Targets

Bioinformatics Plus Comp. Bio BMC bioinformatics and now Nature Scientific Reports.

Introduction

- almost all recent work has been on the heuristic aspects of variant calling [@Alexandrov2013].
- very little attention to the statistical model, either in competition or development
- there is useful biology. . . ..
- Rather than using a constant probability for mutation, as other variant callers do, we convert that to an average or expected mutation probability, and compute the probability conditional on context and genome composition
- Poisson models make similar assumptions about the probability of an allele at a site. Or do they, they are only looking at error rate(Illumina technical note https://www.illumina.com/Documents/products/technotes/technote_somatic_variant_caller.pdf).
- we simulate neutral tumor evolution, and assign vafs using a Beta(1,6) distribution
  - if M(f) is proportional to 1/f, then an exponential distribution is implied (martincorena in mendeley). We choose a beta distribution to achieve a slightly fatter distribution in the 2-5% range in which we are most interested.

Results

Sensitivity and specificity in simulated data

We can look at interactions now that we have settled on an experimental design.

1. Brief description of simulations, see methods
2. Words about the figure
   - What is the linear regime in the Mutect ROC curves about?
   - Is it related to the uniform prior, and does it give a good explanation of the performance difference?

Experiment 2 is a 100X whole genome with ~29000 spiked variants, most of which are under 2% because of the way the simulation works.

Experiment 10 is a 100X whole genome with the same variants as Experiment 2, but with a uniform vaf distribution. Prior method is still better, but in the

uniform scenario there are only a small fraction of the total variants that are challenging to call.

Experiment 9 is a whole exome that was supposed to have 1,7,11, but instead has a random set of signatures due to a bug

Sensitivity in real data

We examined two validation datasets from real tumors. An acute myeloid leukemia whole genome was sequenced to average coverage of 365X, and over 200,000 mutations validated by deep sequencing, generating a set of "platinum" consensus calls for the tumor. In addition to the full dataset we also called mutations on two downsample datasets, one retaining 50% of the original reads and one retaining 25%. ROC curves were generated using the "platinum" calls as cases, and sites where validation sequencing depth was greater than 100X and no variant reads were found as controls. Both algorithms perform similarly and nowhere along the curve is the {what is the name of this thing} method below raw mutect calls. The {method} calls a higher fraction of platinum calls at every odds threshold, and is especially effective at the common threshold of 2:1 odds in favor of the mutation.

Going to need a table of AUROCs in the supplement for this

Effect of odds threshold

This has very little effect, even in an exome, as the figure inserted shows. 1. As threshold goes to infinity you get mutect. 2. As threshold goes to zero you should also get mutect. 3. Observe very little difference in the middle

Effect of number of mutations

We will have this from the difference between exome and wgs on the same vaf distribution and signature. This is likely to have some signature dependence. 1. How to approach this? - At what point does the empirical make more sense than the dirichlet. - I think never, they will converge - What is the stopping point with a low number of high confidence mutations - Implementation of the dirichlet should let us create an estimation of total error between the final empirical at a given threshold and the dirichlet at every point in the process. Maybe a plot of this?

Effect of variant allele frequency distribution

1. TCGA data for different distributions.

- Different cancer types?
- Hypermutators vs. not?
- This should only be related to the number of mutations that are confident and contribute to the prior
- If that is the case, is there an analytical way to better describe this? THE ONLY EFFECT IS ON THE ROC. EASIER DISTRIBUTIONS SHRINK THE EFFECT BECAUSE SO FEW ARE NEAR THE CRITICAL POINT

Methods

100X whole genome and 500X whole exome for each of three signatures

1,7,11 UV (Very concentrated at C>T) 1,4,5 Tobacco (Slight concentation at C>A and C>T) 1,3,5 Breast (diffuse)

