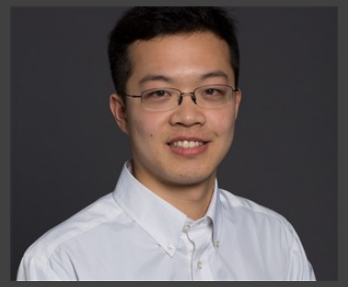
Moss: Enabling high-sensitivity single-nucleotide variant calling from multiple bulk DNA tumor samples

Idoia Ochoa



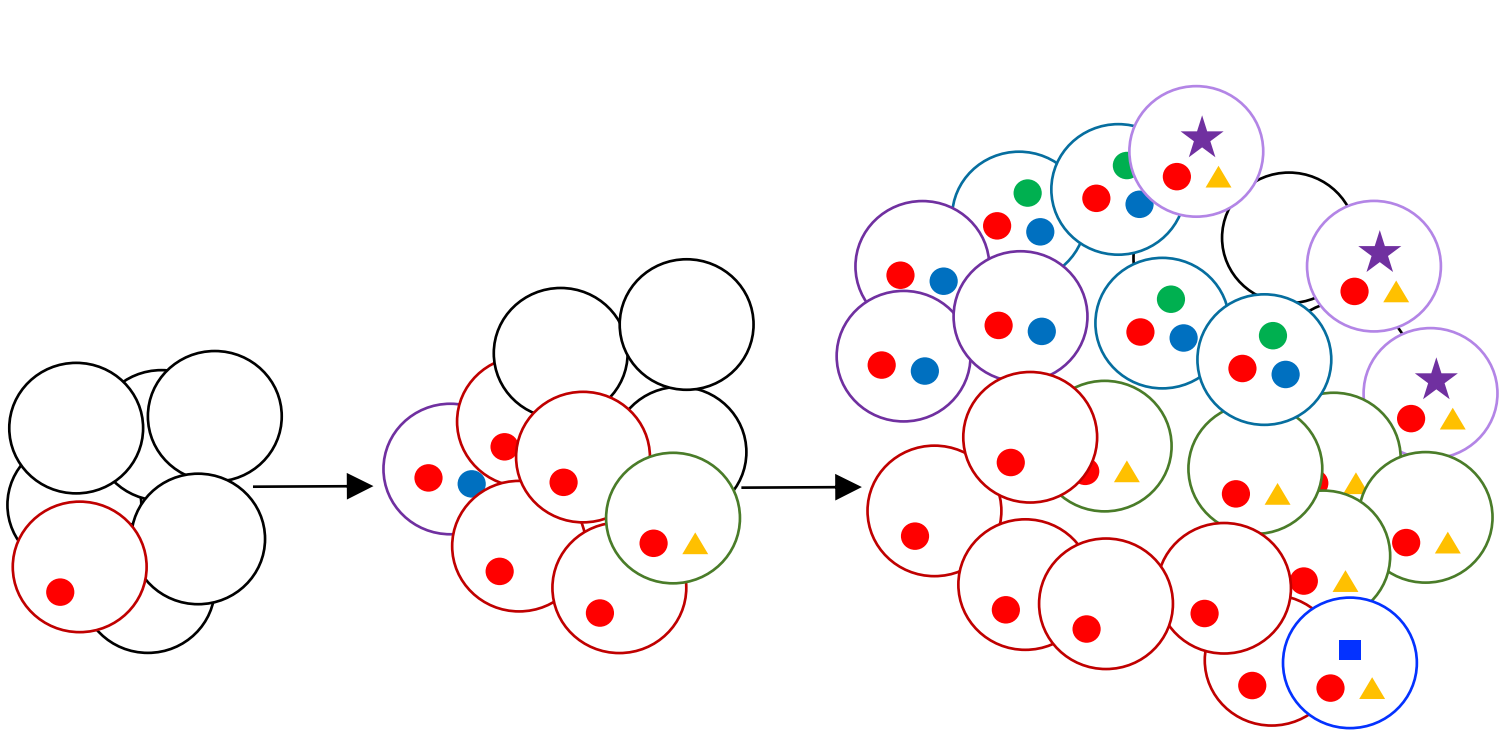
Chuanyi Zhang

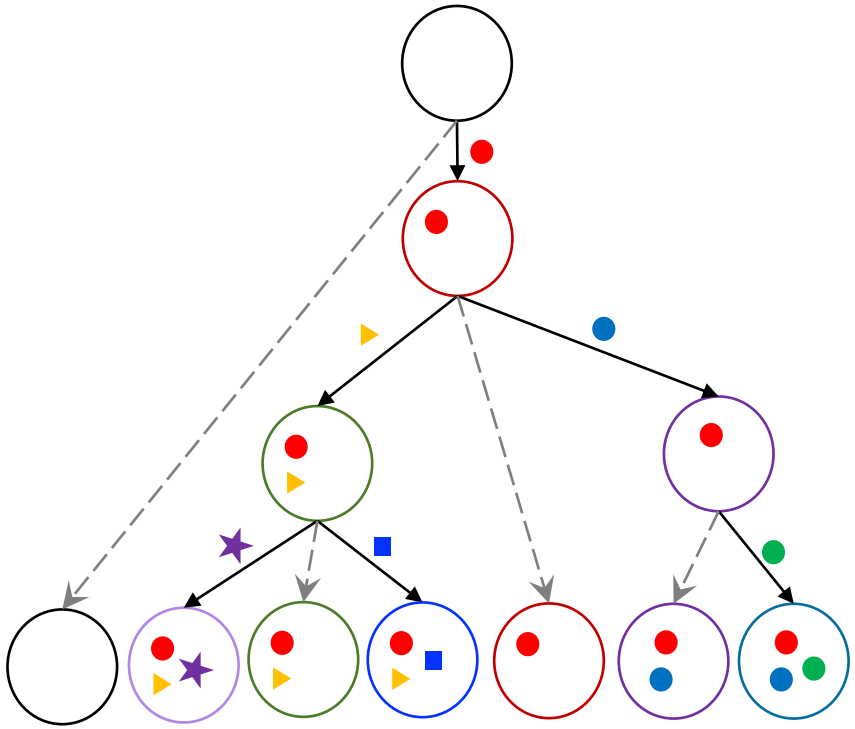
Mohammed El-Kebir

DATAI
February 1st 2023

Background: problem

- Tumor consists of clones with different **somatic variants**
- Important to identify these variants for downstream analysis in cancer genomics





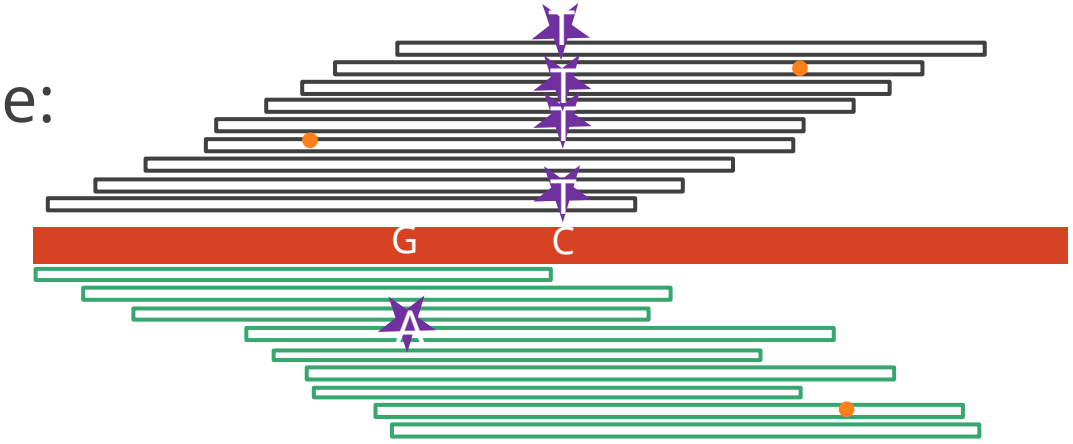
Background: problem

- Somatic vs germline SNVs (Single Nucleotide Variants)

Normal sample:

Reference genome

Tumor Sample

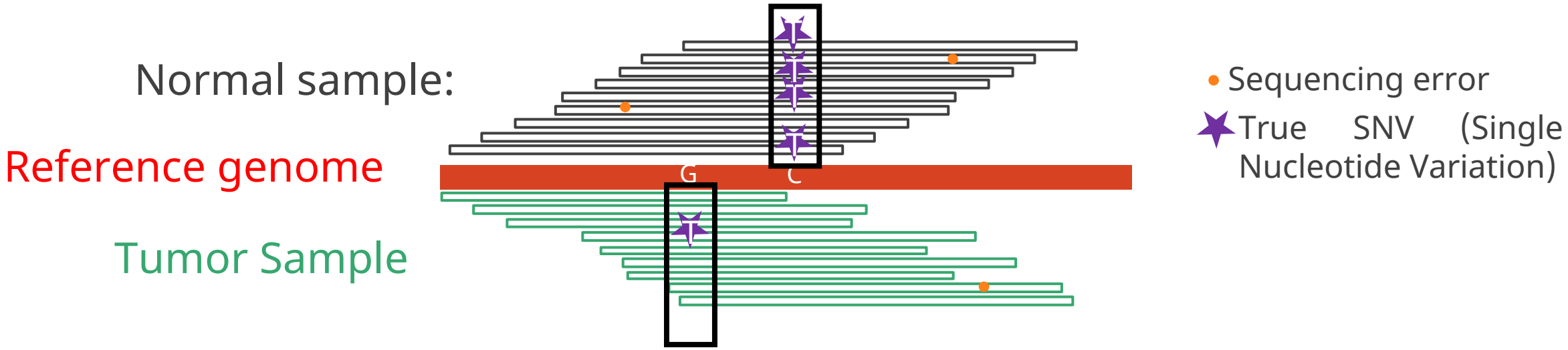


Sequencing error

True SNV (Single Nucleotide Variation)

Background: problem

- Somatic vs germline SNVs (Single Nucleotide Variants)

