# Pathway Methods

## Pathway Selection

Almost all pathways that were used came from either MSigDB1 (v6.2, accessed 3/8/2019) or BioPlanet2 (v1.0, accessed 7/29/2019). Two additional estrogen receptor alpha pathways based on the literature3 were manually entered: “Ryan Estrogen Receptor Alpha Up/Down”. The entire MSigDB set, comprised of 18,674 gene sets, was downloaded as an .xml file and imported into Excel. This was then pared down to the 14,660 gene sets with “Homo sapiens” under the organism heading. Only two types of gene set were kept from MSigDB: 50 Hallmark4 gene sets and 352 gene sets in the C2: chemical and genetic perturbations category with “RESPONSE” in their pathway name. Hallmarks were chosen because they “represent specific well-defined biological states or processes”. The “RESPONSE” gene sets were kept because they generally represent experiments like our own where the response to a specific chemical is being identified. Also, we observed that there were some useful estrogen response pathways present in this set, such as “Dutertre Estradiol Response 6HR UP”5. 1,658 BioPlanet pathways were downloaded in a .csv and combined with the rest to yield a beginning set of 2,062 pathways. We sometimes referred to this as the “bhrr” pathset (short for: BioPlanet, Hallmark, Response, Ryan).

For our purposes, pathways are simply a subset of genes whose expressions are aggregated in some way to derive a pathway score. Since there are roughly 10,000 genes in this experiment, the number of possible pathways is on the order of 210000. It should be expected that every chemical will be active in many of these potential pathways due to random noise, so we were wary of constructing pathways that fit the data. Of course, this could still be done using a validation framework to avoid overfitting, but only in cases where the number of chemicals in the data known to produce a given activity is sufficiently large. So far, only pre-existing pathways have been used to give some confidence that results were reproducible in another context. Ideally, we would use only a small number of trusted pathways, both to reduce processing time and to ease interpretability. Using many overly similar pathways or pathways that are unreliable would increase the number of false positives in the data and potentially swamp the signal. How to choose pathways in a principled way is a deep question that we have not entirely addressed yet.

## Pathway Scoring

The input to the pathway pipeline was a matrix of log2(fold change) (“l2fc”) values where the rows are chemical/concentration combinations and the columns were probes. When more than one probe was assigned to a single gene, we used the maximum of the probe values as the value for the matching gene. This yielded a chemical/concentration by gene matrix of l2fcs. Genes with missing values for all chemical/concentrations were omitted and pathways containing fewer than 10 of the remaining genes were dropped immediately before scoring. During scoring, pathway sizes were calculated that took into account the actual number of non-missing genes used to compute each pathway/chemical/concentration combination. Specific pathway scores calculated with fewer than 10 non-missing genes were then dropped afterwards to ensure pathway sizing consistency. Small pathways were dropped to avoid scores potentially dominated by the noise of a small number of genes.

Three scoring methods were seriously entertained: “FC” (fold change), “MYGSEA” (a modified version of single sample gene set enrichment analysis or “ssGSEA”), and “GSVA6” (gene set variation analysis). The FC method was originally intended to be a simple method with which to compare the others, but it has some merits of its own in the context of concentration response modeling. The FC method is simply

where are the l2fc’s for genes in a given pathway and are the l2fc’s for genes outside of a given pathway.

For the GSVA method, we used the GSVA R package (v1.32.0); the steps of GSVA are briefly summarized here. First, an empirical cumulative distribution (ECDF) with Gaussian kernels is estimated for each gene using all the samples (chemical/concentrations) and individual l2fc’s are replaced by their value in this ECDF. Then the genes are ranked within each sample separately and these ranks are used to calculate a Kolmogorov-Smirnov (KS) like random walk statistic for each chemical/concentration/pathway combination (similar to regular Gene Set Enrichment Analysis7 or “GSEA”), exemplified by Figure 1. The final pathway score is the difference between the absolute maximum and absolute minimum values of this statistic.



Figure 1: KS random walk statistic for GSVA

ssGSEA8 is a straightforward generalization of GSEA to individual samples. Within a given sample, the genes are replaced by their ranks and a KS statistic is computed. The enrichment score is the sum of all values of the KS statistic. Optionally, the score can be normalized by dividing by the range of the enrichment scores produced across all experiments. The scores for each sample are thus computed independently of one another except for a final normalizing factor. MYGSEA was based on the ssGSEA code provided in the R GSVA package. The method was changed in two ways: first, the l2fc values in each sample x were normalized using

where x is the original l2fc and p is the length of the sample. This ensured that both extremes of l2fc were upweighted, instead of only the largest positive value. Secondly, the code was altered to ignore missing l2fc’s, where it previously treated the missing genes as if they were the most active.

Each method has its own theoretical strengths and weaknesses. GSVA has the advantage of following a normal distribution in its output scores due to the use of the Gaussian kernel. However, because it normalizes across all samples before computing a KS statistic, the scores for a given experiment will change according to which other samples are included in the processing. This means that the entire data set must be scored simultaneously in order to yield consistent results; additionally, if any samples are later deemed to be outliers, the entire analysis has to be performed again. The FC method has the advantage of being the fastest computationally and of being the least removed from the original data. Its distribution of scores tends to be somewhat leptokurtic and it is more sensitive to outliers than other methods. MYGSEA falls somewhere in between the other two in terms of: closeness to the original data, outlier sensitivity, output distribution kurtosis, computational speed, and ease of computing scores for data subsets.

## Pathway Noise Estimation

Originally, the noise levels were set to three times the baseline median absolute deviation. The baseline was taken to be the pathway scores of all samples below some concentration under which the activity was assumed to be zero. This concentration was typically set to .2 uM so that the two lowest concentrations were taken to be part of the baseline. However, this approach had two weaknesses. First, some samples were, in fact, active at all concentrations; this effect was most marked in the estrogen pathways because several highly potent estrogens were present in the dataset. Additionally, since different scoring methods had differently shaped distributions, the same multiple of MAD corresponded to different sensitivities for each method. For instance, there might be the same number of active pathways above 2 BMAD in GSVA as above 3.5 BMAD in the FC method.

To estimate noise, we first generated a set of randomized null data. This was done by using the empirical distribution at all concentrations for a given gene to randomly resample that gene for some number, N, of null samples (typically N = 1,000). For each gene, the uniform distribution was sampled N times and these values were fed into the empirical quantile function (using R’s “quantile”) to retrieve a resampled distribution. Then, missing values were added in random locations of the resampling to match the original fraction of missing values. Thus, the null data preserved the distributional properties of each gene, but any correlation between genes was broken. An example of the resulting null probability density is shown in Figure 2. The null data was then scored using the same scoring pipeline as the actual data to estimate the distribution of noise in the data. The cutoff was set to be greater than some percentage (typically 95%, corresponding to a p-value of .05) of the absolute pathway scores; this was referred to as the p-value cutoff.

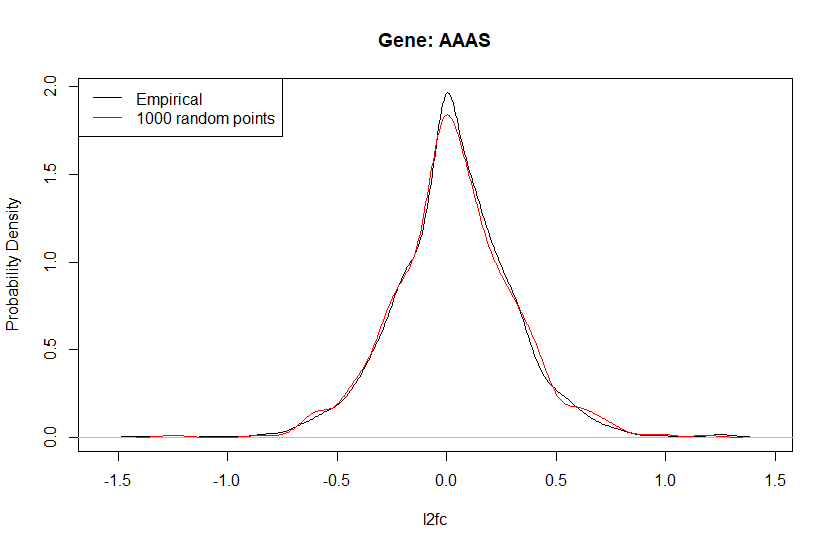


Figure 2: Example of null data probability density (red) for gene AAAS versus probability density across all samples (black).

As an alternate strategy we considered using the False Discovery Rate (FDR). This method set the cutoff at the level at which some percent of the values are expected to be false positives. This was done by calculating the p-value of each pathway response relative to that pathway’s noise, converting those to FDR values, and choosing the cutoff to be halfway between the smallest response meeting the FDR criteria and the largest response that does not meet it. Figure 3 shows an example of the pathway score empirical probability density for Hallmark Estrogen Response Early versus the null distribution and the corresponding FDR and p-value cutoffs that would be selected. A p-value of .05 leads to a cutoff that’s near the ends of the null distribution, irrespective the empirical data. The FDR cutoff, on the other hand, is lower in cases, such as this, where the signal/noise ratio is high, but can be much higher or even infinite when the signal/noise ratio is low. At the FDR 0.25 cutoff, four times as many absolute scores in the actual data exceed the cutoff as do those in the null data.

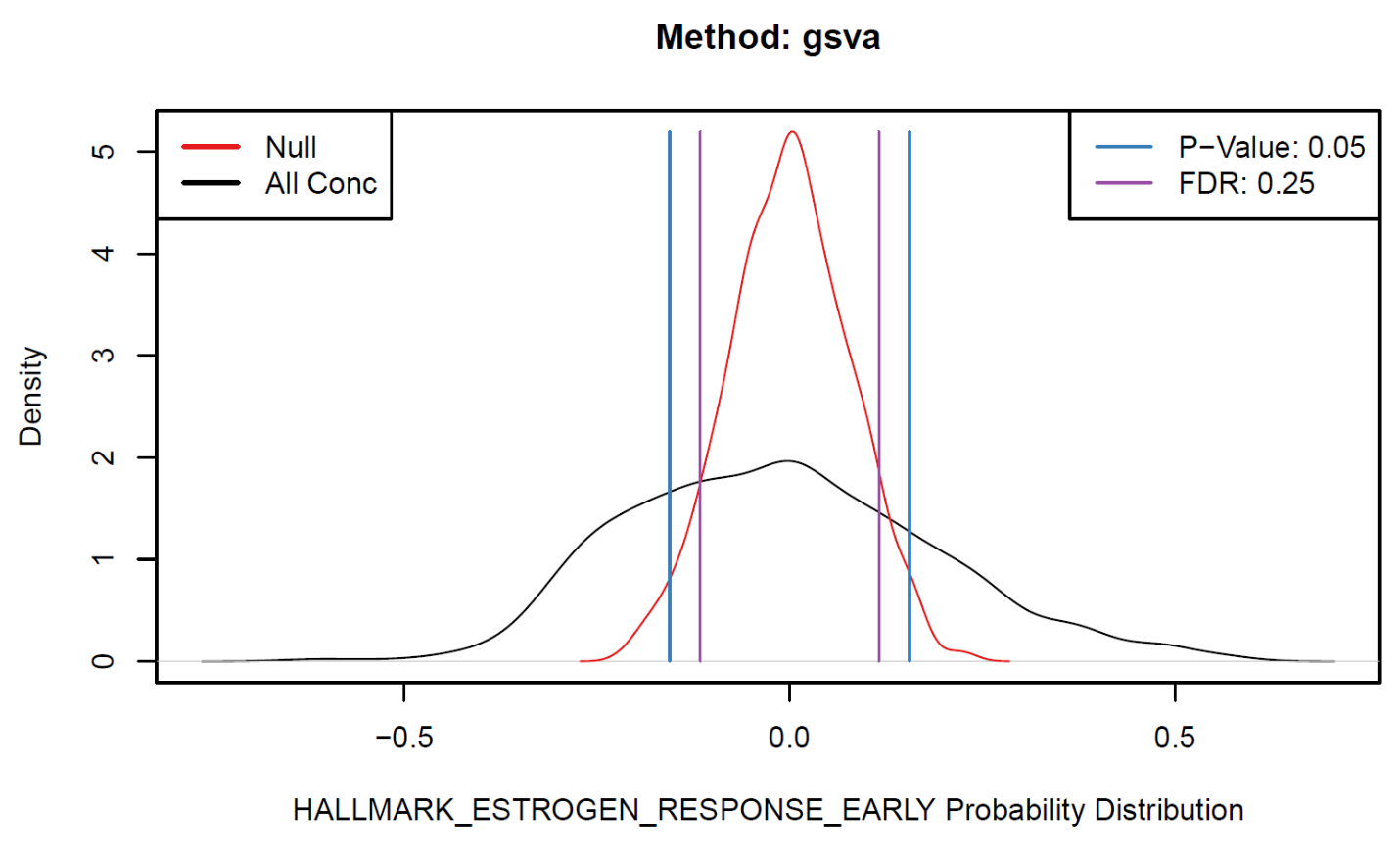


Figure 3: Null score distribution for a given pathway (red) versus actual data score distribution (black). These are used to compute the 0.05 p-value cutoff (blue) and 0.25 FDR cutoff (purple).

# Pathway Concentration Response

## Model Fitting

After estimating the cutoff, the median of each pathway’s null data was subtracted from its corresponding scores. Each individual chemical/pathway combination was fit to each of ten concentration response models. Modeling routines were creating by overhauling a version of the ToxCast pipeline9 (“tcpl”). Data with fewer than four responses were not fit to any models. When using discrete hitcalls, data with no responses above the cutoff were only fit to the constant model, since they were guaranteed not to be hits. When using continuous hitcalls, however, these cases were not filtered out.

In addition to the original tcpl models (constant, hill, and gain-loss), we added all continuous models listed in the BMD technical guidance10: exponential 2/3/4/5, polynomial 1/2, and power. For models where the background , the background was subtracted off, resulting in one fewer parameter in the case of exponential 4/5. Once background was subtracted, we were left with four model behavior categories: constant, gain-loss, hill-type (hill, exponential 4/5), and unbounded growth (exponential 2/3, polynomial 1/2, power). The models are described in further detail in Appendix A.

Models were fit using maximum likelihood estimation; nonlinear constrained optimization was performed using the Nelder-Mead method for all models, except the constant model, which used the Brent method. The value to be maximized was:

where is the t-distribution with degrees of freedom, is the i-th concentration, Is the i-th response, N is the number of concentration/response pairs, is the model curve with parameters to be optimized, and is the error parameter that is concurrently optimized. The 4 degree of freedom t-distribution was used to give a fit that is robust to outliers; however, whether this is preferable to some other distribution in the case of pathway scores is an open question. Additionally, it may also be advantageous to use the noise estimation as the error term instead of allowing it to be optimized.

Initial conditions and bounds were chosen separately for each method beforehand (detailed in Appendix A). Bounds and initial conditions of models with natural “top” and AC50 parameters (Exponential 4/5, Hill, Gain-Loss) were set similarly to those found in tcpl whenever practical. The absolute value of the top for these functions was limited to less than 1.2 times the absolute value of the maximum median response. AC50’s were bounded between one log unit below the lowest concentration and half a log unit above. The minimum distance between the Gain-Loss gain AC50 and loss AC50 was increased to 1.5 log units. Other models were given initial conditions such that the curve passes through the final concentration-response pair. Bounds for these models were set as loosely as possible, while protecting against underflow/overflow errors and ensuring that each function was monotonic.

## Model Selection

Following a similar example in the literature11, a nested likelihood ratio test was used to choose the best polynomial model. The quadratic model was required to have a significantly improved fit () over the linear model, otherwise it was discarded. After this, the Aikaike Information Criterion (“AIC”) was calculated for each and the model with the lowest AIC was selected as the “winning model”. In the case of discrete hits, the constant model was capable of being chosen (giving a hitcall of zero). When computing continuous hitcalls, the constant models was not allowed to be chosen as the winning model, but its AIC relative to the winning model was used to calculate the hitcall. Because of the relatively small number of samples (typically eight), use of the corrected AIC (“AICc”) was also considered:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where is the number of samples and is the number of model parameters. This correction gives a large penalty to models with greater numbers of parameters, causing some to never be selected as shown in Figure 4. However, the eight gene l2fc’s at each concentration are an aggregate of 24 original replicates, three at each concentration, and the pathway scores are an aggregate of many genes themselves. So, whether it is valid to use a higher in the AICc or perhaps reformulate the process so that each of the 24 replicates is scored separately remains an open question. Furthermore, the common AICc formula shown in Equation (1) assumes a linear model, when all but one of the models used were nonlinear.