All vafs will be from the beta(1,6) which is a fat exponential

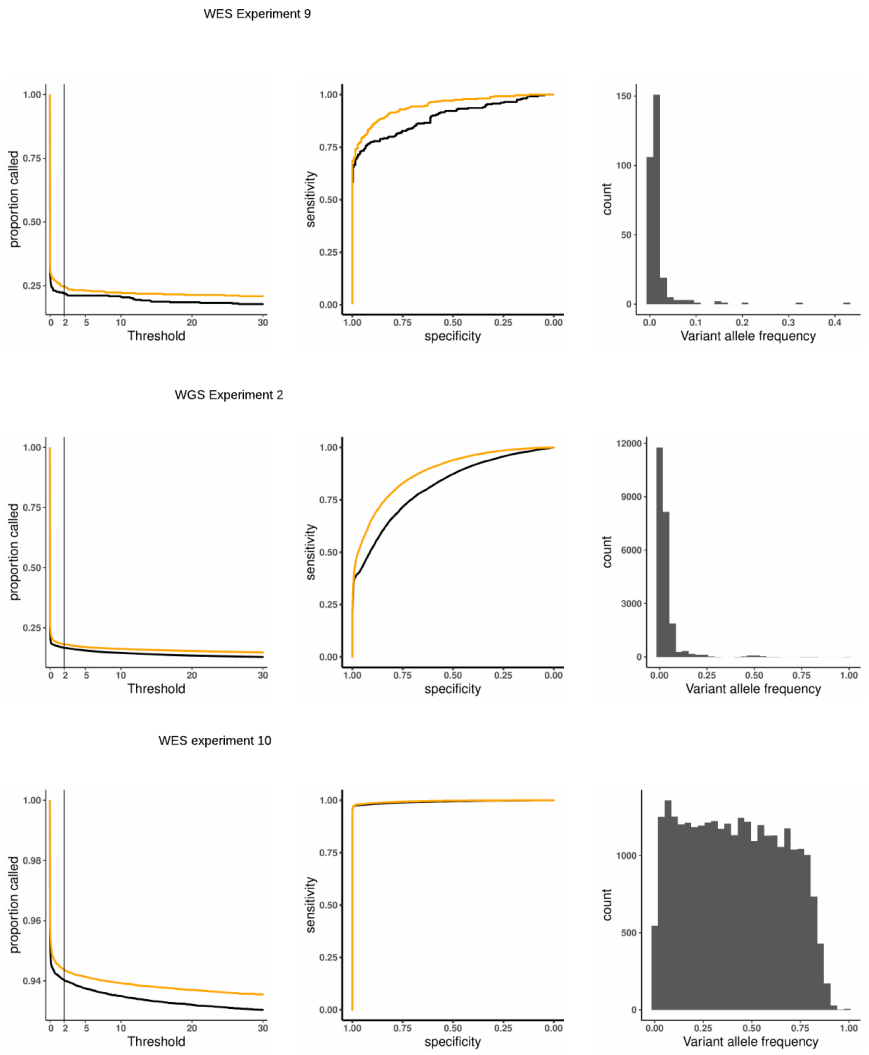
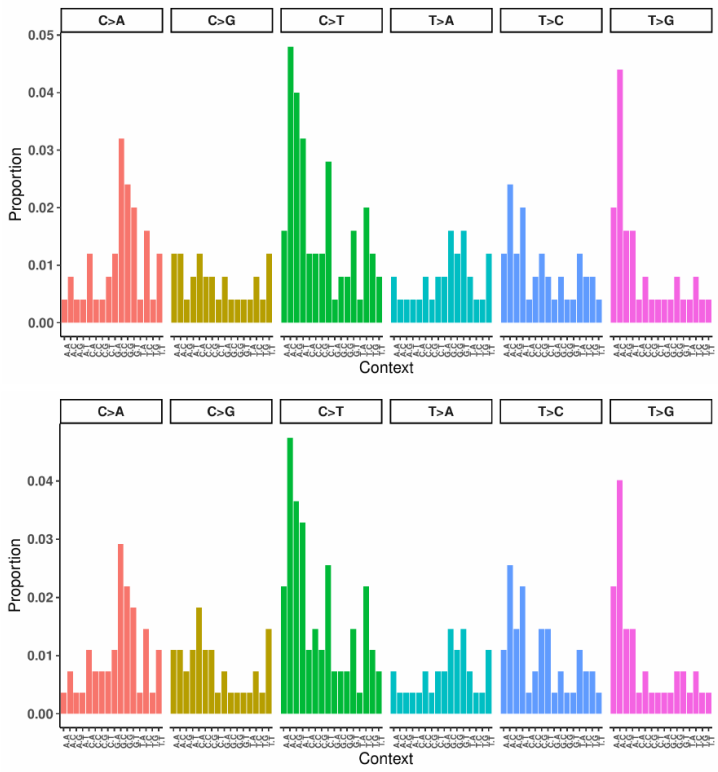