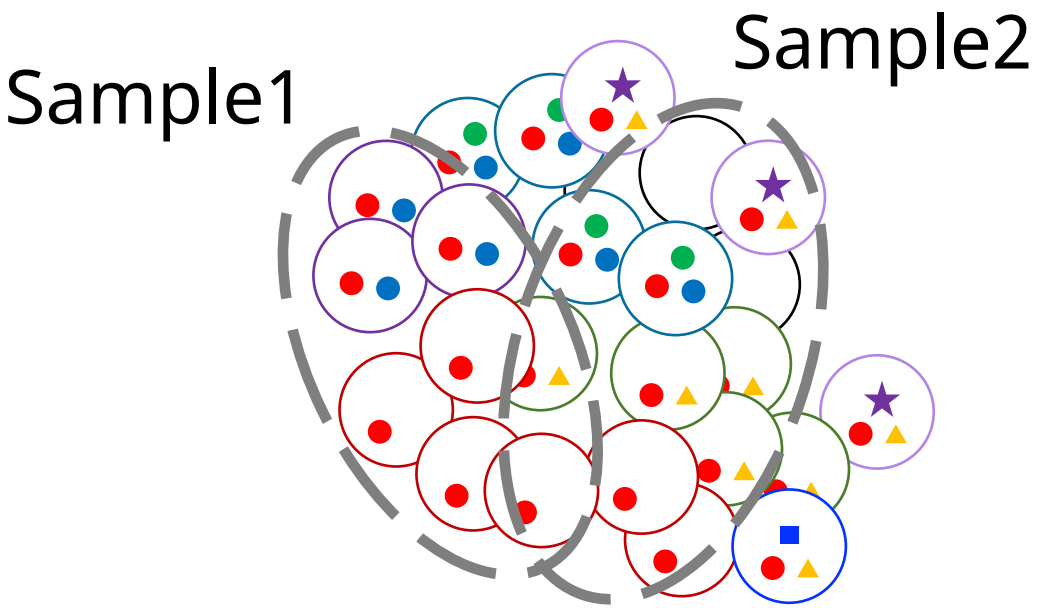


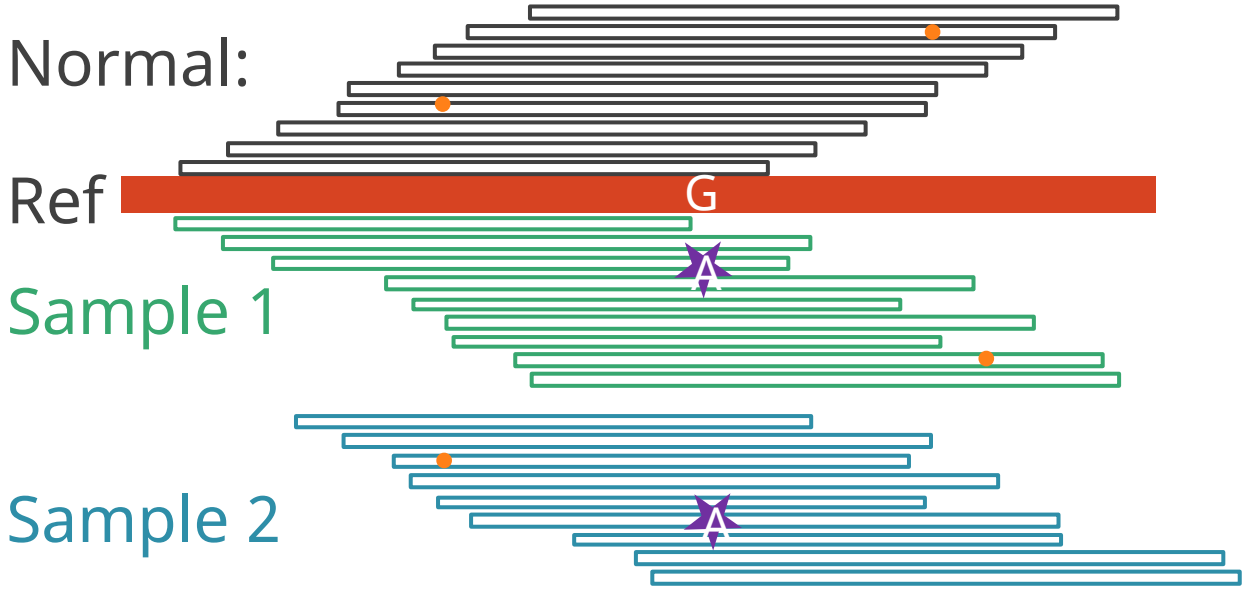
VAF (Variant Allele Frequency):

- Normally 0, 0.5, or 1 for germline SNVs (diploid organisms)
- Any value between 0 and 1 for somatic SNVs -> difficult to distinguish low VAF variants from sequencing errors!

Background: problem

- **Motivation**: Multiple samples from the same patient can help discover low-VAF SNVs





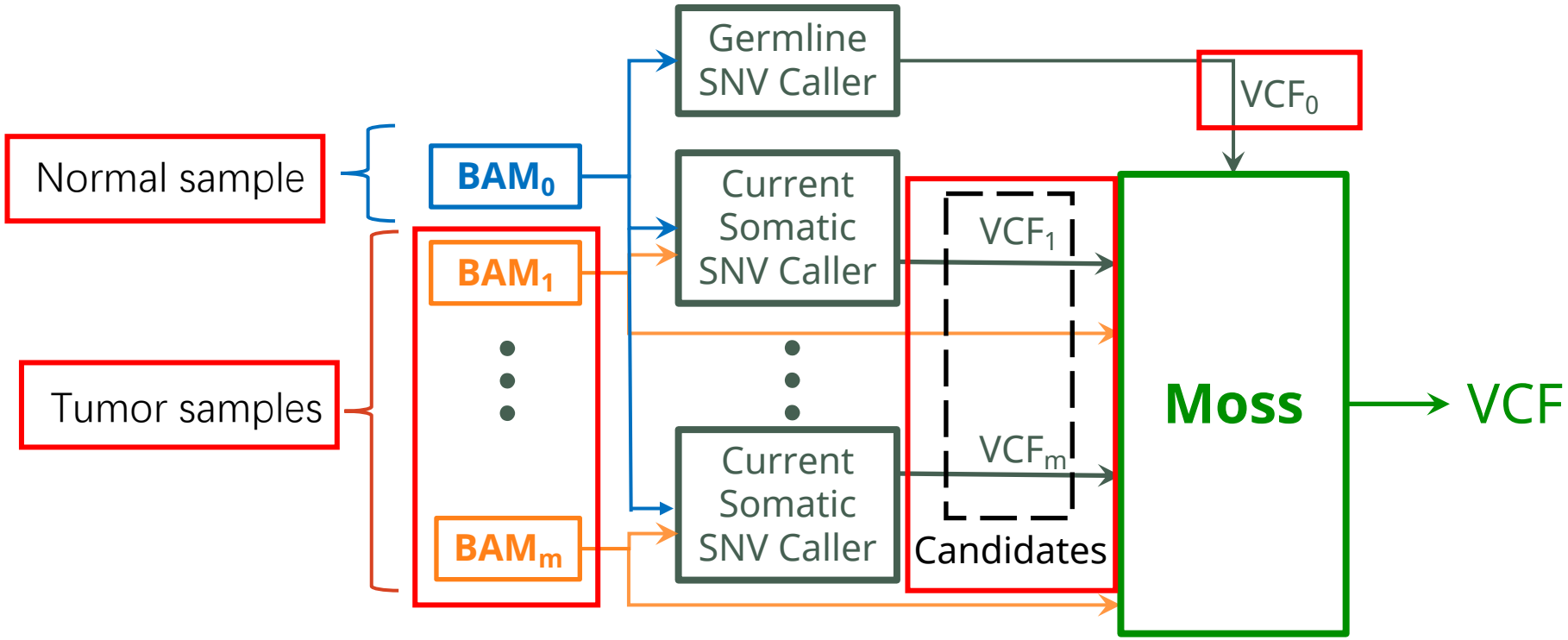
Current somatic SNV callers

Somatic Variant Caller	Support of Multi-Sample?	Authors
Mutect2 (GATK v4.0) [1]	✗	Broad Institute
Strelka2 [2]	✗	Illumina
Mutect2 (GATK v4.1)	✓	Broad Institute
multiSNV [3]	✓	Cambridge
(Octopus)[4]	✓	Univ. of Oxford

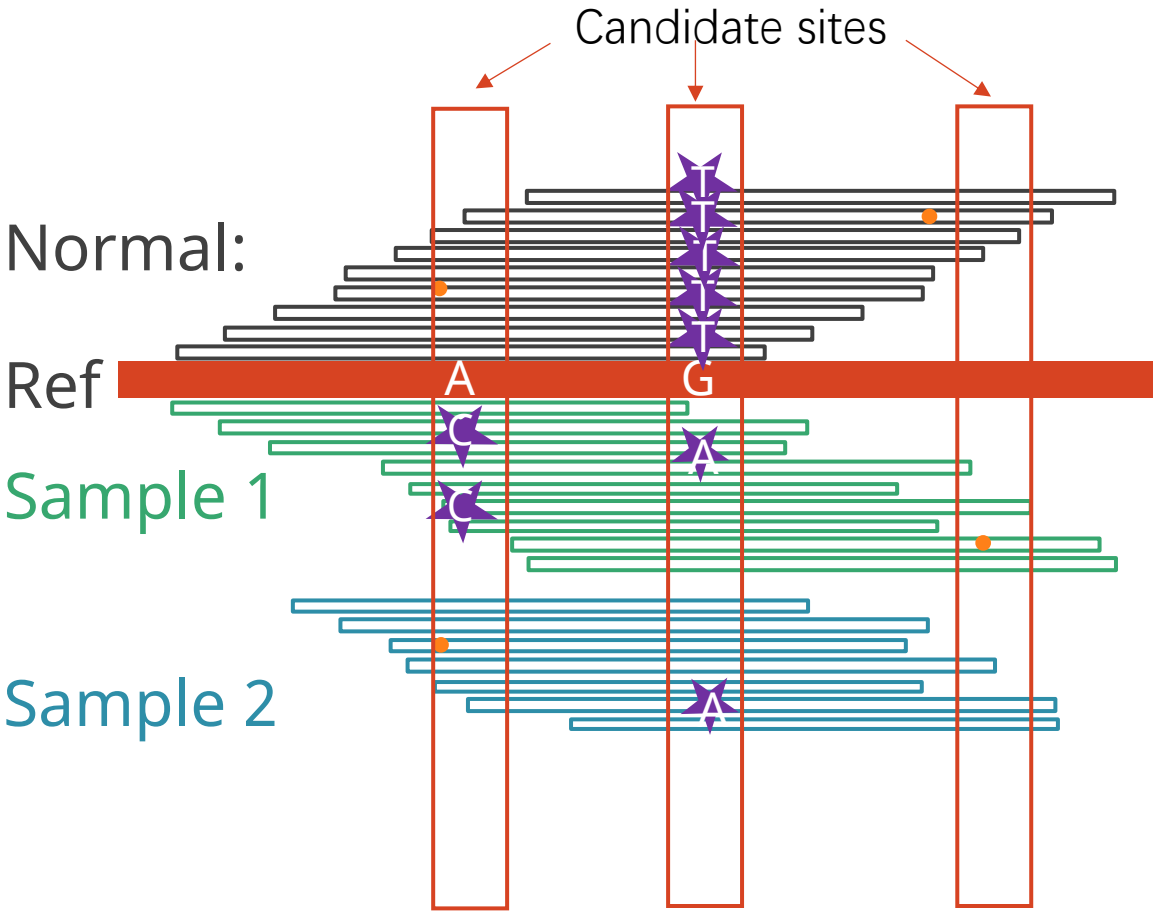
[1]https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_hellbender_tools_walkers_mutect_Mutect2.php
[2] Kim, S., et al. "Strelka2: fast and accurate calling of germline and somatic variants." *Nature methods* 15.8 (2018): 591.
[3] Malvina, J., et al. “multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples”, Nucleic Acids Research (2015)
[4] Cooke, D., et al. "A unified haplotype-based method for accurate and comprehensive variant calling." *Nature biotechnology* 39.7 (2021): 885-892.

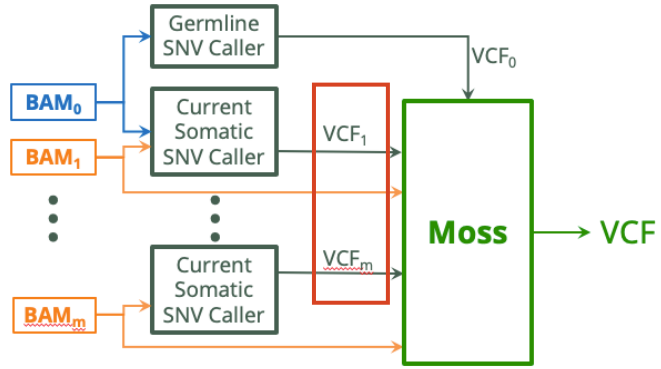
Moss

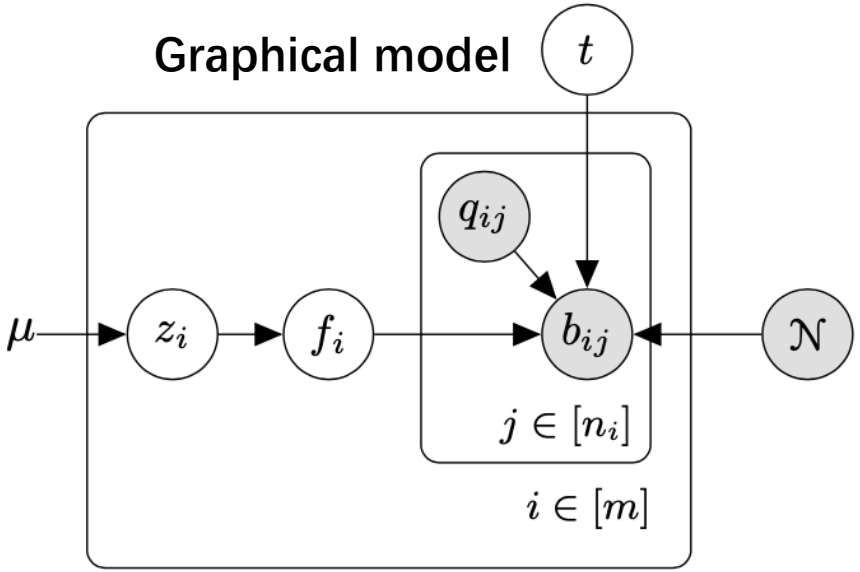
- Extension to existing single-sample variant callers:
  - No need to care about INDELs, structural variants, etc.
- Improves sensitivity while keeping high precision



Key ideas of Moss (Bayesian model)

