Comparision between V6 and V7

Here we compare prediction models based on GTEx V6p (2016-09-08) and V7. While training the V7 models we identified a few issues in the V6 pipeline used for the September and November 2016 release. We identified 3 issues in the V6p pipeline used in the release of September 2016. We describe here the issues, the consequences in prediction performance estimates and downstream S-PrediXcan association, and the fixes.

Issue1: Penalization parameter was estimated in-sample (using all the samples)

V6p old pipeline computed R2 using lambda estimated from all data. This led to overestimates of the performance measures. See Figure 1.

We classified gene/tissue pairs with negative correlation in the red group (to be dropped). This proportion ranges from 6 to 13%, with larger proportions for tissues with smaller sample sizes. If a more conservative cutoff of 0.10 is used, 22% to 29% of genes would be dropped.

Issue2: There was induced negative correlation between training and test sets

This was causing a slight underestimation of performance. In the old pipeline, the correlation between predicted and actual was computed using all folds combined. The sample mean of the training set (which excludes the test fold) is necessarily of opposite sign to the sample mean of the fold (all data mean was 0, so that the test fold average is of opposite sign to the training set's average). This effect leads to a small underestimation of the correlation. Figure 2 shows the correlation estimated within each fold and averaged across folds, which is slightly larger than the correlation from the old pipeline.

Issue3 Sign of correlation (predicted vs actual) was ignored

There were a few genes that had negative correlation between predicted and actual but the old pipeline was accepting them as reliable model because it was only considering the square and the p-value of the correlation.

Consequence in prediction performance measures

The combined consequence of the three issues can be seen in Figure 1 where the correlation between predicted and actual (with in sample lambda for the old pipeline and out of sample lambda for the new pipeline) is shown. This error did not affect well predicted genes (upper right region of the plot). The lower left region shows that the old pipeline was over estimating the correlation (most points to the left of the identity line).

