after standard chemotherapy treatment.

Data: RNA-Seq samples at Diagnosis

The RNA-Seq samples at diagnosis were collected across 2.5 years. The Australian data come from the Australasian Leukemia and Lymphoma Group AMLM19 Clinical Trial and they were sequenced in two batches one year apart from each other. The Canadian cohort was published in Lavallée et al[1], (2016) and is available on GEO.

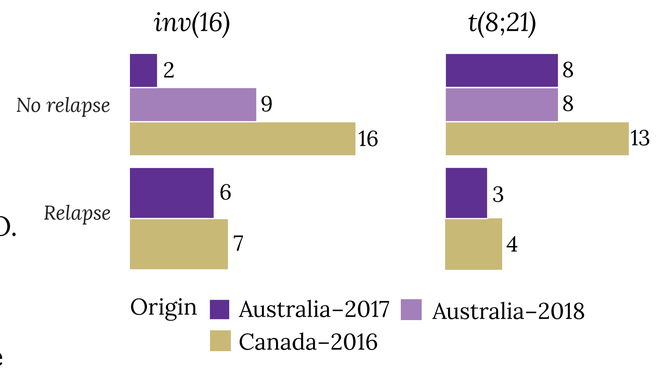


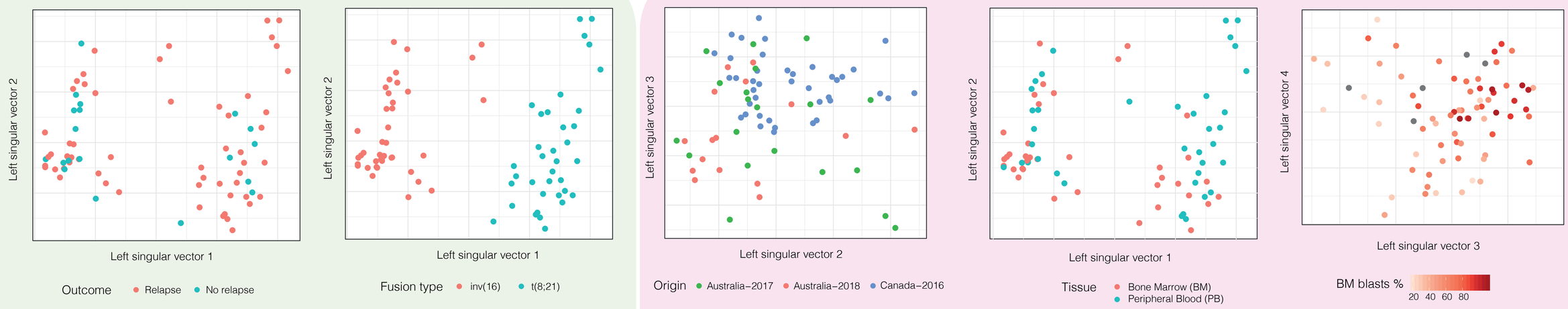
Figure 2. Number of diagnostic RNA-Seq samples available at diagnosis (technical replicates are excluded) by fusion type.

Aim of the study and challenges

Can we overcome the heterogeneity in the data (Fig.3) and the relatively small cohort size (Fig.2) to detect differentially expressed (DE) genes between patients who will relapse and patients who will stay in a long remission?

Heterogeneity in gene expression

Figure 3. Exploring the variability in the cohort in Figure 2 using the Singular Value Decomposition (SVD)



The **wanted heterogeneity** are the factors of interest whose effect we wish to correctly estimate

The **extra heterogeneity (technical and biological)** that needs to be corrected for, to avoid confounding in the estimates for the factors of interest

Illustrative example: dealing with differing tumour content

Within the linear model framework - Model for a single gene

Include tumour content (and other known variables) as covariate

$$Y_{m\times 1} = X_{m\times p}\beta_{p\times 1} + \alpha_1 TumourContent_{m\times 1} + W_{m\times(k-1)}\alpha_{(k-1)\times 1} + \epsilon_{m\times 1}$$

Known covariates with W assumed known

Log_2 gene expression matrix
m samples
1 gene
p **factors of interest**

😍 Easy to use
😥 Tumour content needs to be estimated
😕 A proxy available is the bone marrow blast content (%), missing for some samples

RUV-4: Estimate a correction factor from the expression data

$$Y_{m\times 1} = X_{m\times p}\beta_{p\times 1} + W_{m\times k}\alpha_{k\times 1} + \epsilon_{m\times 1}$$

