CAMBRIDGE Cancer CENTRE

Cambridge science is
changing the way we treat cancer

Somatic SNV Calling

Andy Lynch

CRUK CI

July 2016

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Cambridge University Hospitals NHS
NHS Foundation Trust

Outline

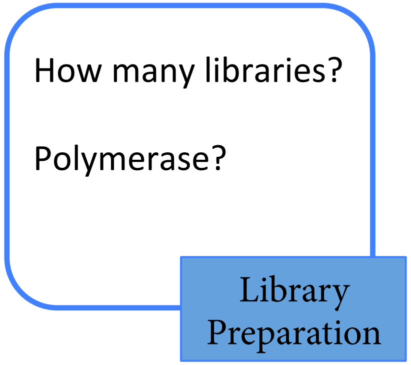
This morning’s session:

- **An example processing pipeline**
- **Some calling tools**
- **How well should you expect a tool to perform?**
- **Some special cases**

Somatic SNV calling

An example processing pipeline

A processing pipeline

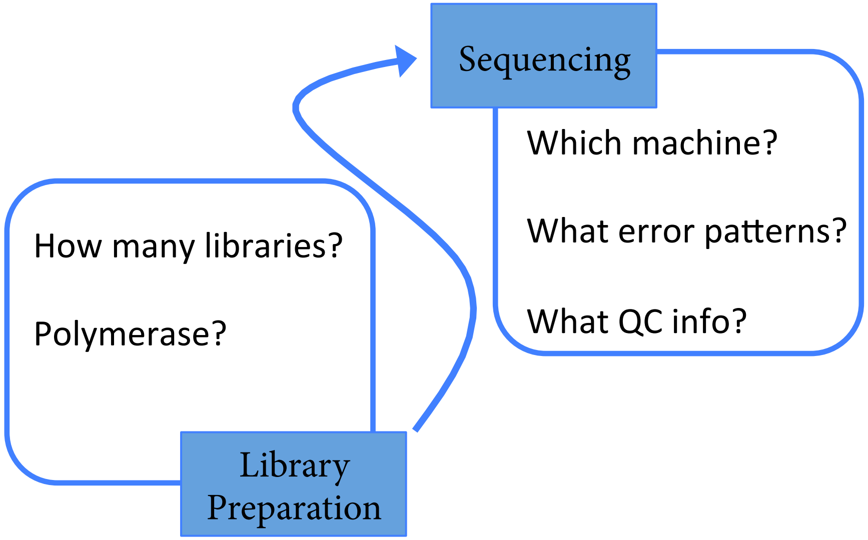


How many libraries?

Polymerase?

Library Preparation

A processing pipeline



Sequencing

Which machine?

What error patterns?

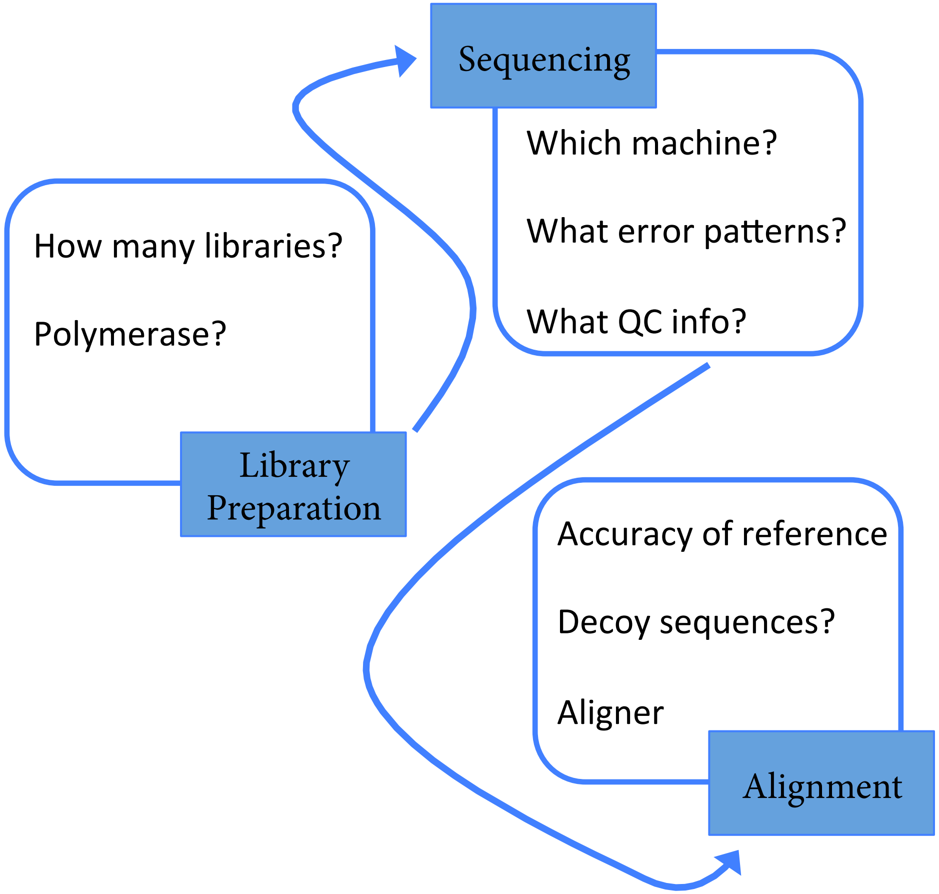
What QC info?