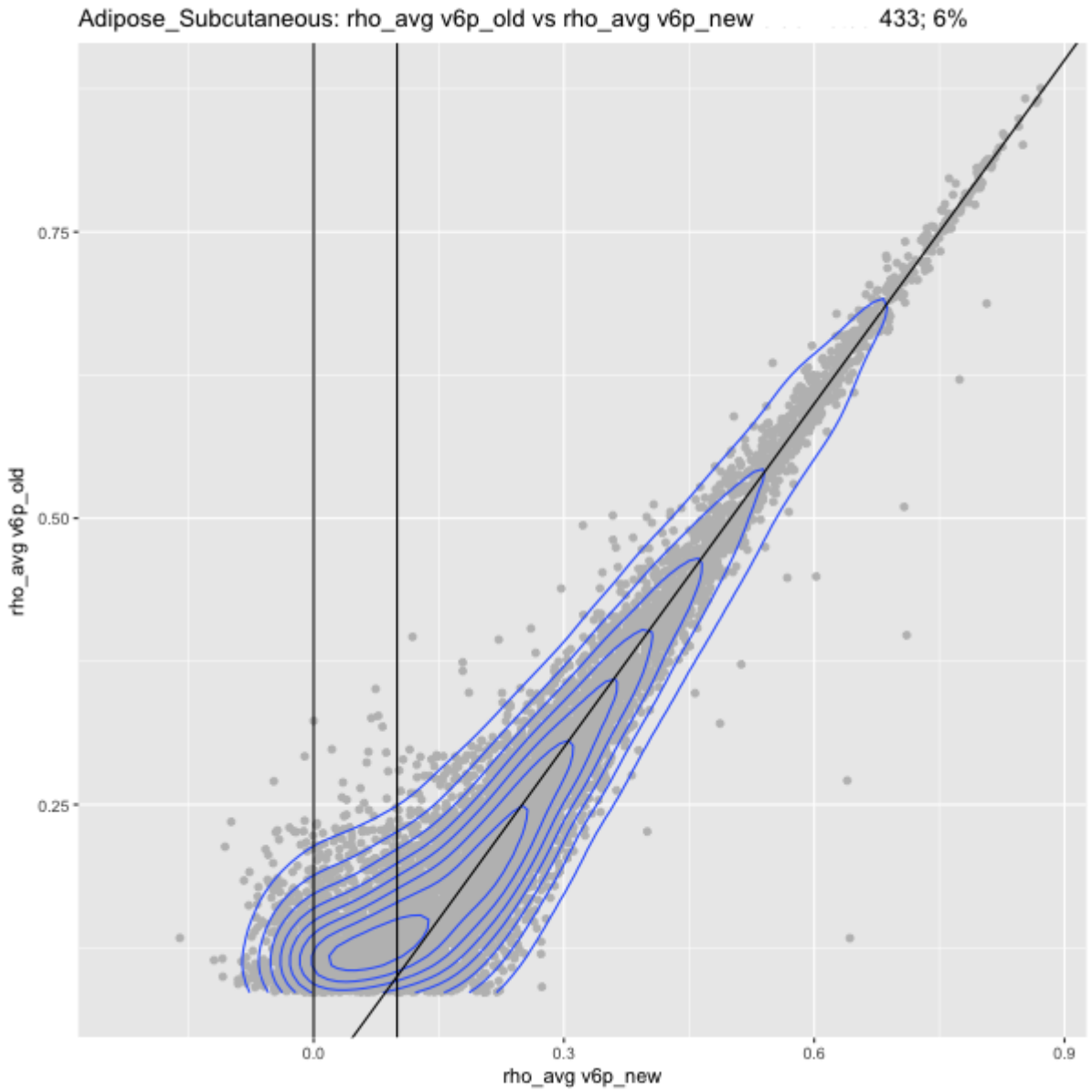

Fig 1: Estimated correlation between predicted and observed expression levels. The old pipeline overestimated (y axis) the correlation between predicted and actual expression levels (x axis). The effect is more notable for at the lower end of correlations.

Figure 2 shows the old pipeline estimates on the y axis, and on the x-axis is the correlation using the in-sample (all data) lambda parameter but computing the correlation within each fold and averaging across folds. Therefore, Figure 2 is showing the effects of issues 2 and 3. There are more points to the left of the identity line, indicating that issue 2 was underestimating the performance.

