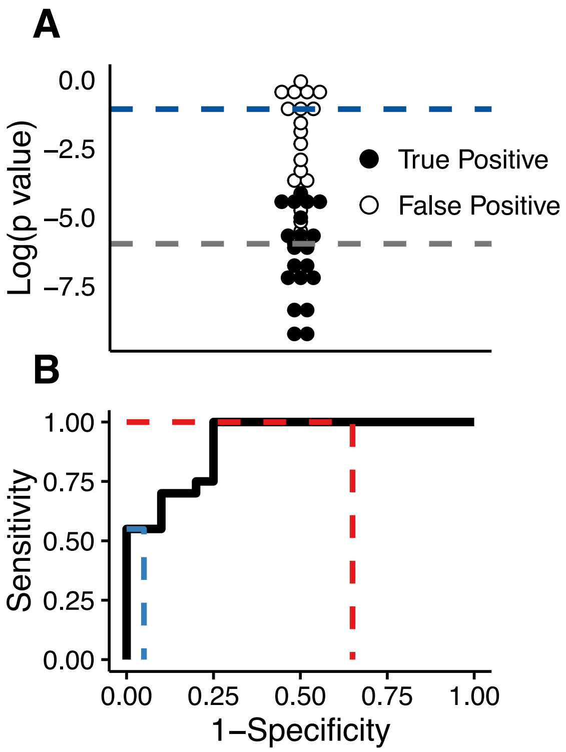
Figures

JT McCrone

February 9, 2015

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

Figure 1

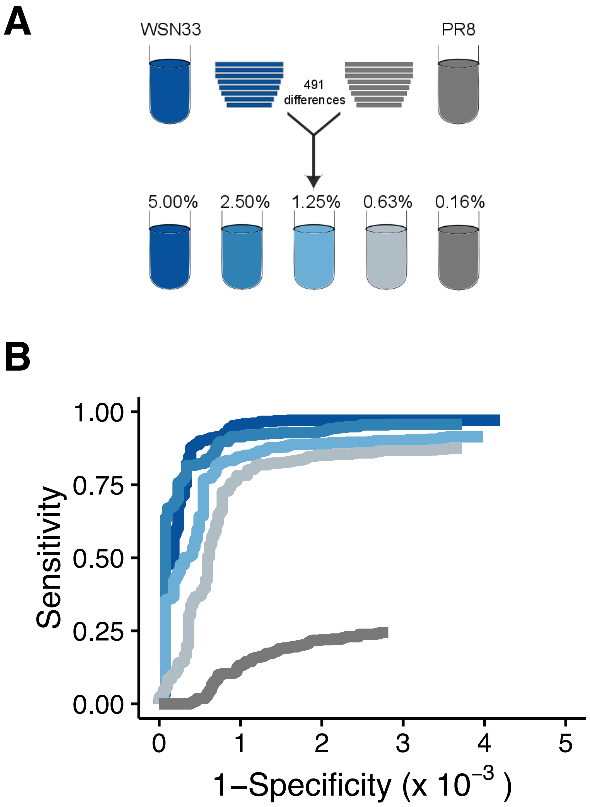


```
## pdf
##   2
```

Figure 1. An example ROC. A) Hypothetical variants are stratified by p value. True variants are shown as filled circles, while those corresponding to RT, PCR, or sequencing errors are shown as open circles. Potential thresholds are indicated as colored lines. B) An ROC made from the data shown in A. The colored lines indicate the points made by the thresholds in A.