How many libraries?

Polymerase?

Library Preparation

A processing pipeline



Library Preparation

How many libraries?

Polymerase?

Sequencing

Which machine?

What error patterns?

What QC info?

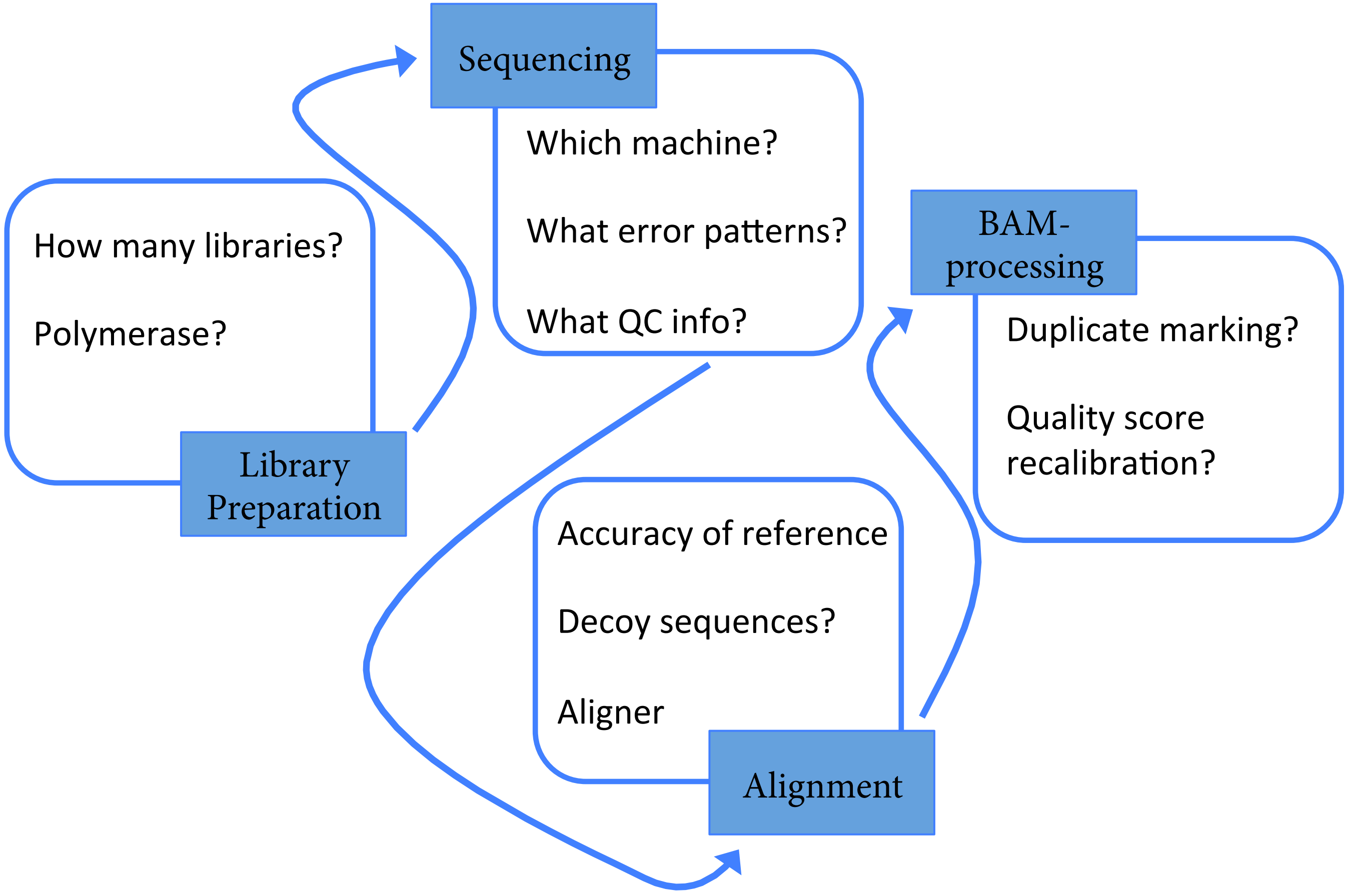
Alignment

Accuracy of reference

Decoy sequences?

Aligner

A processing pipeline



Library Preparation

How many libraries?

Polymerase?

Sequencing

Which machine?

What error patterns?

What QC info?

Alignment

Accuracy of reference

Decoy sequences?