Figure 1: roc curve figure experiment 9



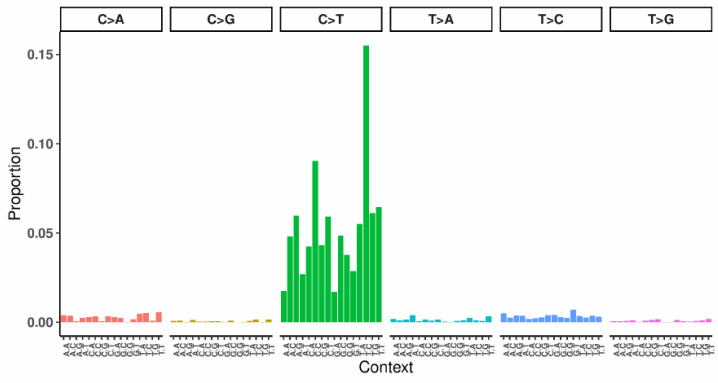


Figure 2: figure experiment 9

Figures

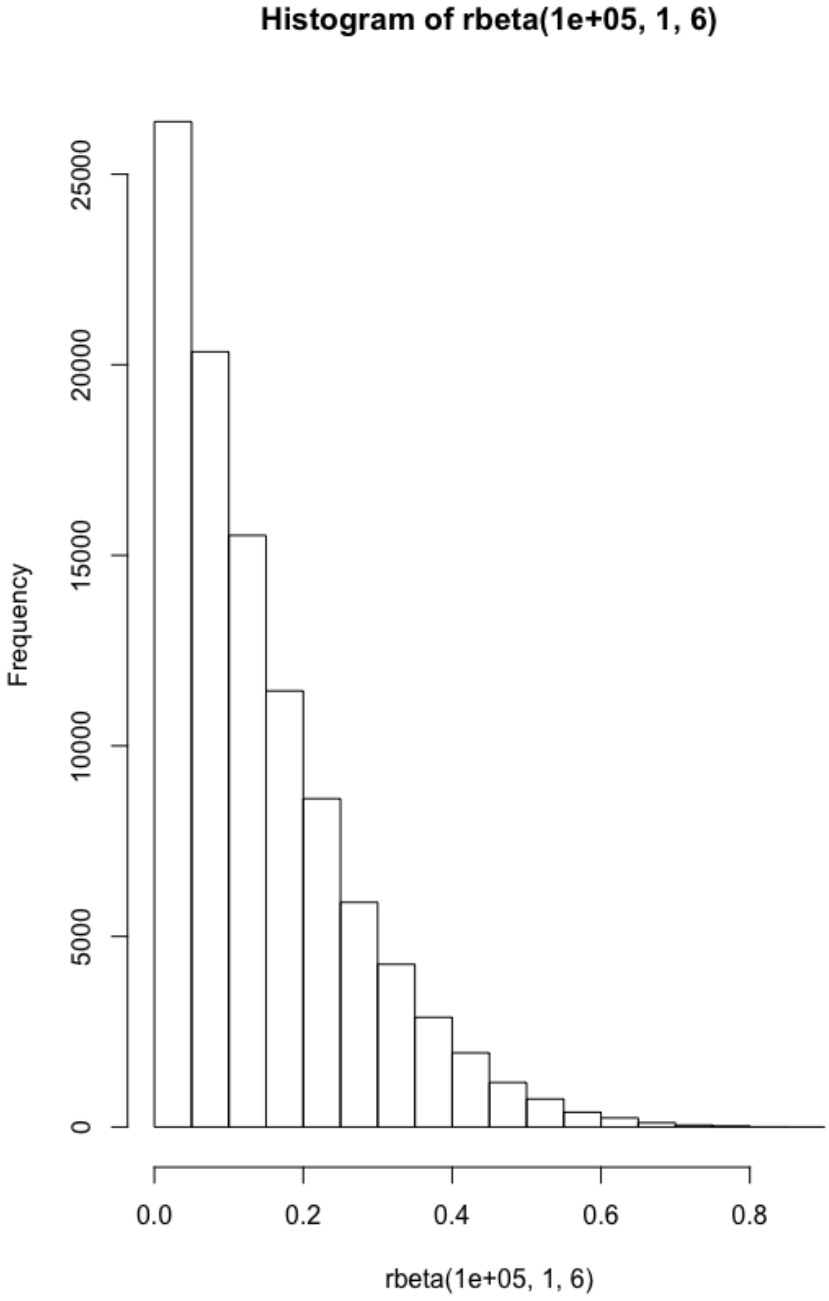

Histogram of rbeta(1e+05, 1, 6)
Frequency
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
rbeta(1e+05, 1, 6)