Figure 2

```
## pdf
##   2
```



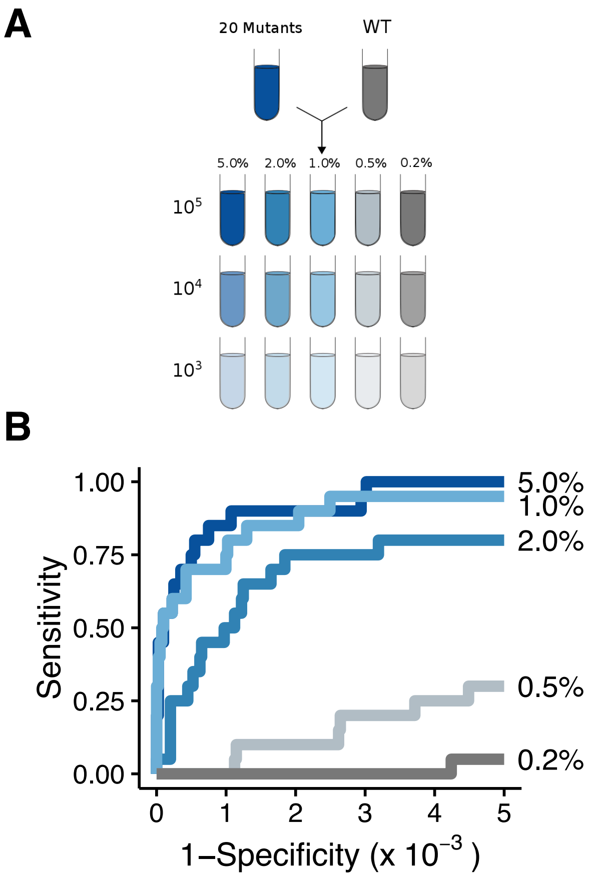
C

Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	0.9715	477	0.9982	68
2.5%	0.9348	459	0.9980	78
1.25%	0.8900	437	0.9981	75
0.63%	0.8330	409	0.9983	66
0.016%	0.1426	70	0.9989	41

Figure 2. Initial DeepSNV accuracy. A) RT-PCR amplified WSN33 genomes were diluted into PR8 at decreasing frequencies and sequenced on an Illumina Miseq. B) An ROC measuring the pipeline's ability to identify WSN33 SNVs relative to PR8 at the indicated frequencies. C) A summary of the DeepSNV accuracy at a p value threshold of 0.01.

Figure 3

```
## pdf
##   2
```



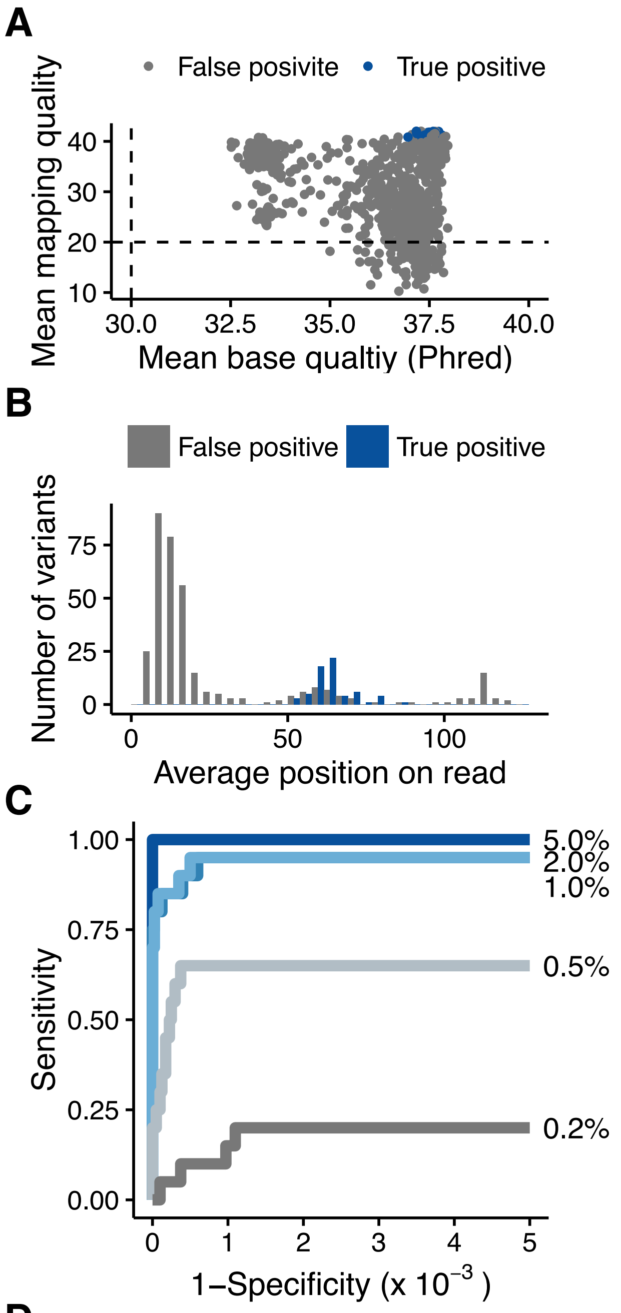
C

Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	1.00	20	0.9938	250
2.0%	0.85	17	0.9942	232
1.0%	0.95	19	0.9975	101
0.5%	0.35	7	0.9948	210
0.2%	0.05	1	0.9937	252

Figure 3. DeepSNV accuracy patient sample conditions A) 20 known single point mutants were diluted into wild type at decreasing frequencies. These were diluted further into viral media to match the genomes concentration found in patent samples (10^5-10^3 genomes/μl) and sequenced on an Illumina Hiseq (2 x 152bp reads. B) An ROC indicating the ability to accurately identify the 20 known variants in the 10 ^{5} genomes/μl samples. C) A table of the accuracy of our variant calling for the 10 ^{5} genomes/μl samples at a p value threshold of 0.01.