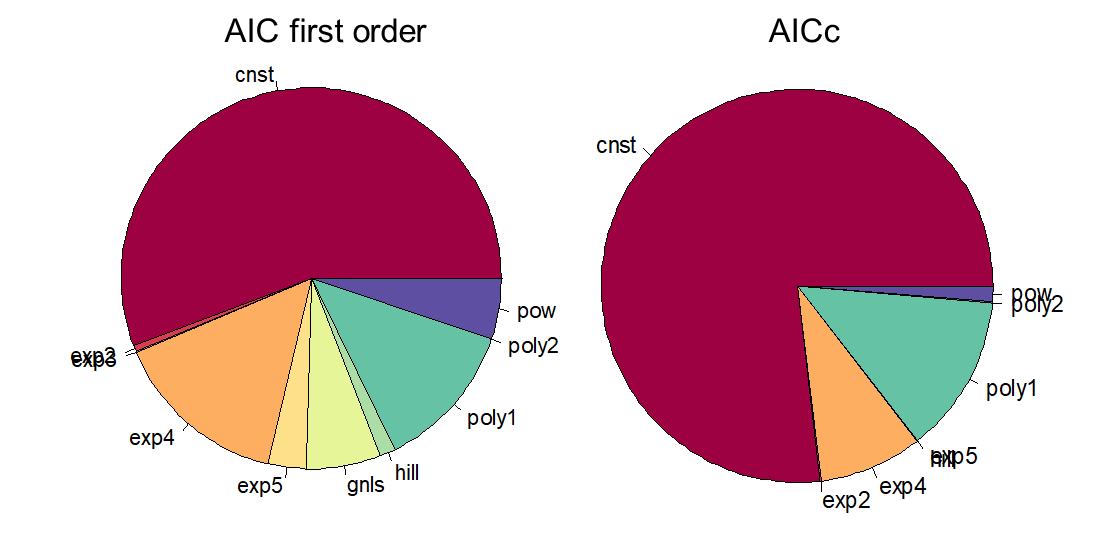


Figure 4: An example of model selection frequencies using first order AIC versus AICc

## Top/AC50/BMD

Unbounded growth type models lacked an explicit “top” parameter, so their top was defined as the model value when it passes through the highest concentration. We expect that these curves would eventually reach some saturation value and stop growing at some higher, untested concentration, so in these cases the top value signifies a lower bound. The AC50 in these cases was defined as the concentration at which half the top value was reached. Benchmark Doses (“BMDs”)12 were calculated as the point at which the curve first crossed the Benchmark Response (BMR). Because both positive and negative responses were allowed, we set the BMR at 1.349 times the standard deviation of the noise estimation11. The BMD was calculated using the mathematical inverse of the model function, except in the case of the Gain-Loss model, for which the inverse had to be approximated numerically. AC50’s were computed in the same fashion when the AC50 was not available as an explicit model parameter.

BMD bounds (i.e. BMDL and BMDU) were computed in accordance with the profile likelihood method13. Each model function was first re-parameterized by performing the appropriate substitution (see Appendix A) for a chosen parameter so that the BMD appears explicitly. Parameters that represented the x-scale or the AC50 were chosen to be substituted whenever possible. Once this was done, the BMD was adjusted until the corresponding likelihood was reduced from the maximum likelihood by where represents the quantile of the distribution corresponding to one degree of freedom and 90% confidence limits. Sometimes, the BMDL and/or BMDU could not be computed because no such solution existed. This typically occurred when the BMD was near the maximum or minimum concentration tested or when the Gain-Loss model top was close to the BMR. Figure 5 gives an example of a model fit with corresponding BMD ranges. Figure 6 shows a case where the BMR is too close to the gain-loss top for a BMDU to be calculated.

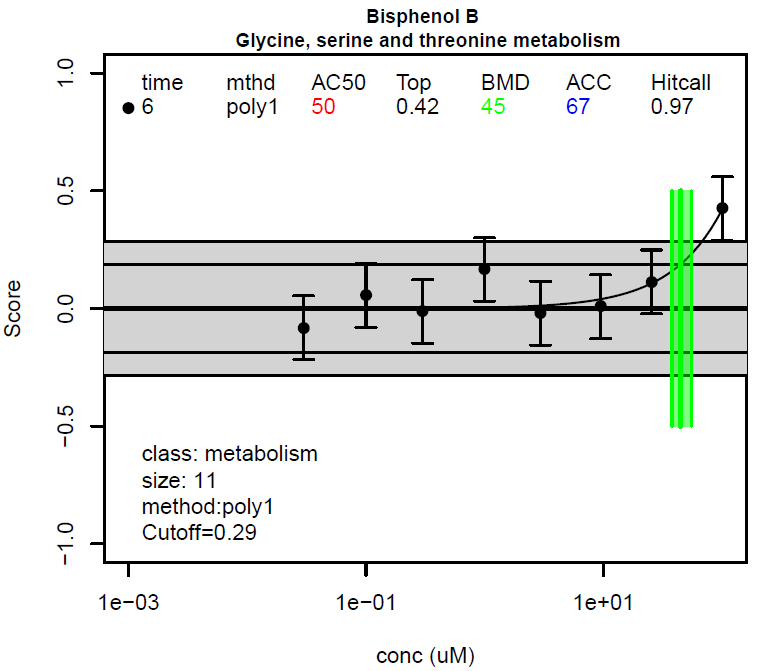


Figure 5: Example model fit with BMD and its 90% confidence interval in green. The gray area is the cutoff range and the black horizontal line inside the gray area is the BMR.



Figure 6: Example fit where BMDU cannot be calculated

## Continuous Hitcalls

Whereas discrete hitcalls classify all assays as either a hit or a miss, continuous hitcalls seek to quantify the strength of hits and identify borderline cases. One excellent, if computationally intensive, approach is to use Monte Carlo methods to find the fraction of times an assay is called a hit when its responses are perturbed in accordance with the expected noise14. Unfortunately, with approximately 2,000 chemicals with hundreds or thousands of pathways each, this was not a practical solution in our case. Thus, we developed a method of calculating a hit index, or “continuous hitcall”, based on statistical properties of the fit.

Previously, a fit was called a hit when the curve met three criteria:

1. At least one median response must be greater than the cutoff.
2. The top must be above the cutoff.
3. The winning AIC must be less than that of the constant model.

For continuous hitcalls, the probability of each criteria occuring was computed and the three values multiplied. The first probability was computed by using the error parameter and t-distribution to calculate the odds of at least one response exceeding the cutoff. The second was computed in a similar method as the BMD bounds, by substituting one parameter with the value that would make the top equal the cutoff (see Appendix A for substitutions) and using the likelihood ratio to compute the one-sided probability of the cutoff being exceeded. The third was set to be the Aikaike weight15 relative to the constant model:

where is the AIC of the winning model and is the AIC of the constant model.

This continuous hitcall is guaranteed to fall between zero and one, but it is not rigorous enough to be interpreted as a probability itself, especially since, for instance, the correlation between the three probabilities was not considered. Figure 7 shows how the continuous hitcall compares to the discrete hitcall for an example run. Note that a continuous hitcall of .5 does not correspond to a case where about half of the assays would be called hits; the boundary is instead closer to .4. However, the plot suggests that there is a reasonable, monotonically increasing relationship between continuous and discrete hitcalls.



Figure 7: Relationship between continuous hitcalls and discrete hitcalls for an example concentration response set

# Plots/Results

## Estrogen Results

In one preliminary study, we investigated whether the pathway method was able to reliably predict estrogenicity among chemicals. Because MCF-7 cells should be sensitive to estrogen and because the estrogen receptor has been studied relatively thoroughly, this provided an idealized test case of whether our methods are generating reasonable conclusions. We used an integrated estrogen receptor model16 as the standard to compare out results against. In that model, each chemical was assigned an agonist and antagonist AUC from 0 to 1 to reflect its potency. Any chemical with either we considered ER active, or “positive”. Chemicals with no activity in any assay were assigned a pseudo AC50 median of 1,000,000; and it was these chemicals we took to be ER inactive or “negative”. We used all 76 ER active chemicals from the screen and randomly selected 76 ER inactive chemicals as well. The pseudo AC50 median was used as a benchmark against which to test our prediction accuracy.

For our estrogen predictions we used unshrunk phase 1 data and a handful of estrogen pathways. The “Dutertre Estradiol Response 6HR UP” pathway consistently outperformed other pathways, so it was chosen as the main benchmark. Multiple plots were generated for different choices of p-value cutoffs: 0.2, 0.1, 0.05, 0.01 0.005, 0.001. Figures 8 and 9 correspond to the default setting: .05, which is also near the peak of balanced accuracy. Using discrete hitcalls, as in Figure 8, the balanced accuracy (“BA”) was calculated in the usual way and the root-mean-square-error (“RMSE”) and coefficient of determination (“R2”) were only calculated for true positives. Only true positives were plotted, and the coloring indicates the winning model.

In the continuous hitcall version, Figure 9, all positives are plotted, with the color indicating a hitcall range with the given upper bound. The lower bound is the next lowest number in the legend (<1 corresponds to , <0.99 corresponds to , etc.).   
The number of true positives was calculated by summing the predicted hitcalls for all positive chemicals while the number of true negatives is calculated by summing for all negative chemicals. These values were then used to compute balanced accuracy. The RMSE used a mean of square errors weighted by hitcall instead of an unweighted mean. R2 likewise used weighted means to compute the total sum of squares and residual sum of squares.

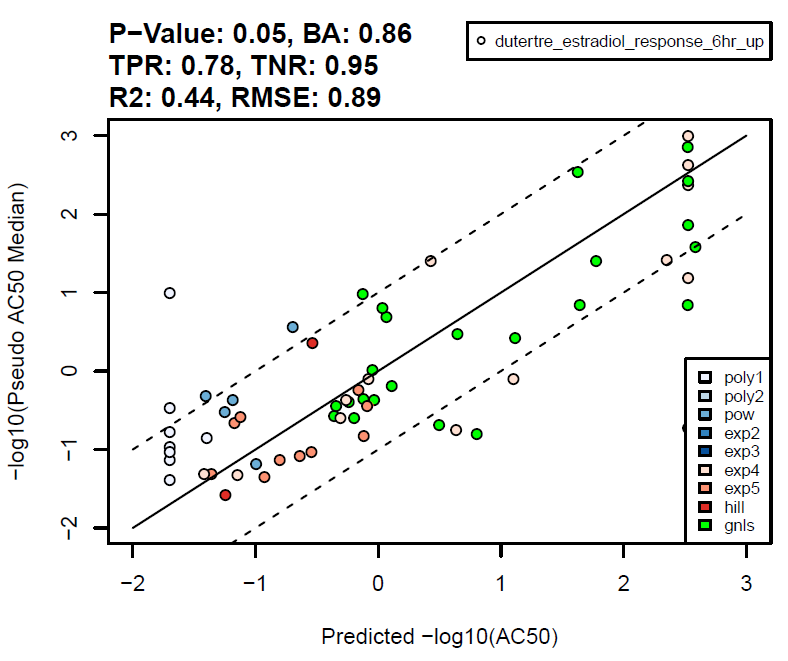


Figure 8: Estrogen detection using discrete hitcalls; only true positives are plotted

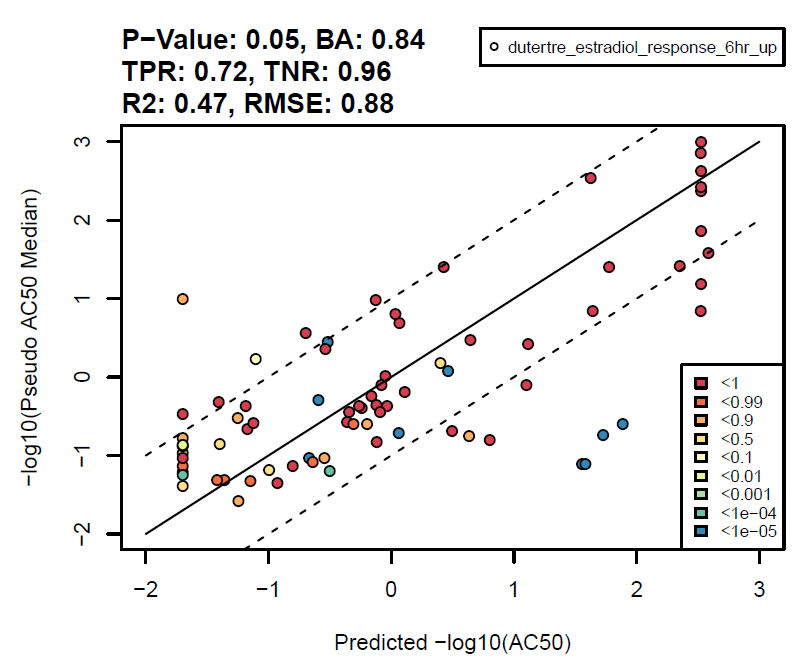


Figure 9: Estrogen detection using continuous hitcalls; only positives are plotted

## BMD Accumulation Plots

Ultimately, we would like to be able to estimate a single point of departure for each chemical that represents the earliest concentration at which activity is expected to occur. We would also like this point of departure to include confidence intervals. To do this, we created a method to aggregate individual pathway BMDs, BMDLs, and BMDUs into an accumulation plot, ultimately yielding an accumulation BMD, BMDL, and BMDU. First, in order to smooth the accumulation curve and account for the uncertainty of individual BMDs, each pathway BMD was modeled as a gaussian with its mean equal to the BMD and its standard deviation set such that the BMDL lied at the 5th percentile. If the BMDL could not be calculated, the standard deviation was set such that the BMDU lied at the 95th percentile. If neither could be calculated, either the 95th percentile was set to be 100 or the 5th percentile was set to .001, according to whichever value was farther from the BMD, resulting in a very wide distribution. The cumulative distribution functions of these individual gaussians were then weighted by their hitcalls and added to form an accumulation plot. The black curves in Figures 10 -13 are examples using discrete hitcalls, so that each hit corresponds to one gaussian of area 1 and non-hits are ignored. The black curve in Figure 14 is an example using continuous hitcalls, where every pathway contributes a gaussian to the plot with an area equal to the hitcall.

We then compared these curves to the curves that would be generated by noise alone. We began by generating null data as described in the Noise Estimation section. Then we randomly assigned the concentrations of actual chemicals to groups of this null data to create null chemicals with similar concentration patterns as the actual data. 125 of these null chemicals with approximately eight concentrations apiece were run through the exact same concentration response process as the actual data to generate BMDs, hitcalls, etc. These null chemicals then each used to generate an accumulation curve; the grouping of 125 was used to determine the mean, 5th percentile, and 95th percentile null accumulation response at each concentration plotting point. These form the red curves in Figures 10 -13, which are the same for every chemical. In order to determine the accumulation BMD, we used a method analogous to that found in the Benchmark Dose Technical Guidance document10: the accumulation BMD was set at the smallest concentration at which the accumulation curve exceeded the mean null curve by 10% of the mean null curve’s maximum value. Likewise, the accumulation BMDL and BMDU were set at the lowest concentrations at which the accumulation curve exceeded the 5% and 95% curves, respectively, by 10% of their maximum values.

Figures 10-14 offer some typical examples of the results of this process; note that the dashed black line denotes the minimum tested concentration, while the maximum tested concentration is typically at 100. Figure 10 is an example of the ideal case, where the accumulation curve grows quickly to a large maximum, resulting in a tight estimation of the accumulation BMD. Figure 11 is a case where the accumulation curve grows along with the noise and then flattens, resulting in an accumulation BMD range the spans the entire range of tested concentrations. Figure 12 is another common case where the accumulation BMD lies roughly at or below the minimum tested concentration. Figure 13, on the other hand, is an example of a chemical that is apparently inactive. Cases such as these are interesting in that the accumulation curve has fewer total hits than any of the null chemicals, implying that the null data might be overestimating the noise. Lastly, Figure 14 is an example of using continuous hitcalls with the same data as in Figure 10. The results are largely the same, except the continuous case registers BMDs above the highest concentration while the discrete case ignores these because none are hits.