Figure 1 - aml31 no downsample roc

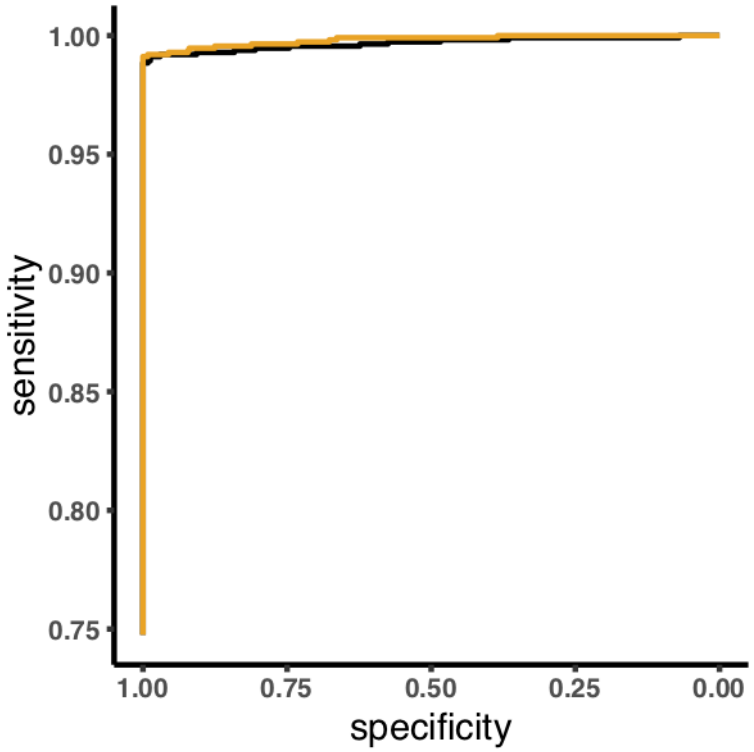


Figure 3: Figure 1 - aml31 no downsample roc

Figure 2 - aml31 no downsample fraction called

Figure 2a - aml31 no downsample vaf

Figure 3 - aml31 50 percent downsample roc

Figure 4 - aml31 50 percent downsample fraction called

