2020

Truncated and Aggregated P Value Test

lei shi

TRUNCATED AND AGGREGATED P VALUE TEST

by

Lei Shi

A Dissertation

Submitted in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

May 2020

Shi, Lei, Ph.D. The University of Memphis, May, 2020. Truncated and aggregated P value test. Major Professor: E. Olusegun George.

High throughput genome screening techniques are enabling researchers to interrogate human genome at single base pair level for their association with outcomes by genome wide association study (GWAS). However, it is usually challenging for GWAS to provide clear statistical conclusions at the gene level when multiple genomic features, either from same platform or different platforms, reside in the same gene. Traditionally a gene is considered as associated with an outcome when at least one genome feature within that gene is significantly associated with the outcome after adjusting for multiple genome-wide tests. Under that framework, only the most significant genome feature is used to determine the gene/outcome association. However adjustments for multiple testing impose a large penalty on single feature from high density arrays such as Affymetrix SNP6 arrays. Here we propose a procedure based on truncated and aggregated P values (TAP) to aggregate individual genome feature P-values within designated allele/gene. We then construct a hybrid permutation test to obtain a single P-value for the allele/gene in order to assess the overall association of the segment with clinical outcome.

Table of Contents

List of Figures

List of Tables

1. Introduction

Abundant genomic data developed and accumulated in the past two decades have provided researchers numerous targets for subsequent laboratory work. Various techniques have facilitated the interrogation of human genome from different aspects, such as single nucleotide polymorphism (by SNP array), copy number variation (by SNP array), gene expression (by expression array or RNAsq), DNA methylation (by methylation array) and mutation (by Whole exon sequencing or RNAsq). Human genome can be screened at base pair level rapidly and efficiently. All the high-throughput platforms can be categorized into two groups: array-based platforms and sequencing based platforms.

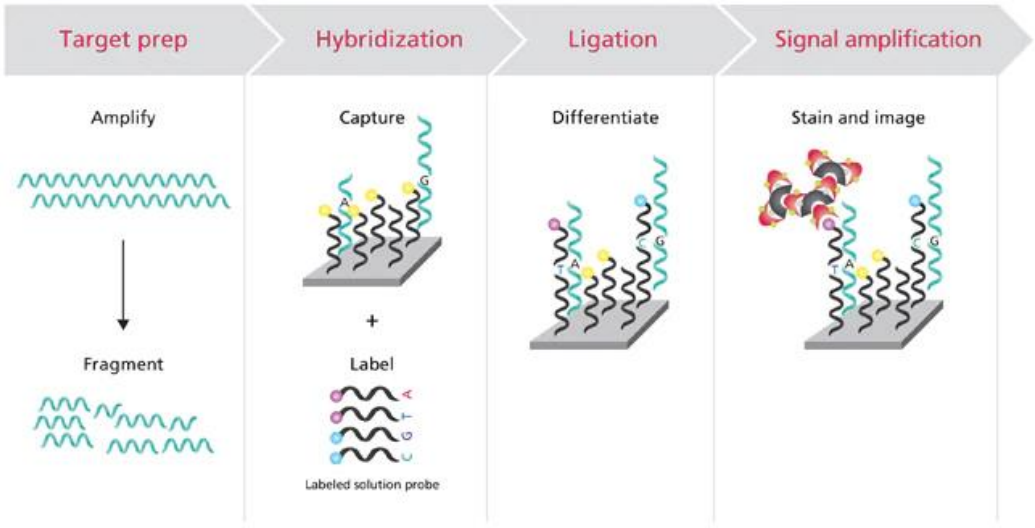

Figure 1. Array-based platform principles.

Source: https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/agrigenomics-solutions-microarrays-gbs/axiom-genotyping-solution-agrigenomics.html

Array-based platforms includes SNP microarray, gene expression array and DNA methylation array and are all based on hybridization of two DNA strands. A collection of microscopic DNA spots is attached to a solid surface, either glass plate or glass bead. Then a sample DNA is

degraded into small pieces by DNA enzyme and presented to the array. When complementary DNA pieces from sample bind to those spots on glass tube, the fluorescence intensity will reflect the relative abundance of certain DNA target in the sample (Figure 1). There could be hundreds of thousands even millions of DNA spots in one microarray, all of which will be examined simultaneously. Array-based platform has fixed number of probes in each array which will cover fixed number of base pairs in whole genome. Table 1, gives list of major commercially available genomic arrays that are used at St Jude Children's Research Hospital.

Table 1. List of major genomic arrays used a St Jude Children's Research Hospital

Product Name	Company	Platform	# of Probes
GeneChip® Human Mapping 50K	Affymetrix	SNP	50,000 SNPs for Xba240 50,000 SNPs for Hind240
GeneChip® Human Mapping 500K	Affymetrix	SNP	262,000 SNPs for 250K Nsp 238,000 SNPs for 250K Sty
Genome-Wide Human SNP Array 6.0	Affymetrix	SNP	946,000 CN probes 906,600 SNP probes
Human Genome U133 Plus 2.0 (U133)	Affymetrix	Expression	37,000 genes
Human Exon 1.0 ST (HuEx)	Affymetrix	Expression	1,400,000 Exon probes
Human 2.0 ST (HuGene)	Affymetrix	Expression	24,838 RefSeq genes
Infinium Omni2.5Exome-8	illumina	SNP	2,618,000 SNPs
Infinium MethylationEPIC	illumina	Methylation	850,000 probes

In comparison to array-based platforms, sequencing-based platforms, such as RNAseq, whole exon sequencing and whole genome sequencing, can cover every base pairs in their targeted genome region. But there is no consensus among researchers how to summarize sequencing data for statistical test. In this project we will focus on array-based platform to illustrate our statistical procedure.

2. Literature Review

2.1 Genome wide association study

Genetic variants, also known as genetic features, observed in a cohort by those high throughput DNA arrays can be used to test association with clinical outcomes. Since hundreds of thousands,

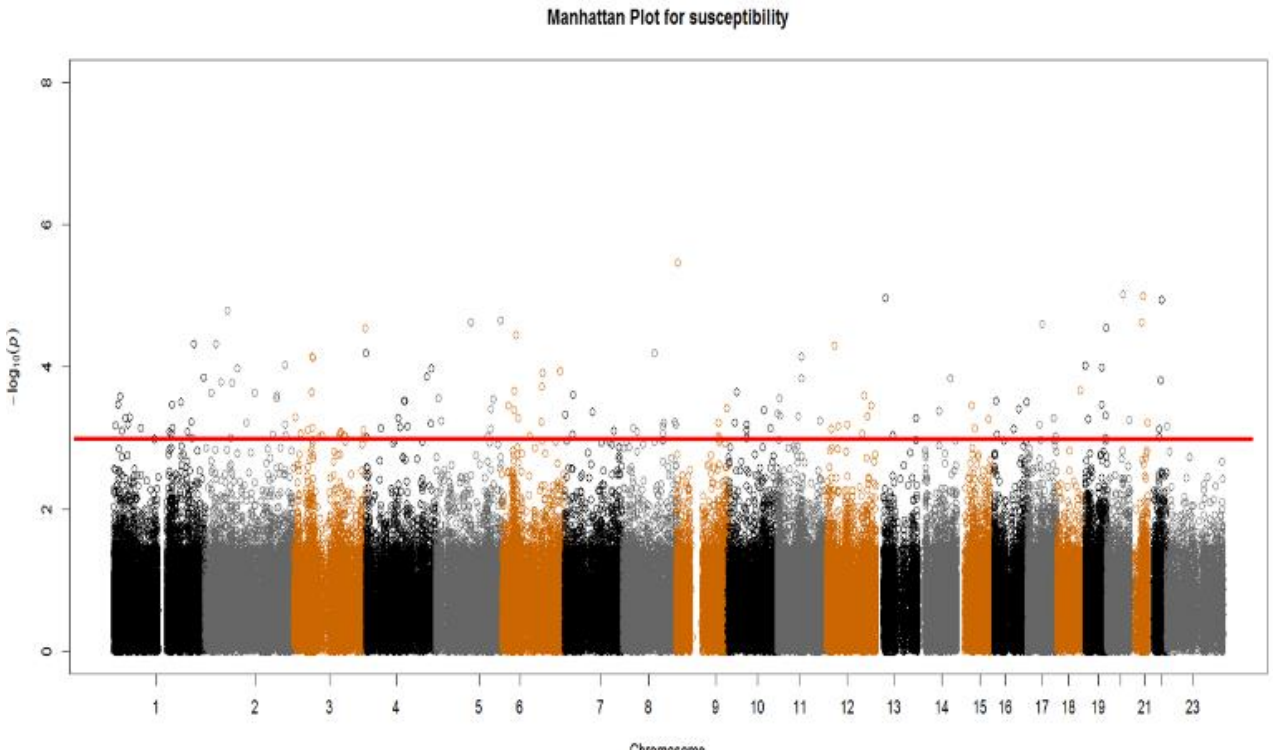

Figure 2. Manhattan Plot.

X-axis is the genome location of each genomic feature. Y-axis is -log_{10} transformed p value for each feature. The red line represents -log_{10}(0.01).

even millions of features are interrogated simultaneously, sample size is usually much smaller than the variables observed. Therefore, it is mathematically impossible to estimate feature effects

in a linear model to cover whole genome area. Dimension reduction method such as principal factor analysis or variable selection methods such as lasso are needed to make data analysis feasible. Furthermore, when multiple outcomes are available for one cohort, neither method is helpful in testing all features and all outcomes in one approach. In genome-wide association studies (GWAS) these features are usually tested individually. Features with P-value that is less than a certain threshold, are selected as potential targets for further research in wet-lab. The P-values can be summarized and visualized using Manhattan plot (Fig 2). The traditional concepts of type I and type II error in single P-value can be applied to multiple tests performed in series to answer one question. In Table 2, n_1 represents the number of type I errors, which is rejecting a true null hypothesis, of the series tests, while n_4 is the type II errors, which is fail to reject a false null hypothesis. Every single test has a small probability to contribute to n_1 which is the type I error. With no adjustments to the P-values, n_1 would increase when N become large leading to an inflated overall type I error and hence an overall test that is not conservative.

Table 2 Test result versus actual result (Hochberg & Benjamini, 1995).

	H_0 is true	H_a is true	Total
Number of significant tests	n_1	n_2	S
Number of non-significant tests	n_3	n_4	N-S
Total	m0	ma	N

Two approaches have been proposed to control type I errors in multiple tests: Control of the family-wise error rate (FWER) and the control of false discovery rate (FDR).

2.1.1 Family-wise error rate (FWER)

The FWER is defined by the probability of making at least one type I error in a series of

hypothesis testing.

$$FWER = \Pr(n_1 \geq 1) = 1 - \Pr(n_1 = 0)$$

FWER control focuses on minimizing FWER. Controlling FWER to a level α usually requires that the level of type I error on individual tests be adjusted. The procedures are considered as conservative with low power although they ensure high specificity.

FWER procedures include:

1. Bonferroni procedure

   Let α be significance threshold for single test. Then we can reject H_i when

   $$p_i < \frac{\alpha}{m}$$

2. Holm's step-down procedure (Holm, 1979)

   First all the P-values are ordered from lowest to highest, $p_{(1)}, p_{(2)}, \dots, p_{(m)}$ with corresponding hypothesis $H_{(1)}, H_{(2)}, \dots, H_{(m)}$. Find minimal k that fits following criteria

   $$p_{(k)} > \frac{\alpha}{m + 1 - k}$$

   Where k is the k^{th} P-value in ordered P values series.

   Then reject all the $p_{(i)}$ that i<k

3. Hochberg's step-up procedure (Hochberg, 1988)

   First all the P-values are ordered from lowest to highest, $p_{(1)}, p_{(2)}, \dots, p_{(m)}$ with corresponding hypothesis $H_{(1)}, H_{(2)}, \dots, H_{(m)}$. Find maximum k that fits following criteria

   $$p_{(k)} \leq \frac{\alpha}{m + 1 - k}$$

   Then reject all the $p_{(i)}$ that i<k

In 1996 Aickin M et al showed that Holm's step-down procedure is uniformly more powerful

than Bonferroni procedure (Aichin & Gensler, 1996). Hochberg's step-up procedure is more powerful than Holm's step-down procedure only if we assume independent (or non-negative dependence) among P-values (Sarkar, 1998). FWER are preferred when researchers require strong evidence to support their findings.

2.1.2 False discovery rate (FDR)

FDR, on the other hand, assumes a certain rate of type I error in multiple tests under null hypothesis. False discovery rate can be defined as

FDR is the expected value of Q

$$FDR = E(Q)\ where\ S \neq 0$$

$$FDR = 1\ when\ S = 0$$

Where

$$Q = \frac{n_1}{S}$$

All FDR procedures focus on ensuring that $E(Q) = \alpha$.

1. Benjamini-Hochberg procedure (Hochberg & Benjamini, 1995)

   For ordered independent m P-values, $p_{(i)}$, find the largest k that satisfies

   $$p_{(k)} \leq \frac{k}{m}\alpha$$

   And reject all $p_{(i)}$ that $i \leq k$

2. Benjamini-Yekutieli procedure (Benjamini & Yekutieli, 2001)

Benjamini-Yekutieli procedure is an adjusted BH procedure to deal with arbitrary stochastic dependence among P-values.

$$p_{(k)} \leq \frac{k}{m * c(m)} \alpha$$

Where

$$c(m) = \begin{cases} 1 & iid\ p_{(i)}\ or\ positive\ dependence \\ \sum_{1}^{m} \frac{1}{i} & positive\ regression\ dependence \end{cases}$$

FDR procedures control the expected type I error rate instead of controlling the probability of not making more than one type I error. Therefore, FDR procedures are less conservative compared to FWER with higher power but less specificity. FDR is chosen when researchers wish to identify candidate genes or genome regions from a large pool of genes such as the high throughput genomic data.

However, neither procedures adequately control type I error rate when the number of tests performed is too large. Bonferroni-type procedures assume independence among all tests, which is likely to be violated when genomic features are spatially close on chromosomes. In 2005, the international HapMap Consortium estimated that there are 150 per 500 kilobase pairs common independent variants in European population. Therefore, they suggested $5X10^{-8}$ as genome wide significant level for SNP array (International HapMap Consortium, 2005). In this thesis we use $1X10^{-7}$, which is a little more liberal than the suggested level, as a P-value threshold for SNP array since that is a FWER style method to control type I error and we want to be a little less stringent. Moreover, FDR procedure performance is only acceptable when total number of tests is modest, or other criteria are used to select candidate targets.

