Coverage Normalization

April 2, 2020

Melanie Kirsche, Srividya Ramakrishnan,
Mike Alonge, Peter Thielen, Tom Mehoke,
Winston Timp, Michael Schatz

Coverage Normalization

Many regions of the genome have very high coverage (>1000x) because of uneven amplification/proximity to primers, but random downsampling can result in losing information at regions with already low coverage (<50x)

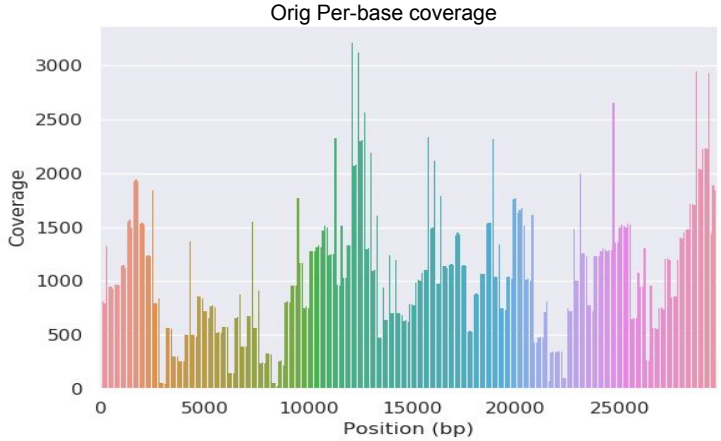
Threshold Sampling Method

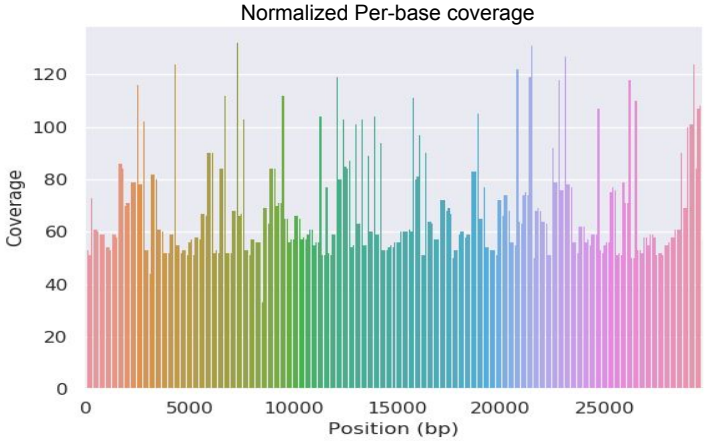
1. Shuffle the reads
2. Iteratively accept a read if it spans some base which has been covered by less than threshold number of reads

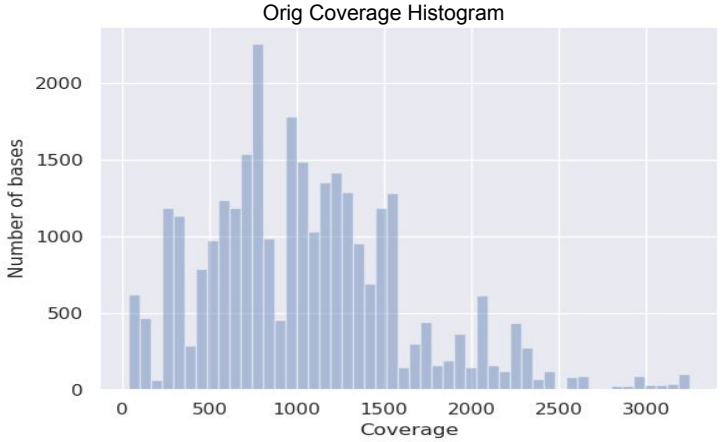
Note: Full coverage is guaranteed across low coverage regions