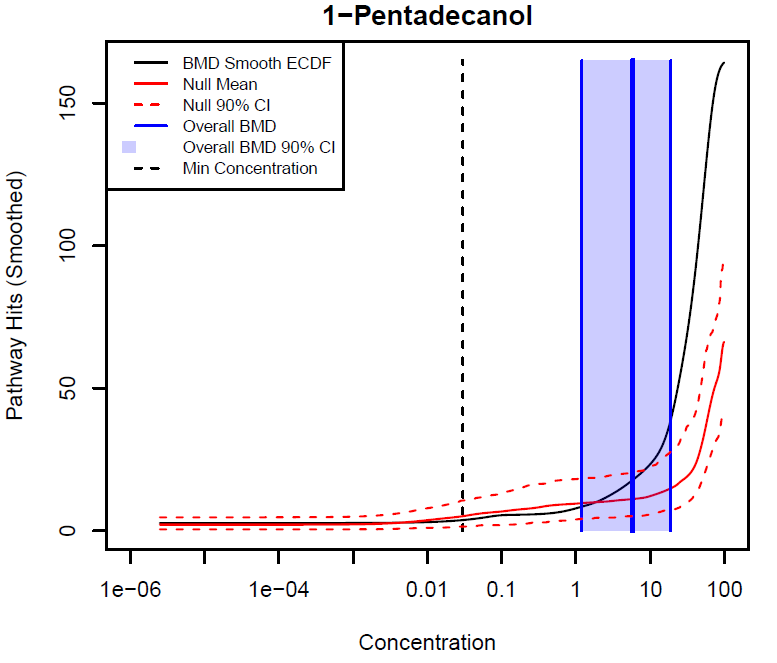


Figure 10: BMD accumulation plot in the ideal case

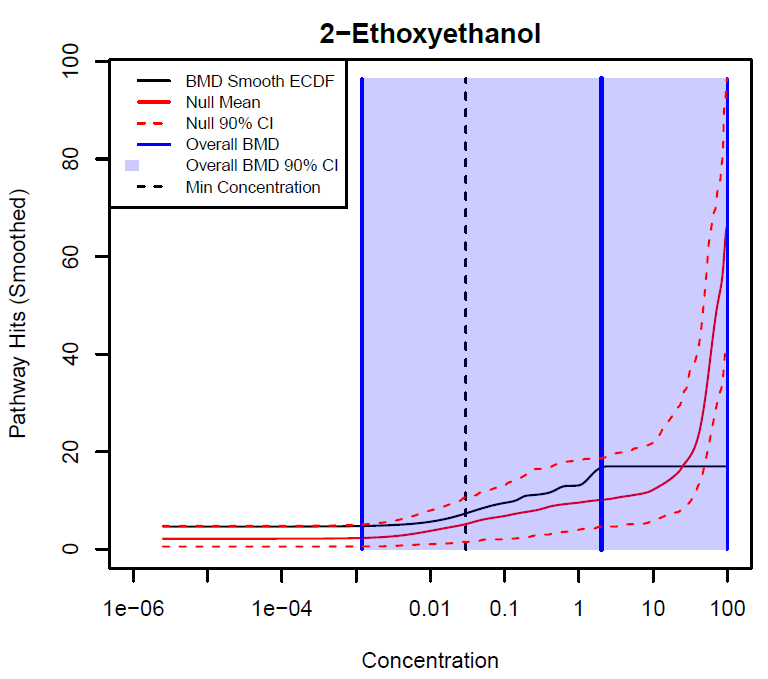


Figure 11: BMD accumulation plot in the inconclusive case

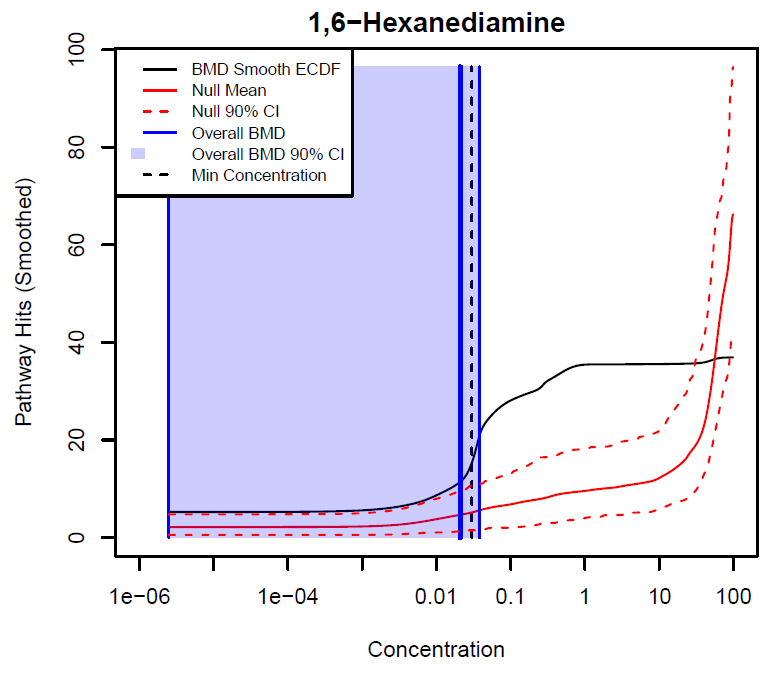


Figure 12: BMD accumulation plot when activity begins at or below the minimum concentration

## 

Figure 13: BMD accumulation plot in the inactive case

## 

Figure 14: BMD accumulation plot using continuous hitcalls in ideal case

## Replication Study

Each of 43 chemicals used in the pilot study were also present in the phase 1 screen, so we were able to compare results across the two studies using different processing techniques. We compared only probes and pathways that met the size requirements in both replicates. We also checked estrogen receptor detection accuracy on the 8 active and 5 inactive chemicals identified in this set using the same criteria as in the Estrogen Results section. The main difference in processing methods had to do with “shrinkage”; namely, which statistical model was used to control outliers and convert probe counts into l2fc. The five methods were: no shrinkage (“none”), an older method of DESeq2 shrinkage (“normal old”), a newer method of DESeq2 shrinkage (“normal”), shrinkage using the ashr package (“ashr”), and shrinkage using the apeglm package (“apeglm”). Additionally, the count flooring was set to either 5 or 10 and plate effect correction was set to either on or off. Testing every combination of these settings yielded 40 different processing methods. After processing, pathways were computed using both the FC and GSVA methods; additionally, concentration response curves were also calculated for individual probes.

Figures 15-26 all use the FC pathway method. Figure 15 shows the fraction of pathways across all chemicals in both replicates (“Combined Hit Fraction”) that exceeds a given continuous hitcall. Continuous hitcall was varied as in a sensitivity versus specificity tradeoff, where anything above that threshold was called a hit and anything below a miss. The legend format is “flooring\_plate effect on/off\_shrinkage method”. The combined hit fraction is steady at around .2 for all methods when hitcalls are shown on a linear scale. Continuous hitcalls are best thought of as an ordering index, rather than as a meaningful number on its own because their values are intrinsically linked to the noise estimation cutoff (here, the p-value was set to 0.05). A higher, more stringent cutoff would cause the hit fraction to level off at a lower value, say 0.1; conversely, a lower cutoff would cause the hit fraction curve to move upwards. Thus, the more natural scale for further plots is the combined hit fraction, which has a clearer meaning, and which we would expect to give more stable answers across different cutoffs. They would be exactly stable if the *order* of hitcalls between different hits was unchanged; in practice, the order can change for different cutoffs, but it should not change dramatically.

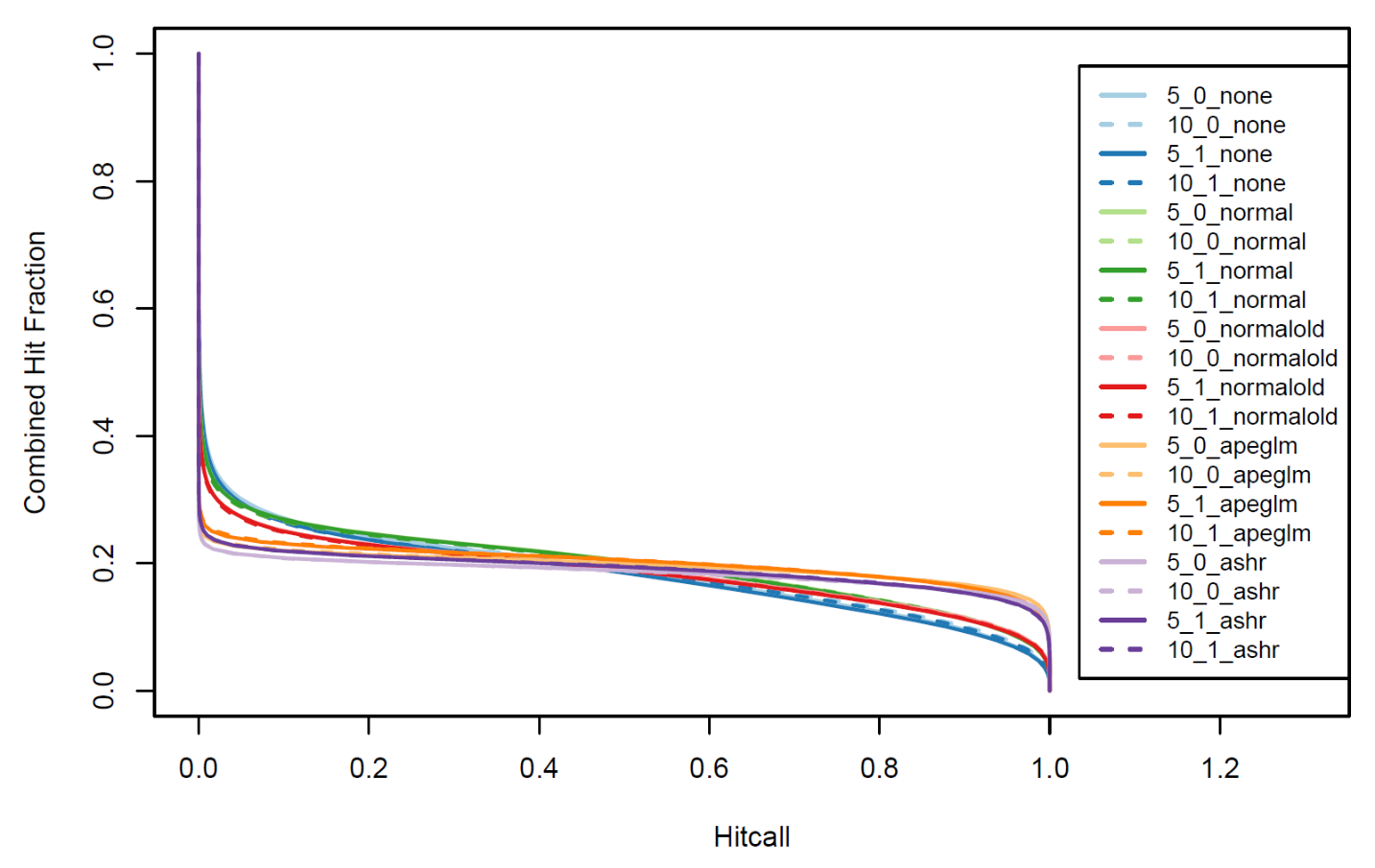


Figure 15: Fraction of pathways in both replicates that are hits when setting the continuous hitcall threshold to a given value

Figure 16 now plots the fraction of hitcalls that match against the combined hit fraction, which is varied by varying the continuous hitcall threshold. For a given combined hit fraction, the fraction of hits and misses that match between replicates are called the “Fraction of Hitcalls Matching”. When the combined hit fraction is 0.0, every curve in both replicates is a miss, so every hitcall matches; likewise, when the combined hit fraction is 1.0 every curve is called a hit and the fraction matching is again 1.0. The black curve in the plot represents the theoretical curve that would result from hits being evenly distributed between replicates and chosen completely at random: . So, the methods perform better than random chance, but no method is clearly better than the others.

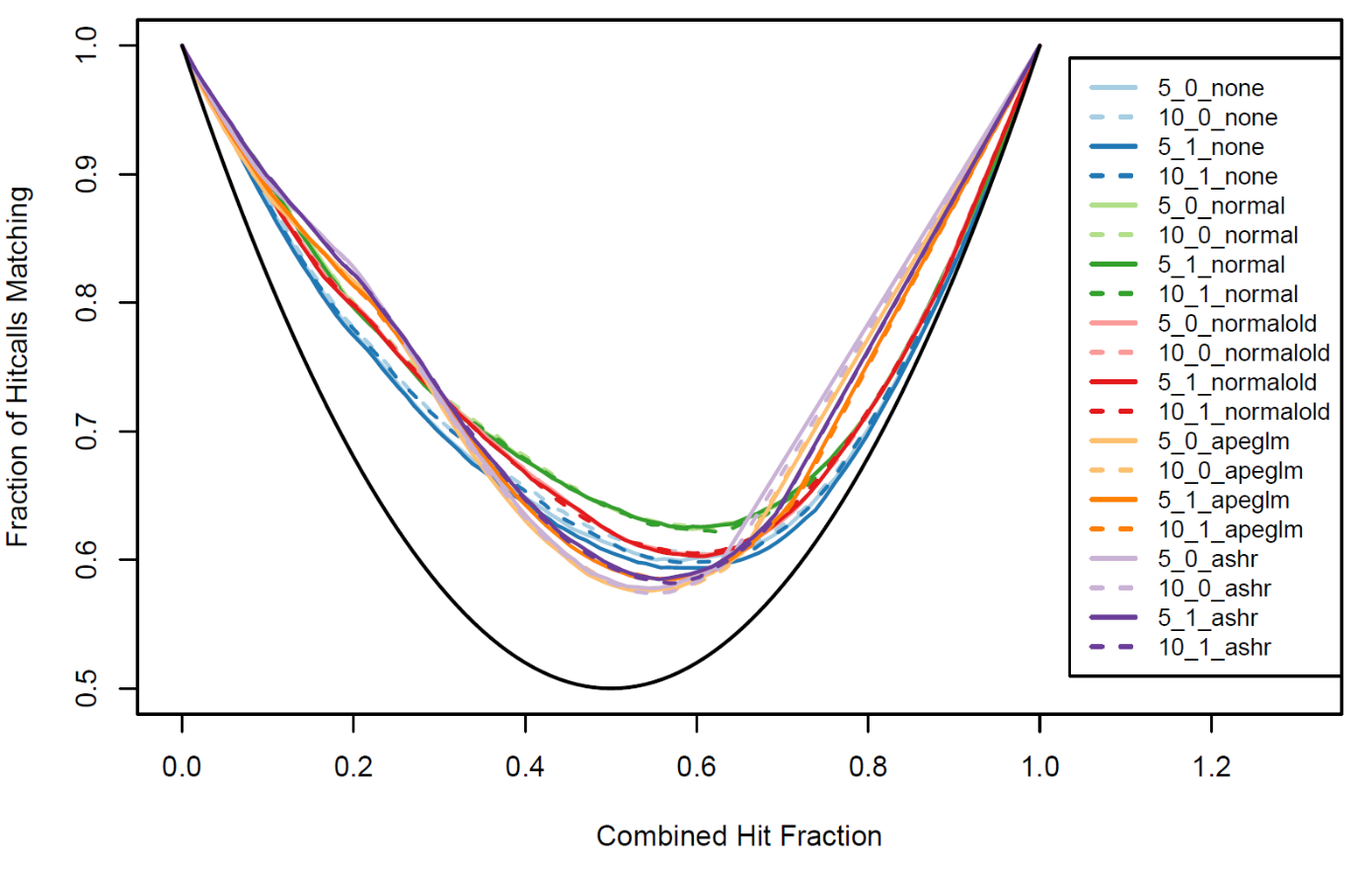


Figure 16: Fraction hitcalls that match between replicates versus combined hit fraction

Figure 17 shows a simple bar graph representing the Pearson correlation between continuous hitcalls in each replicate. This has the benefit of simplicity, but may have only limited value because the calculation is greatly affected by the choice of cutoff. Figure 18 shows the Spearman correlation instead, which appropriately emphasizes the hitcall rank instead of the hitcall value. While the Pearson correlation suggests ashr replicates best, the Spearman correlation instead points to normal shrinkage.