In addition to multiple testing, another concern in the use of GWAS is with its effectiveness for identifying meaningful genomic allele to guide wet-lab research. GWAS only provides significance level and effect size for base pair level features such as SNP, methylation in one CpG island etc, which, in most cases, are not the primary interests for wet-lab researchers. Moreover, when patient DNA are interrogated by multiple genomic platforms and association

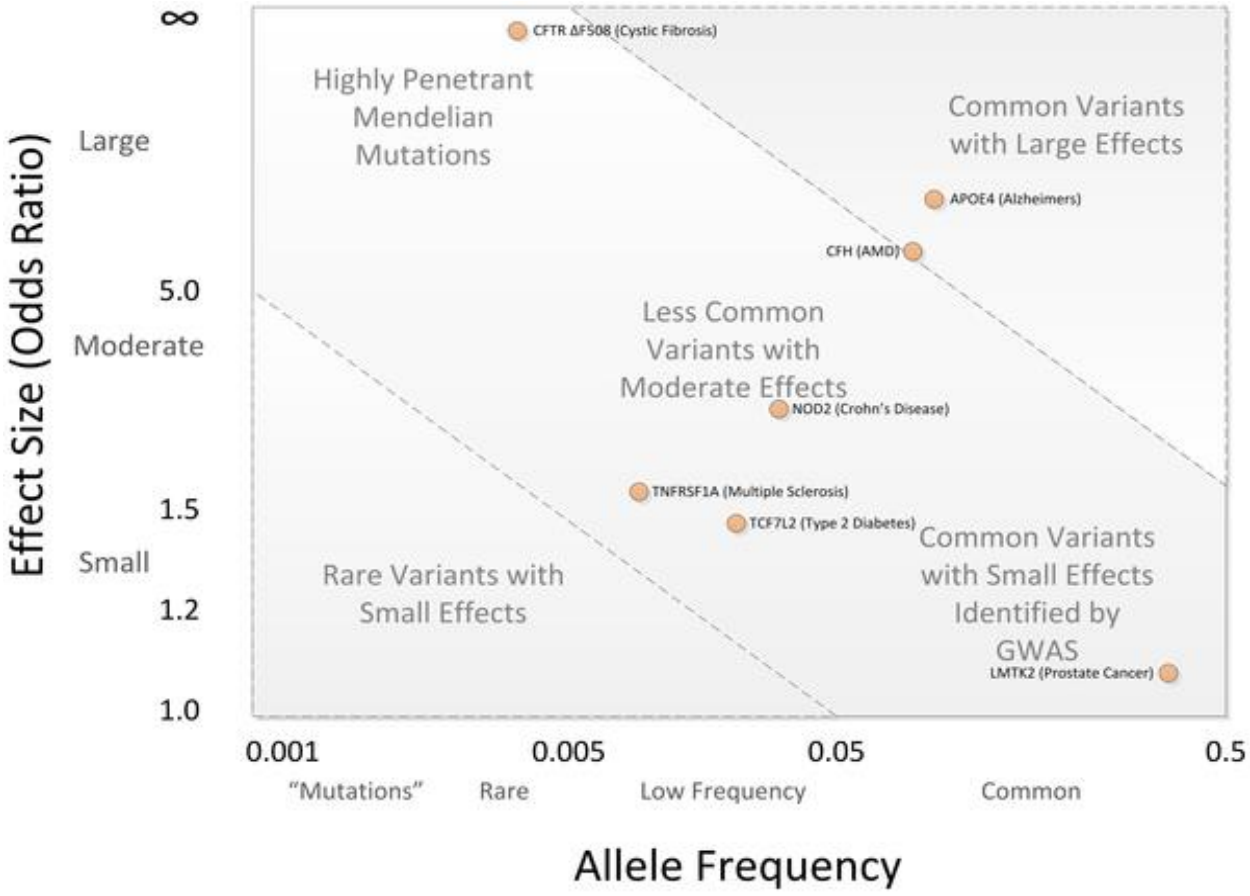

Figure 3. Allele frequency and effect size (Bush WS, 2012)

.

tests are performed for multiple outcomes, it is not well established on how to summarize the results for making meaningful conclusion.

Finally, GWAS results are not reliable for identifying rare genomic variant. This problem is

inherited from the inadequacy of statistical model (mostly linear model) to significantly select rare genome variant covariates, when sample sizes are too small. For example, if a genotype is only observed in 4/500 patients and 4/4 patients with that genotype are deceased after the treatment, the Cox regression model for overall survival could yield a large hazard ratio or very small p-value, which would not make much sense practically (Firth, 1993). This problem is more severe when the effect size of the feature is relatively small (Fig 3). Yet in real study we usually select features with P-value smaller than a certain threshold or k features with smallest P-value. In general, threshold or ranking selection will inflate association estimates (Faye, et al., 2011). This is also known as GWAS selection bias. Re-sampling techniques such as bootstrap is usually implemented to eliminate those bias (Faye, et al., 2011)

2.2 Set-based association analysis

Intuitively a linear model with all genomic features versus multiple phenotypes would be appropriate to estimate the significance and effect size of designated genome allele. Of course, sample size would be a concern since the number of variables in the model may exceed the number of samples. Grouping features may reduce the number of tests, therefore alleviate multiple testing problem. Features are usually grouped by their genome location, such as gene, or by their function, such as pathway. There are several set-based/region-based association analysis published before, such as Burden test, variance-component test and combination of those two. Those tests can be categorized into 2 classes based on their approach to integrate genomic features. One class is collapsing the features into a genome score for subsequent test such as Burden test. The other class is testing the variance of genome features (Table 3). All those

methods assume a complete genomic dataset, which requires no missing data among all subjects in study cohort for every genome feature. Practically that requirement is challenging, especially for multiple platform genomic data. Different subgroup of subjects may be interrogated by different arrays at different time. Therefore, some subjects may miss some genomic data which makes it impossible to perform set-based association test.

Table 3. Summary of statistic test of rare variant association test (Lee S, 2014)

	Description	Advantage	Disadvantage
Burden tests	collapse rare variants into genetic scores	powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants
Adaptive burden tests	use data-adaptive weights or thresholds	more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	often computationally intensive; VT requires the same assumptions as burden tests
Variance-component tests	test variance of genetic effects	powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	less powerful than burden tests when most variants are causal and effects are in the same direction
Combined tests	combine burden and variance-component tests	more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive
EC test	exponentially combines score statistics	powerful when a very small proportion of variants are causal	computationally intensive; is less powerful when a moderate or large proportion of variants are causal

2.3 Meta-analysis

For a dataset with missing data due to miscellaneous reasons, association test by single features would be more appropriate and practical. But the question remains that how to integrate single feature result at gene level to determine its association with outcome. One simple solution is to report that the gene is significantly associated with outcome when at least one feature within that gene is significantly (after multiple test adjustment) associated with the outcome (single feature method). However, this method does not consider the combined effect of multiple marginally significant features from multiple platforms. For example, if there are 100 genomic features interrogated in a gene, and 10 of them are marginally significant, it would be quite evident, in real study, that this gene may be significantly associated with outcome, but the evidence collected is not enough to yield a statistically significant result. Yet the current ranking or threshold selection method to pick the significant features will likely miss those features. In a genome wide study, when millions of features are tested, it is more likely to encounter similar situation such that many features with small effect on outcome are present. It would be beneficial if we can combine all those small effect within a gene so that we can strengthen evidence of the association between that gene and outcome.

One approach to combine all evidence is to integrate association test P-values from individual features into a single unified P-value. It is natural and more convenient. In the meantime, it is going to reduce the number of hypothesis tested and make it easier to control type I error. Furthermore, it is known that combining a series non-significant or marginally-significant P-

value may result in significance (Rosenthal, 1978).

2.3.1 Combining P-values

Let's consider m independent hypothesis tests with p-values p_i (i=1, 2, …m).

Under Null hypothesis, the joint distribution of p_i is

$$H_0: p_i \sim U(0,1)\ (i = 1, 2, \dots, m)$$

Several statistics are commonly used in combining P-values.

1. Fisher's method ((Fisher, 1932) (FCT)

$$-2\sum_{i=1}^{m} \ln(p_i) \sim X_{2m}^2 \qquad (1)$$

2. Pearson's method (Pearson, 1933)

$$-2\sum_{i=1}^{m} \ln(1 - p_i) \sim X_{2m}^2 \qquad (2)$$

3. George's method (Mudholkar & George, 1979)

$$-2\sum_{i=1}^{m} \ln \frac{p_i}{(1-p_i)} \qquad (3)$$

4. Edgington's method (Edgington, 1972)

$$\sum_{i=1}^{m} p_i \qquad (4)$$

5. Stouffer's method (Stouffer, et al., 1949)

$$\sum_{i=1}^{m} \Phi^{-1}(p_i) \sim N(0, m) \qquad (5)$$

6. Tippett's method (Tippett, 1931)

$$\min(p_1, p_2, \dots, p_m) \sim Be(1, m) \qquad (6)$$

7. Lancaster's method (Lancaster, 1961)

$$\sum_{1}^{m} \gamma_{\left(\frac{w_i}{2},2\right)}^{-1}(1-p_i) \sim \chi_d^2$$

Where $d = \sum w_i$ under H_0. $\gamma_{\left(\frac{w_i}{2},2\right)}^{-1}$ is inverse CDF of gamma distribution

Fisher's, Lancaster's and Tippett's method are sensitive to small P-values. While Pearson's method is sensitive to large P-values. The Logit and Stouffer's methods are robust to outliers and regards as balanced when combining tests of significance from multiple studies (Heard, 2018). The other 3 methods, especially Stouffer'e method is balanced (Heard, 2018). All the procedures listed above are monotone as functions of P-values and hence optimal in various settings. However, the Fisher, Lancaster and Logit procedures have asymptotic optimality property in Bahadur efficiency (Mudholkar & George, 1983) (Hedges & Olkin, 1985). In GWAS, gene enrichment analysis and many association studies, where individual tests may be too subtle to detect signals. Fisher's combination method (FCT) is perhaps the best known and widely used. FCT assumes that all P-values are independent and come from continuous underlying test statistics. Under the null hypothesis, indicidual P-value will follow uniform(0,1) distribution and FCT has a Chi-square distribution (Fisher, 1932). It has been widely used on many association studies. Thus for our application in this thesis, we choose Fisher's combining method (FCT) to integrate P-values. In this project when number of combined tests, m, is too large and most of the P-values are not significant, FCT will lose power (Huber, 1977). In this thesis, we propose a solution to this problem that uses the threshold truncated P-values or rank truncated P-values.

2.3.2 Threshold truncated P-values

Zaykin et al (2002). first proposed to combining P-values truncated by a common threshold δ

such as such as $\delta = 0.05$ (TPM). He proposed the use of product of truncated P-values as a statistic.

$$W = \prod_{i=1}^{m} p_i^{I(p_i<\delta)}$$

and derived its distribution for the case where the P-values are independent. For correlated P-values, he used Choleskey decomposition to transform correlated statistics to independent one (Zaykin, 2002) and showed that TPM (with $\delta = 0.05$) has higher power than other methods of combining P-values. He also argued that by choosing appropriate cut-off point such as 0.05, TPM increases power. However he did not provide any evidence in his simulation study. One problem with TPM is that a knowledge of the variance-covariance matrix is needed to obtain Choleskey factor. In practical application the variance-covariance matrix usually is unknown and estimated from data.

2.3.3 Rank truncated P-values

An aleternative approach by Dudbridge et al (2003) is to combine k smallest P-values as test statistic (RTP) (Dudbridge & Koeleman, 2003). They first sorted the P-values from smallest to largest, then the product of first k P-values is used as statistic:

$$W_R = \prod_{i=1}^{k} p_{(i)}$$

where k<m. The power of RPT is sensitive to the choice of k. The choice of k depends heavily on the number of true loci, but not the total number loci selected. This makes sense since only

the first k P-values are used to calculate RPT. However, they showed in simulation study that when a cut-off in TPM (eg. $\delta = 0.05$) is fixed, the power of TPM decreases when ratio of true loci and all loci decreases. Interestingly the author didn't show the power of RPT when the number of true loci is much larger/smaller than k. But it would be expected that when the number of true loci is far much bigger than k, the power of RPT would decrease.

For correlated cases, RTP utilizes information from linkage disequilibrium. Dudbridge proposed a permutation process to adjust the truncation point k and also proposed a blocking analysis by grouping features within LD block to reduce correlation. Yu et al suggested an adaptive rank truncated product method (ARTP) to overcome the power problem in RTP (Yu, et al., 2009) and used permutations to find the optimal k in RPT that maximized the power.

2.3.4 Remarks

All those methods are developed based on the assumption that P-values are independent. For correlated cases they either assumed a variance-covariance matrix or grouping features within LD blocks to reduce correlation. However, neither procedures seem appropriate in handling multi-platform genomic data. The correlation structure between features especially features from different platforms is not clear. The concept of LD blocks can't be applied to multi-platform data. Thus, re-sampling methods such as permutation/bootstrap may be our last resort. Furthermore, neither their simulation study compares the power of their method to traditional GWAS.

Here we propose a truncated and aggregated P-values test (TAP). It combines negative logarithm transformed truncated P-values from genomic features within a gene region as test statistics. Then a limited number of permutations (200 times) are performed to obtain empirical density function. A smooth kernel is used to smooth the density function so that a small TAP P-values can be obtained. The Integrated non-truncated independent P-values follows Gamma(i,1), where i is the number of P-values. It is easy to derive that combined truncated independent p-values follow truncated Gamma. Based on this, we constructed null distribution of TAP by hybrid permutation test.

3. Truncated and aggregated P-value test (TAP)

In this report we propose a comprehensive procedure for genome wide association study that contains following elements.

1. Framework and algorithm to construct statistics for aggregated test results over gene and derive a unified P-value on gene level to indicate the association between gene and relevant traits.
2. Algorithm to efficiently annotate genome features to gene regions.
3. Pipeline based on high performance computing facility to perform bootstraps/permutations on genome wide association test among genomic data and clinical outcomes.

We hope the statistical method, algorithm and framework established in this work can provide a powerful and convenient tool to provide a comprehensive reliable test result of the gene and outcome association.

3.1 TAP statistics

Suppose there are n genes and m_i features observed on gene i (i=1, 2, …., n). First, the features are annotated to the genes on the same chromosome by their genome location. Then each feature is used to perform an association test with certain outcome. For the ith feature, the P-values of the associated tests are then truncated and combined to form the aggregated P-value test statistics (TAP) defined as.

$$T_i = \sum_{k=1}^{m_i} -\ln(p_k) I(p_k \leq \delta) \qquad (2)$$

where δ is the cutoff point of the P-values.

3.1.1 Hybrid permutation test

The design of the permutation scheme is problem-specific. For each gene, the hybrid permutation test estimates the null CDF of the TAP statistic T based on the m (approximately) independent identically distributed (i.i.d) observations of truncated P-values, then computes a P-value from the null cumulative distribution function (CDF) estimator. To obtain a series (approximately) independently identically distributed observations of T, we will shuffle clinical data and genomic data to form new pairs of clinical and genomic data. After B times of permutation, we get a set of observations of T $\{T'_0, \ldots, T'_B\}$ where T_0 represents the T from

original data and T_j is obtained from the jth permutation. The starting point is a template CDF $F_0(\cdot)$ with support $[0, \infty)$. First a heuristic from the probability integral transformation.

