**Metadata extraction**

*Already-published genomes*

The purpose of this section is to describe the metadata information I extracted from already-published yellow or Anubis baboon genomes. There are 6 studies that have published whole-genome data:

* Wall et al. 2016
  + 44 individuals of yellow, Anubis and mixed ancestry, low coverage.
  + Additionally, they generated a new genome assembly for yellow baboon (Pcyn\_1.0).
* Snyder-Mackler et al. 2016
  + 70 individuals in total, but only 62 were used (AMB53-AMB58 were not used, even though they are uploaded to ENA). In the study, they use fecal samples for 54 Amboseli + 8 Guinea baboons, and two additional blood-derived samples for two of the Amboseli individuals.
* Rogers et al. 2019
  + 16 individuals were used in the study, but more sequences are uploaded to ENA: 4 anubis (8 in ENA), 2 cynocephalus (14), 2 papio (2), 2 hamadryas (2), 3 kindae (15), 2 ursinus (2), 1 gelada (1).
  + Additionally, they generated a new genome assembly for Anubis baboon (Panu\_3.0).
* Robinson et al. 2019
  + There is data for 884 individuals, of which only 100 are used in the publication, and 31 are colony founders.
* Vilgalys et al. 2022
  + 430 individuals of yellow or mixed ancestry, mostly low-coverage.
* Sørensen et al. 2023 (as part of Kuderna et al. 2023)
  + 217 baboons from different species (including 92 Anubis and 60 yellow) + 1 gelada.

The metadata for all of these samples was downloaded from ENA and is stored here:

/mnt/primevo/work/iker\_rivas\_gonzalez/01\_baboon\_reference\_panel/metadata

All of the FASTQ files were downloaded from ENA, md5sum-checked, and stored here:

/mnt/primevo/shared\_data/sequencing\_data/baboon\_other\_studies

A first assessment of the downloaded FASTQ files was conducted using “FastQC”.

*Newly-generated Amboseli genomes*

In addition, there are new blood or tissue-derived Amboseli samples for 143 individuals (tentatively, pre-QC), 69 of which have already been sequenced in Vilgalys et al. 2022. Note that one individual (WAJ) has been sequenced in two of the flow cells. These samples are stored here:

/mnt/primevo/shared\_data/sequencing\_data/novaseq\_baboons

*Reference genome*

As a reference, Panubis1.0 is used. This was downloaded from NCBI using “datasets”, and it is stored under:

/mnt/primevo/work/iker\_rivas\_gonzalez/01\_baboon\_reference\_panel/data/reference\_data/newref/GCF\_008728515.1\_Panubis1.0\_genomic.fna.gz

Panu\_3.0 was also downloaded and can be found in the same directory.

**Mapping pipeline**

Reads were mapped following the latest GATK recommendations.

1. FASTQ files were converted to uBAM files using “picard FastqToSam”.
2. Illumina adapters were marked using “picard MarkIlluminaAdapters”.
3. Reads were mapped to reference using a mapping pipe:
   1. “picard SamToFastq”
   2. “bwa mem”
   3. “picard MergeBamAlignment”
4. Coordinates were sorted using “picard SortSam”.
5. All of the mapped reads from different runs but from the same individual were merged using “samtools merge”.
6. Duplicates were removed using “picard MarkDuplicates”.
7. Merged BAMs were modified to replace read names by sample names using “samtools view”.

As part of routine QC, the total read count and the number of mapped reads were computed after steps E and G. Additionally, the average coverage is calculated after step G.

The mapping procedure is unified with a gwf workflow.

**Variant calling**

After mapping, joint calling was performed also following GATK guidelines. From the final merged BAMs after step G in the mapping pipeline, gVCFs were generated per chromosome using “HaplotypeCaller”. Afterwards, gVCFs were combined using “GenomicsDBImport”, and individuals were joint called using “GenotypeGVCFs”. The joint calling pipeline is also unified with a gwf workflow.