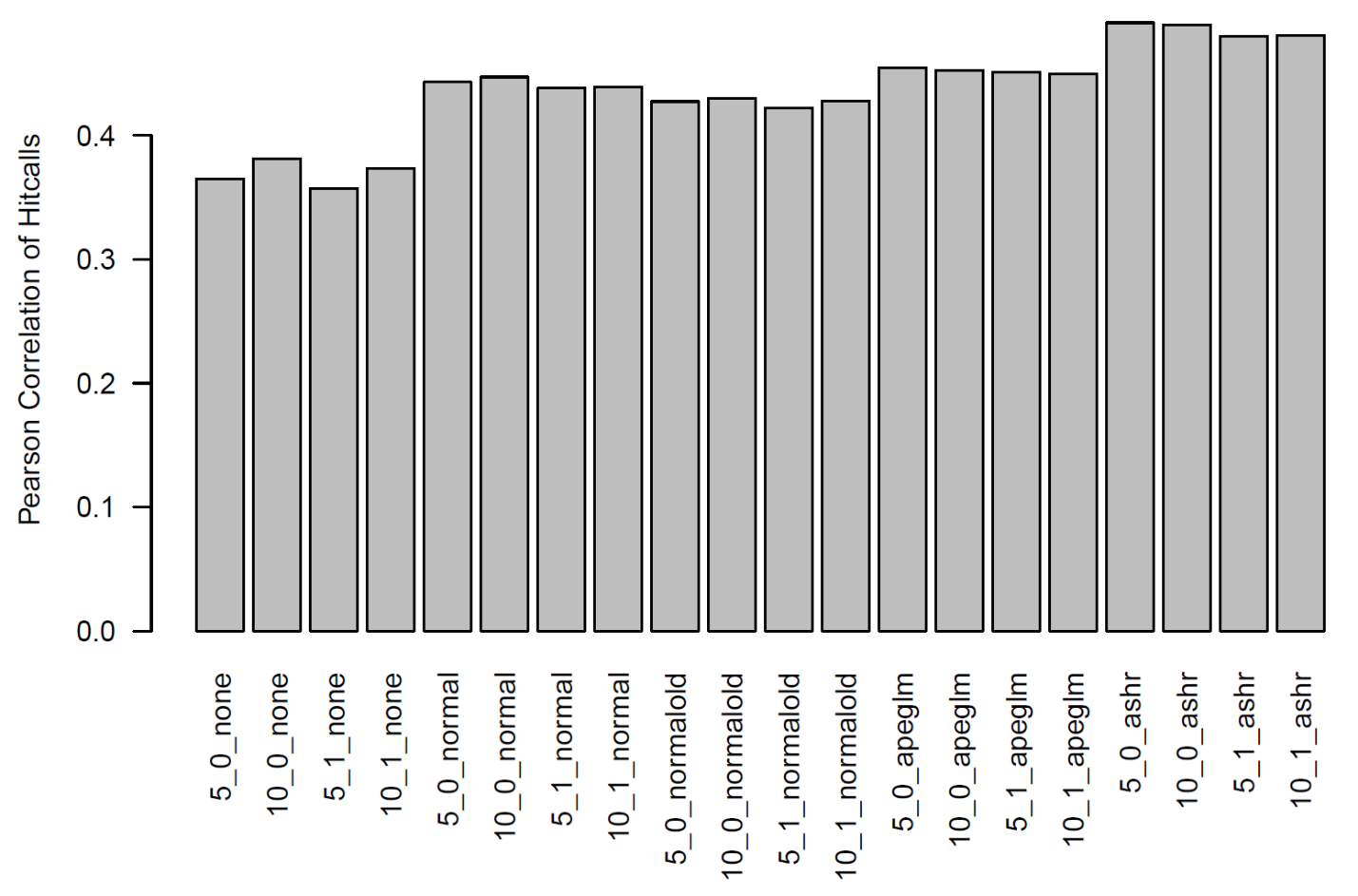


Figure 17: Pearson correlation of hitcalls with FC pathway scoring

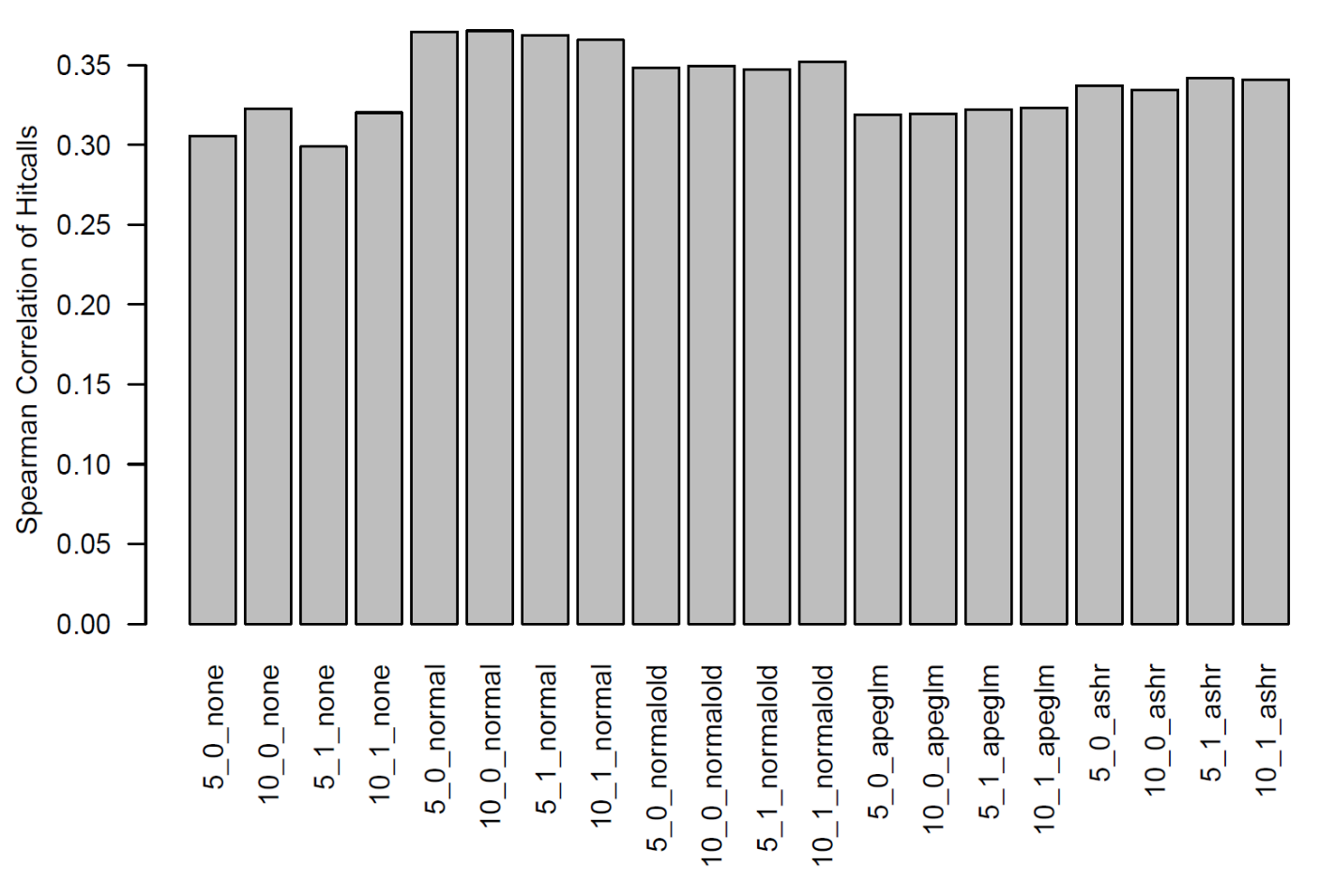


Figure 18: Spearman correlation of hitcalls using FC scoring

Another metric we considered is how well points of departure replicate. In most situations we would not know which hits in a given study would be replicated and which would not, implying that we should consider points of departure that were a hit in either replicate. However, points of departure in curves that are not hits are often pushed to some default minimum or maximum, leading to many computational artifacts. So, in Figures 19 and 20 we only considered curves that were hits in both replicates (for a given combined hit fraction). Ashr and apeglm shows the bets replication in both plots, with normal shrinkage only slightly worse. Flooring and plate effect seemed to have minimal impact on the outcome.

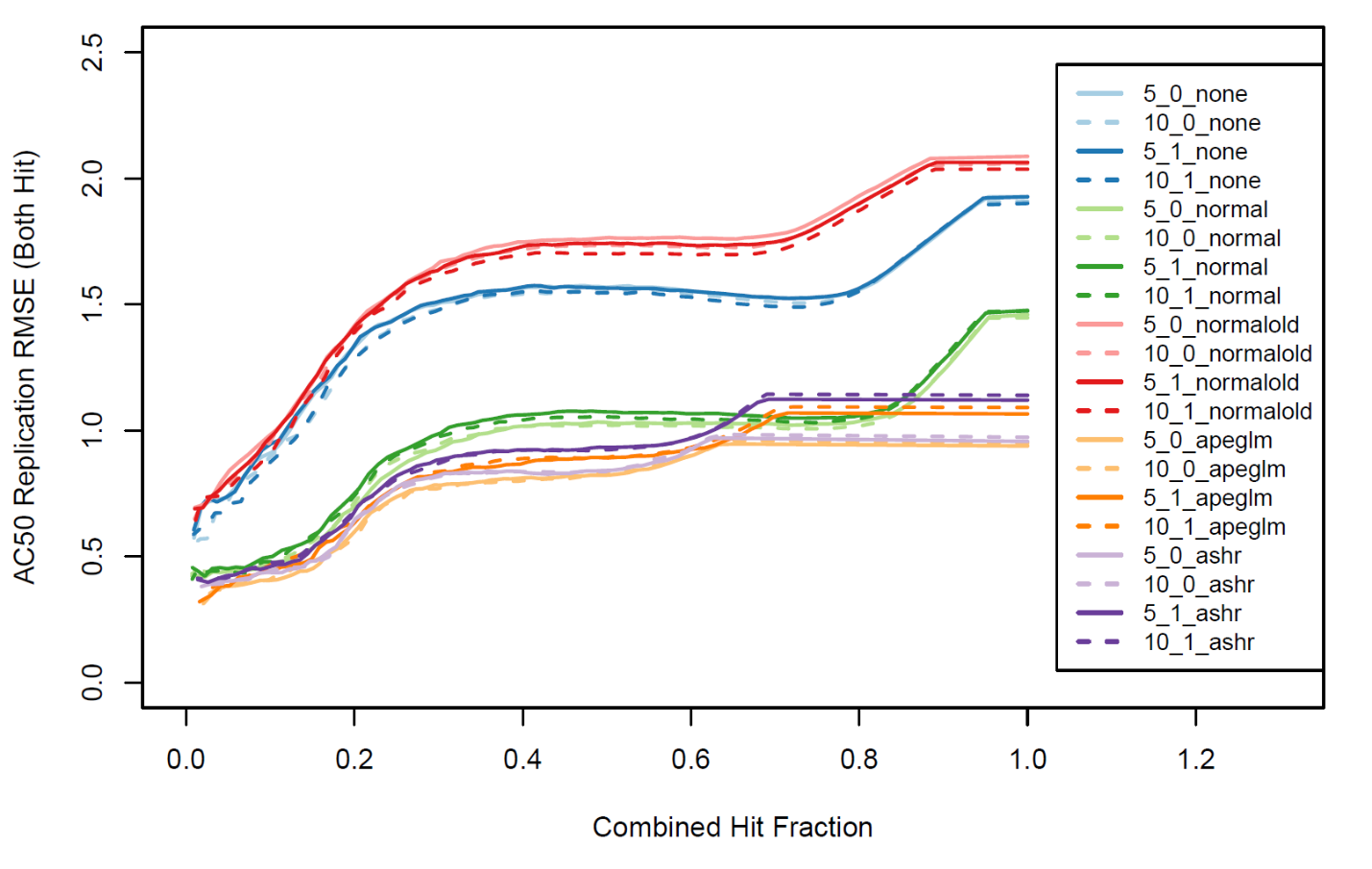


Figure 19: RMSE between AC50s for pathways that hit in both replicates using FC scoring

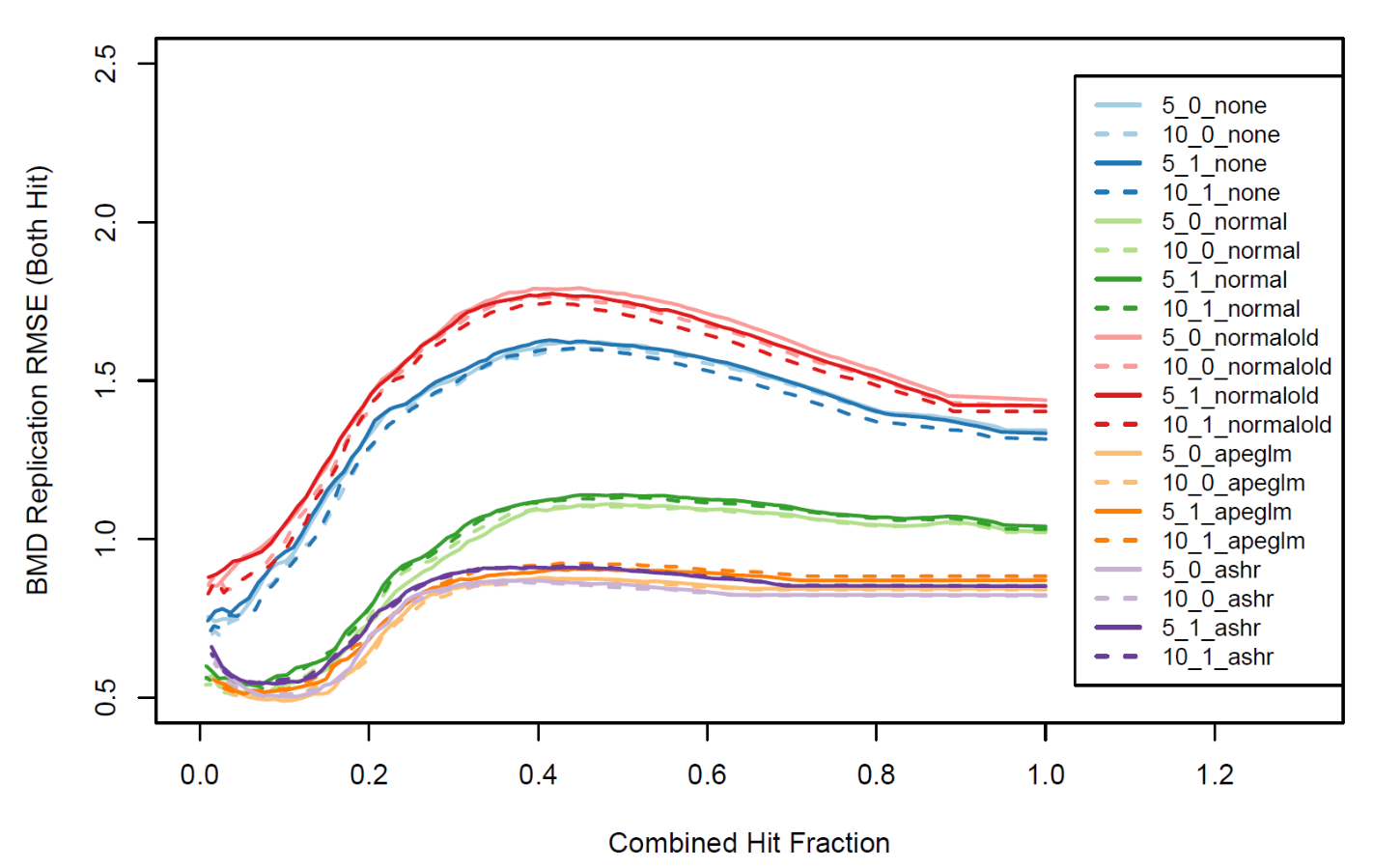


Figure 20: RMSE between BMDs for pathways that hit in both replicates using FC scoring

Estrogen detection again used a sliding continuous hitcall threshold to determine the tradeoffs between sensitivity and specificity. Since there were only 8 active and 5 inactive chemicals present, we used four pathways to make estrogen predictions: Dutertre Estradiol Response 6HR UP, Hallmark Estrogen Response Early, Hallmark Estrogen Response Late, and Ryan Estrogen Receptor Alpha UP. This was done to lessen the impact of any one pathway/chemical combination on the results. Figures 21 and 22 show the Receiver Operating Characteristic (ROC) curves for the pilot and phase 1 screen separately. Interestingly, while all methods did a reasonably good job of distinguishing estrogens in the pilot, ashr and apeglm performed significantly worse than the others in the phase 1 screen. The combined area under curve (AUC) bar plots in Figure 23 summarizes these results. The balanced accuracy of the various methods in the pilot and phase 1 screen are shown in Figures 24 and 25, respectively, and summarized in Figure 26. These plots are in good agreement with the figures based on the ROC curve.

