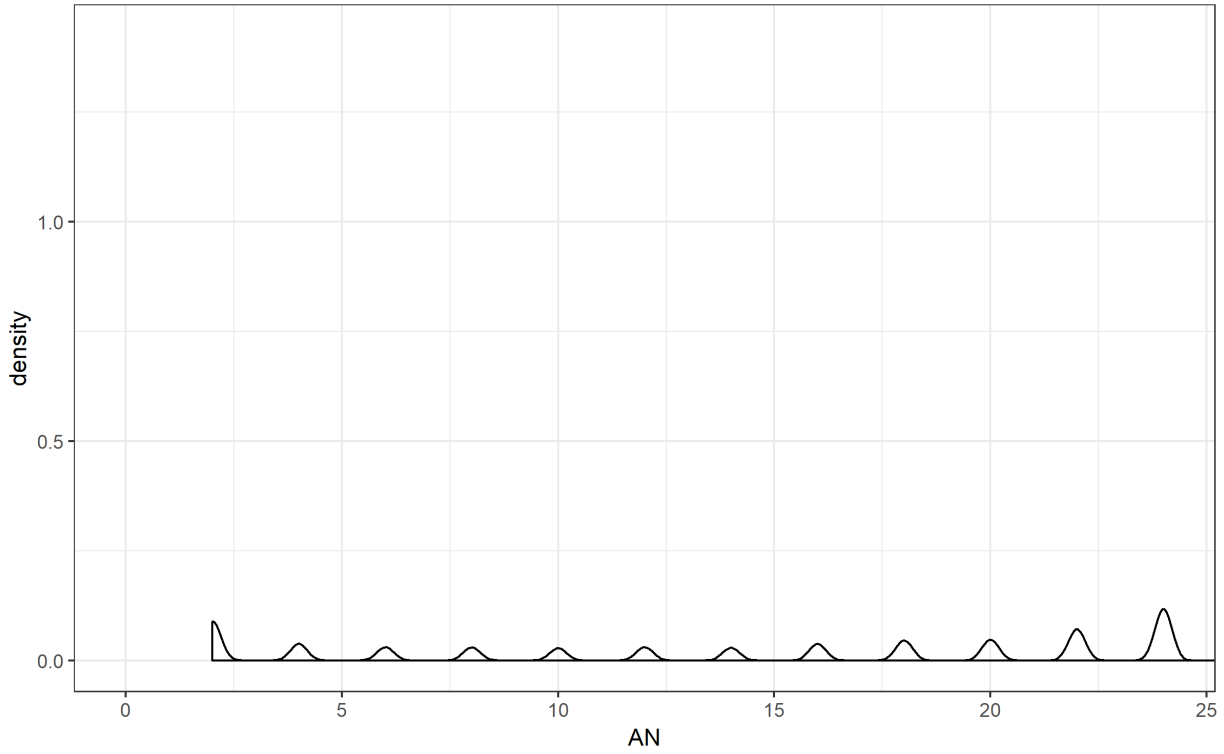
Hard-filter of Cayo exome data on chrX

Number of alleles that are genotyped AN

- There are 14 exomes → the maximum AN is 28, which means that a variant is genotyped across all 14 exomes



