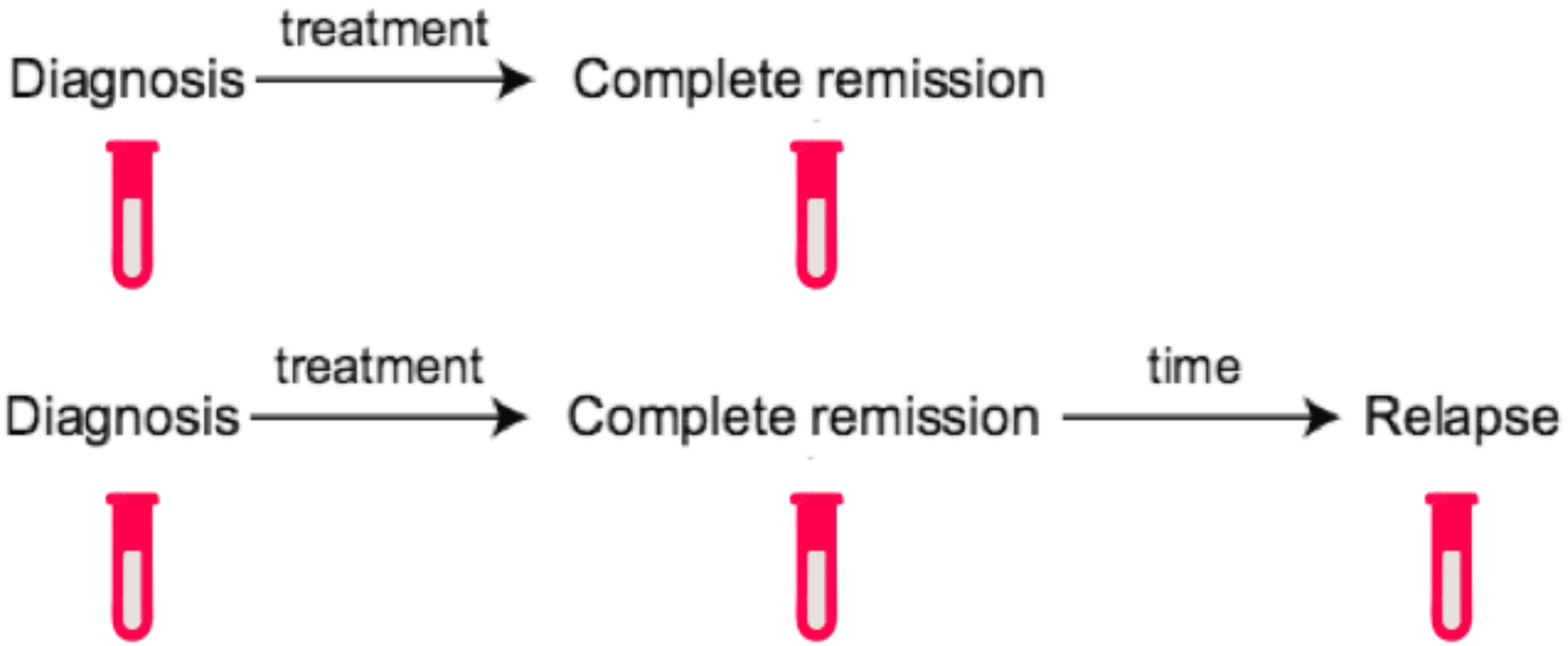
A suite of Bioinformatics tools to call and explore somatic mutations over time from RNA-Seq

Anna Quaglieri
Monash Bioinformatics
3rd April 2019

Why?

RNA-Seq comes as a sequence, so why not?

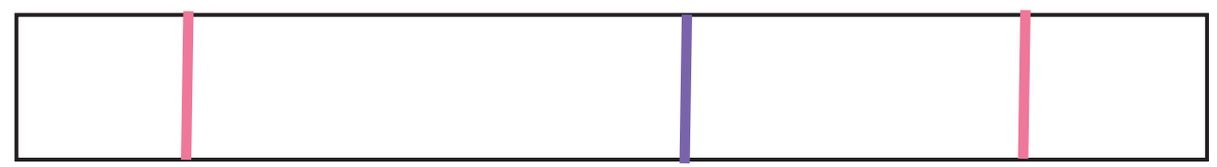
Data available for my project: Leukemia RNA-Seq



We wanted to extract all things from RNA-Seq…

Variant calling

Tumour-normal variant calling



Tumour sequence

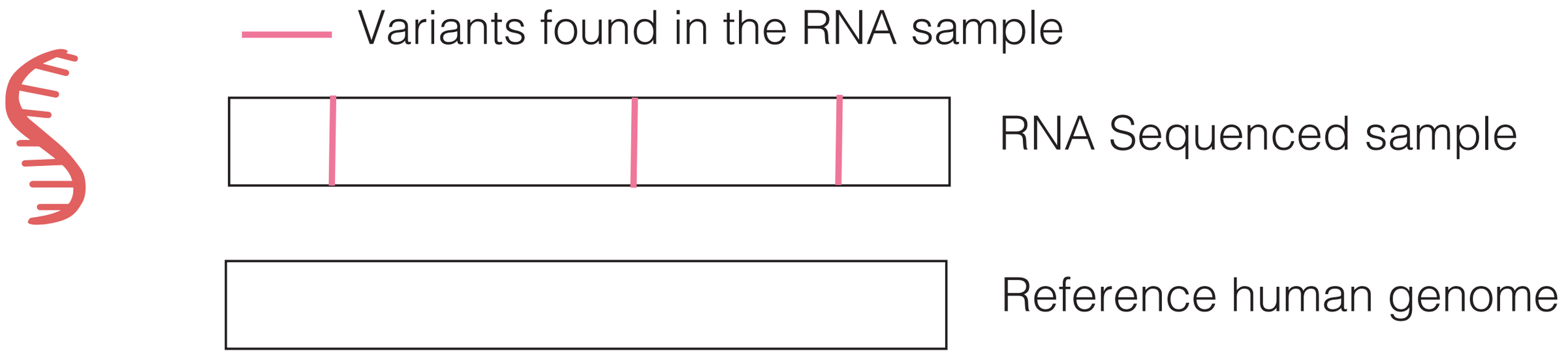
Normal sequence

Reference human genome

Somatic variants found only in the tumour sample

Germline variant found in both the tumour and normal samples

Calling variants in tumour-only RNA-Seq with multiple samples for same patient across time



We need other strategies to classify a variants in **somatic** or **germline**

Challenges

- **How do we need to pre-process the bamfiles?**

- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?

- How do we filter germline variants without a matched normal?

- How do we track changes in mutations or groups of mutations (clones) in time for a patient?

- How to combine, summarise and plot all these results?

Challenges

- How do we need to pre-process the bamfiles?

- **We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?**

- How do we filter germline variants without a matched normal?

- How do we track changes in mutations or groups of mutations (clones) in time for a patient?

- How to combine, summarise and plot all these results?

Challenges

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- **How do we filter germline variants without a matched normal?**
- How do we track changes in mutations or groups of mutations (clones) in time for a patient?
- How to combine, summarise and plot all these results?

Challenges

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- How do we filter germline variants without a matched normal?
- **How do we track changes in mutations or groups of mutations (clones) in time for a patient?**
- How to combine, summarise and plot all these results?

Challenges

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- How do we filter germline variants without a matched normal?
- How do we track changes in mutations or groups of mutations (clones) over time for a patient?
- **How to combine, summarise and plot all these results?**

Let's start…

- **How do we need to pre-process the bamfiles?**

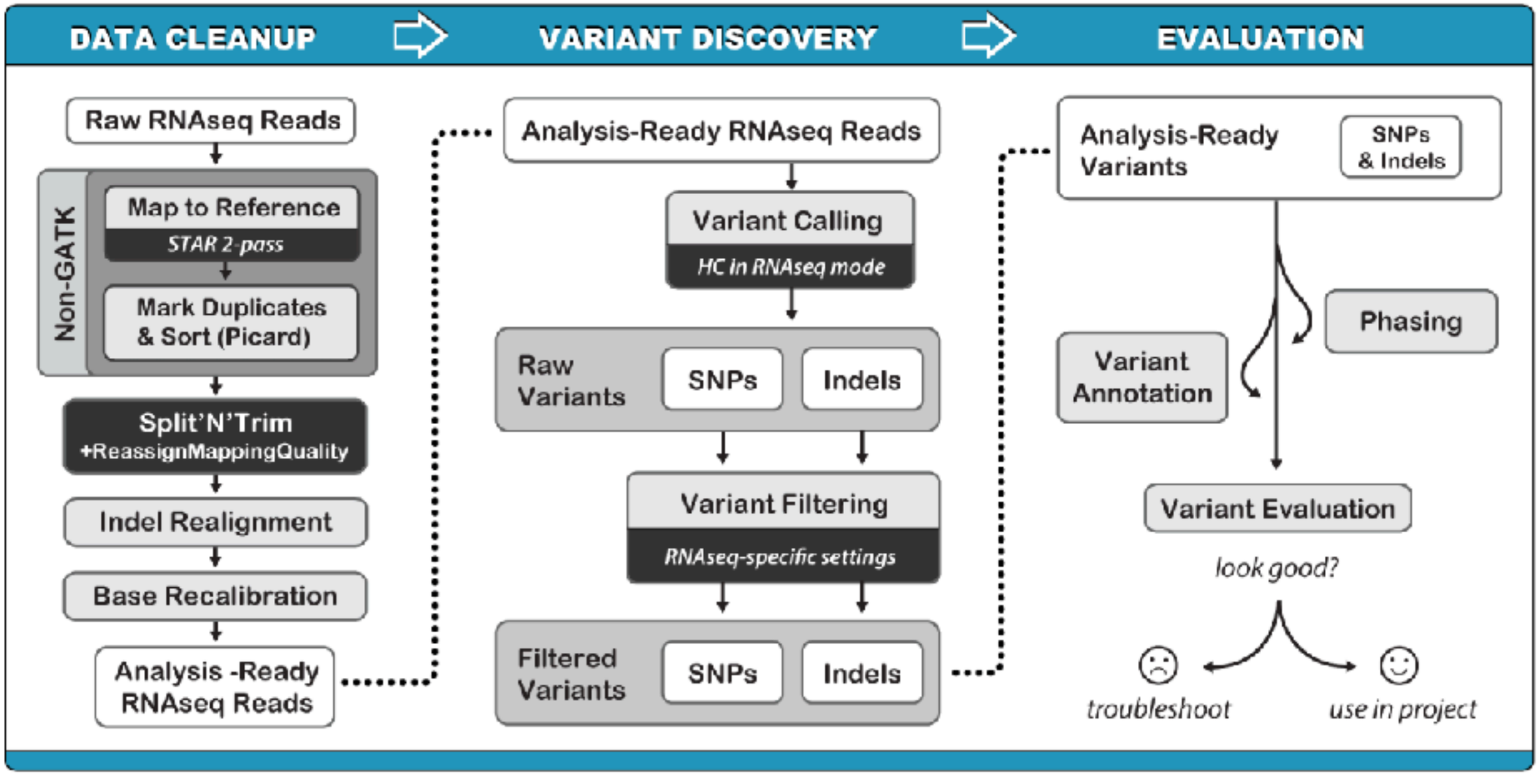
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?

- How do we filter germline variants without a matched normal?

- How do we track changes in mutations or groups of mutations (clones) in time for a patient?

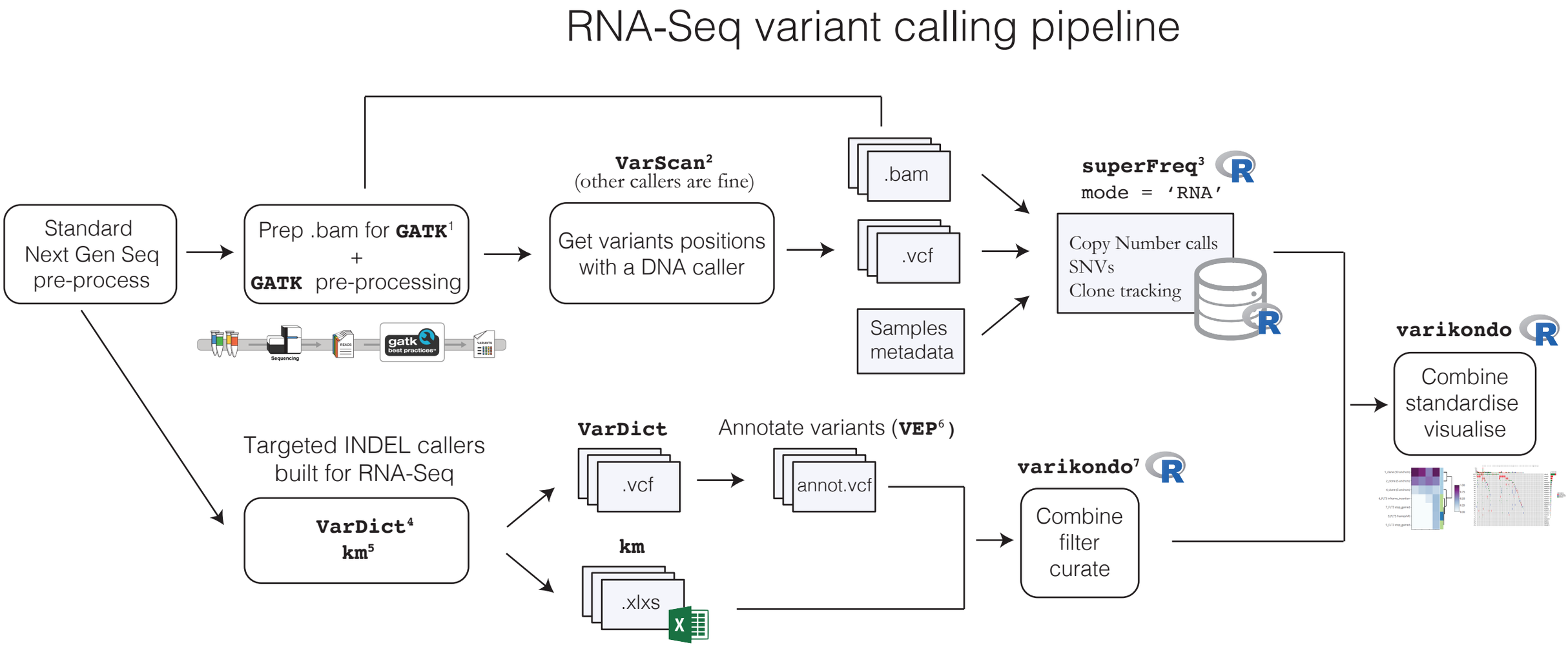
- How to combine, summarise and plot all these results?