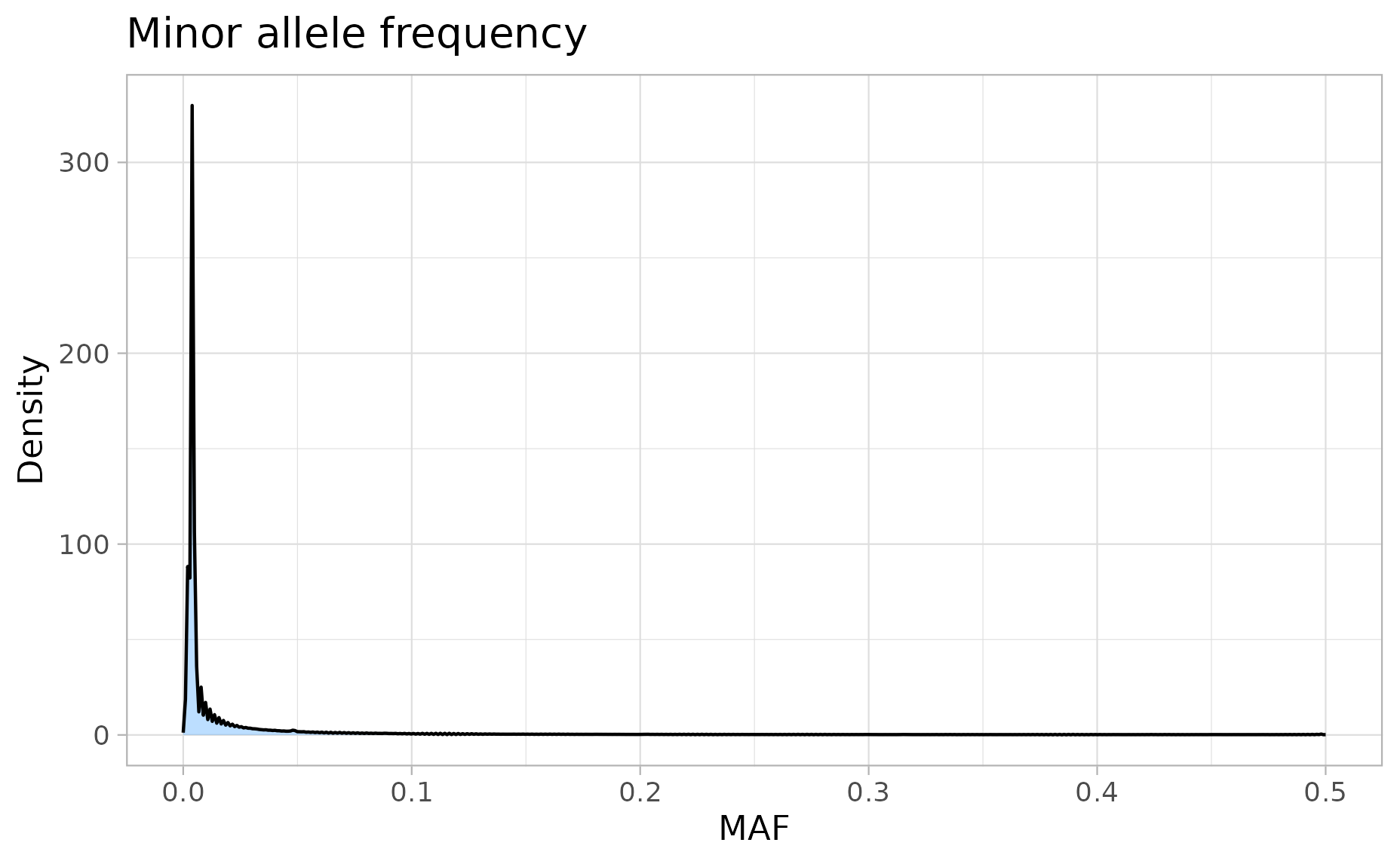