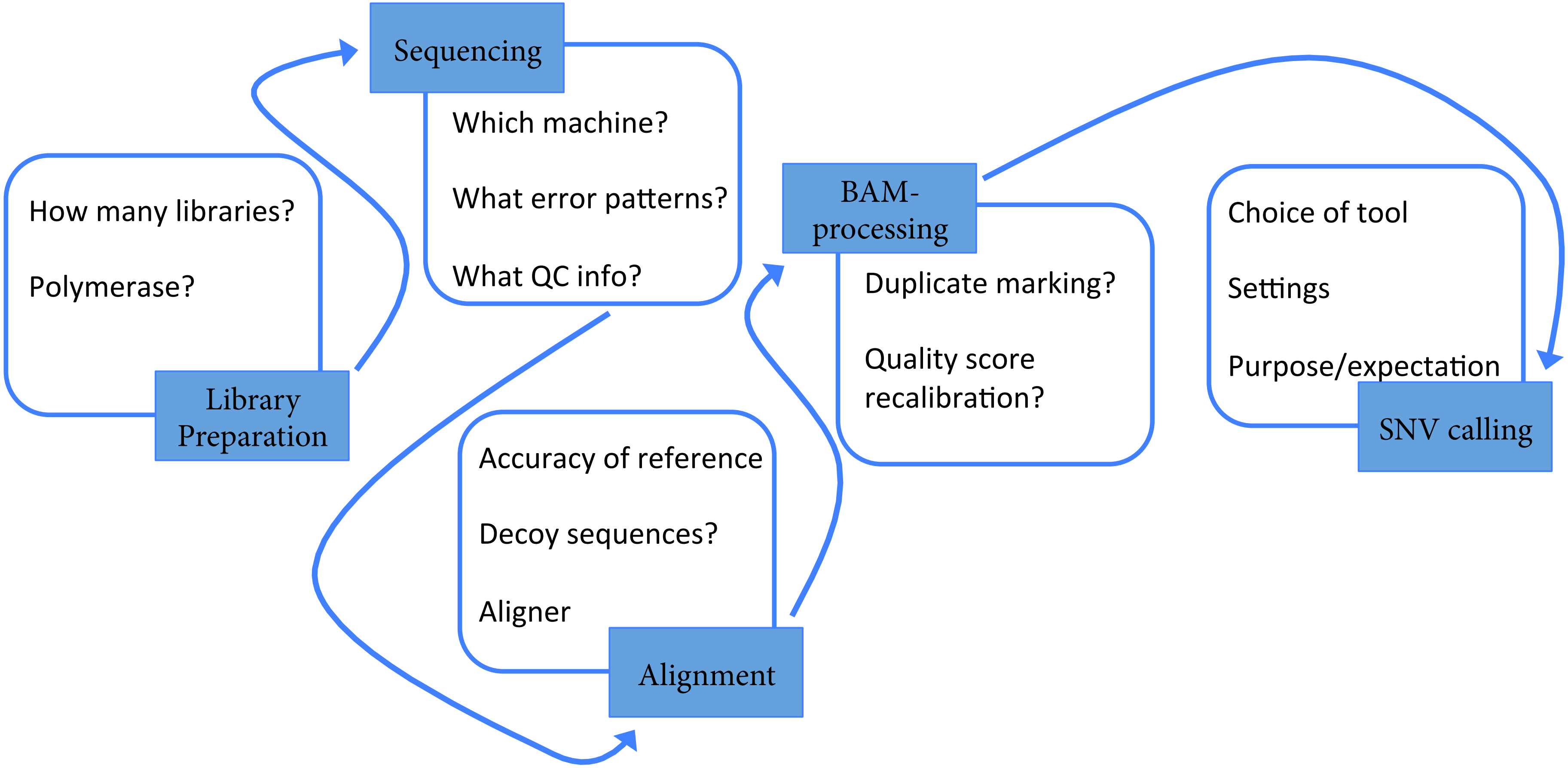
Aligner

BAM-processing

Duplicate marking?

Quality score recalibration?

A processing pipeline



Library Preparation

How many libraries?

Polymerase?

Sequencing

Which machine?

What error patterns?

What QC info?

Alignment

Accuracy of reference

Decoy sequences?

Aligner

BAM-processing

Duplicate marking?

Quality score recalibration?