W unknown
Estimated from **Negative Control (NC) genes**

😍 Available for all samples
😍 Takes into account bone marrow and blood differences when adjusting for blast content (Fig.4 A-B)
😍 Can adjust for several sources of heterogeneity at the same time (Fig.3 C-D)
😥 More challenging to use: choice of dimension **k** and **NC** genes

Results and conclusions

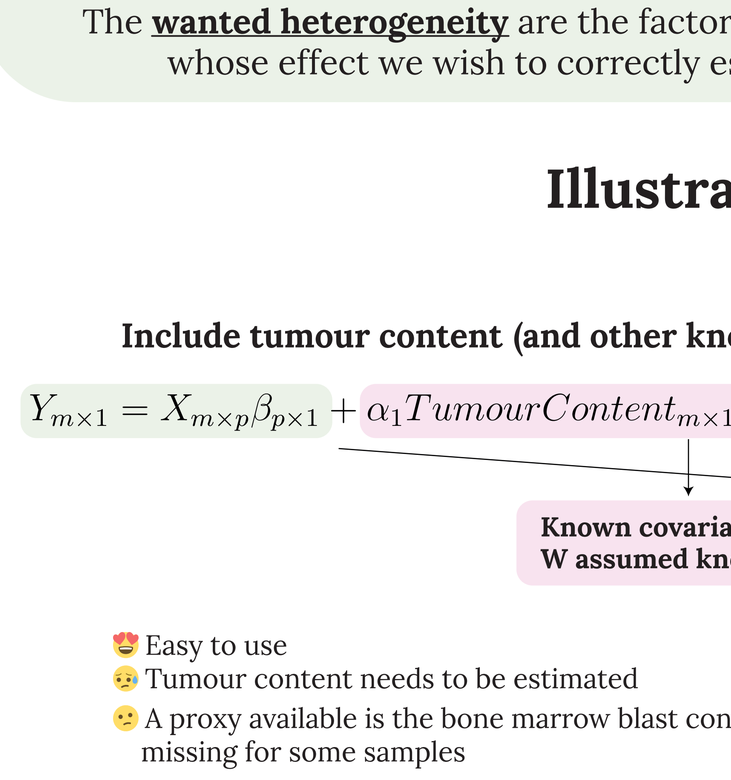


Figure 4. **A)** W_1 estimated by RUV-4 using the Australian cohort vs BM blasts %. **B)** W_2 estimated by RUV-4 using the Canadian cohort vs BM blasts %. Signs in the estimated columns of W do not matter. **C)** and **D)** W_1 and W_2 from RUV-4 applied to the combined cohort. W1 picks up tissue differences and W_2 also picks up cohort differences.

Model fitted to combined or separate Australian and Canadian cohorts

$$Y = \mu + \beta_1 Fusion.Type + \beta_2(inv(16)Rel - inv(16)NoRel) + \beta_3(t(8;21)Rel - t(8;21)NoRel) + W\alpha + \epsilon$$

- Using limma on the combined cohort including blast content, tissue type and centre of origin as covariates, no genes were found DE between *Relapse* and *No Relapse* samples in either fusion types. Four samples were removed due to missing blast content.
- We used **RUV-4** to estimate W with separate and combined models with the Australian and Canadian cohorts. We show that RUV-4 is able to **adjust for the heterogeneity in the cohorts** (**Figure 3 A-D**).
- Using RUV-4 we found 63 up and 91 downregulated genes in inv(16) and 59 downregulated in t(8;21). However, the significance and the logFC shows that the signal is not very strong.
- Some promising genes previously associated with sensitivity to chemotherapy are identified among the top DE genes (PID1, MAPK10).
- However, the small cohort size, the large amount of heterogeneity (tissues, fusion types, origins and tumour content) offer a **serious challenge** and more research is required to assess the improvement provided by RUV-4 as well as the **reproducibility of the signature found**, in an independent cohort.

Bibliography

1. https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=190
2. Lavallée et al, RNA-sequencing analysis of core binding factor AML identifies recurrent ZBTB7A mutations and defines
3. RUV http://www-personal.umich.edu/~johanngb/ruv/index.html

Dr Edward Chew is a recipient of the RCPA Foundation Illumina Cancer Research Grant 2018 from the Royal College of Pathologists of Australasia Foundation