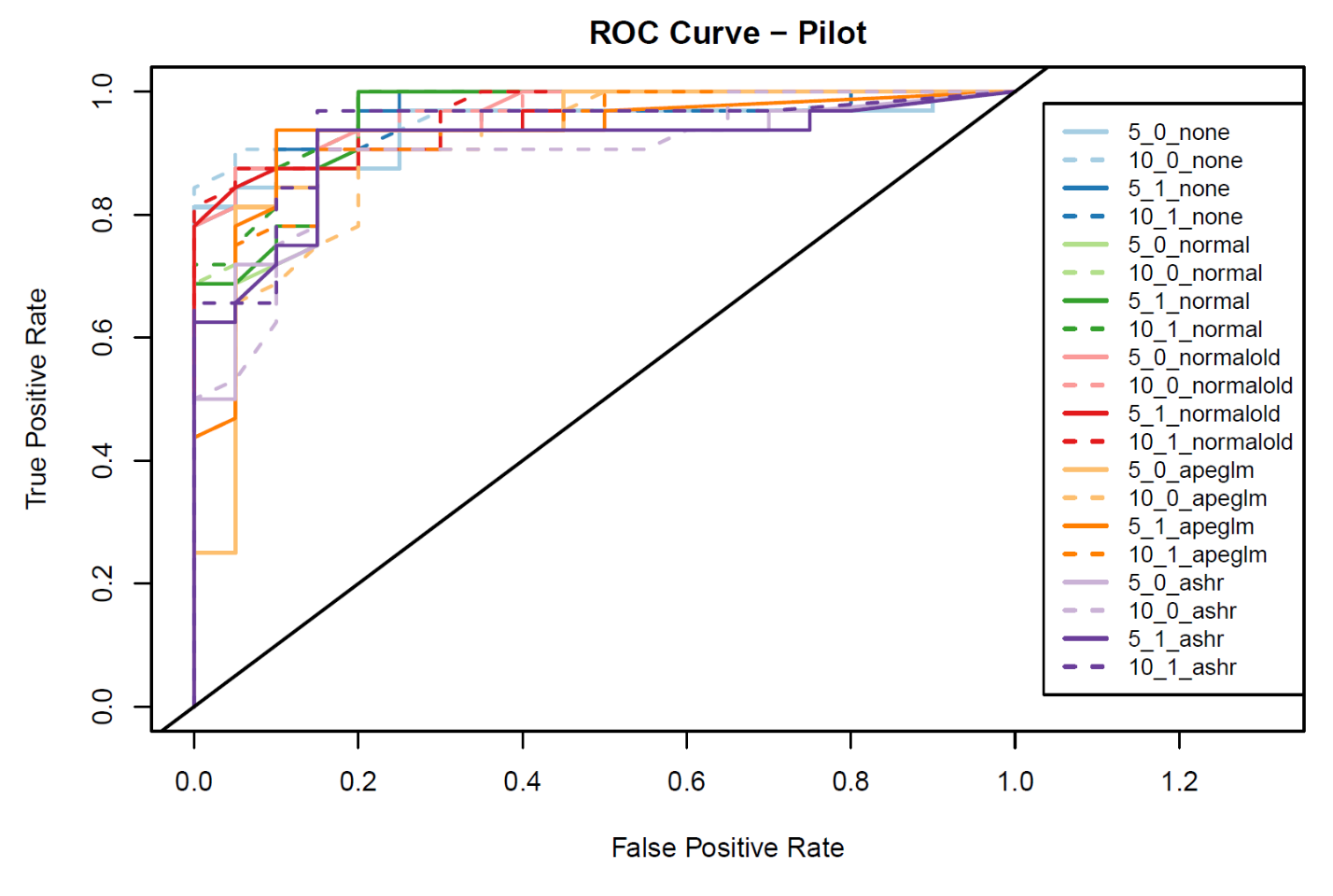


Figure 21: ROC curve for estrogen detection in pilot replicate using FC scoring

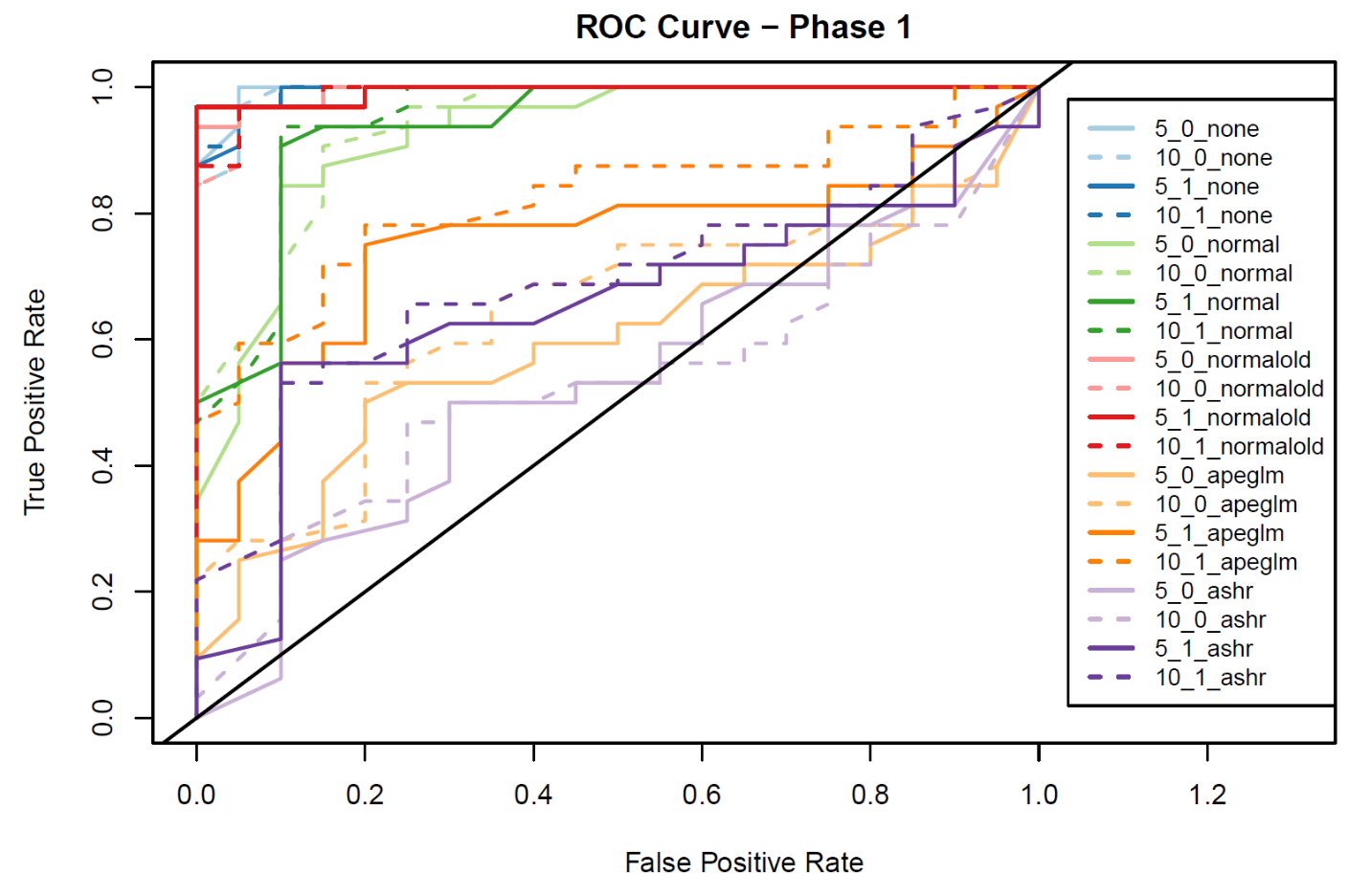


Figure 22: ROC curve for estrogen detection in phase 1 replicate using FC scoring

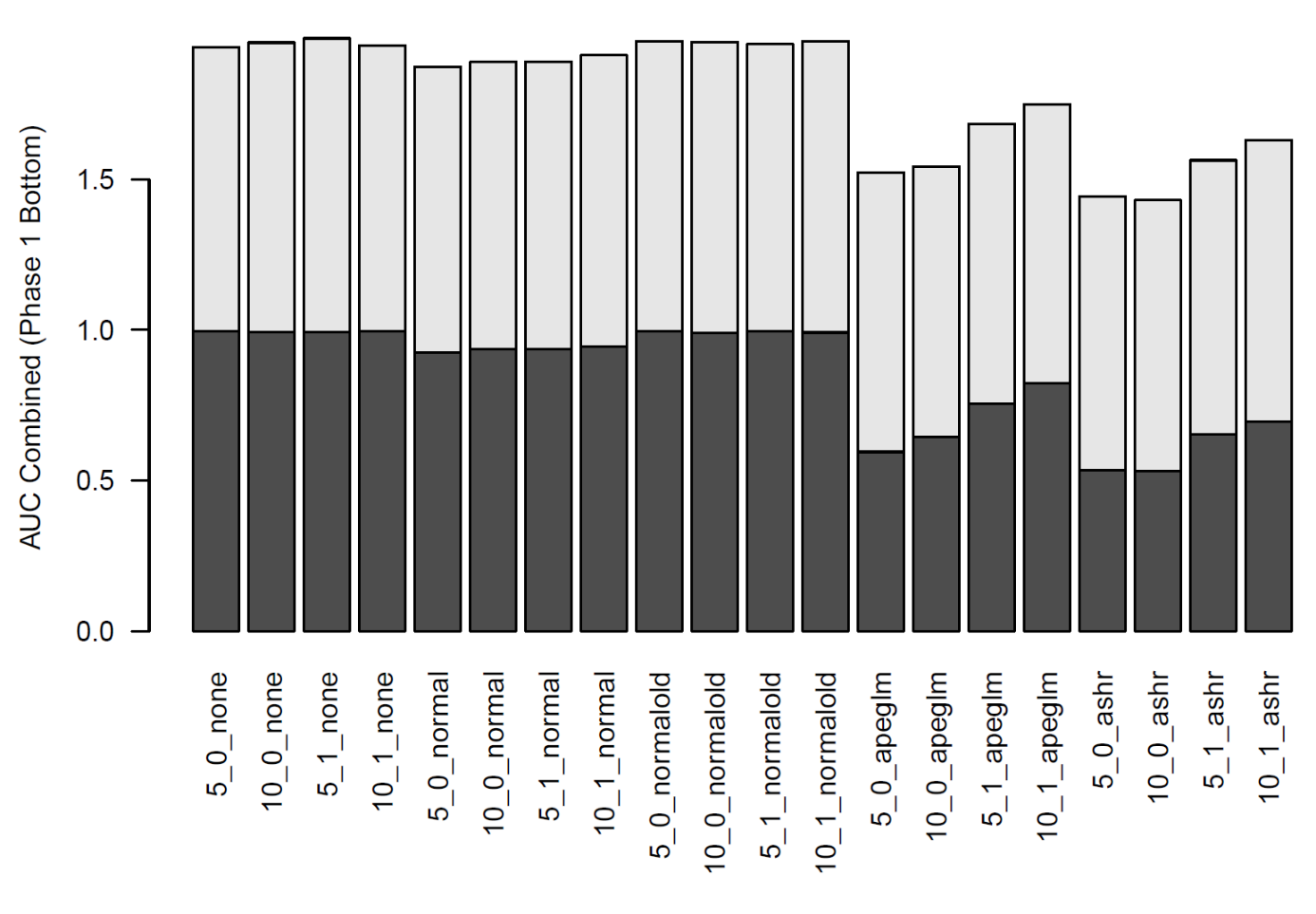


Figure 23: Pilot (top) and Phase 1 (bottom) AUCs for estrogen detection using FC scoring

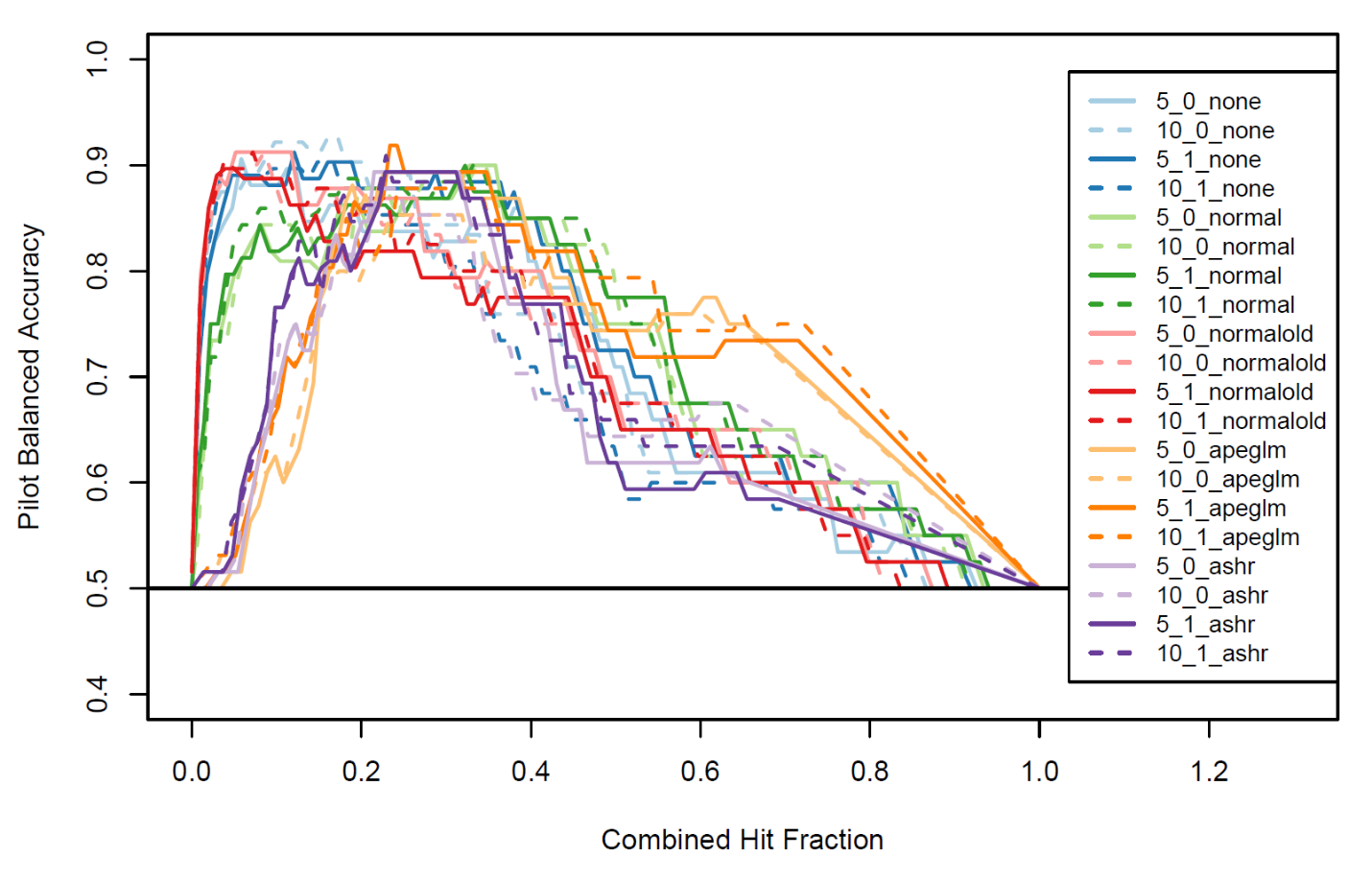


Figure 24: Balanced accuracy of estrogen detection in pilot replicate using FC scoring

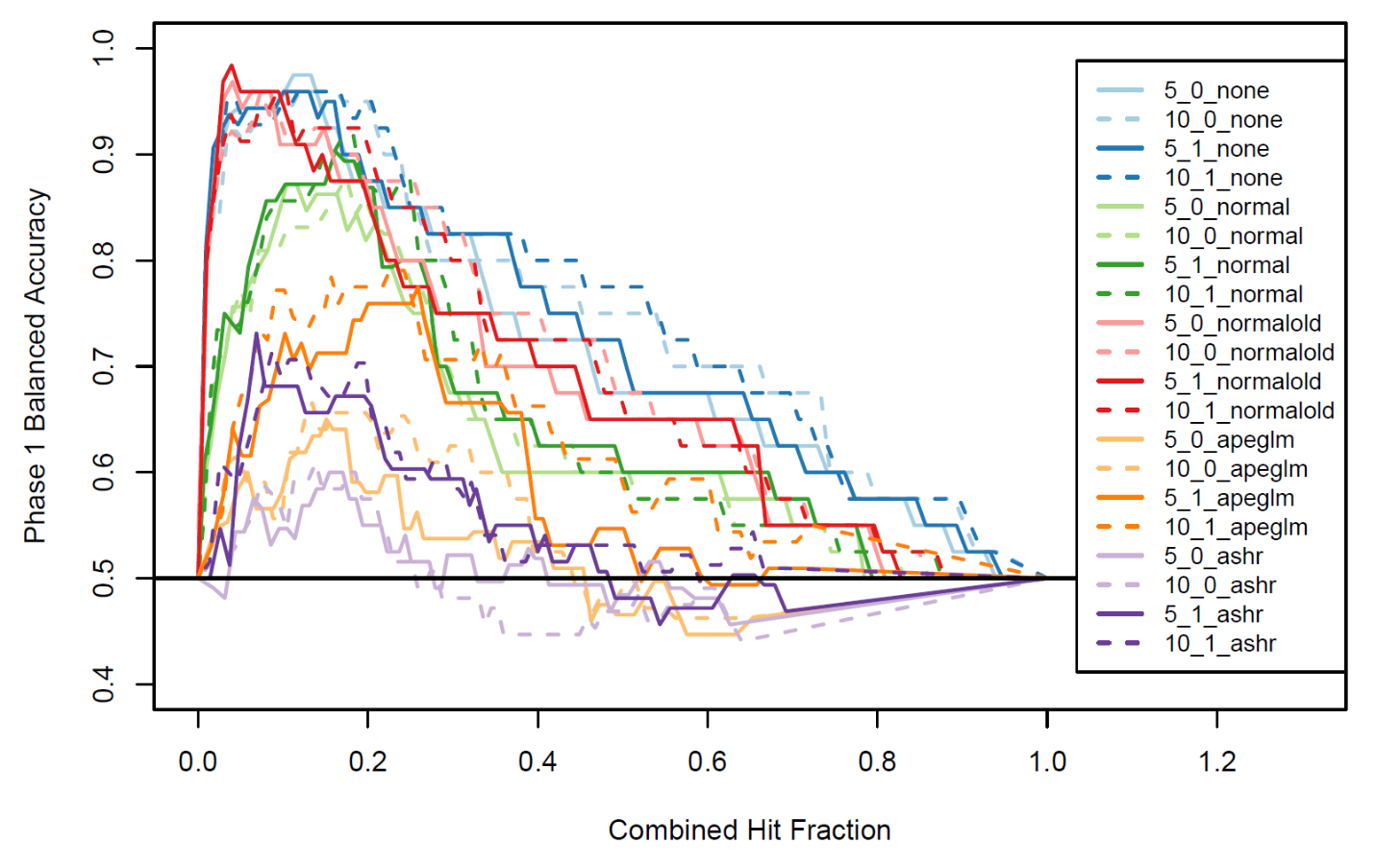


Figure 25: Balanced accuracy of estrogen detection in pilot replicate using FC scoring

