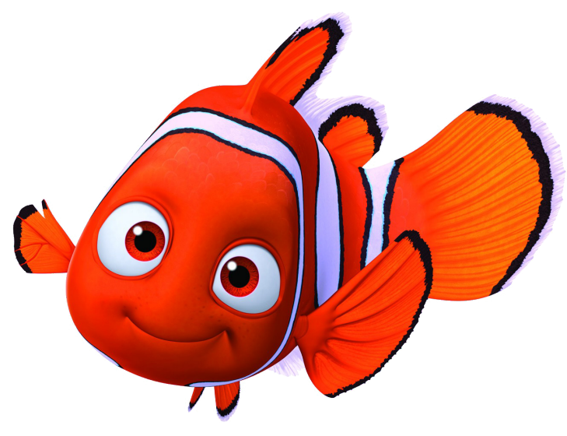
Finding optimal coverage



Anna Quaglieri[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

Background

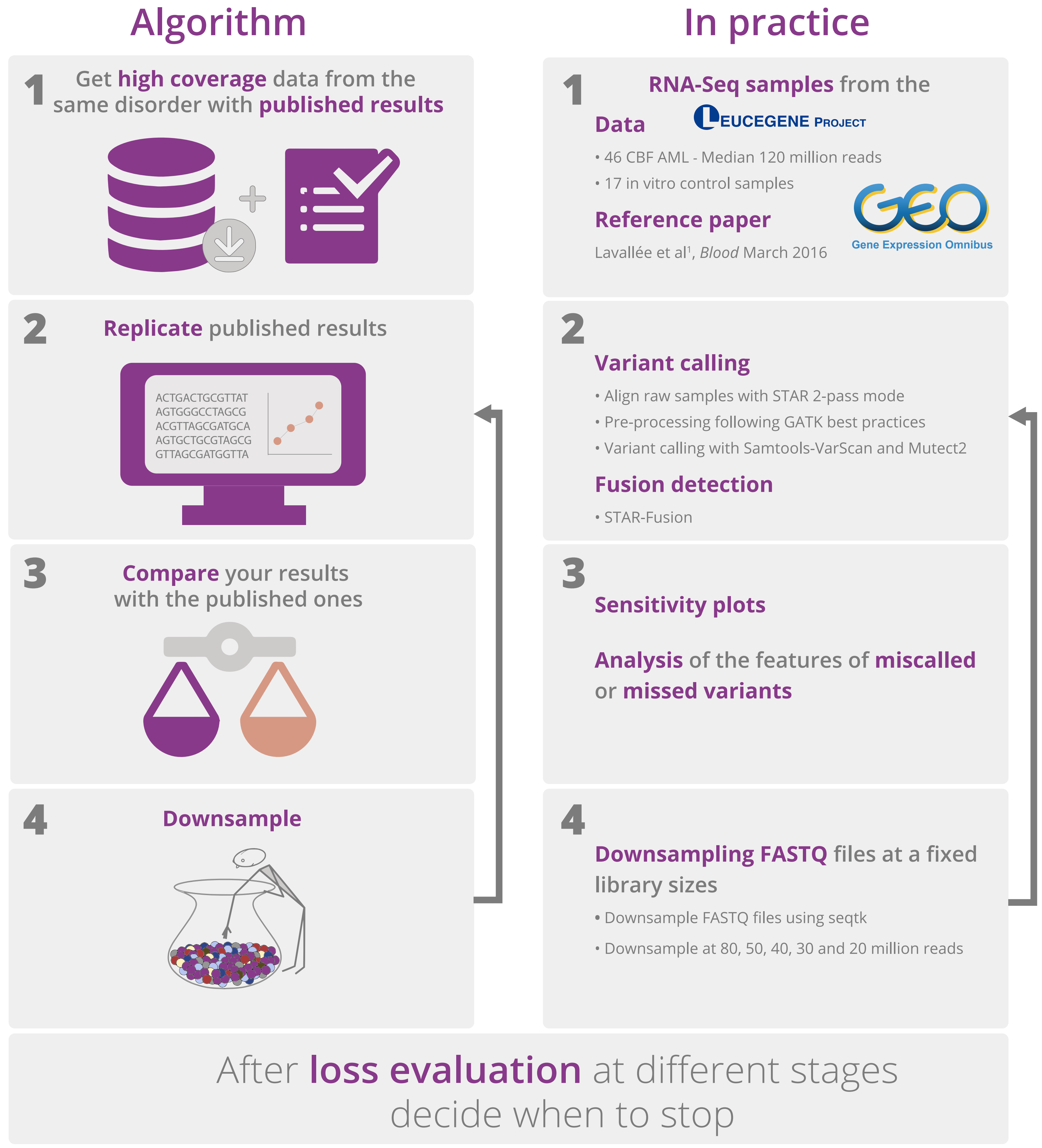
Next-Generation Sequencing (NGS) technologies have become a **critical source of information in the understanding of diseases.** However, experimental design is often overlooked resulting in suboptimal power and high financial costs.

