Analysis of TERT promoter mutations in UK Biobank Whole Genomes

William Dunn

2025-01-09

Data exploration and cleaning

Here, we explore the data generated by performing pileup at each position along the *TERT* promoter. This is the data that was generated by Sruthi Cheloor Kovilakam by running Samtools mpileup across the *TERT* promoter on WGS data in the UK Biobank.

We begin by exploring the dimensions of the table of all of these pileup calls:

```
## [1] 4688015      10
```

There are 4.68 million rows, that is, 4.68 million alternate allele calls (for 489,548 individuals - i.e. almost all individuals in the UKB). This means that across the region of interest (201bp long), each individual has on average 9-10 alternate allele calls - much too high to be true passenger/driver mutations, and already strongly suggestive of a lot of sequencing noise/error in the unfiltered calls.

We next examine the range of depths to see if low depth regions have been filtered - here we survey the deciles of the depth across all positions:

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100% 
##    2   22   26   29   31   34   37   40   43   49  242 
```

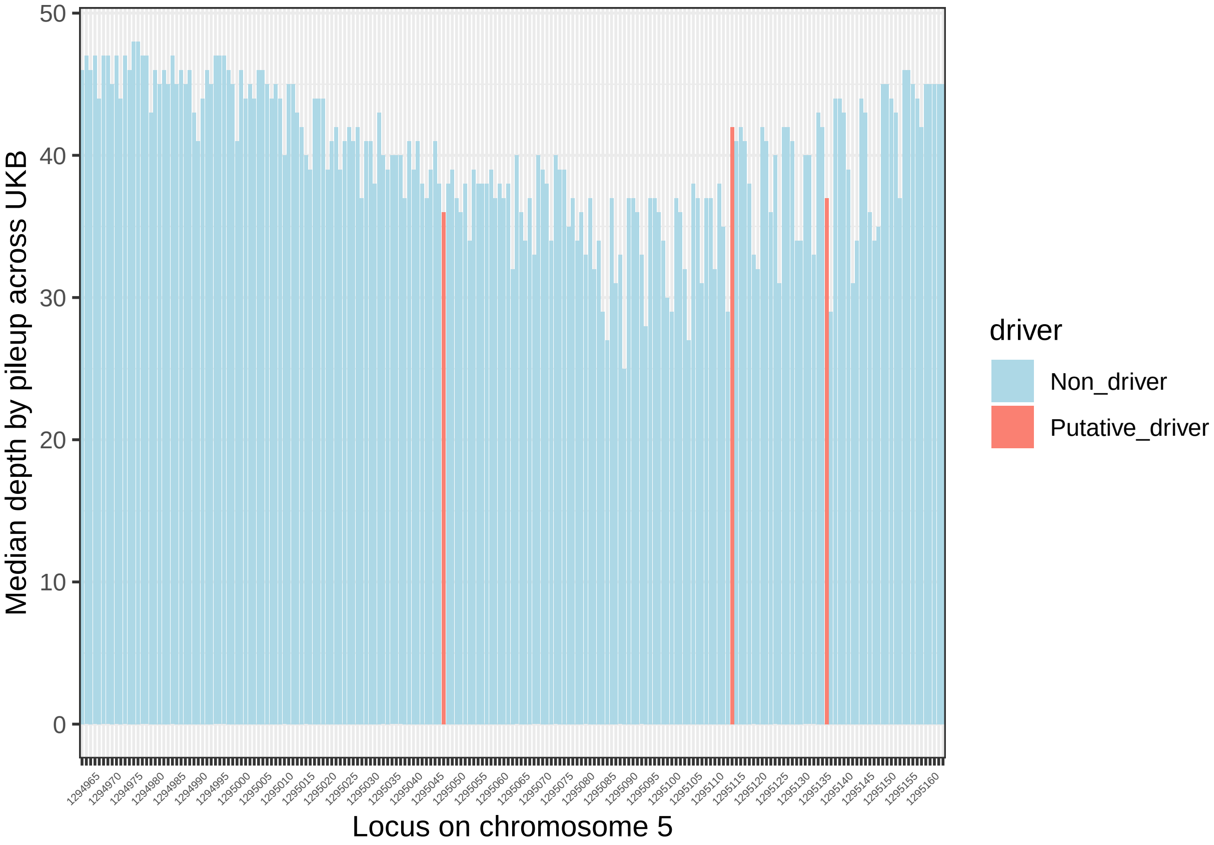
We can see that the depth across all position ranges from 2-242 reads, though most sites have 22-49 reads, and the median depth is 34 reads.

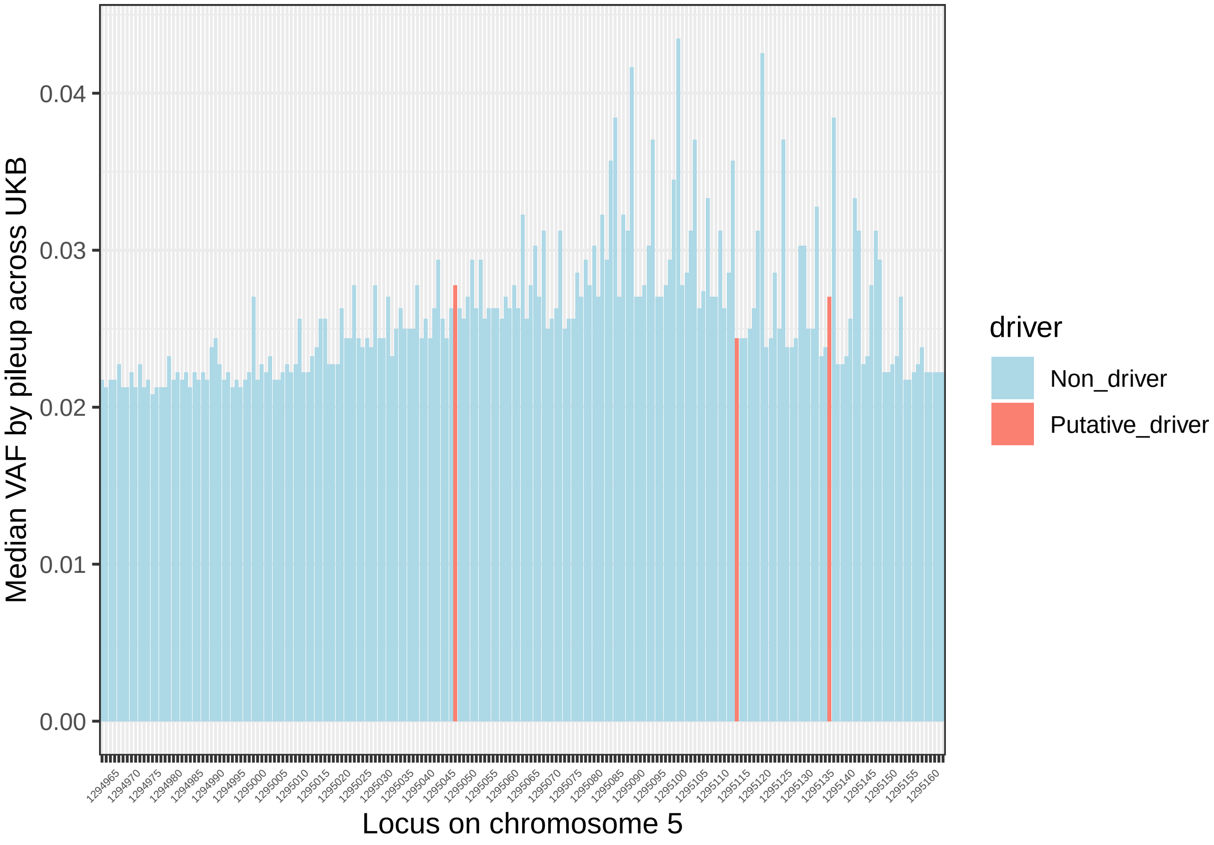
Similarly, we can examine the deciles for the variant allele fraction (VAF):

```
##         0%        10%        20%        30%        40%        50%        60% 
##  0.4132231  2.0833333  2.3809524  2.6315789  2.8571429  3.1250000  3.4482759 
##        70%        80%        90%       100% 
##  3.8461538  4.5454545  5.8823529 72.2222222 
```

We can see that the median VAF is 3.1%, and that most alternate alleles are at low VAF (<6%), however, there are a few sites with very high VAF that might be more consistent with germline variants.

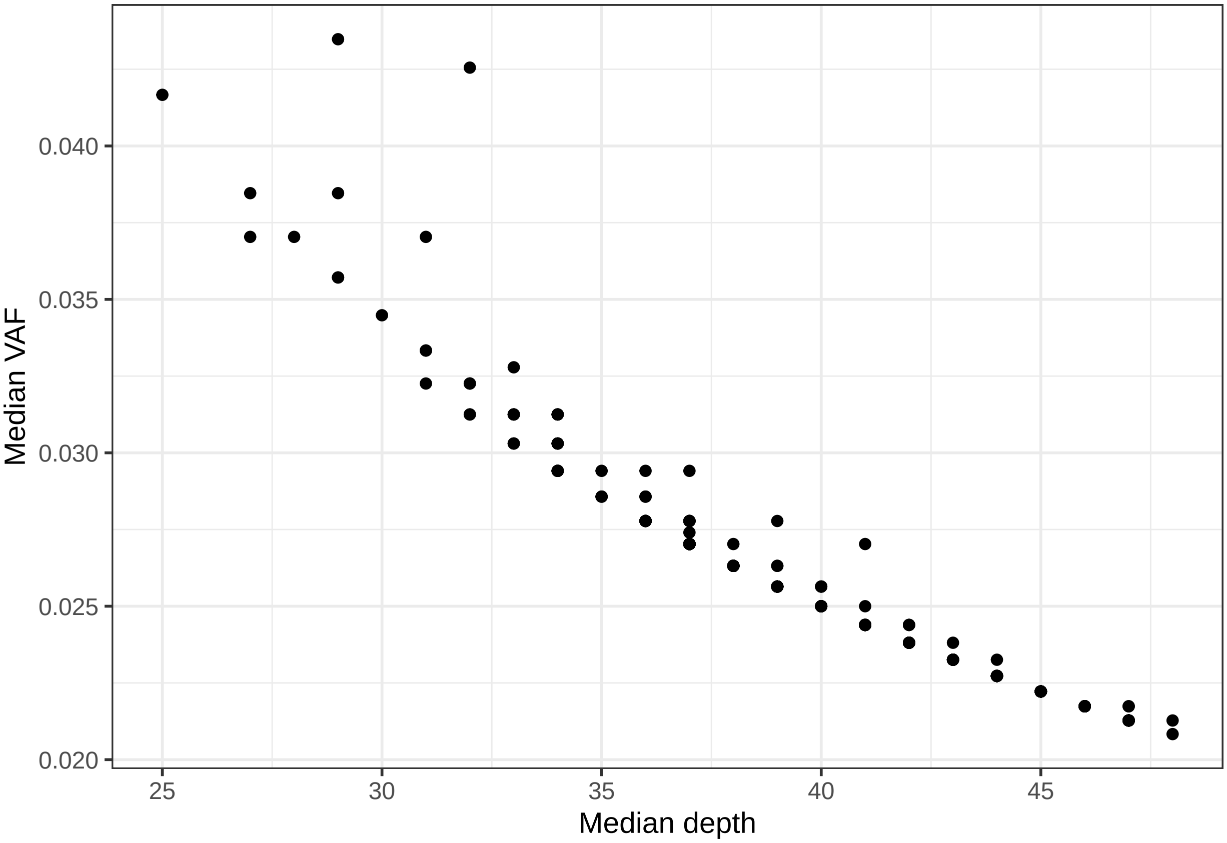
We can visualise depth and VAF across the *TERT* promoter - here, we colour by whether or not the position is one of 3 sites that we propose to be driver mutations based on *a priori* knowledge of the positions of somatic rescue mutations in telomere biology disorders[1] (chr5:1295046:T:G, chr5:1295113:G:A, and chr5:1295135:G:A):



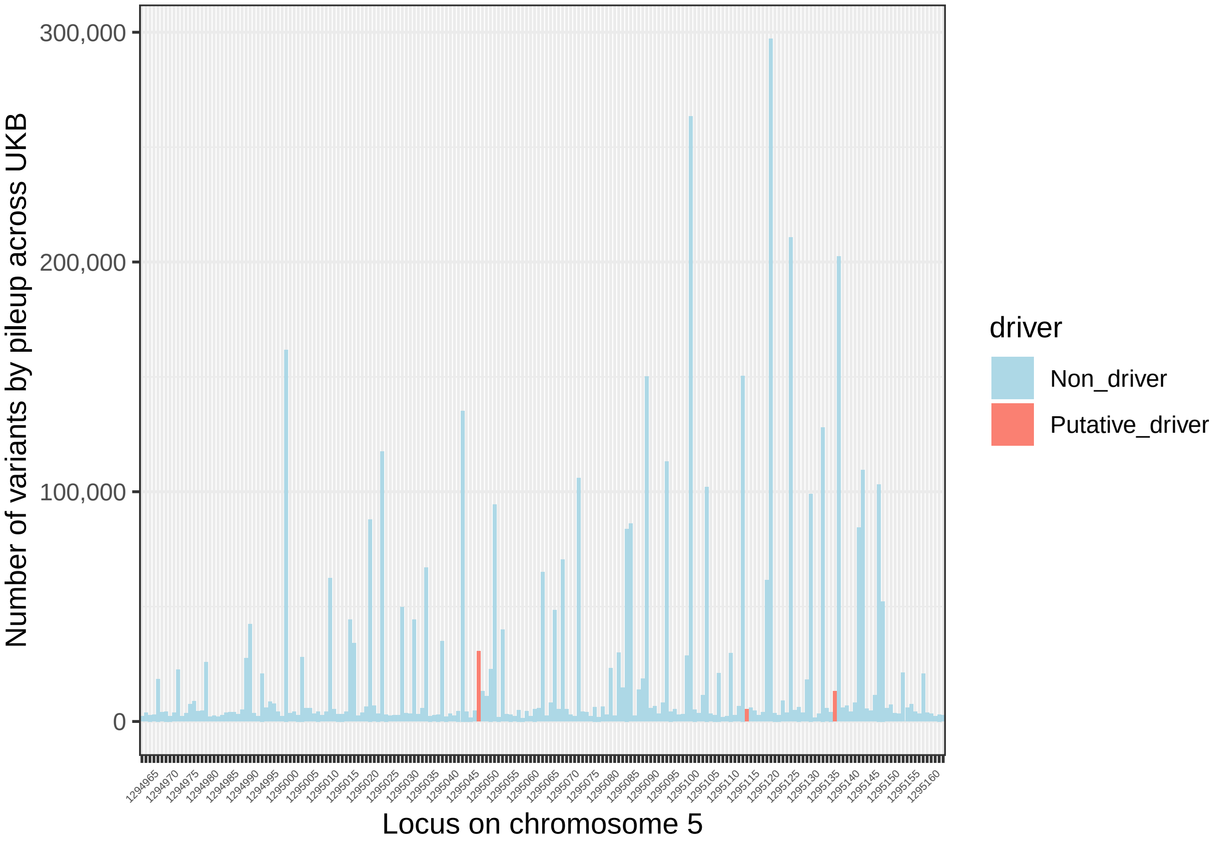


As expected, we see an inverse relationship between the median depth at each position and the median

VAF of the alternate allele calls. We can more easily visualise this here by directly plotting median VAF vs median depth:



Finally, we can examine the number of variants at each position:



We can see that at some positions, there are huge numbers of alternate allele calls (that is, at least 1 alternate allele detected) - with one site having almost 300,000 variant calls across the UKB. Our three sites of interest seem to have a relatively low number of alternate allele calls, although one of the sites (chr5:1295046) has a higher number, suggesting this site might be more prone to error than the others.