I started by looking at GATK best practices



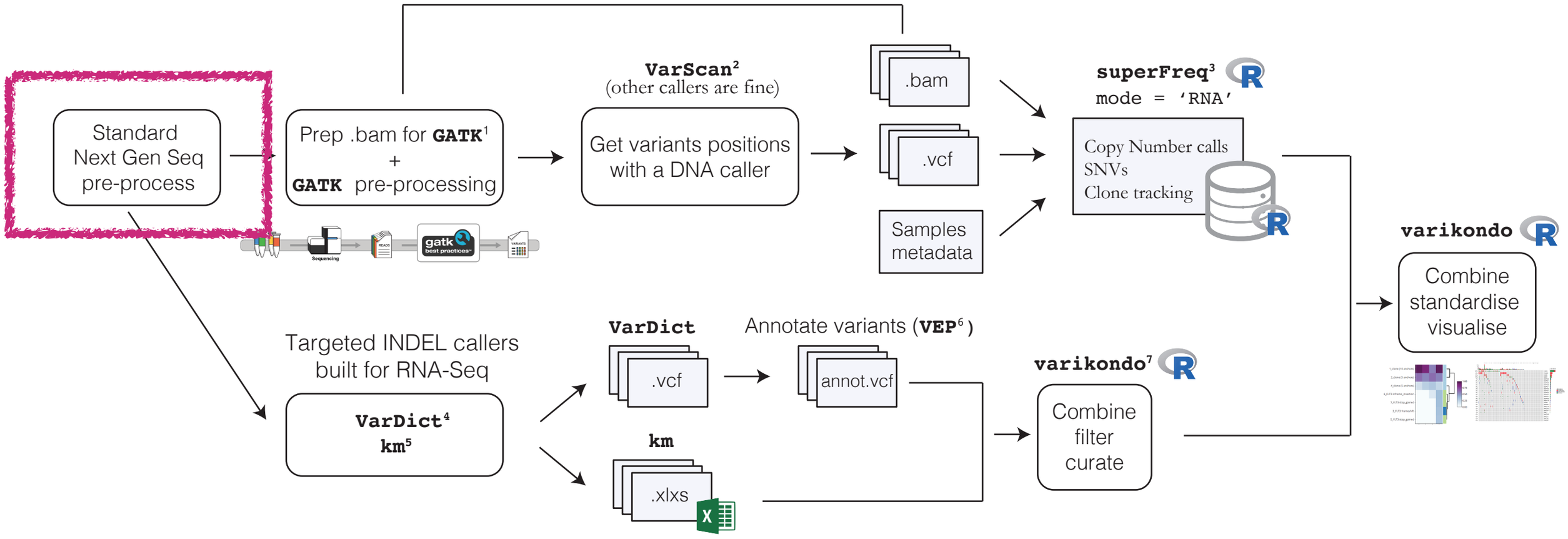
https://software.broadinstitute.org/gatk/documentation/article.php?id=3891

Explanations, code and function of pipeline:
https://rna-mutation-calls.netlify.com/



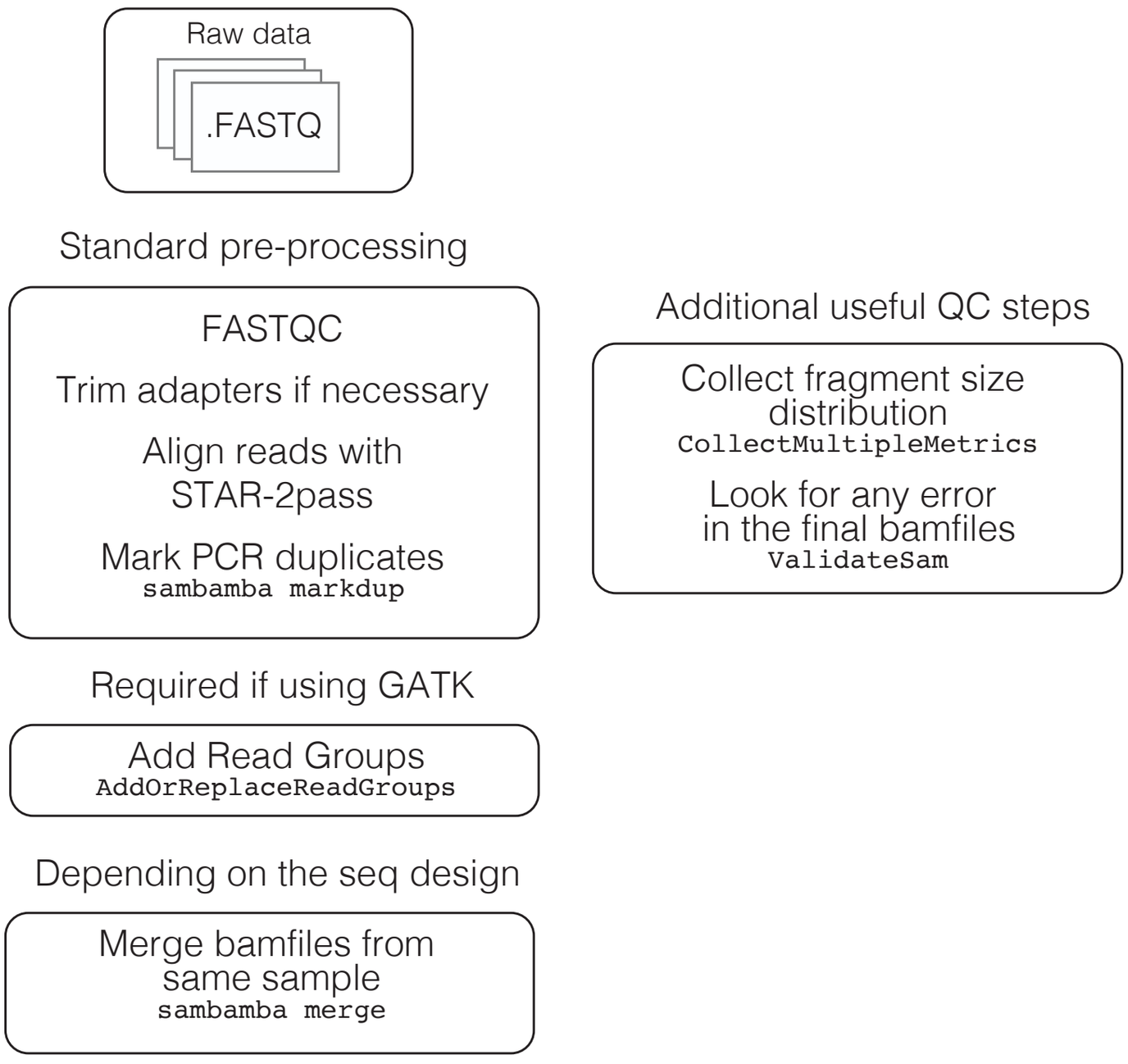
1. `GATK` https://software.broadinstitute.org/gatk/
2. `VarScan` http://varscan.sourceforge.net/
3. `superFreq` https://github.com/ChristofferFlensburg/superFreq
4. `VarDict` https://github.com/AstraZeneca-NGS/VarDict
5. `km` https://github.com/iric-soft/km
6. `VEP` https://asia.ensembl.org/info/docs/tools/vep/index.html
7. `varikondo` https://github.com/annaquaglieri16/lineplots

Standard Seq pre-processing



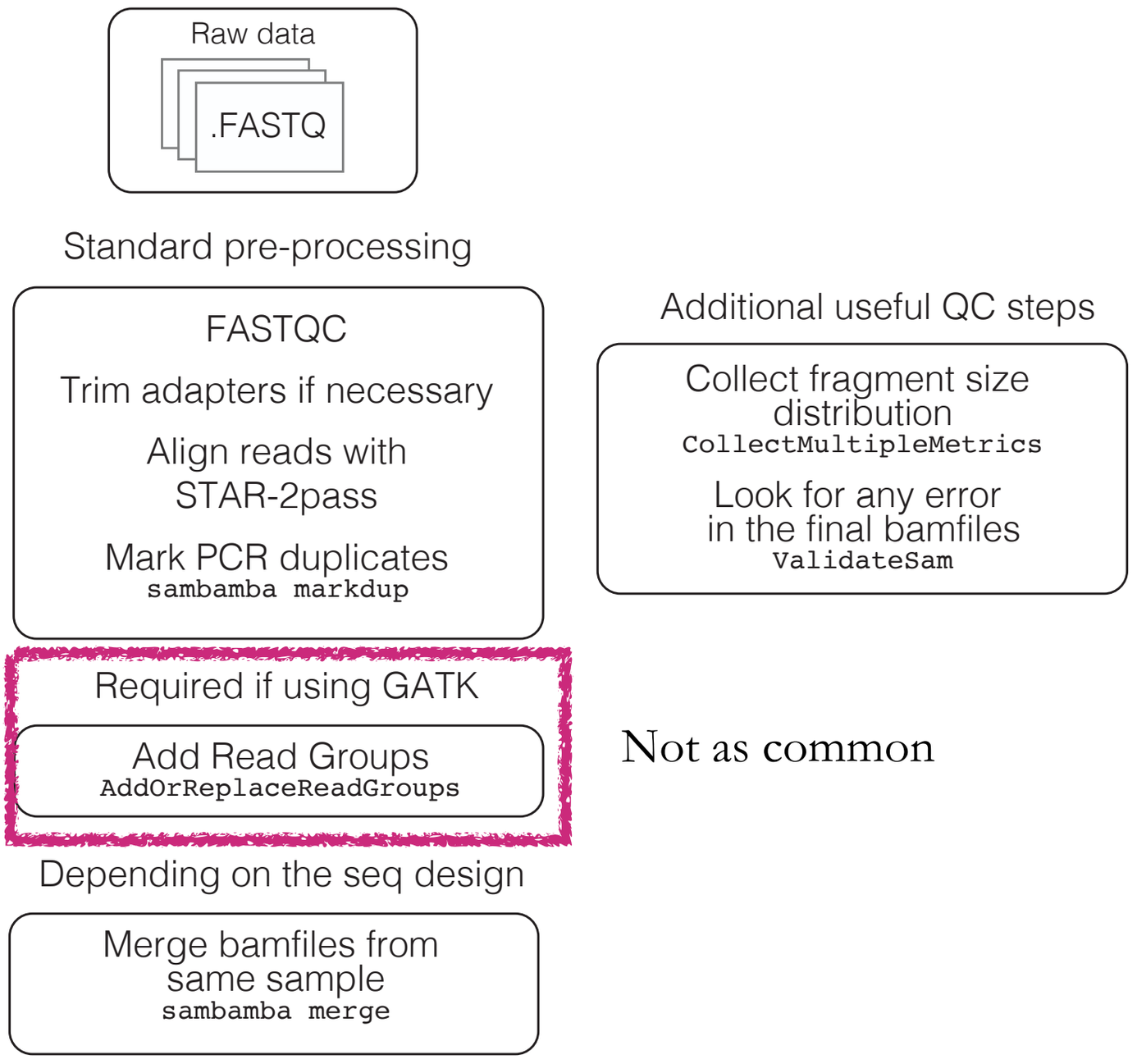
1. `GATK` https://software.broadinstitute.org/gatk/
2. `VarScan` http://varscan.sourceforge.net/
3. `superFreq` https://github.com/ChristofferFlensburg/superFreq
4. `VarDict` https://github.com/AstraZeneca-NGS/VarDict
5. `km` https://github.com/iric-soft/km
6. `VEP` https://asia.ensembl.org/info/docs/tools/vep/index.html
7. `varikondo` https://github.com/annaquaglieri16/lineplots

Standard pre-processing



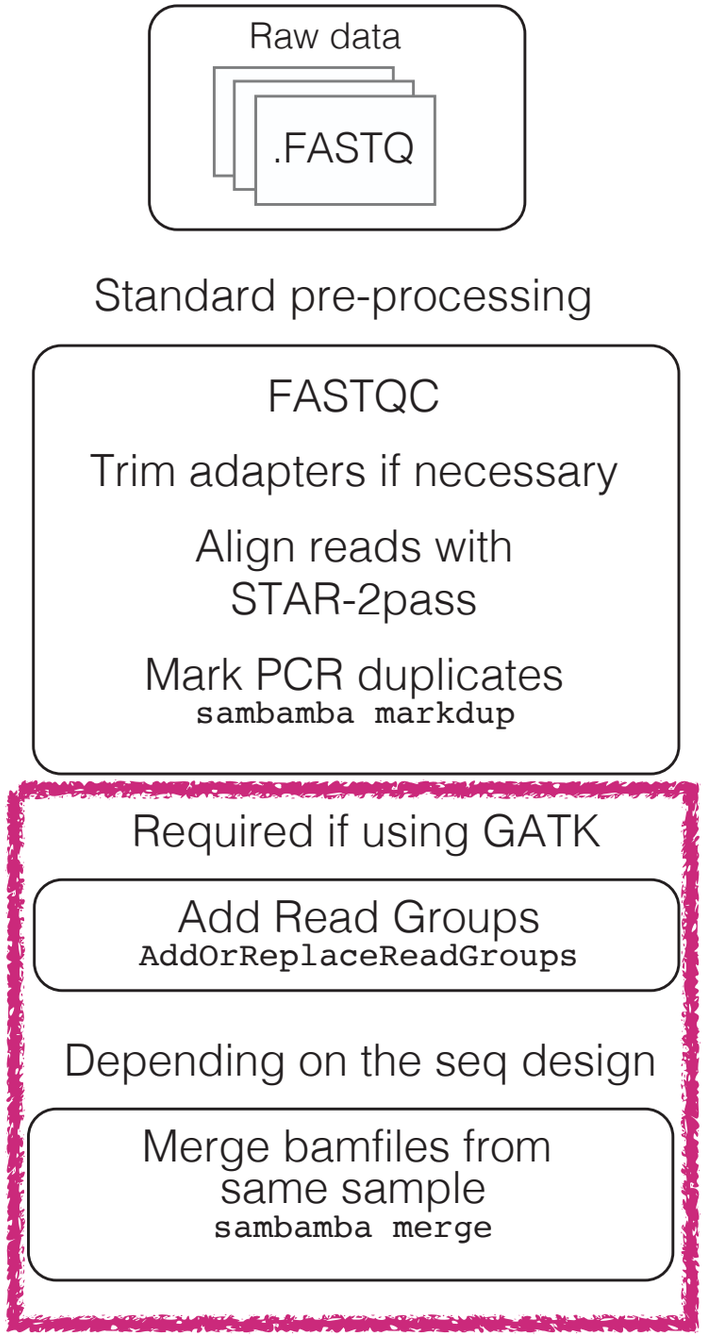
I also then remove duplicate reads with
sambamba markdup input.bam outputDedupl.bam --remove-duplicates!