Let $W = F_0(T)$ and $w = F_0(t)$, then the CDF of W on [0, 1] is given by

$$F_W(w) = \Pr(F_0(T) \leq w) = \Pr\left(T \leq F_0^{-1}(w)\right) = F_T\left(F_0^{-1}(w)\right) = F_T(t)$$

where $F_T(\cdot)$ represents the CDF of T. Hence

$$F_T(t) = F_W\left(F_0(t)\right)$$

The idea then is to estimate $F_W(\cdot)$ from the permutation data

$$w_j := F_0\left(t_j\right) \quad j = 1 \ldots B$$

and plug into the above equation to obtain an estimator of $F_T(\cdot)$ under the null hypothesis. The estimator of $F_W(\cdot)$ can be constructed by general kernel smoothing (Cheng & Parzen, 1997) of the empirical CDF of W from permutations, defined as

$$\widetilde{F_W}(w) := B^{-1} \sum_{j=1}^{B} I\left(W_j \leq w\right):$$

$$\widehat{F_W}(w) = \int_0^1 \widetilde{F_W}(u) d_u K(w, u)$$

A particular choice of the smoothing kernel is the variation diminishing spline built from the B-spline basis (Devore, 1972). It is important to capture the tail behavior, so the internal knots sequence can be constructed by sorting the elements in the set

$$\{0(0.1)0.1\} \cup \{0.8(0.05)1\} \cup \{\text{deciles of } W_i\ (i = 1, \ldots, B\}).$$

The estimator of $F_T(\cdot)$ under the null hypothesis is then

$$\widehat{F_T}(\cdot) := \widehat{F_W}\left(F_0(\cdot)\right)$$

and the P value

$$P = 1 - \widehat{F_T}(T_{obs}).$$

3.1.2 Construction of the template CDF

We assume that each P value follows the uniform (0,1) distribution under the null hypothesis. Heuristically if the P values are also independent, under the null hypothesis, then

$$\Pr(T = 0) = \Pr\left(\sum_{l=1}^{M} I(P_l \leq \delta) = 0\right) = (1-\delta)^M$$

and for t>0

$$\Pr(T \leq t) = \sum_{j=1}^{M} Pr(T \leq t | \sum_{l=1}^{M} I(P_l \leq \delta) = j) \Pr\left(\sum_{l=1}^{M} I(P_l \leq \delta) = j\right)$$

$$= \sum_{j=1}^{M} b(j; M, \delta)\, G_0(t; j, \delta)$$

where $b(j; M, \delta)$ is the $Binomial(j; M, \delta)$ probability mass function (pmf) and $G_0(t; j, \delta)$ is a given CDF on [0,1). Under the null hypothesis, the P-values are U(0,1). Hence, since P-values ae independent, $\sum_{K=1}^{j} -\log(p_K)$ has Gamma(j,1) distribution. Adjusting for the truncation, $G_0(\cdot\,; j, \delta)$ is defined as

$$G_0(t\,; j, \delta) = \frac{F_G(t; j, 1) - F_G(-jlog\delta; j, 1)}{1 - F_G(-jlog\delta; j, 1)} I(t \geq jlog\delta)$$

where $F_G(\cdot\,; j, 1)$ represents the CDF of Gamma(j, 1).

3.2 Simulation study

A set of genome and clinical data were simulated to test reliability and accuracy of the model. We assume that genomic data and outcomes follow central dogma as illustrated in Fig 4. At

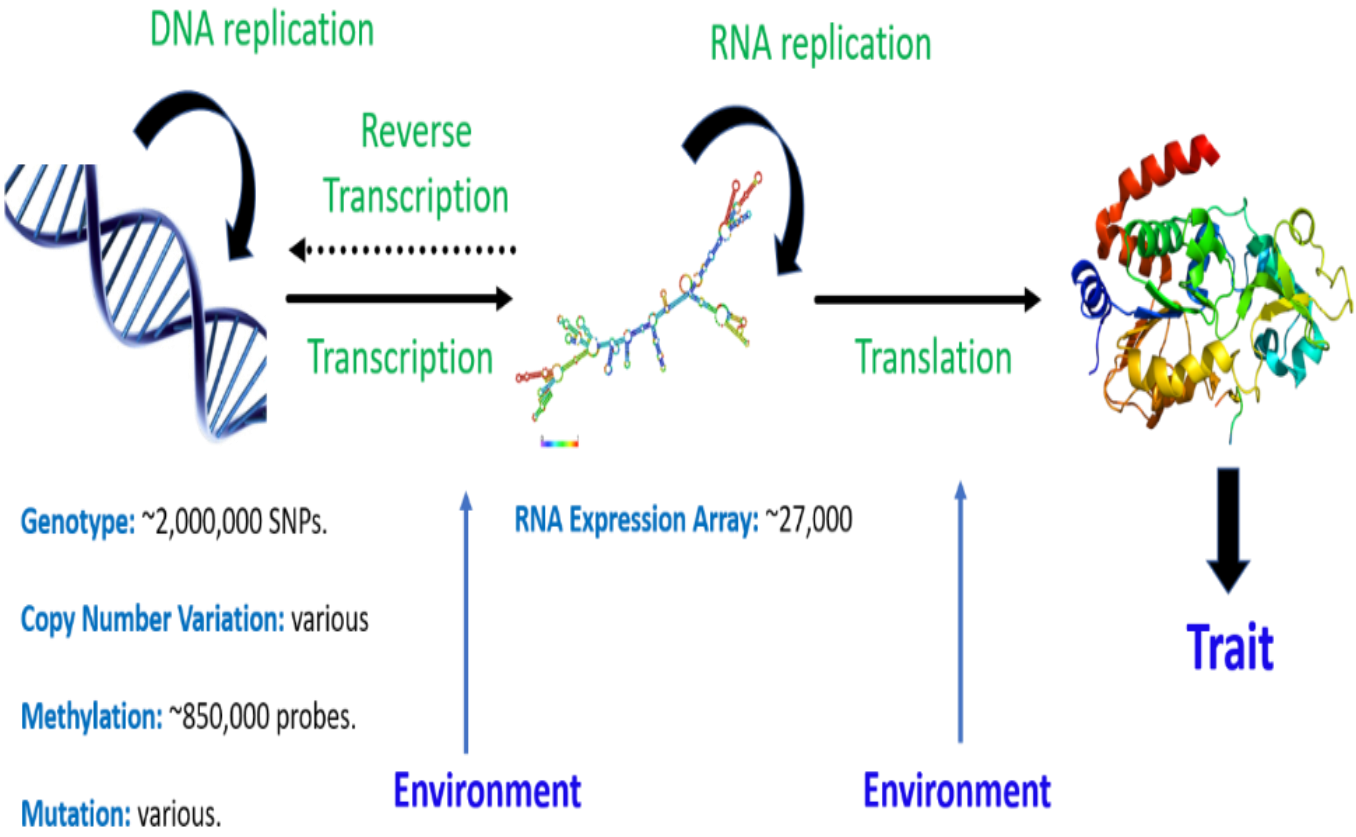

Figure 4. Central dogma and genomic data

DNA level, SNP array, methylation array and next generation sequencing data such as RNA-seq, whole exon-seq and whole genome-seq are used to interrogate status of designated base pairs. We can observe genotypes, methylation level and mutation of certain base pair, and whole genome copy number variation.

3.2.1 Simulate genomic data

We considered NLRP3 gene region with 47 SNP probes, 2 methylation probes, 1 copy number segment and 1 expression probe, among which 10 genotype probes, 2 methylation probes, 1 copy number segment and 1 expression probe are effective features and involved in our hypothetic model. To maintain linkage blocks within SNPs, instead of simulating genotype for individual SNP, we downloaded haplotype of 47 SNPs (included in Affymetrix SNP6 array) within NLRP3

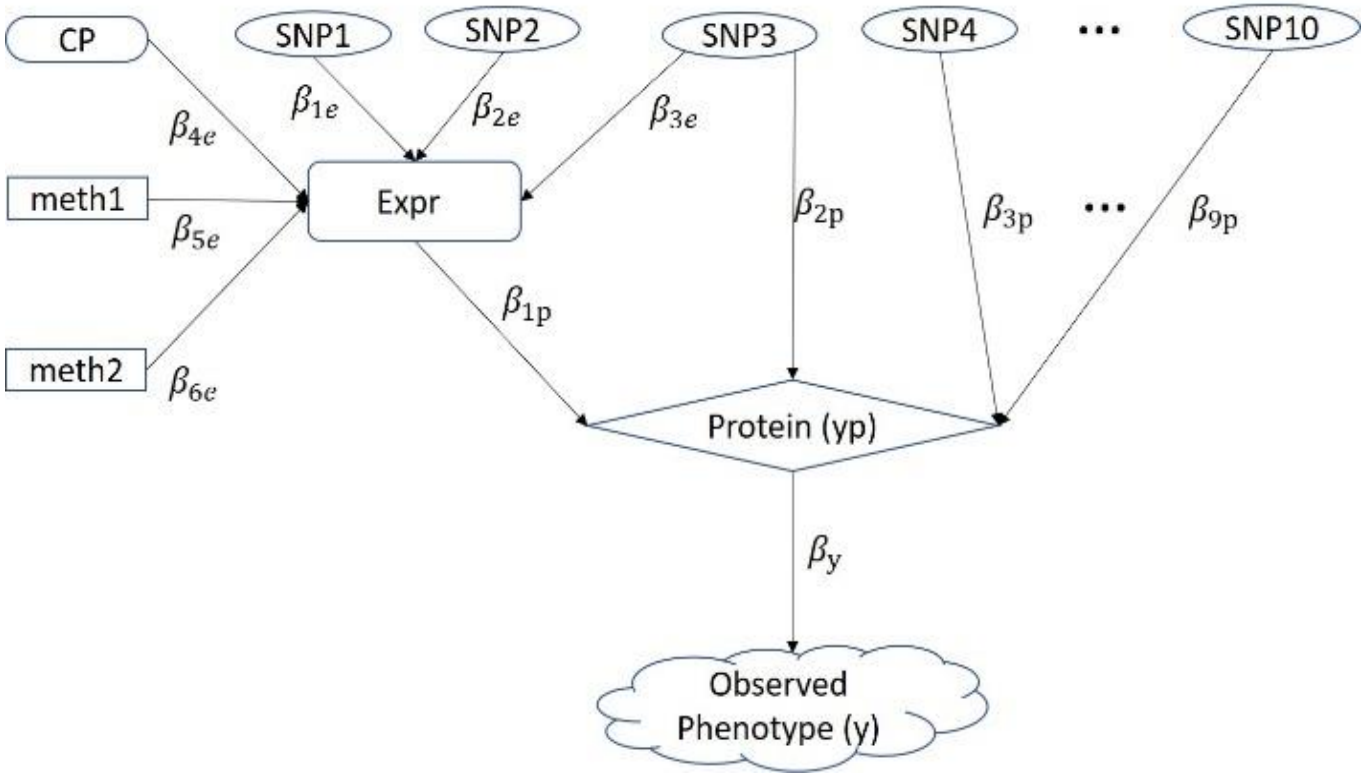

Figure 5. Model of genomic features, protein and observed outcome.

gene from 2004 individuals provided by 1000 genome project (https://www.internationalgenome.org/). Then we sample 2 out of 5008 single strand DNA to form a double strand and determine genotype for each SNP. By sampling the single strand DNA,

the correlation structure among those 47 SNPs are maintained.

Based on Central dogma, we assume that gene expression (Expr) and genotype of certain SNPs (SNP3, SNP4, …, SNP10) may affect the protein concentration (yp). The protein level will affect observable phenotype (y). In the meantime, methylation (meth1, meth2), copy number status (CP) and SNPs (SNP1, SNP2 and SNP3) can affect gene expression (Fig 5). Quantitatively the genotype was assigned value as 0, 1, 2 corresponding to number of the selected reference allele. The copy number status is assigned as 1, 2 corresponding to gain or loss. The methylation status is set as 0, 1, 2 corresponding to low/medium/high. We assume a linear relationship between genomic features and logarithm of gene expression. Then other genomic features and gene expression were linearly correlated with protein function. Finally, protein function is linearly correlated with phenotype. Features that are involved in the model with a non-zero coefficients are called causal features. Expr, yp and y were calculated based on following formulas.

$$\log(expr) = \beta_{1e} * SNP1 + \beta_{2e} * SNP2 + \beta_{3e} * SNP3 + \beta_{4e} * CP + \beta_{5e} * \text{meth1} + \beta_{6e} * meth2 + e$$

$$yp = \beta_{1p} * expr + \beta_{2p} * SNP3 + \beta_{3p} * SNP4 + \beta_{4p} * SNP5 + \beta_{5p} * SNP6 + \beta_{6p} * \text{SNP7} + \beta_{7p} * SNP8 + \beta_{8p} * SNP9 + \beta_{9p} * SNP10 + e$$

$$y = \beta_y * y_p + e$$

where e is the error following standard normal distribution.

A simple linear model between single feature and phenotype (y), among which Expr is logarithm transformed, was fitted to test significance of the association between the feature and phenotype.

The P-values were truncated at 0.05 for TAP statistics calculation ($\delta = 0.05$).

3.2.2 Simulation study setup

300 observations with outcomes and genomic data are simulated based on the model described above. All β_s are fixed except for β_y, which is set from 0 to 1 with 10,000 simulations performed in steps of 0.04 at each β_y when $\beta_y > 0$ and 1,000,000 simulations are performed when $\beta_y = 0$ which is the null hypothesis. In each simulation, traditional method and TAP method were used to determine whether the gene is significantly associated with outcome. 200 permutations were performed for each simulation to obtain the empirical distribution of TAP. In commonly used procedures the gene is considered as significantly associated with an outcome when as least one genome feature within that gene is significantly associated with the outcome. The significant association by single feature is determined by following rules:

1. Features from SNP array are considered as significant when their P-values are less than $1X10^{-7}$. This is more liberal than the commonly the recommended genome wide significant level ($1X10^{-8}$) for SNPs so that the threshold for SNP is not too stringent comparing to other features.
2. In real studies FDR is commonly used to adjust multiple comparison for expression array, methylation array and other arrays when the number of probes is moderate. In this simulation study we follow this tradition. For features from other platforms such as gene expression array, methylation array and copy number variation, we use FDR to determine the significant threshold. More specifically for each calculated feature P-value from every simulation, we add 99 U(0,1) random values to ensure a 1% true positive rate, or 19

U(0,1) random values, to ensure 5% true positive rate. Then an FDR procedure assuming 10% or 20% false discovery rate is applied to those value. If the calculated P-value is significant under the FDR procedure, the corresponding feature is considered significant in the simulation.