We now start applying some fairly liberal arbitrary filters to the dataset to try and remove some of the noise and error and make the data more intelligible. We begin by applying a threshold of a minimum of 3 alternate reads, and also filtering sites with low depth (<15 reads) or a high VAF that might be suggestive of germline contamination (VAF >30%). Finally, where an individual has more than one different alternate allele call at the same position, we consider only the variant with the highest number of alternate alleles (e.g. if an individual has reads for both C>T and C>A at position X with 5 and 4 reads respectively, we retain the C>T call and discard the C>A call). What does this do to the number of alternate allele calls across the UKB?

```
## [1] 105455     10
```

We have now reduced the number of alternate allele calls from 4.68 million to 105,455. This is a substantial reduction - but it is still much higher than might be expected for real somatic mutations, suggesting that erroneous calls remain in the dataset.

After applying these filters, how many variants are there at each position?

```
##
## 1294966 1294971 1294974 1294975 1294978 1294986 1294988 1294989 1294992 1294993
##       7       5       2       3       7       2      14      95       6       2
## 1294994 1294995 1294996 1294998 1295002 1295004 1295009 1295010 1295011 1295014
##       5       4       1    4878      14       1      77       1       1     100
## 1295015 1295018 1295019 1295020 1295022 1295023 1295024 1295027 1295030 1295033
##      33       1     656       2    1706       1       1      92      43     262
## 1295034 1295037 1295042 1295043 1295046 1295047 1295048 1295049 1295050 1295052
##      12      27    2486       1      60       1       4      15     813      66
## 1295053 1295055 1295056 1295060 1295062 1295063 1295065 1295067 1295068 1295071
##       3       1       1       1      71       5      64     179       5    1205
## 1295075 1295079 1295081 1295083 1295084 1295086 1295087 1295088 1295089 1295090
##       2       7      25     719     223       1       8     997       2       1
## 1295093 1295094 1295095 1295096 1295098 1295099 1295100 1295102 1295103 1295106
##     737       1       4       1      33   24563       2       2     485      11
## 1295107 1295109 1295111 1295112 1295113 1295114 1295118 1295119 1295121 1295122
##       4      15       2     736      80       1     242   38357       1       1
## 1295124 1295126 1295128 1295129 1295131 1295132 1295133 1295135 1295136 1295138
##   11681       4       5    1062       5    2274       1      13    6040       1
## 1295141 1295142 1295145 1295146 1295147 1295149 1295152 1295153 1295154 1295155
##     802    1716       1    1322     195       1      11       1       2       1
## 1295157 1295161
##      11       1
```

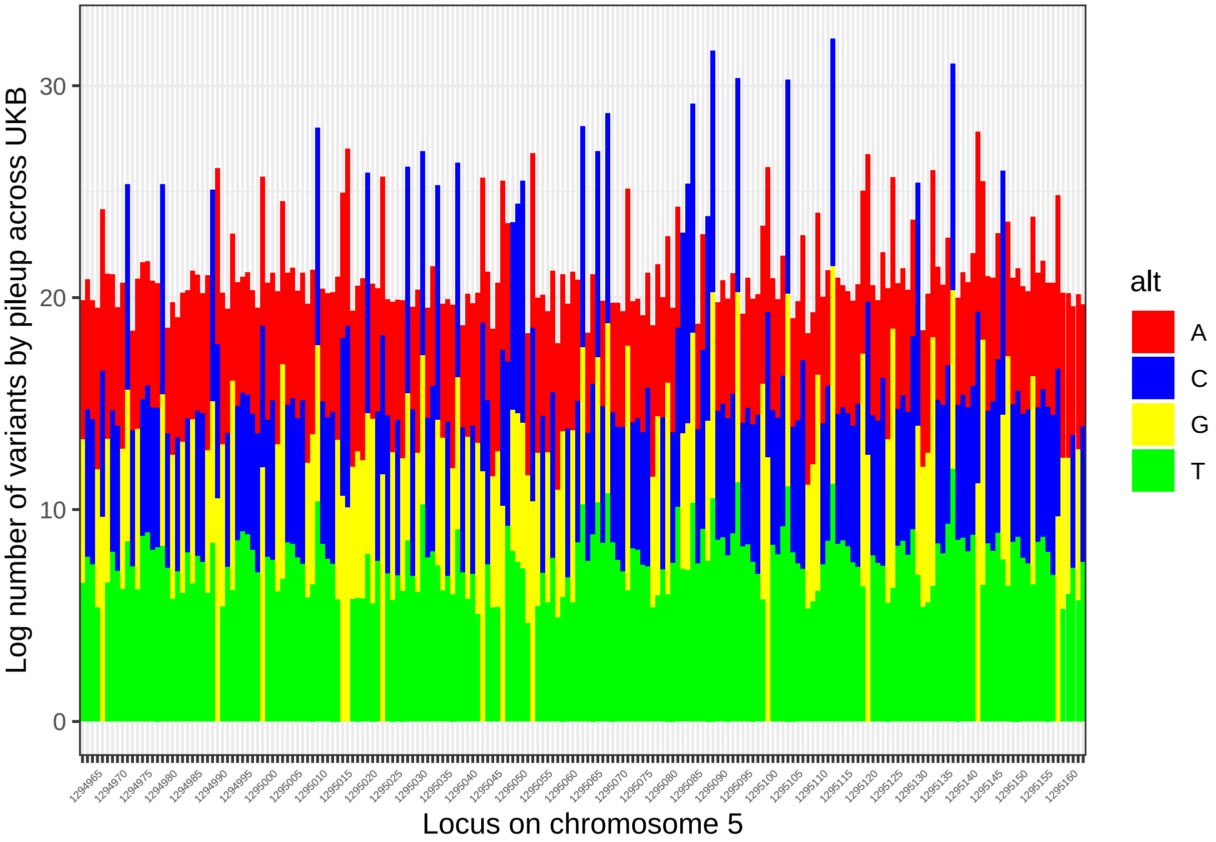
We can see that there are some sites with only a single, or sometimes a handful of alternate allele calls (which could be true passengers), others with tens of thousands, suggestive of error, and some with total numbers in between, representing a grey area where it is more difficult to classify variants as driver, passenger or error.

Specifically, after applying these initial filters, how many variants are there amongst our three putative drivers, and what is the distribution of the alternate allele calls (the base changes are driver-specific: **chr5:1295046:T:G - NM_198253.3:c.-57A>C**, **chr5:1295113:G:A - NM_198253.3:c.-124C>T**, and **chr5:1295135:G:A - NM_198253.3:c.-146C>T**).