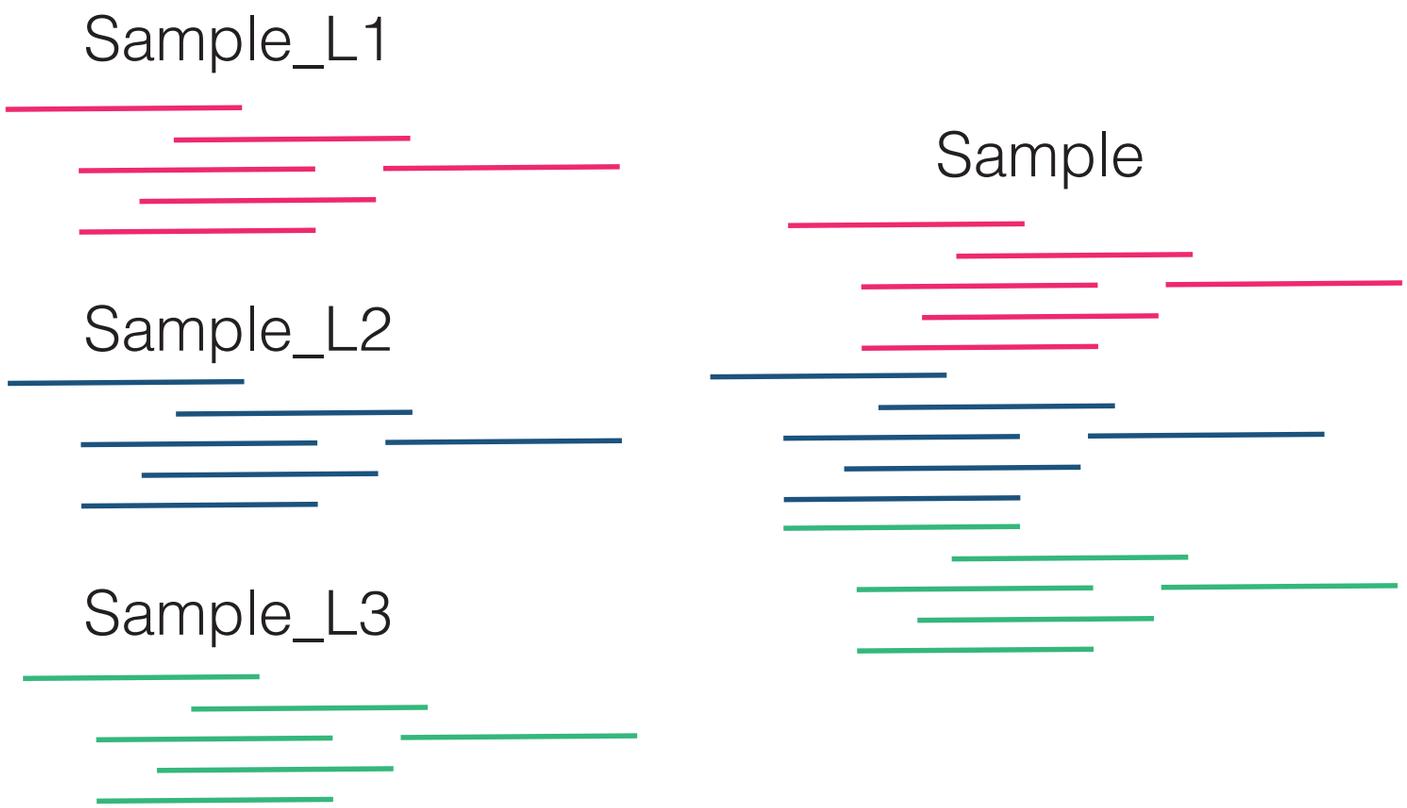
Standard pre-processing



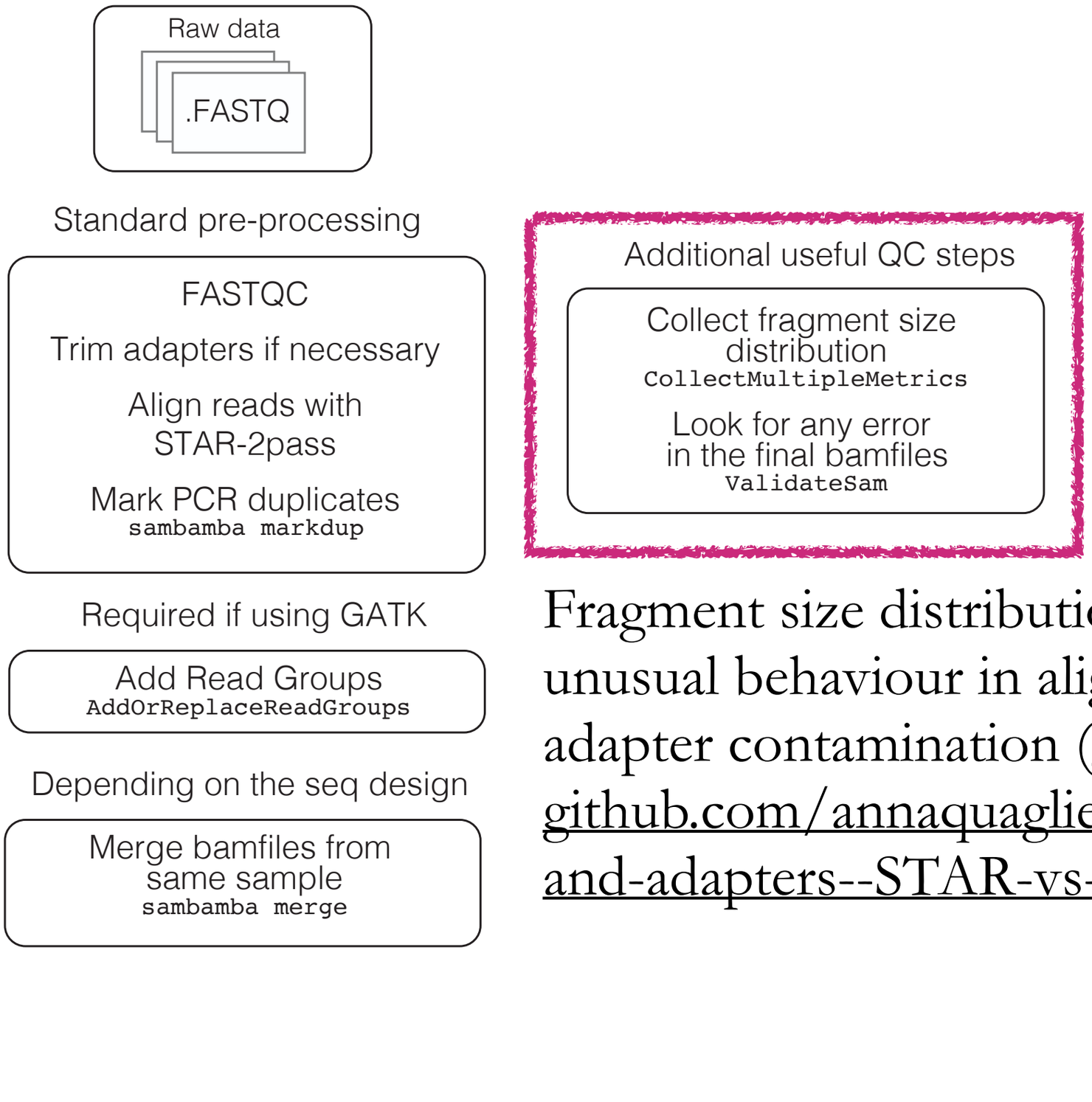
Not as common

Standard pre-processing



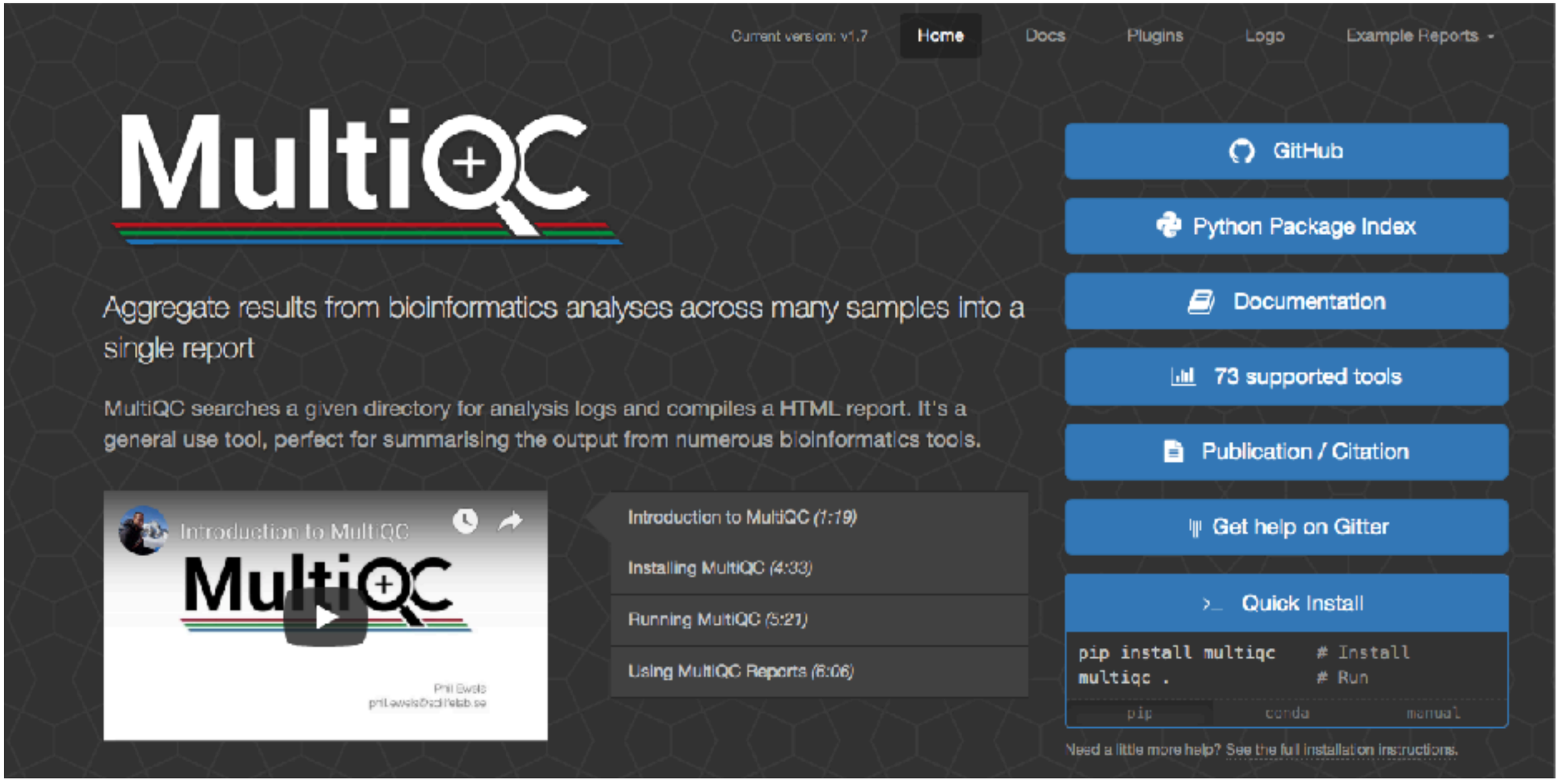


Some tips



Fragment size distribution to detect unusual behaviour in alignment due to adapter contamination (https://github.com/annaquaglieri16/RNA-Seq-and-adapters--STAR-vs-Subjunc)

Use MultiQC!!


Current version: v1.7
Home
Docs
Plugins
Logo
Example Reports
MultiQC
Aggregate results from bioinformatics analyses across many samples into a single report
MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.
Introduction to MultiQC
Introduction to MultiQC (1:19)
Installing MultiQC (4:33)
Running MultiQC (5:21)
Using MultiQC Reports (6:06)
GitHub
Python Package Index
Documentation
73 supported tools
Publication / Citation
Get help on Gitter
Quick Install
pip install multiqc # Install
multiqc . # Run
pip
conda
manual
Need a little more help? See the full installation instructions.

Let's continue…

- How do we need to pre-process the bamfiles?