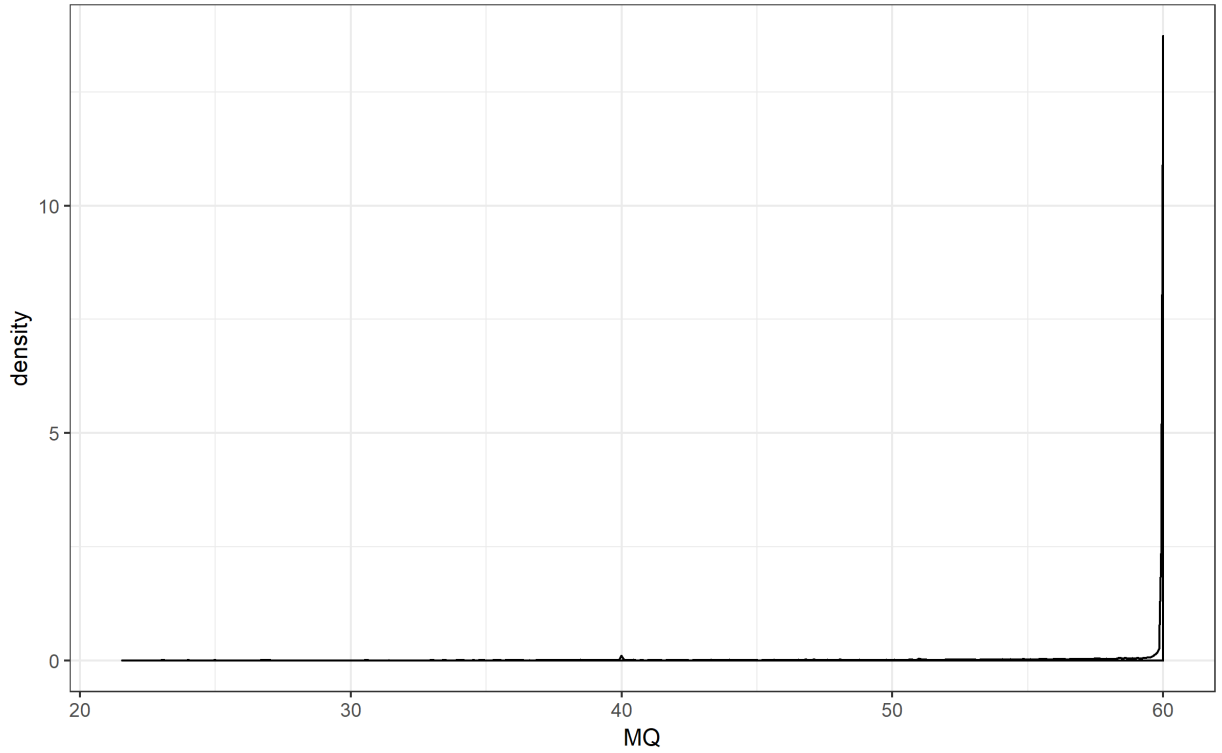
- Should we hard-filter based on AN?
- For example, we only keep variants where AN >= 6 (genotyped in at least 3 exomes → 25%).
  - There are 14,794 out of 15,794 (~94%) variants with AN >= 6

Investigating AN

		Number of variants
AN >= 2	Called in 1 (out of 14) exomes or more	15,794
AN >= 4	Called in 2 (out of 14) exomes or more	15,091
AN >= 6	Called in 3 (out of 14) exomes or more	14,794
AN >= 8	Called in 4 (out of 14) exomes or more	14,556
AN >= 10	Called in 5 (out of 14) exomes or more	14,323
AN >= 12	Called in 6 (out of 14) exomes or more	14,100
AN >= 14	Called in 7 (out of 14) exomes or more	13,862
AN >= 16	Called in 8 (out of 14) exomes or more	13,637
AN >= 18	Called in 9 (out of 14) exomes or more	13,337
AN >= 20	Called in 10 (out of 14) exomes or more	12,980
AN >= 22	Called in 11 (out of 14) exomes or more	12,609
AN >= 24	Called in 12 (out of 14) exomes or more	12,047
AN >= 26	Called in 13 (out of 14) exomes or more	11,131
AN = 28	Called in 14 (out of 14) exomes	0