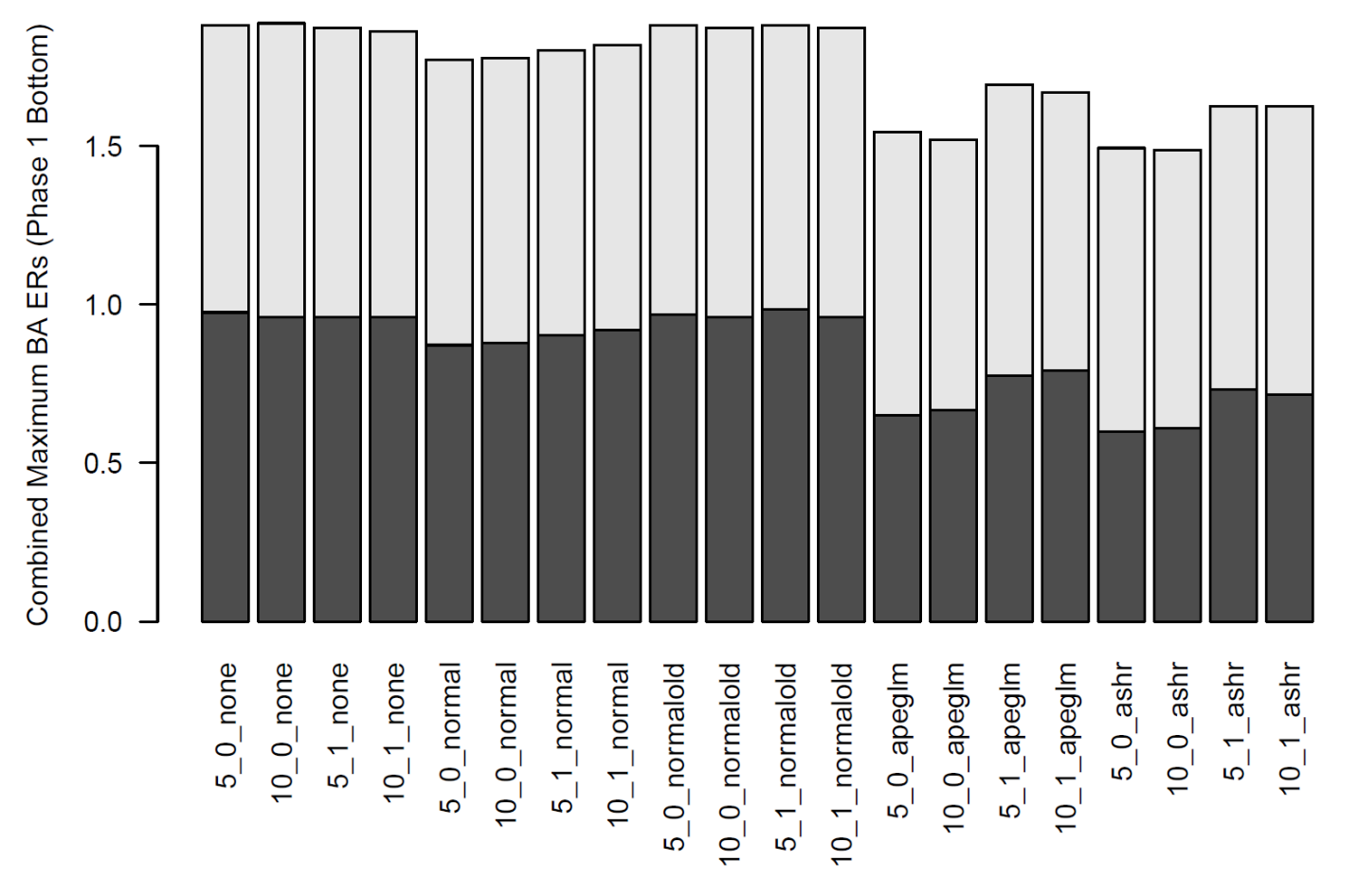


Figure 26: Maximum balanced accuracy for pilot (top) and phase 1 (bottom) replicates using FC scoring

The same plots were generated using the GSVA pathway scoring method and the most important are shown here. Figure 27 shows that the Spearman correlation is around 0.25, lower the maximum for FC, which was about 0.35. The points of departure replication, shown in Figures 28 and 29, verify that GSVA replication is significantly worse across the board for every method. The combined AUCs in Figure 30 show a similar estrogen detection ability as in FC for the best methods: none, normal old, and normal.

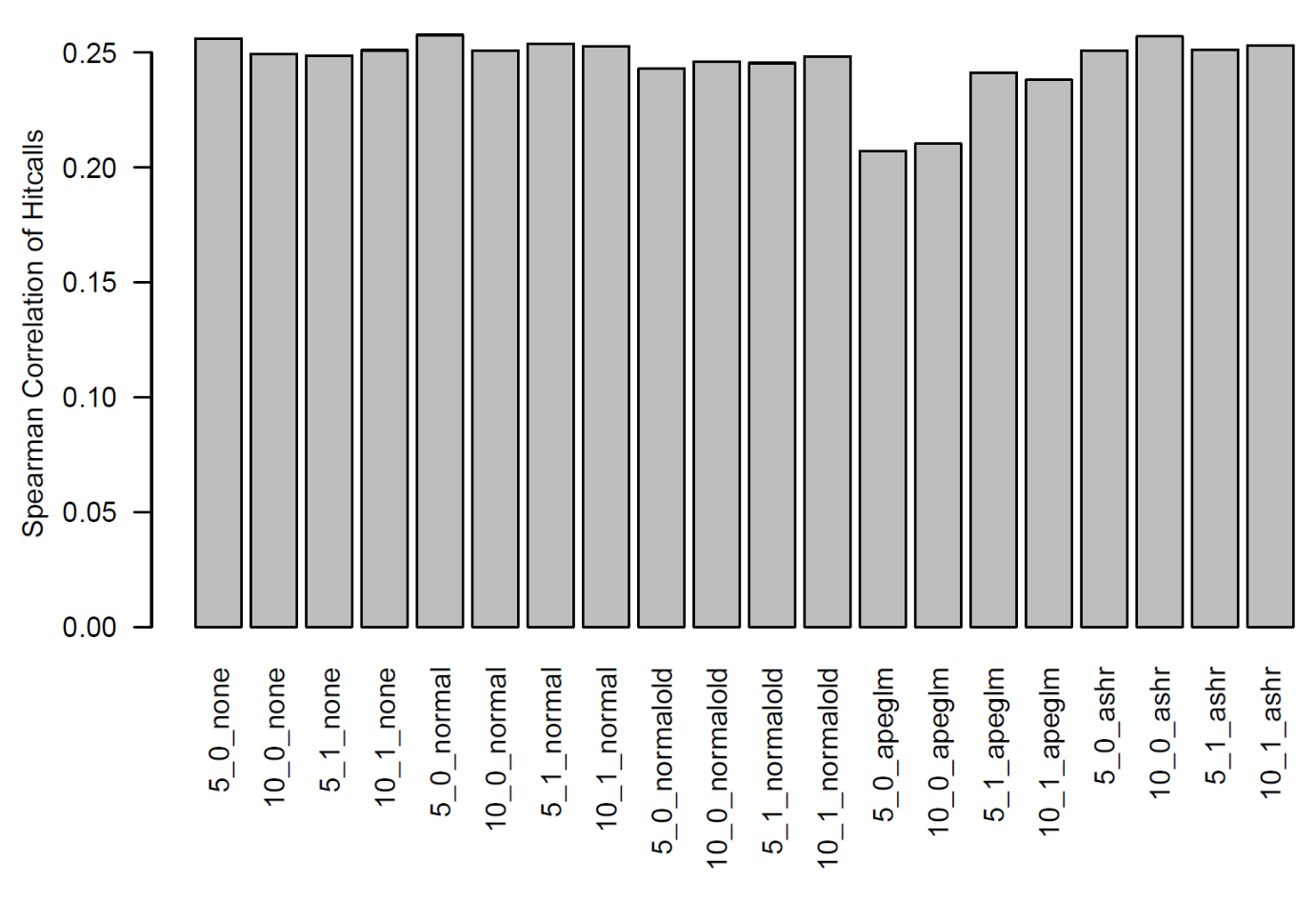


Figure 27: Spearman correlation between continuous hitcalls using GSVA scoring

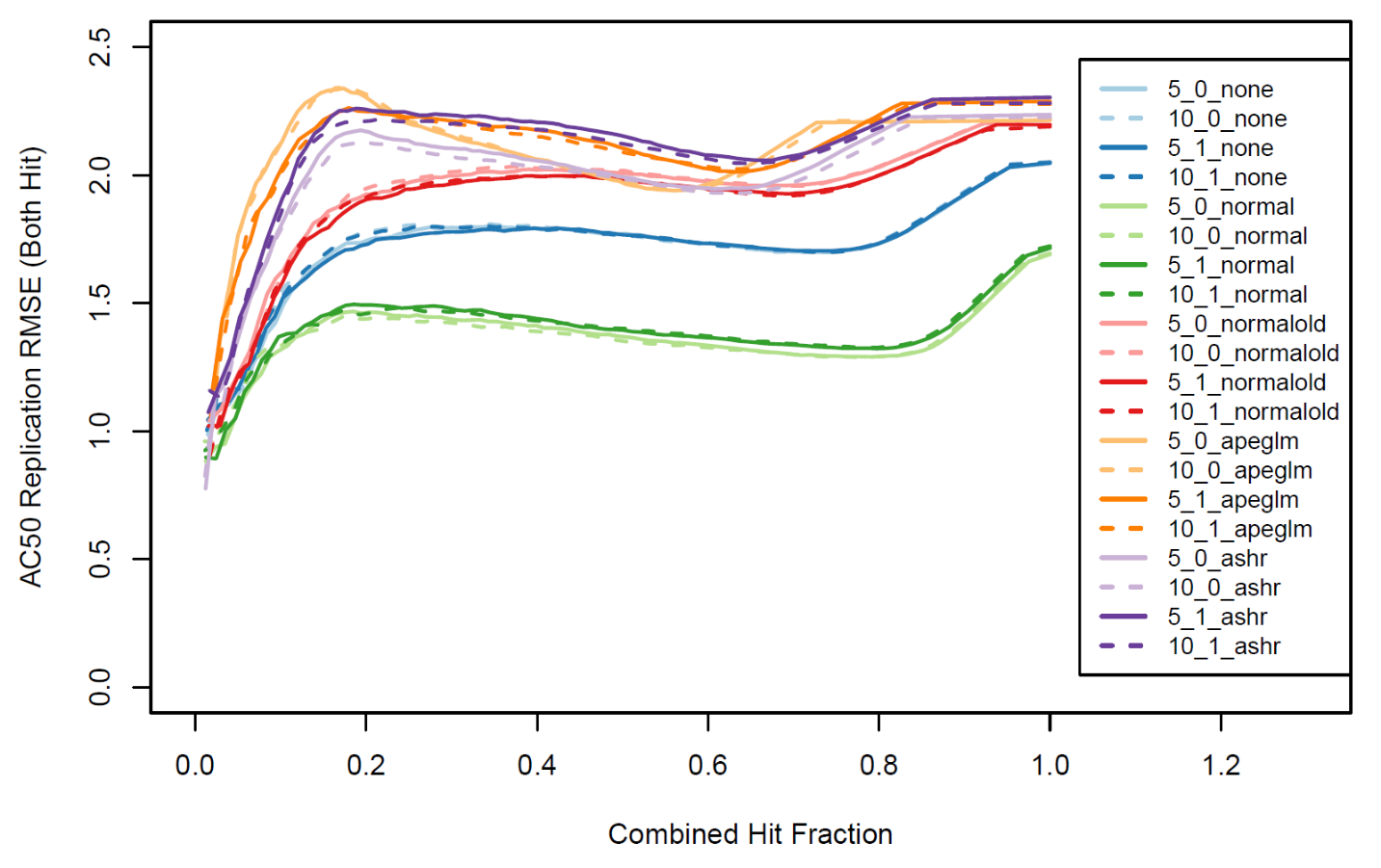


Figure 28:RMSE between AC50s for pathways that hit in both replicates using GSVA scoring

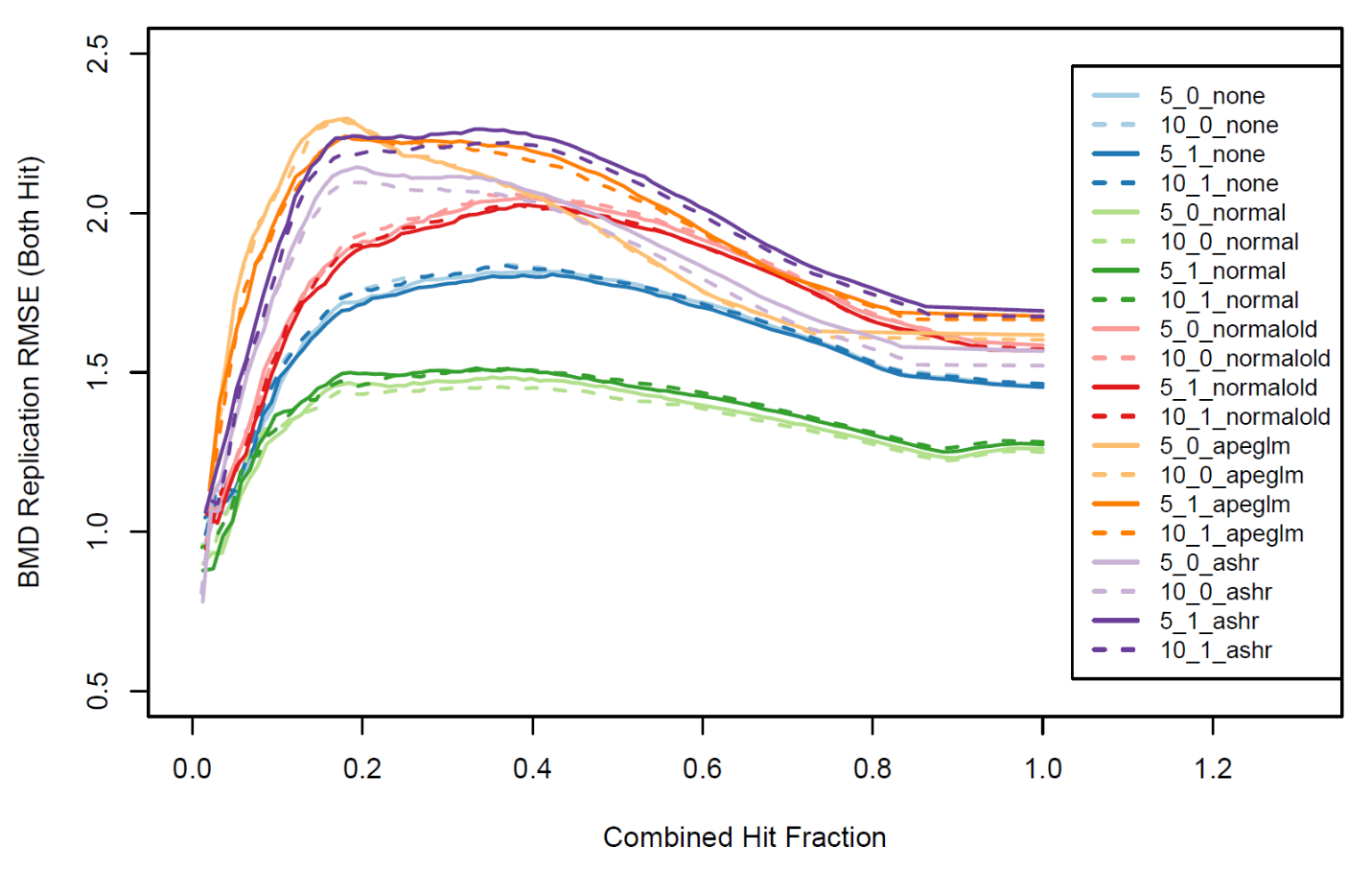


Figure 29:RMSE between BMDs for pathways that hit in both replicates using GSVA scoring

