PNAS
www.pnas.org

Supplementary Information for

A selective inference approach for FDR control using multi-omics covariates yields insights into disease risk

Ronald Yurko, Max G'Sell, Kathryn Roeder and Bernie Devlin

Kathryn Roeder.
E-mail: roeder@andrew.cmu.edu

This PDF file includes:

Supplementary text
Figs. S1 to S28
Tables S1 to S3
Legend for Dataset S1
SI References

Other supplementary materials for this manuscript include the following:

Dataset S1

Supporting Information Text

AdaPT conditional two-groups model

This section provides a more detailed explanation of updating the rejection threshold $s_t(x_i)$ in the AdaPT procedure, expanding on the description from *Methods* in the main manuscript. As in the main text, this is essentially an explanation of the EM approach of (1). Note that for coherence some text is repeated from the main manuscript. (1) use a conditional version of the classical two-groups model (2) yielding the conditional mixture density,

$$f(p|x) = \pi_1(x) f_1(p|x) + 1 - \pi_1(x), \quad [1]$$

where the null p-values are modeled as uniform ($f_0(p|x) \equiv 1$). They proceed to use a *conservative* estimate for the conditional local false discovery rate, $\text{fdr}(p|x) = \hat{f}(1|x)/\hat{f}(p|x)$, by setting $1 - \pi_1(x) = f(1|x)$.

We model the non-null p-value density with a beta distribution density parametrized by μ_i,

$$f_1(p|x_i) = h(p; \mu_i) = \frac{1}{\mu_i} p^{1/\mu_i - 1}, \quad [2]$$

where $\mu_i = \mathbb{E}[-\log(p_i)]$, resulting in a conditional density for a beta mixture model,

$$f(p|x_i) = \pi_1(x_i) \frac{1}{\mu_i} p^{1/\mu_i - 1} + 1 - \pi_1(x_i). \quad [3]$$

In this form, we can model the non-null probability $\pi_1(x_i) = \mathbb{E}[H_i|x_i]$ and the effect size for non-null hypotheses $\mu(x_i) = \mathbb{E}[-\log(p_i)|x_i, H_i = 1]$ with two separate gradient boosted tree-based models. The XGBoost library (3) provides logistic and Gamma regression implementations which we use for $\pi_1(x_i)$ and $\mu(x_i)$ respectively.

There are two categories of missing values in these regression problems: H_i is never observed, and at each step t of the search, the p-values for tests $\{i : p_i \leq s_t(x_i) \text{ or } p_i \geq 1 - s_t(x_i)\}$ are masked as $\tilde{p}_{t,i}$. An expectation-maximization (EM) algorithm can be used to estimate both $\hat{\pi}_1(x_i)$ and $\hat{\mu}(x_i)$ by maximizing the partially observed likelihood. The complete log-likelihood for the conditional two-groups model is,

$$l(\pi_1, \mu; p, H, x) = \sum_{i=1}^{n} \{H_i \log(\pi_1(x_i) + (1 - H_i)\log(1 - \pi_1(x_i))\} + \sum_{i=1}^{n} H_i \log\{h(p_i; \mu(x_i))\}. \quad [4]$$

During the E-step of the $d = 0, 1, \ldots$ iteration of the EM algorithm, conditional on the partially observed data fixed at step t, $(x_i, \tilde{p}_{t,i})_{i\in[n]}$, we compute both,

$$\hat{H}_i^{(d)} = \mathbb{E}_{\hat{\pi}_1^{(d-1)}, \hat{\mu}^{(d-1)}}[H_i \big| (x_i, \tilde{p}_{t,i})_{i\in[n]}] \quad [5]$$

$$\hat{b}_i^{(d)} = \mathbb{E}_{\hat{\pi}_1^{(d-1)}, \hat{\mu}^{(d-1)}}[\mathbf{1}(p'_{t,i} = p_i) \big| (x_i, \tilde{p}_{t,i})_{i\in[n]}, H_i = 1], \quad [6]$$

where $\hat{b}_i^{(d)}$ indicates how likely $p'_{t,i} = \min(\tilde{p}_{t,i})$ equals p_i for non-null hypotheses. The explicit calculations of $\hat{H}_i^{(d)}$ and $\hat{b}_i^{(d)}$ for both the revealed, $\tilde{p}_{t,i} = p'_{t,i}$, and masked p-values, $\tilde{p}_{t,i} = \{p_i, 1 - p_i\}$, are available in the supplementary materials of (1).

The M-step consists of estimating $\hat{\pi}_1^{(d)}$ and $\hat{\mu}^{(d)}$ with separate gradient boosted trees, using *pseudo*-datasets to handle the partially masked data. In order to fit the model for $\pi_1(x_i)$, we construct the response vector $y_\pi^{(d)} = (1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{2n}$ and use weights $w_\pi^{(d)} = (\hat{H}_1^{(d)}, \ldots, \hat{H}_n^{(d)}, 1 - \hat{H}_1^{(d)}, \ldots, 1 - \hat{H}_n^{(d)}) \in \mathbb{R}^{2n}$. Then we estimate $\hat{\pi}_1^{(d)}(x_i)$ using the first n predictions from a classification model using $y_\pi^{(d)}$ as the response variable with the covariate matrix $(x_i)_{i\in[n]}$ replicated twice and weights $w_\pi^{(d)}$. Similarly, for estimating $\hat{\mu}^{(d)}(x_i)$ we construct a response vector $y_\mu^{(d)} = (-\log(p_1), \ldots, -\log(p_n), -\log(1-p_1), \ldots, -\log(1-p_n)) \in \mathbb{R}^{2n}$ with weights $w_\mu^{(d)} = (\hat{b}_1^{(d)}, \ldots, \hat{b}_n^{(d)}, 1 - \hat{b}_1^{(d)}, \ldots, 1 - \hat{b}_n^{(d)}) \in \mathbb{R}^{2n}$, and again take the first n predicted values using the duplicated covariate matrix.

The conditional local fdr is estimated for each $p'_{t,i}$,

$$\text{fdr}_{t,i} = \frac{\hat{\pi}_1(x_i) h(1; \hat{\mu}(x_i)}{\hat{\pi}_1(x_i) h(p'_{t,i}; \hat{\mu}(x_i)) + 1 - \hat{\pi}_1(x_i)}, \quad [7]$$

and we follow the procedure detailed in Section 4.3 of (1) to update the rejection threshold to $s_{t+1}(x_i)$ by removing test $i^* = \arg\max_{i\in\mathcal{R}_t} \text{fdr}_{t,i}$ from $\mathcal{R}_t$. A summary diagram of the EM algorithm is displayed in Figure S1.

SCZ results with independent loci

One potential concern regarding the assessment of performance of AdaPT is the impact of linkage disequilibrium (LD). In the Manhattan plots of Figures 2(B-C), the discoveries visually appear to be located close to one another. However, the visual appearance of genomic positions is somewhat misleading because our initial selection of eSNPs greatly reduces the number of SNPs commonly portrayed in Manhattan plots – many of these SNPs are not very close to each other in the genome and not in high LD, although the format of the Manhattan plot makes this feature hard to see. To take this analysis further we follow common practice for GWAS results by identifying the "best" or "lead" SNPs in a LD block/cluster, using a similar approach as (4), for each of the set of discoveries presented in Figure 2:

1. order the SNPs by the AdaPT -$\log_{10}$(q-value) in descending order,
2. starting with the SNP with the largest value for the AdaPT -$\log_{10}$(q-value),
    - remove all SNPs with $r^2 \geq 0.1$ within a 500kb window,
    - move on to next SNP that is still remaining,
3. return the retained SNPs as the LD-independent SNPs in low LD ($r^2 < 0.1$). (Remark: this approach excludes SNPs whose contribution to the GWAS signal is partially independent of the lead SNP, but it has the advantage of simplicity.)

We use the reference European sample genotype data from the 1000 Genomes project (5) to compute the r^2 values between SNPs. In the GWAS setting this LD clumping procedure is typically applied to the reported SNP p-values, but because the ordering of SNPs varies between the different sets of discoveries (intercept-only versus use of covariates) we perform the operation separately with their respective q-values. For each of the different set of covariates considered, this results in reducing the 25,076 selected eSNPs down to the following number of "independent loci":

- Intercept-only: 3,958
- BD z-stats: 3,966
- BD z-stats + eQTL slopes: 3,962
- BD z-stats + eQTL slopes + WGCNA (w/ interactions): 3,963
- BD z-stats + eQTL slopes + WGCNA (w/o interactions): 3,959
- WGCNA: 3,954

The differences in counts are due to the different number of ties that take place between the resulting q-values for each considered set of covariates. Next, for the identified set of "lead" SNPs we observe how many have q-values less than the target FDR level $\alpha = 0.05$ (i.e. associations detected at $\alpha = 0.05$). The results are displayed in the Figure S2, including Manhattan plots Figures S2(A-B) of the q-values for the AdaPT intercept-only and BD z-stats + eQTL slopes + WGCNA (w/ interactions) results, rather than using the actual p-values. The lead SNPs in each of the Manhattan plots are denoted by an X shape. In conjunction with Figures S2(C-D), the relative improvement in the set of independent loci within the discovery sets from AdaPT is analogous to the results presented in Figure 2, emphasizing the advantage of accounting for covariates and their interactions via gradient boosted trees. Additionally, Figure S3 further emphasizes that the improvement in power is not restricted to a particular section of the genome. As seen in Figure S4, we observe a similar improvement in the number of independent loci when ordering the SNPs with the observed 2014-only studies SCZ p-values.

While we maintain FDR control on the original set of discoveries (see Figure 3 in *Results*), we do not retain any guarantees regarding the detected independent loci presented in Figure S2. In order to maintain FDR control on the set of discovered independent loci, an alternative approach or adjustment to the AdaPT algorithm is required. A simple alternative is to first apply LD pruning/clumping as initial step prior to applying AdaPT to a reduced set of lead SNPs. However, this encounters the challenge of defining lead SNPs without data "snooping" based on using the observed p-values. Future work will explore modifications for AdaPT, potentially exploring recent developments (6), to maintain FDR control on an independent subset of SNPs.

SCZ variable importance and partial dependence

We explore further the variable relationships from the gradient boosted trees. First, Figure S5 displays the change in variable importance for the non-null effect size (μ) at each model fitting iteration, with the top variables in the final model highlighted. The variable importance measures are relatively stable across all model iterations with the BD z-statistics and eQTL slope measures maintaining the highest level of importance. Figure S6 displays the partial-dependence plot at each AdaPT model fitting iteration for the estimated marginal relationship between the BD z-statistics and the non-null effect size μ, evaluated at the $0, 2.5\%, 5\%, \ldots, 100\%$ percentiles. The estimates reveal an increasing effect size as the BD z-statistics grow in magnitude, which is relatively stable across the model iterations. Figures S7(A-C) display the relationships for the probability of non-null model, while (D-F) display relationships for the effect size under the alternative. Although the partial dependence plots show considerable variability due to the high dimensional of the model, we can still see general trends consistent with the variable importance plots from Figure 3(A) and Figure S5.

In Figure S8 we display the p-value distributions comparing the enrichment for membership in the different WGCNA modules reported by (7). While many of the WGCNA modules lack clear evidence or contain too few eSNPs, as denoted by their respective y-axes, the *cyan* and *salmon* modules display noticeable enrichment. Additionally, as mentioned previously, membership in the *gray* module displays a lack of enrichment versus no associated cis-eQTL gene affiliated with the unassigned WGCNA module.

As additional context for the improved performance from using all covariates with interactions, Figures S9(A-B) display the change in partial dependence between the BD z-statistics and probability of being non-null π_1 across the AdaPT search for

the AdaPT results using (A) BD z-statistics only and (B) all covariates without interactions. When compared to the results using all covariates with interactions in Figure Figure 3(B), we see that both versions of these results display relatively flat relationships near the end of the AdaPT search. This provides evidence of the importance of the interactions between other covariates and the BD z-statistics in retaining discriminatory power of the eSNPs near the end of the AdaPT search.

Replication simulations

We use simulations to empirically assess the observed nominal replication rate, percentage of discoveries with p-values less than 0.05 in holdout *2018-only* studies, of 55.2% for the 843 SCZ discoveries from the *2014-only* studies at target FDR level $\alpha = 0.05$. We use the final non-null effect size model returned by the AdaPT, $\hat{\mu}^*$, to generate simulated p-values $\boldsymbol{p}^{sim}$ and nominal replication rates to compare the observed rate against. For the simulations, we assume that all 843 SCZ discoveries from the *2014-only* studies are truly non-null, and we use the actual eSNPs, their observed standard errors σ_{14}, σ_{18} from the *2014-only* and *2018-only* studies respectively, as well as their actual covariates for generating $\boldsymbol{p}^{sim}$. A single iteration of the simulation proceeds as follows:

- For each of the $R_{\text{SCZ}} = 843$ discoveries $i \in \mathcal{R}_{\text{SCZ}}$:
    1. Assume test status is non-null: $H_i = 1$.
    2. Generate effect size using final AdaPT model as truth:
$$-\log p_i^{sim} | x_i^{\text{SCZ}} \sim \text{Exp}\big(1/\hat{\mu}^*(x_i^{\text{SCZ}})\big). \quad [8]$$
    3. Transform effect sizes to p-value p_i^{sim}.
    4. Convert simulated p-value to z-statistic $z_i^{sim} = \left|\Phi^{-1}(p_i^{sim}/2)\right|$.
    5. Calculate updated z-statistic to reflect observed reduction in standard error for *2018-only* studies relative to *2014-only*,
$$z_i^{*,sim} = z_i^{sim} \cdot \frac{\sigma_{14}}{\sigma_{18}}. \quad [9]$$
    6. Convert updated z-statistic to p-value:
$$p_i^{*,sim} = 2 \cdot \Phi(-|z_i^{*,sim}|). \quad [10]$$
- Calculate nominal replication rate using $\boldsymbol{p}^{sim} = (p_i^{*,sim}, \ldots, p_{R_{\text{SCZ}}}^{*,sim})$,
$$\text{Nominatl replication rate } = \frac{|\{i : p_i^{*,sim} \leq .05\}|}{R_{\text{SCZ}}}. \quad [11]$$

We repeat this process to generate ten-thousand simulated values for the nominal replication rate. The distribution of the simulated values ranges from approximately 51% to 63%, with an average and median of $\approx 57\%$, close to the observed rate of 55.2%. Obviously, assuming that all of the 843 rejections are truly non-null is an overtly optimistic assumption given the use of FDR error control. Thus, the average simulated nominal replication rate of $\approx 56.6\%$ is reassuringly close to the observed rate and likely higher than what would be expected if false discoveries were accounted for among the 843 considered eSNPs.

SCZ results with *all 2018* studies

We generate the AdaPT results using the SCZ p-values from *all-2018* studies to the same set of $n_{\text{SCZ}} = 25,076$ eSNPs with the same covariates x_i^{SCZ}. As a comparison to the results displayed in Figure 2 using the *2014-only* studies, Figures S10(A-D) display the same figures but with the results from *all 2018* at target FDR level $\alpha = 0.05$. In contrast to before, we see that due to the increase in power from the study size, the use of modeling the auxiliary information provides a much smaller increase in power with just an approximately 19% increase in discoveries from the intercept-only results (1,865 discoveries) to using all twenty-four covariates with interactions (2,228 discoveries).

For comparison, we additionally examine the change in variable importance and partial dependence plots returned by AdaPT using *all 2018* studies. Similar to before, Figures S11(A-B) display the change in variable importance plots for both the probability of being non-null π_1 and effect size under alternative μ models using the SCZ p-values from *all 2018* studies respectively. The results are similar to before, but with the complete sample eQTL slopes possessing the highest importance. The BD z-statistics are again highly important for *all 2018* studies, displaying the similarly increasing relationships across the AdaPT models as seen in the partial dependence plots in Figures S12(C-D). The change in partial dependence plots for the different eQTL slopes summaries are seen in Figures S13(A-F). Figure S14 displays the levels of SCZ enrichment for *all 2018* studies, revealing modules that are consistent with the *2014-only* studies such as *cyan* and *salmon*.