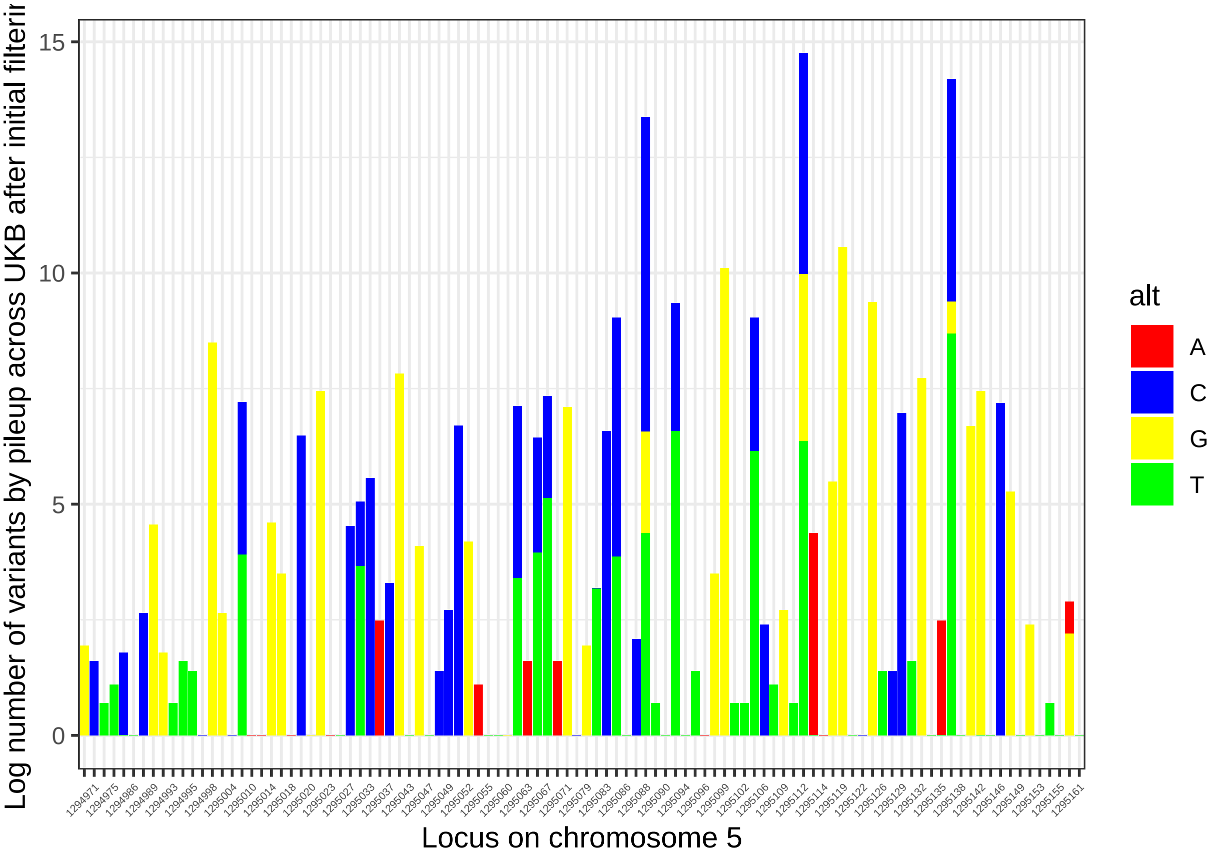
```
##
##            A  G  T
##   1295046  0 60  0
##   1295113 79  0  1
##   1295135 12  0  1
```

At these sites, almost all the variants represent the expected driver mutation seen in telomere biology disorders. We can contrast this with the rest of the variants and try to visualise it with a stacked bar chart. Note: the y axis is on a log scale so that we can see the infrequently varying sites alongside the very frequently varying sites. This can be misleading as relatively modest differences in the proportions on log scale can reflect large differences on a linear scale.

Firstly, we examine the data before applying our filters (such as choosing the maximum at sites with more than one alternate allele call):



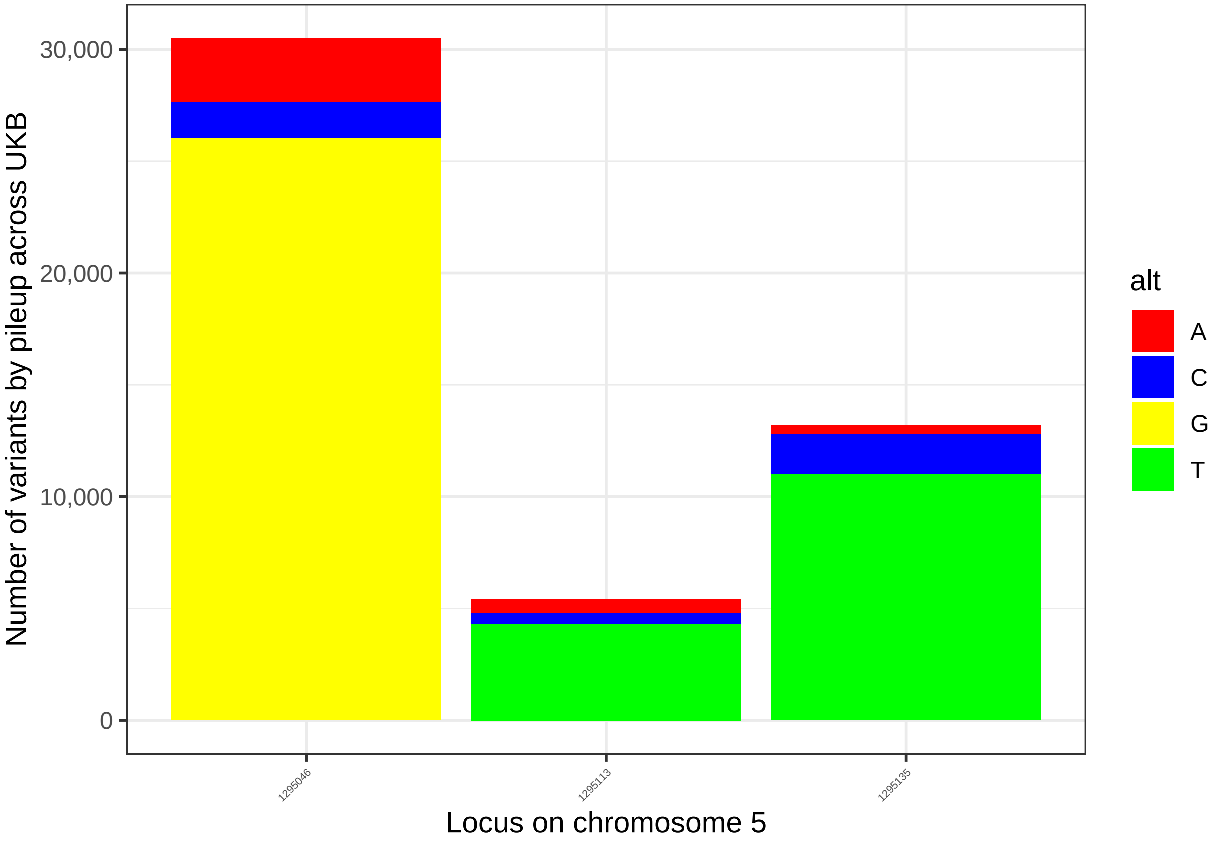
Next, we examine the data after filtering:



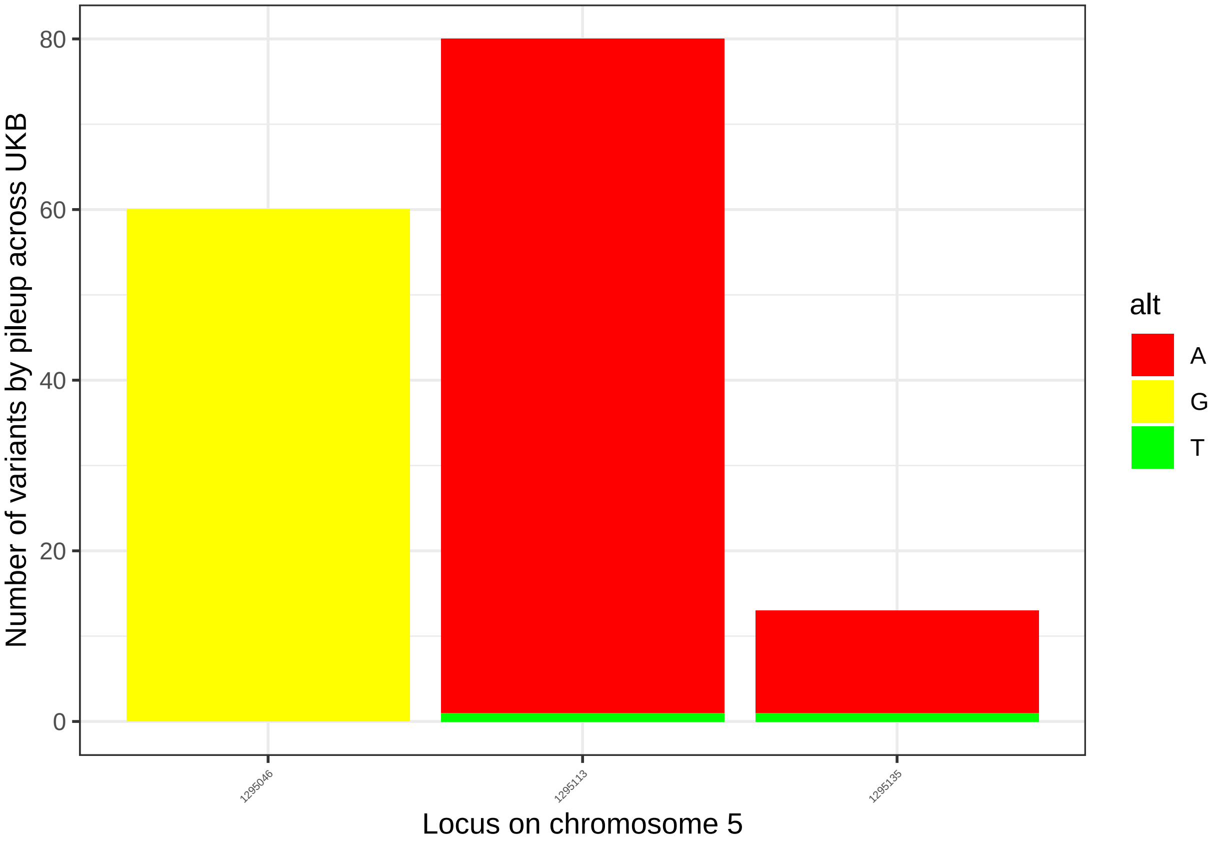
There are many sites where just a single alternate allele dominates after filtering by the minimum number of alternate reads and taking the alternative allele with the highest number of reads.

We now focus on our three sites of interest and the proportion of alternate allele calls before and after filtering. Because the numbers are more comparable here, we now use a linear scale on the y axis.