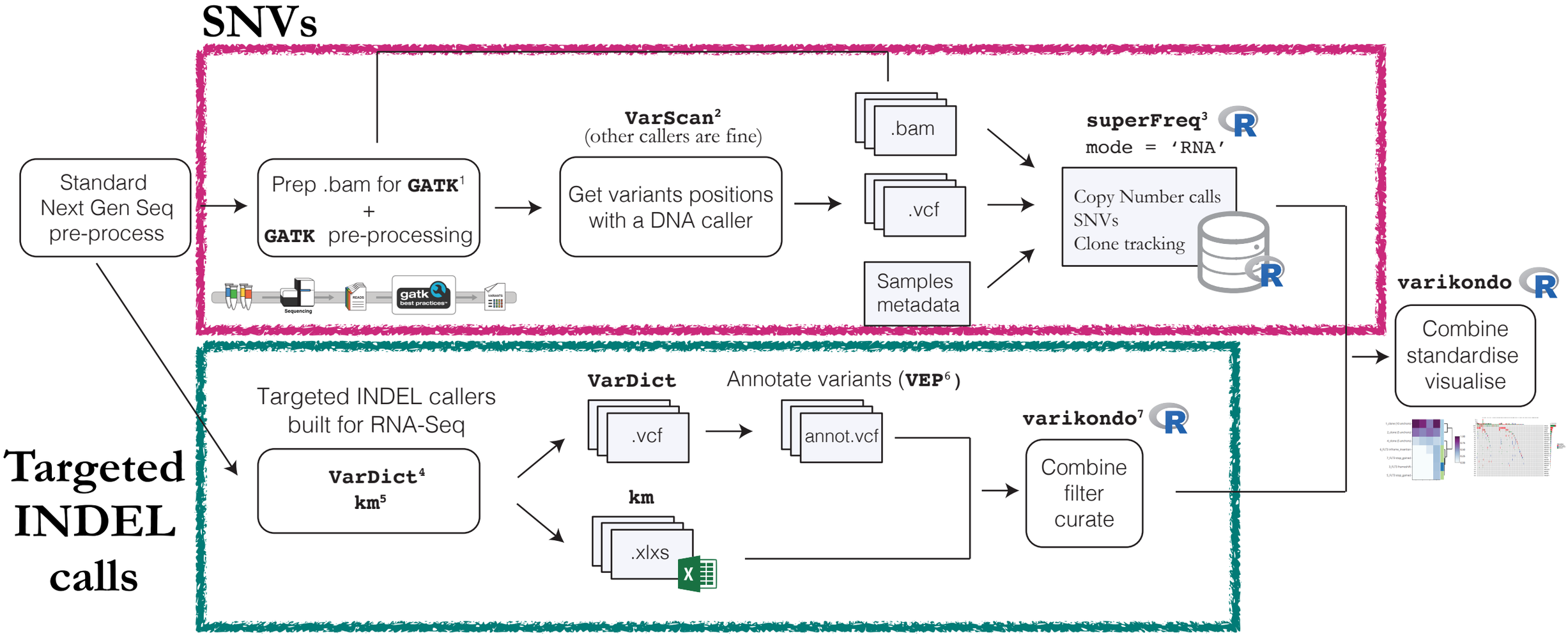
- **We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?**

- How do we filter germline variants without a matched normal?

- How do we track changes in mutations or groups of mutations (clones) in time for a patient?

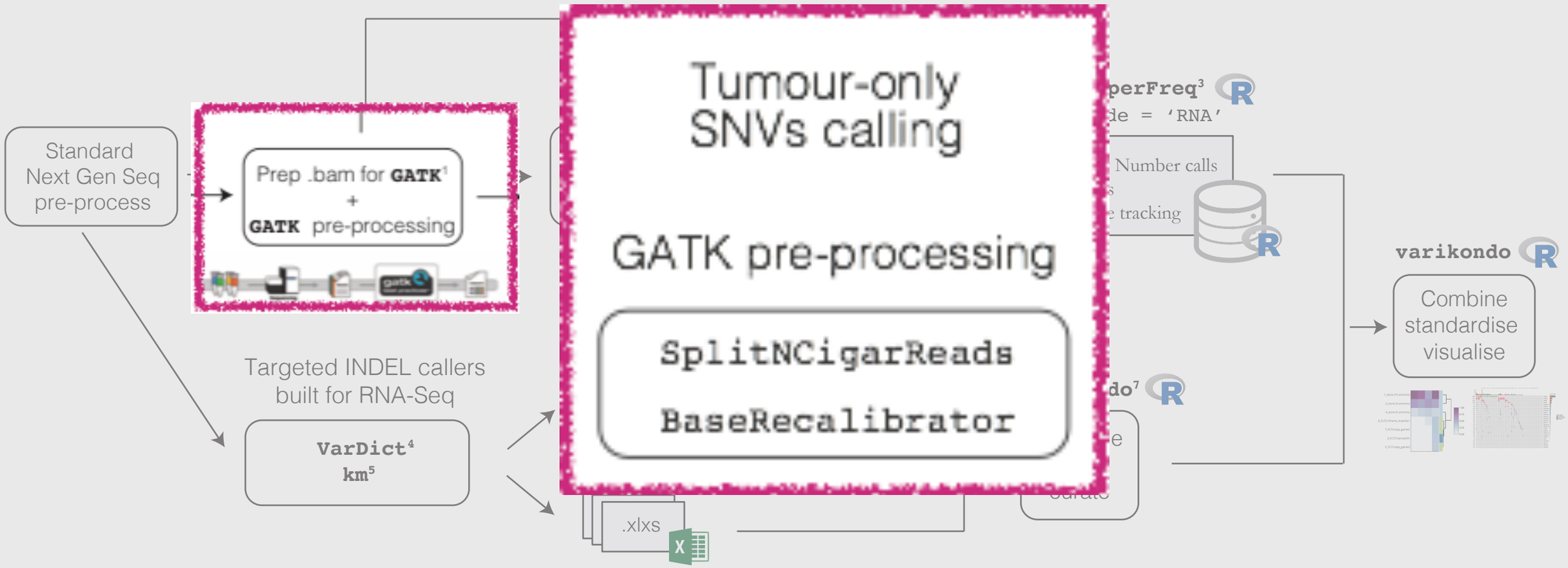
- How to combine, summarise and plot all these results?