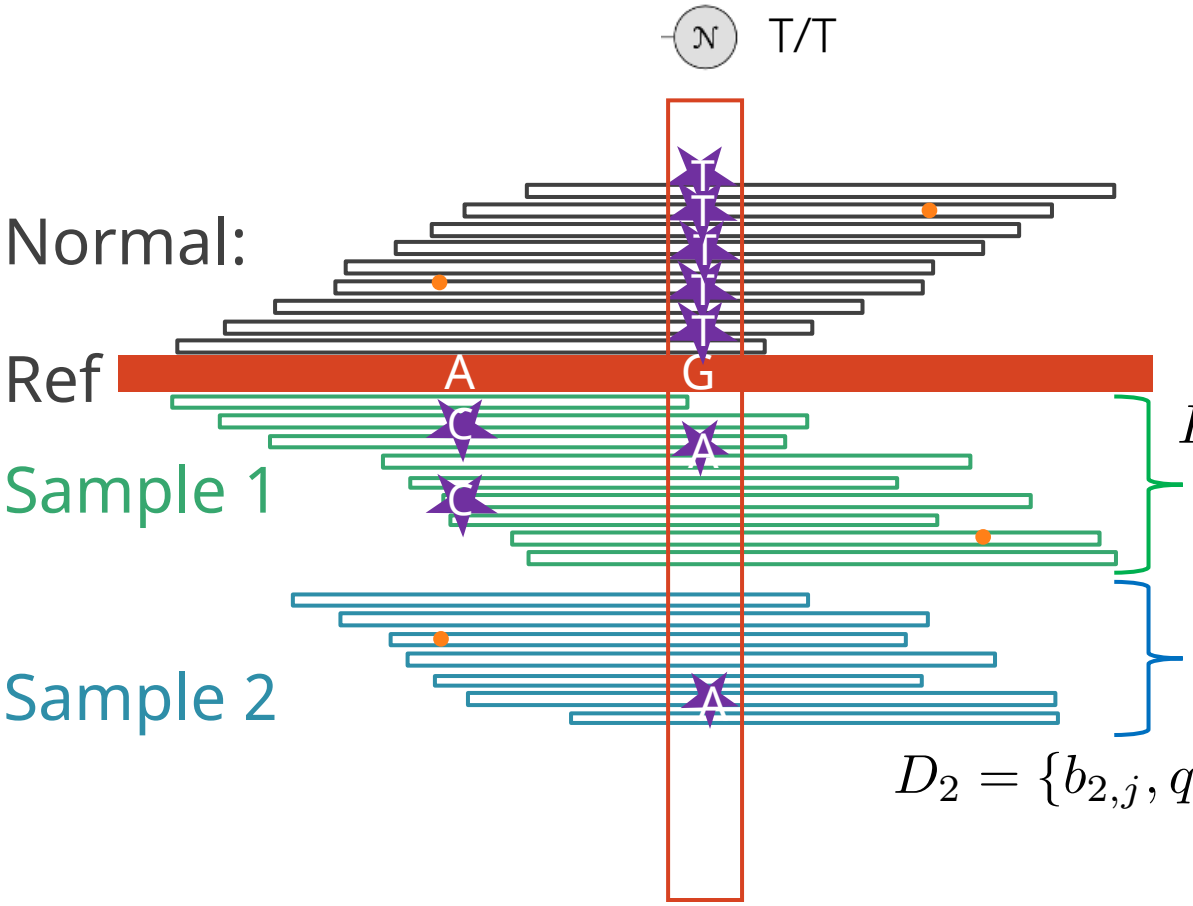






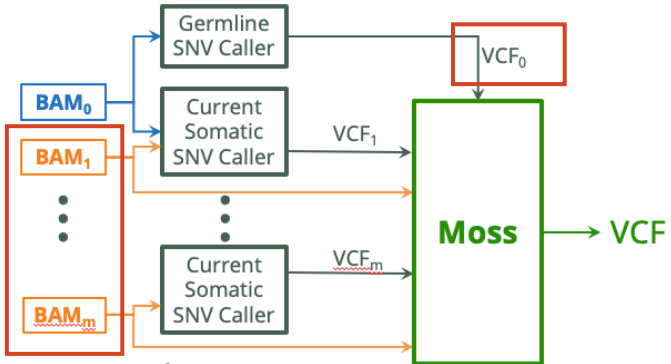
Moss analyzes each candidate site in isolation

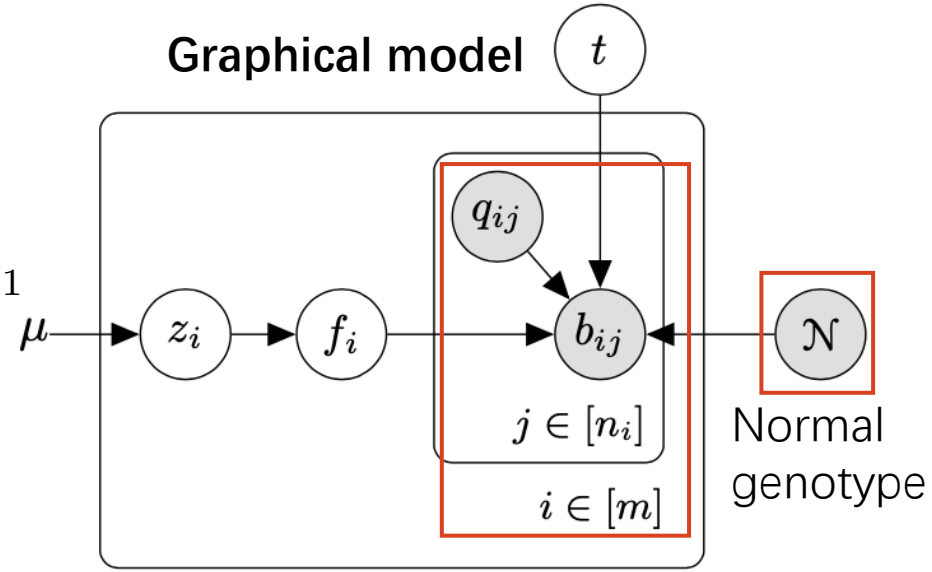
Key ideas of Moss (Bayesian model)