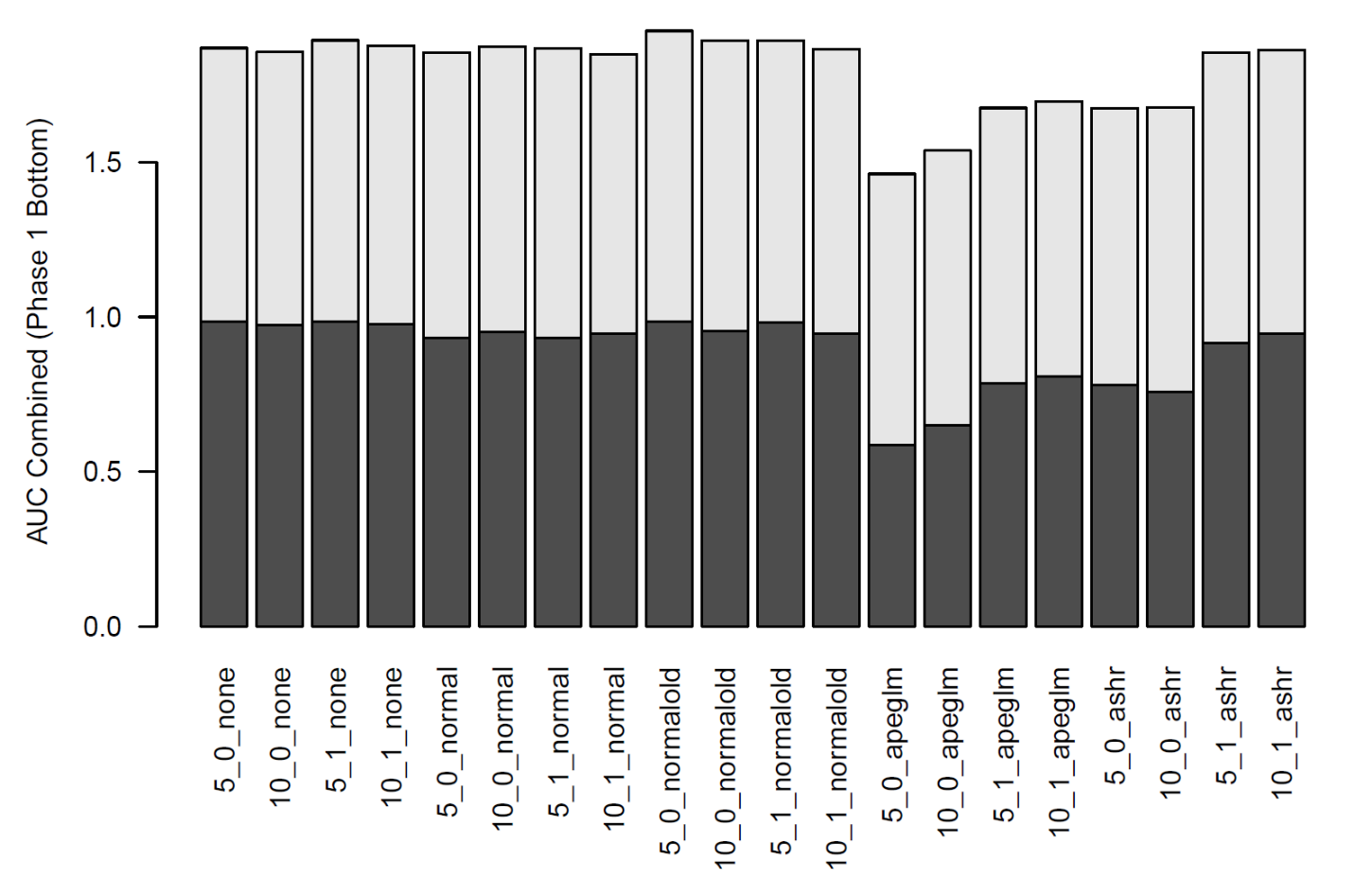


Figure 30:Estrogen detection AUCs for pilot(top) and phase 1(bottom) using GSVA scoring

Lastly, individual probe concentration responses were considered. Figure 31 shows that hitcall spearman correlation is down to a maximum of 0.20 versus 0.35 for FC. Figures 32 and 33 show that point of departure replication is quite like that of the FC method when both replicates are hits. For estrogen detection we used all the probe ids corresponding to genes found in the Dutertre Estradiol Response 6HR UP pathway. The ideal scenario is that all these genes would be hits for an estrogen active chemical and misses for an inactive chemical. It’s not entirely reasonable to expect that this would occur, which is why pathway scores were developed in the first place; nonetheless, the results were instructive. Figures 34 and 35 show the ROC curves for the pilot and phase 1 data, respectively, and Figure 36 summarizes the AUCs. The phase 1 screen again has a more difficult to detect signal, and none of the methods were better than randomly guessing; however, the pilot had a strong enough signal to achieve AUCs up to 0.7 for some methods.