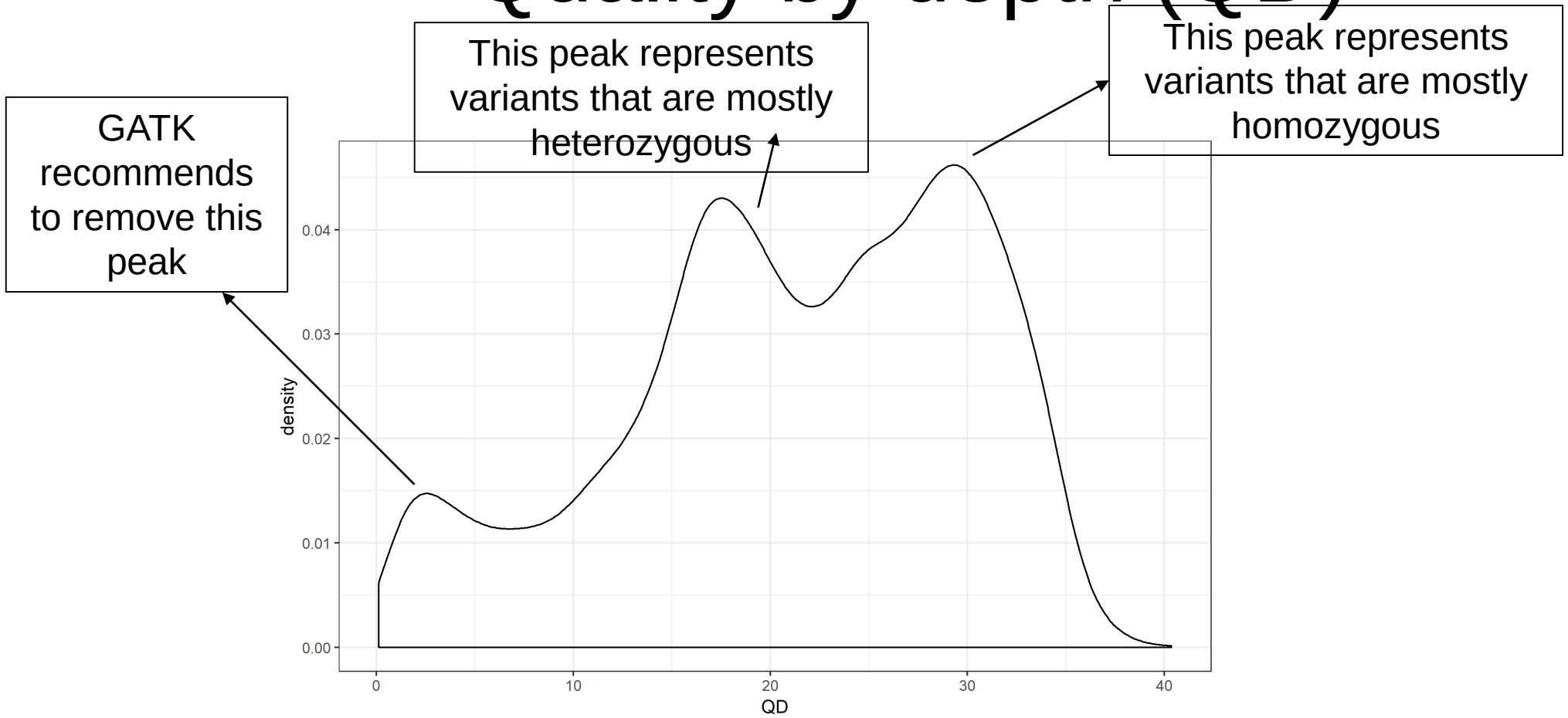
RMSMapping quality (MQ)

- From GATK: “This is the root mean square mapping quality over all the reads at the site. Instead of the average mapping quality of the site, this annotation gives the square root of the average of the squares of the mapping qualities at the site.” and “When the mapping qualities are good at a site, the MQ will be around 60.”