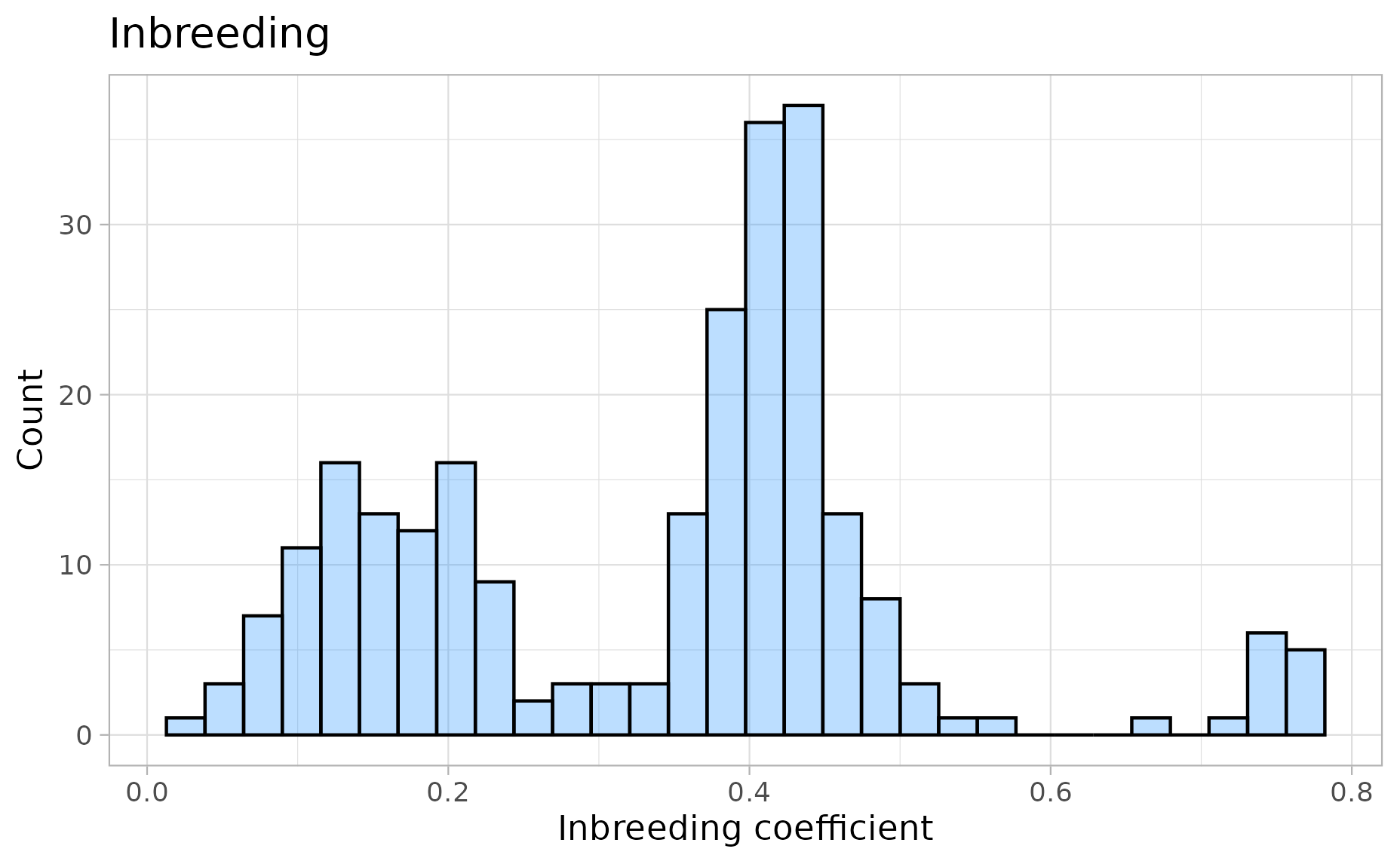
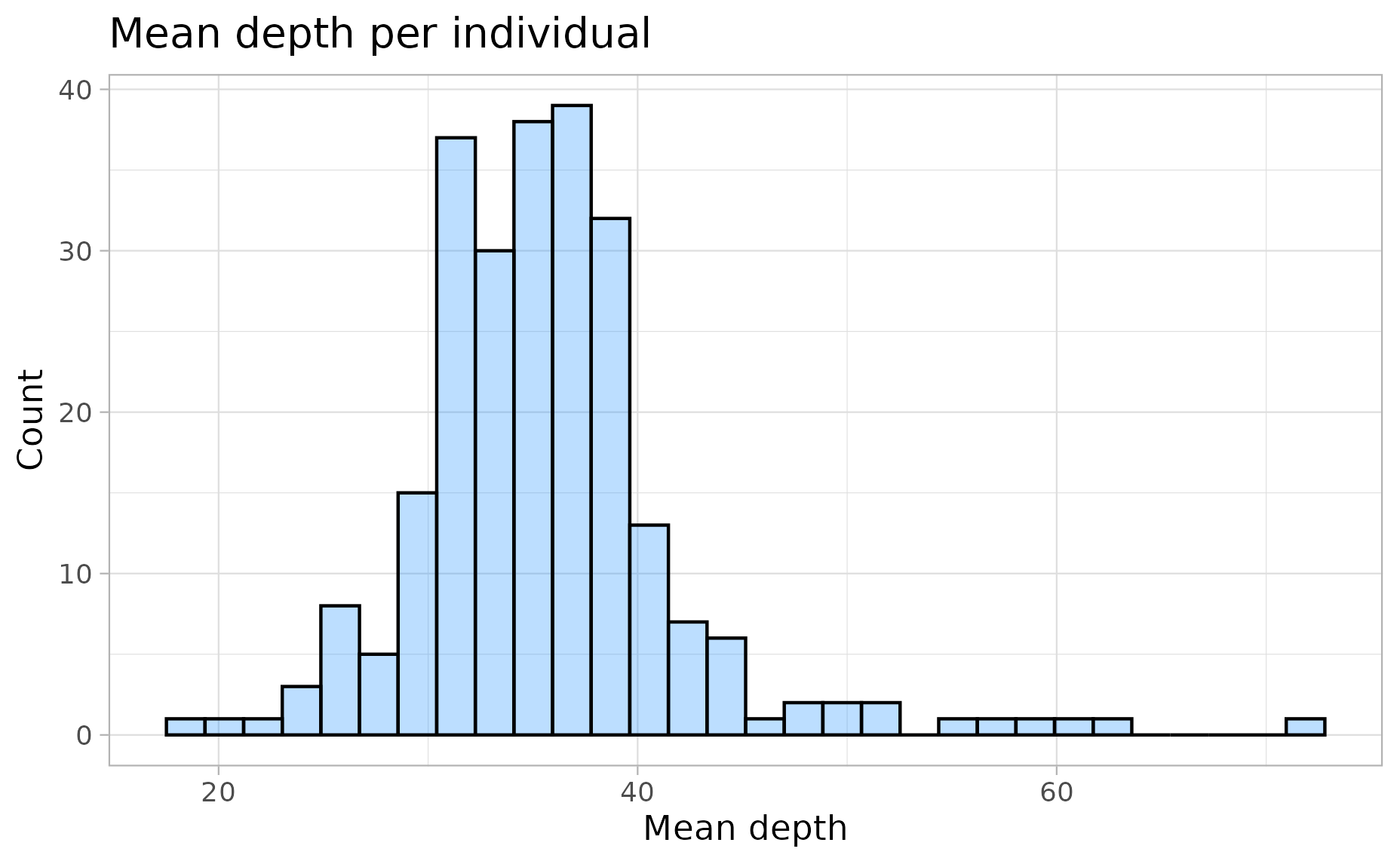