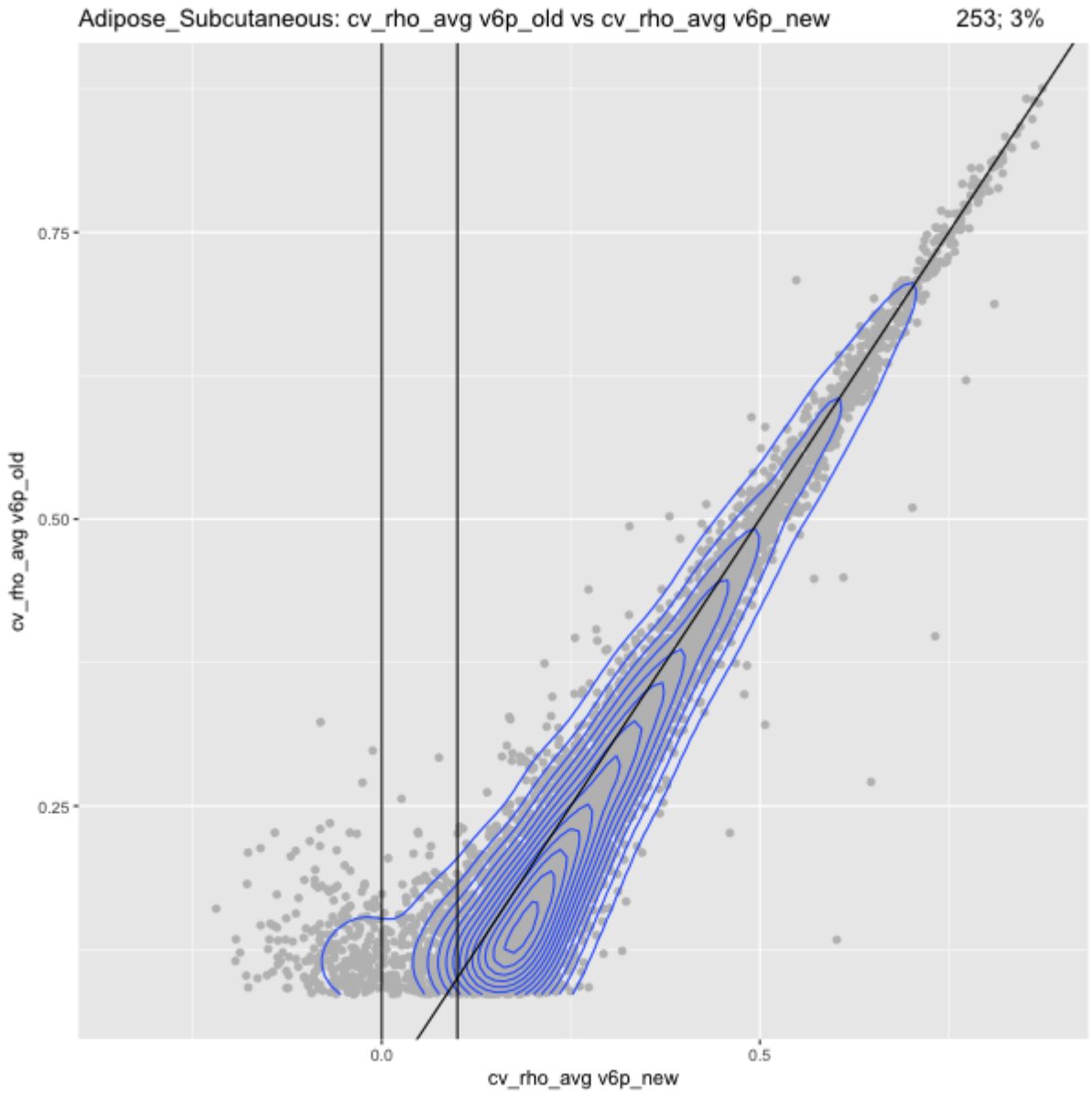

Fig 2: (in-sample lambda) cross validated correlation between predicted and actual expression level. The corrected pipeline's correlation estimates tend to be larger (more points below the identity line) than the old pipeline's estimates.

Definition of color coding by reliability

- **Red:** gene/tissue pairs for which the correlation between predicted and actual expression (nested cross validation) levels were negative with the new pipeline with nested cross validation (all parameters including the penalization parameter lambda is computed without using the data to be predicted).
- **Green:** gene/tissue pairs that pass the new more stringent performance criterion (nested cross validated correlation > 0.10, pvalue<0.05, correlation of old and new pipeline prediction > 0.9)
- **Yellow:** remaining gene/tissue pairs.

Consequences in S-PrediXcan results

Generally speaking, S-PrediXcan Z-scores remain mostly around the identity line. Figure 3 shows V7 vs V6 PrediXcan Z-scores for the height phenotype from the GIANT consortium. We also examined

similar figures for over 40 phenotypes and found that - Genes in common yield similar predixcan Z-scores - Genes in V7 and not in V6p tend to be slightly more significant than genes in V6p and not in V7.