Figure 4

```
## pdf
##   2
```



D

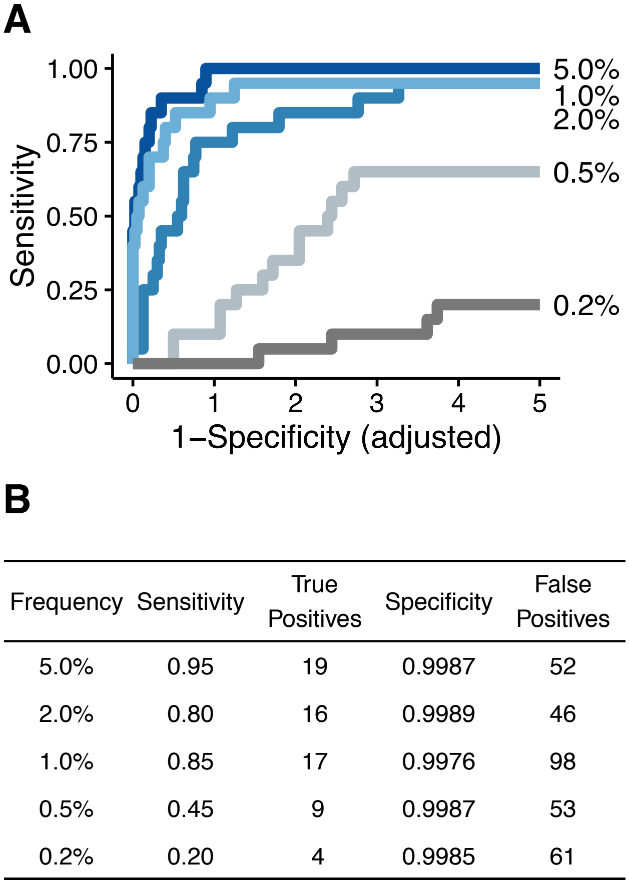
Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	1.00	20	>0.9999	2
2.0%	0.85	17	0.9999	6
1.0%	0.95	19	0.9995	20
0.5%	0.35	7	0.9999	5
0.2%	0.05	1	0.9997	11

Figure 4. Investigating the quality of variants in the the 10^{5} genomes/μl samples. A) All called variants from

the 10 ^{5} genomes/μl samples with p<0.01 stratified by the mean mapping quality of the reads containing the variant and the mean Phred scores of the variant bases. Dashed lines indicate common cutoffs of 20 and 30 for mapping quality and Phred respectively. B) A histogram of all called variants from the 10 ^{5} genomes/μl samples with p<0.01 binned by the average position of the variant relative to the Hiseq reads on which it is found. C) An ROC of the accuracy of the variant caller when the following quality cut offs are applied : mean mapping quality of 30, mean Phred of 35, and an average read position between 32 and 94 (the middle 50% of the read). D) A table summarizing the ROC in C at a p value threshold of 0.01.

Figure 5

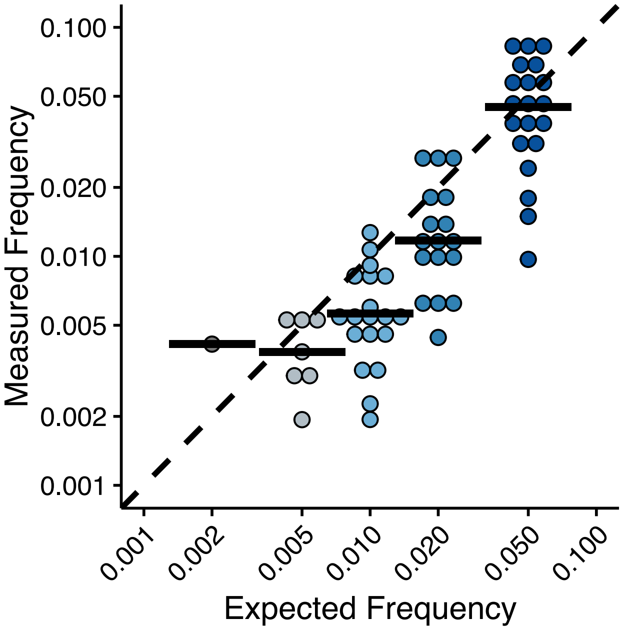
```
## pdf
##   2
```



Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	0.95	19	0.9987	52
2.0%	0.80	16	0.9989	46
1.0%	0.85	17	0.9976	98
0.5%	0.45	9	0.9987	53
0.2%	0.20	4	0.9985	61

Figure 5. The accuracy of Lofreq. The variant caller Lofreq was applied to the 10^{5} genomes/μl samples. A) An ROC of the accuracy of Lofreq using standard parameters. The specificity of the caller was scaled to account for the same number of tests as the DeepSNV algorithm. Lofreq provides quality scores instead of p values. These were used to set cutoffs for the ROC. B) A table summarizing the accuracy of the ROC A using the standard cut offs applied by the algorithm.