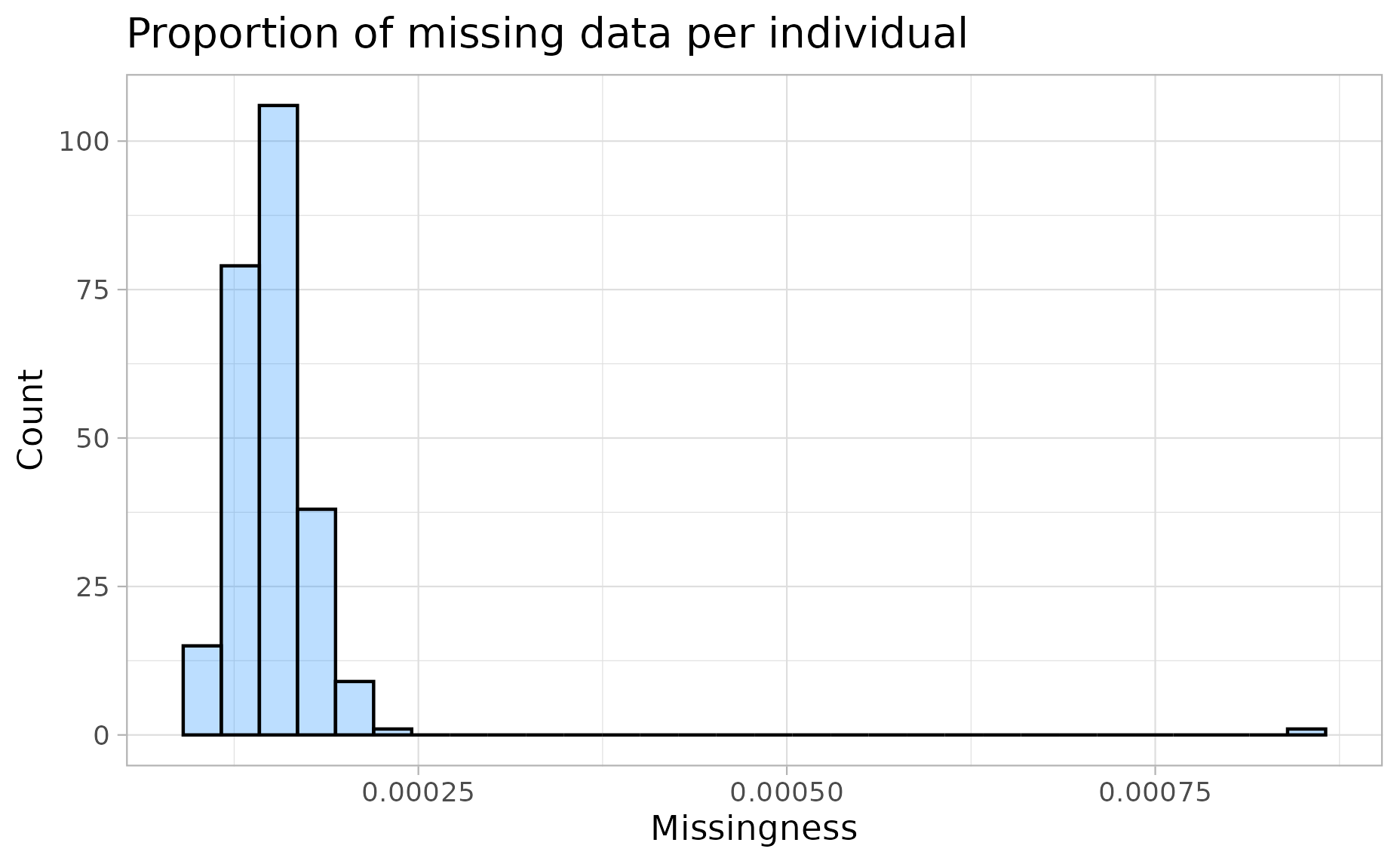