Figure 4a - aml31 50 percent downsample vaf

Figure 5 - aml31 25 percent downsample roc

Figure 6 - aml31 25 percent downsample fraction called

Figure 6a - aml31 25 percent downsample vaf

Figure 7 - cell paper roc

Figure 8 - cell paper fraction called

Figure 8a - cell paper vaf

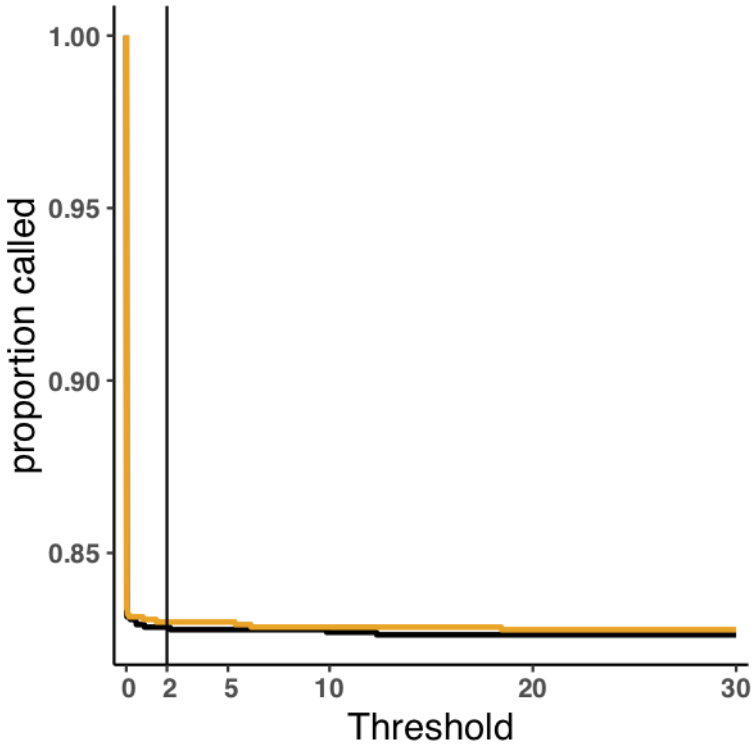