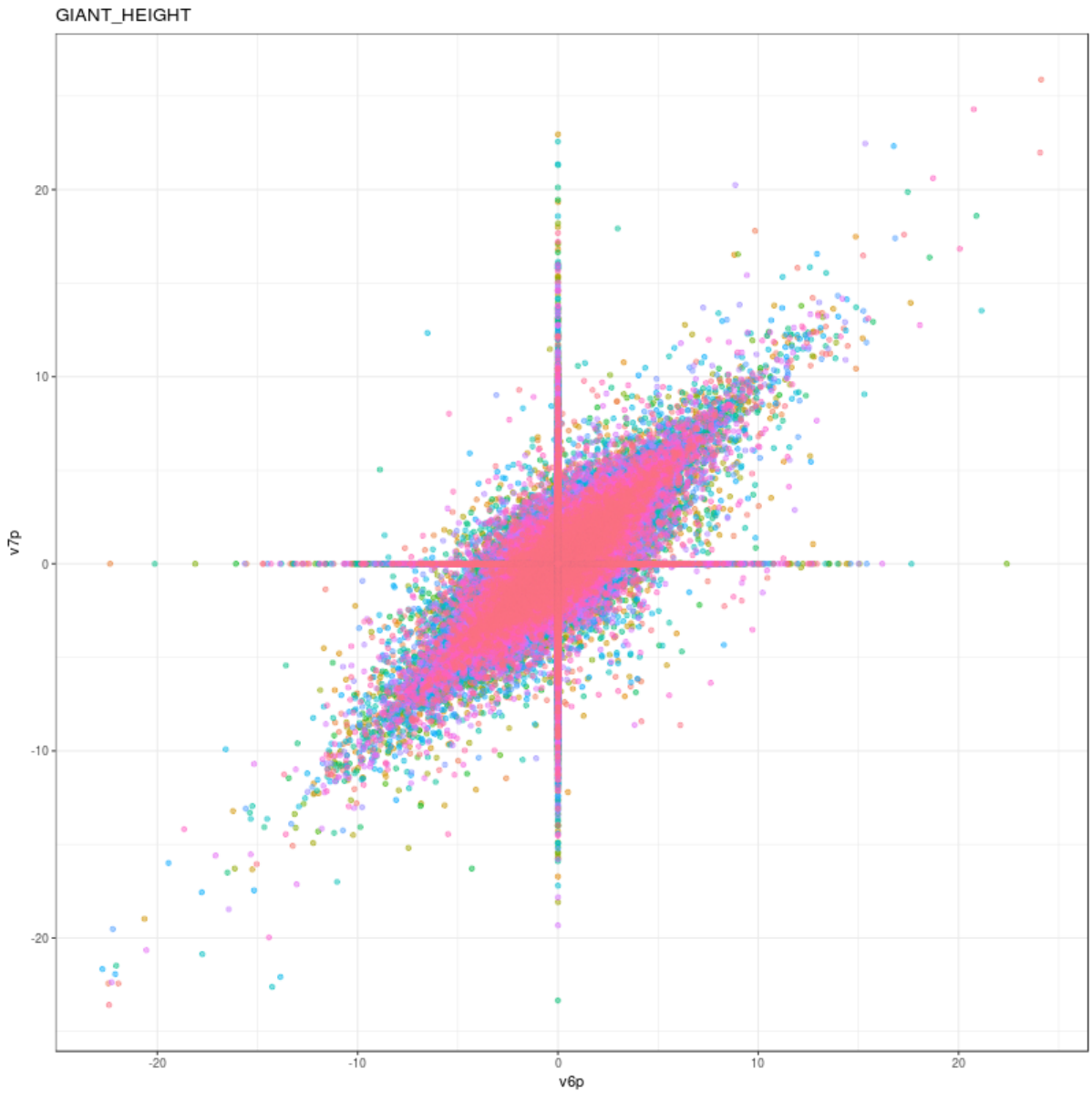

Figure 3: Comparison of S-PrediXcan Z-score using V7 (y-axis) vs V6p (old pipeline). Most gene tissue pairs cluster around the identity line. Points strictly in horizontal axis and vertical axis correspond to gene/tissue pairs that were missing in the other version.

Another change in the criterion for filtering out genes

Before we were using FDR$<$0.05 to filter out genes (using genes within one tissue at a time). Genes for which no convergence was achieved were being ignored, causing another source of bias. One workaround was to assign a uniformly distributed p-values to non converging genes. After some thought, we preferred not to introduce another source of variability and decided on using the following criteria to filter out genes:

- nested cross validated correlation between predicted and actual levels > 0.10 (or equivalently R2>1%) and

- p-value of the correlation test < 0.05.

Prediction performance p-value and R2 will be provided for users who want to use a more stringent criterion.

Should you switch to version 7?

We chose to fit GTEx version 7 data using only Europeans. This has the benefit of allowing us to better estimate LD and making S-PrediXcan more robust, thus avoiding some uncommon but extreme false positive associations. When the correlation between SNPs is not properly estimated, S-PrediXcan may consider that for example two modest associations to be independent evidence of association and thus yield a very significant value when in fact in the study population the SNPs are highly correlated and the two signals should be considered to be the same.

If S-PrediXcan will be used, then V7 is provide more robust associations and fewer instances of extreme false positive p-values. However, if individual level data are available, no extern al LD information is needed since the regression with individual level data will implicitly use the study population's LD.

Figure 4 shows that in general, V7 has better performance measures (larger nested cross validated correlation between predicted and actual). This is probably due to the small increase in sample size (30, about 10% for adipose subcutaneous tissue as shown) and the homogeneity of the population (only Europeans were used in V7 training). However, there is a sizable number of genes with low performance in V7 but relatively good performance in V6p. It is possible that this improved performance in V6p is due to the more diverse population representation in the V6p samples but further comparisons are needed to test this.