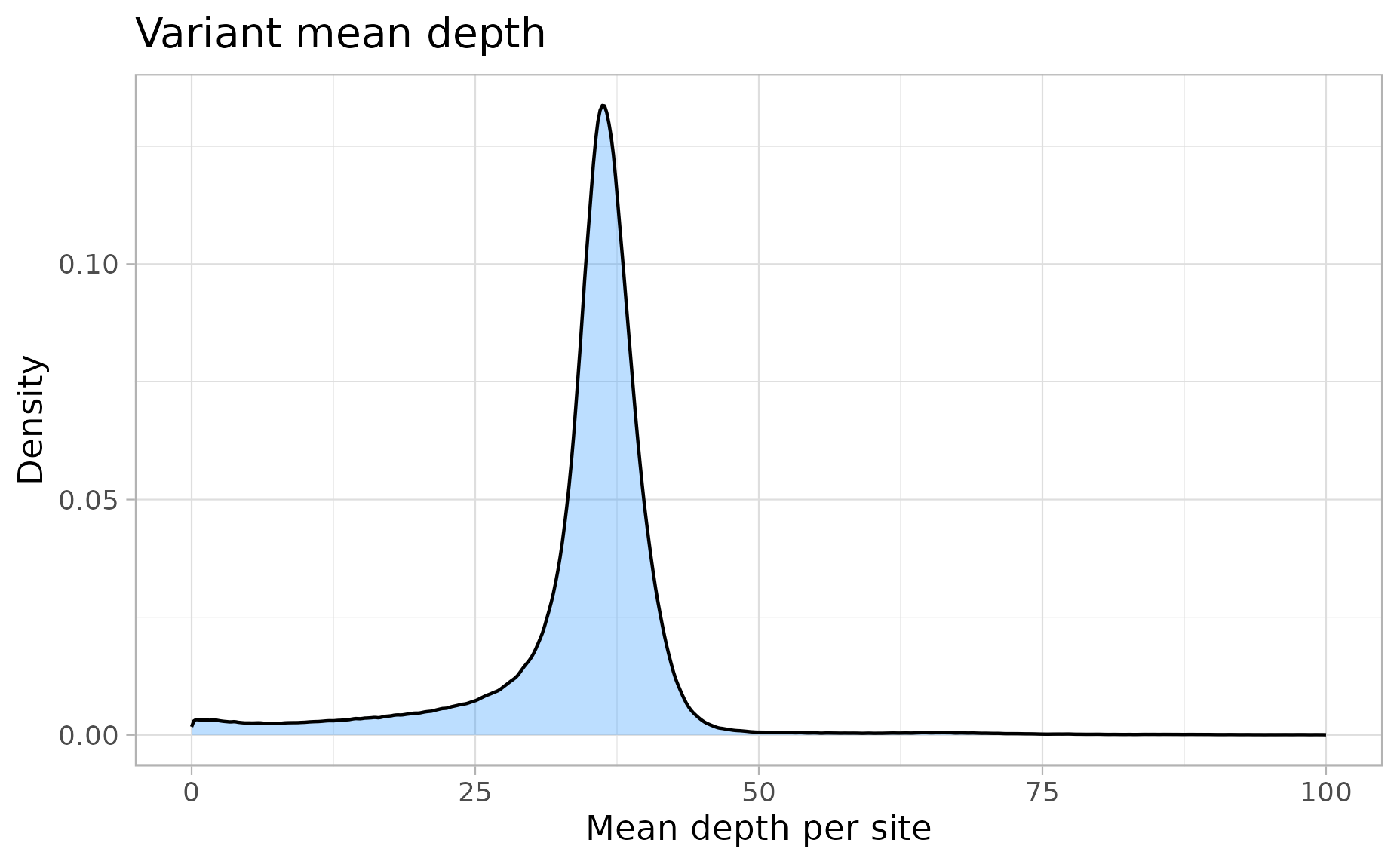