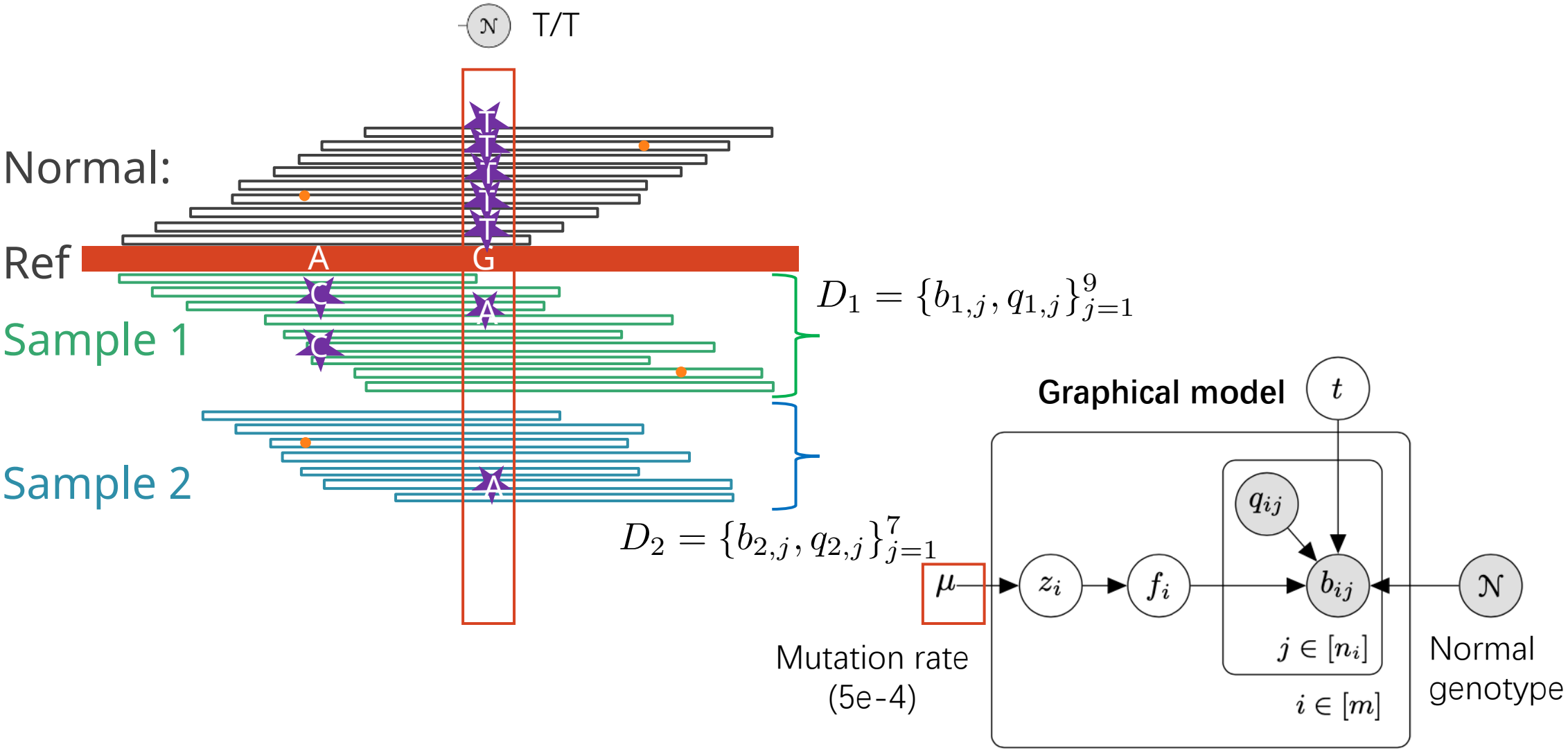
$D_1 = \{b_{1,j}, q_{1,j}\}_{j=1}^{9}$

$D_2 = \{b_{2,j}, q_{2,j}\}_{j=1}^{7}$

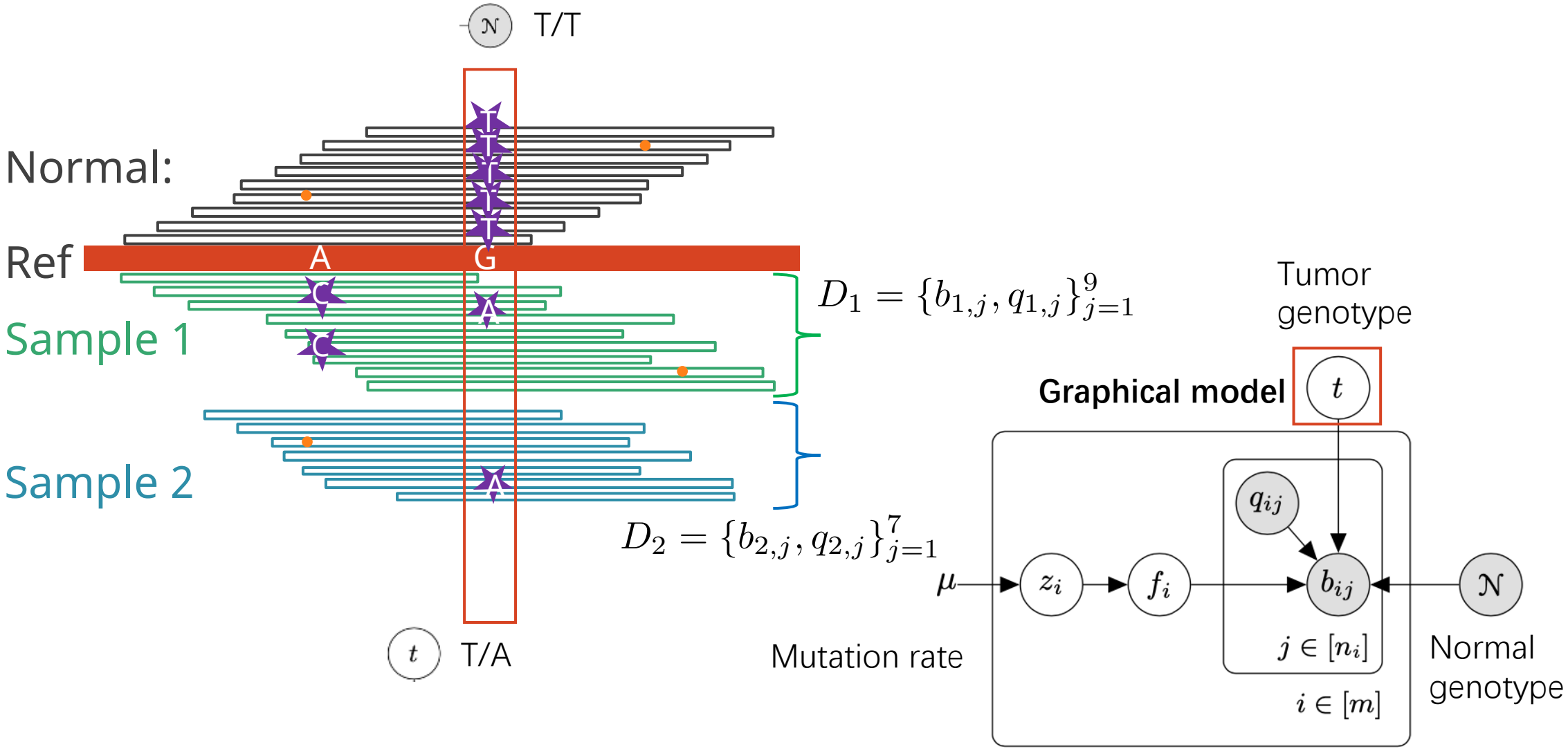




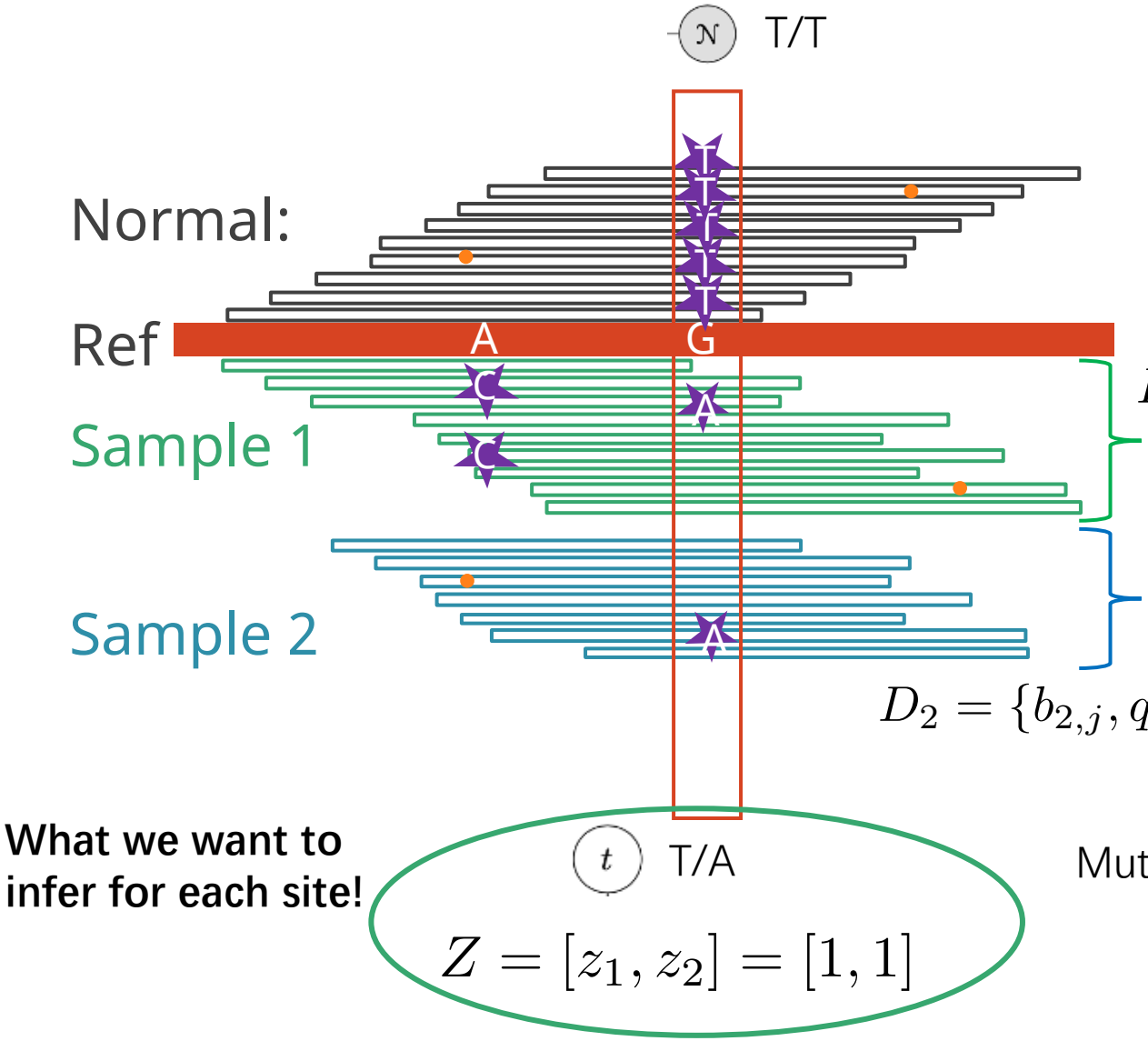
Key ideas of Moss (Bayesian model)

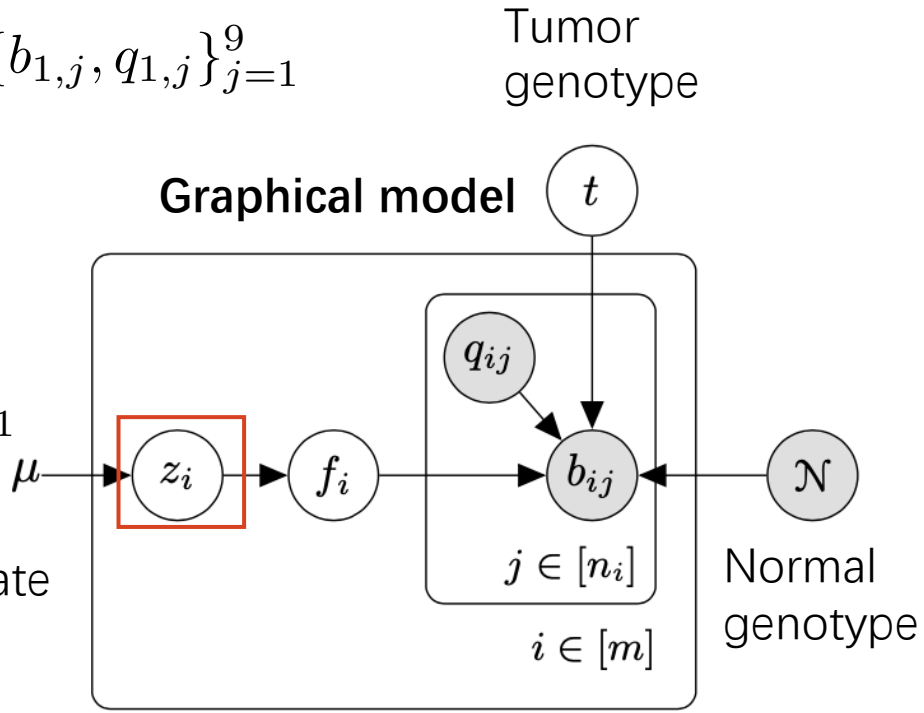


Key ideas of Moss (Bayesian model)

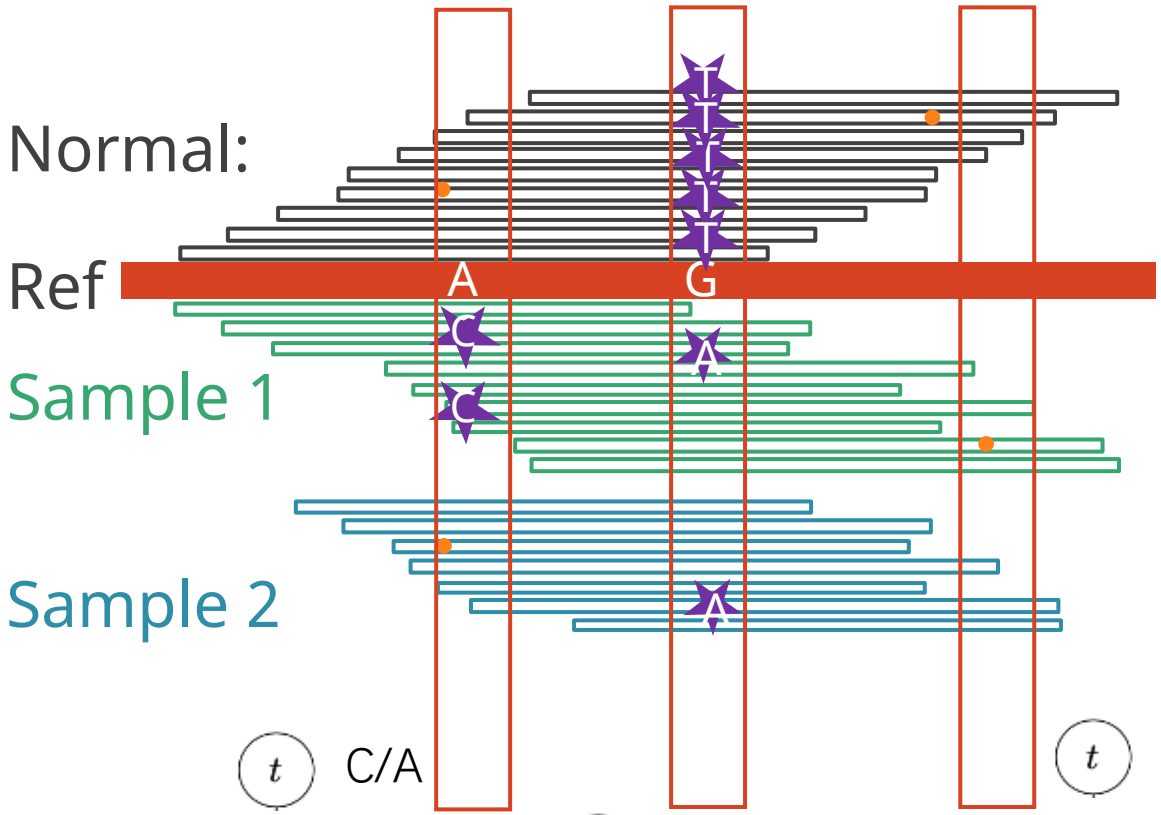


Key ideas of Moss (Bayesian model)





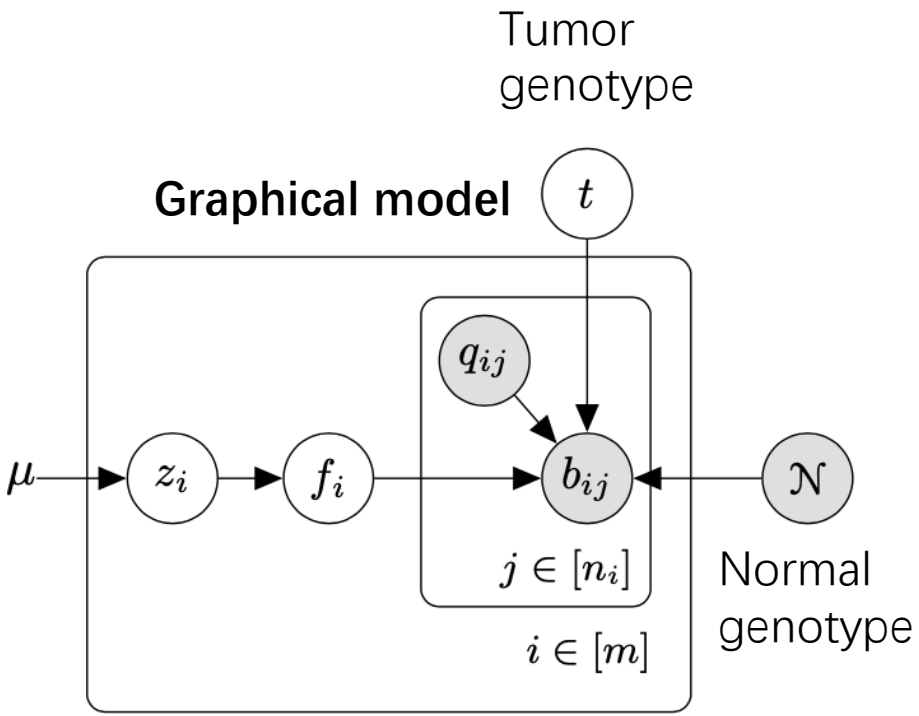
Key ideas of Moss (Bayesian model)