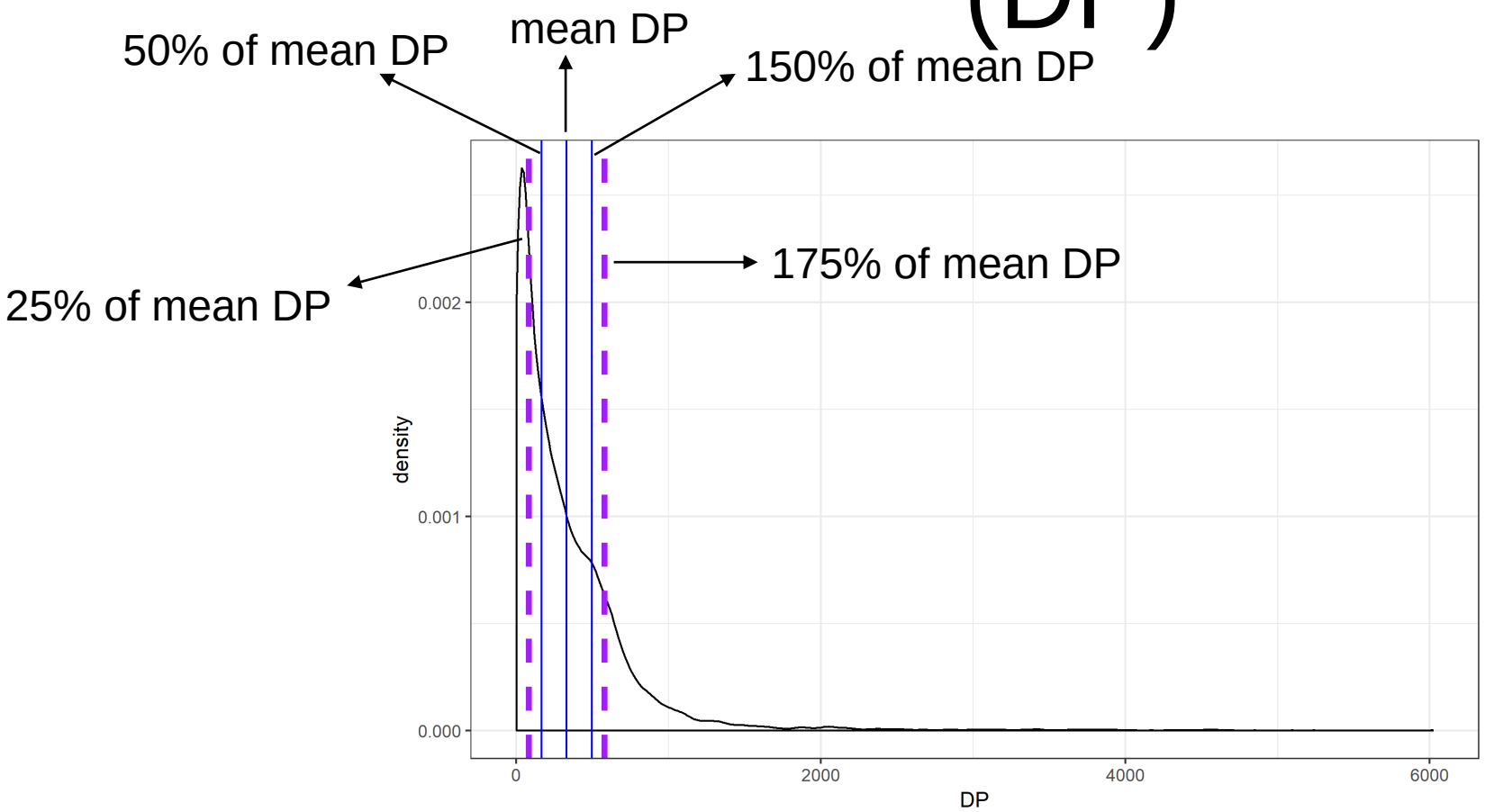
- GATK’s hard-filter suggests to remove variants with MQ < 40
- 15,261 out of 15,794 variants with MQ > 40

Quality by depth (QD)



- Use QD threshold of 5
- There are 14,721 out of 15,794 variants with QD > 5

Total depth of coverage over all sample (DP)



There are 5,457 out of 15,794 variants (~34%) whose DP is between 50% and 150% of the mean.
Is this too strict?

There are 8,699 out of 15,794 variants (~55%) whose DP is between 25% and 175% of the mean.
Do you think that this threshold is more reasonable?

Investigating DP

- Peak = 44
- Median = 215

	Number of variants (out of 15,794 total)
DP > 500X	3,488
DP > 600X	2,405
DP > 750X	1,387
DP > 1000X	701

Investigating DP

	Number of variants
0-25X	2,053
25-50X	1,244
50-100X	1,677
100-150X	1,410
150-200X	1,186
200-250X	1,008
250-300X	910
300-350X	804
350-400X	705
400-450X	646
450-500X	648
500-1000X	2,800
>1000X	703