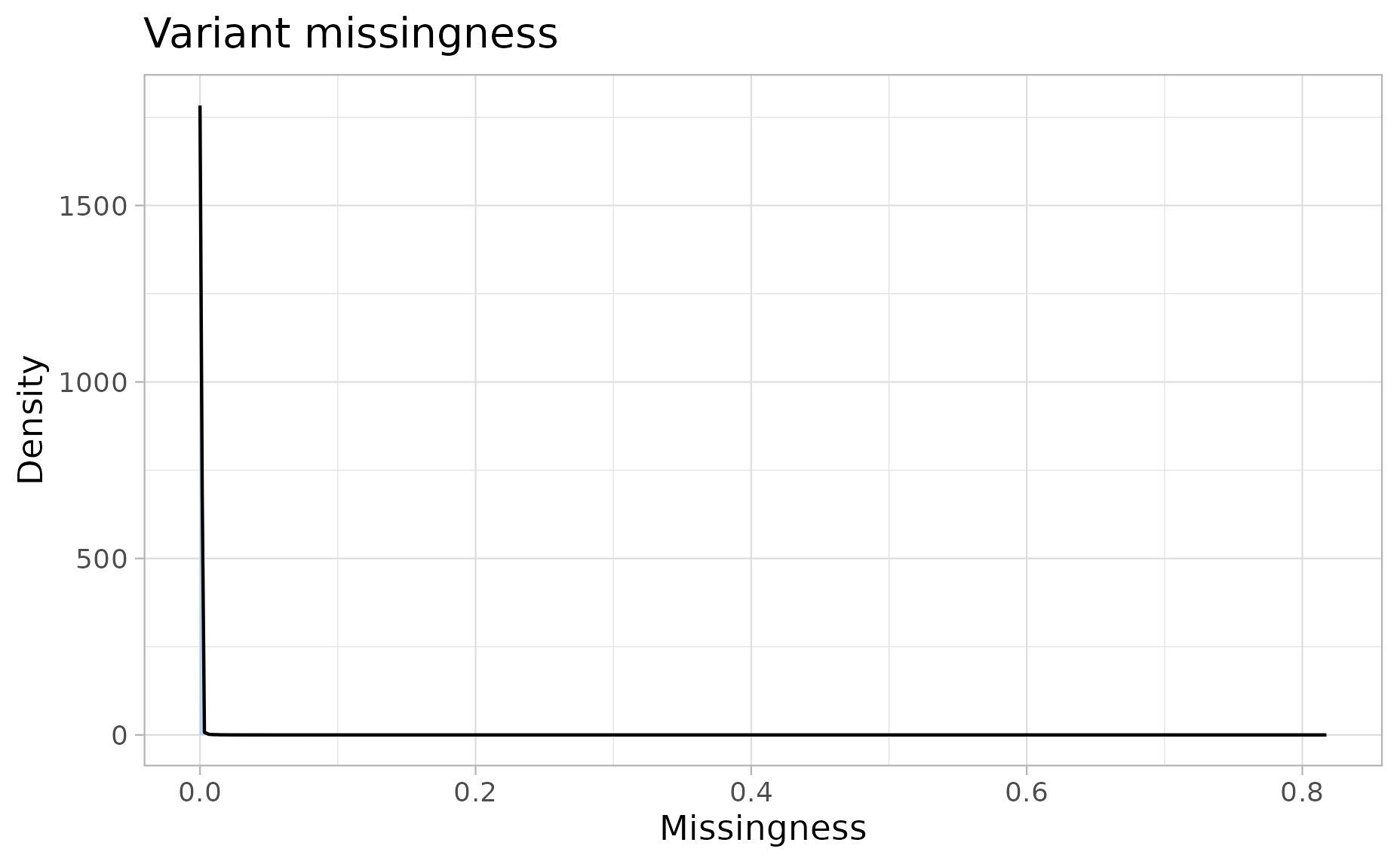