Figure 4: Figure 2 - aml31 no downsample fraction called

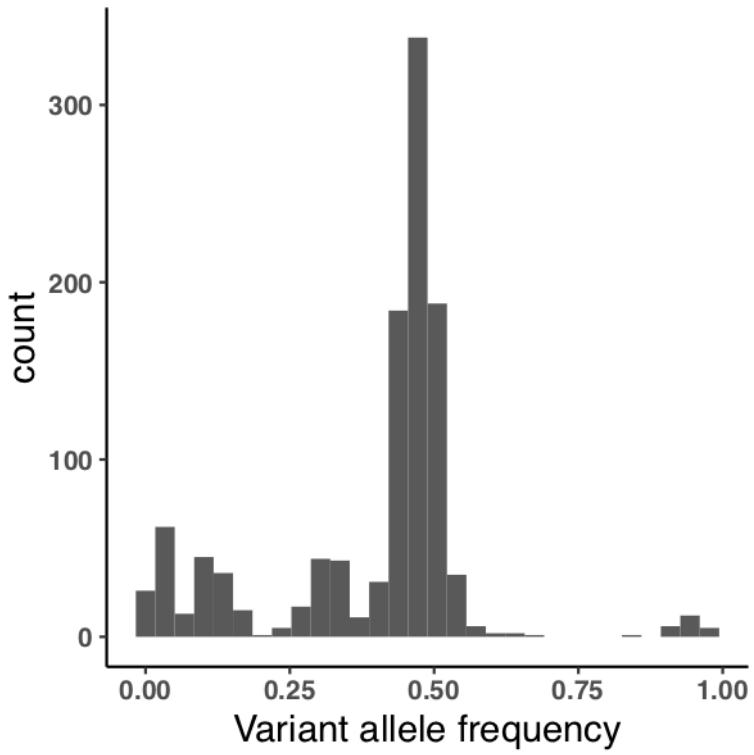


Figure 5: Figure 2 - aml31 no downsample vaf

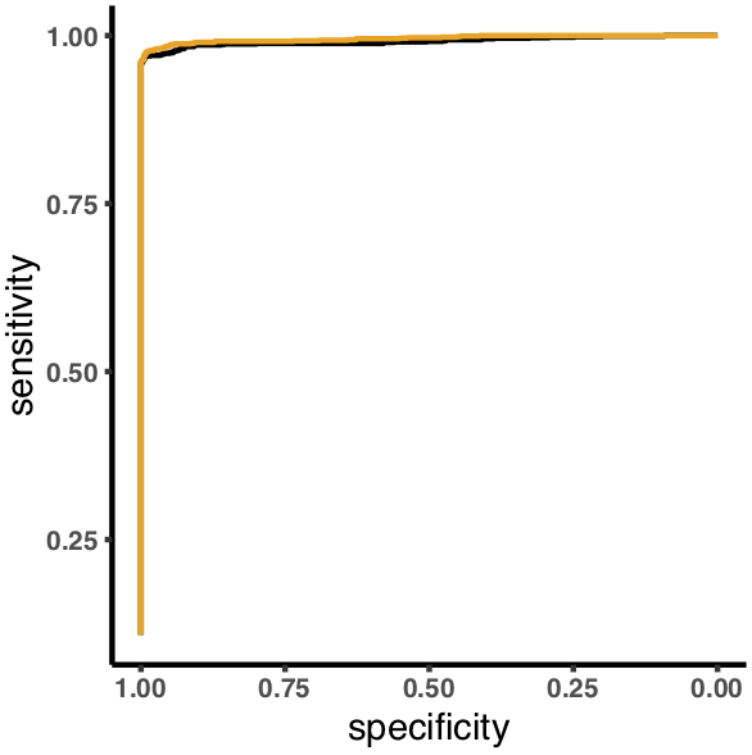


Figure 6: Figure 3 - aml31 50 percent downsample roc

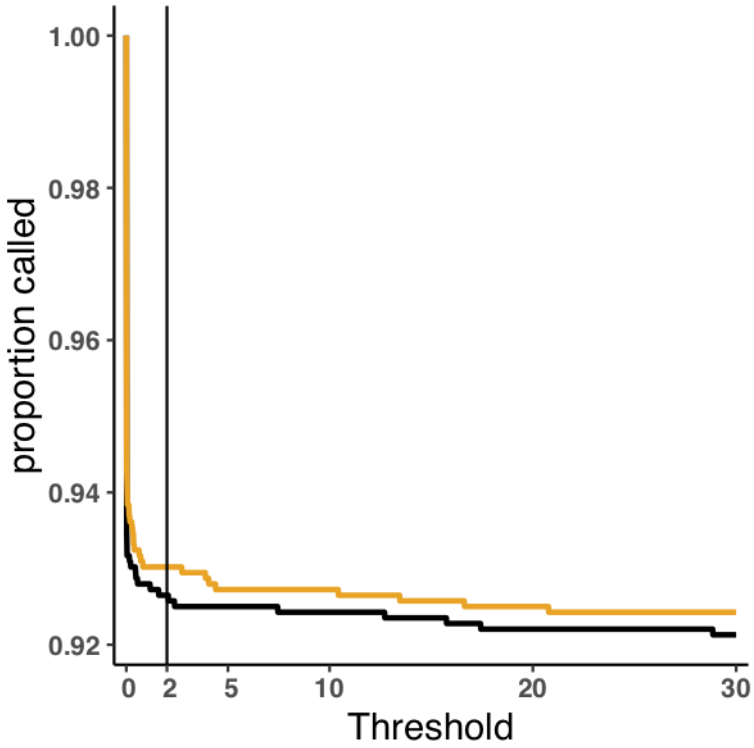


Figure 7: Figure 4 - aml31 50 percent downsample fraction called