Before filtering:



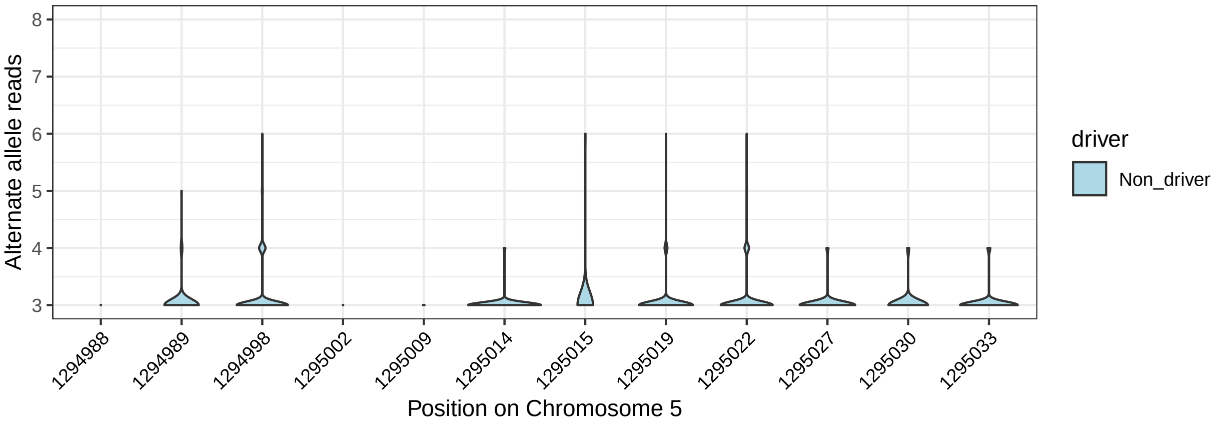
Note the very large number of calls before filtering. After filtering, the number of calls is substantially reduced:

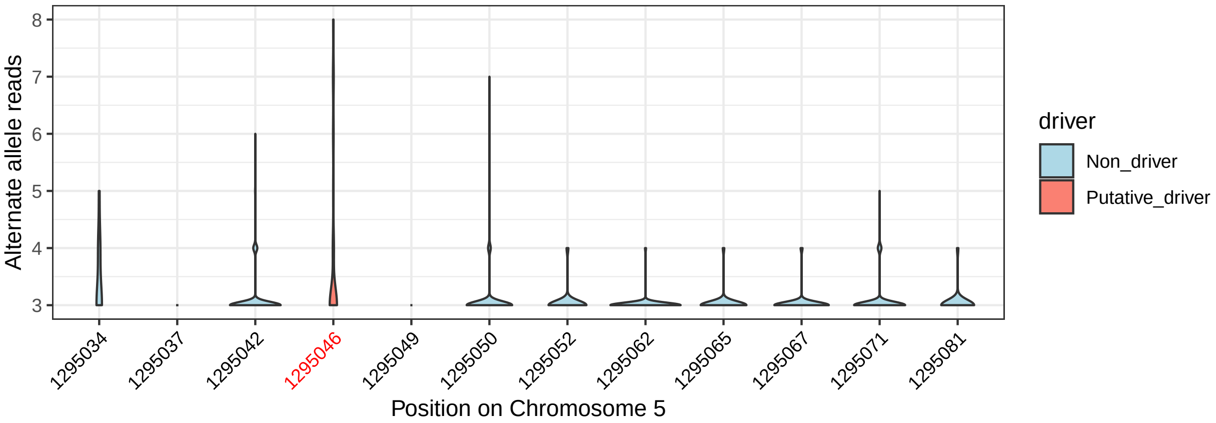


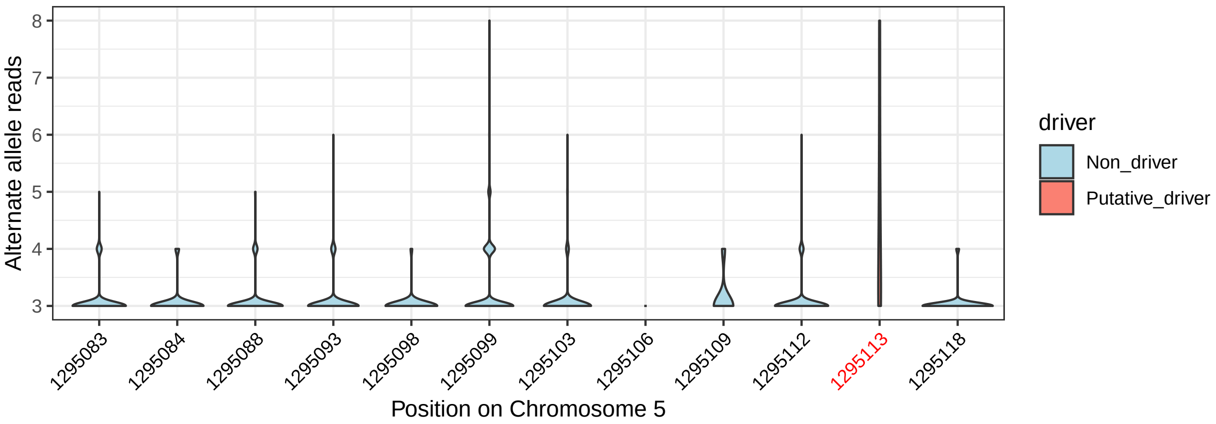
By applying our filters across the *TERT* promoter, we have enriched for what we believe to be “real” drivers at our sites of interest. It is worth noting again though, that at position chr5:1295046, unlike the other two positions, the “real” driver comprised the majority of alternate allele reads prior to filtering, so once again, we should pay special attention to this site, as it may be more likely than the other two positions to be contaminated by sequencing error.

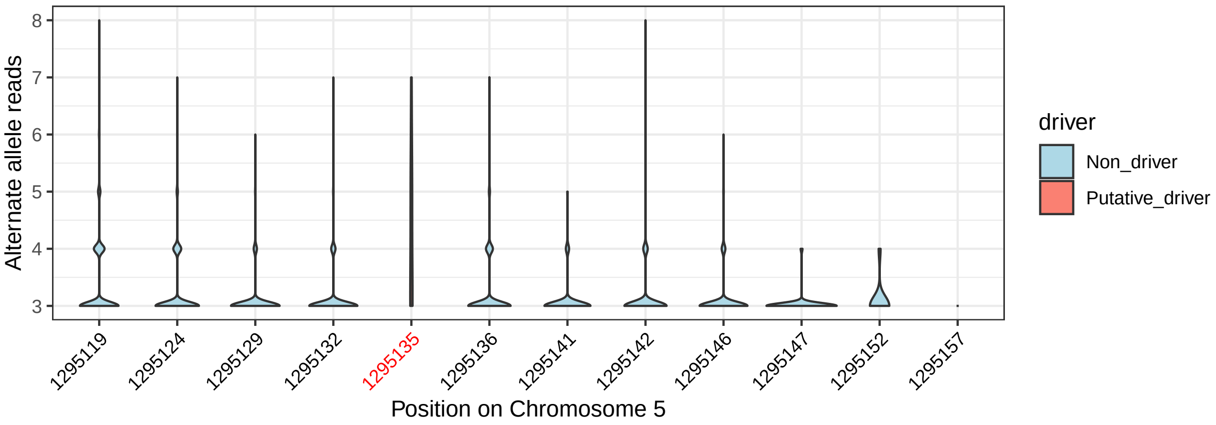
Note that as a final filter at this stage, we have removed the two variants at our sites of interest that don’t have the expected alternate allele.

We will now move on to examining the distribution of the number and VAF of reads at each position in our filtered dataset. As we saw earlier, there are lots of sites with only a handful of alternate reads after filtering; since we cannot discern a “distribution” with so few data points, I have arbitrarily excluded those sites with <= 10 variants for this visualisation step only. As there are a lot of data points, I have elected to make violin plots as visualising individual data points makes for unintelligible graphs due to over-plotting. Here, we examine the distribution of the number of alternate reads at each site (note that at sites where all variants have 3 alternate reads, there is no distribution to plot):


Alternate allele reads
8
7
6
5
4
3
1294988
1294989
1294998
1295002
1295009
1295014
1295015
1295019
1295022
1295027
1295030
1295033
Position on Chromosome 5
driver
Non_driver


Alternate allele reads
8
7
6
5
4
3
1295034
1295037
1295042
1295046
1295049
1295050
1295052
1295062
1295065
1295067
1295071
1295081
Position on Chromosome 5
driver
Non_driver
Putative_driver


Alternate allele reads
8
7
6
5
4
3
1295083
1295084
1295088
1295093
1295098
1295099
1295103
1295106
1295109
1295112
1295113
1295118
Position on Chromosome 5
driver
Non_driver
Putative_driver


Alternate allele reads
8
7
6
5
4
3
1295119
1295124
1295129
1295132
1295135
1295136
1295141
1295142
1295146
1295147
1295152
1295157
Position on Chromosome 5
driver
Non_driver
Putative_driver

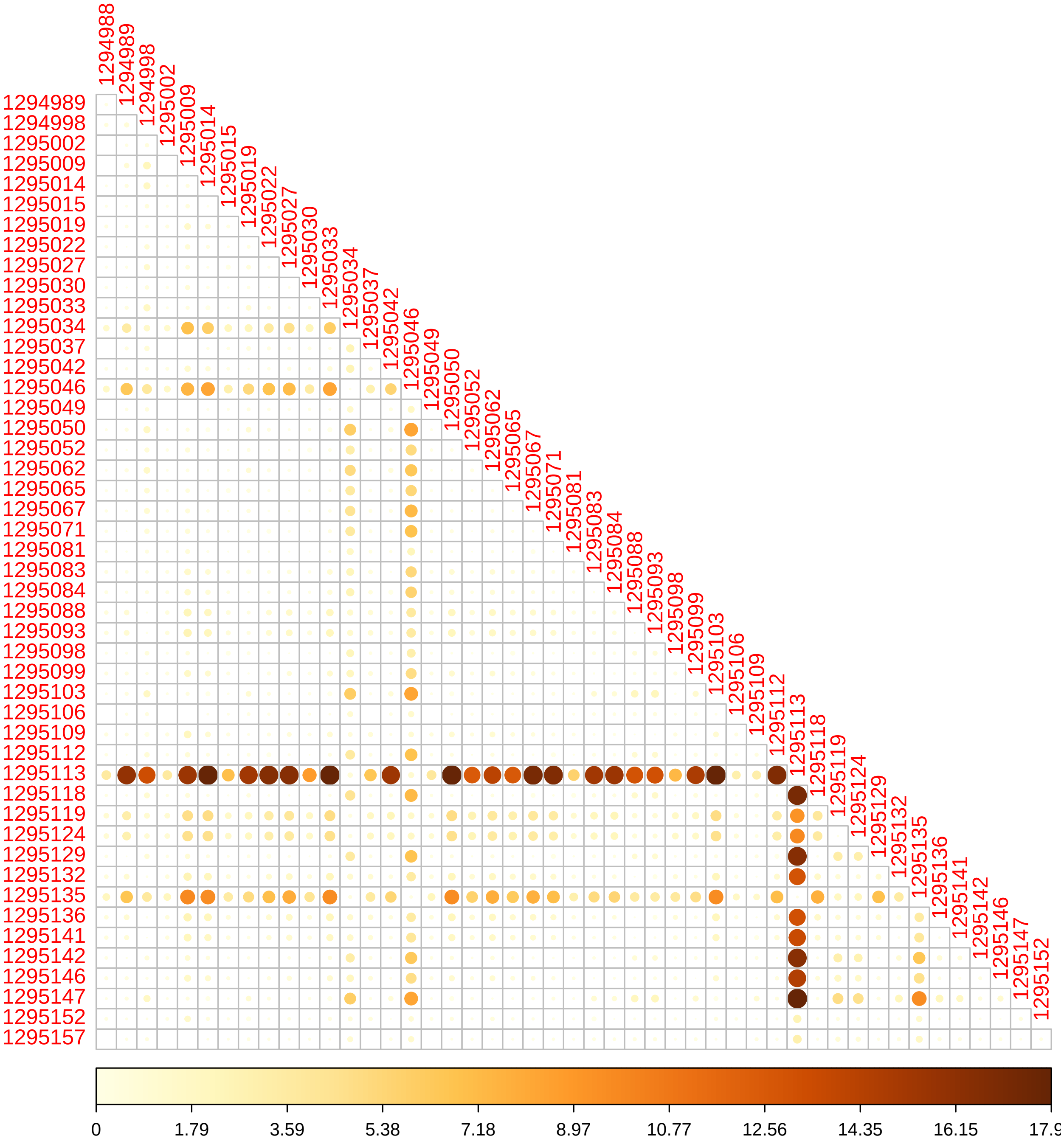
We can see that the distribution of the number of alternate reads is distinctly different at our three sites of interest, consistent with this being "real" and the other sites being largely sequencing noise. Of note, the distribution of the number of alternate reads at chr5:1295034 is very similar to our sites of interest, suggesting that this position might also be worth considering as a putative driver.

What are the alternate alleles at chr5:1295034?