Coverage, seen as the average number of times that a base of a genome is sequenced, and the **number of samples** are fundamental factors affecting both the costs and the results of an experiment.

The choice of coverage is especially critical in **cancer genomics** where data are more noisy and mutations may appear with a low frequency.

Here we describe the approach we took to design the **sequencing** of a set of **RNA** samples from a cohort of **Core Binding Factor Acute Myeloid Leukemia (CBF-AML) patients** collected by the Australasian Leukemia and Lymphoma Group.

Methods & Data



Algorithm

1 Get **high coverage** data from the same disorder with **published results**

2 **Replicate** published results

3 **Compare** your results with the published ones

4 **Downsample**

In practice

1 **RNA-Seq samples** from the

Data LEUCEGENE PROJECT

- 46 CBF AML - Median 120 million reads
- 17 in vitro control samples

Reference paper

Lavallée et al[1], *Blood* March 2016

GEO Gene Expression Omnibus

2 **Variant calling**

- Align raw samples with STAR 2-pass mode
- Pre-processing following GATK best practices
- Variant calling with Samtools-VarScan and Mutect2

Fusion detection

- STAR-Fusion

3 **Sensitivity plots**

Analysis of the features of **miscalled** or **missed variants**

4 **Downsampling FASTQ** files at a fixed library sizes

- Downsample FASTQ files using seqtk
- Downsample at 80, 50, 40, 30 and 20 million reads

After **loss evaluation** at different stages decide when to stop

Features of the data

Variant type	Frequency
[1]Composite	20
[2]Long Insertions	3
Short Deletions	2
Short Insertions	14
Single Base (SNVs)	58

The CBF-AML RNA-Seq libraries have a median library size around **100 million fragments** (PE, 100 bases reads).

CBF-AML patients show one of two known recurrent gene fusions:
- traslocation between chromosomes 8 and 21, also t(8;21), which originates the fusion gene **RUNX1-RUNX1T1**;
- inversion on chromosome 16, also inv(16), which creates the fusion gene **CBFB-MYH11**.