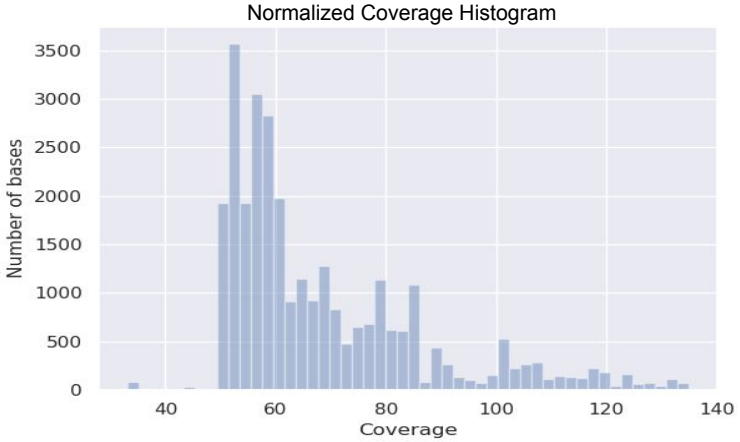
Software available: https://github.com/mkirsche/CoverageNormalization

Coverage Before and After Normalization

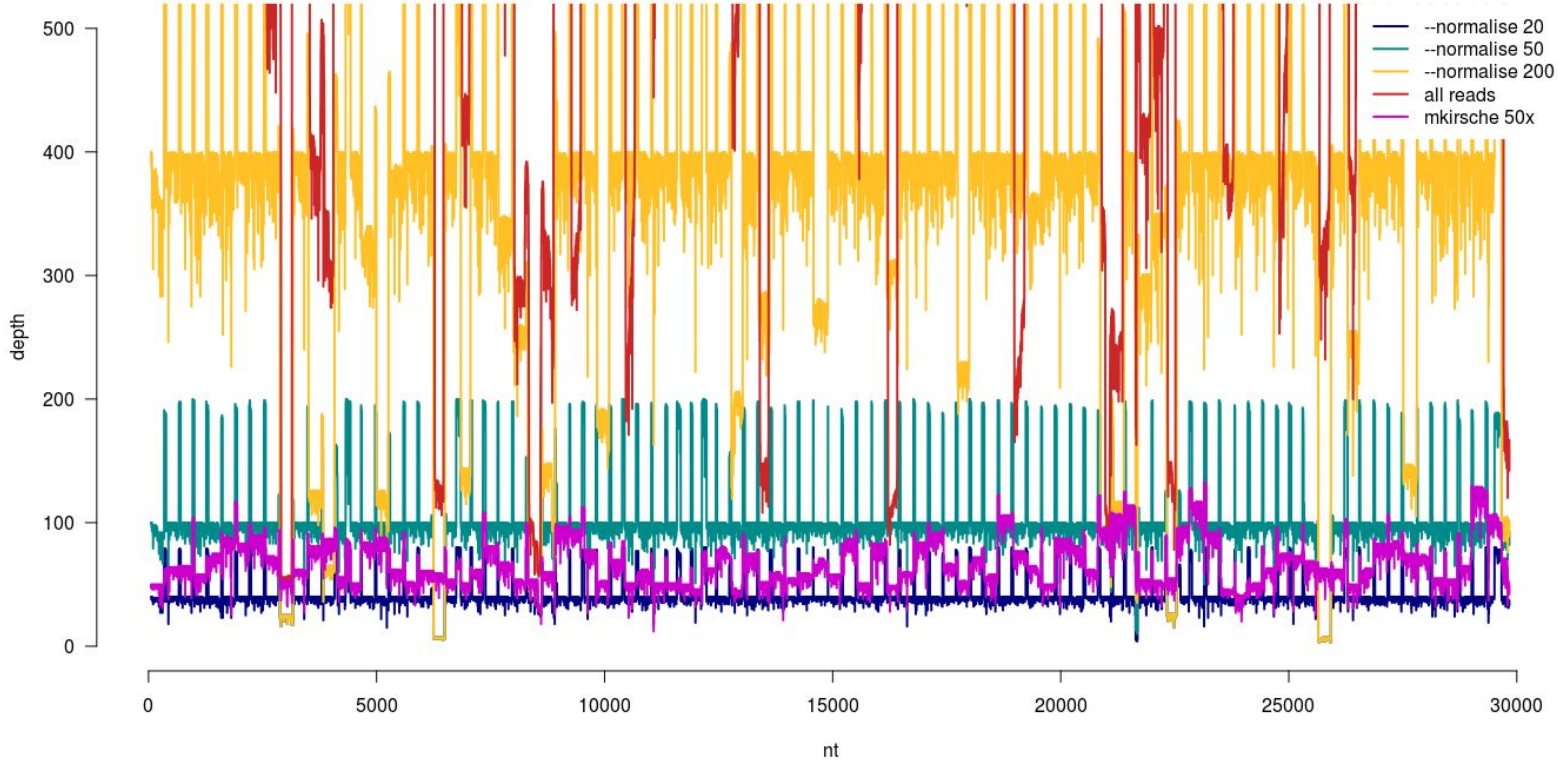






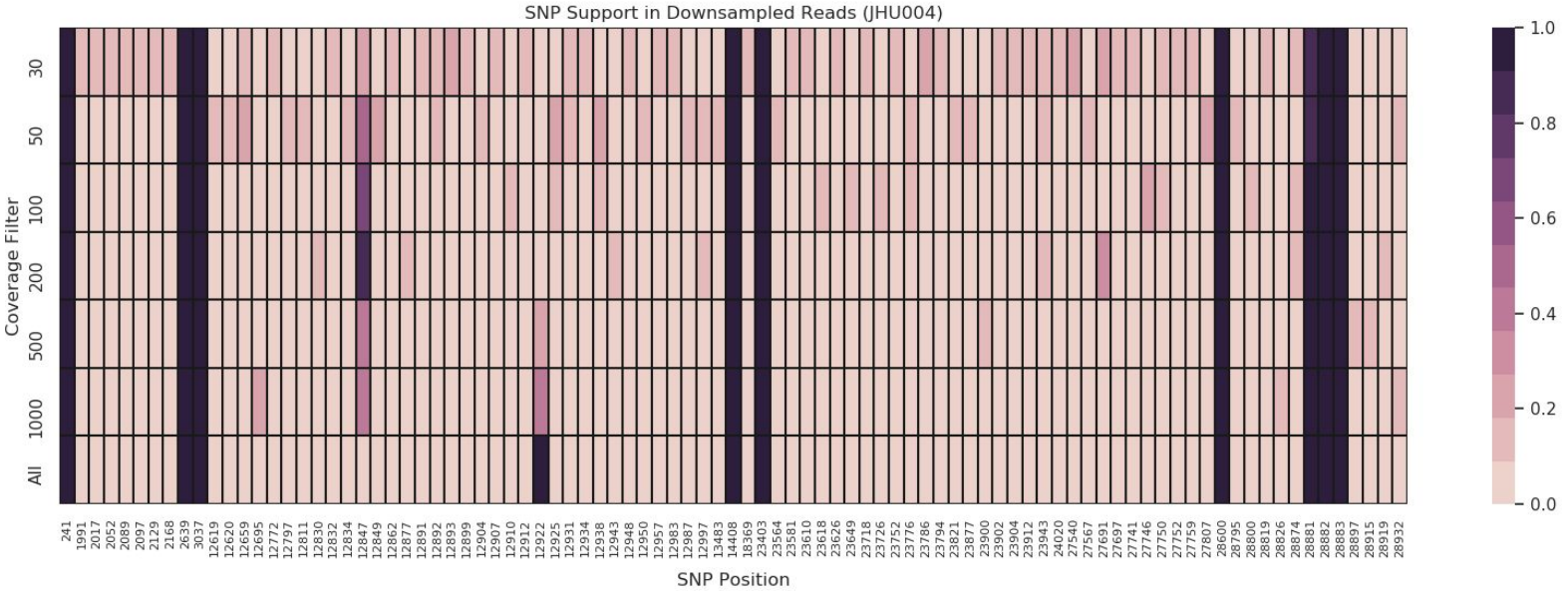


Comparison to artic --normalise



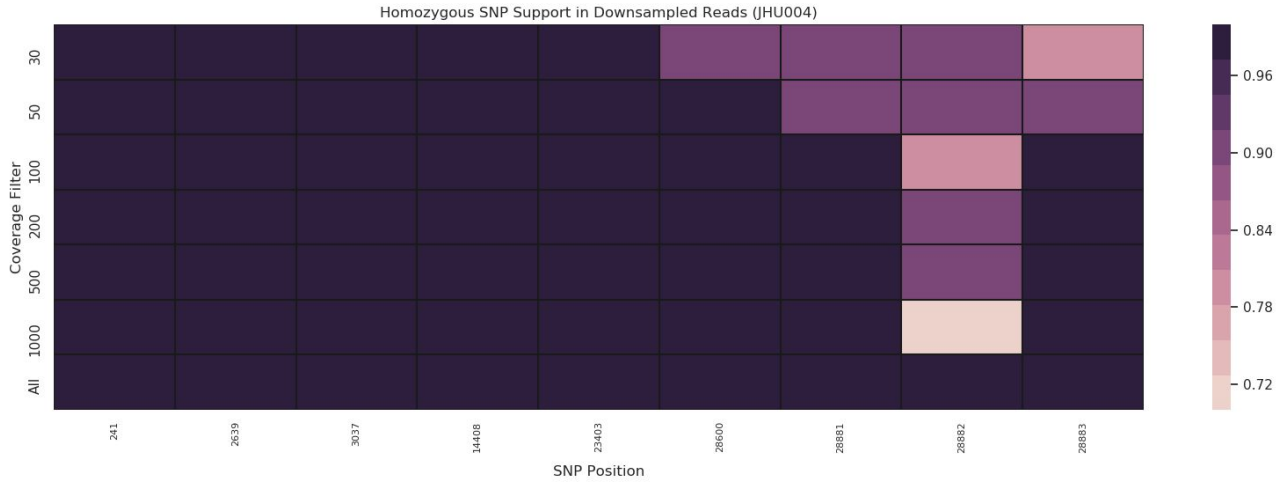