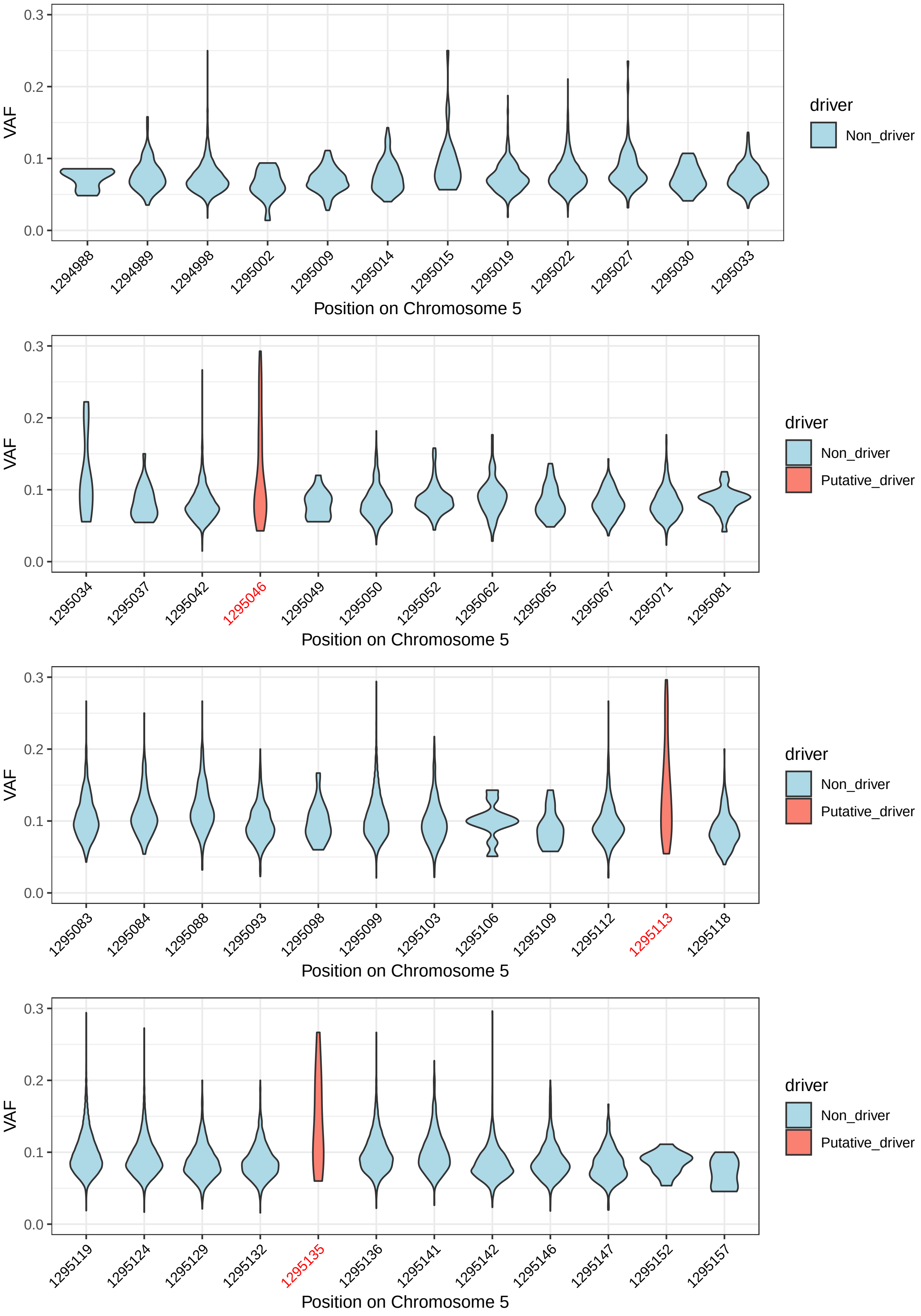
```
##
##            A
##   1295034 12
```

They are all C>A mutations, and there are 12 individuals with these mutations.

We can also visualise the differences between distributions of alternate allele reads more formally by using the Mann-Whitney U test, computing n by n comparisons, then plotting the -log10 of the adjusted p values for these n by n comparisons in a heatmap. Since the computed p value is in part a function of sample size, we take a random sample of n=100 from sites where there are >100 variants passing filters (otherwise, at noisy sites with e.g. 10,000 variants passing our liberal filters, these have very large p values driven largely by their sample size). We see that our three sites of interest have the largest p values, in addition to the fourth site that we speculate may be an unreported driver:



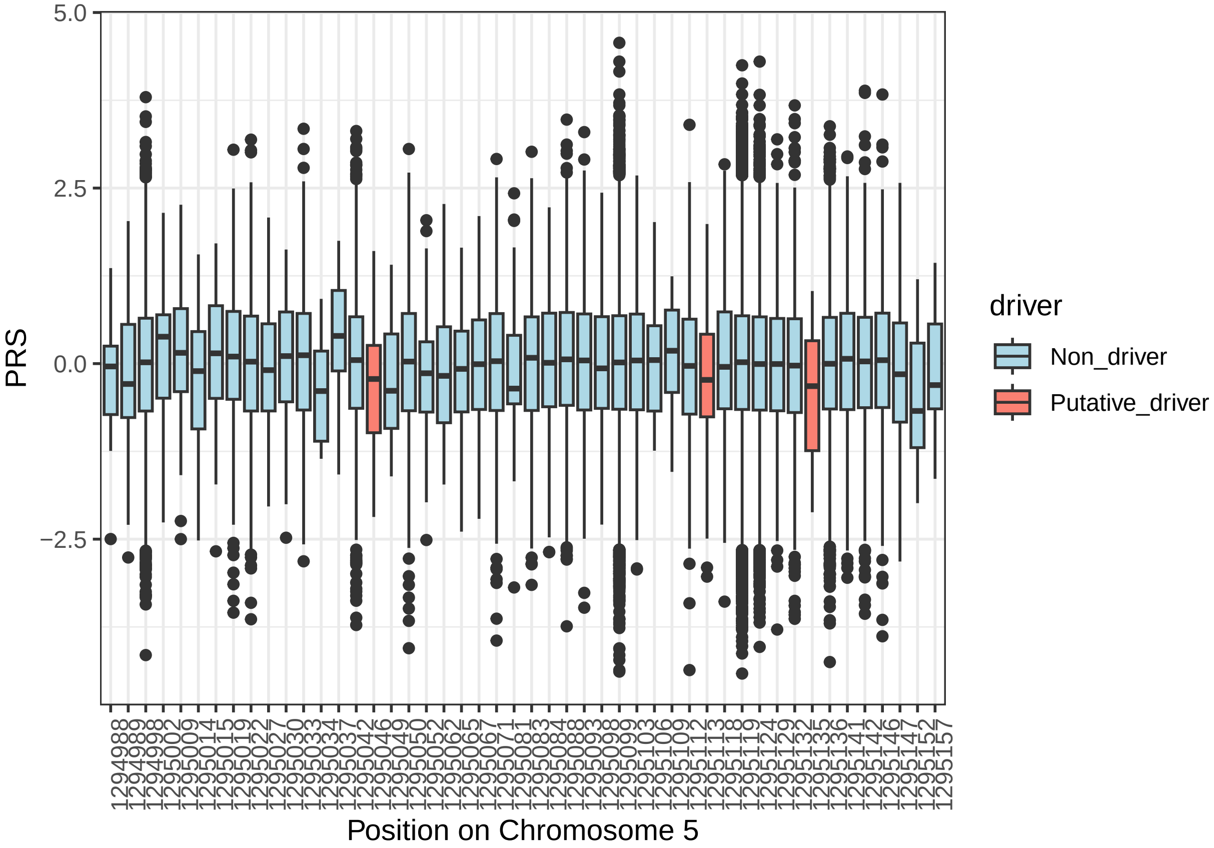
We can also visualise the distribution of VAFs across sites:


VAF
0.0
0.1
0.2
0.3
1294988
1294989
1294998
1295002
1295009
1295014
1295015
1295019
1295022
1295027
1295030
1295033
Position on Chromosome 5
driver
Non_driver
VAF
0.0
0.1
0.2
0.3
1295034
1295037
1295042
1295046
1295049
1295050
1295052
1295062
1295065
1295067
1295071
1295081
Position on Chromosome 5
driver
Non_driver
Putative_driver
VAF
0.0
0.1
0.2
0.3
1295083
1295084
1295088
1295093
1295098
1295099
1295103
1295106
1295109
1295112
1295113
1295118
Position on Chromosome 5
driver
Non_driver
Putative_driver
VAF
0.0
0.1
0.2
0.3
1295119
1295124
1295129
1295132
1295135
1295136
1295141
1295142
1295146
1295147
1295152
1295157
Position on Chromosome 5
driver
Non_driver
Putative_driver

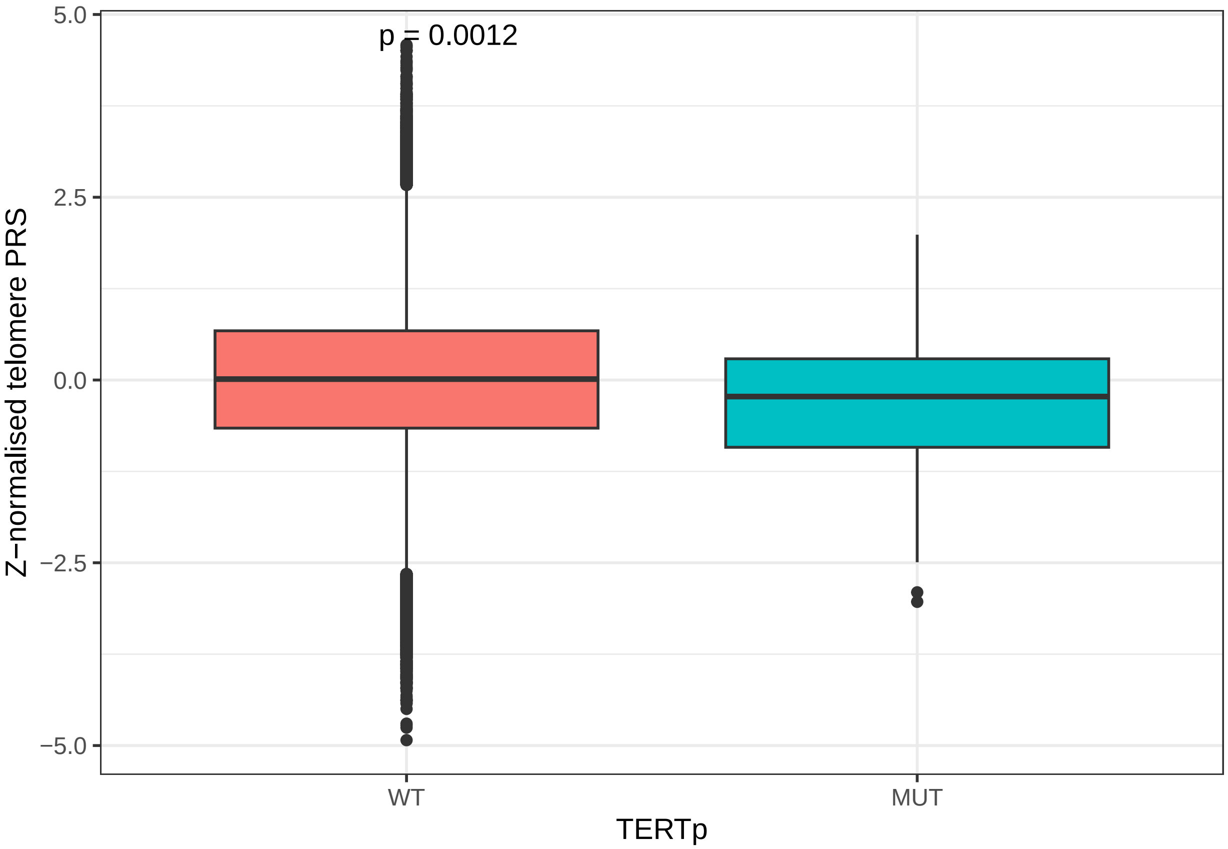
Again, we see that the distribution of VAFs are our site of interest looks distinct from the rest, and once again, the additional position of chr5:1295034 looks to have a distribution of VAFs that bears resemblance to our putative driver sites.