Figure 6



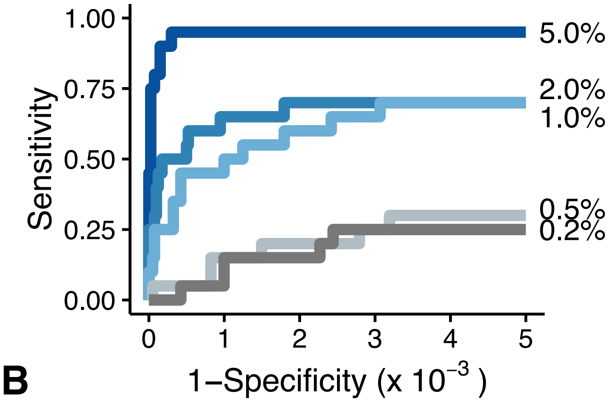
```
## pdf
##   2
```

Figure 6. The accuracy of the frequency measurements for the known variants in the 10 ^{5} genomes/μl samples. The median of each distribution is shown as a black bar.

Figure 7

```
## pdf
##   2
```

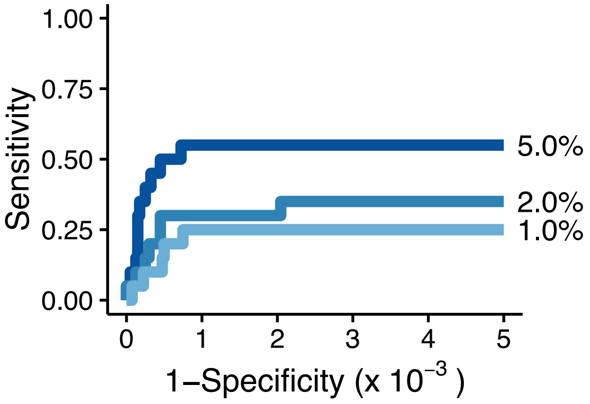
A



B

Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	0.95	19	0.9990	41
2.0%	0.65	13	0.9985	60
1.0%	0.55	11	0.9983	67
0.5%	0.20	4	0.9983	69
0.2%	0.15	3	0.9983	70

C



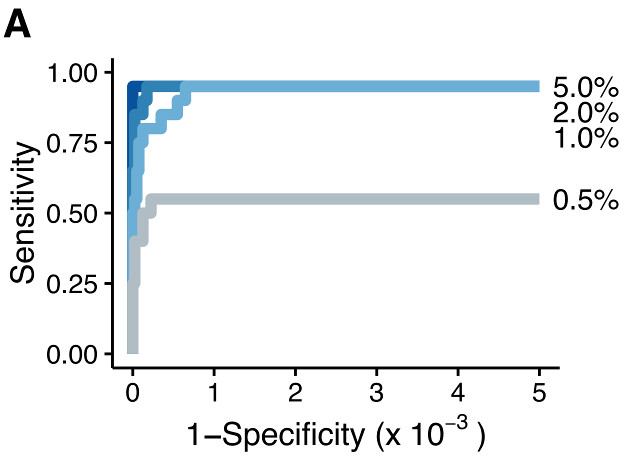
D

Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	0.55	11	0.9990	42
2.0%	0.30	6	0.9981	78
1.0%	0.25	5	0.9978	87

Figure 7. The accuracy of DeepSNV at lower input levels. A) An ROC depicting the accuracy of our method when applied to the 10 ^{4} genomes/μl samples. B) A table summarizing the accuracy at 10 ^{4} genomes/μl with a p value threshold of 0.01. C) An ROC depicting the accuracy of our method when applied to the 10 ^{3} genomes/μl samples. B) A table summarizing the accuracy at 10 ^{3} genomes/μl with a p value threshold of 0.01

Figure 8

```
## pdf
##   2
```



B

Frequency	Sensitivity	True Positives	Specificity	False Positives
5.0%	0.95	19	0.9999	4
2.0%	0.90	18	0.9999	6
1.0%	0.80	16	0.9998	10
0.5%	0.40	8	0.9999	3

Figure 8. The accuracy of DeepSNV at lower input levels when run in duplicate. A subset of the 10 ^{4} genomes/μl samples were processed in duplicate to control for RT-PCR errors. A) An ROC of 10 ^{4} genomes/μl samples processed in duplicate. A variant was required to be found in both duplicates to be considered. B) A table summarizing the accuracy when applied to duplicate samples with 10 ^{4} genomes/μl input, $p<0.01$.