Our normalization leads to fewer low coverage regions and fewer reads overall even when targeting the same amount of coverage

How does normalization affect variant calling?



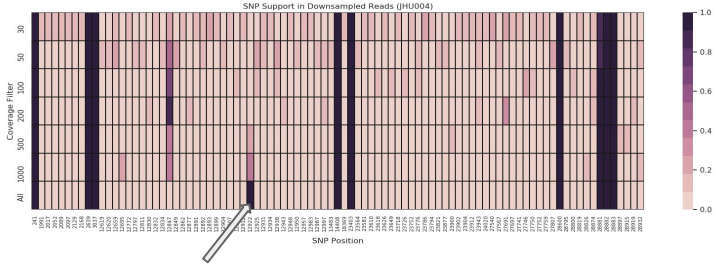
For each coverage level, create 10 samples with that threshold and measure the proportion of samples in which medaka found each variant.

Affect on Homozygous Variant Calls



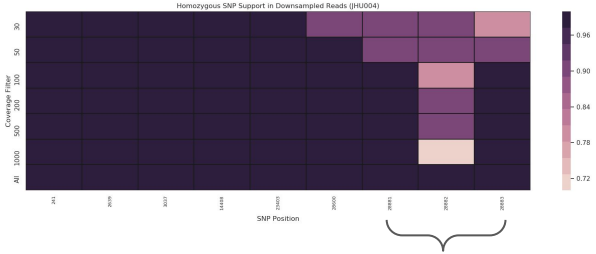
We repeated the experiment in the previous slide, focusing on homozygous calls from each sample.

False heterozygous SNP at full coverage (12922)




```
MN908947.3  12922  .  A  C  14.0  PASS  primary_call=A;primary_prob=0.910;ref_prob=0.910;secondary_call=C;secondary_prob=0.050  GT:GQ  0/1:14.000
```