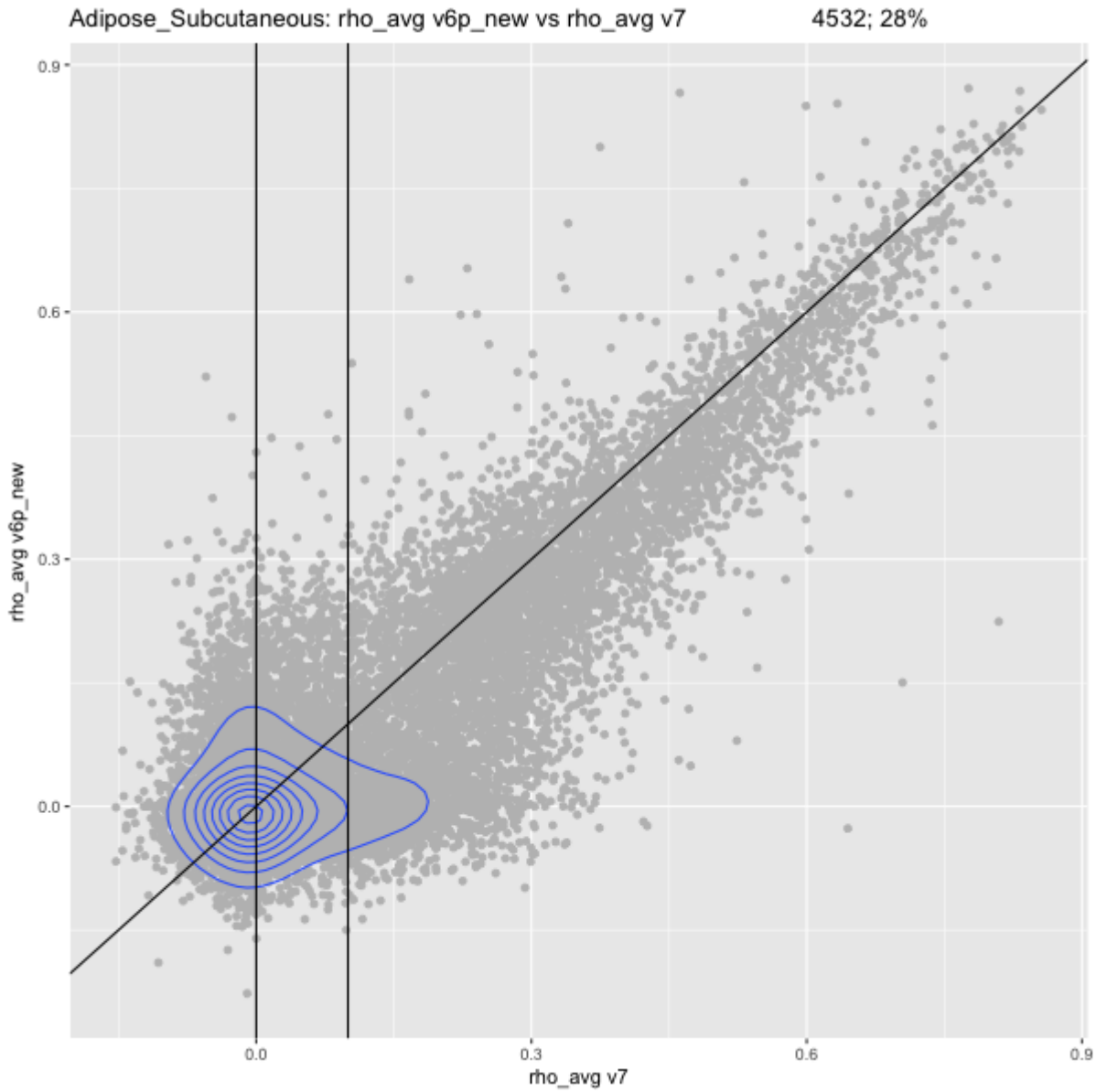

Figure 4: Comparison of performance between V6p and V7 both with corrected pipeline. Nested cross validated correlation between predicted and actual expression levels in V6p vs V7 (adipose subcutaneous shown). Genes in V7 tend to have better performance, i.e. more points are located below the identity line. However, there is also a sizable number of genes that were relatively well predicted in V6p but have correlation near 0 in V7.