Type 2 diabetes results

Using GWAS summary statistics for type 2 diabetes (T2D), unadjusted for BMI, available from Diabetes Genetics Replication And Meta-analysis (DIAGRAM) consortium (8), we applied our full pipeline outlined in Figure 1. Of the initial set of over twenty-three million SNPs available, we identified 176,246 eSNPs from eQTL variant-gene pairs from any GTEx tissue sample using the definition of the GTEx eSNPs explained in *Data*. Figure S15 displays the enrichment for these GTEx eSNPs compared to the original set of SNPs from the T2D GWAS results.

We create a vector of covariates x_i^{T2D} summarizing expression level information from GTEx for pancreas, liver, and two adipose tissues, *subcutaneous* and *visceral (omentum)*. Specifically, we calculate $\tilde{\beta}_i^{r^{\text{T2D}}}$ for each r^{T2D} in the set of tissues: pancreas, liver, adipose - subcutaneous, adipose - visceral (omentum). Additionally, we generate WGCNA module assignments using protein coding genes for pancreas samples from GTEx. To generate the WGCNA results, we only consider protein coding genes identified using the `grex` package in `R` (9, 10). Additionally, all genes with expression levels of zero for over half of the provided samples were removed. This resulted in fourteen different module, including the unassigned *gray* module. Unlike the SCZ application, we do not use independent GWAS results from another phenotype.

Using x_i^{T2D} defined above, we applied AdaPT to the 176,246 GTEx eSNPs. However, we encountered an issue for this data where we were unable to discover any hypotheses at target FDR level $\alpha \leq 0.05$. This was due to the fact that 640 eSNPs had p-values *exactly* equal to one. While this can understandably occur with publicly available GWAS summary statistics, p-values equal to one will then *always* contribute to the *pseudo*-estimate for the number of false discoveries A_t during the AdaPT search (see *Methodology overview*). With a relatively high number of p-values equal to one, AdaPT is unable to search through rejection sets for lower α values. To overcome this challenge, we draw random replacement p-values for the 640 eSNPs from a uniform distribution between 0.97 and $1 - 1\text{E}^{-15}$, a value strictly less than one, to allow some leeway. We refer to this set of p-values as *adjusted*, while the original observed p-values are *unadjusted*. For comparison, Figure S16 shows the difference in the number of discoveries for the *adjusted* and *unadjusted* p-values across different target α values. Due to the similarity in performance for α values greater than 0.1, we use results for the *adjusted* p-values moving forward.

At target FDR level $\alpha = 0.05$, AdaPT yields 14,920 T2D discoveries using the *adjusted* p-values with covariates x_i^{T2D} (compared to 14,693 intercept-only discoveries). The change in variable importance for the T2D AdaPT models are displayed in Figure S17. This set of eSNPs is associated with 5,970 cis-eQTL genes for which we then applied gene ontology enrichment analysis to (11, 12), identifying the gene enrichment for biological processes displayed in Figure 5.

BMI results

We also applied our pipeline of analysis to BMI, unadjusted for waist-to-hip ratio (WHR), using GWAS results for individuals of European ancestry available from the GIANT Consortium. Specifically, we approached BMI in the same manner as SCZ: apply AdaPT to GWAS results from earlier studies with a sample size of 322,154 individuals (13); then compare the nominal replication results on recently conducted studies with a sample size of approximately 700,000 individuals (14). As before, all of the *2015-only* studies from (13) were included as a subset of *all 2018* studies (14). Because both (13) and (14) use the inverse variance-weighted fixed effects approach for meta-analysis, we then compute statistics for the studies exclusive to *2018-only* studies in (14). Additionally, to make this example more comparable to the SCZ use, we also use GWAS results for WHR (15) as a covariate (analogous to BD for SCZ). Following pre-processing steps (matching SNPs across studies and effect alleles in both WHR and BMI), we identified 47,690 GTEx eSNPs from a set of nearly two million SNPs, based on the definition explained in *Data*. Figure S18 displays the enrichment for the GTEx eSNPs compared to the original set of pre-processed SNPs for the *2015-only* studies.

Based on previous knowledge of BMI tissue expression associations (13), we create a vector of covariates x_i^{BMI} summarizing expression level information from GTEx for brain and adipose tissues (both *subcutaneous* and *visceral (omentum)*). Specifically, we calculate $\tilde{\beta}_i^{r^{\text{BMI}}}$ for each $r^{\text{BMI}} \in \{$GTEx brain tissues, adipose - subcutaneous, adipose - visceral (omentum)$\}$, where we consider the following brain tissues: (1) *amygdala*, (2) *anterior cingulate cortex BA24*, (3) *caudate basal ganglia*, (4) *cerebellar hemisphere*, (5) *frontal cortex BA9*, (6) *hippocampus*, (7) *hypothalamus*, (8) *nucleus accumbens basal ganglia*, (9) *putamen basal ganglia*, (10) *spinal cord cervical c-1*, and (11) *substantia nigra*. We do not consider the available *cerebellum cortex* tissue samples from GTEx as these are duplicates of *cerebellar hemisphere* and *frontal cortex BA9* respectively. We instead only use the samples taken the same time as the other brain sub-regions at the University of Miami Brain Endowment Bank, preserved by snap freezing (see GTEx FAQs).

We also created an aggregate across $\mathcal{G}_i^{r^{\text{nc}}}$, all cis-eQTL genes associated with eSNP i for each non-cerebellar hemisphere brain tissue region r^{nc},

$$\bar{\beta}_i^{\text{nc}} = \frac{1}{|\mathcal{G}_i^{r^{\text{nc}}}|} \sum_{g \in \mathcal{G}_i^{r^{\text{nc}}}} |\beta_{i,g}^{r^{\text{nc}}}|. \quad [12]$$

We did not include the cerebellum tissue samples in this aggregate due to the reported distinctness of the cerebellum relative to other brain tissue samples (16). Similarly, we computed an average across the two adipose tissues. As before, when calculating the various eQTL slopes summaries, if eSNP i was not an eQTL for a particular region then we impute a value of zero reflecting the lack of associated expression.

Furthermore, WGCNA module assignments were generated using protein coding genes for three different sets of tissues: (1) all non-cerebellar hemisphere brain tissues, (2) cerebellar hemisphere only tissue, and (3) adipose tissues (using same

settings described previously in *Type 2 diabetes results*). Together with the WHR z-statistics and covariates accounting for the associations and WGCNA module indicators, x_i^{BMI} contained 110 variables.

For BMI eSPS, 376 have p-value exactly equal to one, leading to the same problem as we encountered in the T2D analysis. Again, we proceed by randomly drawing replacement p-values for these 376 eSNPs from a uniform distribution between 0.97 and $1 - 1\text{E}^{-15}$. Figure S19 shows how AdaPT fails to obtain any discoveries across the various α levels without making an adjustment to the p-values. With this limitation recognized, we proceed to focus on the discoveries returned by AdaPT using the adjusted p-values at $\alpha = 0.05$.

Unlike SCZ and T2D, AdaPT using all of the covariates (with the same tuning parameters as SCZ) detected fewer discoveries: 1,383 eSNPs compared to 1,624 eSNPs discovered by the intercept-only AdaPT model at target FDR level $\alpha = 0.05$. With further boosting regularization, beyond what is considered here, one could achieve the intercept-only results with gradient boosted trees. Of these 1,383 discoveries, approximately 83% (1,140 eSNPs) were nominal replications with p-values less than or equal to 0.05 in the independent *2018-only* studies. Figure S20 displays the increasing smoothing spline relationship between the *2018-only* p-values and the resulting *2015-only* q-values from the AdaPT search on the $\log_10$ scale. The much higher observed nominal replication rate is not surprising given the well powered size of the BMI studies, as indicated by the y-axis of Figure S20, which reflects the level of enrichment for the *2018-only* studies.

Additionally, gene ontology enrichment analysis for the 1,383 discoveries using all covariates revealed no significant biological process enrichment at target FDR level $\alpha = 0.05$. One concern is that a model with 110 variables is excessive, because the variable importance plots for the BMI AdaPT models in Figures S21(A-B), along with the partial dependence plots in Figures S22(A-B), emphasize the relative importance of the WHR z-statistics compared to other covariates. To test this conjecture, we explored two simpler models using (1) WHR z-statistics only and (2) WHR z-statistics with eQTL slope summaries. These produced 1,324 and 1,351 discoveries at the 0.05 level, respectively. We conclude that the available covariates do not provide sufficient additional information beyond the signal available with this immense sample and consequently including covariates in the AdaPT model does not increase the power of the procedure.

CV tuning for SCZ, T2D, and BMI results

Rather than fixing the parameter settings for the XGBoost gradient boosted trees, we use the CV algorithm (detailed in *Methods*) at two steps of the search to tune the models (see the following section for justification of using two CV steps). For our search space, we evaluate a small range of values for the number of trees P and limit the maximum tree depth D to result in reasonably shallow trees (referred to as `nrounds` and `max_depth` in the `xgboost` package (17)).

First, for SCZ analysis, when exploring the improvement in discovery rate for the eSNPs by incrementally including more information, we used the following XGBoost settings:

- BD z-stats: Combinations of $P \in \{100, 150\}$, $D \in \{1, 6\}$,
- BD z-stats + eQTL slopes: Combinations of $P \in \{100, 150\}$, $D \in \{3, 6\}$,
- BD z-stats + eQTL slopes + WGCNA: Combinations of $P \in \{100, 150\}$, $D \in \{2, 3\}$,
- WGCNA only: Combinations of $P \in \{100, 150\}$, $D \in \{1, 2, 3\}$.

We explored different settings for the different possible covariates to address the types of variables included. For instance, when using the BD z-statistics only, we considered both single-split "stumps" as well as more depth with six splits to potentially handle the variable's symmetric relationship. Once we have all three types of covariates (BD z-statistics, eQTL slope summaries, and WGCNA results), we limit the maximum depth to be at least two to ensure possible interactions can be captured.

The selected number of trees P and maximum depth D for each of these sets of covariates is displayed in Table S1. When using only the BD z-statistics, as well as only including the eQTL slopes, the single-split settings were selected in the first CV step while the higher depth was selected in the second CV step. When using all covariates, the most complex settings (largest number of trees and largest depth) are selected in both CV steps. This agreement in selection is not surprising given the choice of the low starting threshold $s_0 = 0.05$, which differs from the results displayed in Table S3 of the next section using $s_0 = 0.45$. We evaluated the same possible settings for the various *all 2018* results displayed in Figures S10(C-D): the same choices for P and D displayed in Table S1 were selected in both CV steps.

For the T2D and BMI results with their full set of covariates, we evaluated four combinations: (1) $P = 100,\ D = 2$, (2) $P = 150,\ D = 2$, (3) $P = 100,\ D = 3$, and (4) $P = 150,\ D = 3$. For the BMI results using only WHR z-statistics, we varied over $P \in \{100, 150\}$ and $D \in \{1, 6\}$; for the results using WHR z-statistics with the eQTL slopes, we used combinations of $P \in \{100, 150\}$, $D \in \{3, 6\}$. The selected number of trees P and maximum depth D for each of these sets of AdaPT results at both CV steps is displayed in Table S2.

Selection of s_0 and number of CV steps

To justify the selection of both the starting threshold s_0 and number of CV steps for the AdaPT search, we generated simulations from the first AdaPT models returned from the SCZ *2014-only* results. While these models are based on AdaPT results with a starting threshold of $s_0 = 0.05$ following one CV step, they are only from the first model and are not explicitly parametrized by s_0 and the number of CV steps. We know, however, that these first models are the result of using $P = 150$ trees with a maximum depth of $D = 3$, as indicated in Table S1 of the previous section.

Let $\hat{\pi}_1^*$ and $\hat{\mu}^*$ be the first models for the probability of non-null and effect size under the alternative that AdaPT returns for the eSNPs using all covariates x_i^{SCZ}. We use these models as the "truth" for generating data, in which a single iteration of the simulation proceeds as follows:

- For each eSNP $i \in [n^*_{\text{SCZ}}]$
    1. Generate test status: $H_i | x_i^{\text{SCZ}} \sim \text{Bernoulli}\big(\hat{\pi}_1^*(x_i^{\text{SCZ}})\big)$.
    2. Generate simulated effect sizes:

$$-\log p_i | H_i, x_i^{\text{SCZ}} \sim \begin{cases} \text{Exp}\big(1\big) \text{ if } H_i = 0, \\ \text{Exp}\big(1/\hat{\mu}^*(x_i^{\text{SCZ}})\big) \text{ if } H_i = 1. \end{cases} \quad [13]$$

    3. Transform to p-values p_i.
- Apply AdaPT to simulated study p-values with specified s_0 and v CV steps with two candidate settings:
    1. number of trees $P = 100$ and maximum depth $D = 2$,
    2. number of trees $P = 150$ and maximum depth $D = 3$.
- Compute observed power and FDP at range of target FDR α values.

We generate one-hundred simulations this way for each possible threshold $s_0 \in \{0.05, 0.25, 0.45\}$ and $v \in \{1, 2, 5\}$ CV steps. Figure S23 displays the average difference in power between the different starting threshold values by the number of CV steps. Although the differences are small, we see that using $s_0 = 0.05$ results in higher power, on average, than both 0.25 and the recommended 0.45 value. Using this low starting threshold of $s_0 = 0.05$, we then directly compute the difference in power between the different number of CV steps displayed in Figure S24. Unsurprisingly, while again the differences are small, only one CV step results in the lowest power, on average. Since the computational cost of AdaPT with CV tuning is reduced by only using two CV steps instead of a higher number, such as five, and the simulations demonstrate on average no difference in power at both α values of 0.05 and 0.10, we use the starting threshold of $s_0 = 0.05$ with two CV steps in our applications of AdaPT.

In the previous section, Table S1 displayed the selections in both CV steps with $s_0 = 0.05$. For comparison, Table S3 displays the selections using $s_0 = 0.45$. Instead of selecting the same settings in both steps, the higher initial threshold selects the least complex settings (smallest number of trees and minimum depth) in the first CV step before flipping to the most complex settings in the second step. Intuitively, the higher initial threshold means more information is masked from the models, so it is not surprising to see less complex settings chosen. This further reinforces the use of the lower initial threshold $s_0 = 0.05$: it starts with more revealed information and selects model settings corresponding to improved CV performance for tests with lower p-values of interest.

Dependent p-value block simulation

To demonstrate the performance of AdaPT in the presence of dependent tests, we construct simulations with a block-correlation scheme to emulate LD structure for SNPs. We consider a setting with two independent covariates,

$$x_i = (x_{i1}, x_{i2}),$$
$$\text{where } x_{i1}, x_{i2} \sim \text{Uniform}(0, 1).$$

For each test $i \in [n]$, we define a linear relationship for the log-odds of being non-null using these covariates,

$$\text{logit}(\pi_{1,i}(x_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

Then, the resulting status of the test H_i is a Bernoulli random variable based on the probability $\pi_{1,i}(x_i)$ where $H_i = 1$ indicates the test i is non-null while $H_i = 0$ indicates a true null,

$$H_i \sim \text{Bernoulli}(\pi_{1,i}(x_i)).$$

Given this test status, a vector of true effect sizes $\boldsymbol{\mu} = c(\mu_i, \ldots, \mu_n)$ is also generated as a function of the covariates,

$$\mu_i(x_i) = \begin{cases} \max\{\mu_{floor},\ \gamma_1 x_{i1} + \gamma_2 x_{i2}\} \text{ if } H_i = 1, \\ 0 \text{ otherwise.} \end{cases}$$

To simulate observed effect sizes, we construct an $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ with B blocks of equal size $\frac{n}{B}$. Each block $b \in [B]$ has constant correlation ρ between all tests *within* the block, while each block is independent of each other. This results in constructing individual block covariance matrices, Σ_b, with ones along the diagonal and ρ for the off-diagonal elements. Each of these individual matrices are placed along the diagonal of $\boldsymbol{\Sigma}$, with the remaining off-diagonal elements set to zero so

blocks are independent of each other. As an example, if each block contained only two tests they would be constructed in the following manner,

$$\Sigma_b = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{\Sigma} = \begin{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \dots & \dots & \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \end{bmatrix}$$

Using this block-wise construction of the covariance matrix, we then proceed to generate the vector of observed effect sizes $\boldsymbol{z} = (z_i, \dots, z_n)$ from a multivariate Gaussian distribution,

$$\boldsymbol{z} \sim \text{Normal}(\boldsymbol{\mu}, \mathbf{\Sigma}).$$

We compute the resulting two-side p-value $p_i = 2 \cdot \Phi(-|z_i|)$ for each test's observed effect size.