$Z = [z_1, z_2] = [1, 0]$

$Z = [z_1, z_2] = [0, 0]$

$Z = [z_1, z_2] = [1, 1]$

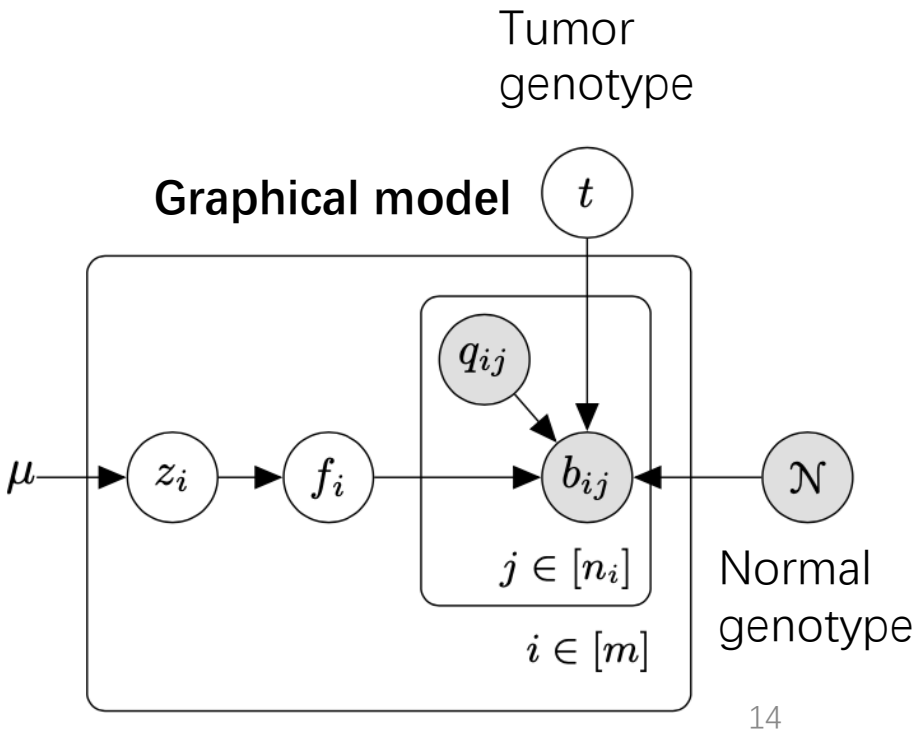


Computation

- Somatic probability

  $P(\mathbf{z} != \mathbf{0}|N, \mathbf{b}, \mathbf{q})$ = 1 - $P(\mathbf{z} = \mathbf{0}|N, \mathbf{b}, \mathbf{q})$

- Tumor genotype t
  - Find t that minimizes $P(\mathbf{z} = \mathbf{0}|N, \mathbf{b}, \mathbf{q}, t)$
- $\mathbf{z} = [z_1, z_2, \ldots, z_m]$
- **Filtering** to remove false positives

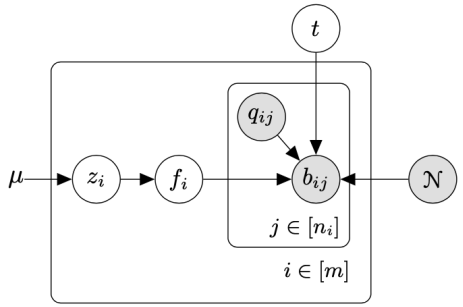


Computations

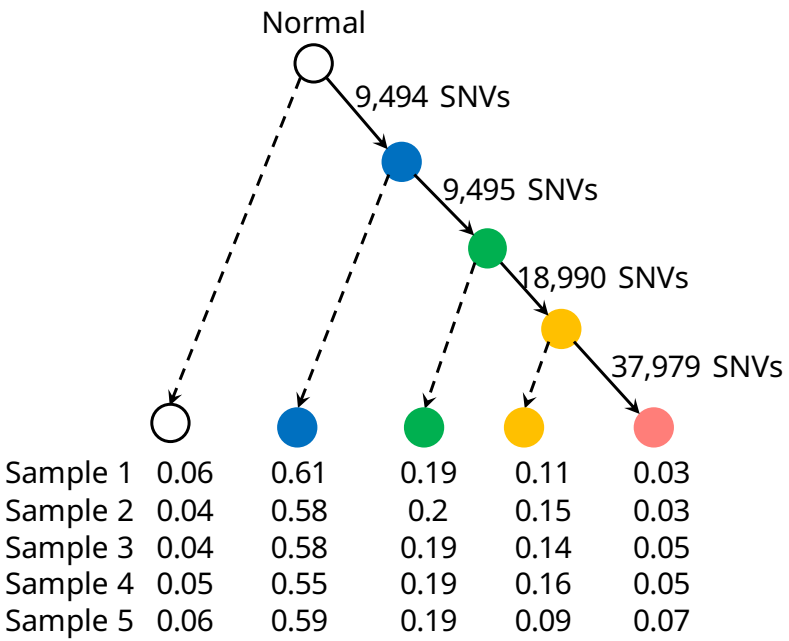
$$\mathbf{P}(\mathbf{Z}=\mathbf{0}\,|\,\mathcal{N},\mathbf{b},\mathbf{q}) = \frac{\mathbf{P}(\mathbf{b}\,|\,\mathbf{Z}=\mathbf{0},\mathcal{N},\mathbf{q})\mathbf{P}(\mathbf{Z}=\mathbf{0})}{\sum_{t\notin\mathcal{N}}\sum_{\mathbf{z}\in\{0,1\}^m}\mathbf{P}(\mathbf{b},\mathbf{Z}=\mathbf{z},t\,|\,\mathcal{N},\mathbf{q})}$$

$$= \frac{\sum_{t\notin\mathcal{N}}\mathbf{P}(\mathbf{b}\,|\,\mathbf{Z}=\mathbf{0},t,\mathcal{N},\mathbf{q})\mathbf{P}(\mathbf{Z}=\mathbf{0})}{\sum_{t\notin\mathcal{N}}\sum_{\mathbf{z}\in\{0,1\}^m}\mathbf{P}(\mathbf{b}\,|\,\mathbf{Z}=\mathbf{z},t,\mathcal{N},\mathbf{q})\mathbf{P}(\mathbf{Z}=\mathbf{z})}$$

$$t^* = \operatorname*{argmax}_t \sum_{\mathbf{z}\in\{0,1\}^m} \mathbf{P}(\mathbf{b}\,|\,\mathbf{z},t,\mathcal{N},\mathbf{q})\mathbf{P}(\mathbf{z}),$$



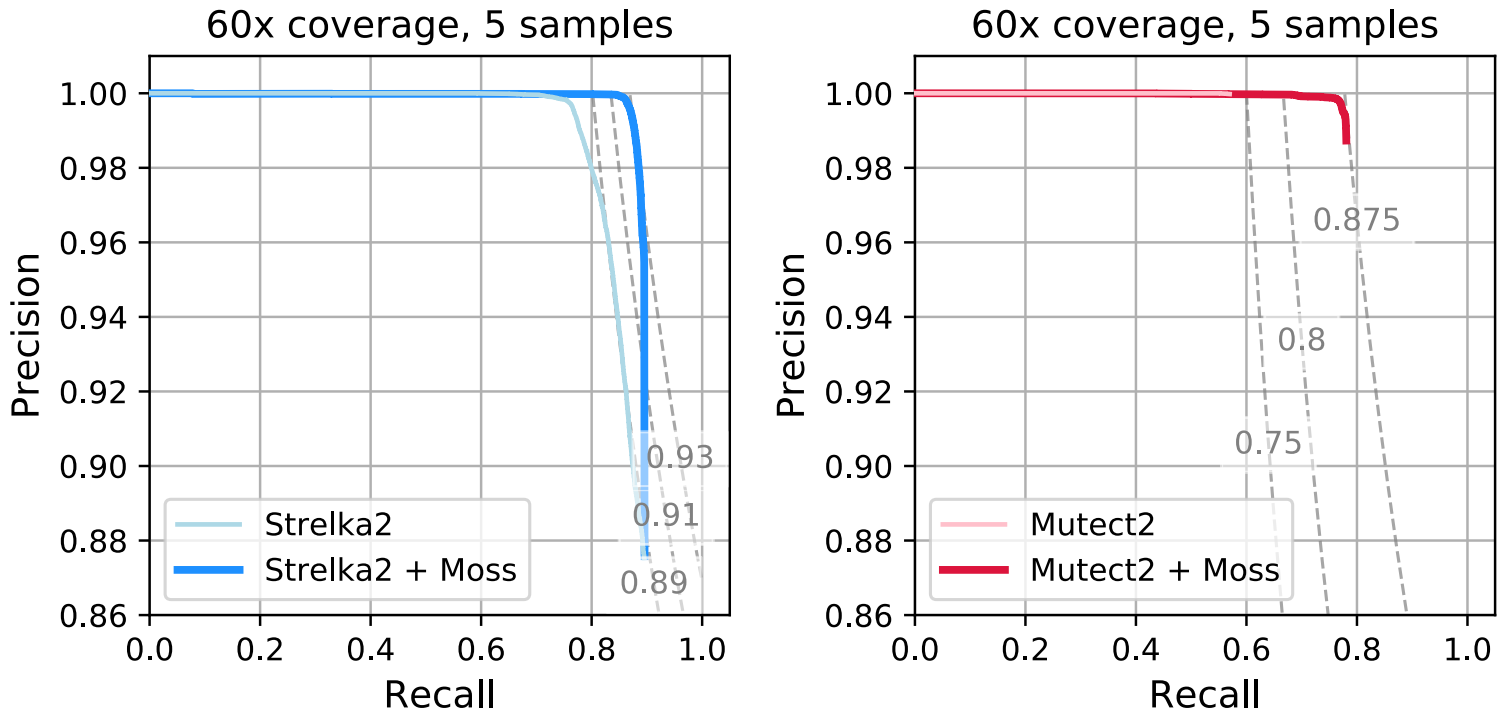
Moss improves accuracy in simulated data



Synthetic data generation:

- Illumina sequencing of 5 samples of chr. 20
- Matched normal sample: chr. 20 with germline SNPs from dbSNP
- Add ~75K somatic mutations following a simple linear phylogeny tree with 4 clones
- Generate 5 samples with different mixture rations and either 30x or 60x average coverage

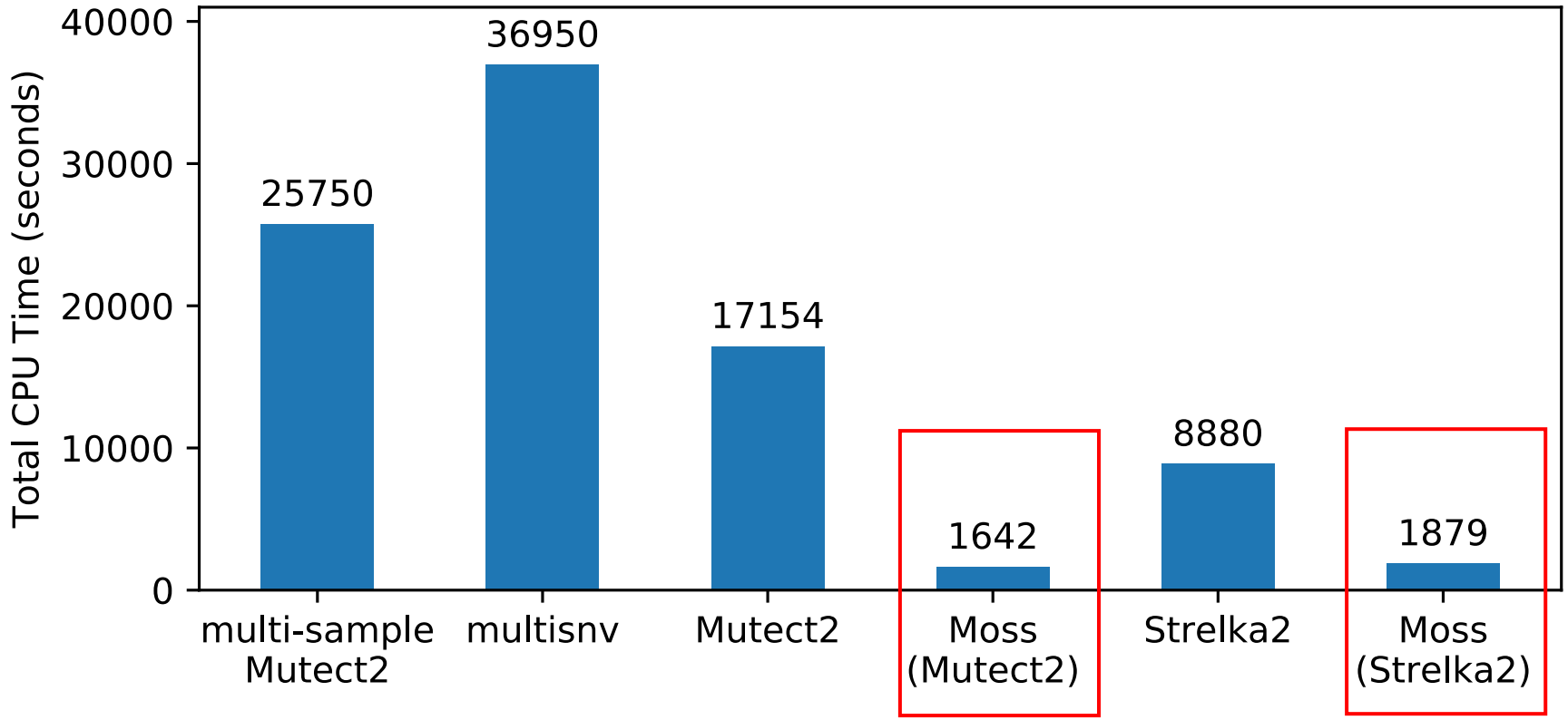
Moss improves accuracy in simulated data