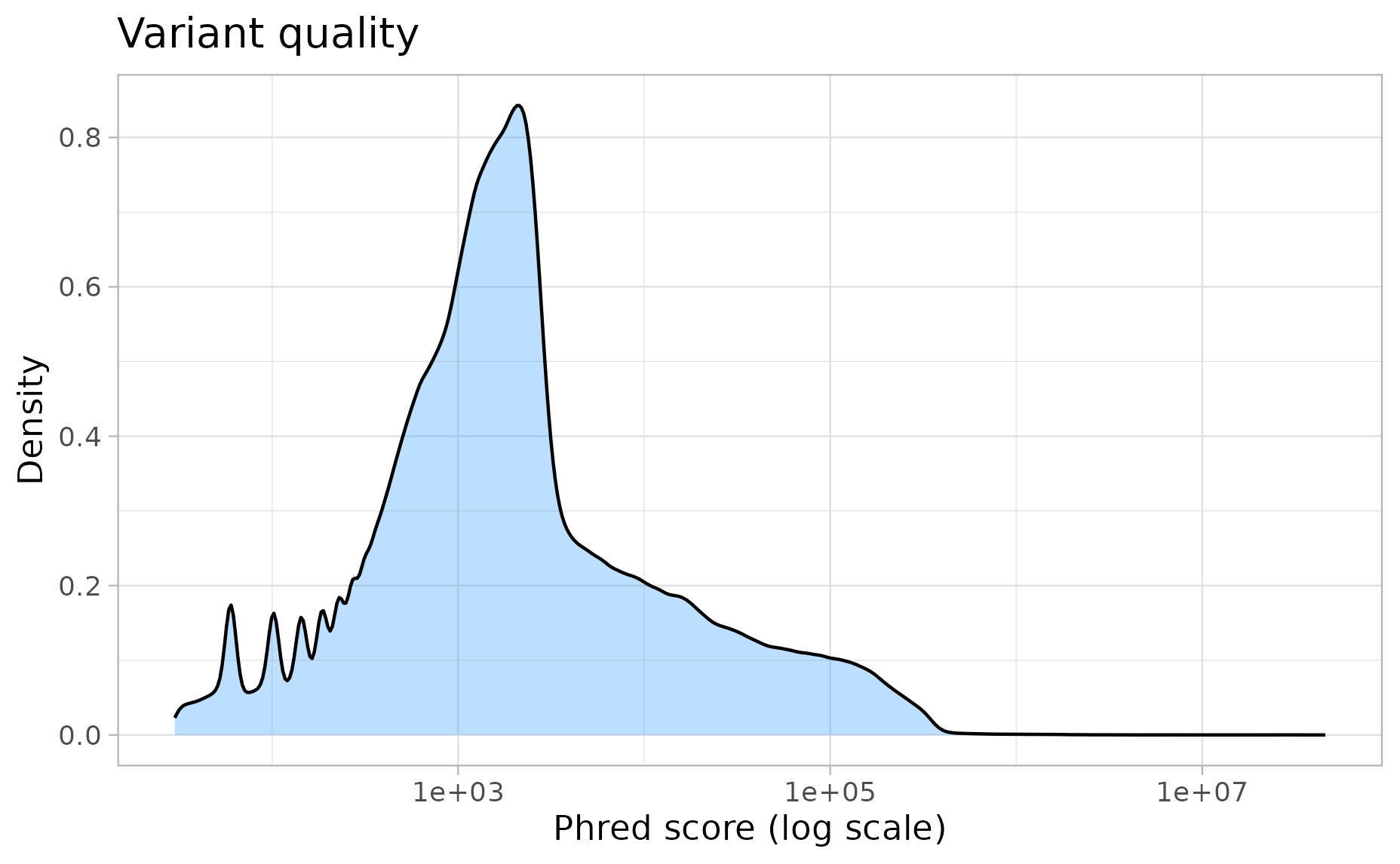
**Quality control**