For each dataset generated using this process above, we compute both the observed FDP and power for the classical BH procedure and two different versions of AdaPT:

1. intercept-only,
2. gradient boosted trees with covariates: $x_i = (x_{i1}, x_{i2})$.

We fix both $n = 10{,}000$ and $B = 500$ blocks, resulting in 500 blocks of twenty tests each. Rather than force all non-nulls together in the same blocks, we first calculate the minimum number of blocks required to hold all non-null tests, $B_A^* = \lceil |\{i : H_i = 1\}|/20 \rceil$. The non-null tests are then randomly assigned to $B_A = \lceil (500 + B_A^*)/2 \rceil$ blocks, ensuring that there will be blocks containing both null and non-null tests. The $|\{i : H_i = 0\}|$ tests are randomly assigned to available spots within the B_A blocks as well as the remaining $500 - B_A$ strictly null blocks.

In our simulations, we fix $\beta_0 = -3$ and require that both $\beta_1 = \beta_2$ and $\gamma_1 = \gamma_2$. We vary the following settings in our simulations:

1. block correlation $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$ where each block has the same value for ρ,
2. $\beta_1, \beta_2 \in \{1, 2, 3\}$,
3. $\mu_{floor} \in \{0.5, 1, 1.5\}$,
4. $\gamma_1, \gamma_2 \in \{0.5, .75, 1\}$.

We generate 100 simulations using the data generating process above, computing both the FDP and power for BH and the two different versions of AdaPT. For the covariate-informed version of AdaPT, we use gradient boosted trees via XGBoost with $P = 100$ trees and maximum depth $D = 1$. For both versions of AdaPT results, we start with the initial threshold of $s_0 = 0.45$ and update the model ten times throughout the search (rather than the recommended twenty for computational speed).

Figures S25, S26, and S27 display points for the average observed FDP and power across the 100 simulations with plus/minus two standard errors bars for μ_{floor} =0.5, 1, and 1.5 respectively, with target FDR level $\alpha = 0.05$. The columns in each figure correspond to the different values considered for $\gamma_1 = \gamma_2$, while the rows correspond to $\beta_1 = \beta_2$. The x-axis for the figures displays the increasing block correlation ρ. Regardless of the simulation setting, we see that the AdaPT results when accounting for covariates (x_{i1}, x_{i2}) maintains valid FDR control at 0.05 similar to BH. This holds in the settings with greater effect sizes, as well as when the covariate information displays the best performance in terms of observed power (the bottom right panels of each figure). We can see that the intercept-only approach fails to achieve FDR control under block settings with perfect correlation, while the use of covariate information appears to inhibits such behavior. Our focus on positive correlation values is synonymous with the setting faced in genomics regarding LD structure. Further exploration of AdaPT's performance in settings with arbitrary dependence structure presents an opportunity for future work, as well as accounting for covariate information that predict observed correlated noise.

Simulations demonstrating effects of overfitting

It is possible that flexible methods like gradient boosted trees can be overfit, especially on small data sets. This could potentially lead to concerns about their incorporation in AdaPT. To assess the effects of overfitting the gradient boosted trees in AdaPT, we constructed simulated datasets using the first models returned by AdaPT on the SCZ GWAS results, $\hat{\pi}_1^*$ and $\hat{\mu}^*$, with the actual covariates x_i^{SCZ} for each of the $n_{\text{SCZ}}^* = 25{,}076$ eSNPs. We then simulated data using these models in the same manner previously explained for choosing s_0 and the number of CV steps, and computed the observed power and FDP over a range of number of trees $P \in \{100, 300, 500, 700, 900\}$.

Figure S28(A) displays the distributions for fifty simulations of the observed FDP as the number of trees in the gradient boosted model increases. Regardless of the number of trees, we still maintain valid FDR control. However, Figure S28(B) shows as the number of trees increases, the method will overfit, resulting in a reduction in power. This reinforces that, although good model tuning can be important for power, the AdaPT method continues to maintain FDR control even as the model breaks down.

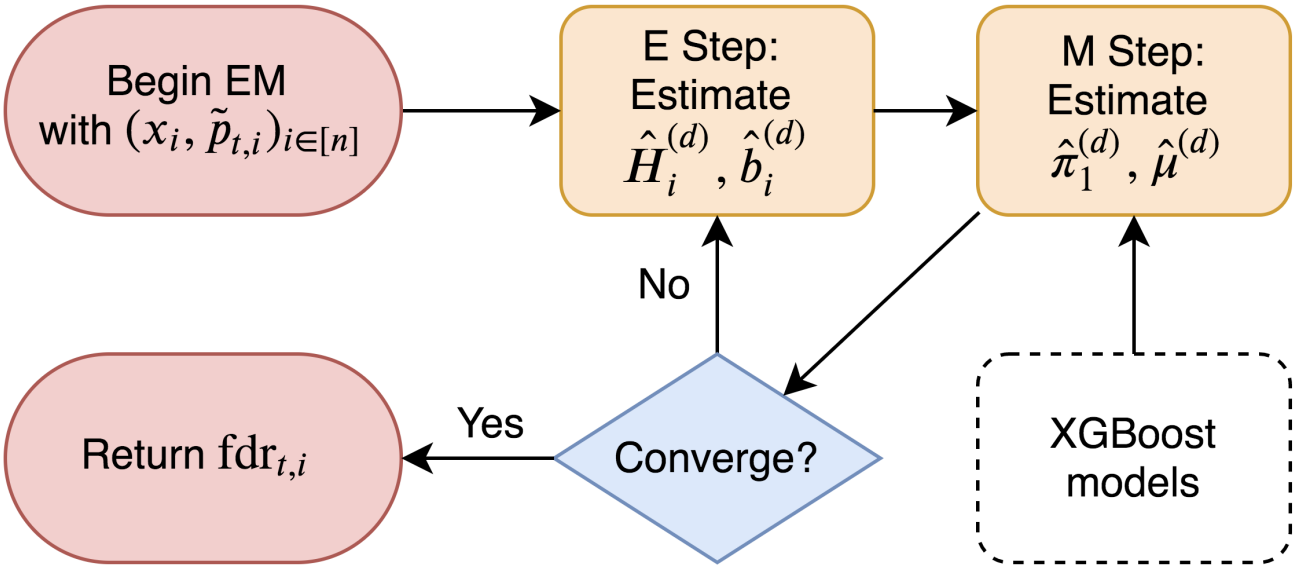

Fig. S1. Summary of AdaPT EM algorithm.

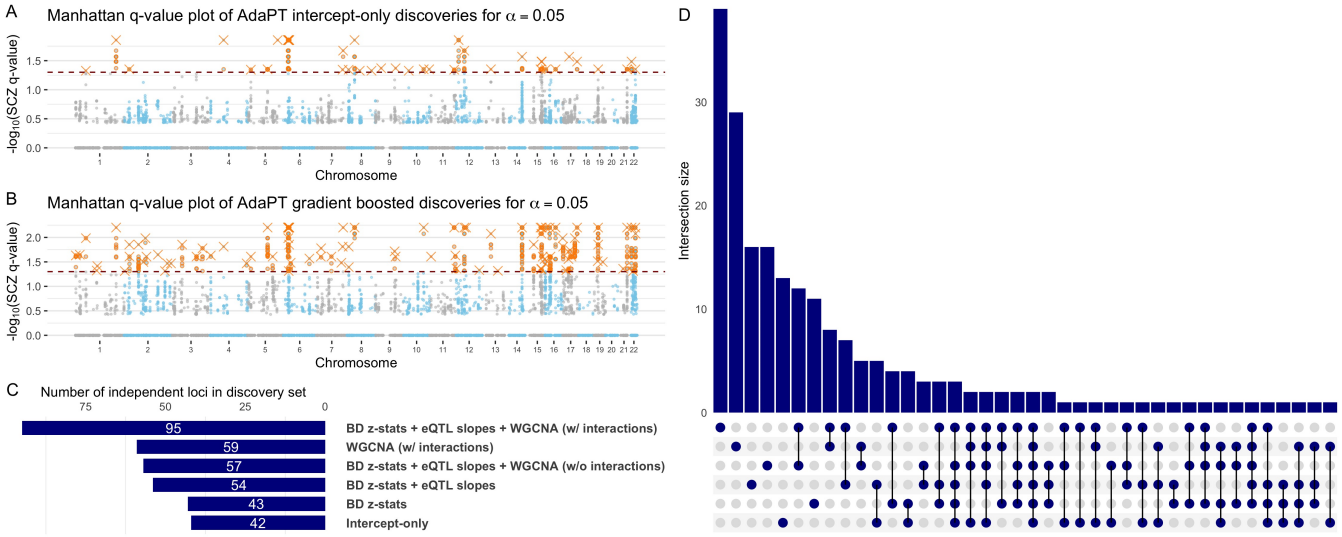

Fig. S2. Manhattan q-value plots of SCZ AdaPT discoveries (orange) using *(A)* intercept-only model compared to *(B)* covariate informed model at target $\alpha = 0.05$, with lead SNPs for independent loci denoted by Xs. *(C)* Comparison of the number of independent loci for each discovery set at target $\alpha = 0.05$ based on LD pruning with the respective AdaPT q-values and *(D)* their resulting discovery set intersections.

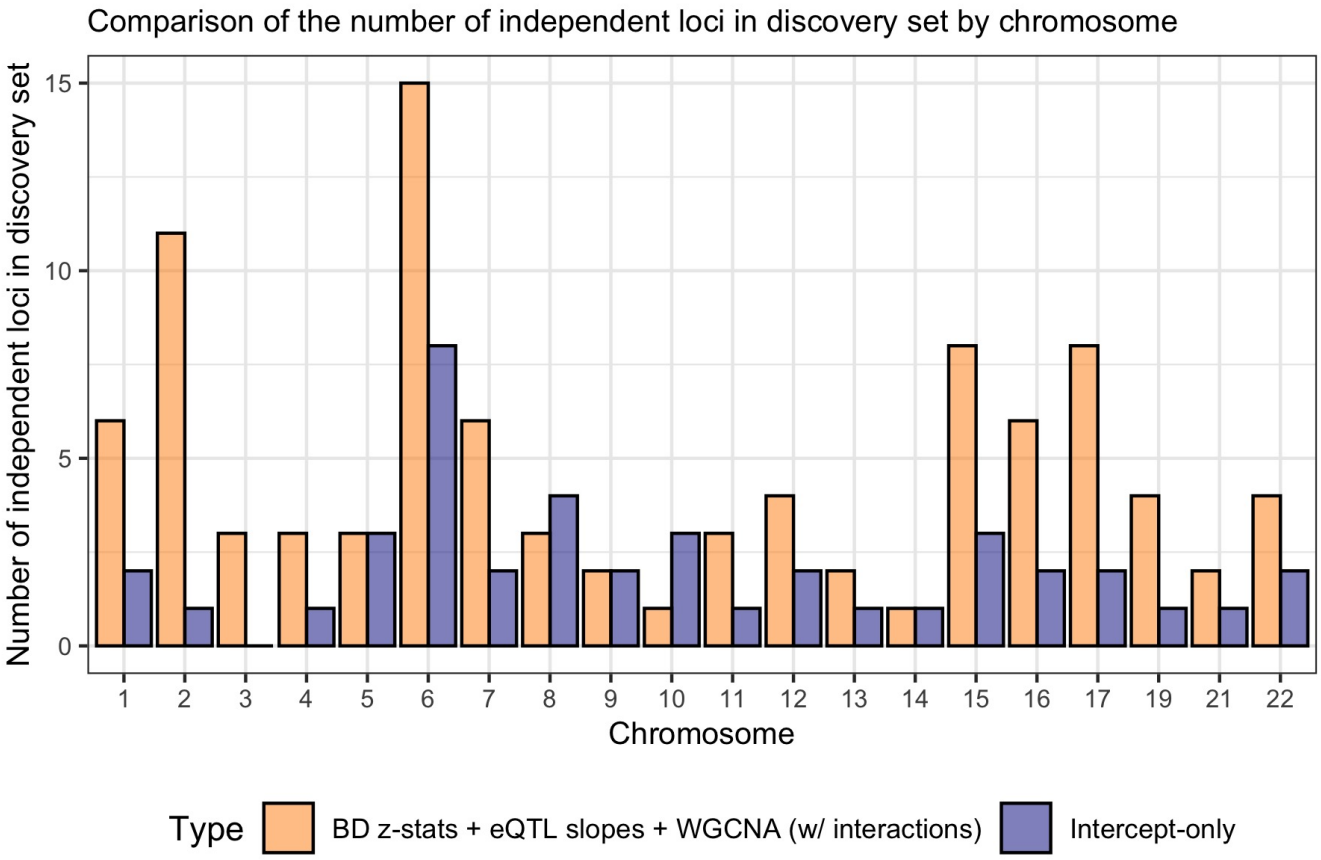

Fig. S3. Comparison of the number of independent loci in the AdaPT discovery sets by type for each chromosome.

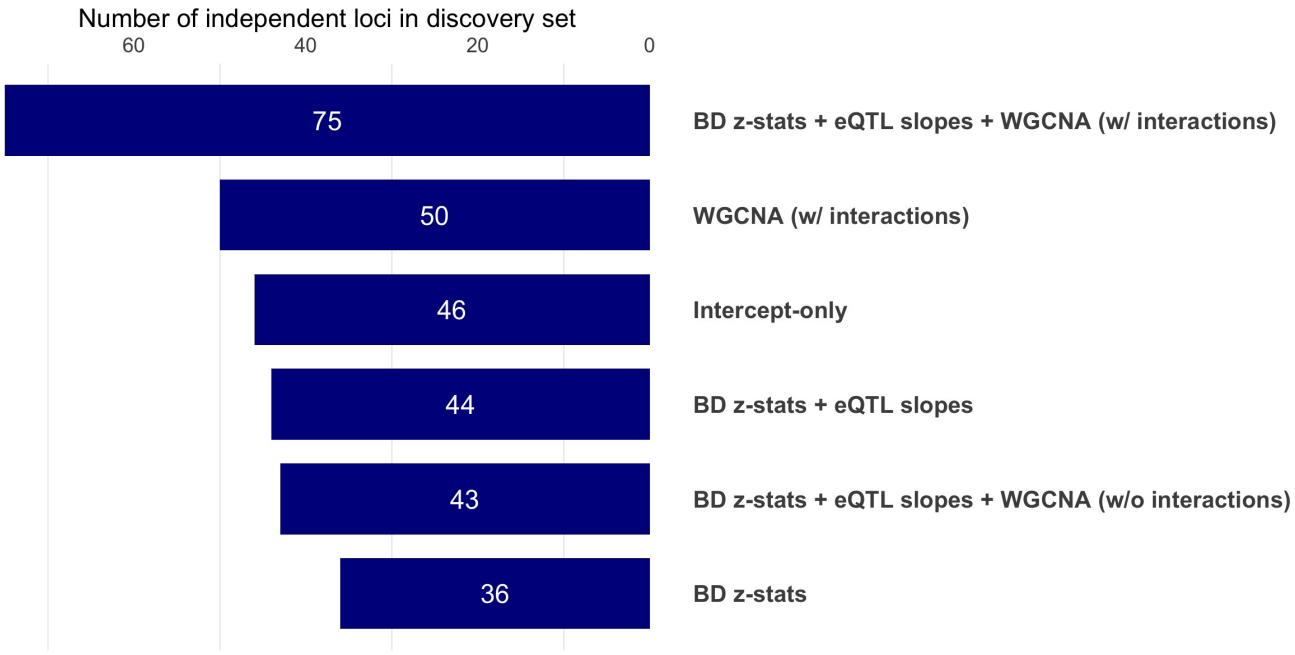

Fig. S4. Comparison of the number of independent loci for each discovery set at target $\alpha = 0.05$, based on LD pruning with the with *2014-only* SCZ p-values.