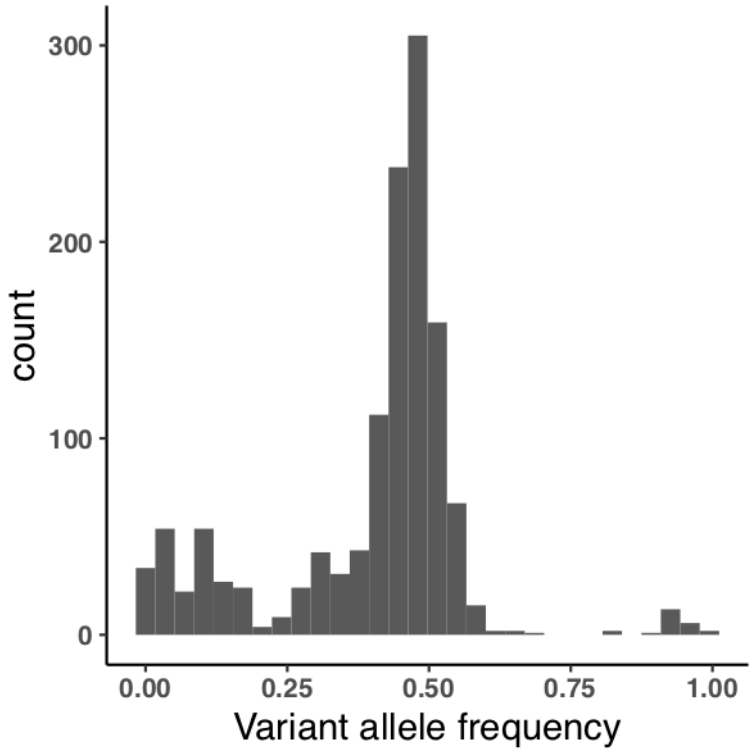


Figure 8: Figure 4 - aml31 50 percent downsample vaf

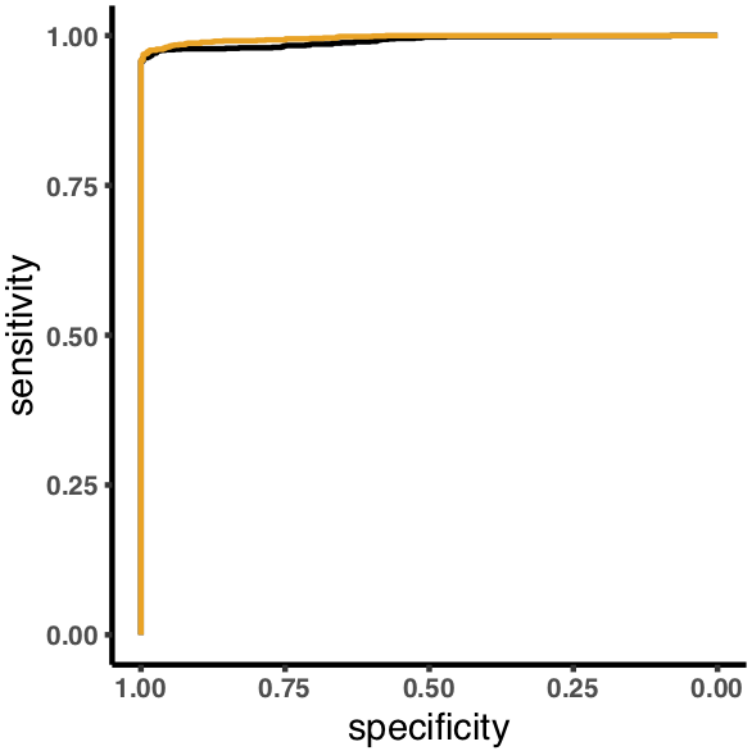


Figure 9: Figure 5 - aml31 25 percent downsample roc

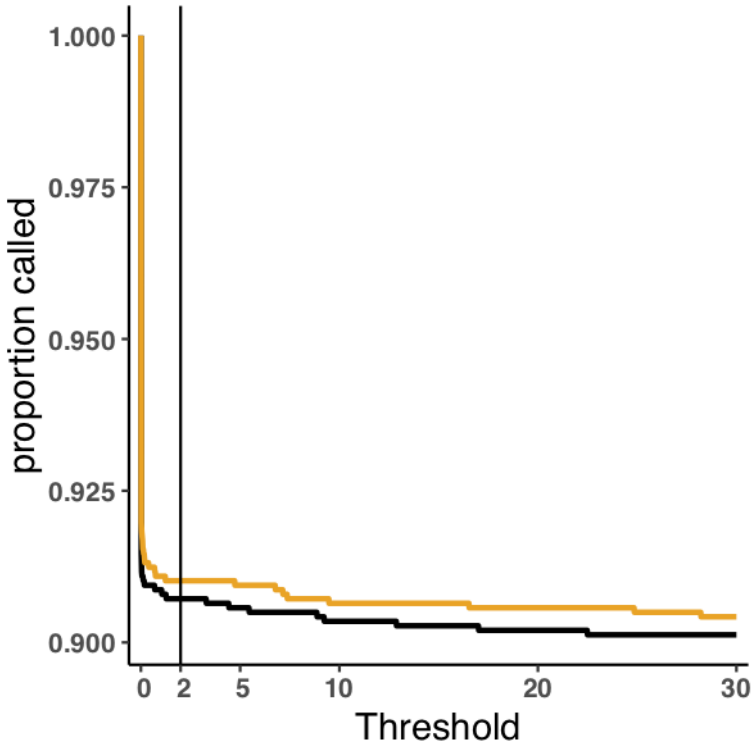


Figure 10: Figure 6 - aml31 25 percent downsample fraction called

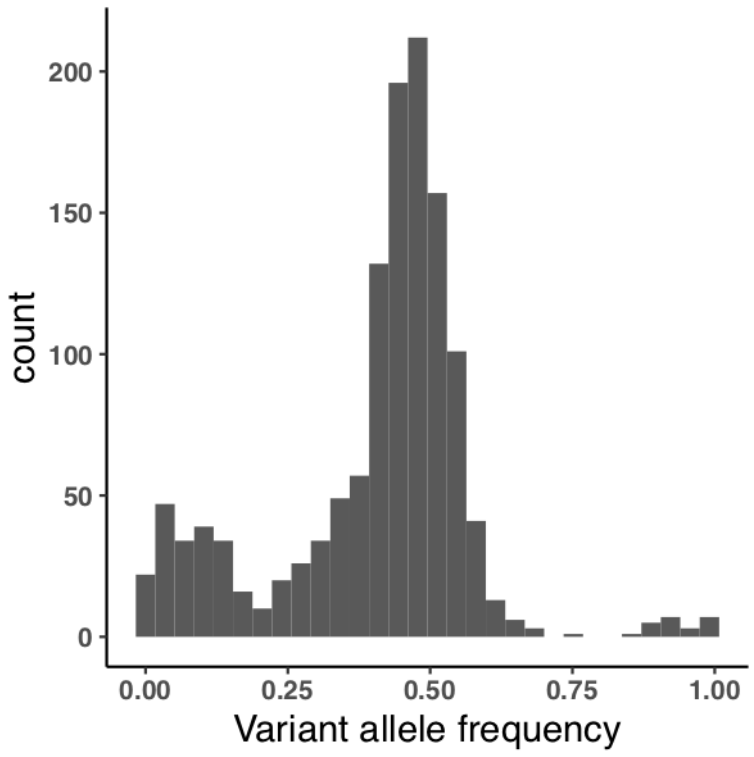