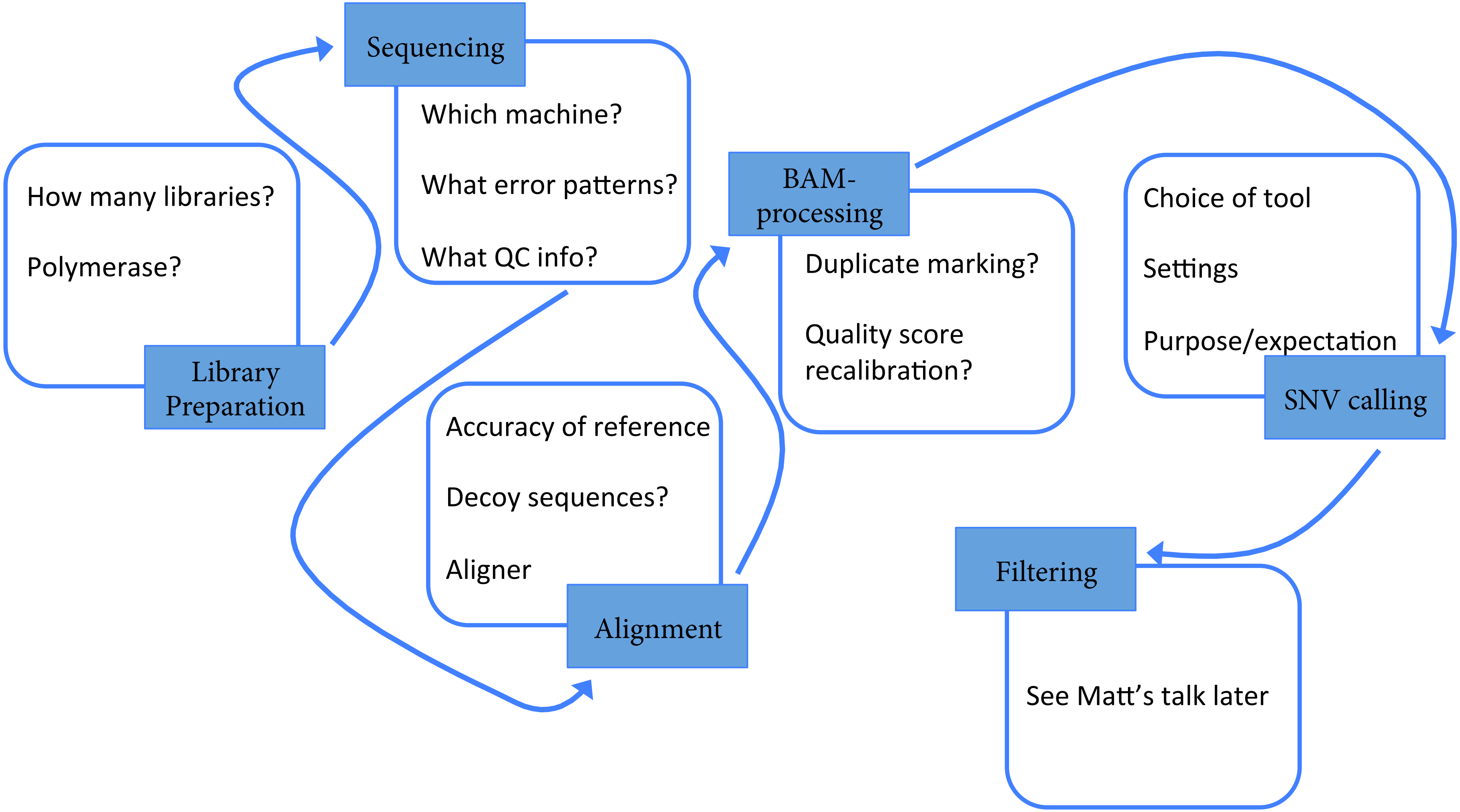
SNV calling

Choice of tool

Settings

Purpose/expectation

A processing pipeline



Library Preparation

How many libraries?

Polymerase?

Sequencing

Which machine?

What error patterns?

What QC info?

Alignment

Accuracy of reference

Decoy sequences?

Aligner

BAM-processing

Duplicate marking?

Quality score recalibration?

SNV calling

Choice of tool

Settings

Purpose/expectation

Filtering

See Matt’s talk later

Somatic SNV calling

SNV-calling tools

The SNV caller is not the only concern

We use CaVEMan here.

Caveat – full details of CaVEMan are not explicitly reported anywhere, and I am not going to go through the code (java). So all a bit of a black box

- **Seems to have a sensible Bayesian model**
- **Considers base quality, read position, lane, and read orientation**
- **Can make use of copy number profiles**
- **Associated filters**

One could argue that any sensible caller would do the job. The secret is in the filtering.

Other tools

Several tools worth considering:

The detail of

- **MuTect2** – Combines a good quality caller with haplotype reassembly. Built in filters and the ability to take in a panel of normal samples. Can also return indels.
- **VarScan2 (Koboldt 2012)** – Uses a basic statistical test rather than a full Bayesian model, but will probably be followed by filtering anyway. A portable java program.
- **Strelka (Saunders 2012)** – A hierarchical model of allele frequencies. Also returns indels.
- **SMuFin (Moncunill 2014)** – A reference free variant caller with high specificity.