Two pipelines

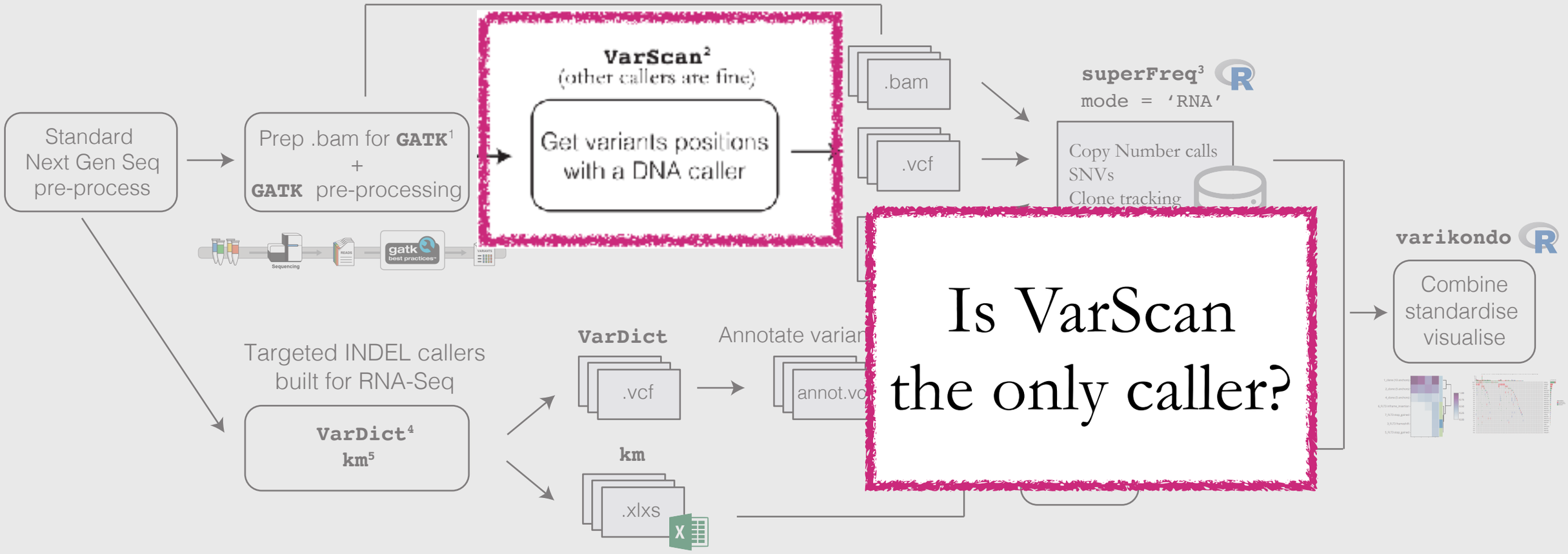


SNV calling

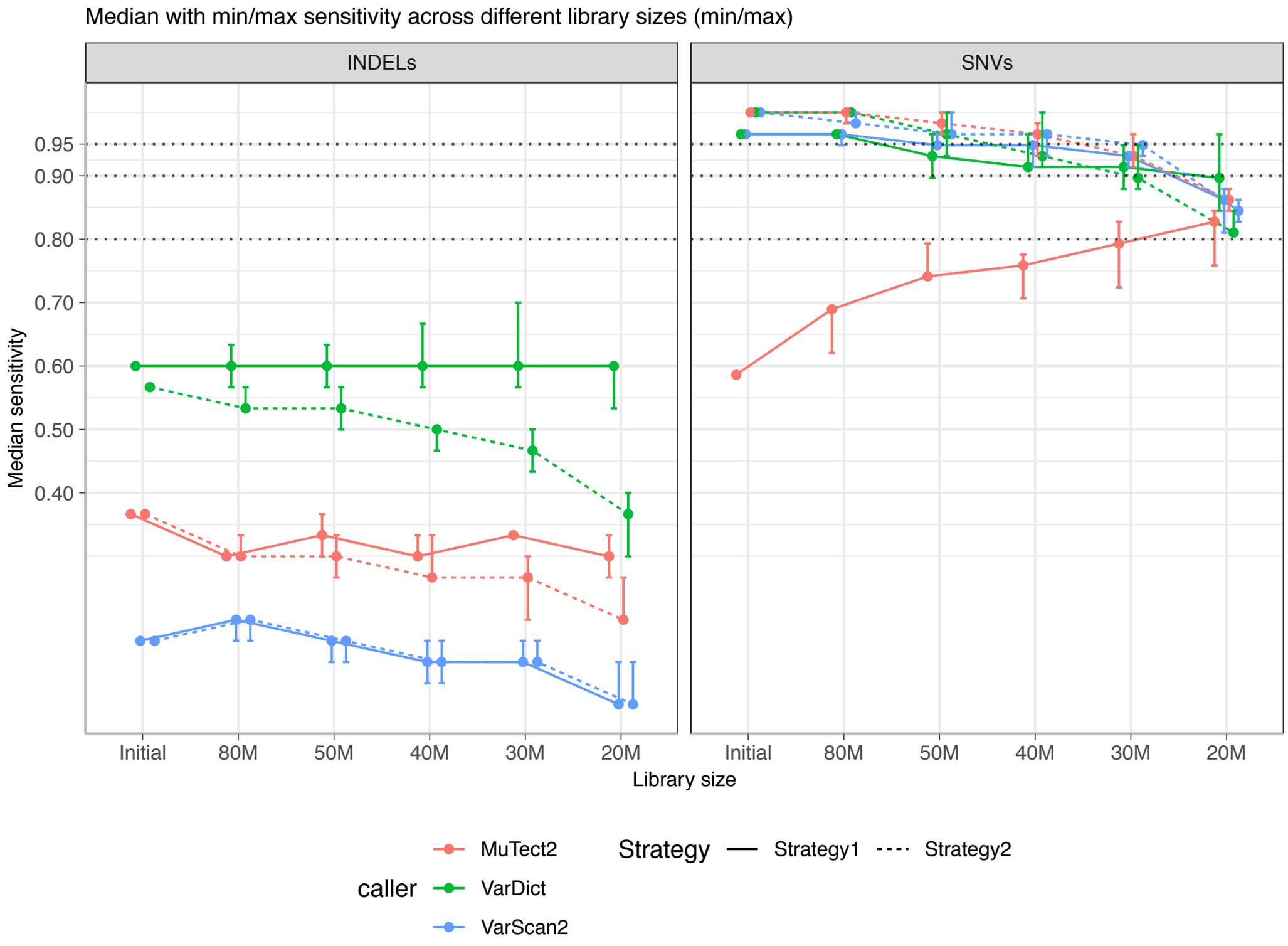
GATK pre-processing



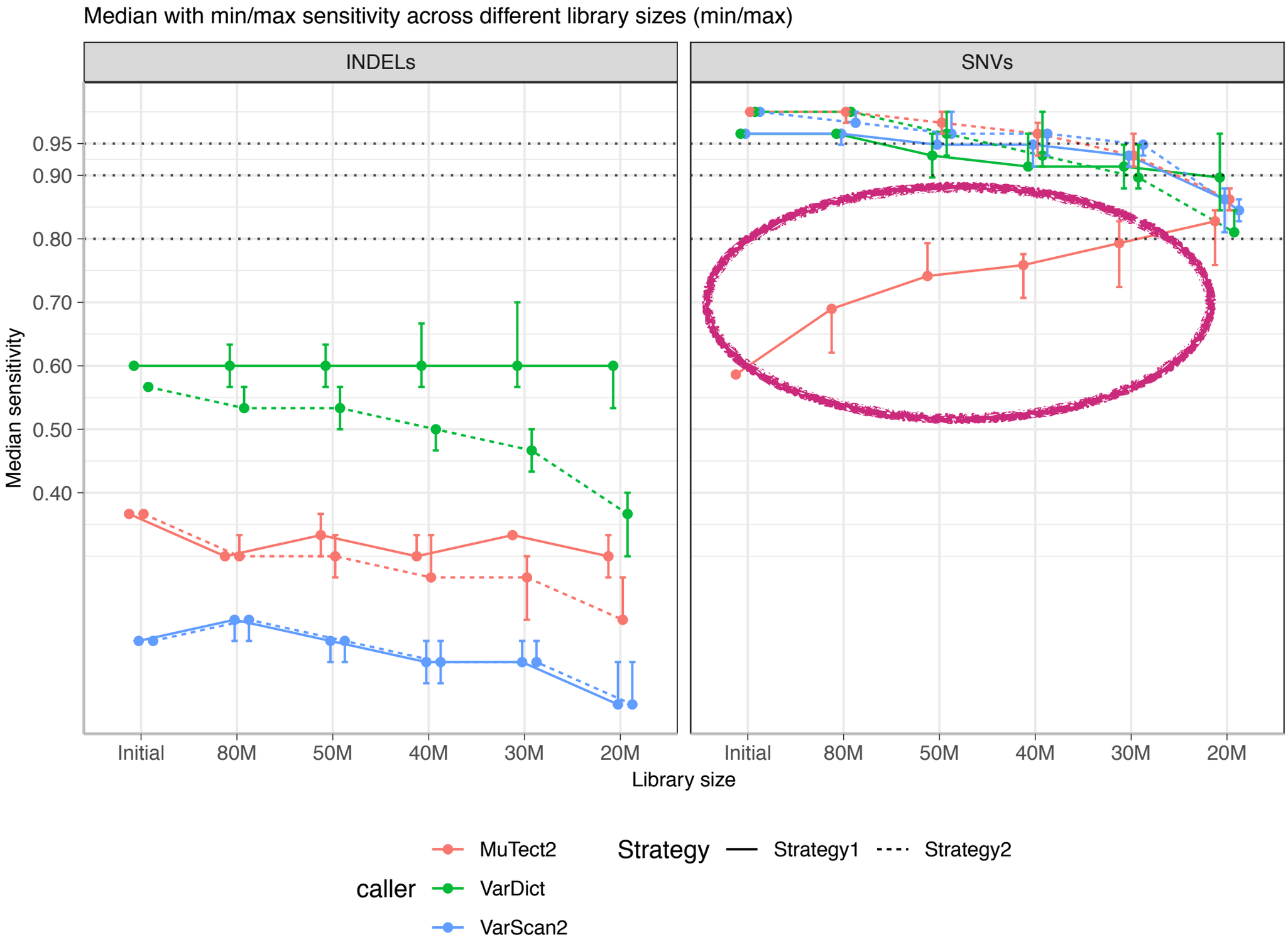
Get SNVs locations



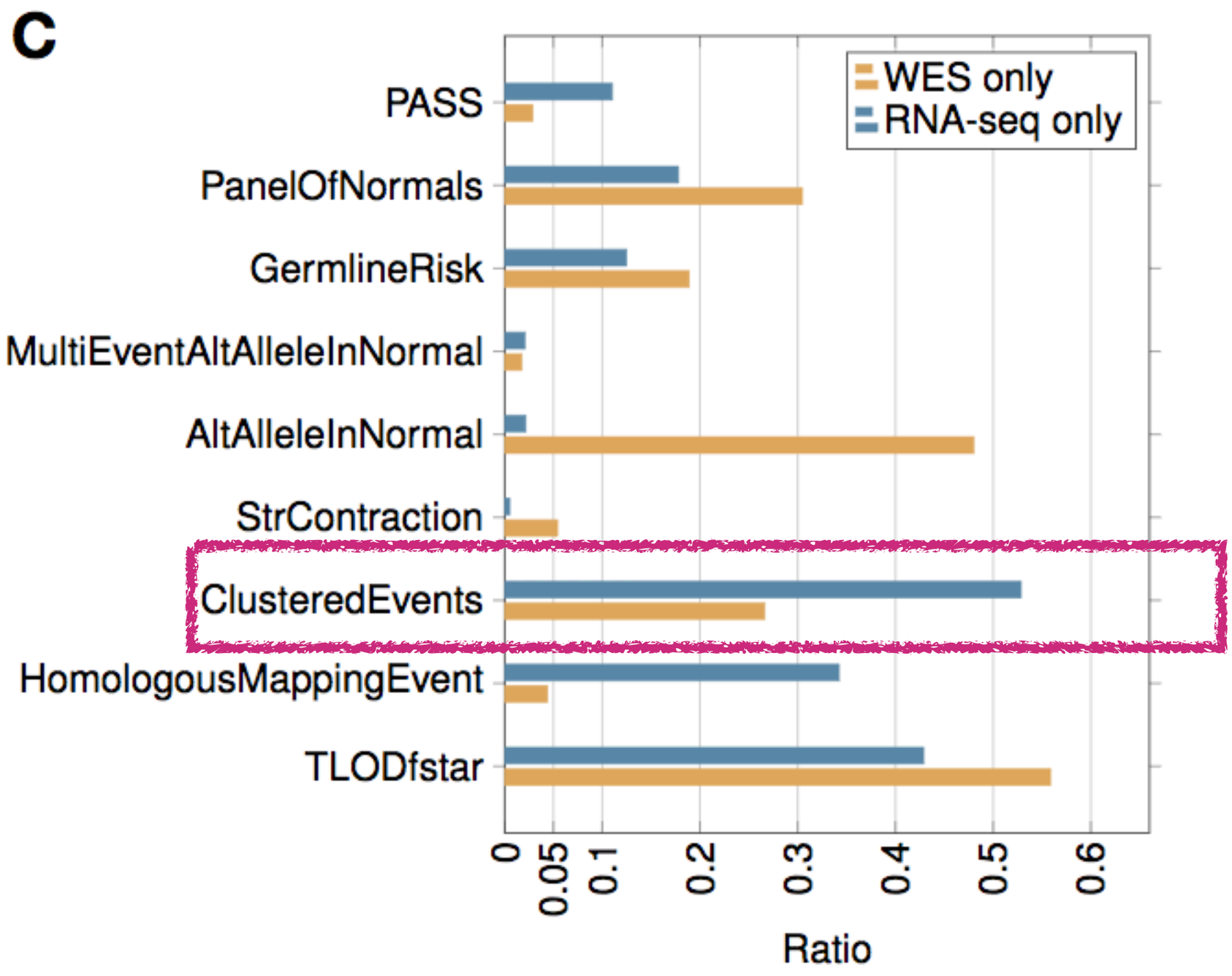
Similar performance for SNVs with other DNA callers



Getting better as library size decreases??



Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data (Coudray et al 2018)

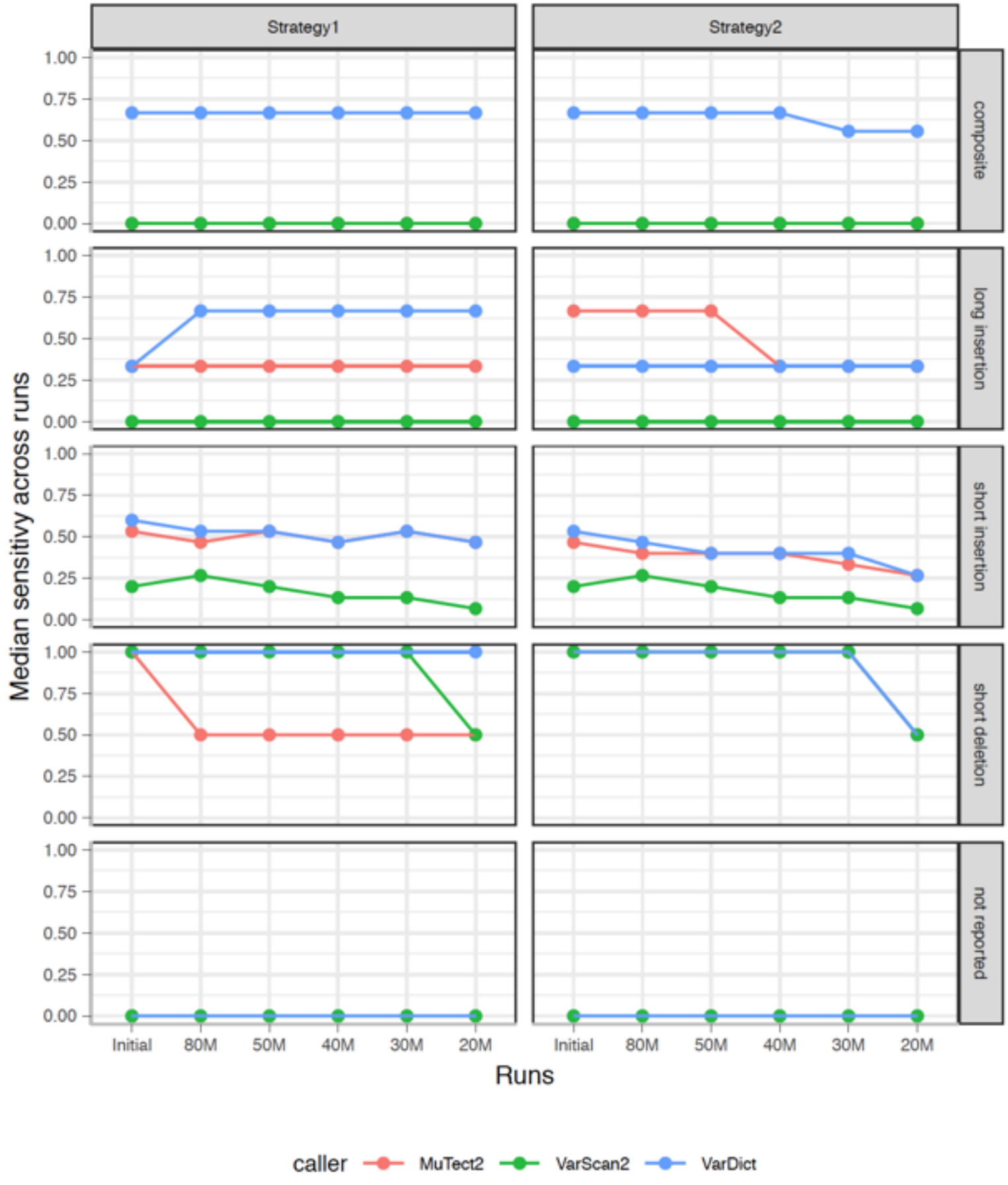


Mutation type	Min VAF	Mean VAF	Max VAF	No. Variants
Composite	0.06	0.25	0.56	9
Long Insertion	0.41	0.50	0.64	3
Short Deletion	0.09	0.24	0.38	2
Short Insertion	0.07	0.33	0.84	15
SNVs	0.05	0.37	0.97	58
Not reported	0.84	0.84	0.84	1

Table 1: Variants detected in Lavallée et al. 2016. A long INDEL involv

variants includes both insertions and deletions at the same time.

Sensitivity Indels by type



INDEL calling

Published online 7 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 11 ***e108***
doi: 10.1093/nar/gkw227

VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research

Zhongwu Lai[1,*], Aleksandra Markovets[1], Miika Ahdesmaki[2], Brad Chapman[3], Oliver Hofmann[3,4], Robert McEwen[2], Justin Johnson[1], Brian Dougherty[1], J. Carl Barrett[1] and Jonathan R. Dry[1]

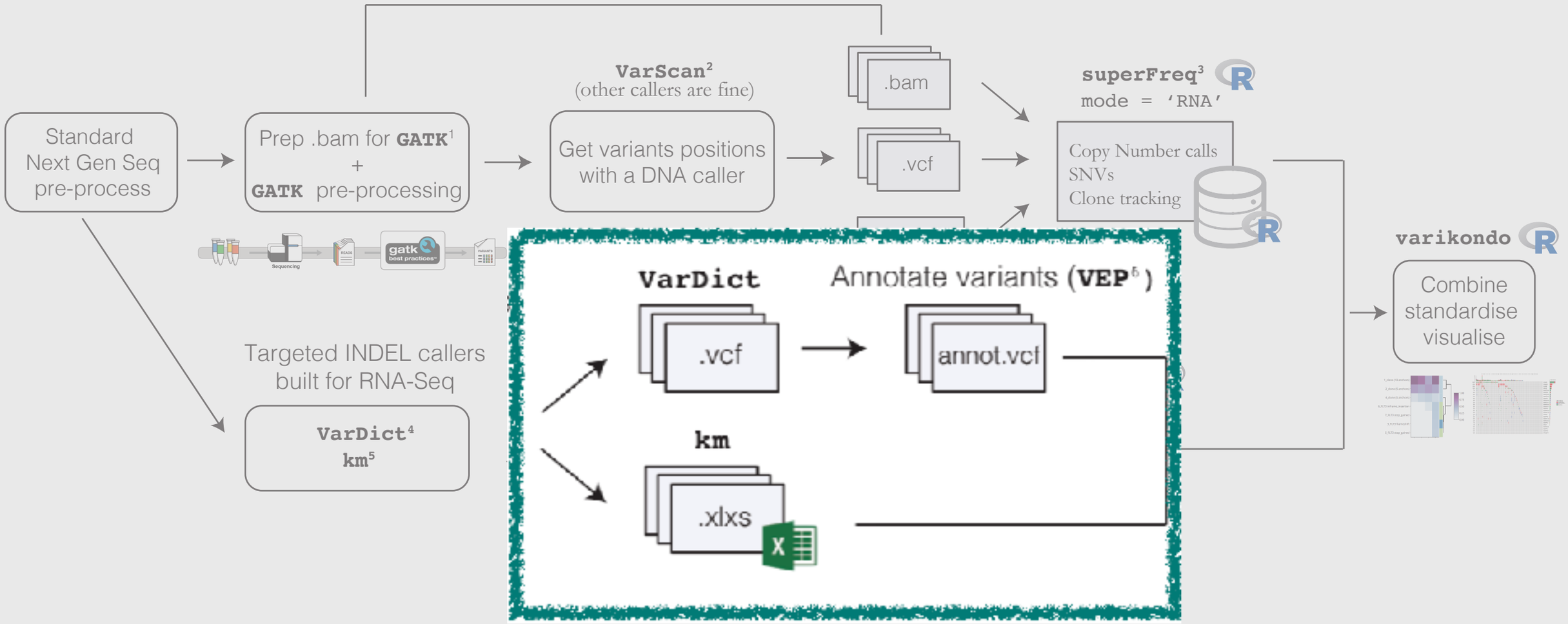
[1]Oncology iMed, AstraZeneca, Waltham, MA 02451, USA, [2]Oncology iMed, AstraZeneca, Cambridge, CB2 0RE, UK, [3]Bioinformatics Core, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA and [4]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Bearsden Glasgow, G61 1QH, UK

Received December 5, 2015; Revised March 4, 2016; Accepted March 22, 2016

Targeted variant detection in leukemia using unaligned RNA-Seq reads

Eric Olivier Audemard[1], Patrick Gendron[1], Vincent-Philippe Lavallée[1,2], Josée Hébert[1,2,4,5], Guy Sauvageau[1,2,4,5], Sébastien Lemieux[1,3*]