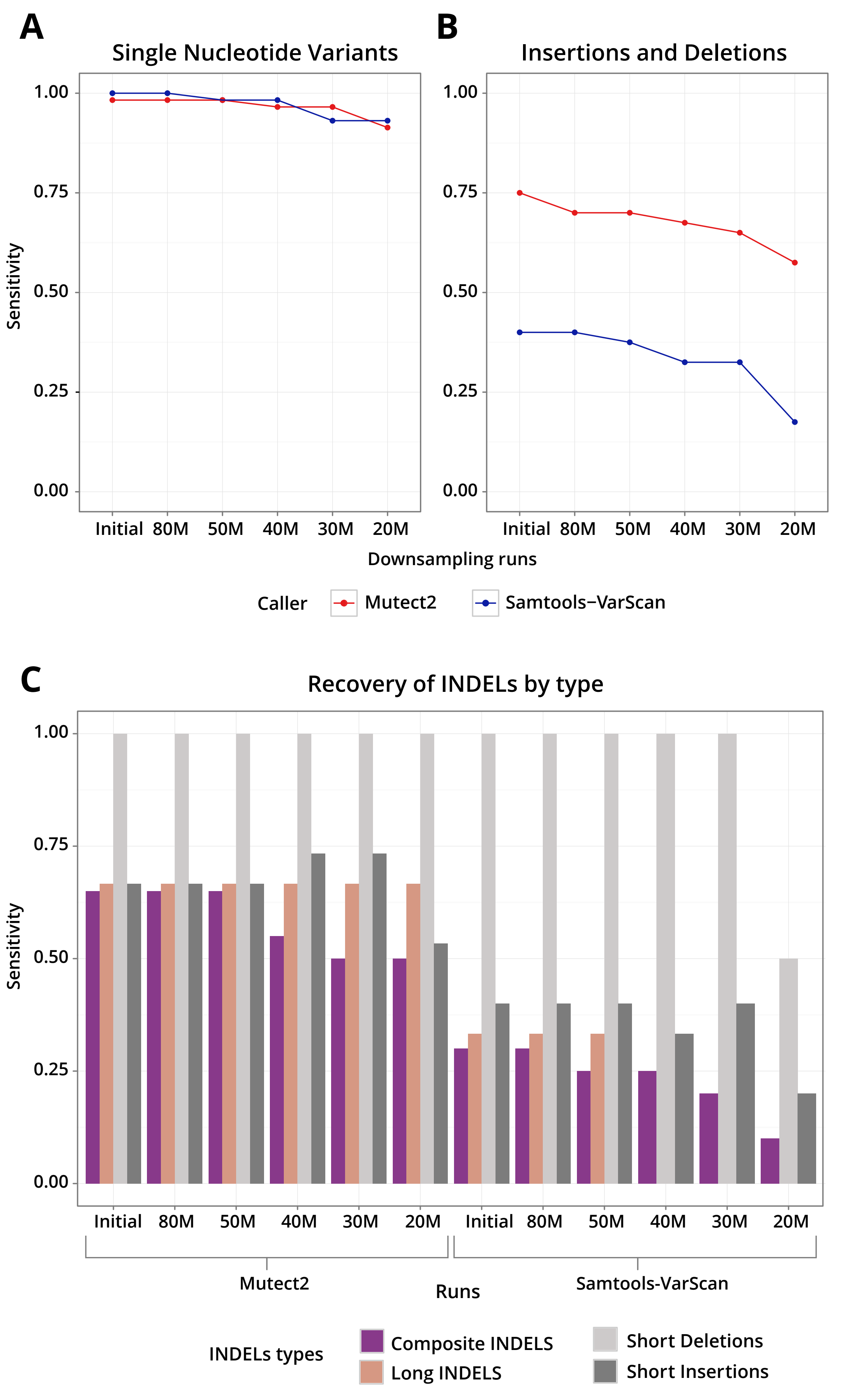
Table 1
Variants detected in Lavallée et al[1].
[1]A **long INDEL** involves more than 10 base pairs.
[2]**Composite variants** include both insertions and deletions at the same time. They include 2 long and 8 short deletions and 10 short insertions.

Bibliography

1. Lavallée et al, RNA-sequencing analysis of core bindign factor AML identifies recurrent ZBTB7A mutations and defines RUNX1-CBFA2T3 fusion signature. *Blood.* March 2016
2. https://bitbucket.org/iric-soft/km

Results

Recovery of variants



SNVs lost	Mutect2		VarScan	
Runs	N	VAF (median)	N	VAF (median)
Initial	1	0.80	0	-
Down 80M	1	0.87	0	-
Down 50M	1	0.94	1	0.1
Down 40M	2	0.94	1	0.1
Down 30M	2	0.94	4	0.06
Down 20M	5	0.1	4	0.06

Table 2
Number of SNVs lost by Mutect2 and Samtools-VarScan at every downsampling step. Their median Variant Allele Frequency (VAF) is also reported.

Figure 1
Sensitivity of SNVs and INDELSs at every downsampling step.
Proportion of SNVs **(A)** and INDELs **(B)** from **Table 1** called by either Mutect2 or Samtools-VarScan.
(C) Sensitivity by type of INDEL across the different downsampling runs and by caller.

