Heterozygosity Experiment

This manuscript was automatically generated on June 14, 2021.

## Authors

# Heterozygosity Experiment

|  |  |
| --- | --- |
| Field | Value |
| Project | Plague Denmark |
| Date | 2021-06-14 |

## Objectives

**1. Why do Denmark samples have high counts of heterozygosity?**

Conclusions: - Sites flagged as heterozygous by snippy core primarily have a low genotype quality. - Low quality can occur because the Alternate Allele has low counts. - One explanation is DNA damage (ex. deamination of cytosines).

**2. How does this compare to other Second Pandemic samples?**

Conclusions: - All Danish samples have less homozygous sites than heterozygous sites - The number of heterozygous sites in Danish samples is equal to or less than other Second Pandemic samples.

## Results

Two characteristics are being investigated:

1. Are there more Heterozygous variants than Homozygous variants?
   * A haploid organism (ie. plague) is expected to have more Homozygous variants.
   * More Heterozygous variants may indicated molecules from multiple strains/species.
2. Are the distributions of depth similar between homozygous and heterozygous sites? (peak and spread)
   * A similar distribution depth may indicate the molecules derive from a singular source.

### Baseline

A selection of samples from the Second Pandemic.

* The number of Heterozygous sites reported by Snippy (in this table) is erroneous. This number includes low quality variants which should not be considered ‘true’ heterozygosity.
* Note that the number of Heterozygous SNPs is not proportion to the mean coverage (put a pin in this thought).

!MultiQC Heterozygosity Second Pandemic.png

* Because, the heterozygosity counts in the previous table are informative, Homozygous and Heterozygous sites were extracted directly from the snippy pairwise alignments.
* Two samples are visualized here to show the true number of homo/set sites is very small (ie. not in the 1000s).
* Black Death 8291 is an example of a GOOD sample.
* STN021.A is an example of a SUSPICIOUS sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Homo | Het | Homo/Het | Graph |
| Black Death 8291 | 105 | 64 | 1.64 | |400  |400 |
| STN021.A | 159 | 247 | 0.64 | |400  |400 |

### Denmark Samples

* Note that the number of Heterozygous SNPs )(ie. low quality variants) is proportional to the mean coverage in Danish samples. It is unclear why.

!MultiQC Heterozygosity Denmark.png

* All the Denmark samples have higher counts of homozygous sites and similar distributions to the heterozygous sites.
* All samples are categorized as GOOD.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample | Mean Depth (X) | Homo | Het | Homo / Het | Graph |
| D51 | 9.2 | 132 | 32 | 4.13 | |400  |400 |
| D62 | 3.9 | 59 | 15 | 3.93 | |400  |400 |
| D71 | 23.0 | 119 | 39 | 3.05 | |400  |400 |
| D72 | 6.1 | 121 | 17 | 7.12 | |400  |400 |
| D75 | 17.7 | 158 | 31 | 5.10 | |400  |400 |
| P187 | 5.2 | 112 | 31 | 3.61 | |400  |400 |
| P212 | 7.1 | 112 | 35 | 3.20 | |400  |400 |
| P387 | 6.9 | 110 | 36 | 3.06 | |400  |400 |
| R36 | 24.2 | 115 | 45 | 2.55 | |400  |400 |

## Methods

### Variant Calling (Pairwise)

* Note\*: This is pseudo code extracted from an automated pipeline.

snippy \  
 --prefix SAMPLE \  
 --reference GCA\_000009065.1\_ASM906v1\_genomic.fna \  
 --outdir SAMPLE \  
 --bam SAMPLE.bam \  
 --mapqual 30 \  
 --mincov 3 \  
 --minfrac 0.9 \  
 --basequal 20 \  
 --force \  
 --cpus 10 \  
 --report 2> SAMPLE.log; \

* Multiqc was run on the output directories of Snippy for all samples.

### Plot Site Distributions

head -n `awk 'END{print NR - 1}' results/snippy\_multi/all/snippy-multi.txt` results/snippy\_multi/all/snippy-multi.txt | tail -n+2 | cut -f 1 | while read sample;  
do  
 in\_vcf=`ls results/snippy\_pairwise/\*/$sample/${sample}.raw.vcf`;  
 homo=${in\_vcf%%.\*}.homo.txt;  
 het=${in\_vcf%%.\*}.het.txt;  
 echo $sample;  
 bcftools query -i 'TYPE="snp" & GT="1/1" & QUAL>=100' -f '%DP\n' $in\_vcf | sort -h > $homo;  
 bcftools query -i 'TYPE="snp" & GT="0/1" & QUAL>=100' -f '%DP\n' $in\_vcf | sort -h > $het;  
 /home/poinarlab/Projects/Plague/Denmark/scripts/plot\_homo\_het.py \  
 --homo $homo \  
 --het $het;  
done  
  
mkdir results/heterozygosity  
mv results/snippy\_pairwise/{sra,local}/\*/\*.jpg results/heterozygosity/