INDEL calling outputs

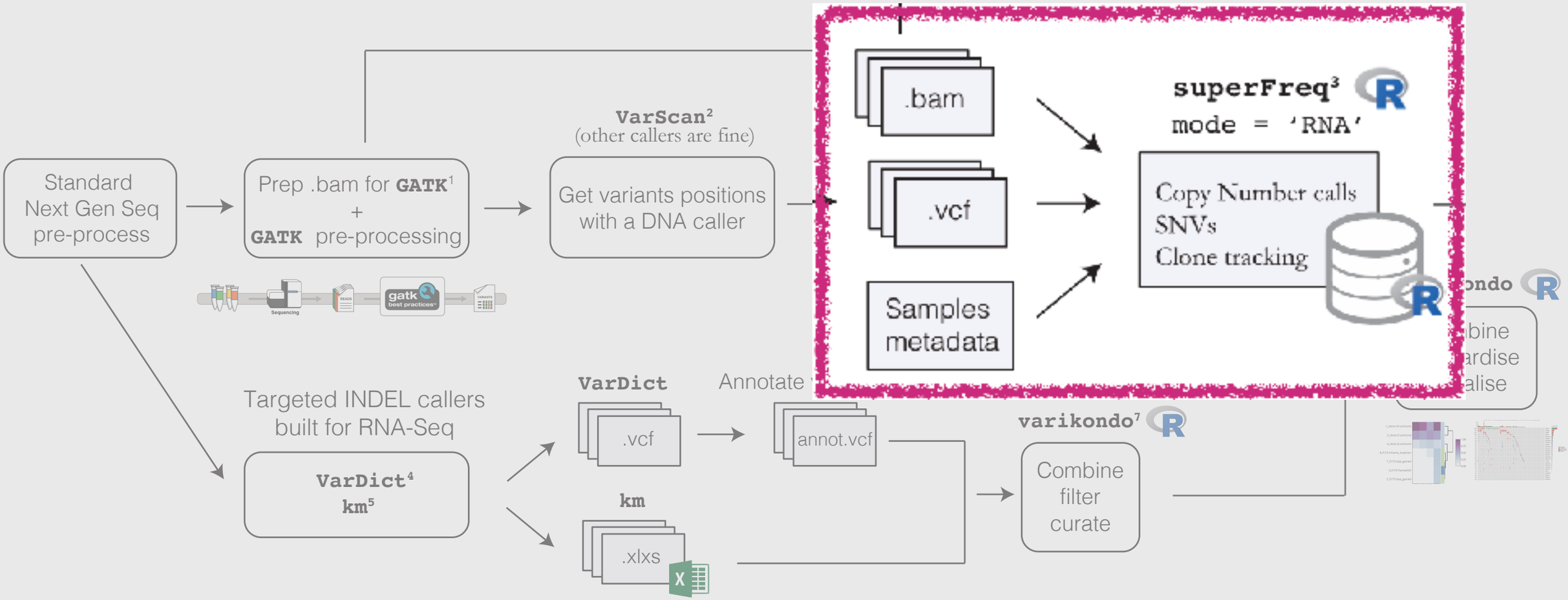


Then…

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- **How do we track changes in mutations or groups of mutations (clones) in time for a patient?**
- How do we filter germline variants without a matched normal?
- How to combine, summarise and plot all these results?

Analyse mutations over time

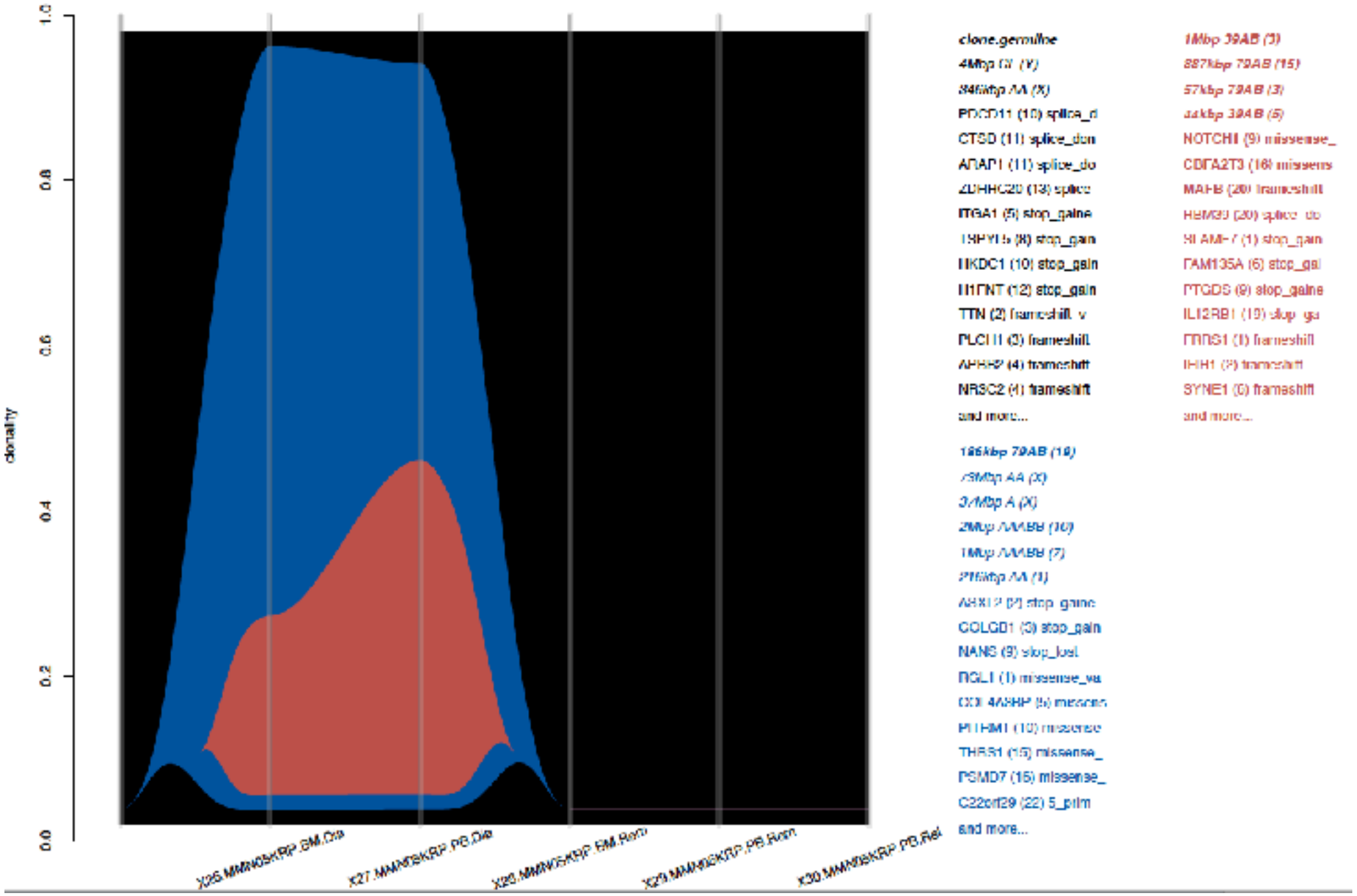
https://github.com/ChristofferFlensburg/superFreq



SuperFreq: Integrated mutation detection and clonal tracking in cancer

Christoffer Flensburg, Tobias Sargeant, Alicia Oshlack, Ian Majewski

doi: https://doi.org/10.1101/380097



Not as easy as for SNVs

Example:

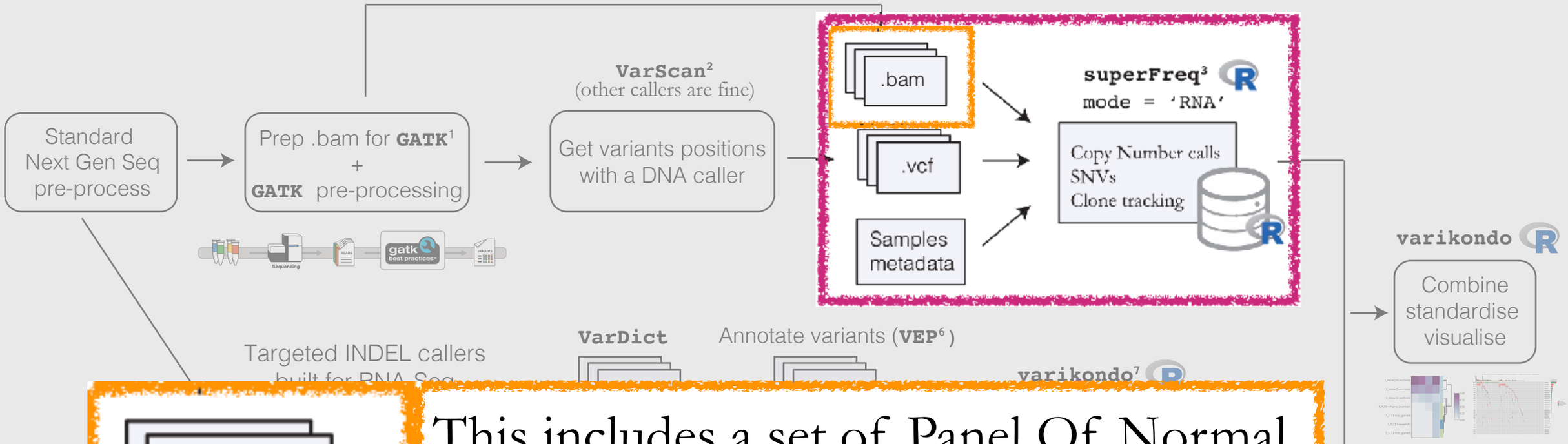
```
            mutation_det                                                                mutation_key SYMBOL
1 NPM1 frameshift_variant chr5_171400152_GTGATGATGATGTGATGACGATGATGATGAAGAGGATGATGATGAAGA   NPM1
2 NPM1 frameshift_variant                  chr5_171400152_GTGATGATGTGAAGAGGATGATGATGAAGA   NPM1
3 NPM1 frameshift_variant                                                          <NA>   NPM1
  D2.Screen.Diag.R1.B2.Resp D2.Cyc1.Rem.R1.B2.Resp
1                    0.7692                    NA
2                        NA                   0.3
3                    0.0000                   0.0
```

Then…

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- How do we track changes in mutations or groups of mutations (clones) in time for a patient?
- **How do we filter germline variants without a matched normal?**
- How to combine, summarise and plot all these results?

Analyse mutations over time

https://github.com/ChristofferFlensburg/superFreq



This includes a set of Panel Of Normal (PON) RNA-Seq from same cell type as my cancer samples

Panel of Normals (PON)

- PON Bamfiles should go through same pre-processing as analyses samples

- Since we are working in RNA-Seq, should be of the same tissue type as what you need to analyse

- SuperFreq removes variants found in PON samples to remove artefacts

More on filtering variants

- Filtering is tricky if only one sample per patient is present, but time-course data could be useful

- superFreq, like other methods, annotates variants found in known databases (dbSNP, ExAC, COSMIC etc..) and these could be used for filtering

- Define filters based on quality threshold like VAF, total depth, quality

  - http://bcb.io/2016/04/04/vardict-filtering/

  - Reliable identification of genomic variants from RNA-seq data (Piskol R et al 2016)

Lastly…

- How do we need to pre-process the bamfiles?
- We are interested in both point mutations (SNVs) and insertions and deletions (INDELs). What callers shall we use?
- How do we track changes in mutations or groups of mutations (clones) in time for a patient?
- How do we filter germline variants without a matched normal?
- **How to combine, summarise and plot all these results?**

Standard
Next Gen Seq
pre-process

Prep .bam for **GATK**[1]
+
GATK pre-processing

VarScan[2]
(other callers are fine)

Get variants positions
with a DNA caller

.bam

.vcf

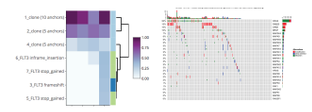
Samples
metadata

superFreq[3]
`mode = 'RNA'`

Copy Number calls
SNVs
Clone tracking

varikondo

Combine
standardise
visualise



Targeted INDEL callers
built for RNA-Seq

VarDict[4]
km[5]

VarDict

.vcf

Annotate variants (**VEP**[6])

annot.vcf

km

.xlxs

varikondo[7]