Precision-Recall curves:

- Moss improves recall without loss of precision in both cases
- Similar results for m=2,3,4 and 30x coverage
- Multi-sample Mutect2: similar performance to Mutect2 + Moss but higher running time
- multiSNV: worst recall and very high running time

Moss incurs little overhead in running time

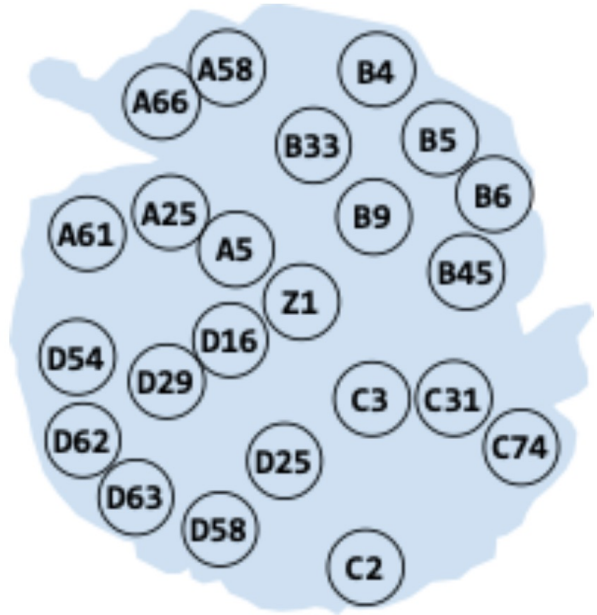


Evaluating Moss on real data

- Hepatocellular carcinoma (HCC) dataset
- Colorectal Cancer (CRC) Dataset
- Acute Myeloid Leukemia (AML) dataset

Hepatocellular carcinoma (HCC) dataset

- Published by BGI*
- 268 samples taken from a slice of HCC tissue
- WES conduced on 23 samples
- Average coverage of 74.4x
- Includes WES of matched normal sample



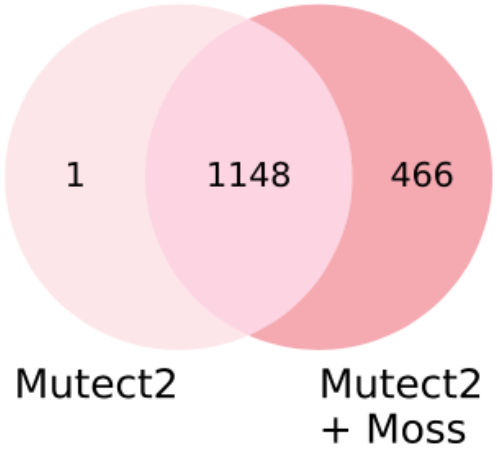
* Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proceedings of the National Academy of Sciences* 112, E6496–E6505 (2015).

Results

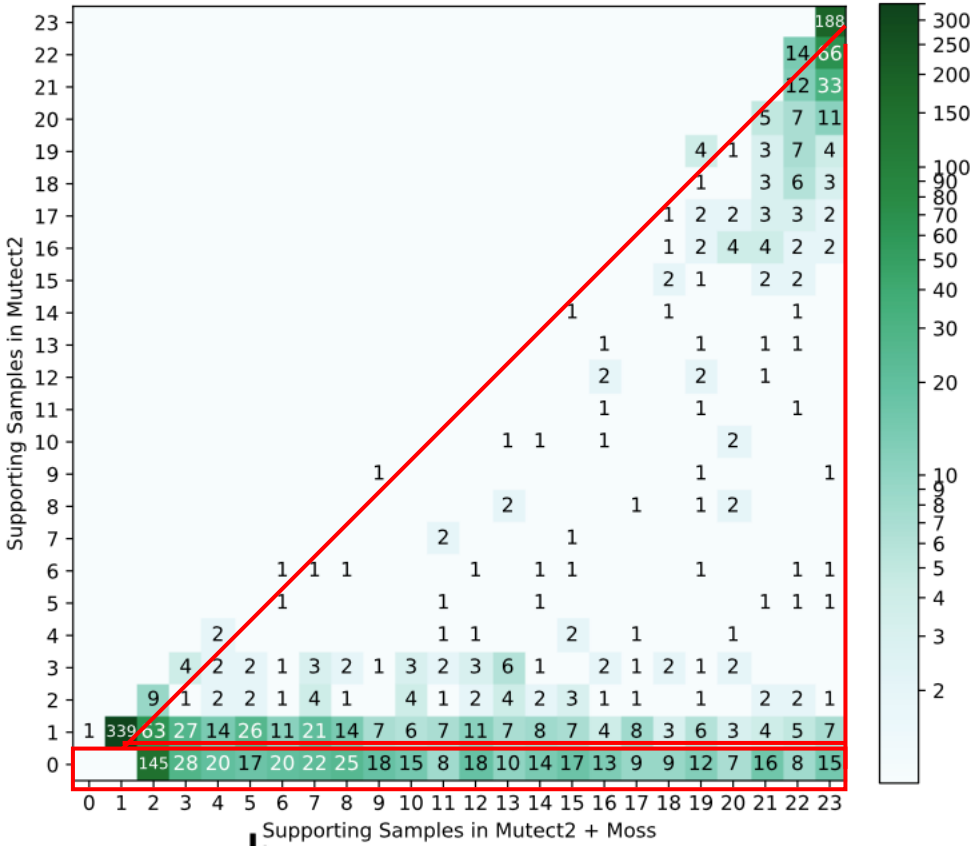
- Multi-sample callers multiSNV and Mutect2 unable to run
- Mutect2 vs Mutect2 + Moss (similar results for Strelka2)

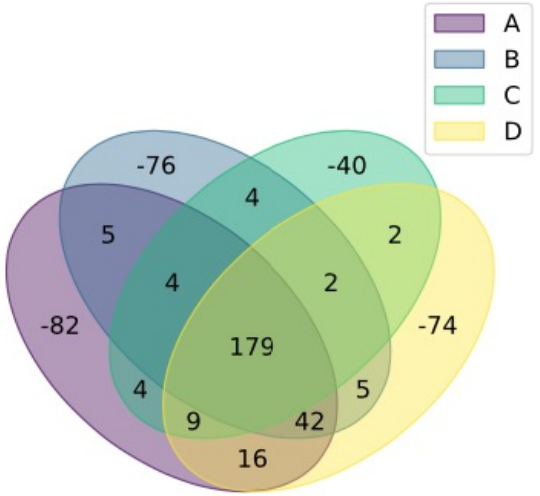
Results

- Multi-sample callers multiSNV and Mutect2 unable to run
- Mutect2 vs Mutect2 + Moss (similar results for Strelka2)



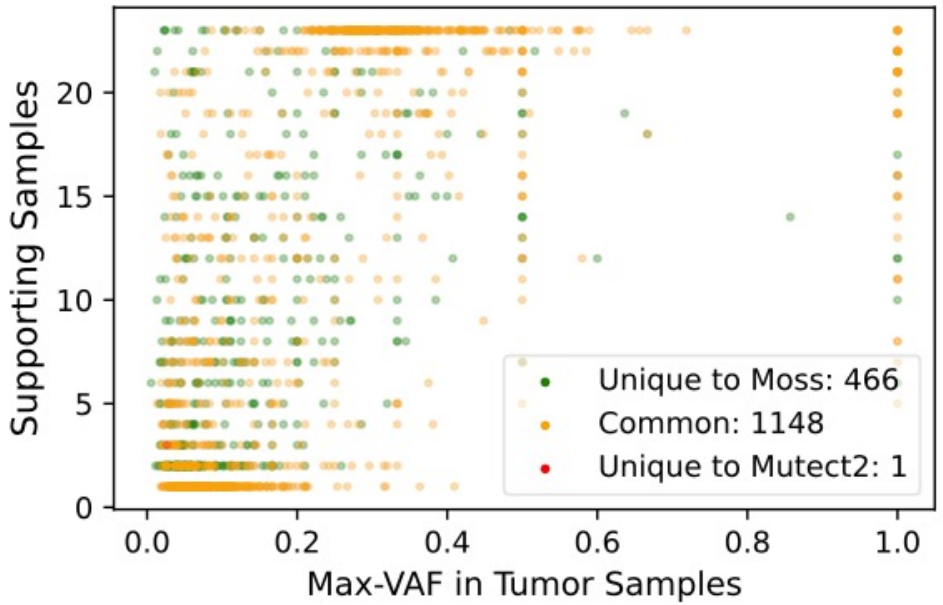
Most new variants:
- VAF < 0.3 on tumor samples
- VAF < 0.06 on normal sample





For 36% of common variants, we find more supporting samples. We reduce the number of samples specific of one spatial location.

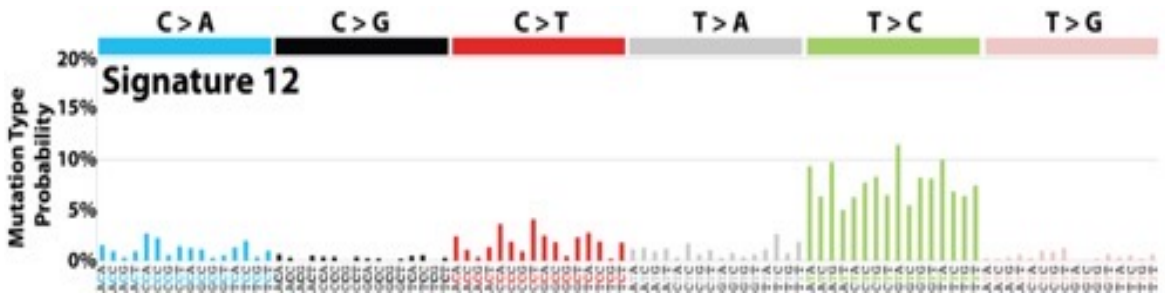
Validation

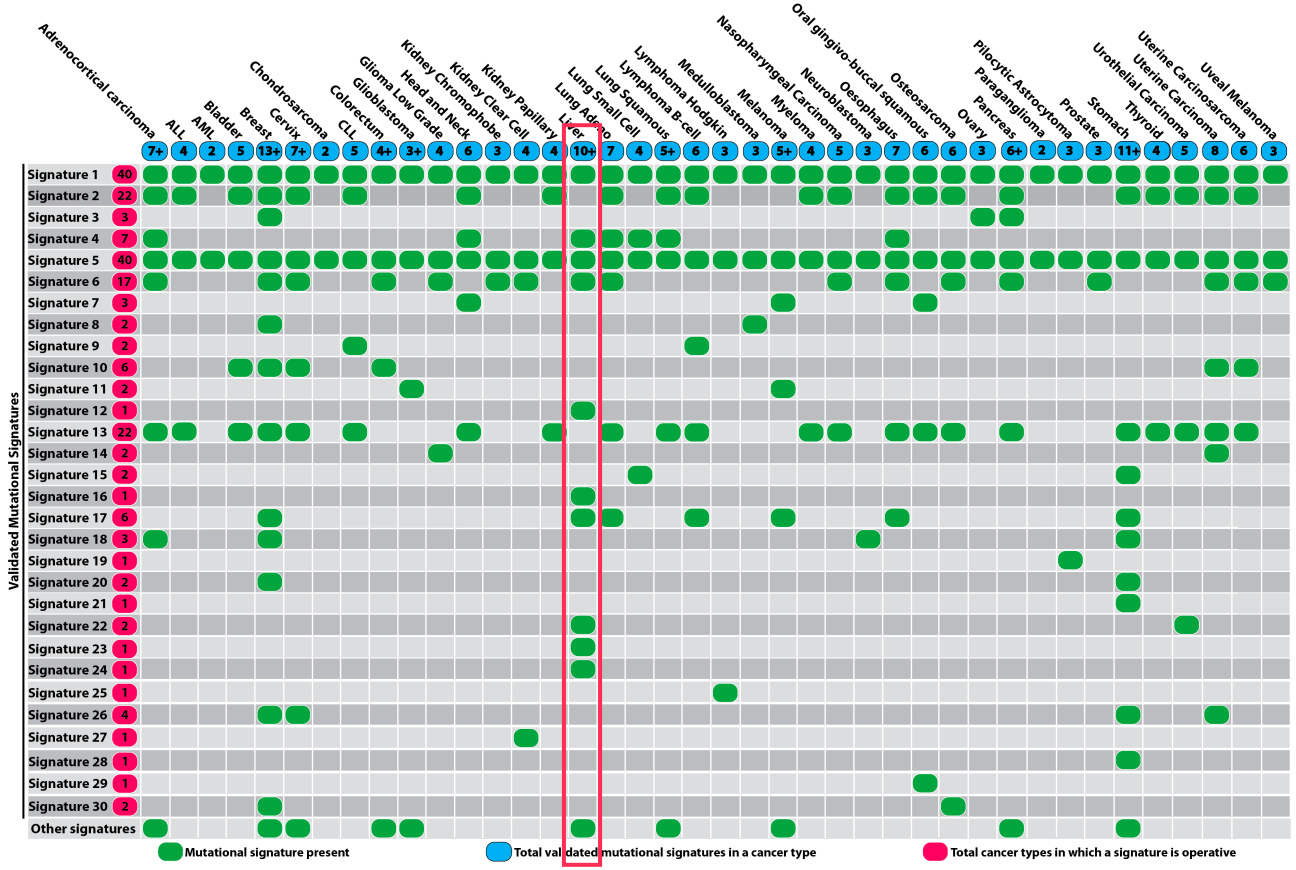


Most unique to Moss variants have VAF < 0.3, and normal VAF < 0.06.