A gene is considered detected if at least one feature within a gene significantly associated with outcome. Since we do not have a close-form expression for $\Pr(T \leq t|\delta)$ even assuming independent P-values, we can't derive an analytic expression of relationship between the power of TAP and δ. Instead we illustrate the relationship by simulation with different δ to show the change of power of TAP. The δ value used are 0.01, 0.05 and from 0.1 to 1 in 0.1 increment.

3.2.3 Simulation result

To assess the sensitivity and specificity of TAP, we compare the power to detect a gene (as defined above) by TAP and single feature. When the gene is not associated with outcome, we expect the TAP P-value to be distributed as U(0,1). Thus, the empirical quantile of TAP P-value under null hypothesis is compared to the quantile of U(0,1).

3.2.3.1 TAP P-value under null hypothesis ($\beta_y = 0$)

First, we check how the TAP P-value behavior when there is no association between gene and outcome. Under null hypothesis ($\beta_y = 0$), we expect that TAP P-value will follow U(0,1) distribution. In Figure 6, Figure 7 and Figure 8 the empirical distribution function of TAP and U(0,1) are compared. The three figures show different ranges of uniform quantile: (0,1), (0,0.1) and (0,0.01) versus TAP P-values under null hypothesis. Under this assumption, we expected TAP P-value to follow U(0,1) distribution. Overall, this expectation is roughly upheld until TAP

P-value exceeds 0.5. We disregard those large TAP P-values since they don't correspond to a region of significance level. Thus, we enlarge the curve between (0,0.1), (0,0.01) in the following two panels to focus on the relevant region of P-values. As illustrated in plots, when $p<0.005$, TAP P-value is more conservative than U(0,1), that is, p is greater than uniform quantile. On the other hand, when quantile is within (0.005,0.04) and (0,0.0005), TAP P-value is more liberal than U(0,1), that is, p is less than uniform quantile. Table 4 gives a summary of expected and actual type I error in TAP under null hypothesis.

Table 4. TAP expected vs actual type I error under null hypothesis

Nominal Type I Error Probability	Estimated Type I Error Probability
0.0010	0.000670
0.0050	0.004685
0.0075	0.012765
0.0100	0.016480
0.0250	0.029659
0.0500	0.048911
0.0750	0.073258
0.1000	0.099135

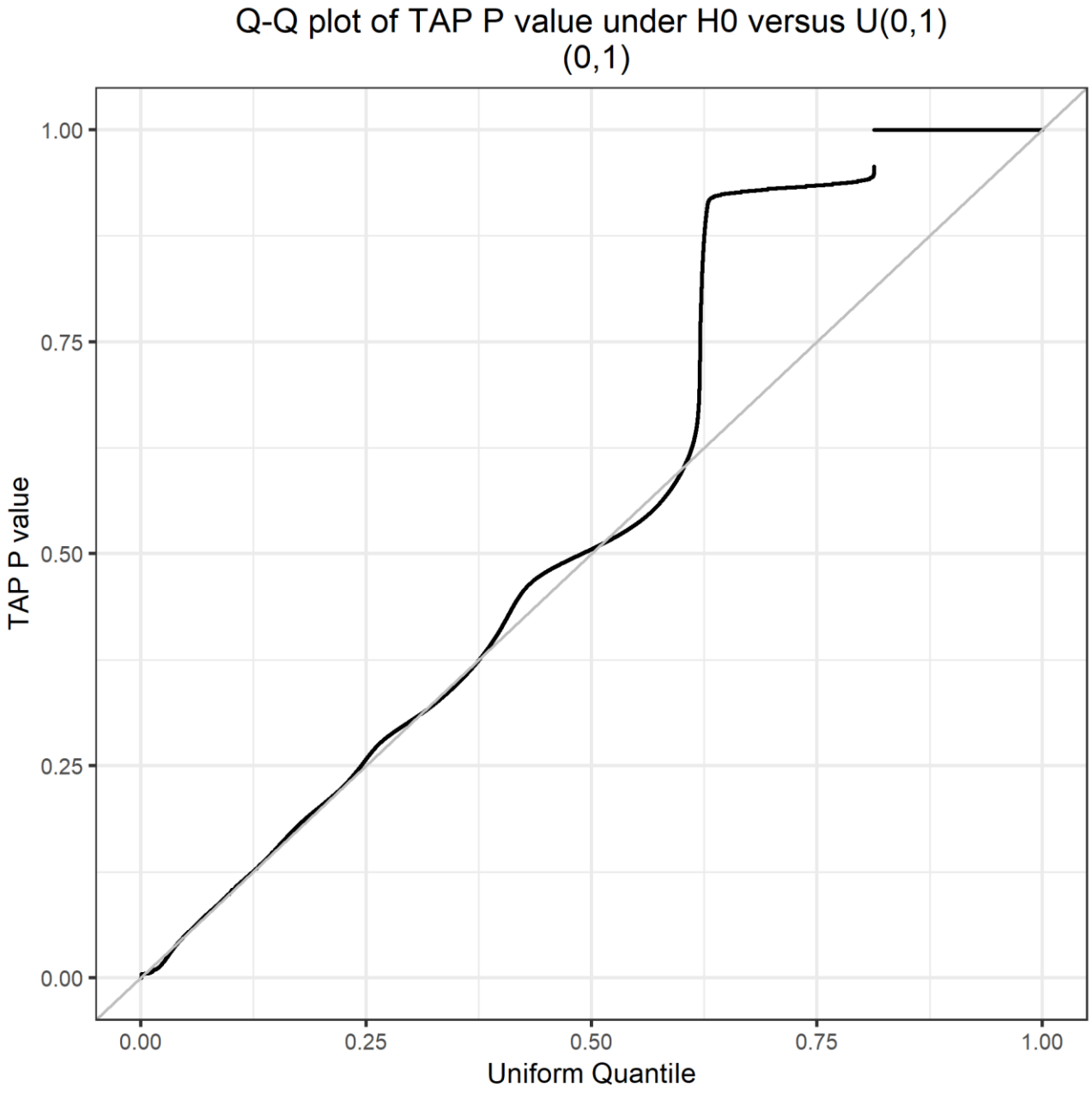

Figure 6. Q-Q plot of TAP P values and U(0,1) ($\delta = 0.05$), range (0,1)

Tests are performed under null hypothesis ($\beta_y = 0$) with 1,000,000 simulations. Figure shows uniform quantile range (0,1).

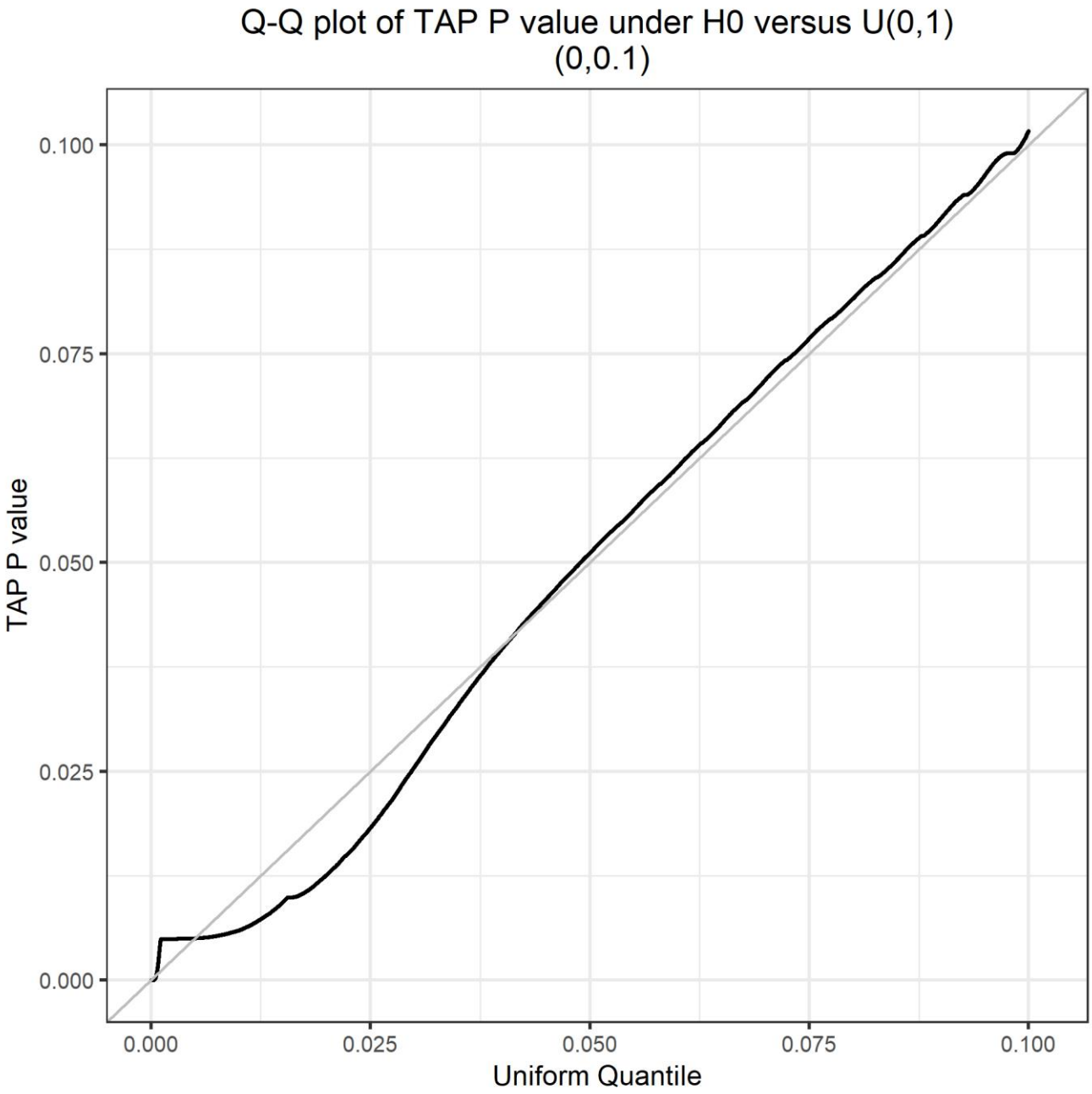

Figure 7. Q-Q plot of TAP P values and U(0,1) ($\delta = 0.05$). range (0,0.1)

Tests are performed under null hypothesis ($\beta_y = 0$) with 1,000,000 simulations. Figure shows uniform quantile range (0,0.1).

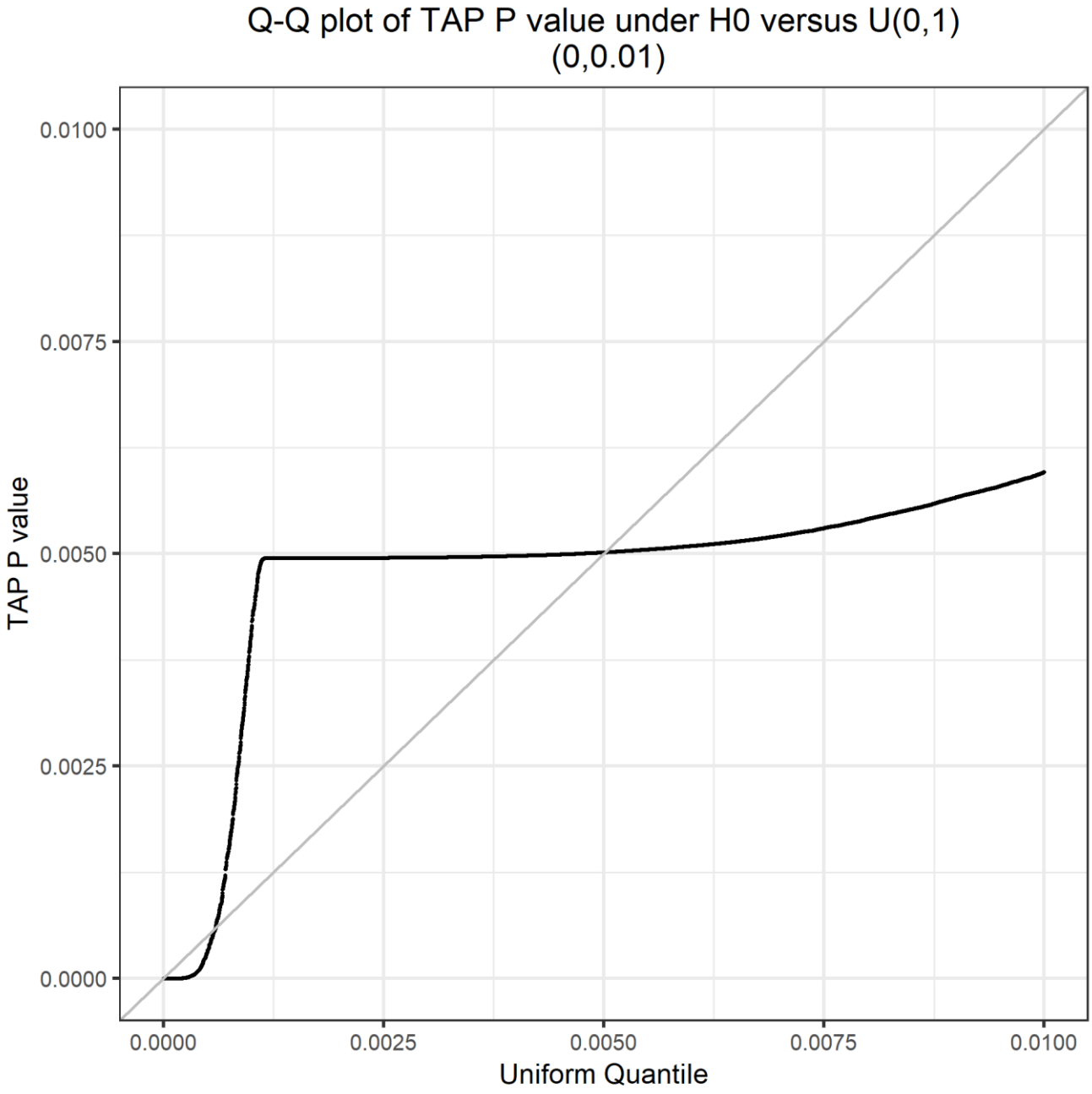

Figure 8. Q-Q plot of TAP P values and U(0,1) (δ=0.05). range (0,0.01)

Tests are performed under null hypothesis ($\beta_y = 0$) with 1,000,000 simulations. Figure shows uniform quantile range (0,0.1).