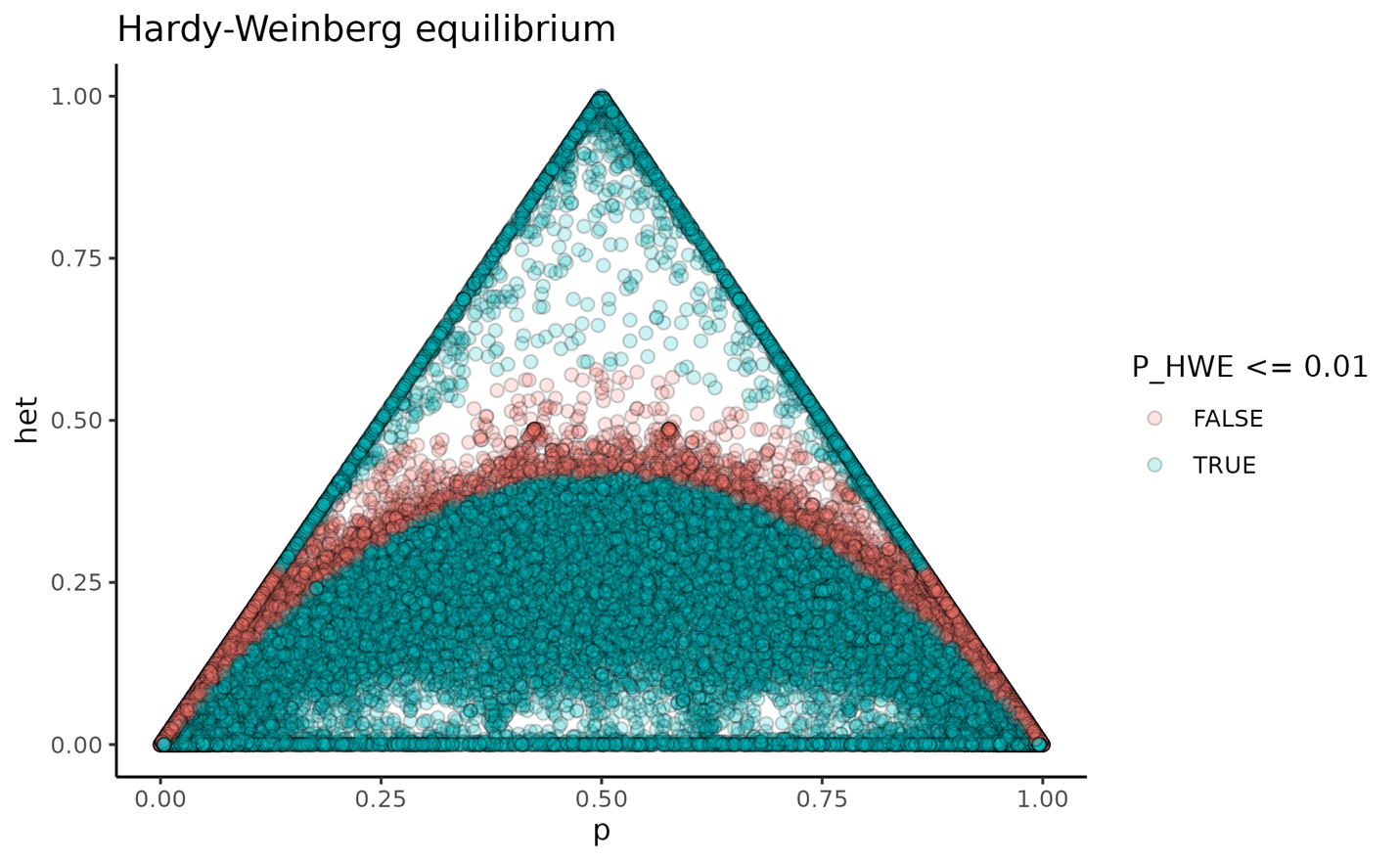
QC was performed on the final VCFs using “vcftools”. More specifically:

* Allele frequency per site was calculated using “freq2”.
* Average depth per individual was calculated using “depth”.
* Average depth per site was calculated using “site-mean-depth”.
* Quality score per site was calculated using “site-quality”.
* Proportion of missing data per individual was calculated using “missing-indiv”.
* Proportion of missing data per site was calculated using “missing-site”.
* Heterozygosity and inbreeding coefficient were calculated per individual using “het”.
* Hardy-Weinberg equilibrium was calculated using “hardy”.

Additionally, a PCA was run using “plink”.

*Robinson et al. 2019 (founders) + Sørensen et al. 2023 (all)*





*Amboseli (Vilgalys et al. 2022 + newly-generated)*