Figure 11: Figure 6 - aml31 25 percent downsample vaf

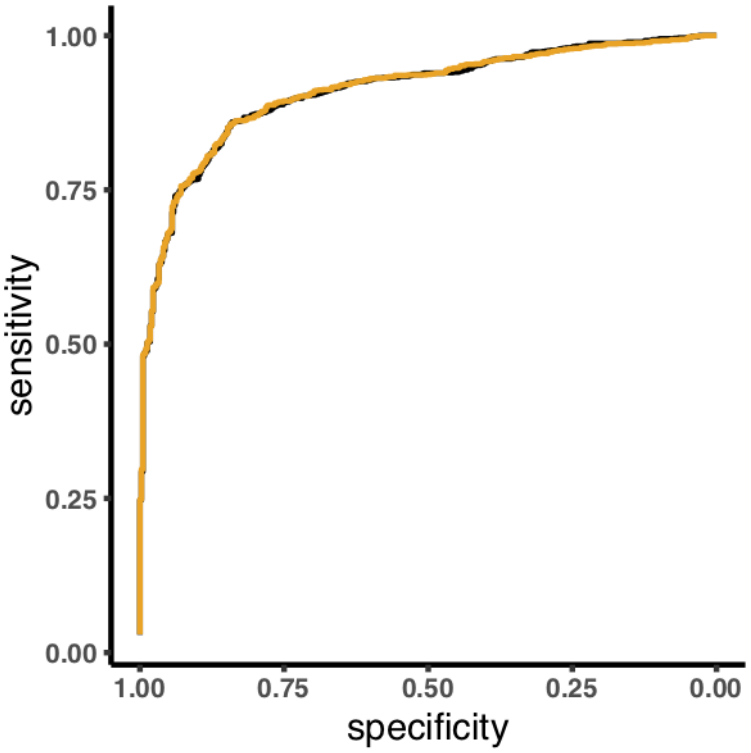


Figure 12: Figure 7 - cell paper roc

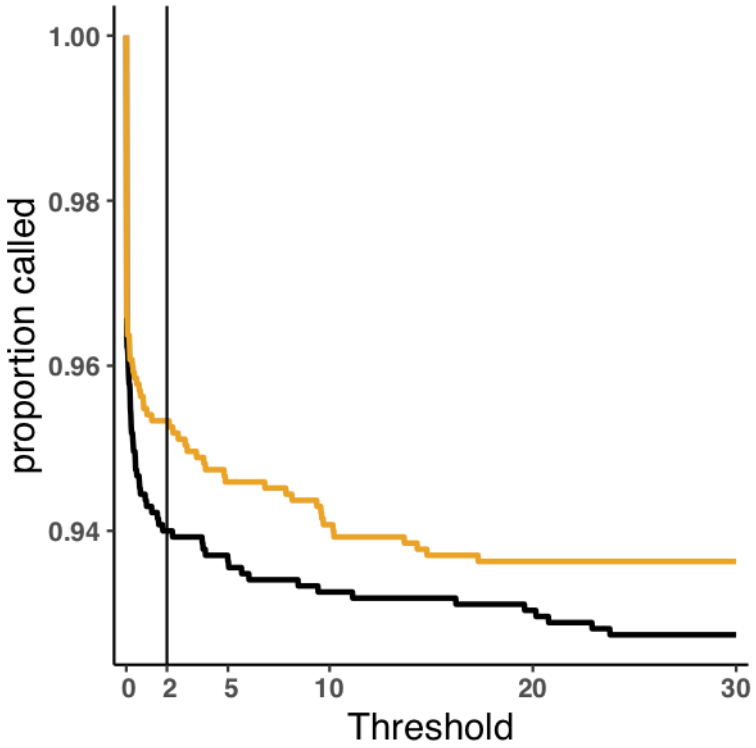


Figure 13: Figure 8 - cell paper fraction called

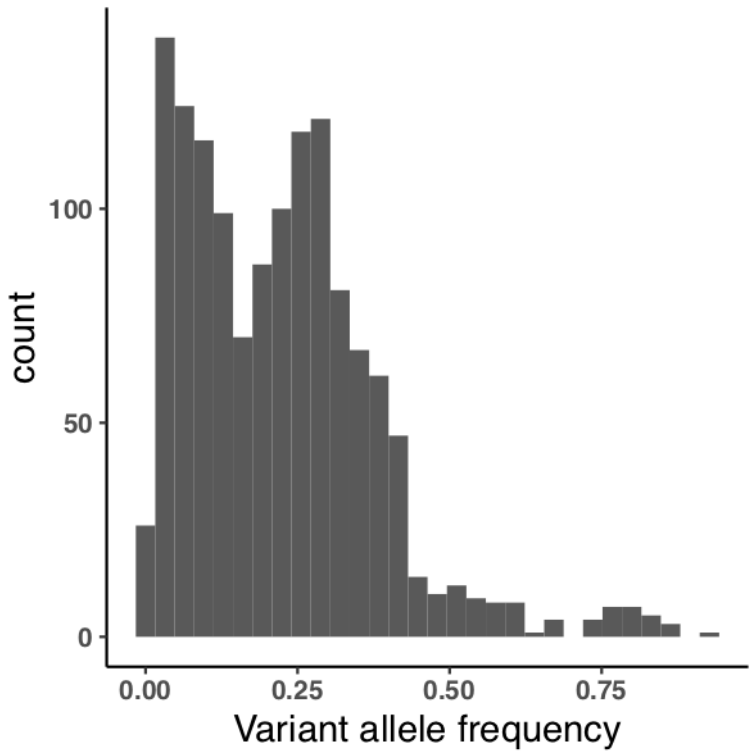


Figure 14: Figure 8 - cell paper vaf

References