Combine
filter
curate

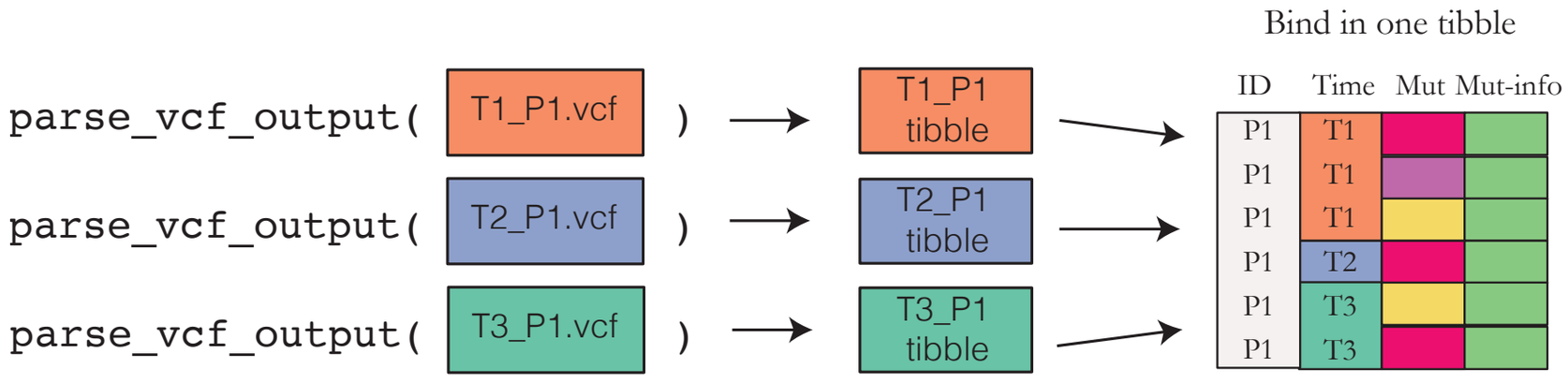
Simply…

`varikondo`

annaquaglieri16.github.io/varikondo/

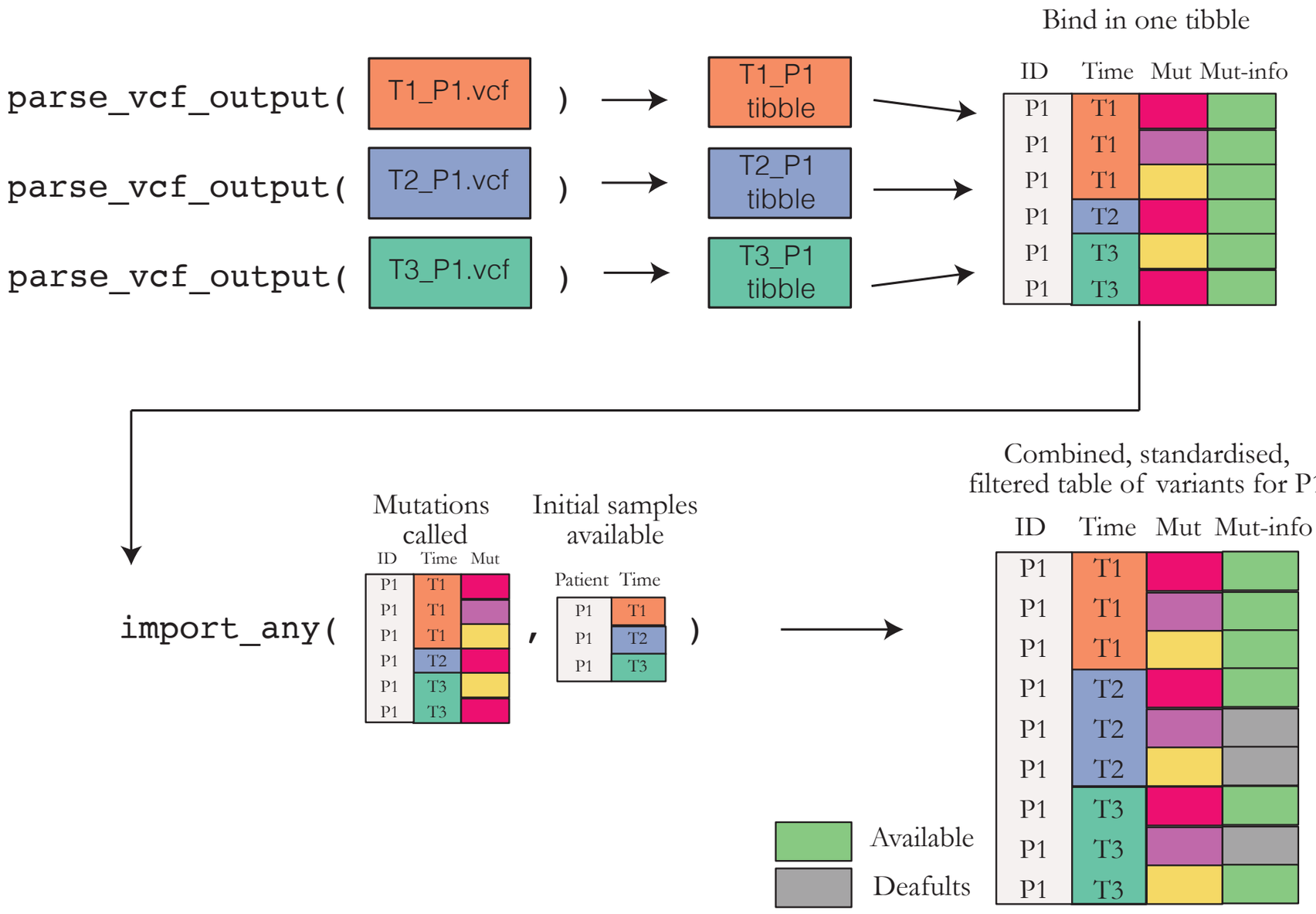
varikondo

Importing and de-cluttering VCF output



varikondo

Importing and de-cluttering VCF output

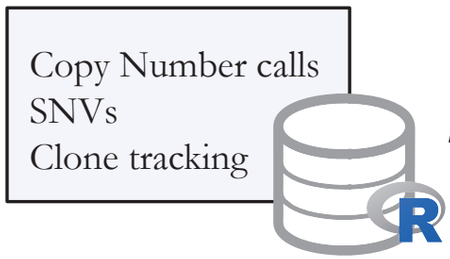


varikondo

Importing and de-cluttering superFreq output

```
import_goi_superfreq(
```

Path to superFreq R folder




```
,
```

studyGenes

```
,
```

Metadata available

Patient	Time
P1	T1
P1	T2
P1	T3

```
,
```

PatientID

```
)
```

varikondo

Importing and de-cluttering superFreq output

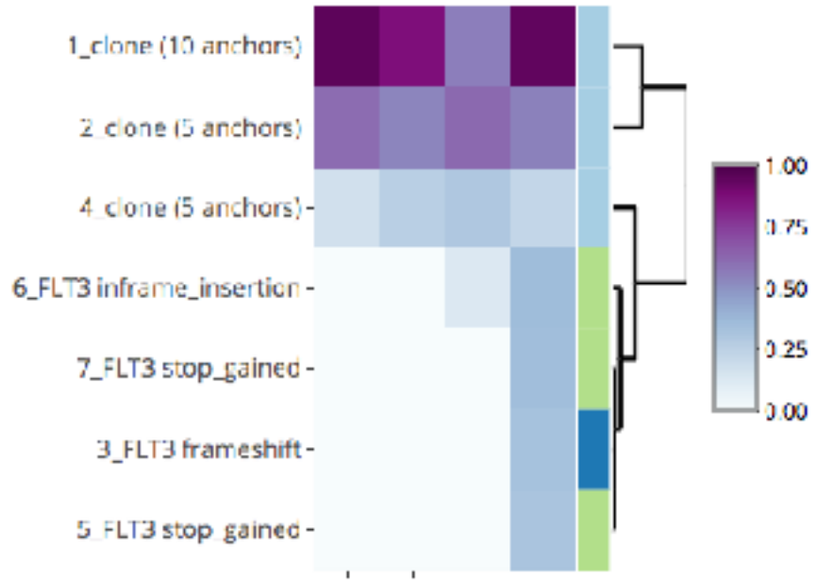
`import_goi_superfreq(` Path to superFreq R folder: Copy Number calls, SNVs, Clone tracking `,` studyGenes `,` Metadata available (Patient, Time: P1 T1, P1 T2, P1 T3) `,` PatientID `)`

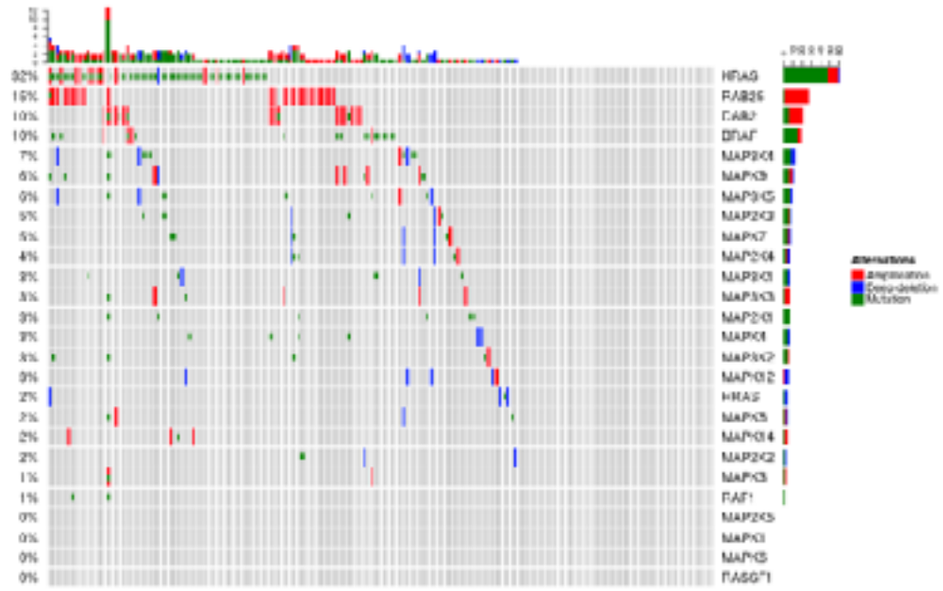
Combined, standardised, filtered table of variants for P1

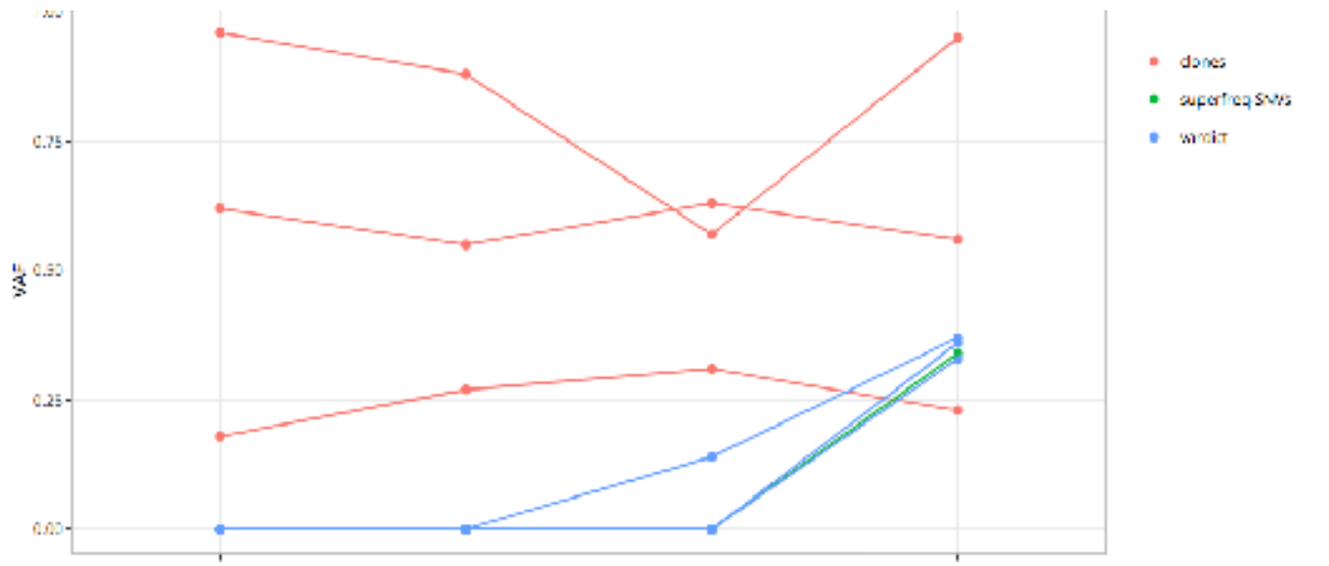
ID	Time	Mut	Mut-info
P1	T1	SNV	
P1	T1	CNV	
P1	T1	Clone	
P1	T2	SNV	
P1	T2	CNV	
P1	T2	Clone	
P1	T3	SNV	
P1	T3	CNV	
P1	T3	Clone	

And finally spark joy!

ID	Time	Mut	Mut-info
P1	T1		
P1	T1		
P1	T1		
P1	T2		
P1	T2		
P1	T2		
P1	T3		
P1	T3		
P1	T3		
P1	T1	SNV	
P1	T1	CNV	
P1	T1	Clone	
P1	T2	SNV	
P1	T2	CNV	
P1	T2	Clone	
P1	T3	SNV	
P1	T3	CNV	
P1	T3	Clone	







Limitations

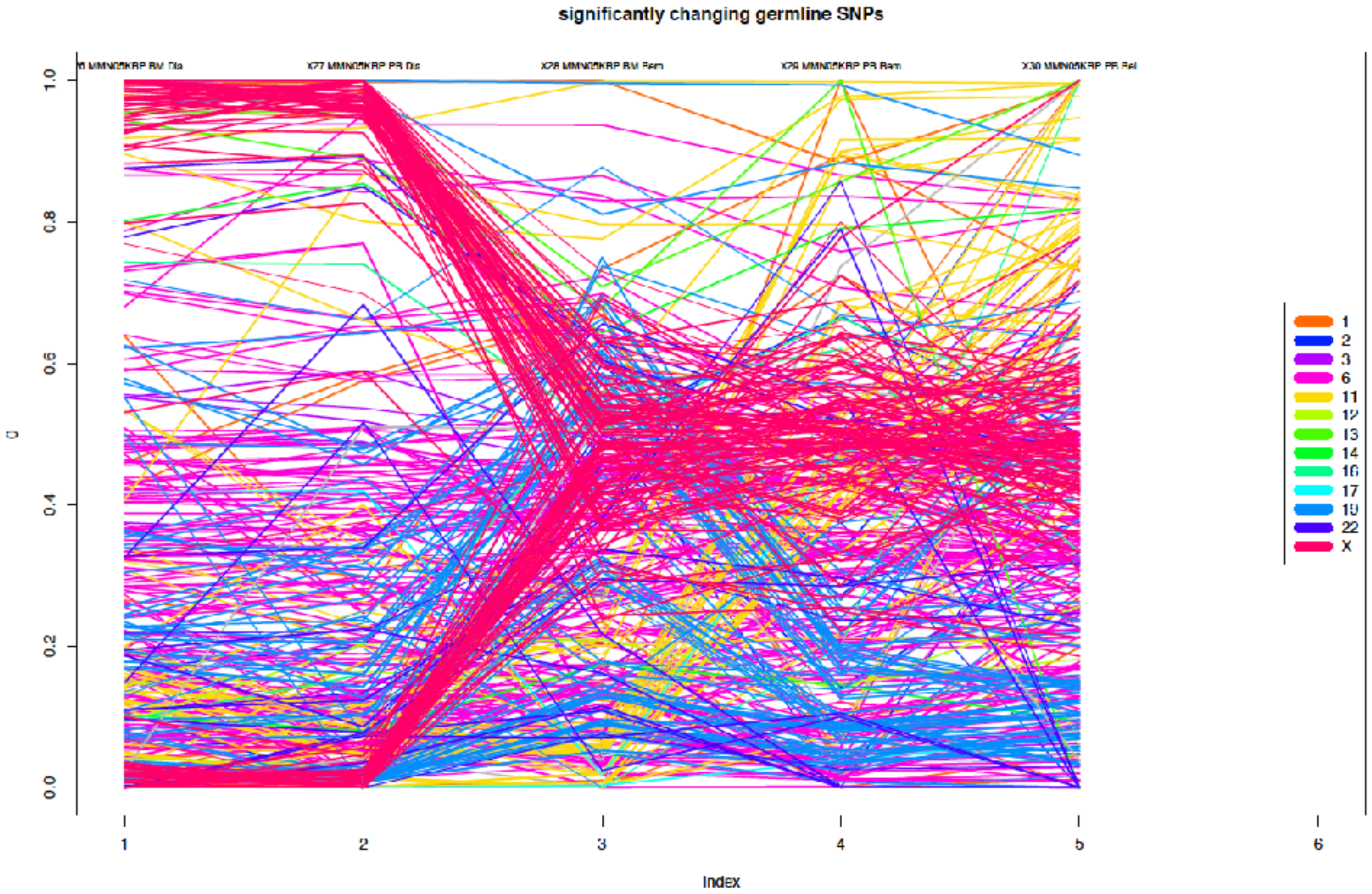
- Expression varies on genes and we cannot detect mutations in non-expressed genes

- Non-sense mediated decay mutations, RNA gets "eaten" very quickly, decays and we can't detect the mutations!