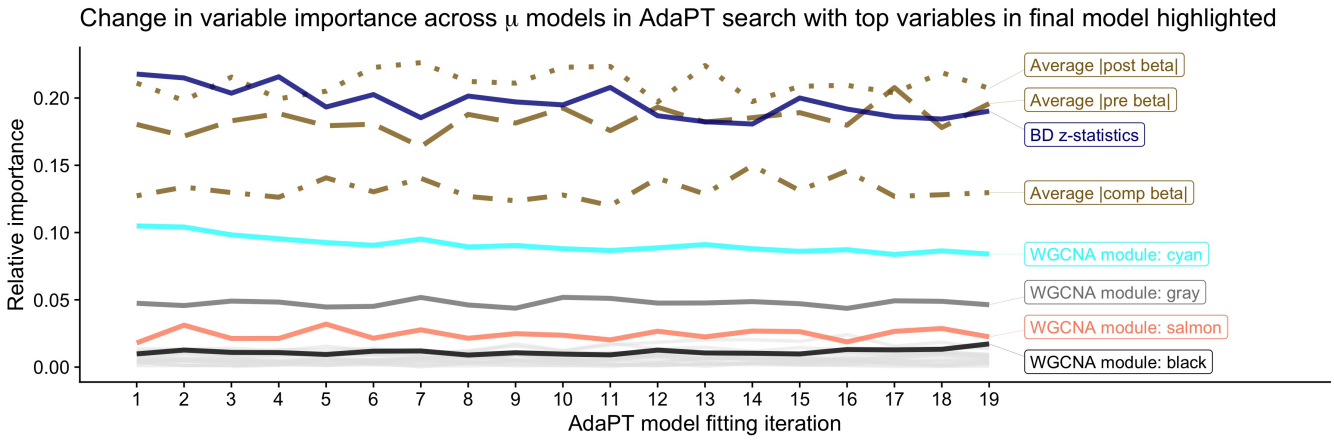

Fig. S5. Change in variable importance for AdaPT non-null effect size μ model across search, with top variables in final model highlighted.

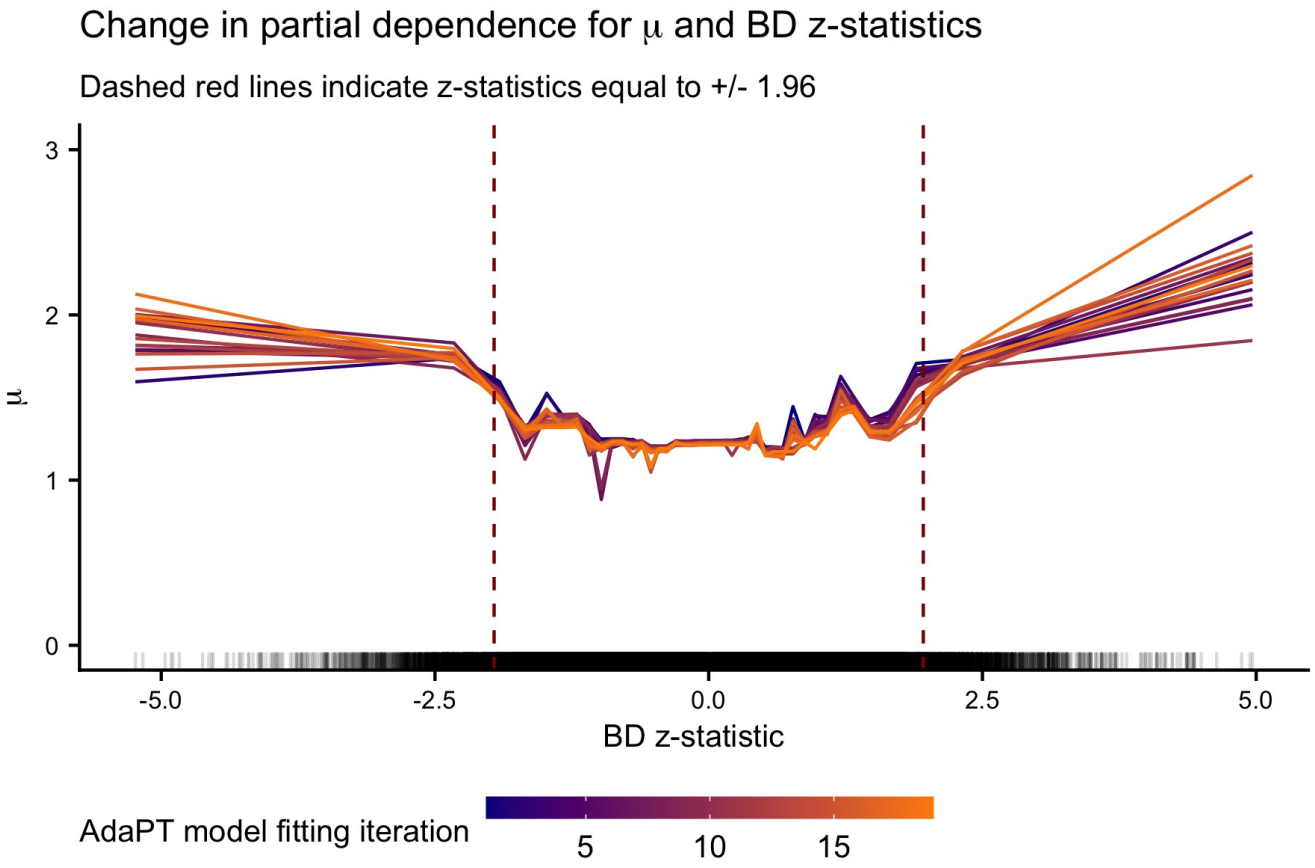

Fig. S6. Change in partial dependence for non-null effect size μ and BD z-statistics across μ models in AdaPT search.

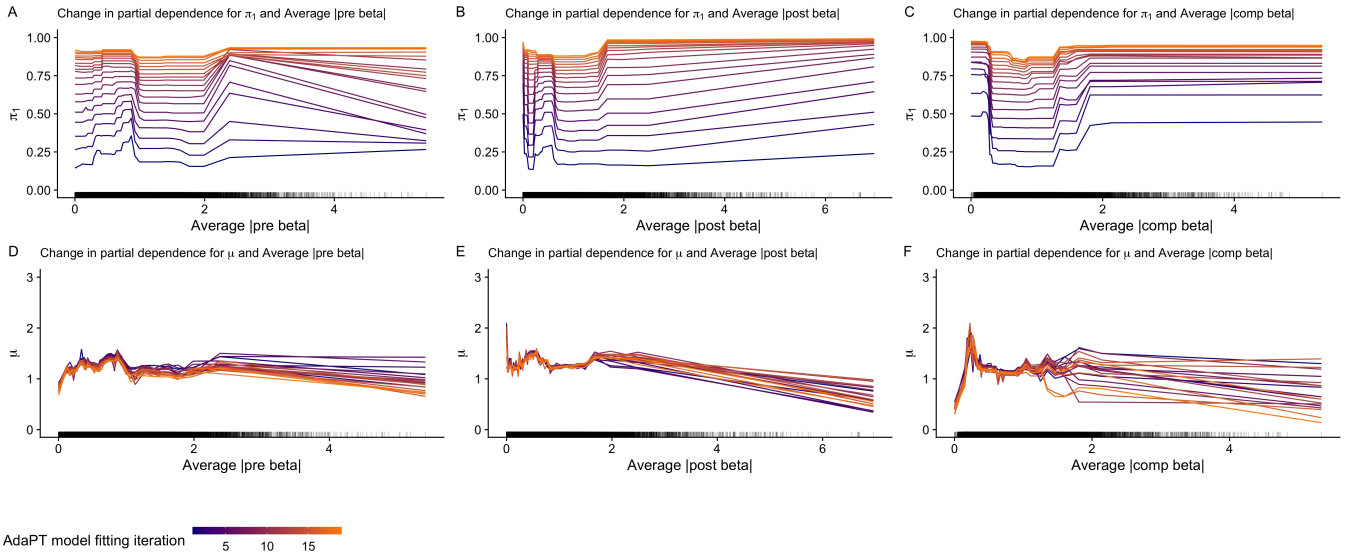

Fig. S7. Change in partial dependence plots for probability of being non-null π_1 in *(A-C)*, and the effect size under alternative μ in *(D-F)*, for each type of eQTL slope. Rugs along x-axis denote distribution of values for each variable.

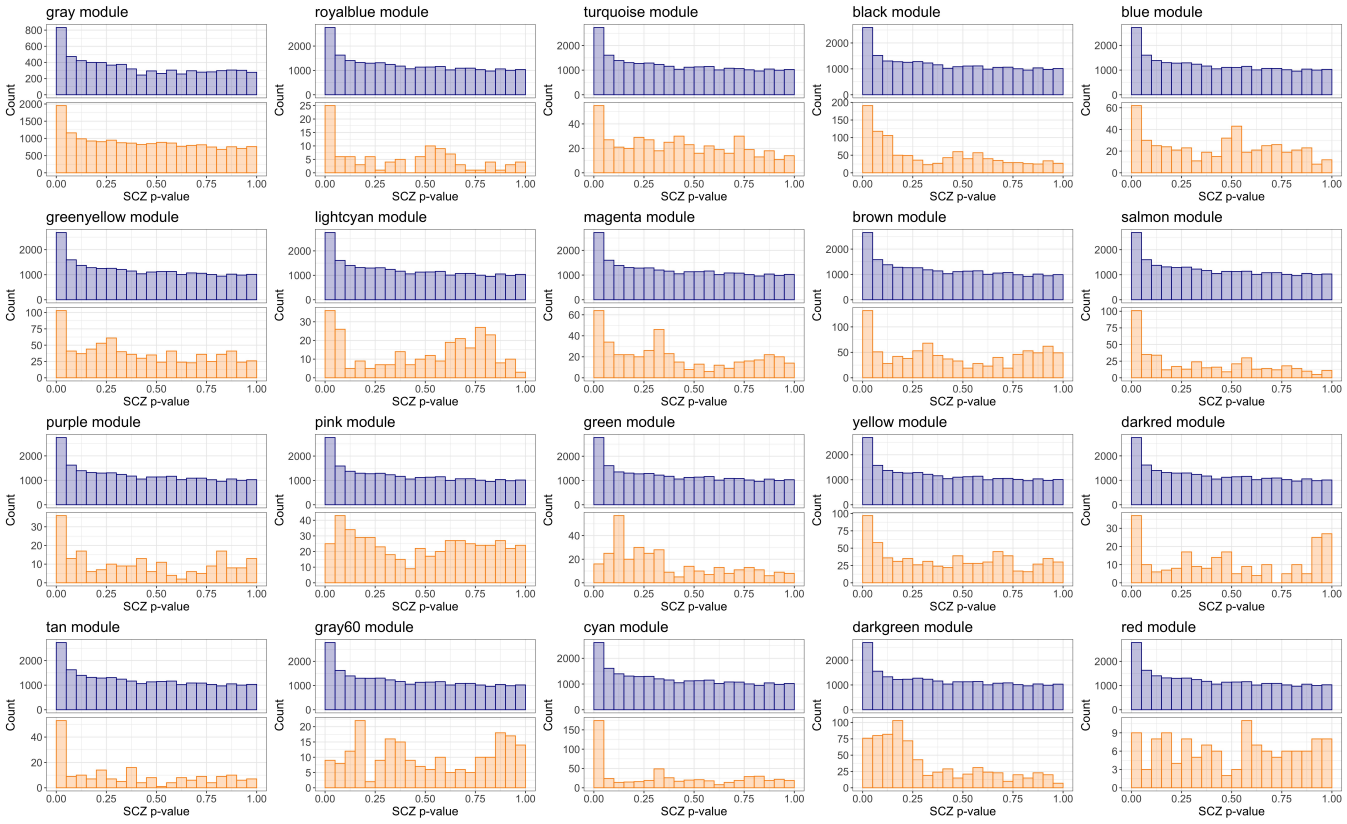

Fig. S8. Comparison of SCZ p-value distributions from *2014* studies by whether or not the eSNP had an associated cis-eQTL gene in the module.

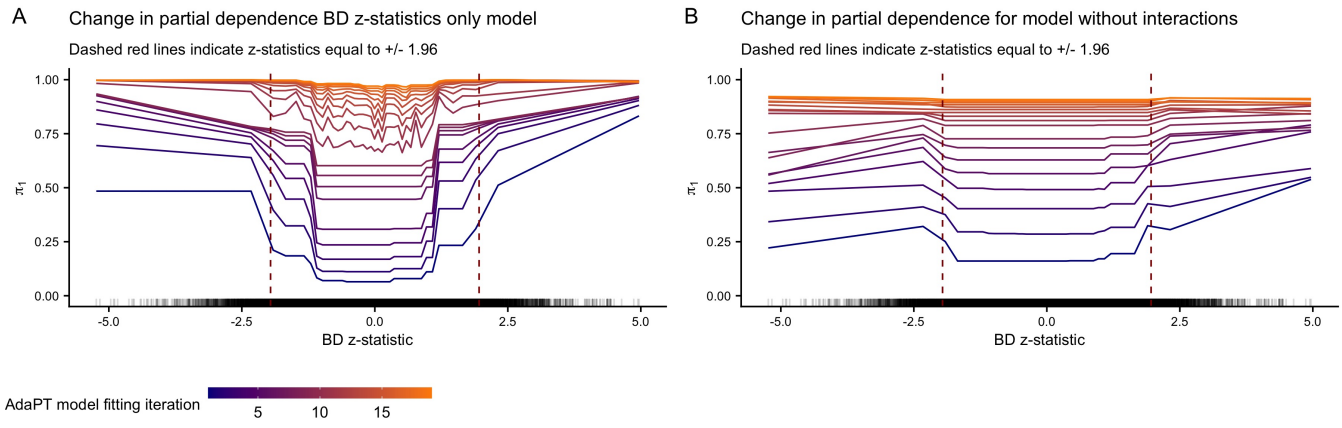

Fig. S9. Change in partial dependence for BD z-statistics and probability of being non-null π_1 for the AdaPT results using *(A)* only BD z-statistics and *(B)* all covariates without any interactions.