Association with polygenic risk score for leukocyte telomere length

When we visualise the distribution of polygenic risk scores at each position on the *TERT* promoter, our three sites of interest seem to have lower PRS. Of note, the previously discussed site chr5:1295034 once again shows a similar trend (though it is worth stressing there are only 12 individuals with mutations passing filters at this site):

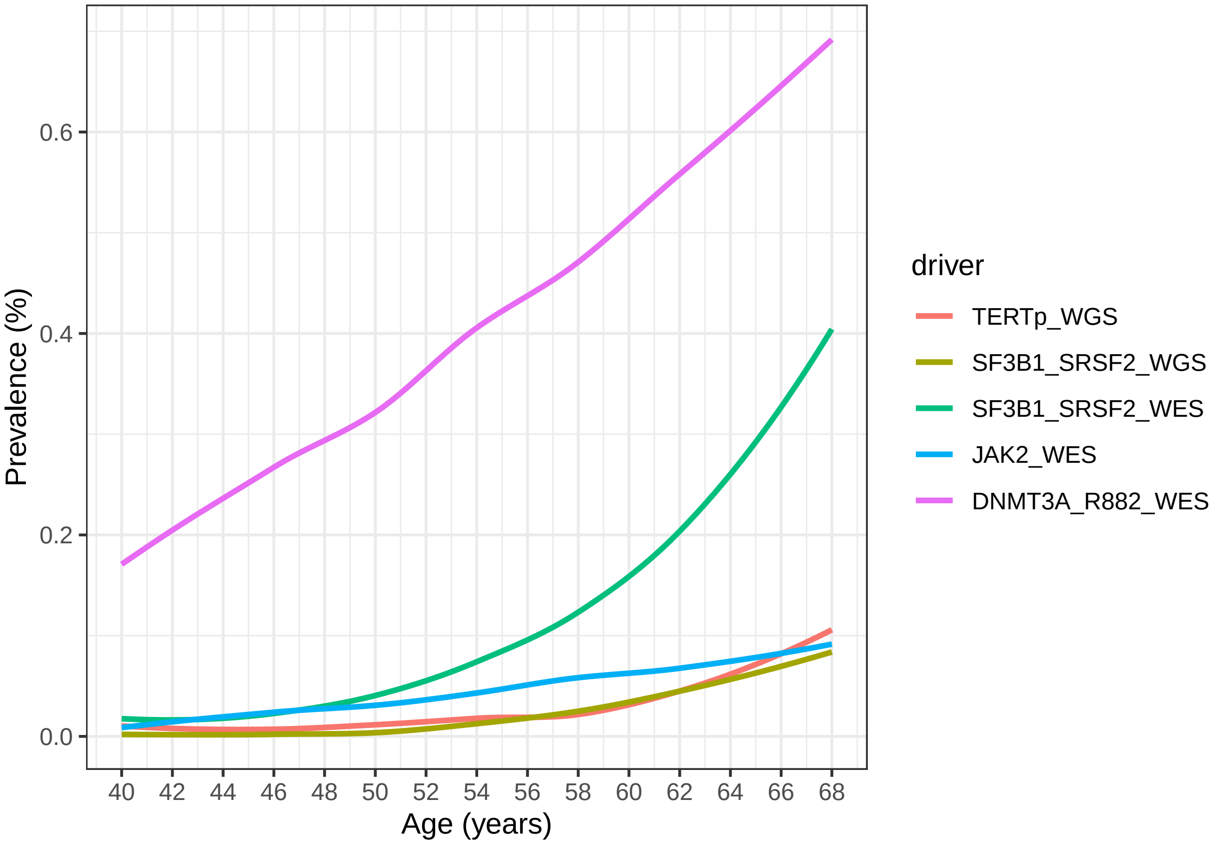


We can also make a pairwise comparison of the PRS of individuals with putative *TERT*p driver mutations with the rest of the UKB:

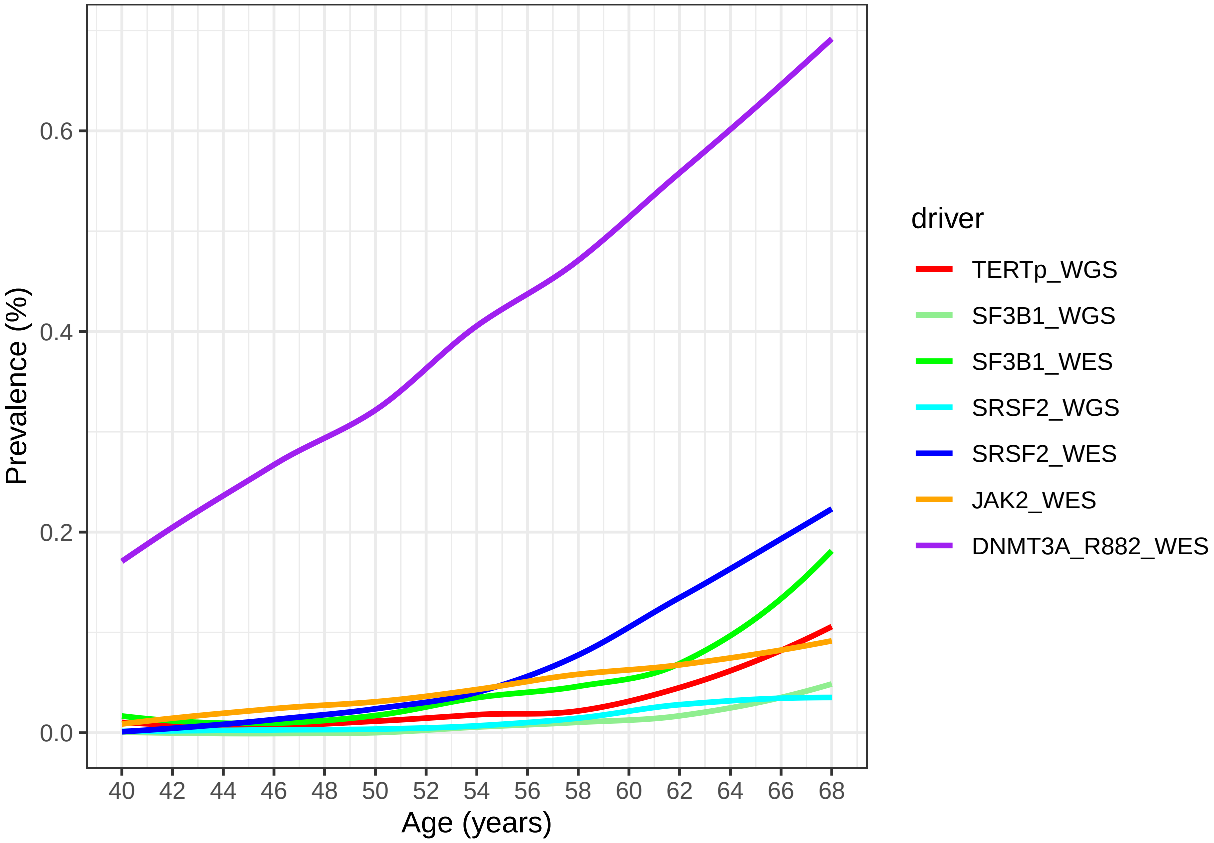


Age-related prevalence of ***TERT*p** putative drivers versus splicing mutations

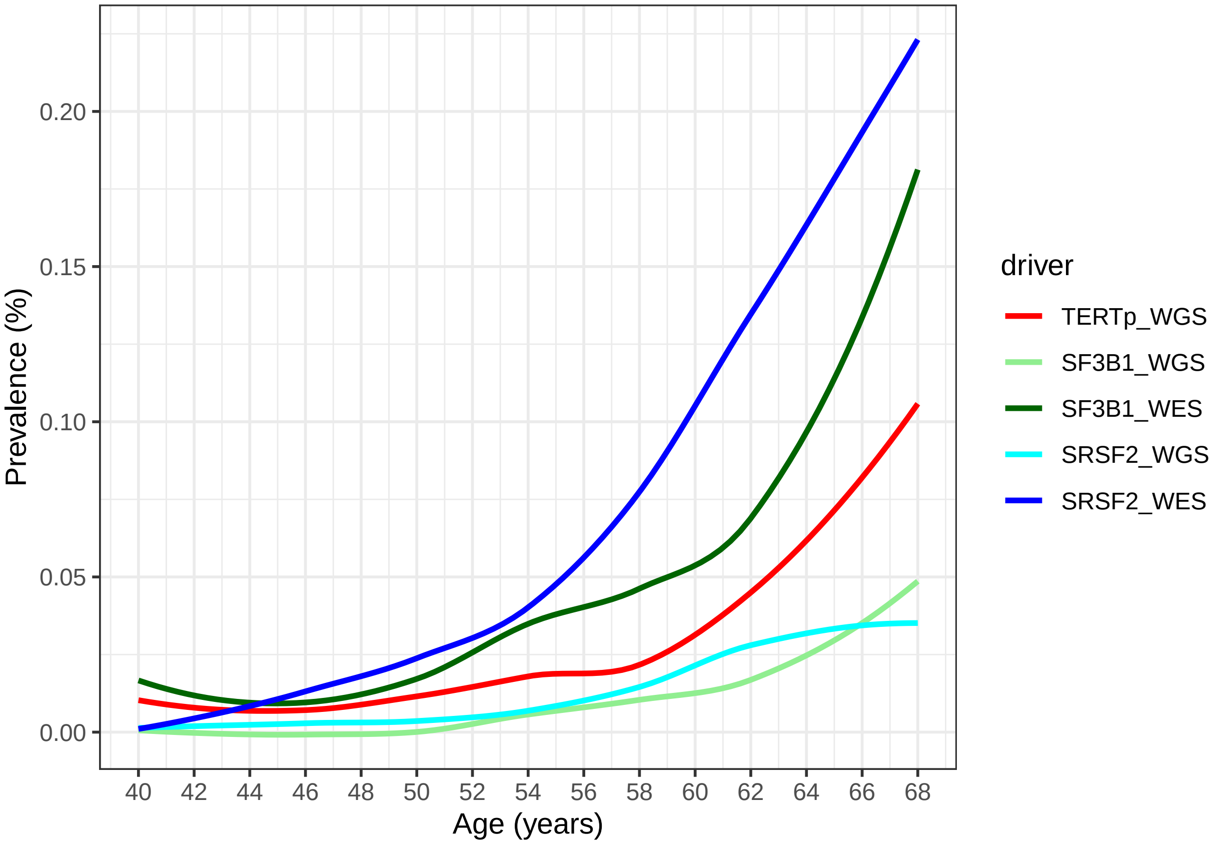
We can also compare the prevalence of the *TERT*p driver mutations we have identified here with the prevalence of splicing factor mutations called from WGS data (using the thresholds and filters specified here), and contrast this with splicing factor, *DNMT3A* and *JAK2* mutations called from WES data using a more conventional Mutect2-based approach. Here, we are only comparing the combined prevalence of the *SF3B1* (R625, K666 and K700) and *SRSF2* (P95) hotspot mutations that we called using pileup, so that we are comparing like with like:



We can also reproduce the plot but distinguishing between *SRSF2* and *SF3B1*:



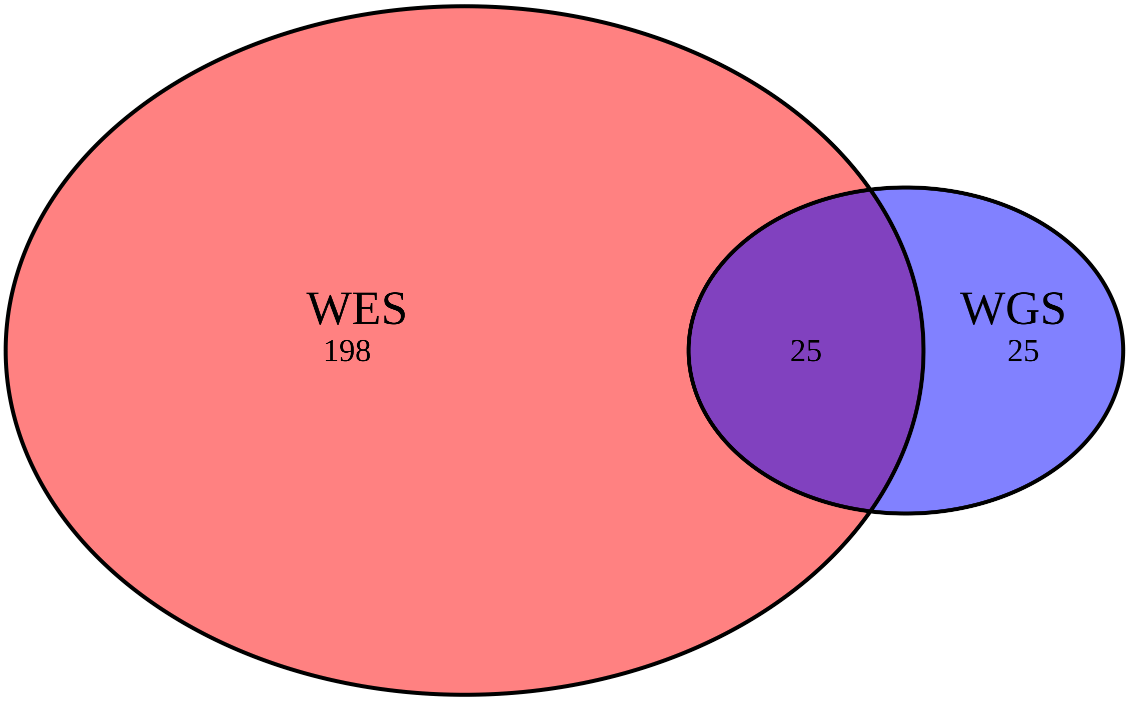
For clarity, we can also re-plot this removing the comparison with *DNMT3A* and *JAK2* hotspot mutation prevalence and colouring in different shades of green/blue for *SF3B1*/*SRSF2* respectively:



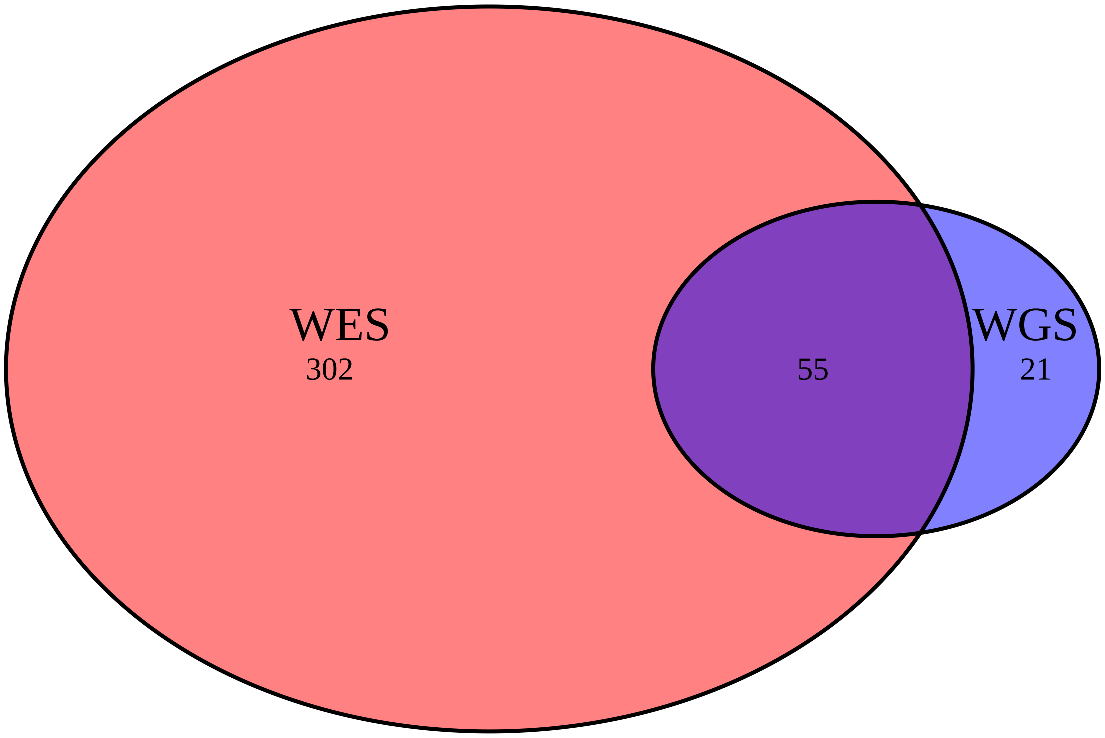