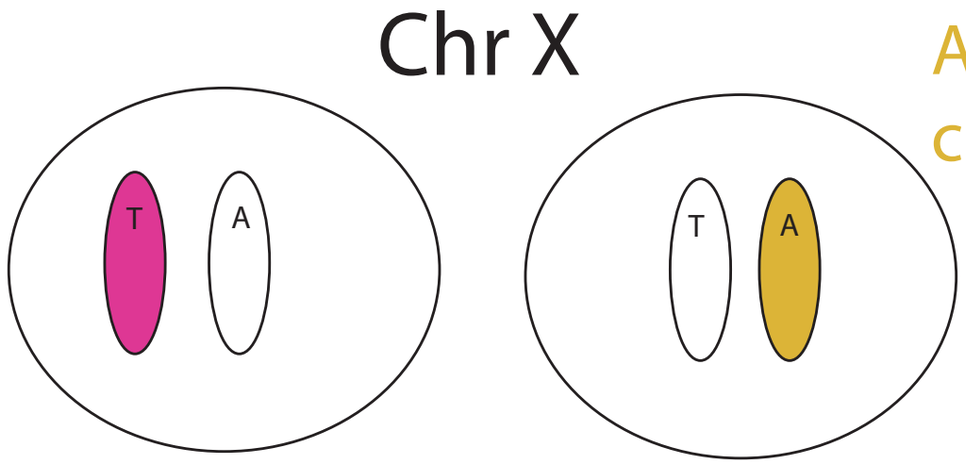
- Allele specific expression. There is a DNA mutation but only one allele is expressed

- Less precise estimate of the variant allele frequency for one mutation

Curiosity!


significantly changing germline SNPs
X27 MMN05KRP PR Dis
X28 MMN05KRP RV Rem
X29 MMN05KRP PR Rem
X30 MMN05KRP PR Rel
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Index
1
2
3
6
11
12
13
14
16
17
19
22
X

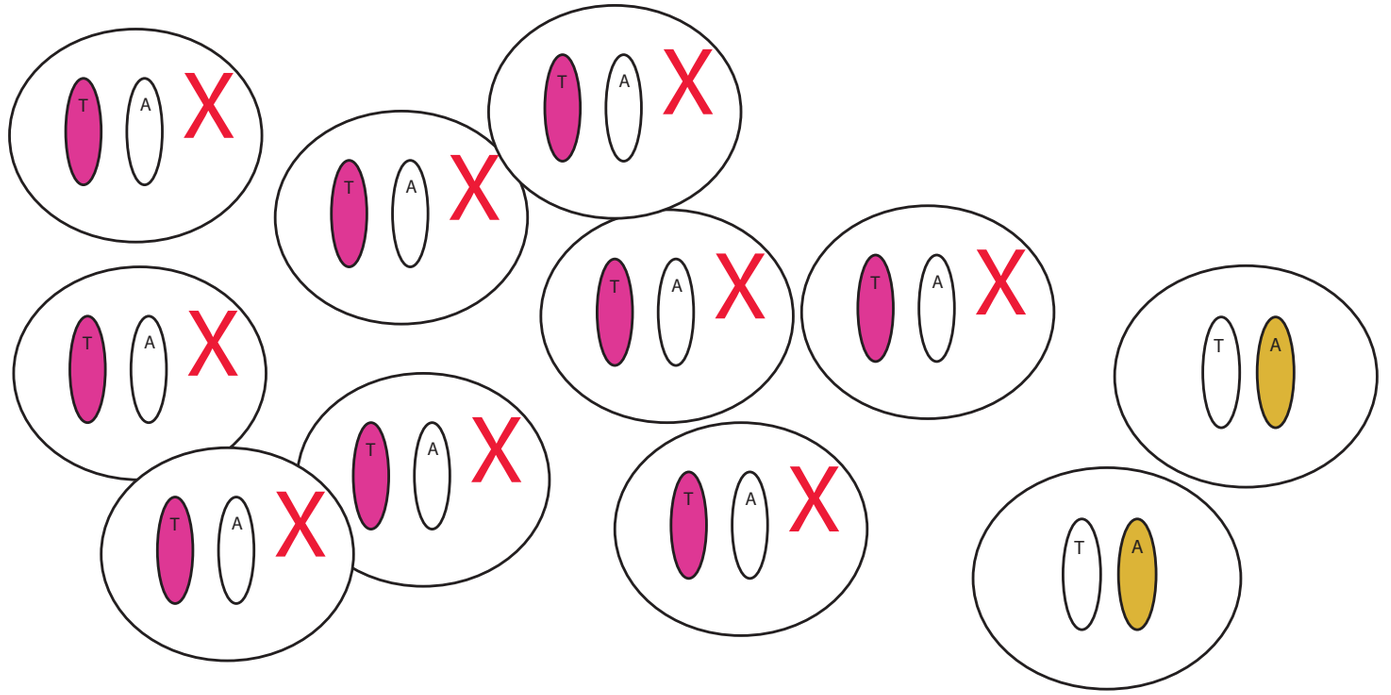
Activated chromosome



Activated chromosome

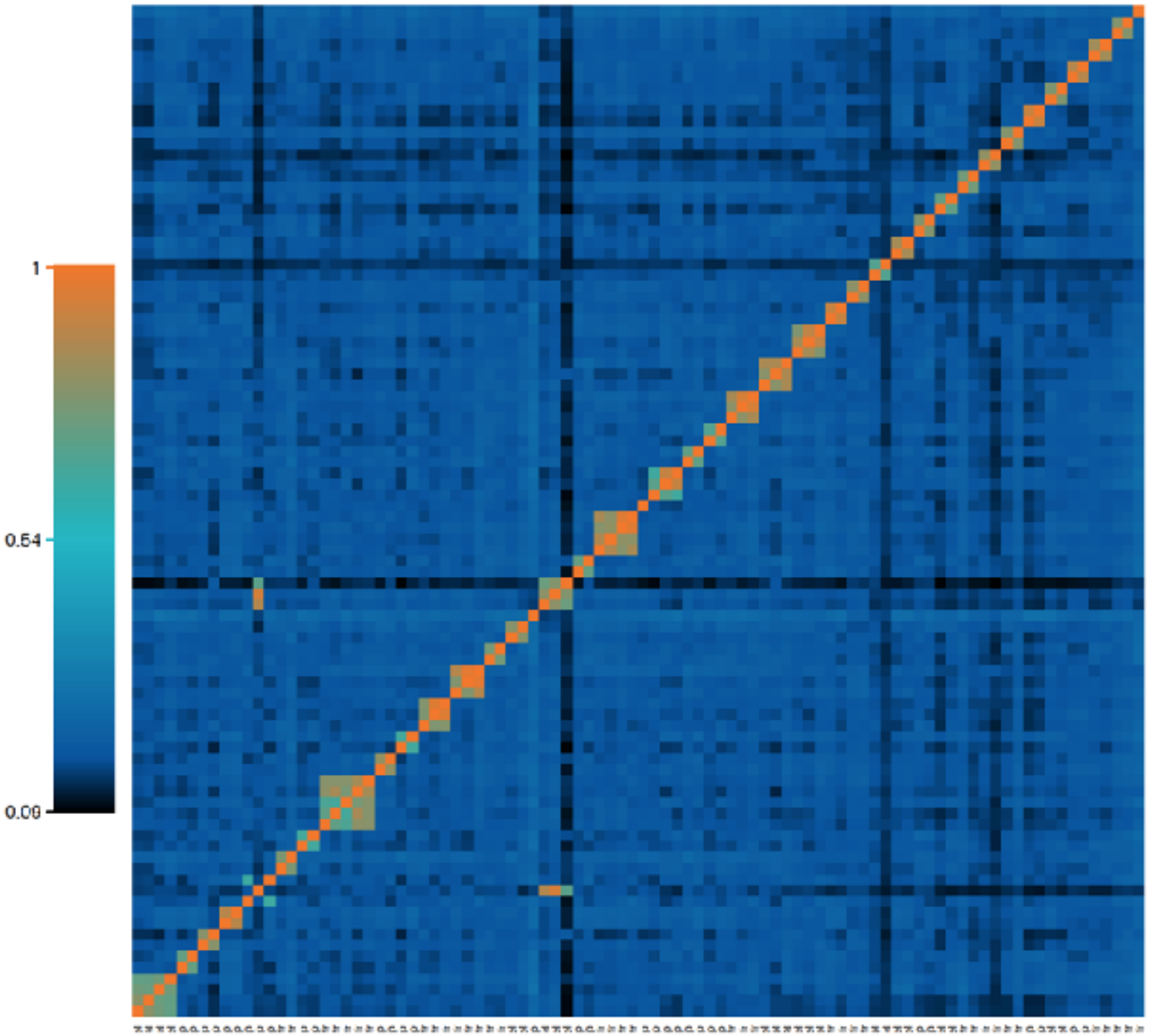
T A
T A
T A
T A
T A

Uncontrolled division
Cancer clone



T A
T A
T
T
T
T
T
T
T
T
T
T
T
T
T

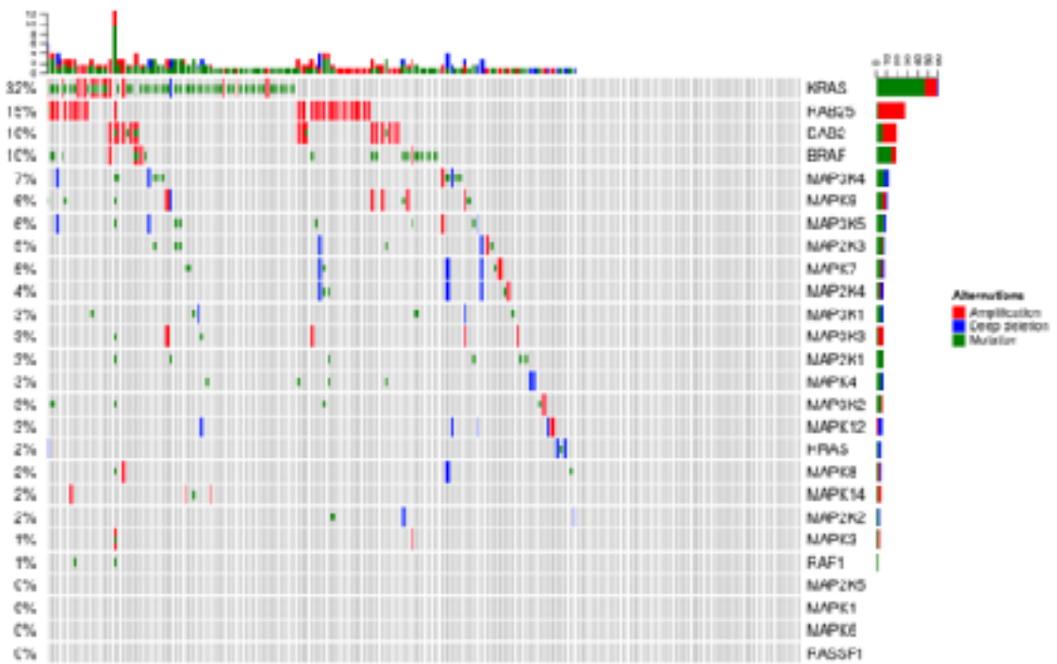
Check relatedness using SNPs



Further additions

- Finish writing a Shiny App included in `varikondo` that allows quick exploration of the de-cluttered variants

- Add a function that transforms the tidy formatted variants into an input for oncoprint to easily produce publication ready plots mutations including the whole cohort



Variant calling in single cells normal RNA

Suggestions:

- Cardelino: clone and donor identification from single-cell data
- Davis McCharty: The effects of DNA variation on gene expression in single cells (Friday, 12pm Agar Theatre, BioSciences 4 Building, The University of Melbourne)

Acknowledgements

- Christoffer Flensburg
- Terry Speed
- Ian Majewki
- Stuart Lee
- Dharmesh Bhuva

- Earo Wang

superFreq RNA mode

- The two main differences are:
    - sometimes in exomes some regions have a strangely high number of reads and these regions would excluded. In RNA this is not as easy to define strange regions since it could be simply due to very high expression
    - using so called HK genes to define CNVs. HK genes are highly expressed and fairly constant in the normals
    - in CN calls with limma, puts more confidence on HK genes and less on the more variable ones (larger CI)
        - can this be a problem when using cancer samples? difference expression from normals? normals were processed differently.. there could in theory be a massive batch effect