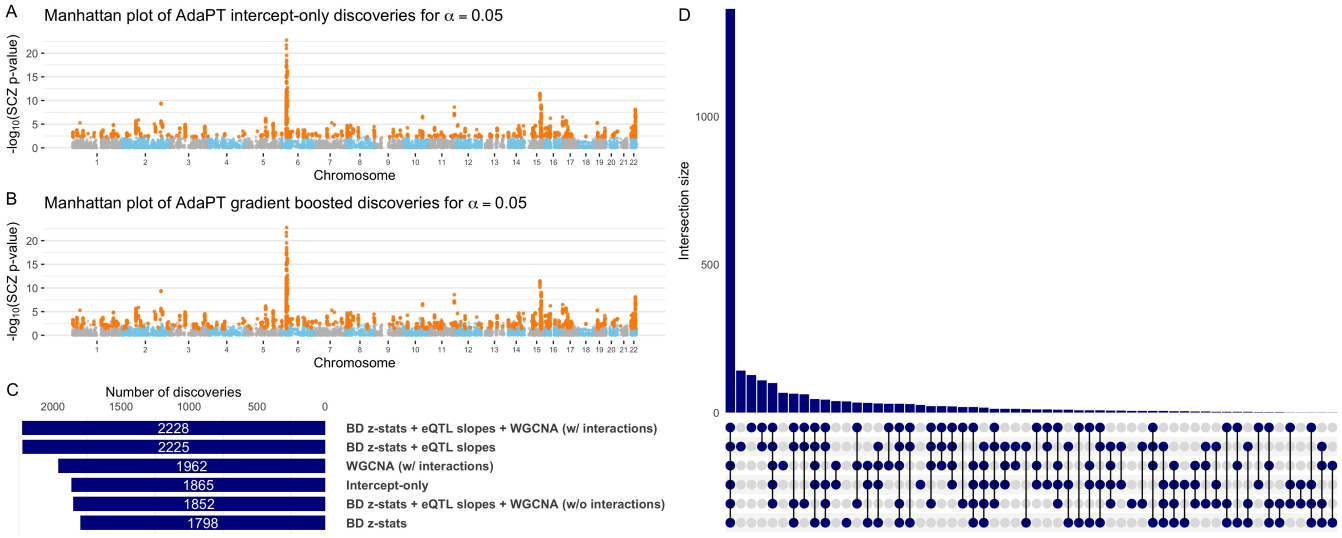

Fig. S10. Manhattan plots of SCZ AdaPT discoveries (in orange) with *all 2018* studies using *(A)* intercept-only model compared to *(B)* covariate informed model at target $\alpha = 0.05$. *(C)* Comparison of the number of discoveries at target $\alpha = 0.05$ for AdaPT with varying levels of covariates and *(D)* their resulting discovery set intersections.

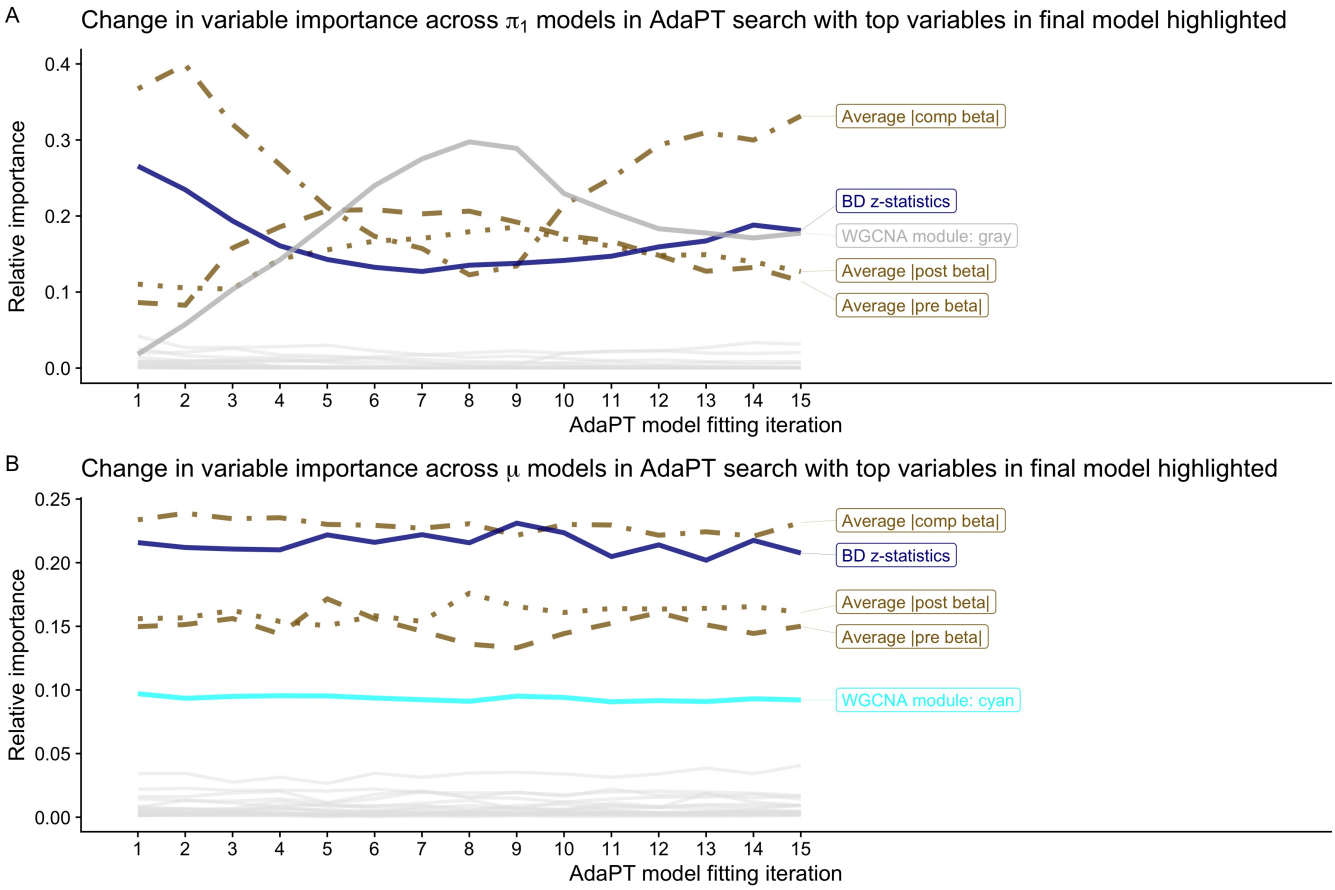

Fig. S11. Using *all 2018* studies: change in variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

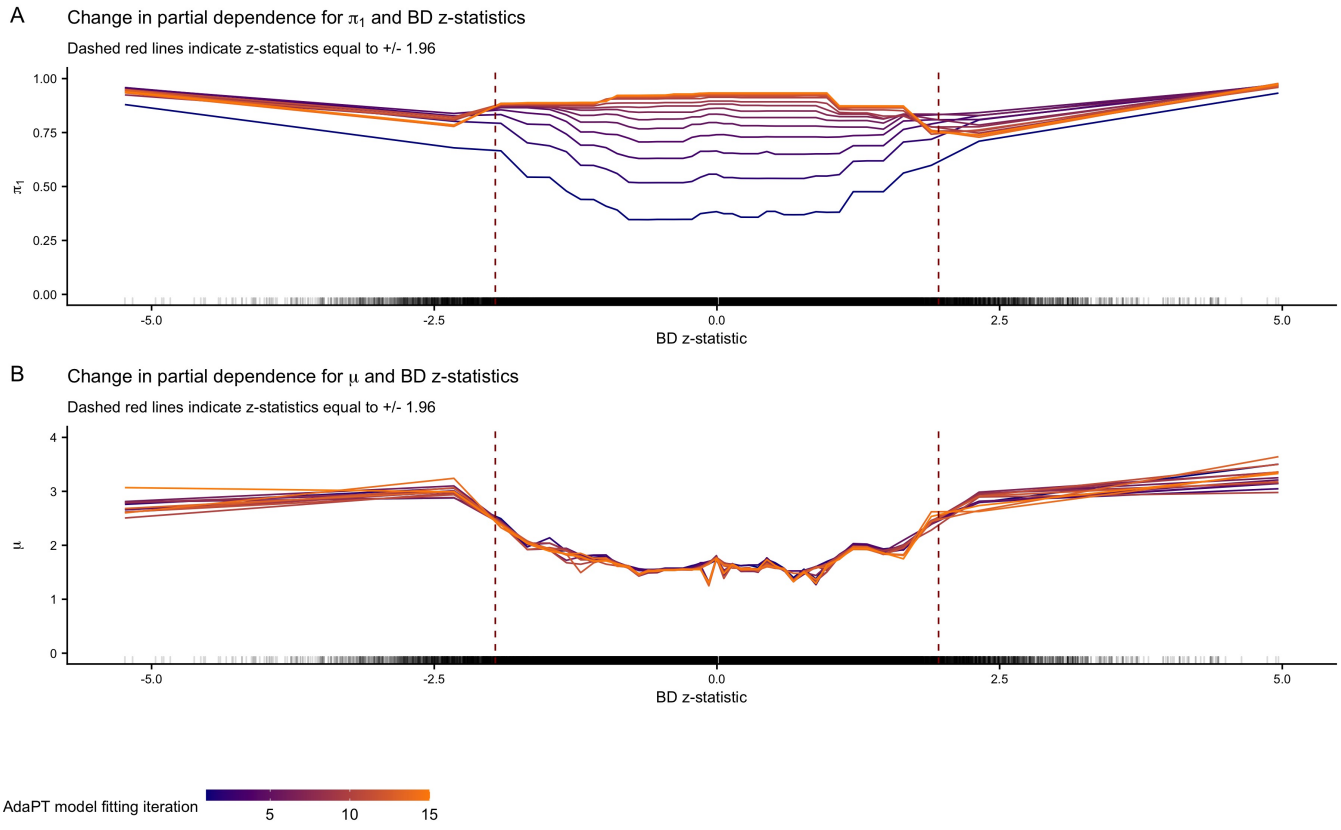

Fig. S12. Using *all 2018* studies: change in partial dependence for BD z-statistics and AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search.

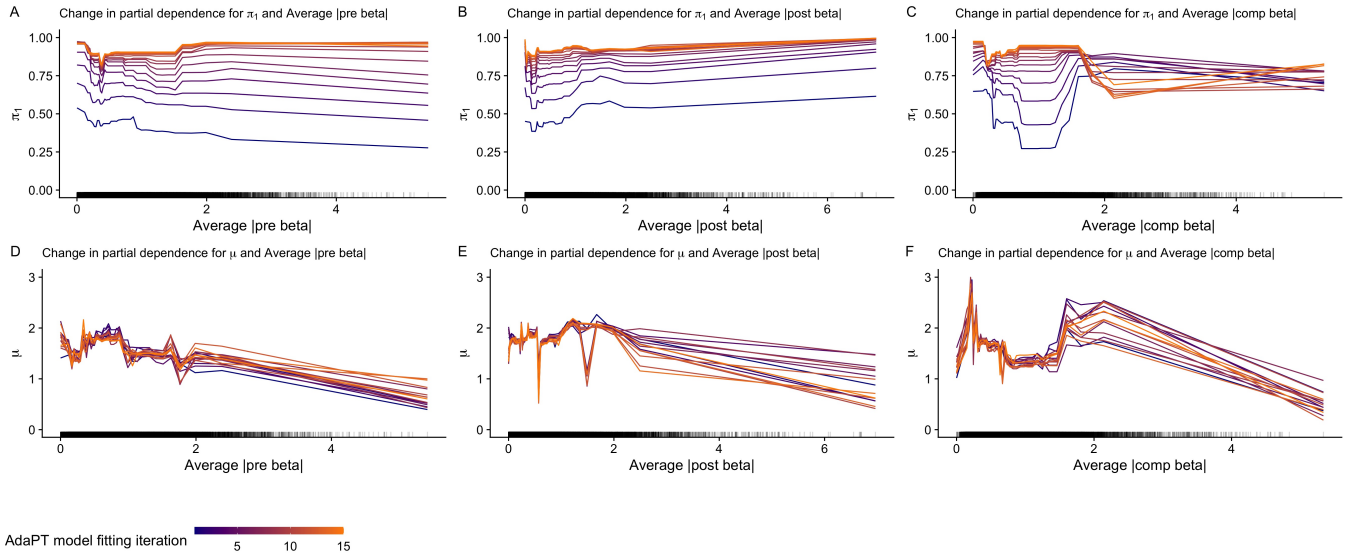

Fig. S13. Using *all 2018* studies: change in partial dependence plots for probability of being non-null π_1 in *(A-C)*, and the effect size under alternative μ in *(D-F)*, for each type of BrainVar eQTL slope. Rugs along x-axis denote distribution of values for each variable.

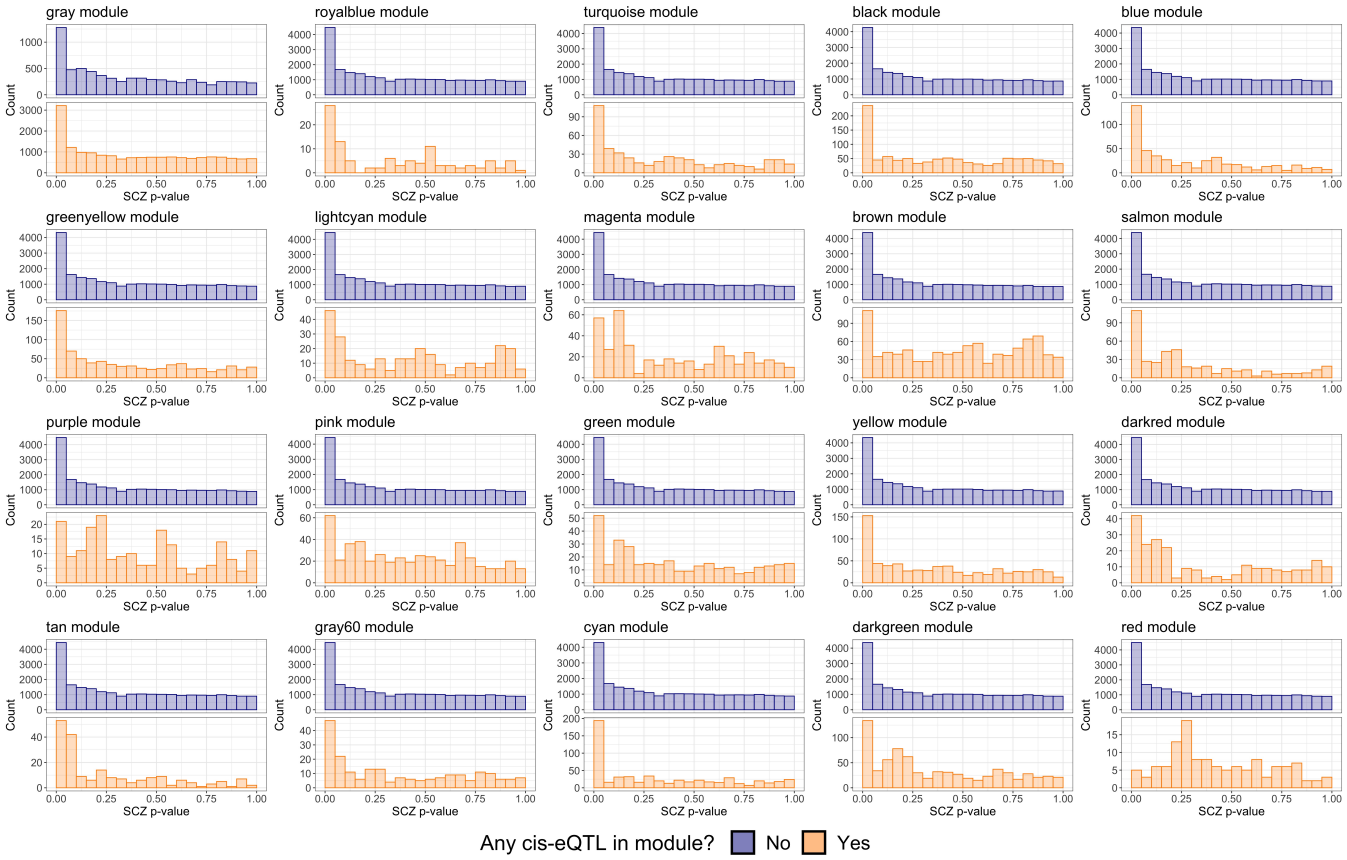

Fig. S14. Using *all 2018* studies: comparison of SCZ p-value distributions from *2014* studies by whether or not the eSNP had an associated cis-eQTL gene in the module.