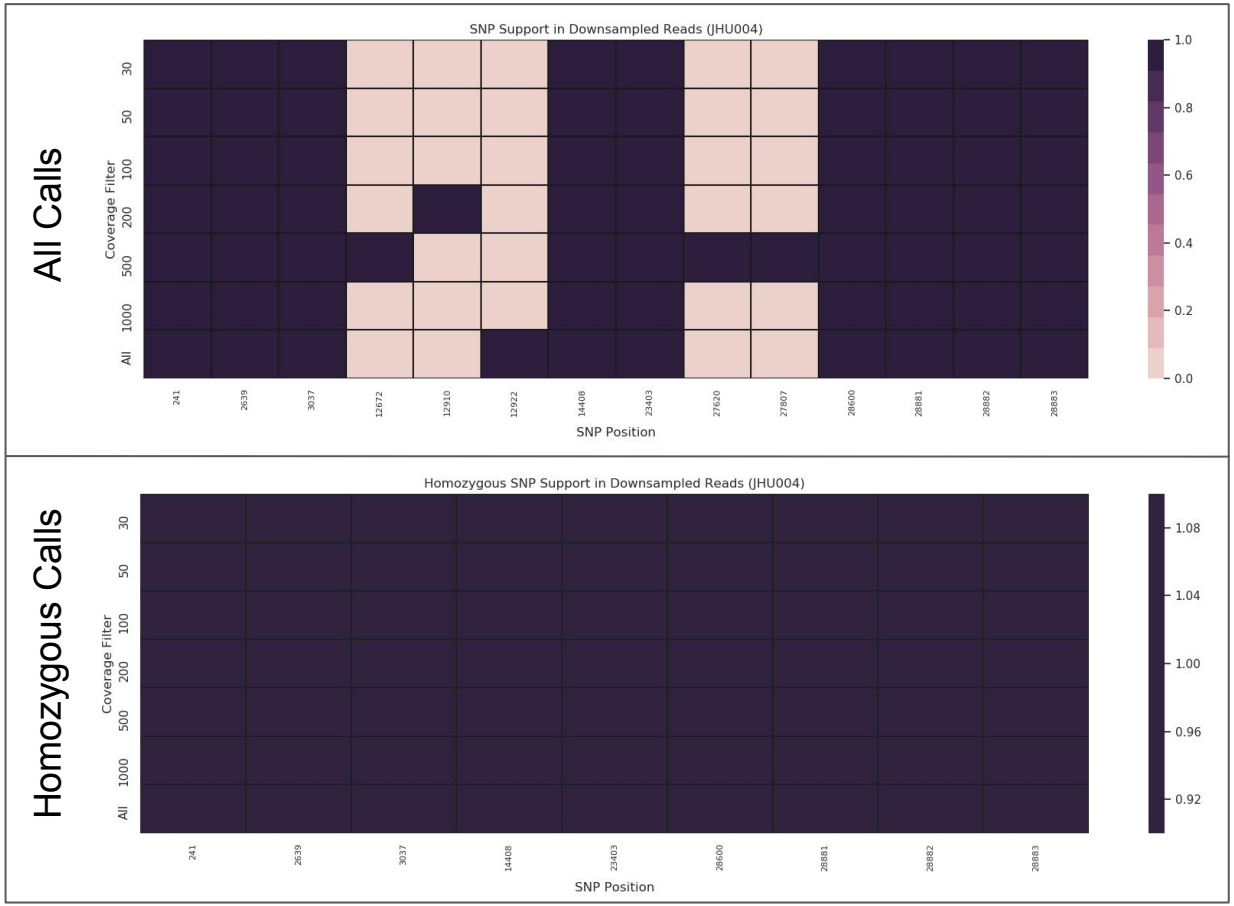
Some homozygous SNPs called as heterozygous at lower coverage



Clear mutation signatures, but also frequent 200 bp deletion, so actual coverage may be lower than expected

Normalization based on alignment accuracy

To avoid sampling many reads with poor alignments, we added an option to consider the reads in order of alignment accuracy rather than in a random order.



Normalization Advantages

- ➔ Faster processing with similar variant call results
- ➔ Variant calls from multiple samples can highlight problematic regions
- ➔ Lower memory for archiving with little information loss
- ➔ Repackaging Fast5 files to the downsampled reads (using ont_fast5_api) reduces space requirements from ~92 GB per sample to 0.34 GB per sample.
  - ◆ Makes it feasible to archive 1000s of samples for signal level analysis at a future date