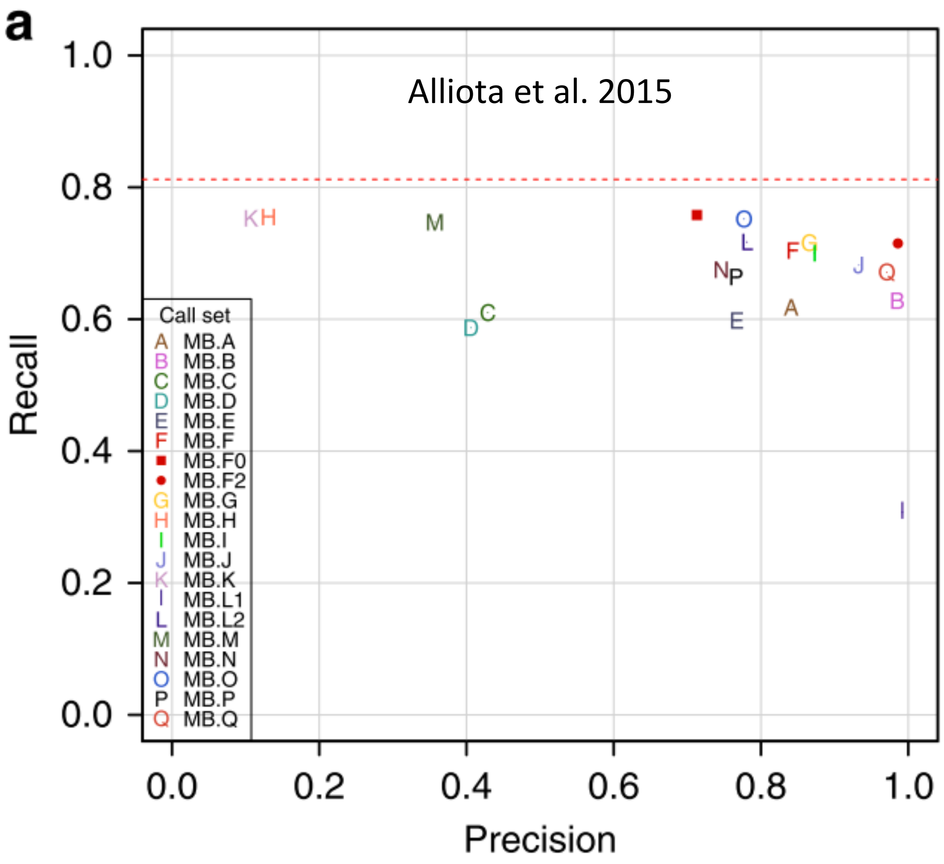
Somatic SNV calling

Anticipated performance

How well should we expect a tool to perform?

Precision/recall is a function of the biology, the sample/data quality, and the calling method.

- **When you see a tool promising a particular precision/recall, how will it perform for you?**
- **The precision and recall are averaged over many SNVs and typically over many samples (although not in this illustration)**
- **The two things affecting precision and recall are the ability to deal with artefacts, and the genuine power of the study**
- **Typically we have the power to detect SNVs that are present in two or more copies in every cell, but those in one copy, or sub-clonal, are more often missed**
- **A sample with relatively more of these recent events will have lower recall/precision**
- **A sample with generally low power will have lower recall/precision**


Alliota et al. 2015

Somatic SNV calling

Some special cases

What if you have multiple tumour samples

Exploring heterogeneity

Datasets that are more that Tumour-Normal are increasingly common. What can be done for variant calling in them?

Theoretically, we can draw strength from related samples to improve our sensitivity.

Still require a filtering regime afterwards.

- **VarScan2 (Koboldt 2012)** offers the ability to call over multiple samples, but doesn’t appear to make the best use of structure in those calls

- **FreeBayes (Garrisson 2012)** can be applied to this task, but the set-up is not optimized for this scenario

- **Platypus (Rimmer 2014)** can be applied to this task. Although it is primarily a germline caller, it does a good job

- **multiSNV (Josephidou 2015)** was designed specifically for the task. It works particularly well in combination with Platypus

What if you have RNA-seq data?

Things get trickier.

We need to stop worrying about recall – there will be a lot missed, and splicing activity and post-transcriptional modifications will introduce artefacts that require new filters.

Nevertheless, there are data to be interrogated…

- **Tophat (Kim et al. 2013) + Isaac (Raczy 2013) variant caller**. Isaac not specifically designed for RNA-seq.

- **MAP-RSeq (Kalari et al. 2014).** Tophat + GATK-based approach**.** Large suite of tools – not a nimble solution.

- **RNASEQR (Chen et al. 2011).** A Bowtie-based approach that takes several passes at the alignment to remove splice-site driven artefacts. Low precision?

- **SNPiR (Piskol et al. 2013).** More expensive aligner to address the problems. Not really designed for somatic variants. See also SNVQ.

- **GLMVC (Sheng et al. 2016).** Specifically for somatic. Addresses cycle bias, but this could be filtered later.

What if you have no matched normal

Obvious strategies:

- **Treat the sample as if it were a normal sample in which you were calling variants. Cellularity allowing, it is probable that somatic events will look like germline heterogeneous SNPs.**

- **Use a relative's, or ethnically-matched, normal sample and run as a T:N pair.**

Either approach will lead to an excess of a couple of million calls, so filtering is required