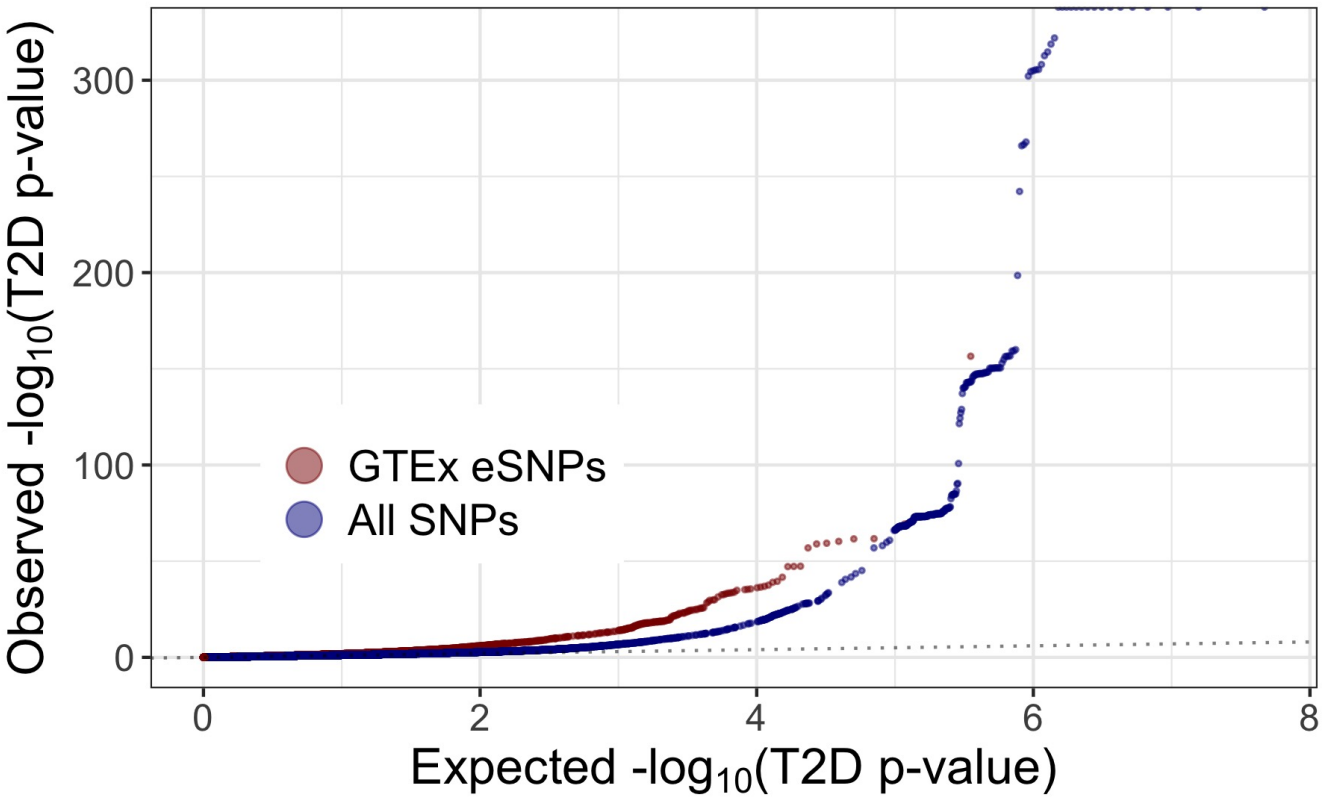

Fig. S15. A comparison of qq-plots revealing T2D enrichment for GTEx eSNPs compared to full set of SNPs.

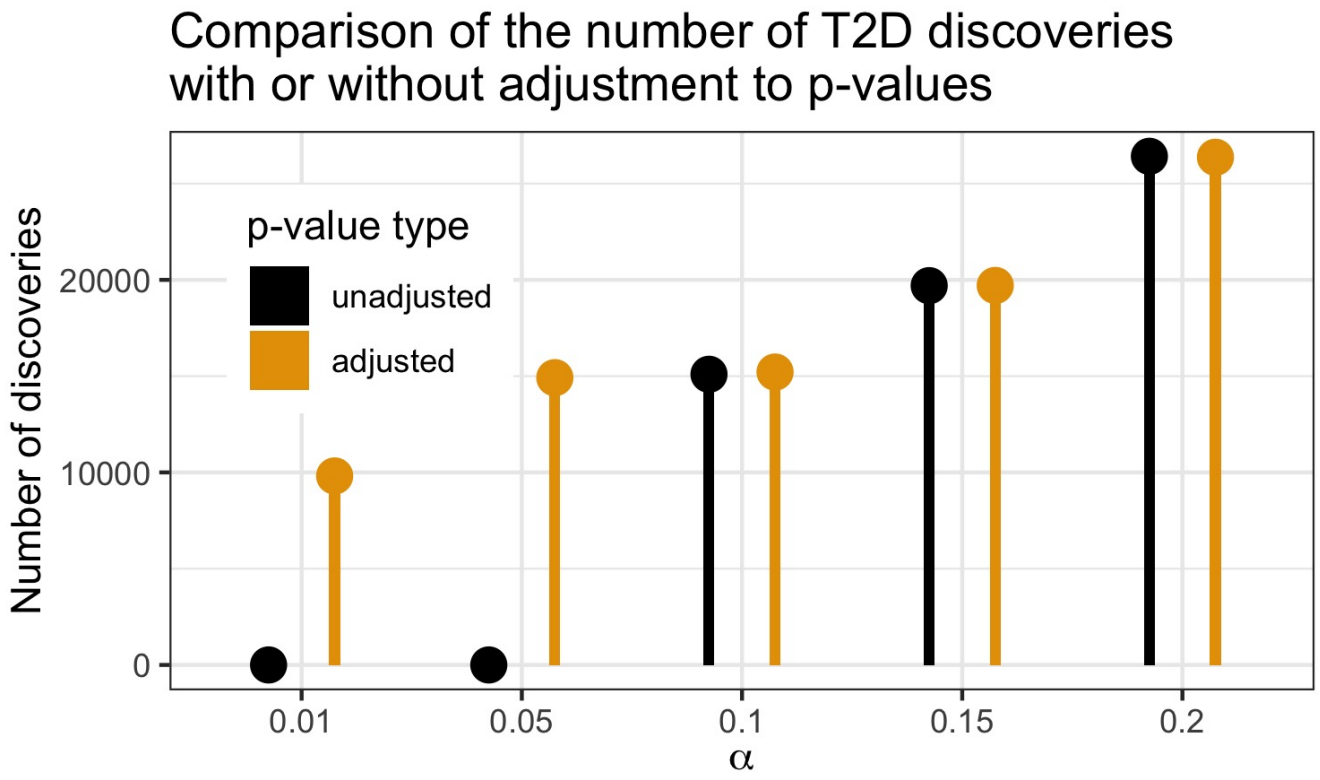

Fig. S16. Comparison of the number of discoveries by AdaPT for T2D by whether or not the adjusted or unadjusted p-values were used.

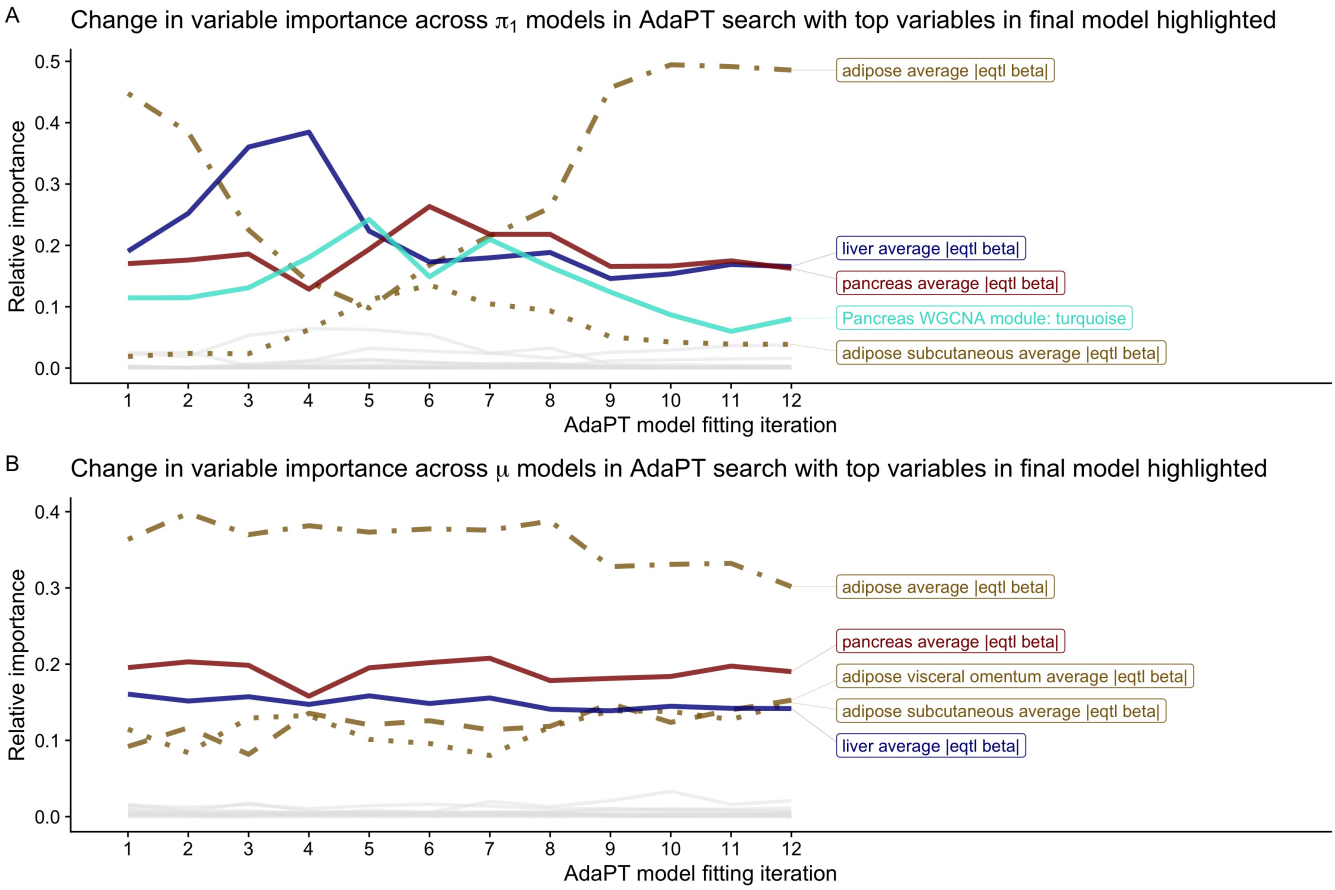

Fig. S17. Change in T2D variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

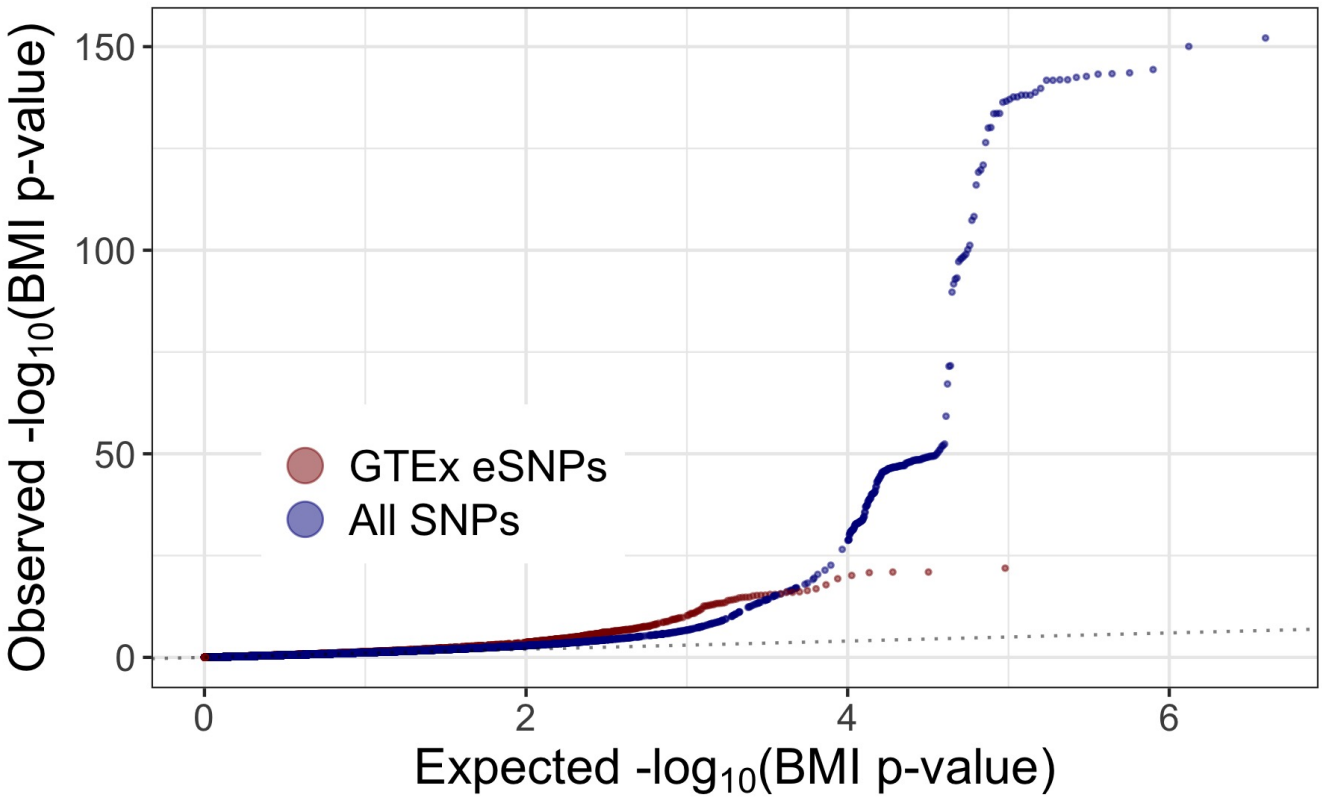

Fig. S18. Comparison of qq-plots revealing BMI enrichment for GTEx eSNPs compared to full set of SNPs.