Next, we explore the effect of δ on empirical distribution function of TAP P-value when $\beta_y = 0$. Overall when $\delta \geq 0.1$ the empirical distribution of TAP P-value fits well with U(0,1) (Fig 9). When $\delta = 0.01$ and 0.05, as we show above, the TAP P-value is generally more conservative than U(0,1) except at small P-values. When p<0.03, all TAP P-value quantile lines have the same trend and relationship with U(0,1) as described above. All lines converged around 0.005 which is probably due to the fact that only 200 permutations were performed in each simulation (Fig 10). Without smoothing, the smallest empirical TAP P-value is 0.005. With smoothing we expect smaller TAP P-value. Interestingly we only see TAP P-value<0.005 when $\delta = 0.01, 0.05, 0.1, 1, 0.9, 0.8$ (Fig 11). The reason is not immediately clear. When $\delta < 0.2$ or $\delta > 0.8$ the empirical TAP P-value lines are further away from U(0,1) line than the empirical TAP P-value lines when $0.2 < \delta < 0.8$. Therefore, when δ is moving close to 1 or 0, the TAP P-value becomes more liberal/conservative than expected U(0,1). This suggests that we may want to choose a δ that is close 0.5 so that TAP P-value can achieve expected type I error under null hypothesis.

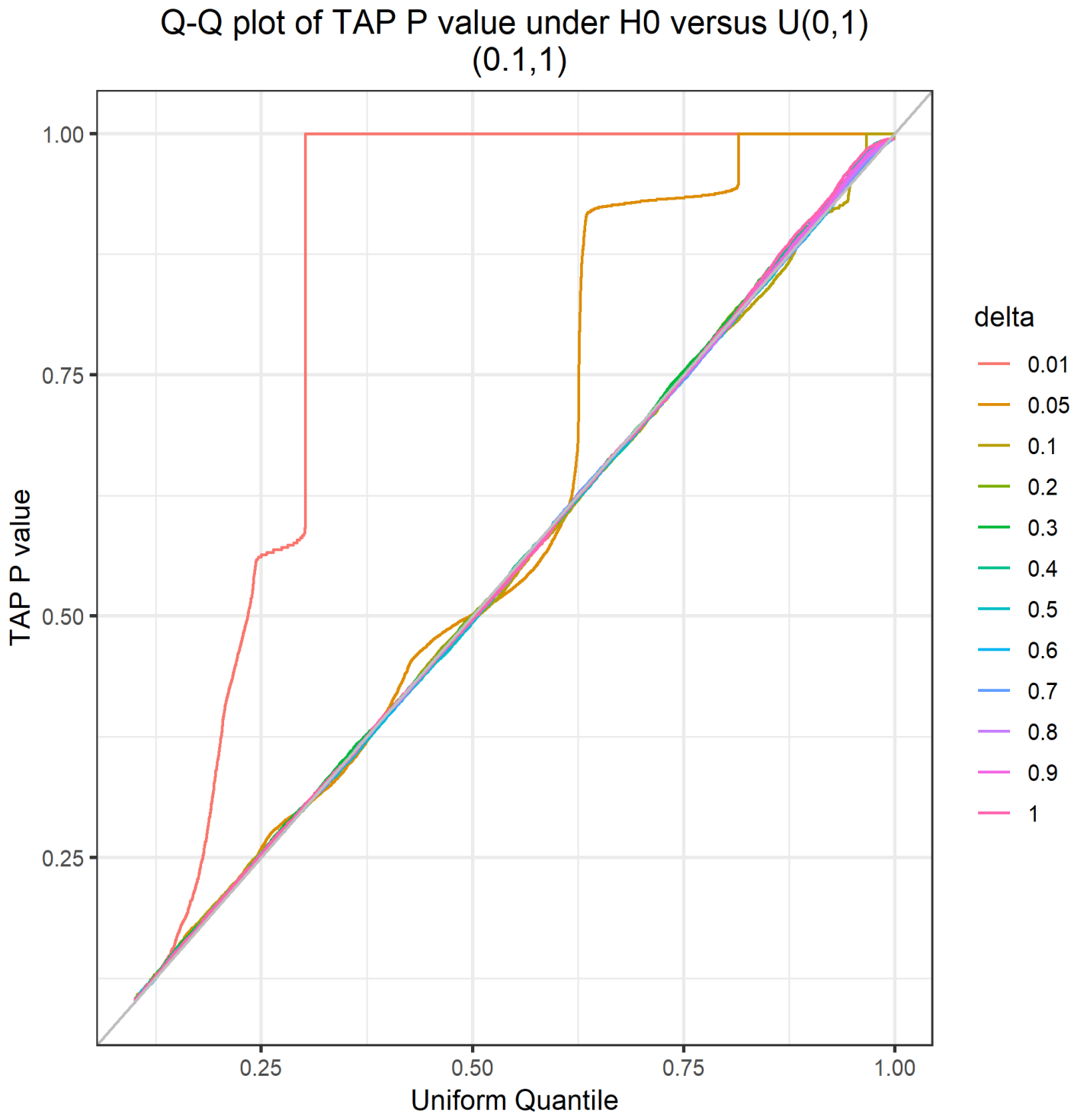

Figure 9. Q-Q plot of TAP P-value versus U(0,1) with various P-value cutoffs (0,1)

Tests are performed under null hypothesis ($\beta_y = 0$) with 10,000 simulations. Figure shows uniform quantile range (0,1).

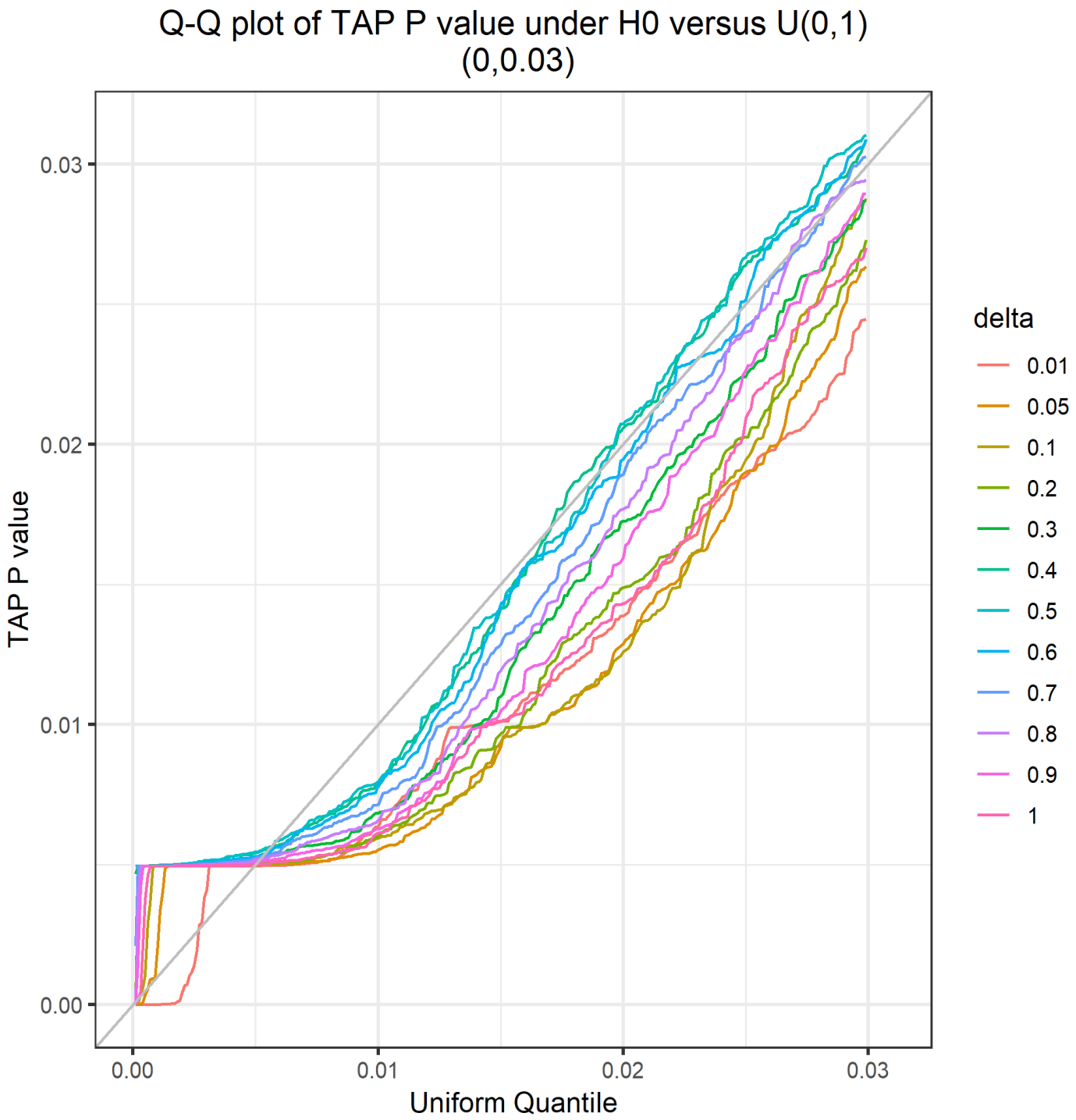

Figure 10. Q-Q plot of TAP P-value versus U(0,1) with various P-value cutoffs (0,0.03)

Tests are performed under null hypothesis ($\beta_y = 0$) with 10,000 simulations. Figure shows uniform quantile range (0,0.03).

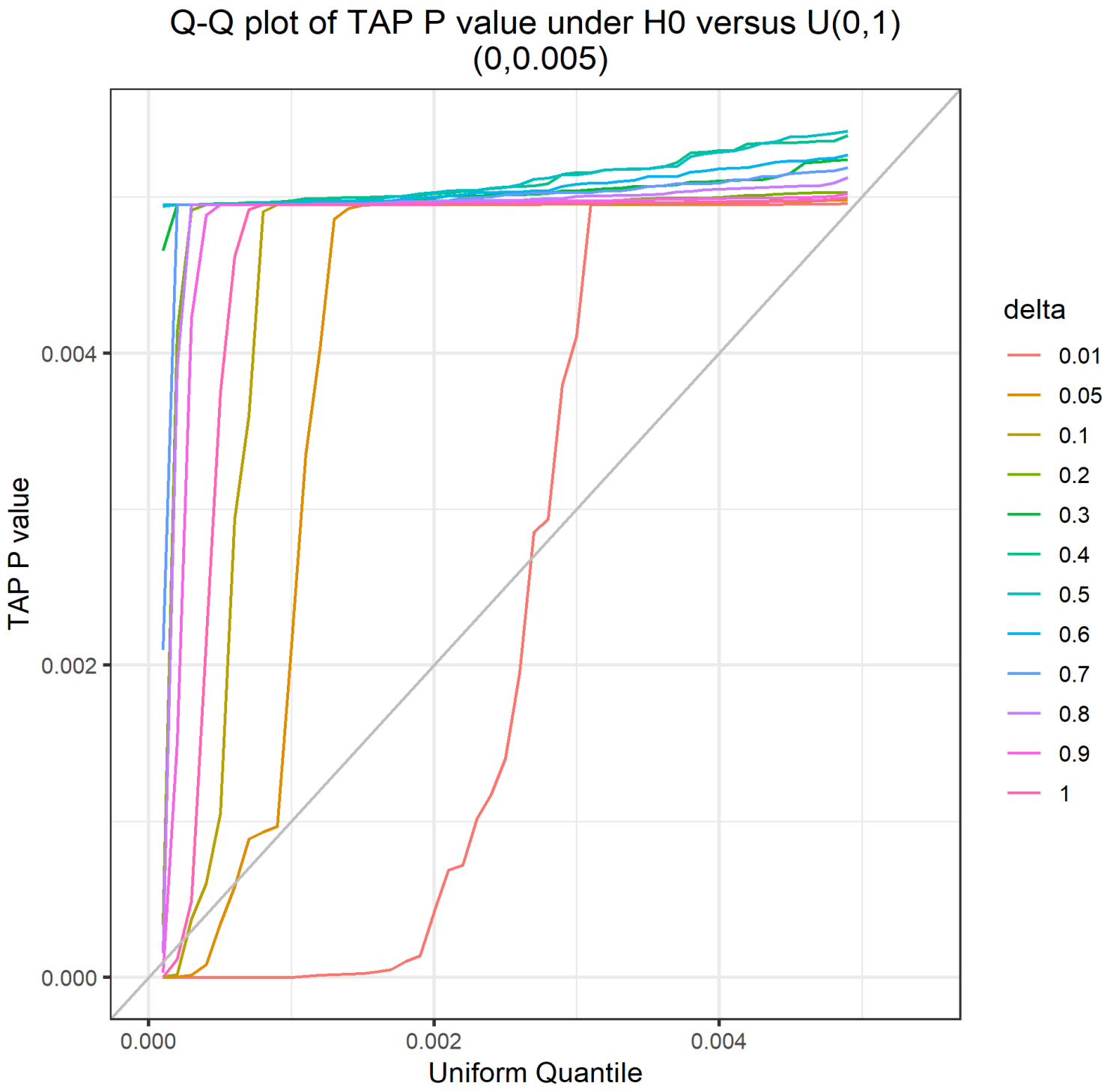

Figure 11. Q-Q plot of TAP P-value versus U(0,1) with various P-value cutoffs (0,0.005)

Tests are performed under null hypothesis ($\beta_y = 0$) with 10,000 simulations. Figure shows uniform quantile range (0,0.004).

3.2.3.2 Gene detection power of TAP and single feature

In this section we compare power to detect the gene by TAP and other features. First, we compare the gene detecting curve by TAP and in several platforms, including SNP array, gene expression array, methylation array and CNV. A gene is defined as detected in a platform if at least one feature in the platform is significantly associated with an outcome. As the real association between gene and outcome becomes stronger (β_y become larger), TAP gene detection probability is either comparable or superior to that of gene expression, which is the strongest single feature to detect gene in our simulated model. When true positive feature rate (pi_1) is 0.01, TAP crosses above the curve of expression probe and becomes superior in detecting gene than any other genome features when $\beta_y > 0.35$. Larger pi_1 and/or larger FDR will move this crossing point towards 0 along x axis, suggesting higher detecting probability by TAP than other genomic features overall for higher true positive feature rate or FDR rate (Fig 12, Fig 13, Fig 14).