These plots show that, when we apply the same methodology to call splicing and *TERT*p hotspot mutations from WGS data, the prevalence of *TERT*p mutations appears similar to the combined prevalence of CH driven by *SF3B1* and *SRSF2* mutations.

To benchmark our ability to call hotspot mutations from WGS using the liberal filtering method described here, we can compare the overlap between *SF3B1*/*SRSF2* called using our described filtering of WGS pileup, versus calling this using Mutect2 on WES (note here that we are using the subset of UKB participants who have both WES and WGS, as ~490k have WGS c.f. only ~450k have WES).

Overlap of SF3B1 calls between WGS (pileup) and WES (Mutect2)

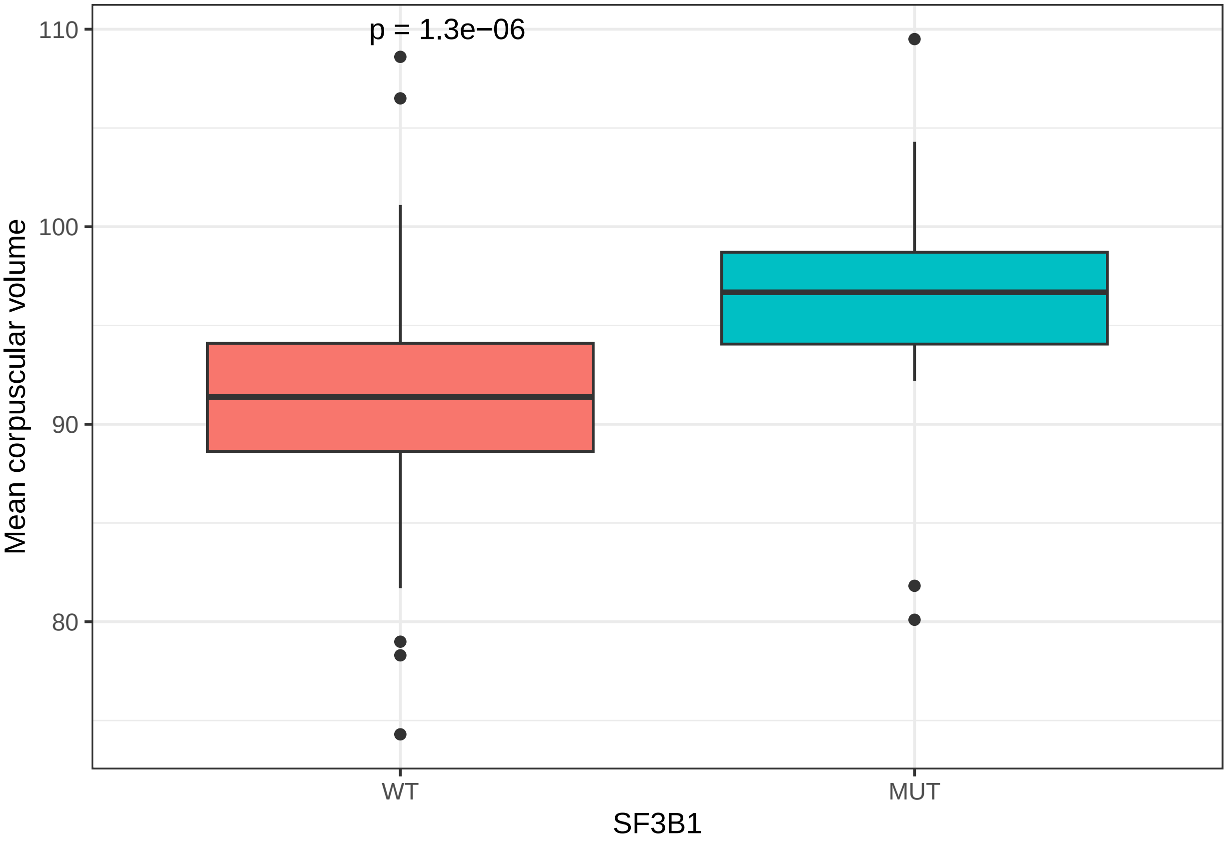


Overlap of SRSF2 calls between WGS (pileup) and WES (Mutect2)

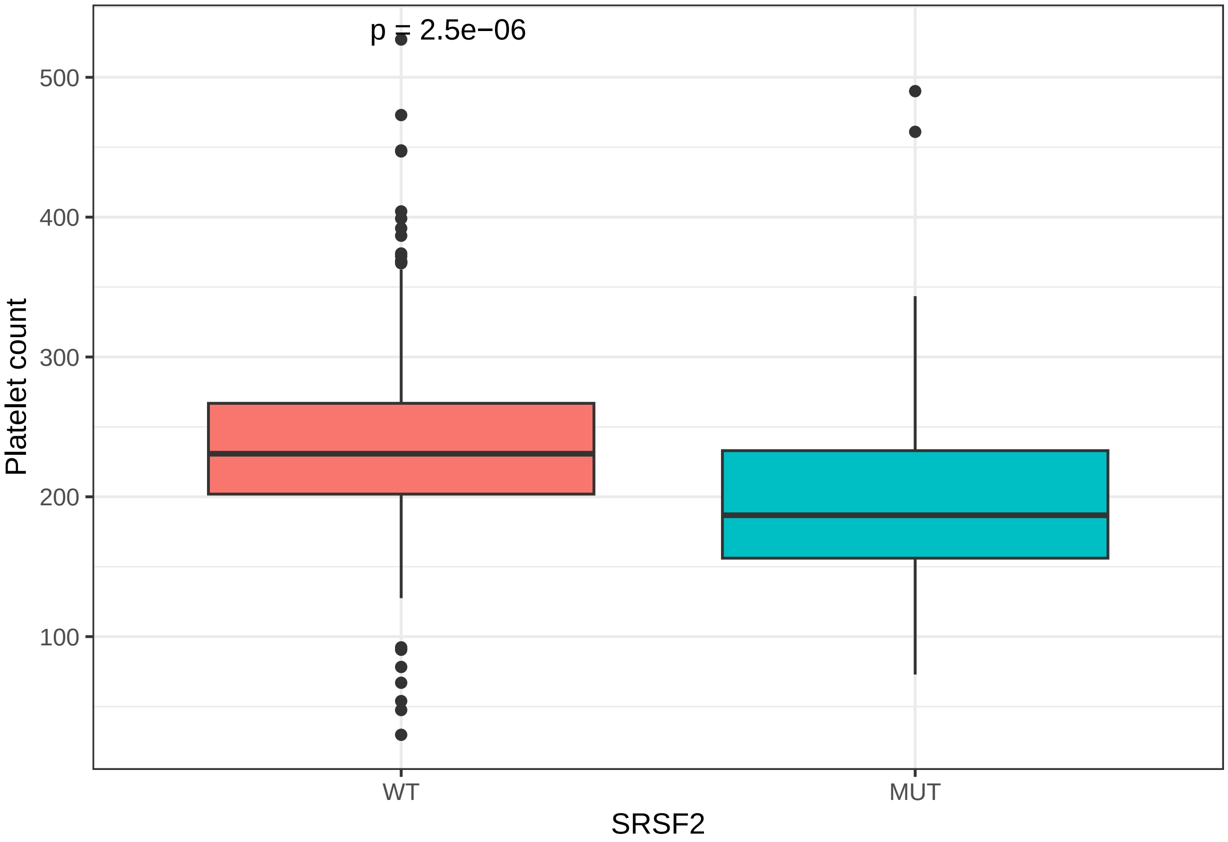


We see that there are a number of mutations that are *only* detected by pileup on WGS, despite the lower depth of WGS. To examine whether or not these appear to be "real" calls (versus error), we next examine whether these WGS-exclusive calls associate with macrocytosis (*SF3B1*) and thrombocytopenia (*SRSF2*) (that is, do they exhibit known blood count phenotypes associated with these mutations). The reference (leftmost) group in both cases is a cohort of age and sex matched controls where the sample size is ten-fold greater than the number of *SF3B1*/*SRSF2* cases respectively:

We can see that our pileup calls for *SF3B1* have a higher MCV than an age- and sex-matched cohort:



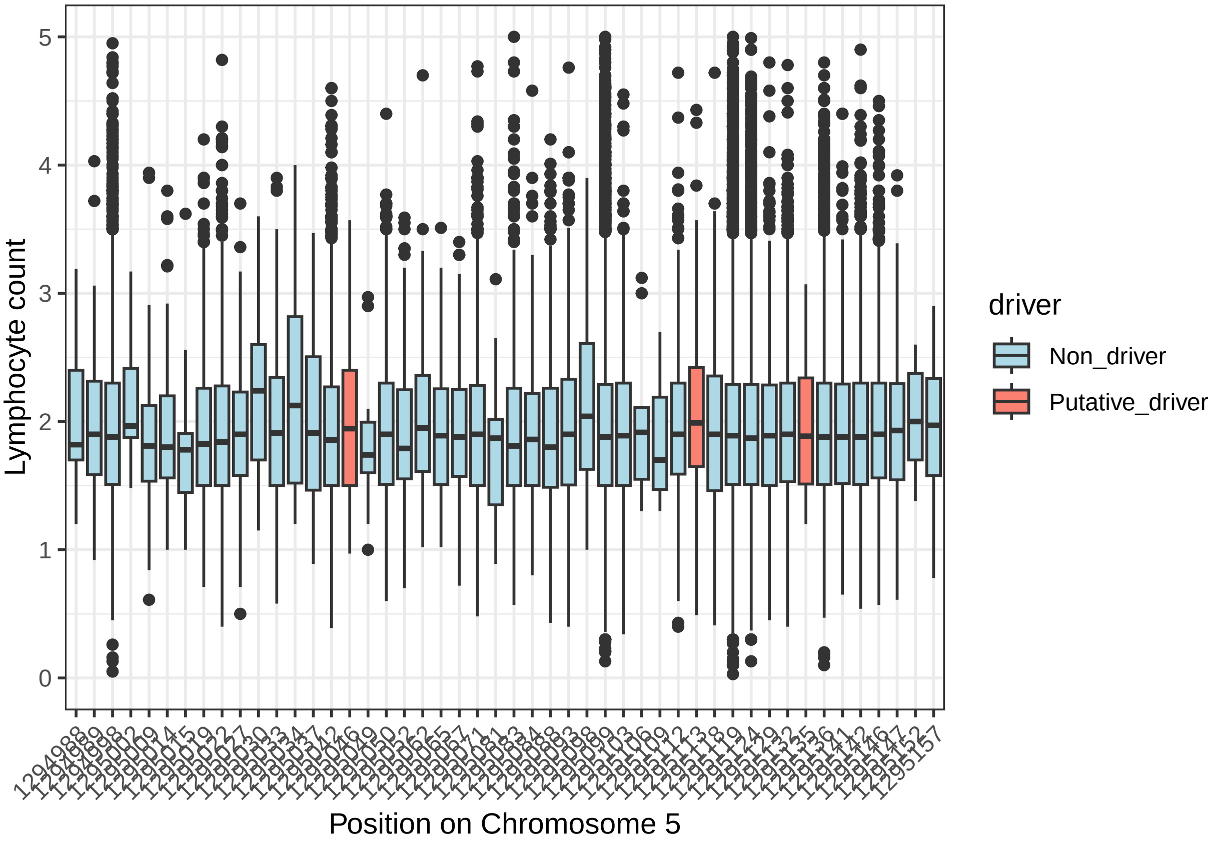
And our *SRSF2* calls by pileup have lower platelet counts:



Association between TERTp variants and lymphocyte count

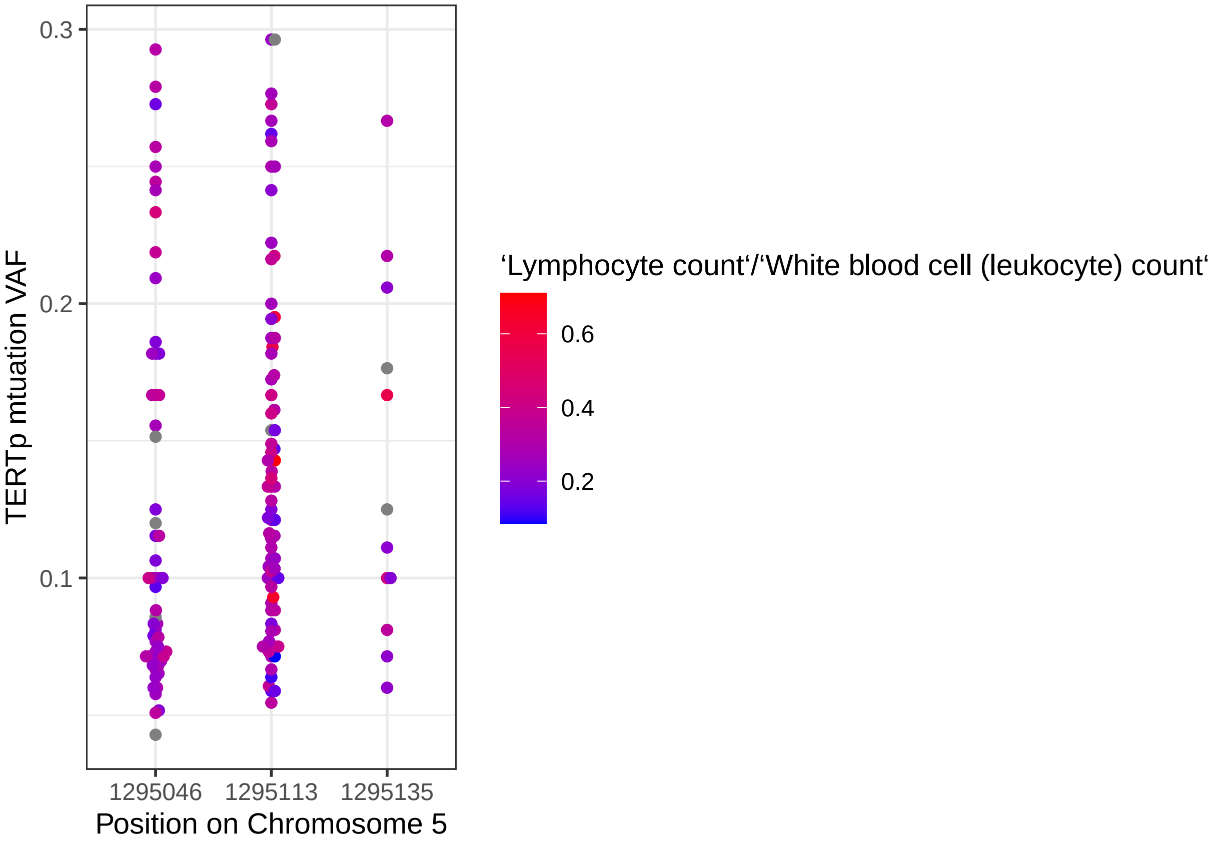