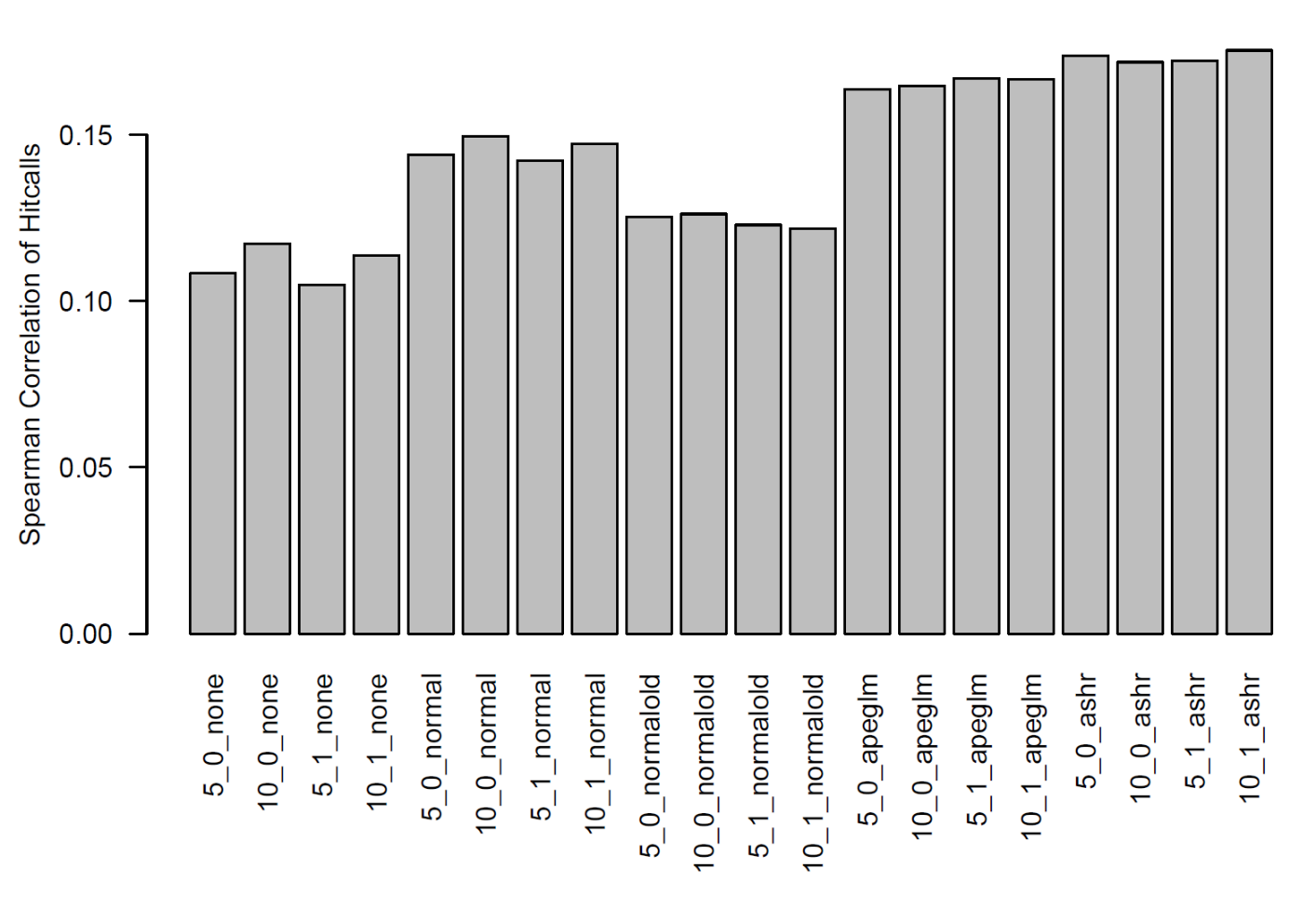


Figure 31: Spearman correlation of continuous hitcalls for individual probes

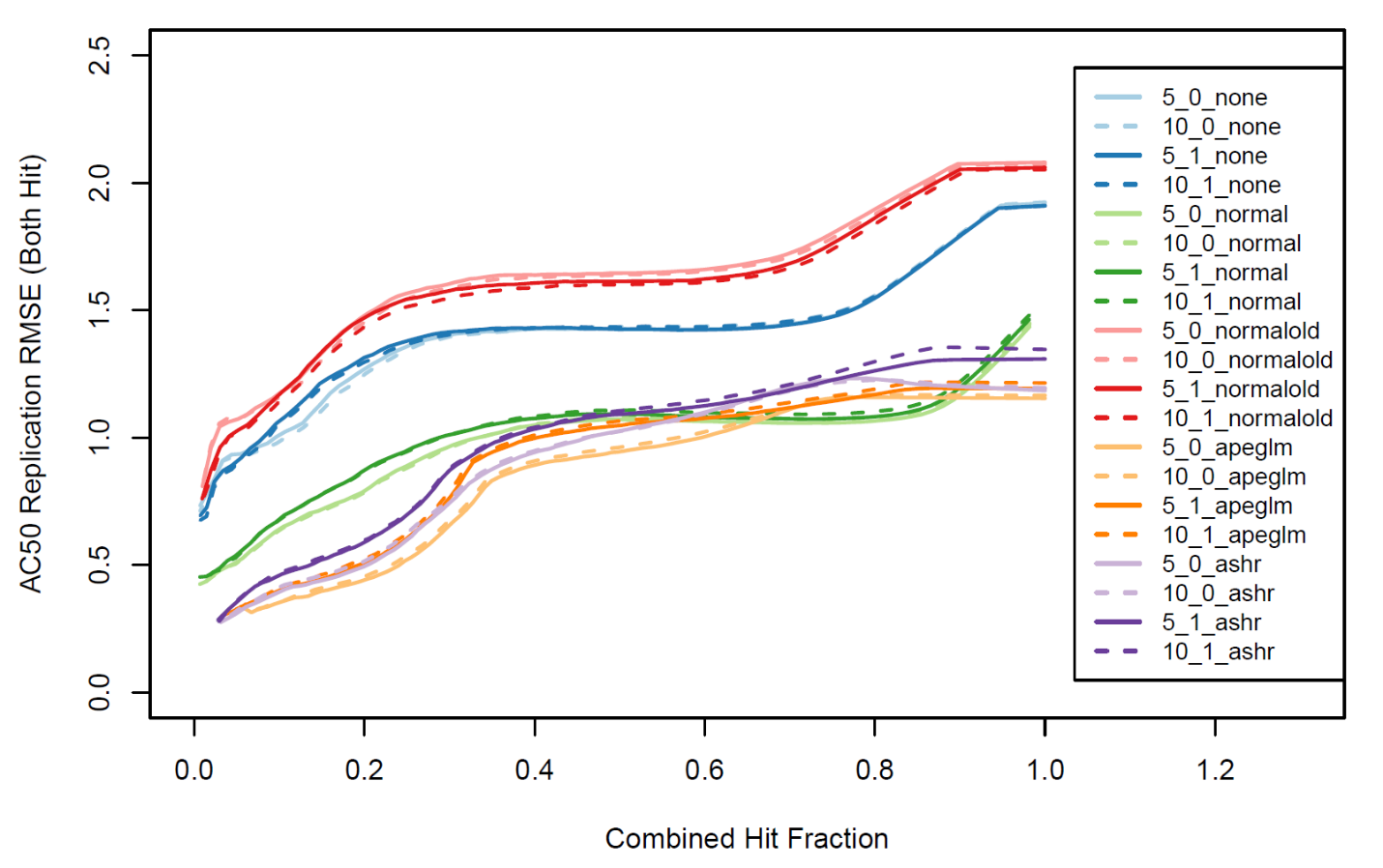


Figure 32:RMSE between AC50s for probe curves that hit in both replicates

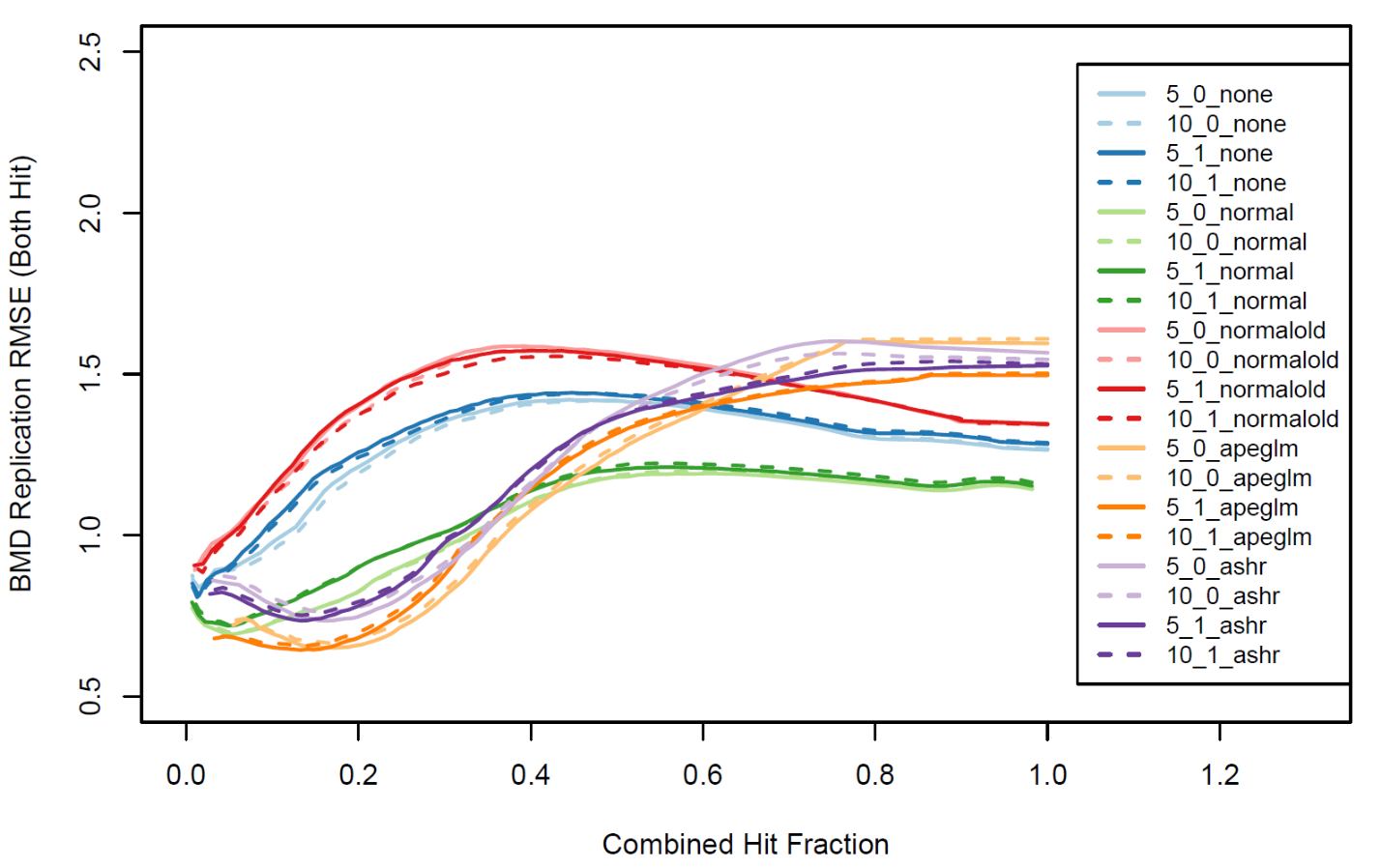


Figure 33:RMSE between BMDs for probe curves that hit in both replicates

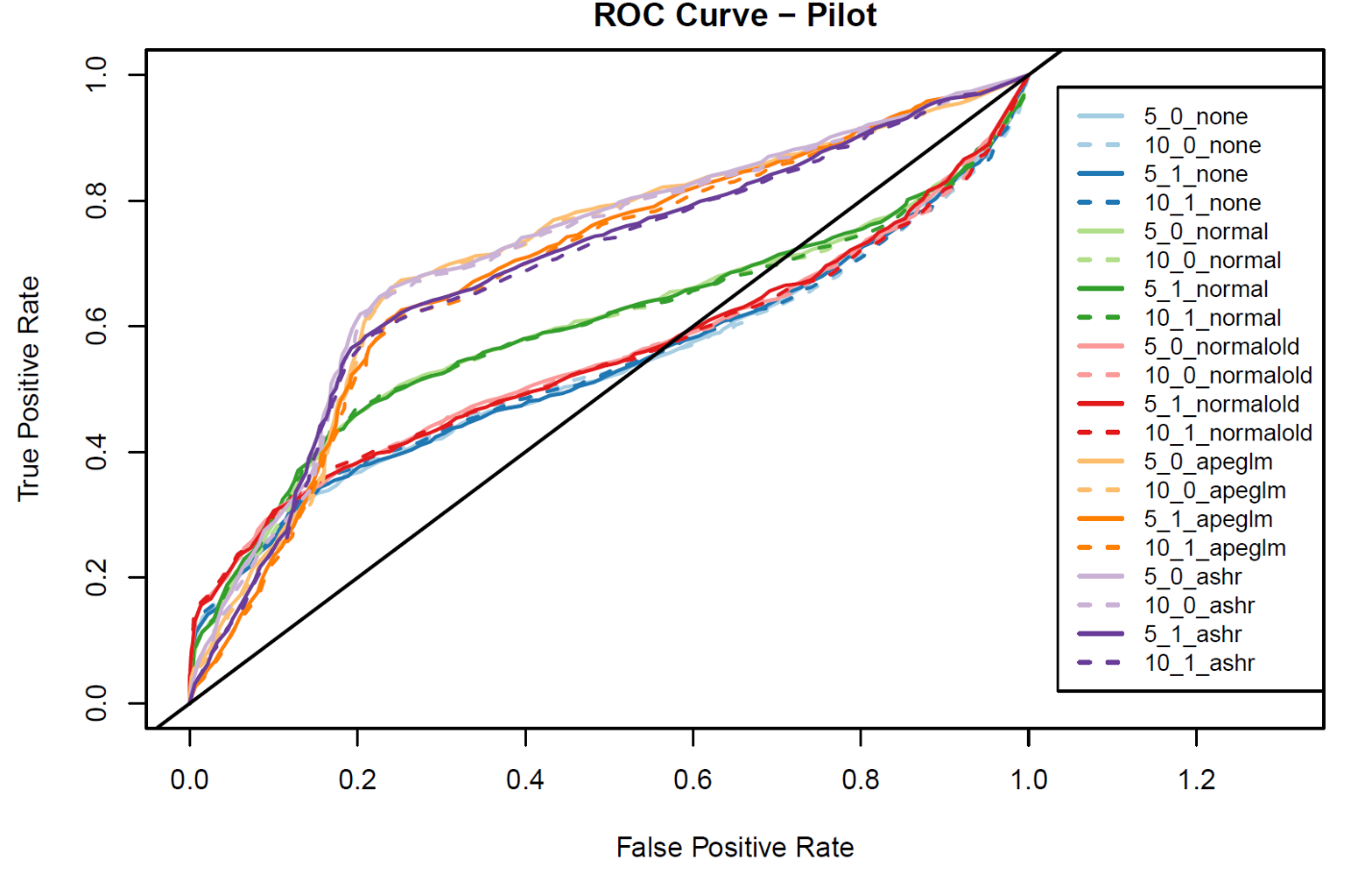


Figure 34: Estrogen detection ROC curves for pilot replicate using individual probes

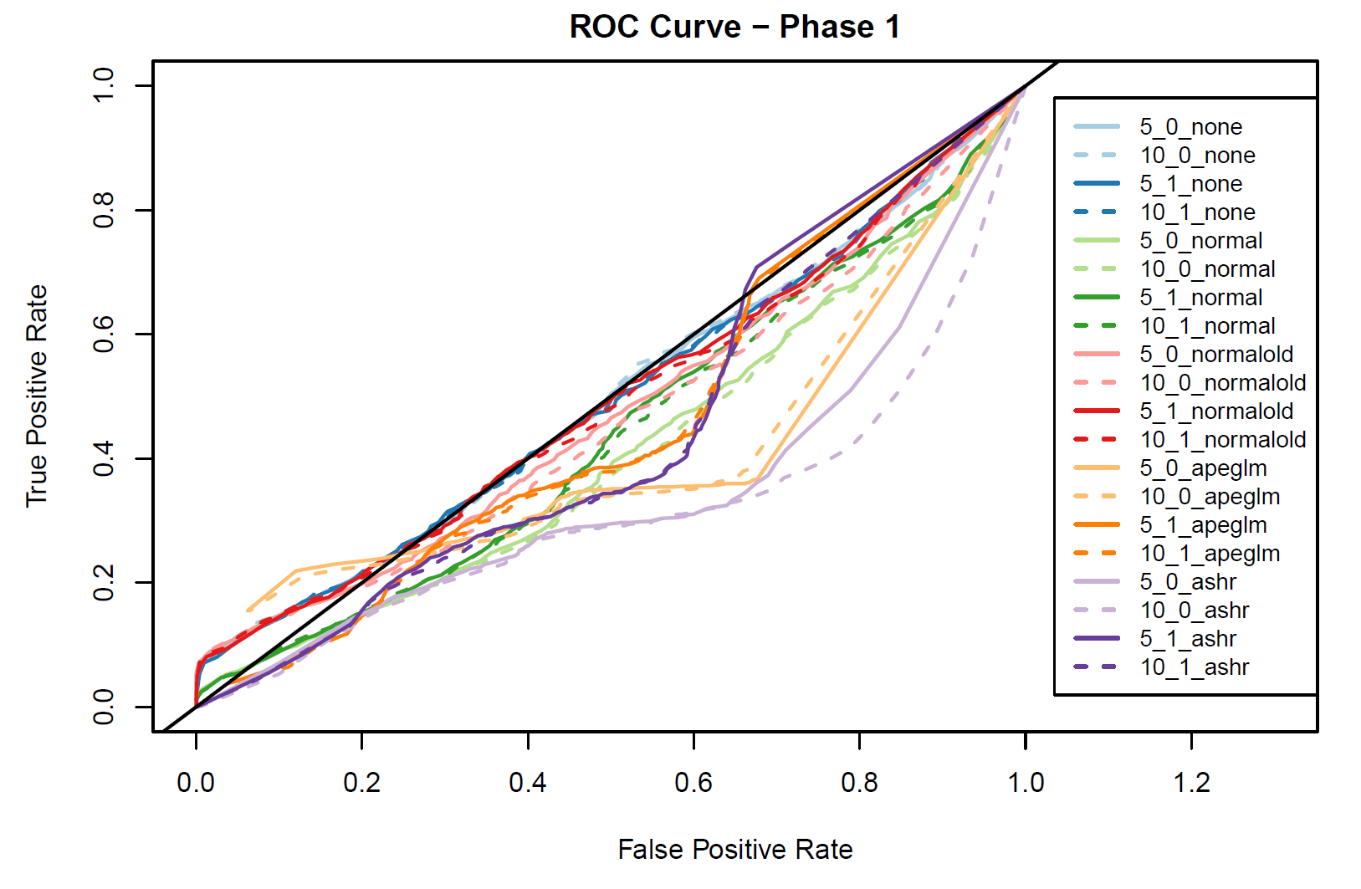


Figure 35:Estrogen detection ROC curves for phase 1 replicate using individual probes

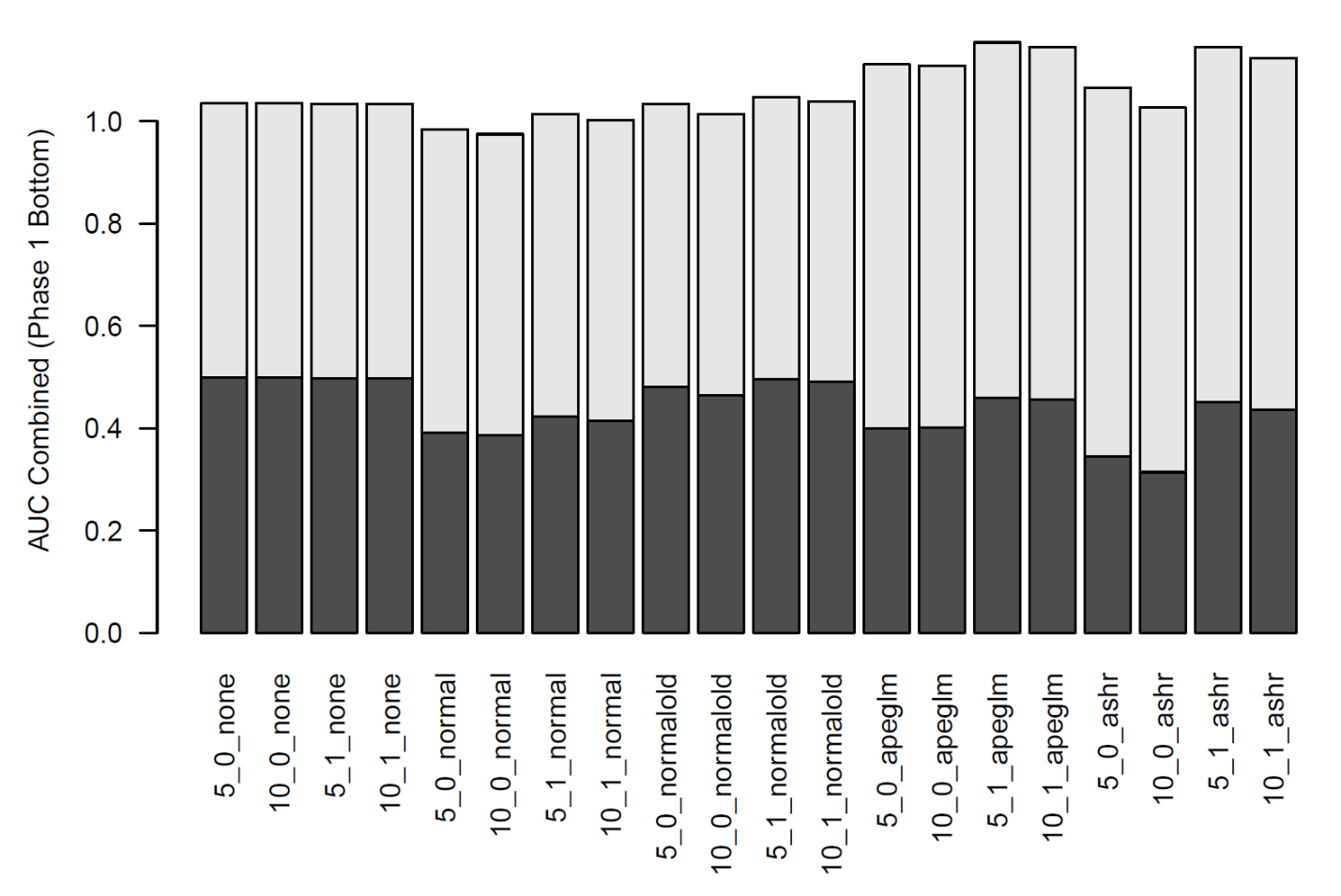


Figure 36:Estrogen detection AUCs for pilot (top) and phase 1 (bottom) replicates using individual probes

Overall, these results were conclusive. First, the FC method had the highest Spearman correlations (and Pearson correlations) between hitcalls for every processing method, implying it is superior at hitcall replication. The FC method and individual probes had a similar ability to replicate points of departure, but GSVA was inferior. FC and GSVA had a similar ability to detect estrogens, but individual probes struggled without further aggregation. The first conclusion is that FC pathway scoring is as good or better than the others by every metric. Within the FC method, we considered individual processing methods. For hitcall replication, ashr had the highest Pearson correlation, but normal was best in the Spearman correlation. For points of departure replication, ashr and apeglm were the best, with normal coming close behind. For estrogen detection, normal, normal old and none excelled, while ashr and apeglm struggled with the phase 1 data. Overall, we concluded that normal shrinkage offers the best balance between replicability of hitcalls and estrogen detection. As for flooring and shrinkage, these barely affected the normal shrinkage results, though they were sometimes consequential in ashr and apeglm.

# Appendix A

## Concentration Response Model Details

“Original form” is the formula given in the BMD technical guidance document10, when applicable. The “model form” is the form that was used for modeling. For convenience, parameters with similar roles in multiple models are given the same name. Some models have been re-parameterized so that the role of each parameter is clearer. All models are treated as using regular units of concentration for simplicity; however, the Hill and Gain-Loss models are internally fit with x in log10 units and the parameters converted to regular units immediately after. Parameter meanings and other conventions are as follows:

* f(x): model function
* x: concentration (in regular units, not log units)
* y: median of responses at each concentration
* a: y scaling factor (in polynomial 1, it’s a slope instead)
* b: x scaling factor
* p: power of x
* q: power of x (gain-loss model loss component)
* tp: model top (following tcpl naming); defined as
* ga: model AC50 (following tcpl naming); defined as
* la: gain-loss model loss AC50 (following tcpl naming); that is, the AC50 for the loss component
* er: log error parameter ; unless otherwise stated,
* x’: concentration with the largest absolute median response
* y’: median response at x’ (can be positive or negative)
* bmr: benchmark response level
* bmd: model benchmark dose; defined as
* C: cutoff

## Constant

|  |  |
| --- | --- |
| Model Form |  |
| Bounds |  |

## Exponential 2

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Exponential 3

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Exponential 4

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Exponential 5

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Gain-Loss

|  |  |
| --- | --- |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Hill

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Polynomial 1

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Polynomial 2

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

## Power

|  |  |
| --- | --- |
| Original Form |  |
| Model Form |  |
| Parameter Initial Conditions |  |
| Bounds |  |
| BMD Substitution |  |
| Top Substitution |  |

1. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J. P., Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011,** *27* (12), 1739-1740.

2. Huang, R.; Grishagin, I.; Wang, Y.; Zhao, T.; Greene, J.; Obenauer, J. C.; Ngan, D.; Nguyen, D.-T.; Guha, R.; Jadhav, A., The NCATS BioPlanet–an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Frontiers in pharmacology* **2019,** *10*, 445.

3. Ryan, N.; Chorley, B.; Tice, R. R.; Judson, R.; Corton, J. C., Moving toward integrating gene expression profiling into high-throughput testing: A gene expression biomarker accurately predicts estrogen receptor α modulation in a microarray compendium. *Toxicological Sciences* **2016,** *151* (1), 88-103.

4. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J. P.; Tamayo, P., The molecular signatures database hallmark gene set collection. *Cell systems* **2015,** *1* (6), 417-425.

5. Dutertre, M.; Gratadou, L.; Dardenne, E.; Germann, S.; Samaan, S.; Lidereau, R.; Driouch, K.; de la Grange, P.; Auboeuf, D., Estrogen regulation and physiopathologic significance of alternative promoters in breast cancer. *Cancer research* **2010,** *70* (9), 3760-3770.

6. Hänzelmann, S.; Castelo, R.; Guinney, J., GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **2013,** *14* (1), 7.

7. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **2005,** *102* (43), 15545-15550.

8. Barbie, D. A.; Tamayo, P.; Boehm, J. S.; Kim, S. Y.; Moody, S. E.; Dunn, I. F.; Schinzel, A. C.; Sandy, P.; Meylan, E.; Scholl, C., Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009,** *462* (7269), 108.

9. Filer, D. L.; Kothiya, P.; Setzer, R. W.; Judson, R. S.; Martin, M. T., tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics* **2016,** *33* (4), 618-620.

10. David Gaylor, J. G., Karen Hogan, Jennifer Jinot, Carole Kimmel, R. Woodrow Setzer, Michael Broder, Diane Henshel, Benchmark Dose Technical Guidance. Agency, U. S. E. P., Ed. 2012.

11. Thomas, R. S.; Allen, B. C.; Nong, A.; Yang, L.; Bermudez, E.; Clewell III, H. J.; Andersen, M. E., A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicological Sciences* **2007,** *98* (1), 240-248.

12. Crump, K. S., Calculation of benchmark doses from continuous data. *Risk Analysis* **1995,** *15* (1), 79-89.

13. Crump, K., Benchmark analysis. *Wiley StatsRef: Statistics Reference Online* **2014**.

14. Watt, E. D.; Judson, R. S., Uncertainty quantification in ToxCast high throughput screening. *PloS one* **2018,** *13* (7), e0196963.

15. Anderson, D.; Burnham, K., Model selection and multi-model inference. *Second. NY: Springer-Verlag* **2004,** *63*.

16. Judson, R. S.; Magpantay, F. M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D. M.; Filer, D. L.; Houck, K. A.; Martin, M. T.; Sipes, N.; Richard, A. M.; Mansouri, K.; Setzer, R. W.; Knudsen, T. B.; Crofton, K. M.; Thomas, R. S., Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicological Sciences* **2015,** *148* (1), 137-154.