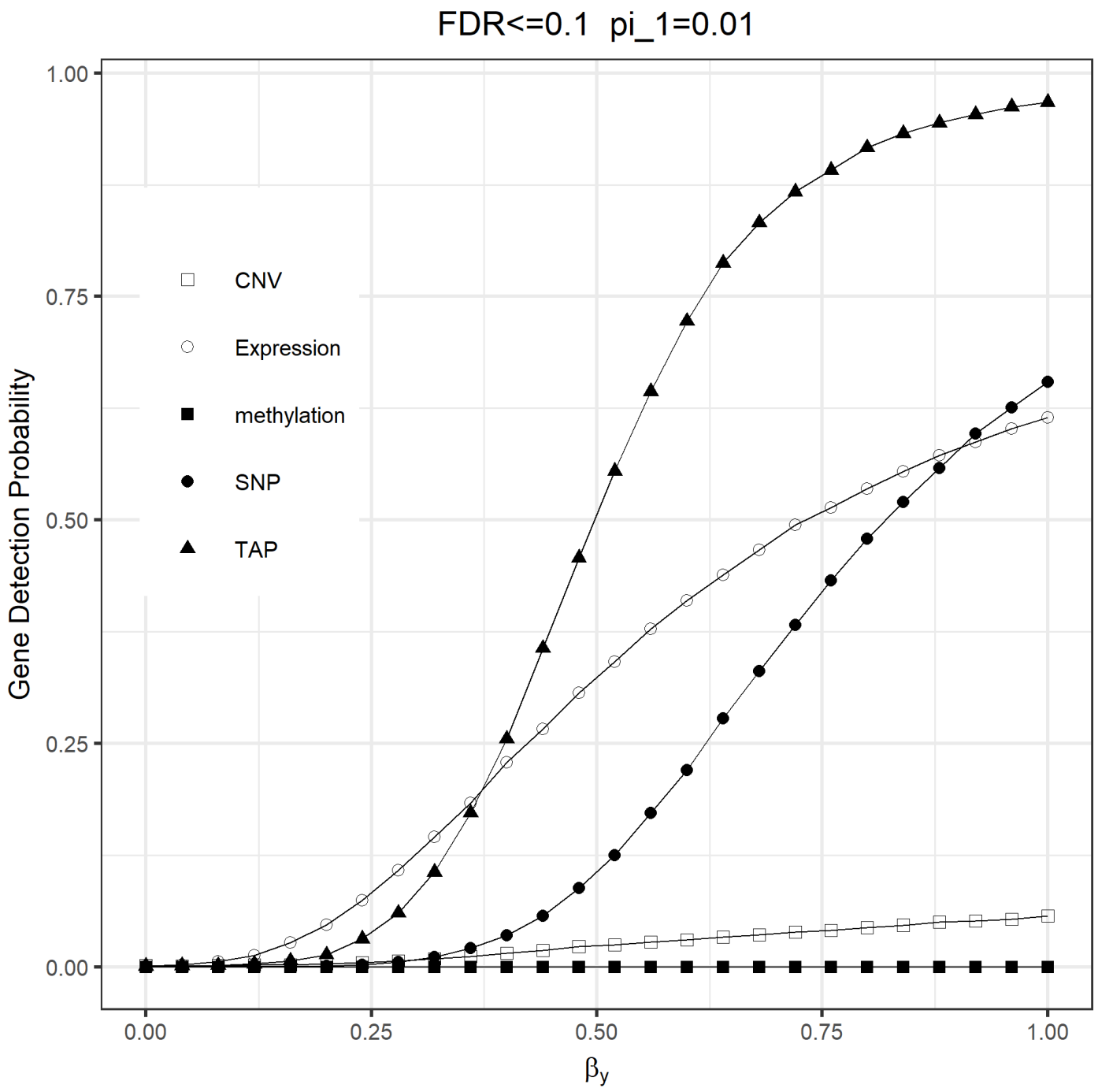

Figure 12. TAP detecting power versus other features. (FDR<=0.1, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 10% (FDR<=0.1).

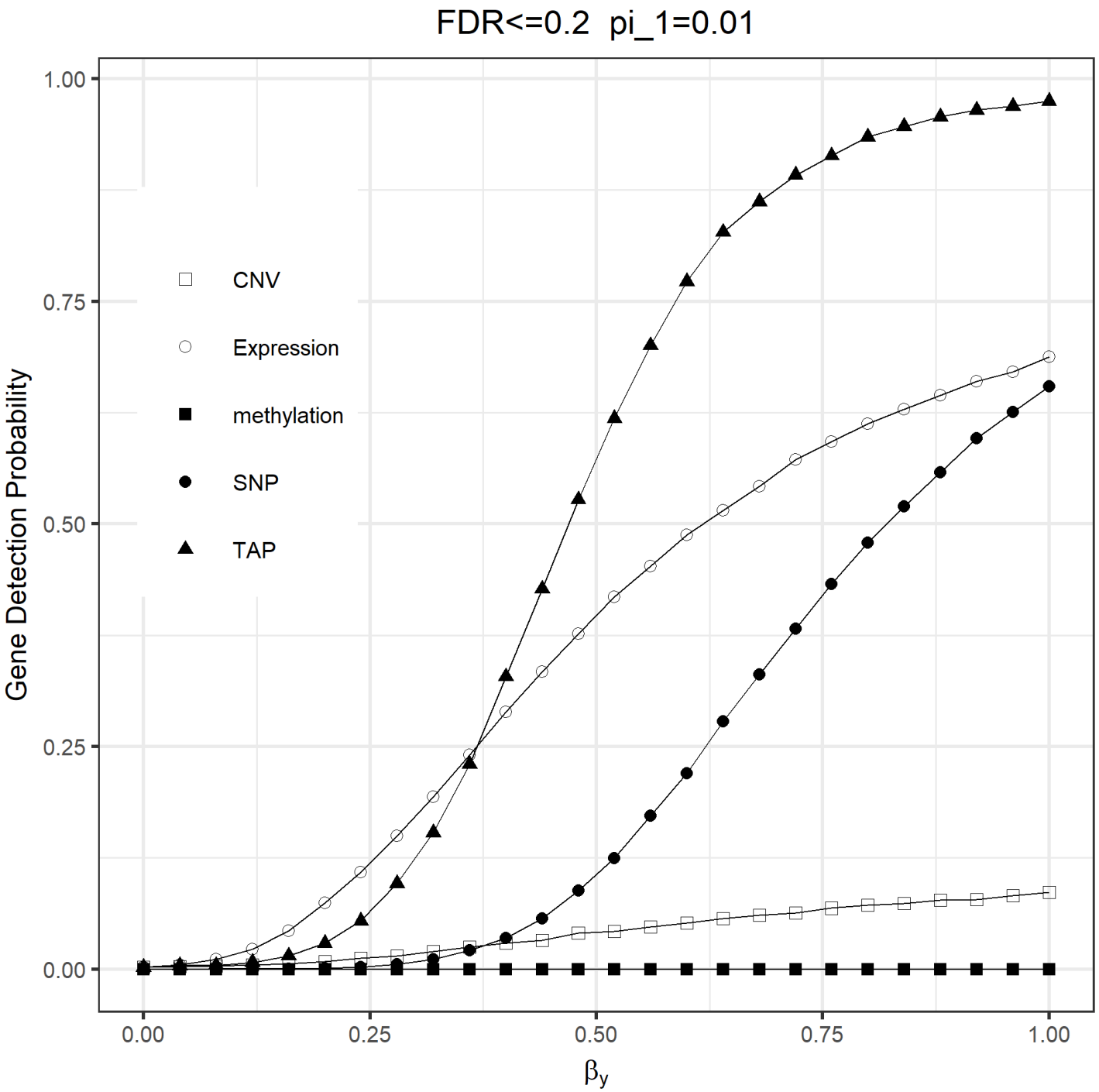

Figure 13. TAP detecting power versus other features. (FDR<=0.2, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 20% (FDR<=0.2).

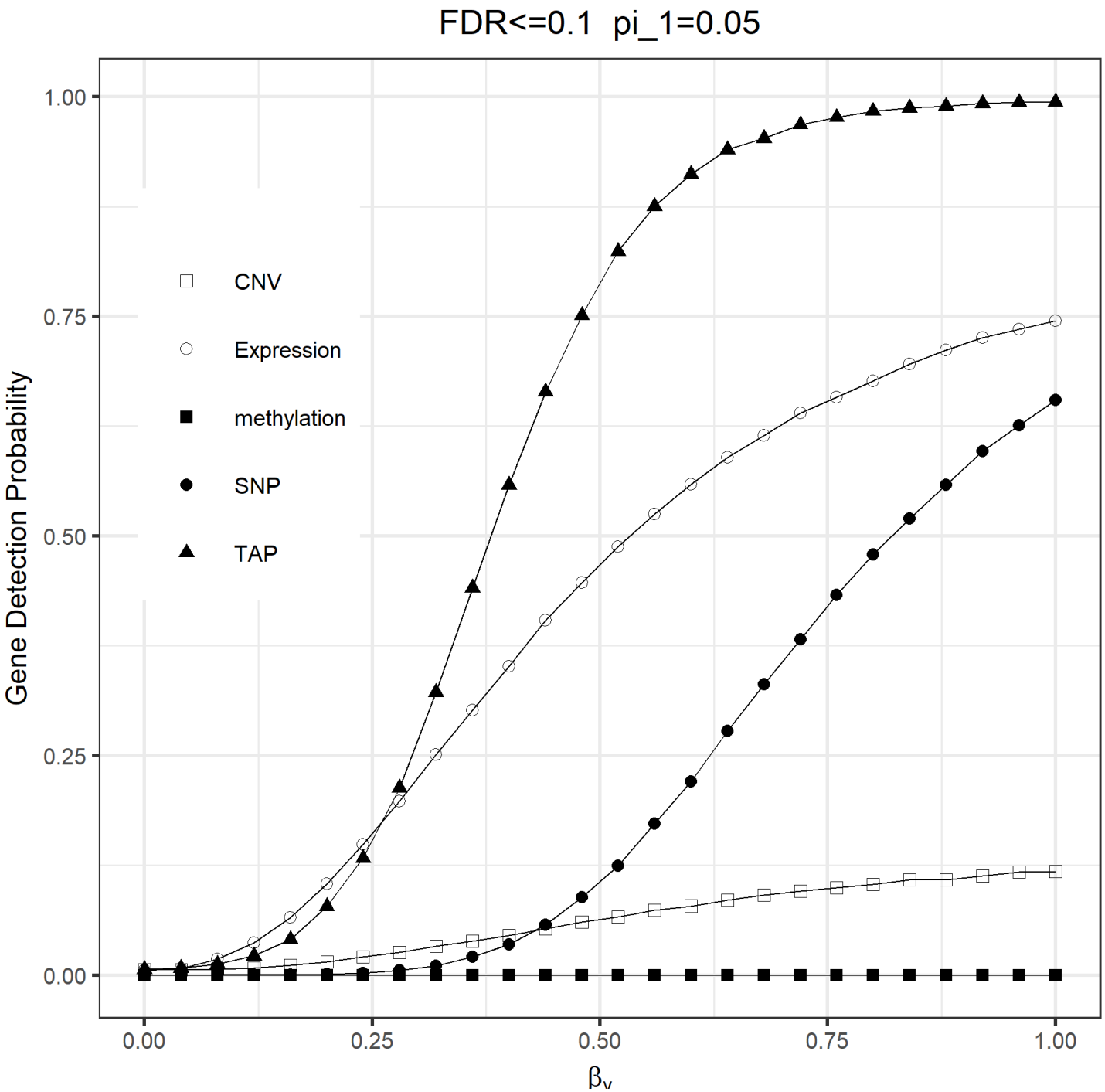

Figure 14. TAP detecting power versus other features. (FDR<=0.1, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 10% (FDR<=0.1).

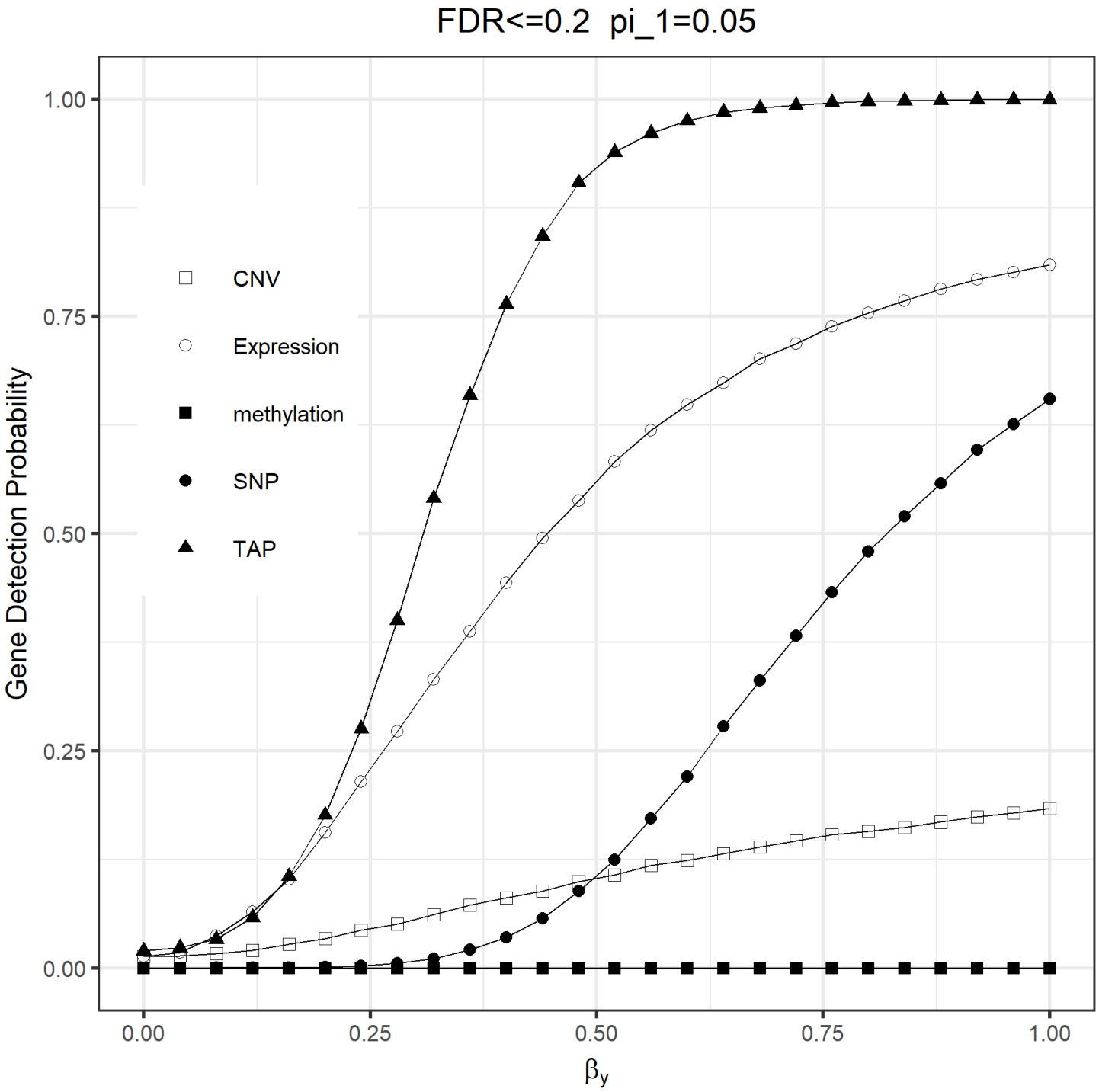

Figure 15. TAP detecting power versus other features. (FDR<=0.2, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 20% (FDR<=0.2).