Mutational pattern*

- 96 classes of SNVs
  - 6 types of substitutions
  - 4 types of 3’ bases
  - 4 types of 5’ bases

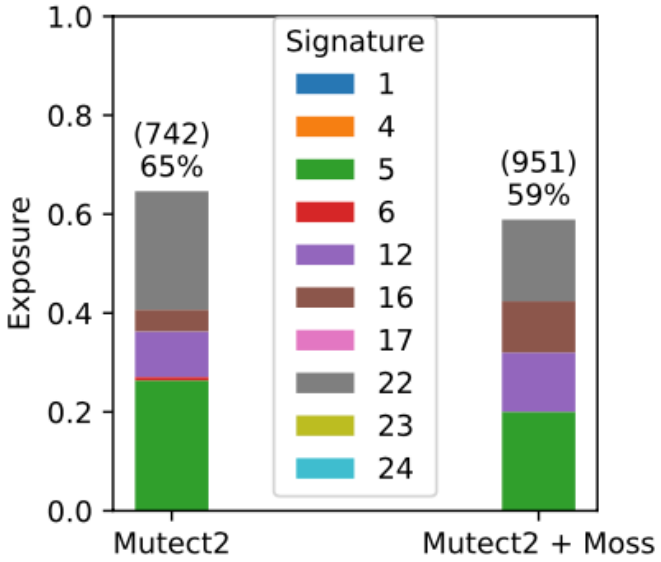




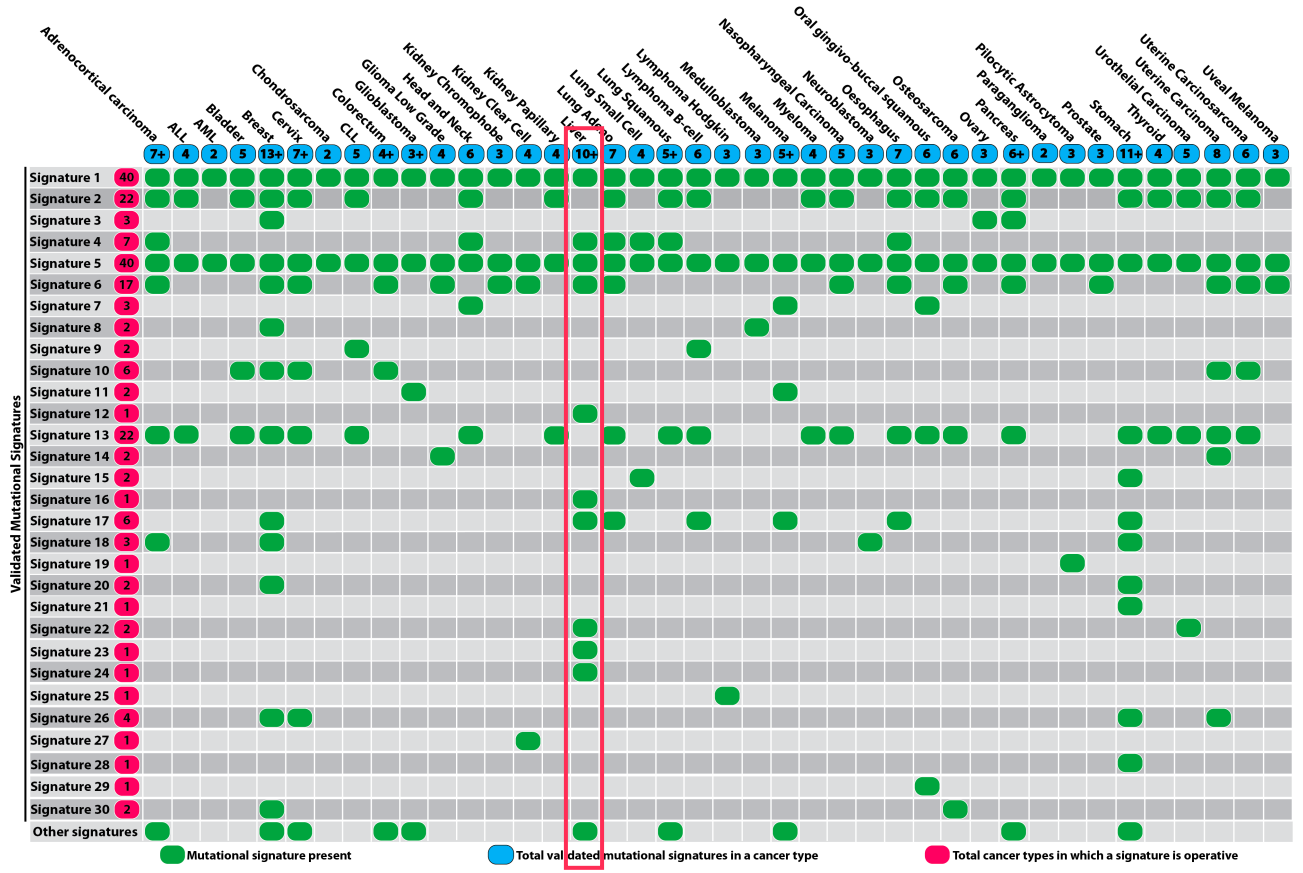
*Alexandrov, Ludmil B., et al. "Clock-like mutational processes in human somatic cells." *Nature genetics* 47.12 (2015): 1402.

Mutational pattern

Exposure percentages of liver cancer signatures



Moss: lower total exposure but more variants explained



Alexandrov, Ludmil B., et al. "Clock-like mutational processes in human somatic cells." *Nature genetics* 47.12 (2015): 1402.

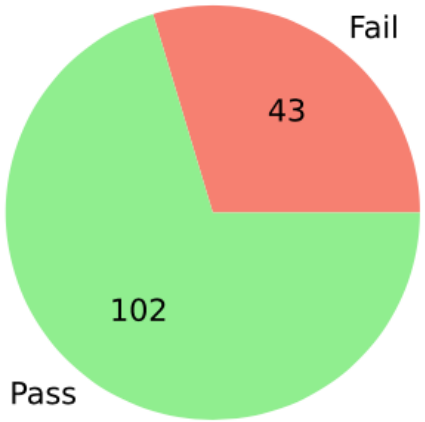
Manual review

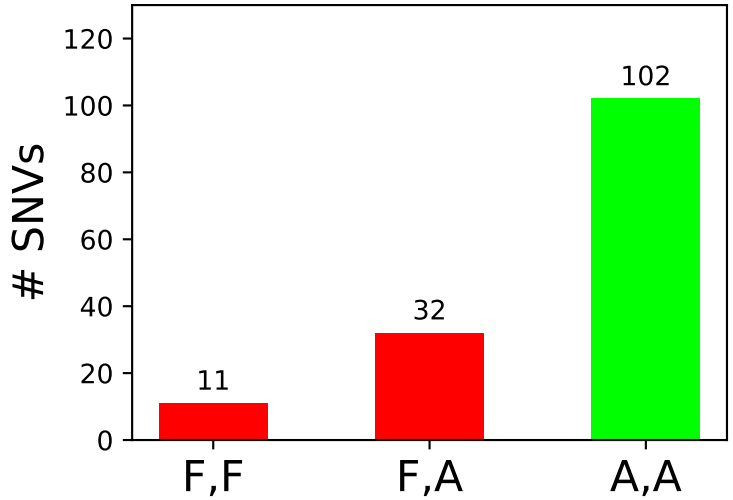
- We follow the procedure of Barnell et. al. *
- Variants unique to Moss and called in exactly 2 samples

*Barnell, E. K. et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genet. Med.* **21**, 972–981 (2019).

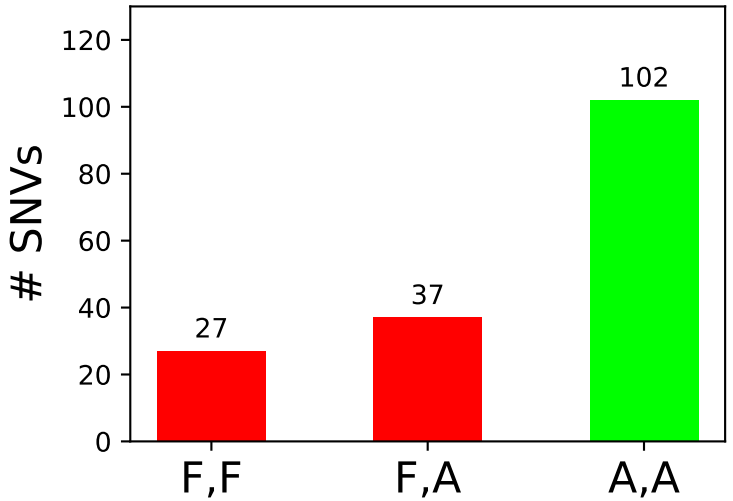
Manual review

- We follow the procedure of Barnell et. al. *
- Variants unique to Moss and called in exactly 2 samples