- **dbSNP**

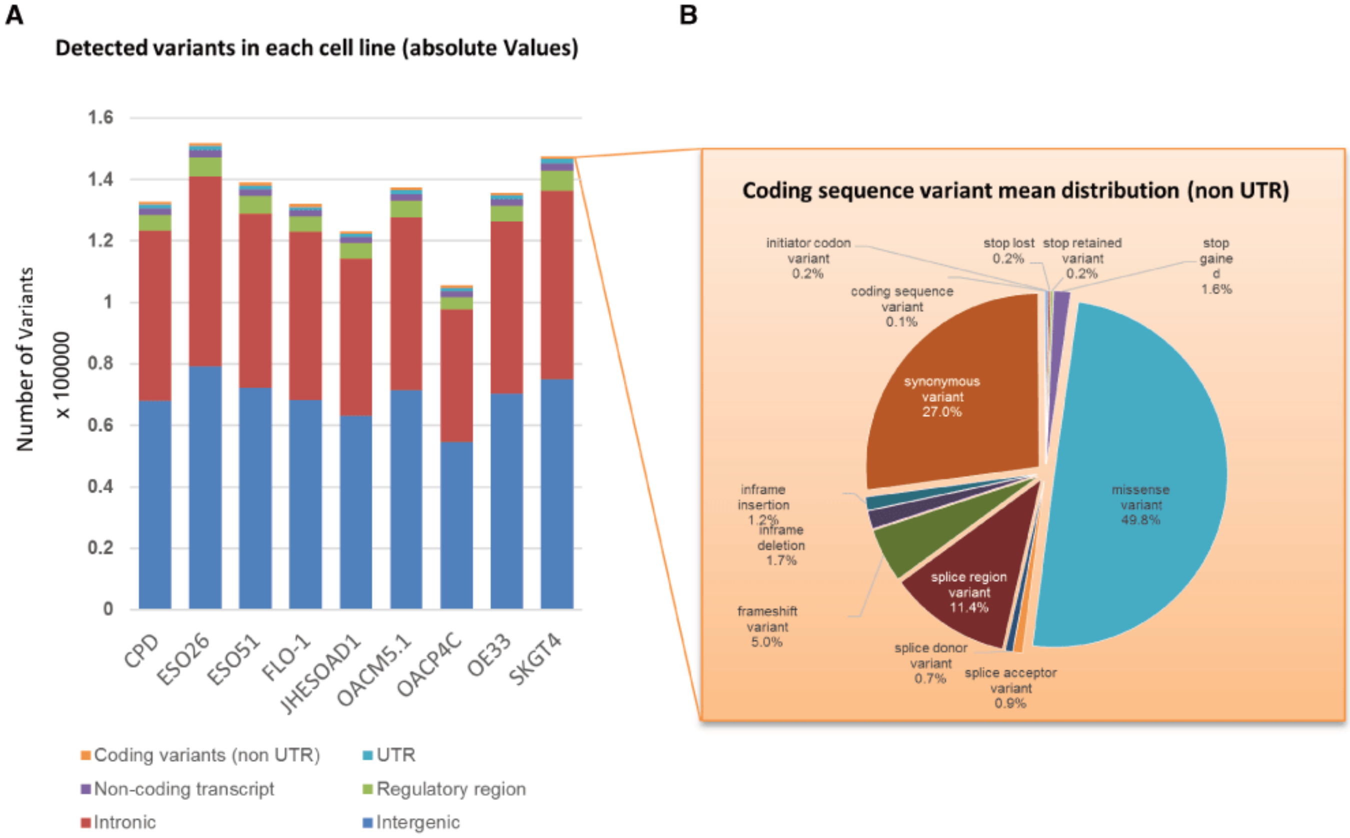
- **Cellularity-driven distinctions in allele-fraction may help**

These should reduce the numbers substantially, but there will still be an excess

What if you have cell-lines

Generally won't have matched normal or cellularity

- **Recently in this situation with OAC cell lines**
- **Clearly too many variants being called**
- **Exonic regions give a feel for the overall performance**



Contino et al. 2016

References

- Alioto T. S. et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nature Communications, 6, 10001.
- Kim D. et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14(4), R36.
- Raczy C. et al. (2013). Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms. Bioinformatics, btt314.
- Kalari K. R. et al. (2014). MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. BMC Bioinformatics, 15(1), 224.
- Chen L. Y. et al. (2012). RNASEQR-a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Research, 40(6), 1–12.
- Piskol R. et al. (2013). Reliable identification of genomic variants from RNA-seq data. American Journal of Human Genetics, 93(4), 641–651.
- Sheng Q. et al. (2016). Practicability of detecting somatic point mutation from RNA high throughput sequencing data. Genomics, 107(5), 163–169.
- Contino G. et al. (2016). Whole-genome sequencing of nine esophageal adenocarcinoma cell lines, 1336(633974), 1–8.
- Koboldt D. C. et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, 22(3), 568–576

References

- Saunders C. T. et al. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England), 28(14), 1811–7.
- Moncunill V et al. (2014). Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nature Biotechnology, 32(11), 1106–1112.
- Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012
- Rimmer A. et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), 912–918.
- Josephidou M. et al. (2015). multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. Nucleic Acids Research, 1–9.