When multiple genomic data are available, a more straight forward approach to combine gene detecting power is by requiring at least one genomic feature within a gene region to be significantly associated with an outcome (single genome feature method). As shown in Figure 16, Figure 17, Figure 18 and Figure 19, we compared gene detection probability by single genome feature and TAP under specified false discovery rate and true positive rate. When $\beta_y > 0.35$, TAP is more powerful than single genome feature, indicating TAP can accommodate gene detecting power from features in different platforms more efficiently. The difference in gene detecting probability between TA P and genome feature is largest when β_y is within [0.6, 0.8]. The true positive rate (pi_1) and assumed false discovery rate (FDR) affect both curves the same away as they do when TAP is studied in single platform.

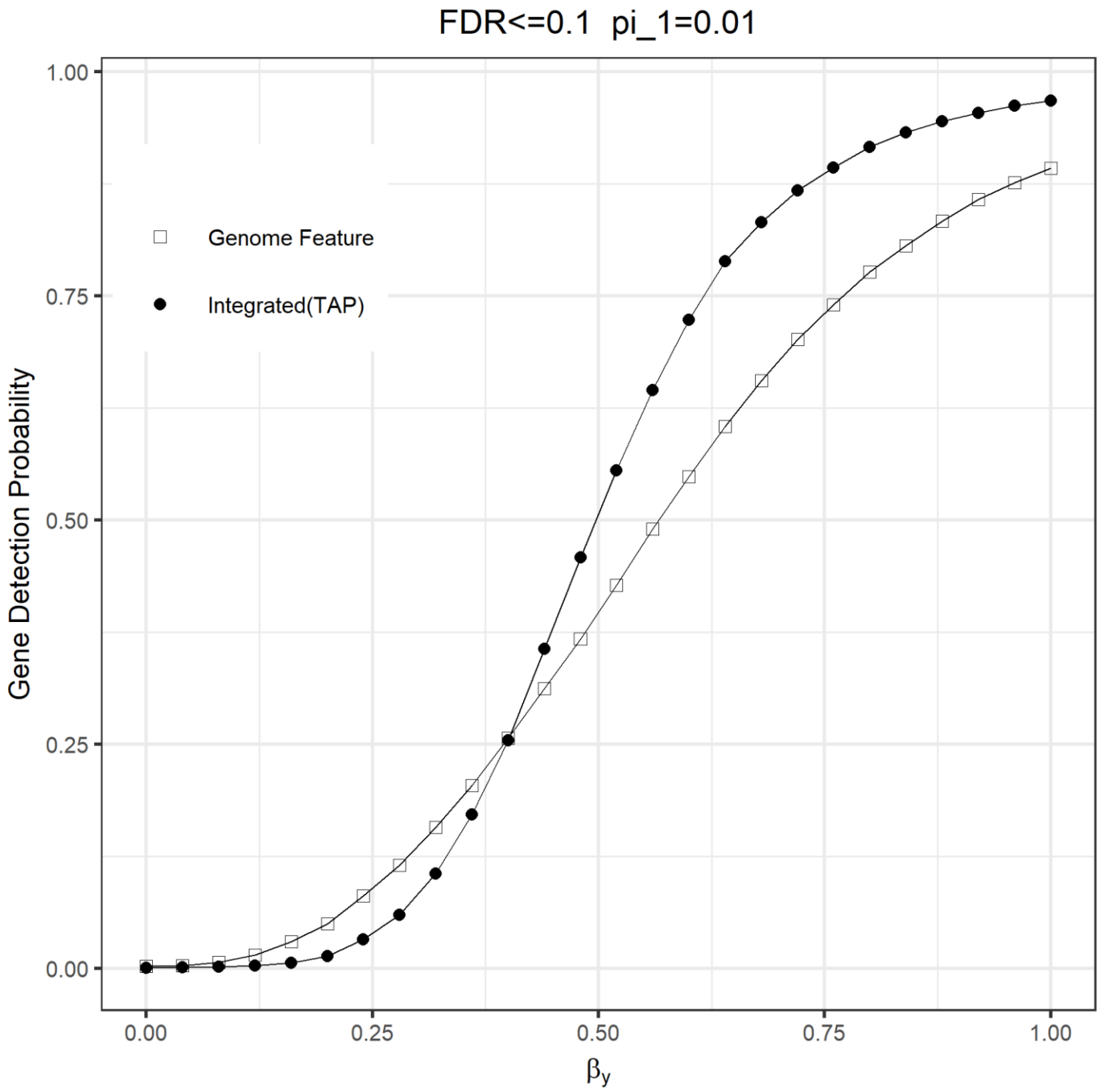

Figure 16. TAP detecting power versus any single feature. (FDR<=0.1, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 10% (FDR<=0.1).

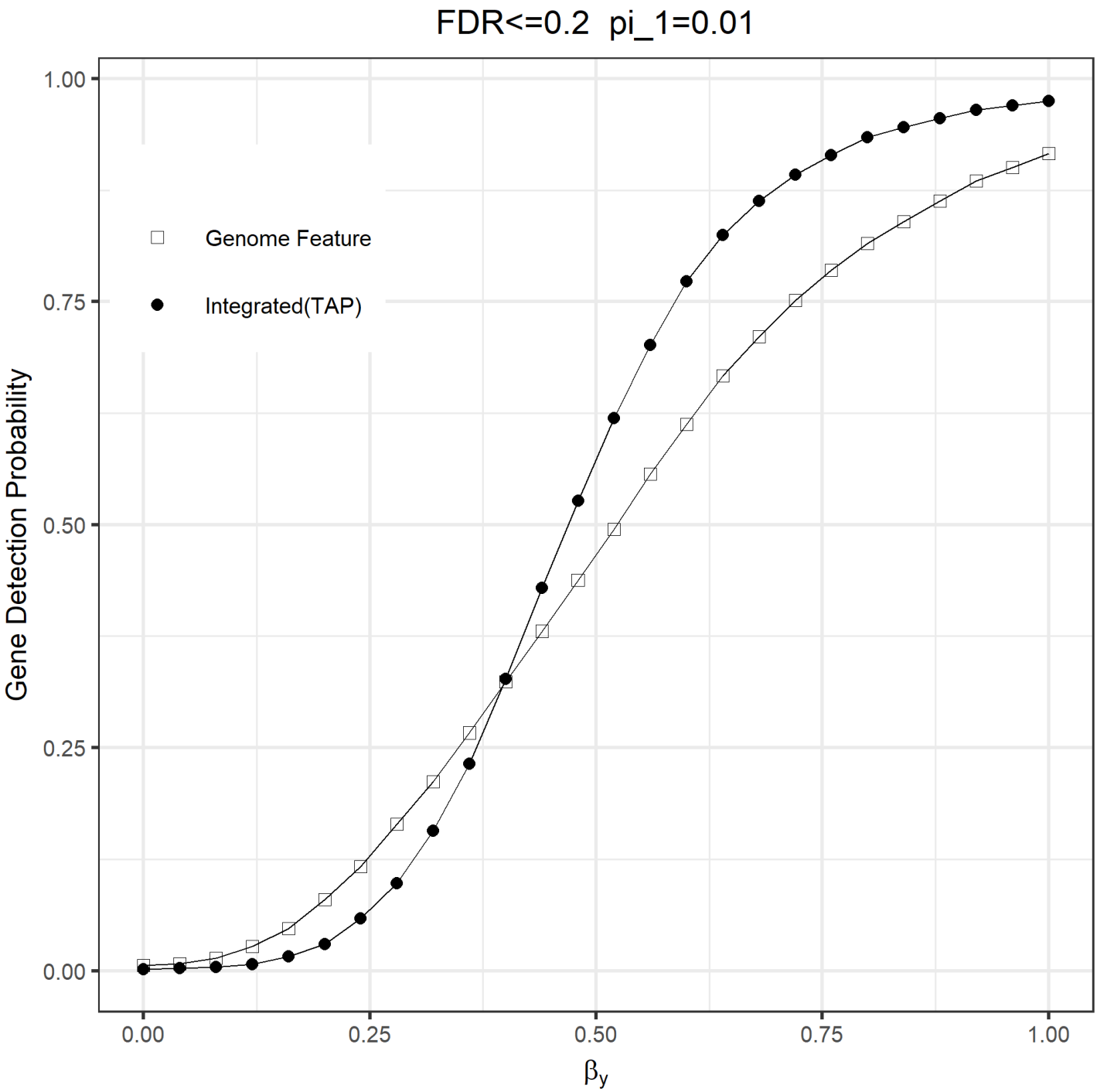

Figure 17. TAP detecting power versus any single feature. (FDR<=0.2, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 20% (FDR<=0.2).

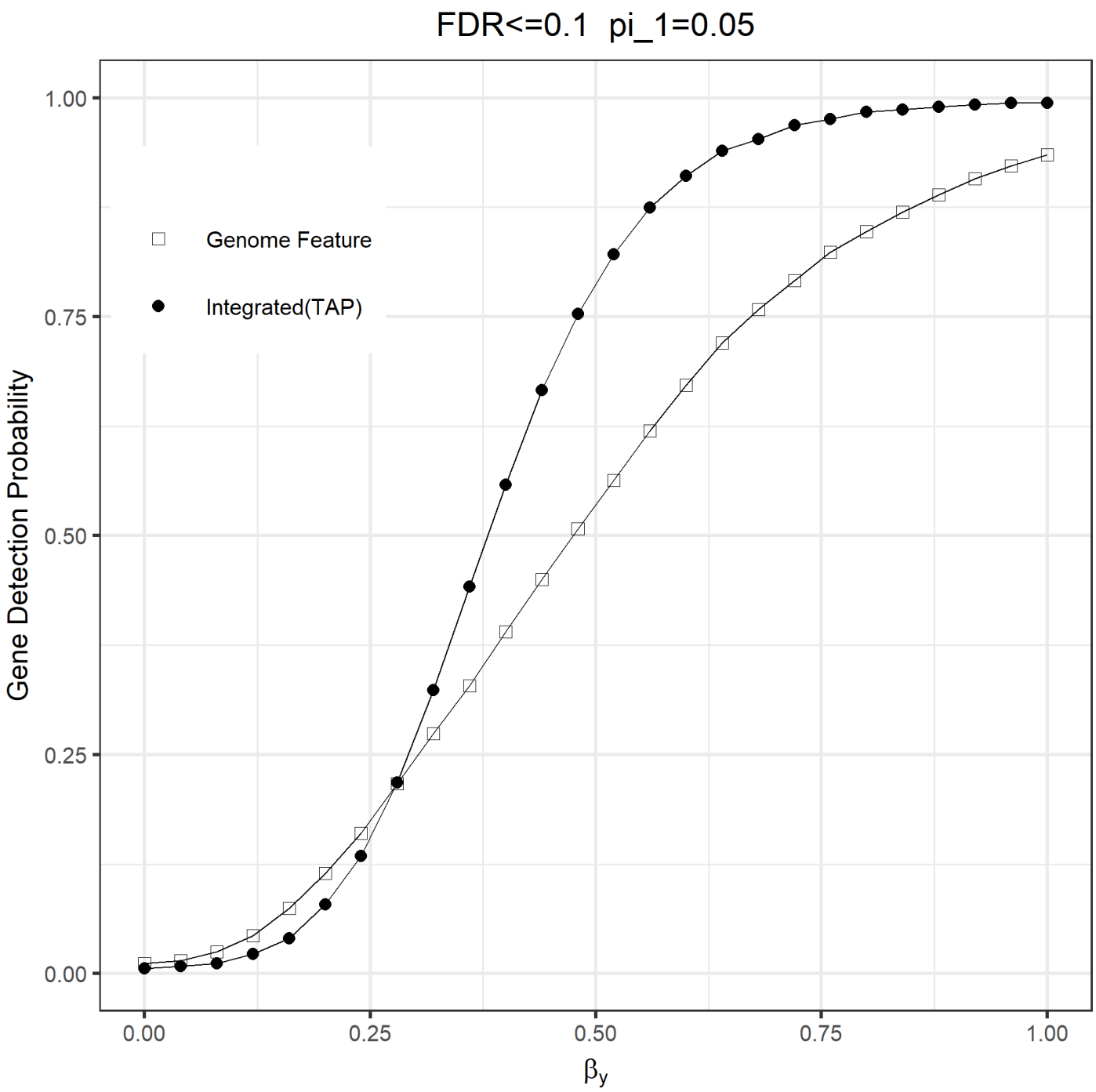

Figure 18. TAP detecting power versus any single feature. (FDR<=0.1, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 10% (FDR<=0.1).

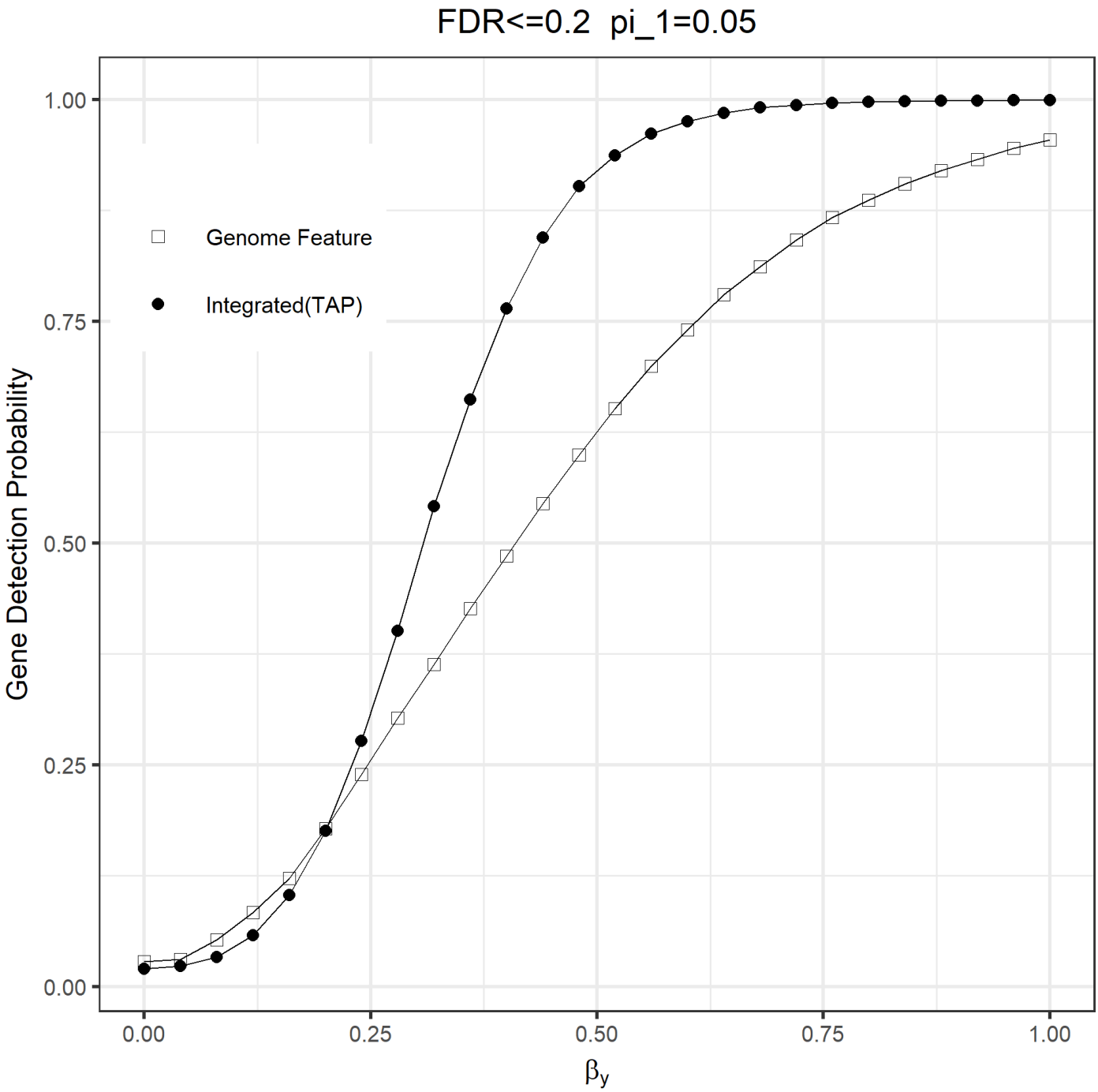

Figure 19. TAP detecting power versus any single feature. (FDR<=0.2, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 20% (FDR<=0.2).