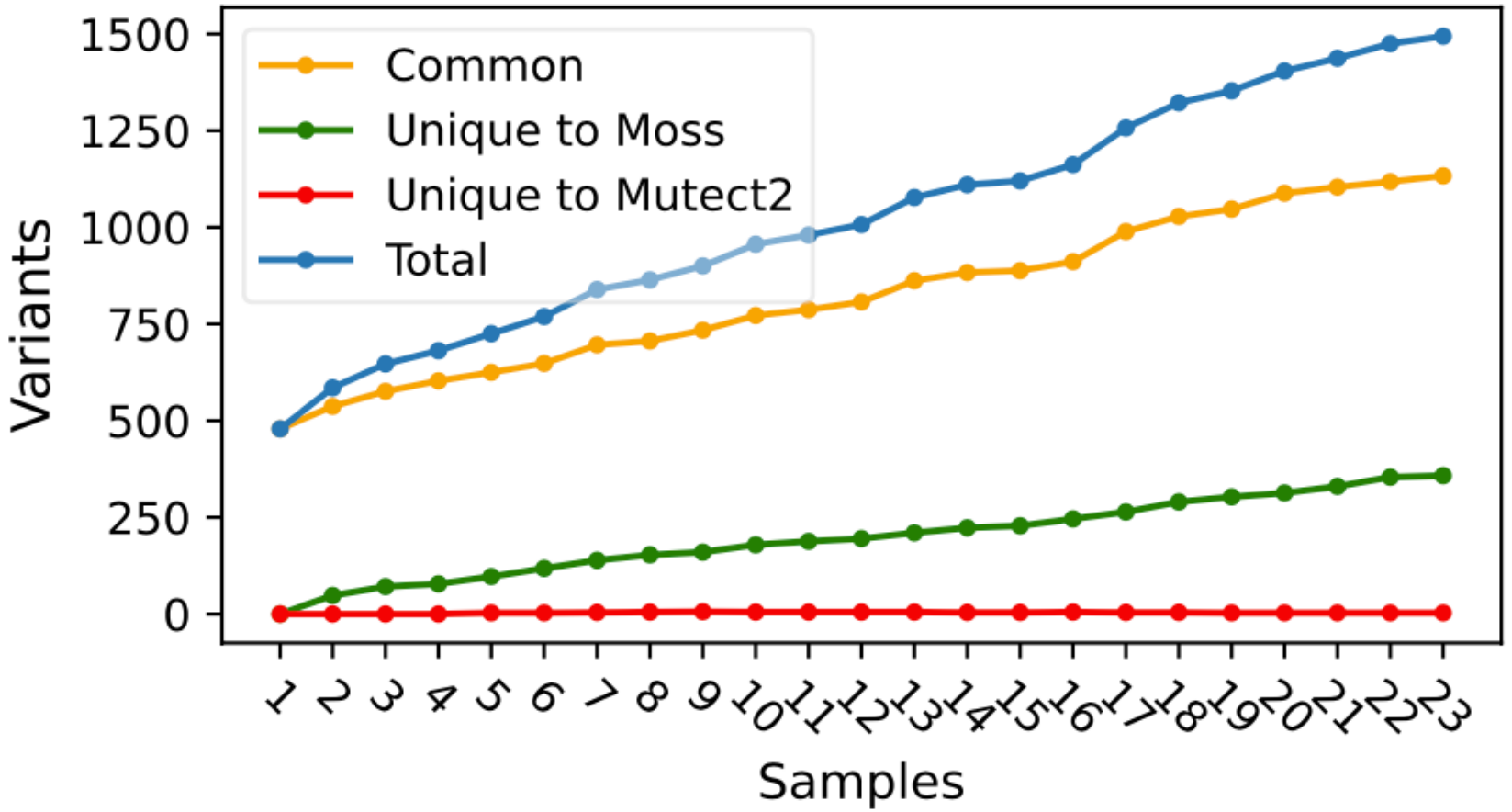


Before Moss filters



*Barnell, E. K. et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genet. Med.* **21**, 972–981 (2019).

Benefit of Moss even with fewer samples



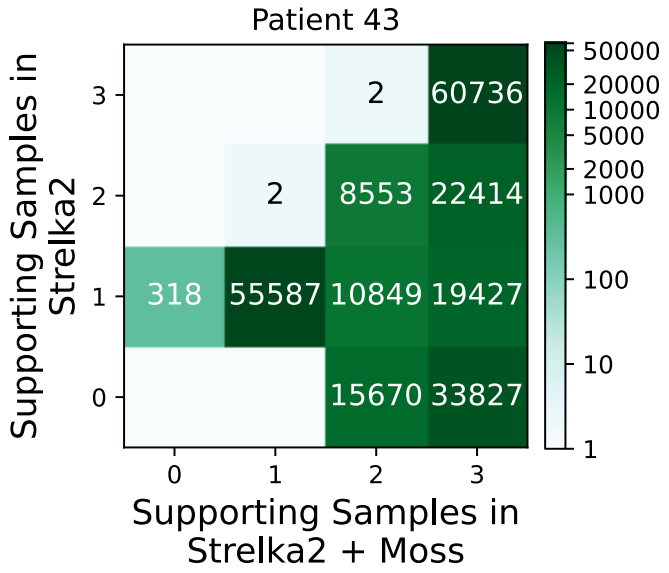
Same trend with data downsampled to 30x, 20x and 10x (from original~75x)

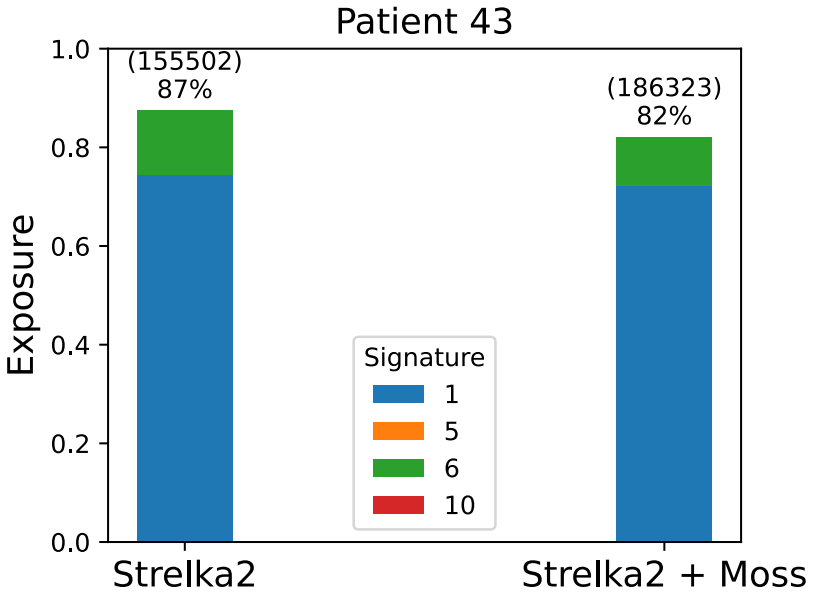
Colorectal Cancer (CRC) Dataset

- 2 patients (denoted as **43** and 45)
- For each patient:
  - 3 tumor samples
  - 1 matched normal (from blood)

Colorectal Cancer (CRC) Dataset

- 2 patients (denoted as **43** and 45)
- For each patient:
  - 3 tumor samples
  - 1 matched normal (from blood)





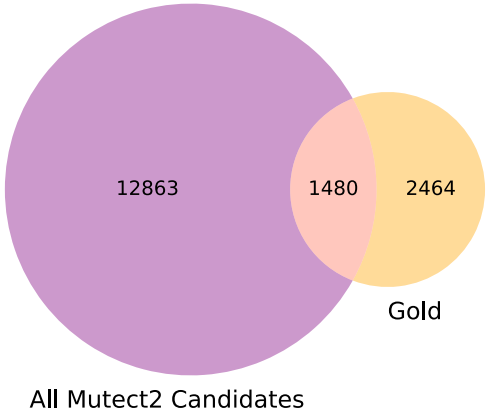
Number of supporting samples increased for 23% of variants

Exposure to same signatures, with 27% more variants

Acute Myeloid Leukemia (AML) dataset* with curated set of SNVs

- Normal sample
- Primary tumor sample
- Relapse sample

WGS, 312x
WES, 433x
Custom targeted capture, 1500x → Gold set of SNVs (3944)

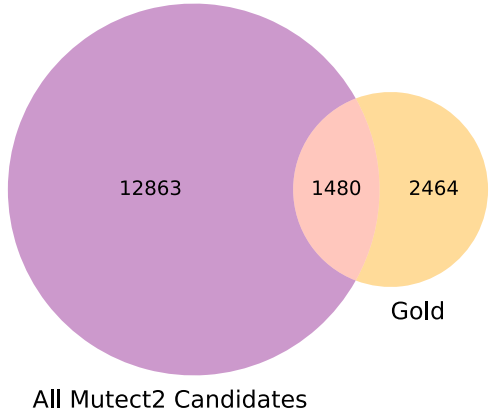


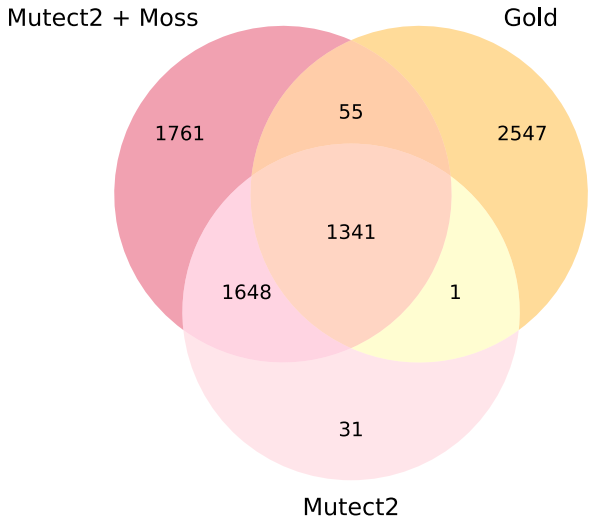
*Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).

Acute Myeloid Leukemia (AML) dataset* with curated set of SNVs

- Normal sample
- Primary tumor sample
- Relapse sample

WGS, 312x
WES, 433x
Custom targeted capture, 1500x → Gold set of SNVs (3944)



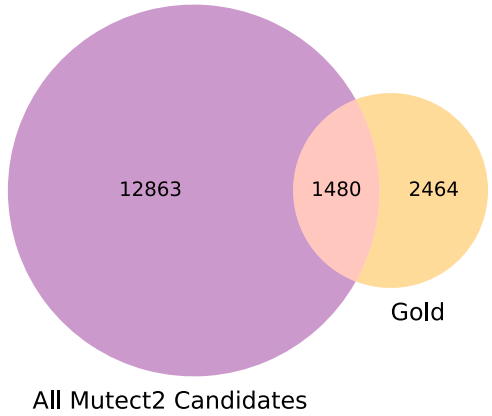


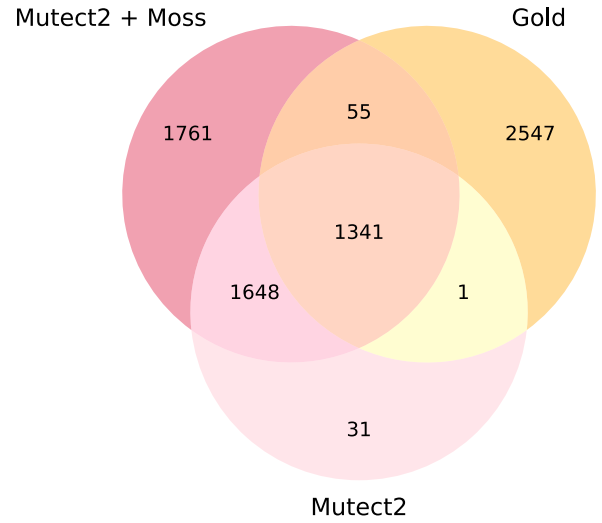
*Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).

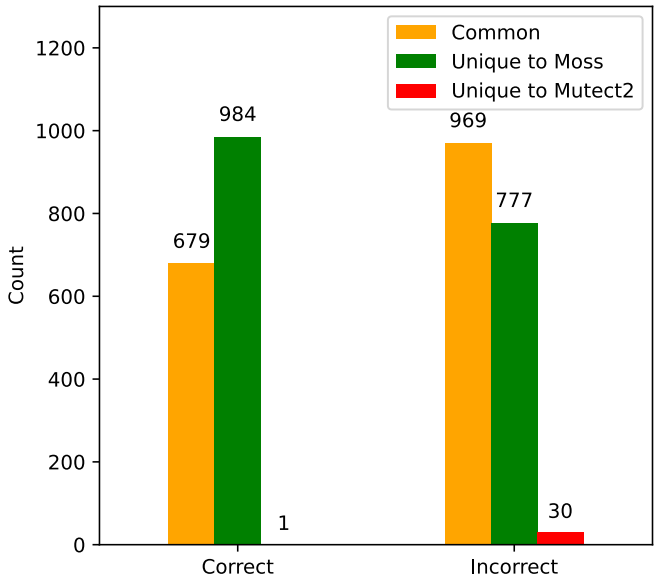
Acute Myeloid Leukemia (AML) dataset* with curated set of SNVs

- Normal sample
- Primary tumor sample
- Relapse sample

WGS, 312x
WES, 433x
Custom targeted capture, 1500x → Gold set of SNVs (3944)







Correct (custom capture):
- VAF<0.05 in normal sample
- At least 5 reads with variant allele in one tumor sample

*Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).

Conclusion: Moss*

- Light-weight versatile multi-sample somatic variant caller
- Transforms any single-caller to multi-sample
- Recovers low-VAF variants maintaining exposure to tumor-specific mutational signatures
- Benefits even with few samples and low coverage
- Increases sensitivity with no (or little) loss of precision
  - Manual review of identified variants for high-quality set
  - Useful in hypothesis-generating context

*Zhang, Chuanyi, Mohammed El-Kebir, and Idoia Ochoa. "Moss enables high sensitivity single-nucleotide variant calling from multiple bulk DNA tumor samples." *Nature communications* 12.1 (2021): 1-10
https://github.com/elkebir-group/Moss

Future directions

- Adapt Moss to
  - single-cell DNA sequencing data (scDNA-Seq) → scMoss
  - hybrid datasets: single-cell and bulk DNA sequencing data
  - long read technologies
  - detect small INDELs and large structural variants
  - incorporate additional information

- Tumor phylogeny reconstruction from scDNA-Seq data*

*https://www.biorxiv.org/content/10.1101/2022.04.18.488655v1.full