Recovery of fusions

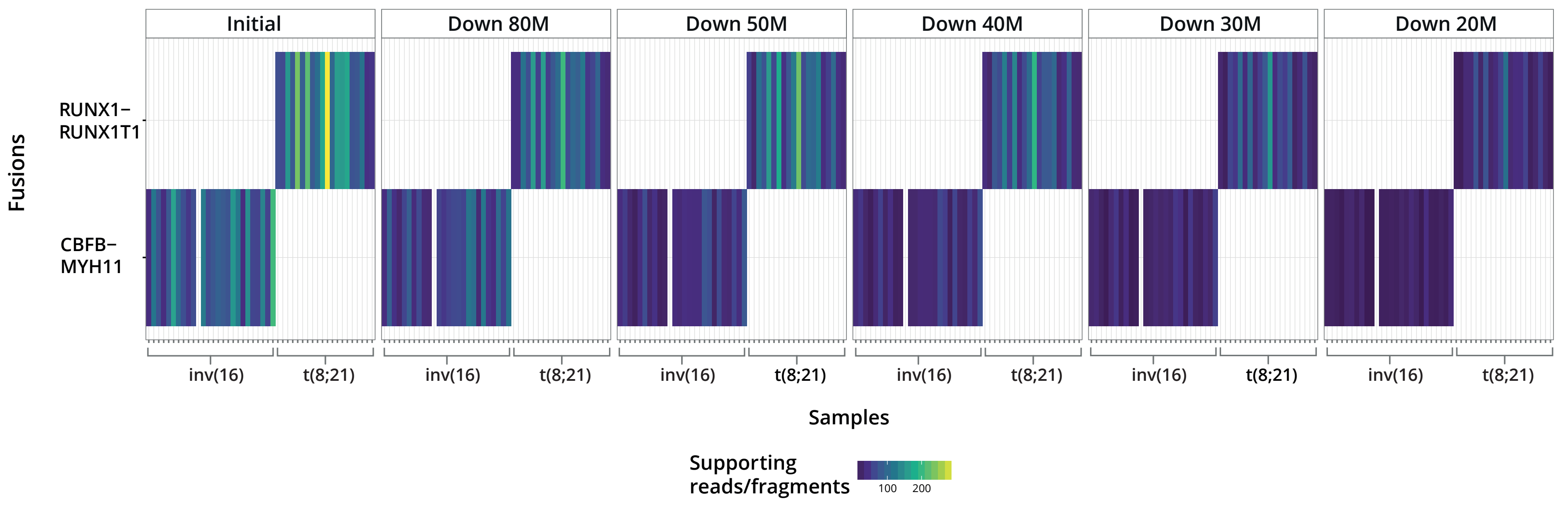


Figure 2
Recovery of the two known CBF-AML recurrent fusions.
A fusion event is called if has at least five supporting reads or fragments. The only fusion not recovered across all the runs is due to a parameter in the STAR aligner which can be tweked for further analysis. Other fusions have been found but mainly ruled out as false positives.

Conclusions

- The above results suggest that a **library size** of **at least 30 million fragments** is advisable, obtaining an approximate **coverage of 83x** using the definition of coverage as (Read Length x 2 x Library Size)/(Num. Genes x Mean Gene Length).
- At 20 million Mutect2 starts missing SNVs with low frequency (Figure 1A) as well as short INDELs (Figure 1B).
- The advisable **library size should be increased for samples with lower tumour content.**
- More downsampling runs are needed to compute error bars around the estimated sensitivity in Figure 1A and 1B.
- **Mutect2** and **VarScan** show **similar power in detecting SNVs** (Figure 1A). However, **Mutect2 is largely better for INDELs** (Figure 1B) while **VarScan is slightly better** in calling variants with **high VAF** (Table 1 and Figure 1A).
- More specialised tools should be used to detect INDELs (km[2] was used in Lavallée et al[1]) but this goes beyond the scope of this analysis.
- The **known recurrent fusions are detected** up to the lowest library sizes (Figure 2).

If you have comments or suggestion you can find me on Twitter! @annaquagli