We wanted to ensure that individuals with *TERT*p mutations did not have a higher relative lymphocyte count (relative to granulocytes), since this might suggest that these mutations were present in a lymphoid progenitor, rather than a HSC or myeloid progenitor.

Firstly, we plot the absolute lymphocyte count vs position on the *TERT* promoter, to check whether or not these individuals have a phenotype suggestive of unannotated CLL or MBL:

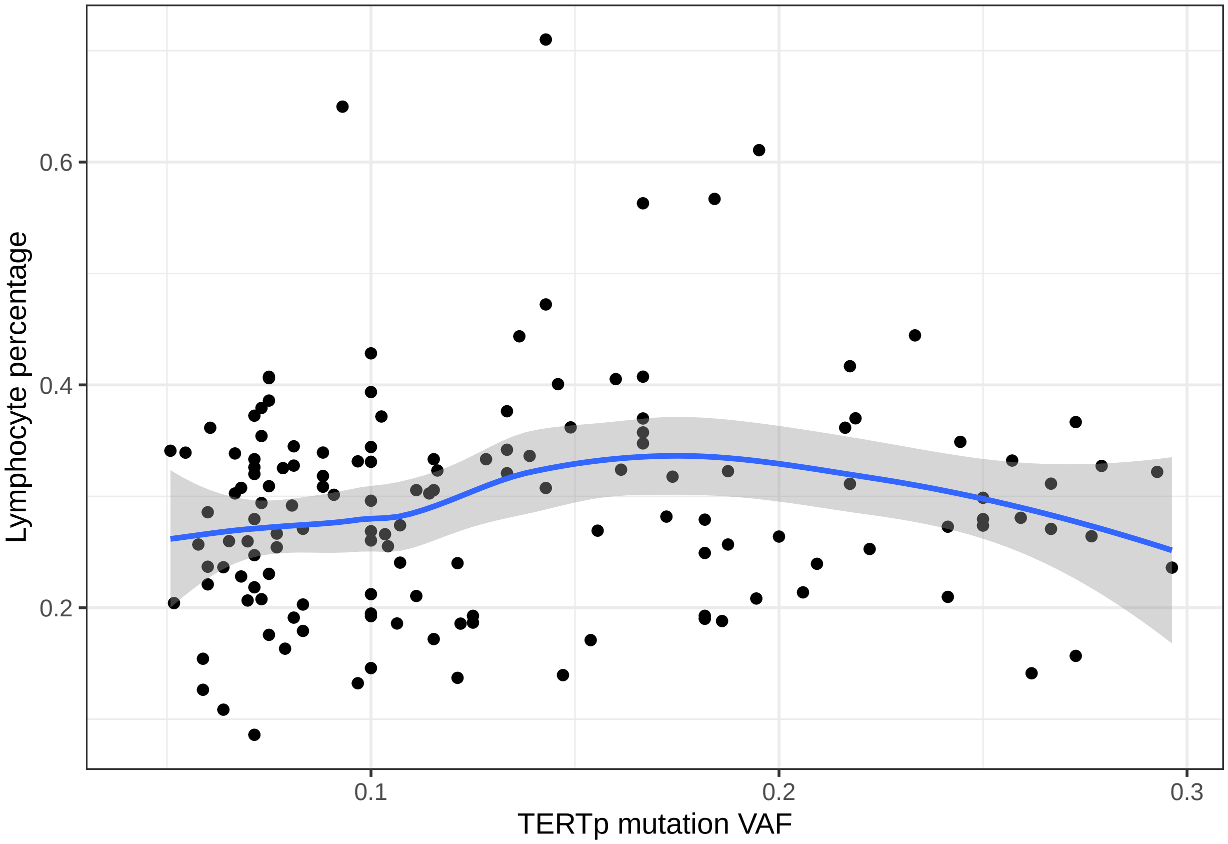


Here, we don't see any strong association between the absolute lymphocyte count and the presence of these putative *TERT*p driver mutations.

We can also look at the VAF and colour by the lymphocyte percentage to examine whether those with a high VAF have a high lymphocyte percentage (consistent with the *TERT*p mutations being present in a lymphocyte progenitor):



Alternatively, we can directly plot VAF vs lymphocyte percentage across the three sites:



We can see that there is no apparent association between TERTp mutation VAF and lymphocyte percentage, and therefore no evidence to suggest that these mutations are present in a lymphoid population.

References

1. Gutierrez-Rodrigues F, Groarke EM, Thongon N, Rodriguez-Sevilla JJ, Catto LFB, Niewisch MR, Shalhoub R, McReynolds LJ, Clé DV, Patel BA, Ma X, Hironaka D, Donaires FS, Spitofsky N, Santana BA, Lai TP, Alemu L, Kajigaya S, Darden I, Zhou W, Browne PV, Paul S, Lack J, Young DJ, DiNardo CD, Aviv A, Ma F, De Oliveira MM, de Azambuja AP, Dunbar CE, Olszewska M, Olivier E, Papapetrou EP, Giri N, Alter BP, Bonfim C, Wu CO, Garcia-Manero G, Savage SA, Young NS, Colla S, Calado RT. Clonal landscape and clinical outcomes of telomere biology disorders: somatic rescue and cancer mutations. Blood. 2024 Dec 5;144(23):2402-2416. doi: 10.1182/blood.2024025023. PMID: 39316766.