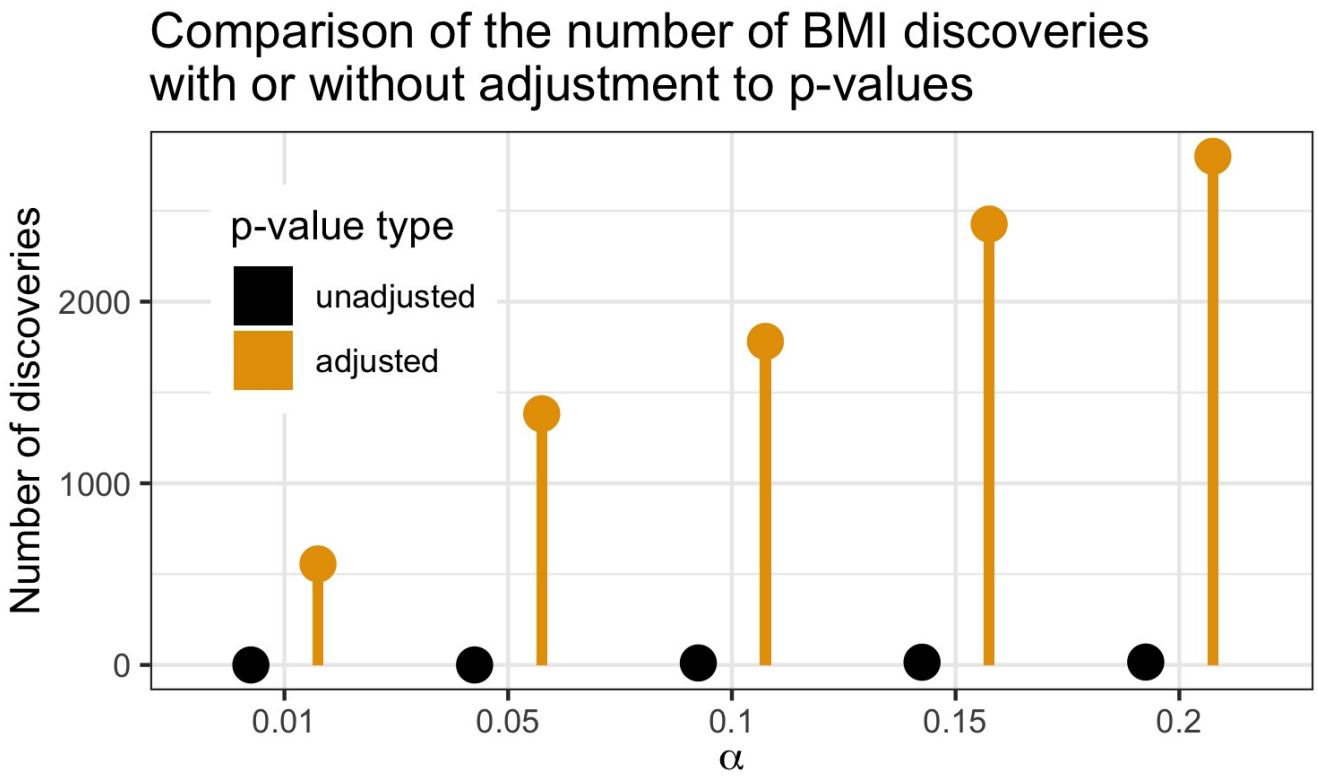

Fig. S19. Comparison of the number of discoveries by AdaPT for BMI by whether or not the adjusted or unadjusted p-values were used.

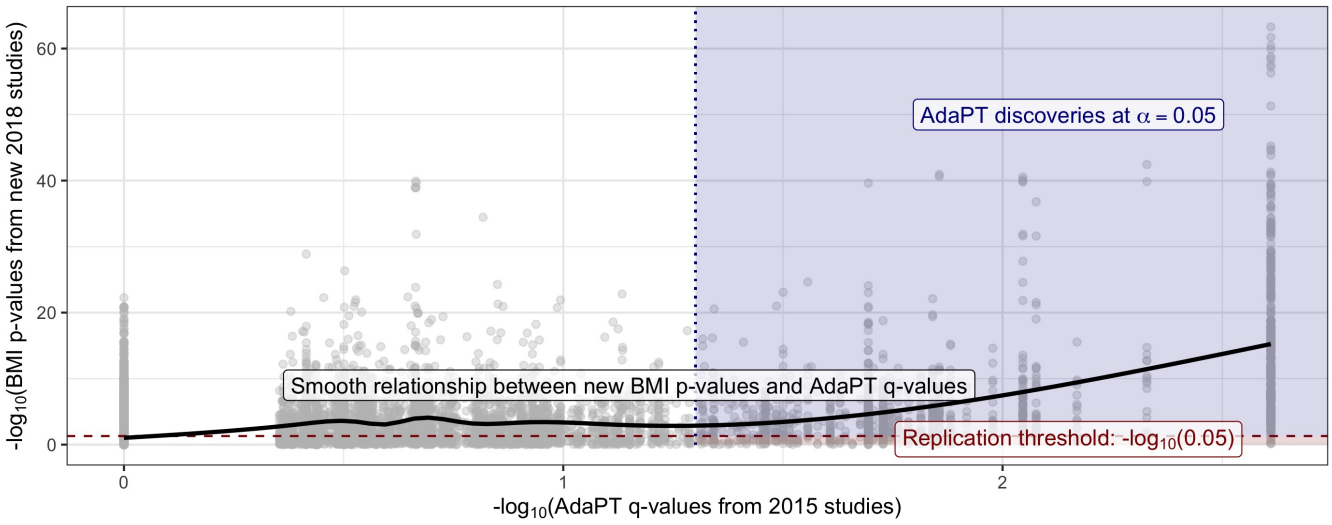

Fig. S20. Black line displays smooth relationship between BMI p-values from *2018-only* studies and the AdaPT q-values from the *2015-only* studies. Blue-shaded region indicates AdaPT discoveries at $\alpha = 0.05$ that are nominal replications, p-values from the *2018-only* studies < 0.05 while red denotes discoveries which failed to replicate.

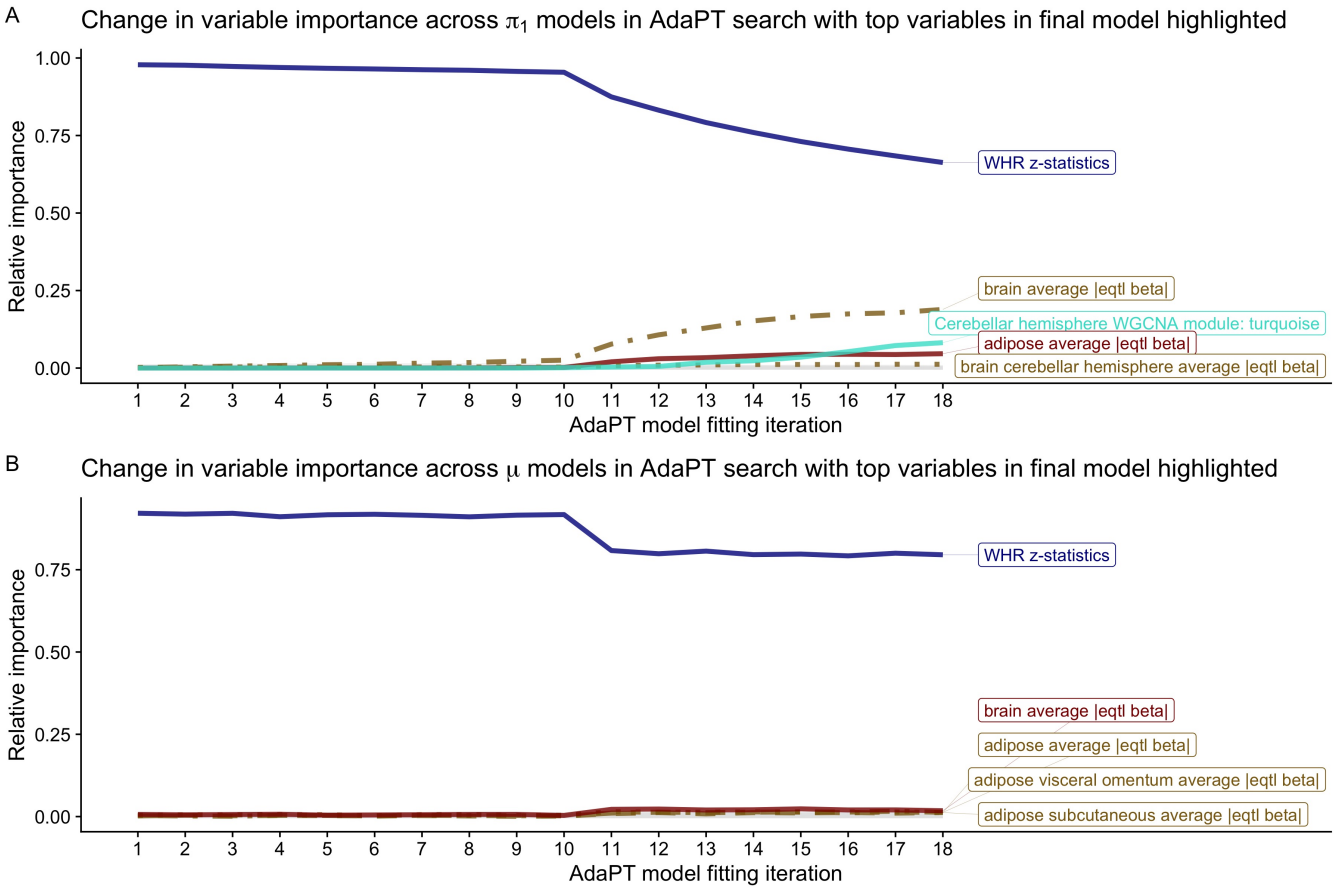

Fig. S21. Change in BMI variable importance for AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search, with top variables in final model highlighted.

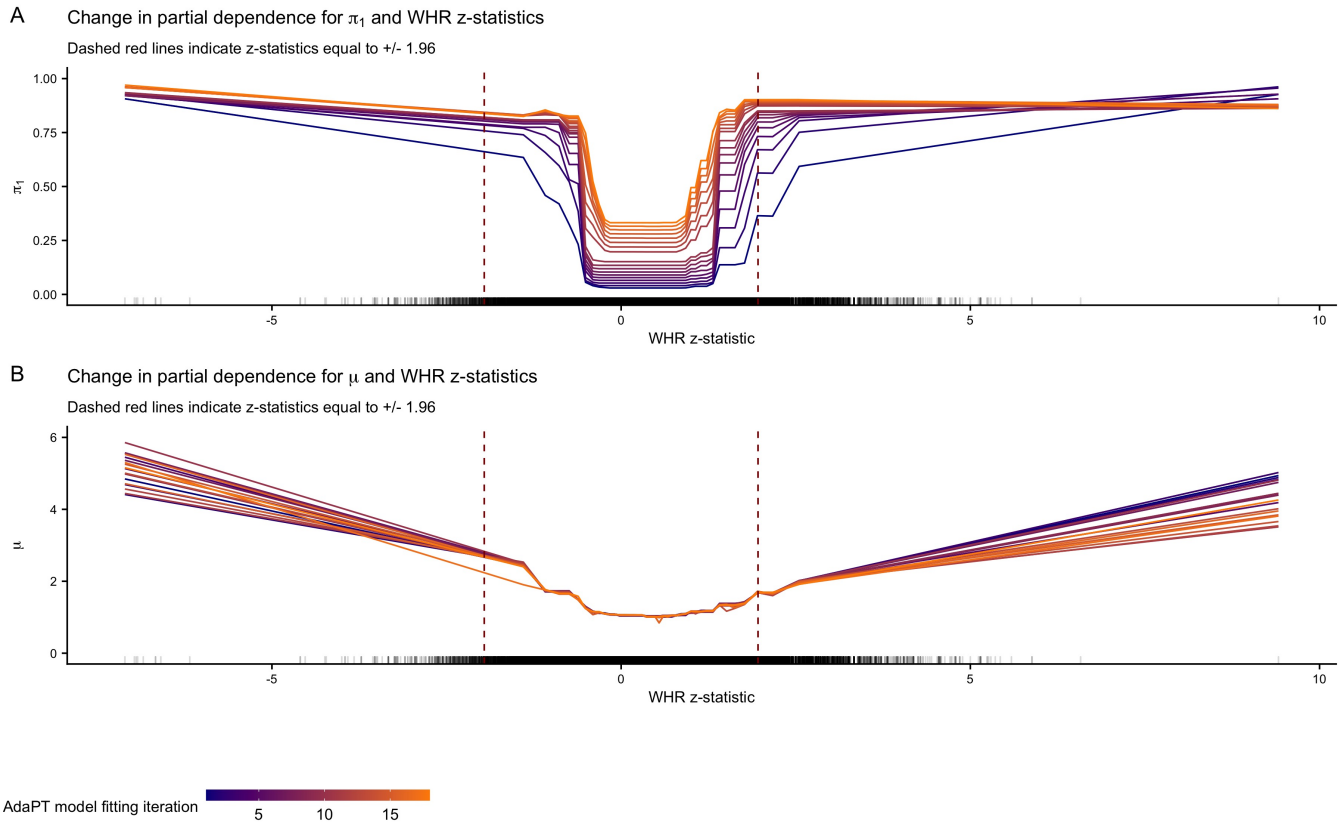

Fig. S22. Change in BMI partial dependence for WHR z-statistics and AdaPT *(A)* probability of being non-null π_1 and *(B)* effect size under alternative μ models across search.

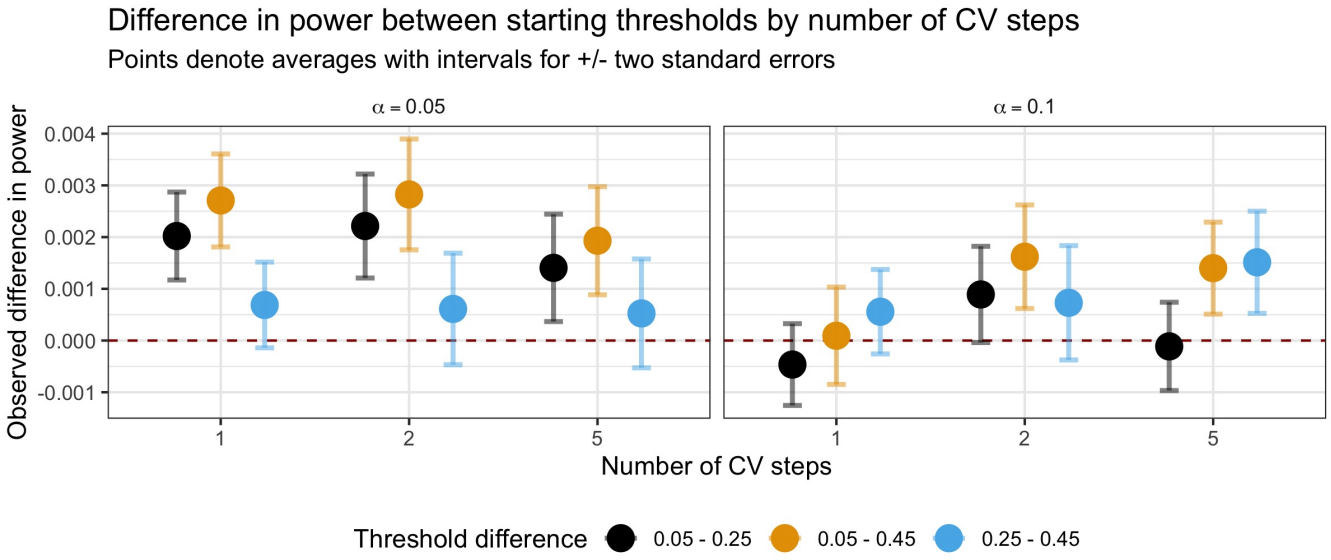

Fig. S23. Difference in simulation power between different initial thresholds s_0 for AdaPT search by number of CV steps. Points denote averages with plus/minus two standard error bars.

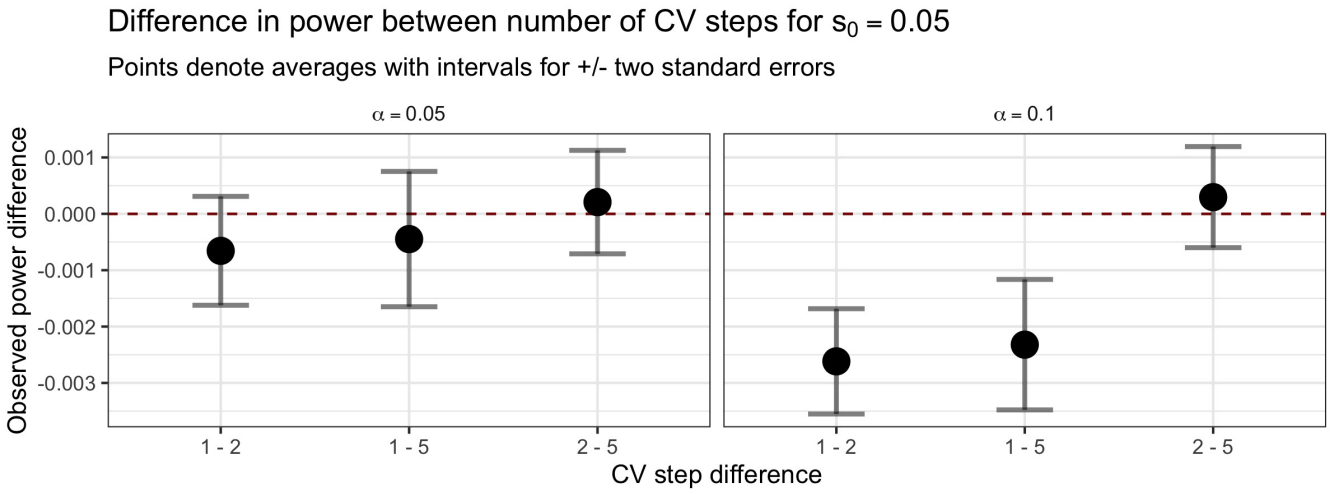

Fig. S24. Difference in simulation power between the number of CV steps with $s_0 = 0.05$. Points denote averages with plus/minus two standard error bars.

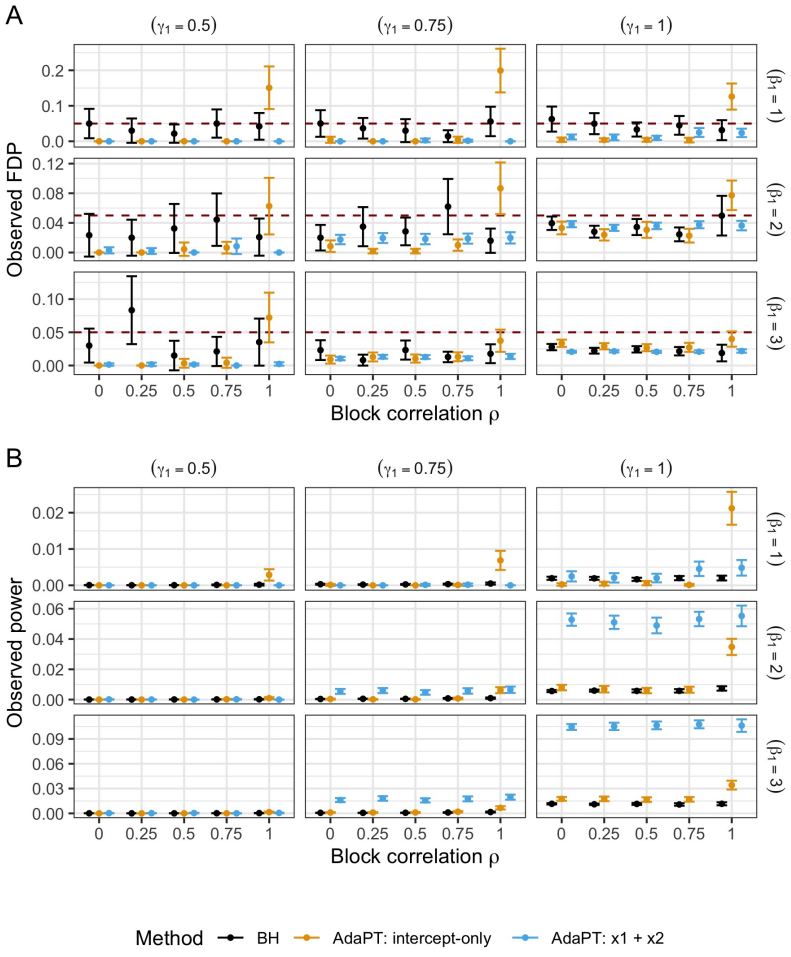

Fig. S25. Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 0.5$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

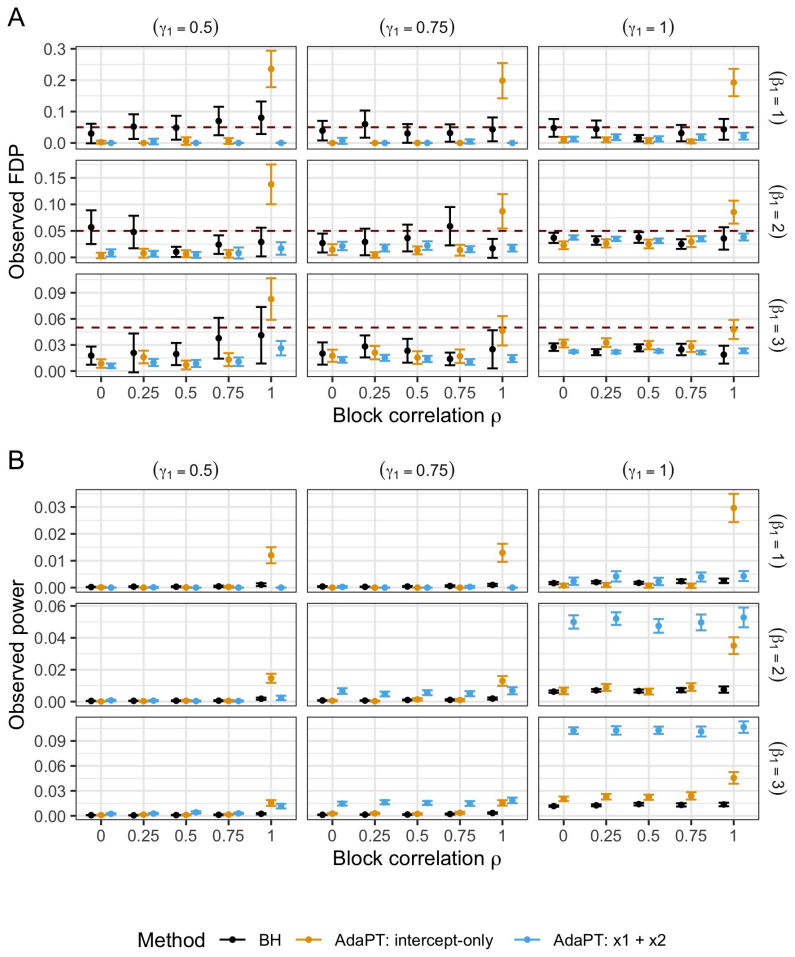

Fig. S26. Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 1$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

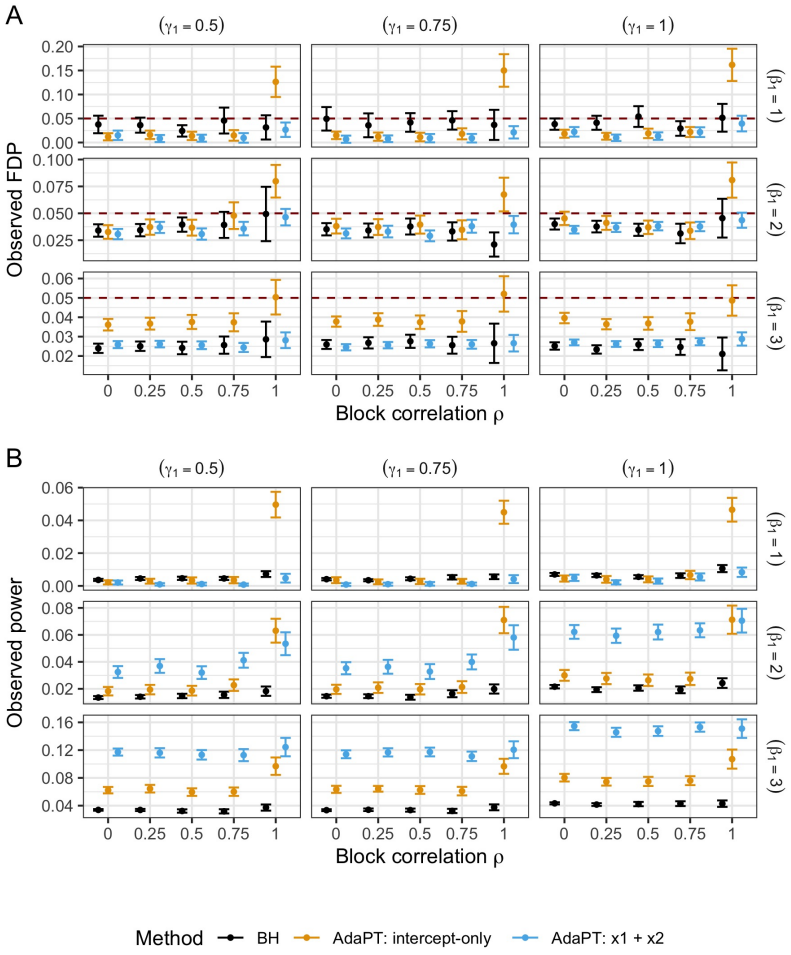

Fig. S27. Comparison of average (A) FDP and (B) power with plus/minus two standard error bars for 100 simulations with $\mu_{floor} = 1.5$, and varying values for β_1 (rows) and γ_1 (columns) and block correlation ρ.

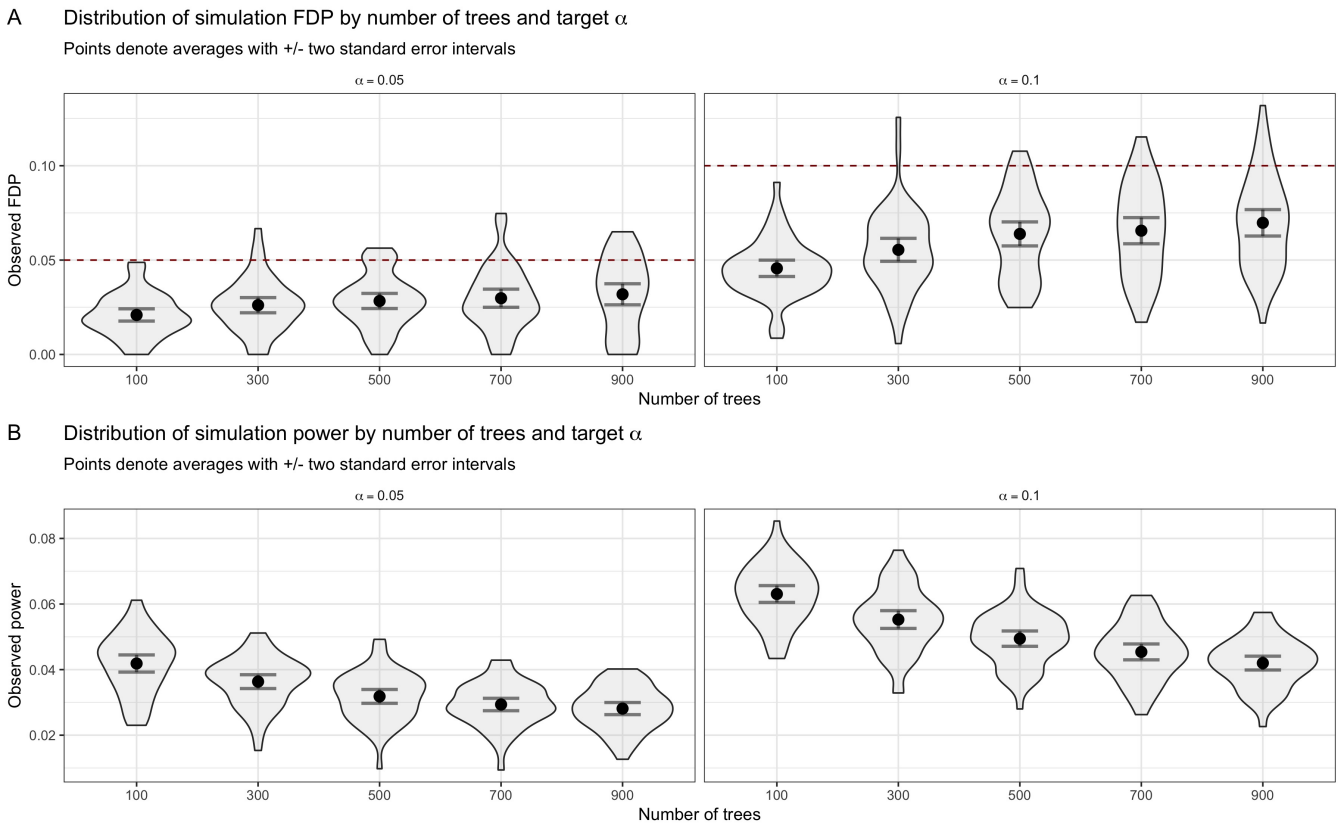

Fig. S28. Distributions of observed *(A)* FDP and *(B)* power for simulations as the number of AdaPT gradient boosted trees increases by target FDR level α. Points denote averages with plus/minus two standard error intervals.

Table S1. Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by covariates for eSNPs in each CV step.

Covariates	m_1^*	m_2^*
BD z-stats	$P = 150,\ D = 1$	$P = 150,\ D = 6$
BD z-stats + eQTL slopes	$P = 150,\ D = 3$	$P = 150,\ D = 6$
BD z-stats + eQTL slopes + WGCNA	$P = 150,\ D = 3$	$P = 150,\ D = 3$
WGCNA only	$P = 150,\ D = 3$	$P = 150,\ D = 3$

Table S2. Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by GWAS results in each CV step.

GWAS results	m_1^*	m_2^*
T2D	$P = 100,\ D = 2$	$P = 150,\ D = 3$
BMI (all covariates)	$P = 100,\ D = 2$	$P = 150,\ D = 3$
BMI (WHR z-stats only)	$P = 150,\ D = 1$	$P = 150,\ D = 1$
BMI (WHR z-stats + eQTL slopes)	$P = 100,\ D = 3$	$P = 150,\ D = 3$

Table S3. Selected boosting settings for number of trees P and maximum depth D with AdaPT CV algorithm by covariates for eSNPs with $s_0 = 0.45$.

Covariates	m_1^*	m_2^*
BD z-stats	$P = 50,\ D = 1$	$P = 150,\ D = 1$
BD z-stats + eQTL slopes	$P = 100,\ D = 1$	$P = 150,\ D = 2$
BD z-stats + eQTL slopes + WGCNA	$P = 100,\ D = 2$	$P = 150,\ D = 3$
WGCNA only	$P = 150,\ D = 3$	$P = 150,\ D = 3$

SI Dataset S1 (adapt_gwas_results.xlsx)

AdaPT discoveries using gradient boosted trees at target FDR level $\alpha = 0.05$ for SCZ, T2D, and BMI. Sheets contain each unique combination of eSNP and associated cis-eQTL gene.

References

1. L Lei, W Fithian, Adapt: an interactive procedure for multiple testing with side information. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **80**, 649–679 (2018).
2. B Efron, R Tibshirani, JD Storey, V Tusher, Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
3. T Chen, C Guestrin, Xgboost: A scalable tree boosting system in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. (ACM, New York, NY, USA), pp. 785–794 (2016).
4. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
5. 1000 Genomes Project Consortium and others, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).
6. Z Ren, E Candès, Knockoffs with side information. *arXiv preprint arXiv:2001.07835* (2020).
7. DM Werling, et al., Whole-genome and rna sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *bioRxiv* (2019).
8. A Mahajan, et al., Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
9. N Xiao, G Wang, L Sun, *grex: Gene ID Mapping for Genotype-Tissue Expression (GTEx) Data*, (2018) R package version 1.8.
10. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2018).
11. M Ashburner, et al., Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
12. The Gene Ontology Consortium, The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* **47**, D330–D338 (2018).
13. AE Locke, et al., Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197 EP – (2015).
14. L Yengo, et al., Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
15. D Shungin, et al., New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518** (2015).

16. GTEx Consortium, The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
17. T Chen, et al., *xgboost: Extreme Gradient Boosting*, (2019) R package version 0.81.0.1.