Next, we explored power curve of TAP with different P-value cutoff (Fig 20, Fig 21, Fig 22 and Fig 23). When the true positive rate (pi_1) is set at 0.05, TAP is more powerful than single feature method when $\beta_y > 0.35$ with FDR<=0.1 or $\beta_y > 0.2$ with FDR<=0.2. Different δ didn't change the TAP power curve regardless FDR value. On the other hand, when true positive rate is 0.01, TAP power curve will spread around single feature power curve (label as genome feature). When δ is close to 0.5, such as 0.3, 0.4, 0.5, 0.6 and 0.7, TAP power curve falls below the single feature power curve in most part of β_y range. When δ is close to the left and right tail (such as 0.01, 0.05, 0.1, 0.2, 0.8, 0.9 and 1) TAP power curve is above the single feature power curve for most values of β_y. The closer of δ to 0 or 1, the more powerful is TAP procedure with more power when δ is close to 0. Therefore, we may need to set a smaller δ to gain more power in TAP. However, smaller δ leads to more conservative TAP P-value under null hypothesis as seen from previous result, which means higher false negative rate during discovery.

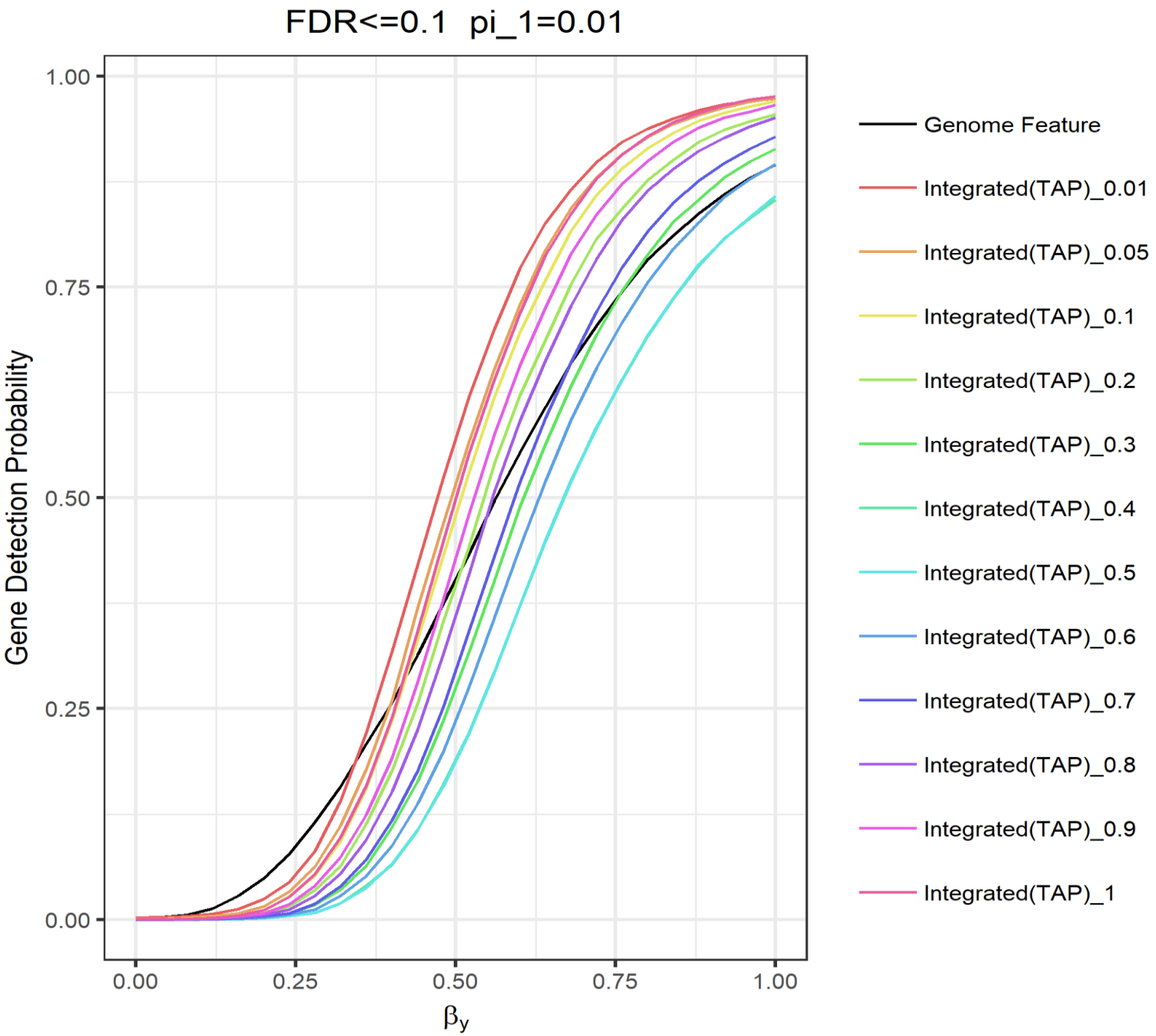

Figure 20. TAP detecting power versus any single feature. (FDR<=0.1, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 10% (FDR<=0.1). TAP P-value were calculated at different δ indicated.

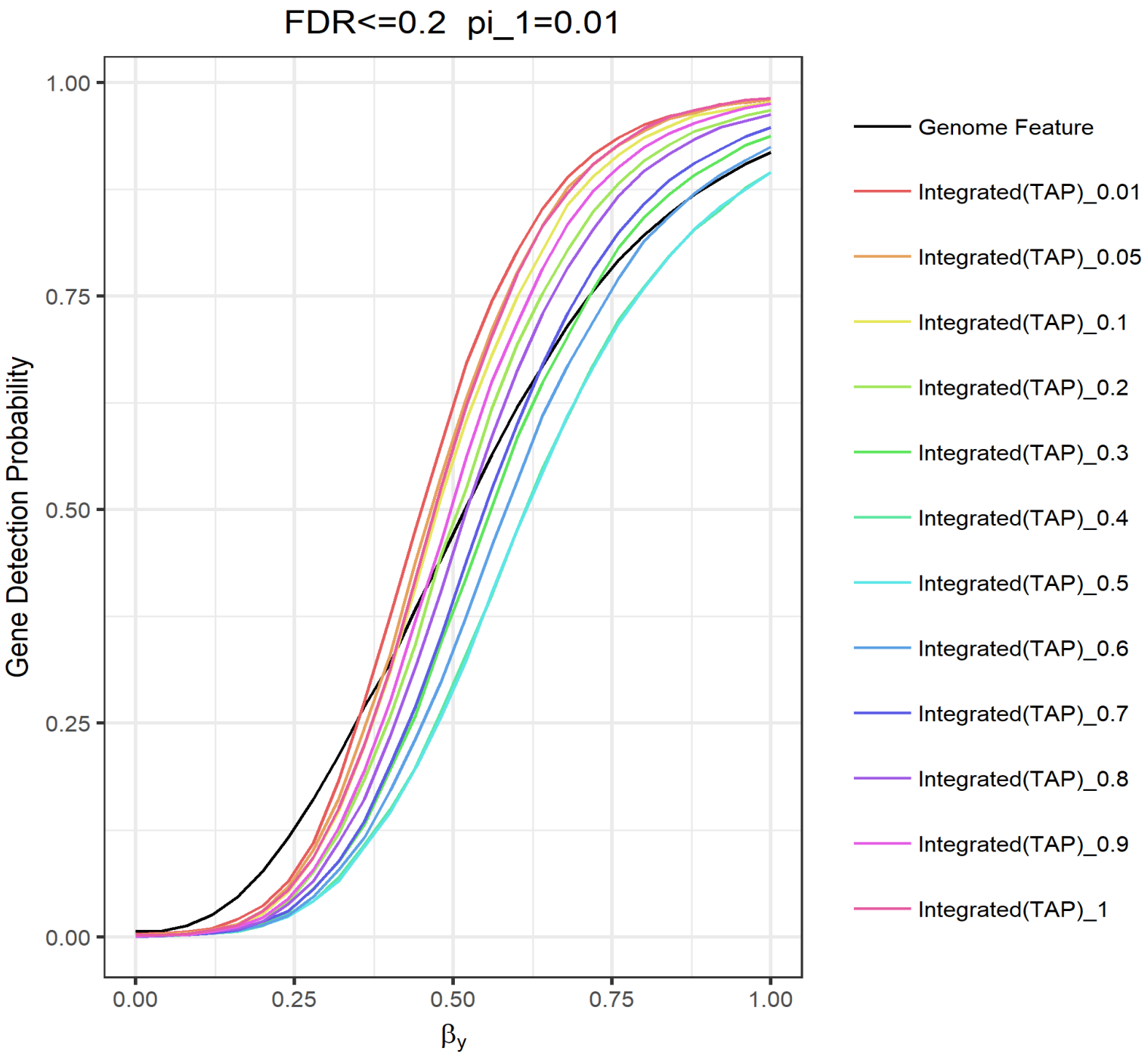

Figure 21.TAP detecting power versus any single feature. (FDR<=0.2, pi_1=0.01)

True positive rate was set to 1% (pi_1=0.01) and the assumed false discovery rate is set at 20% (FDR<=0.2). TAP P-value were calculated at different δ indicated.

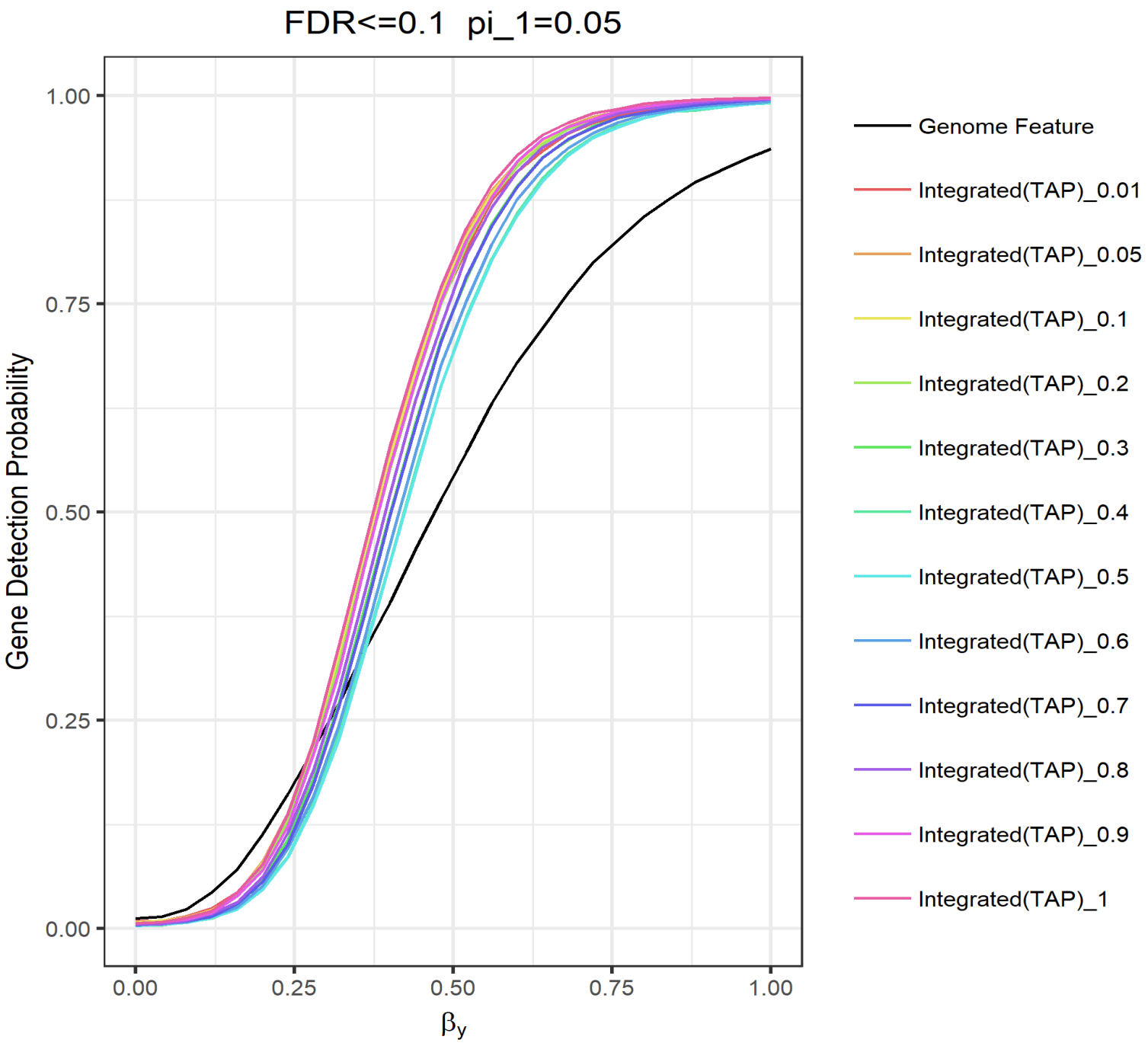

Figure 22. TAP detecting power versus any single feature. (FDR<=0.1, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 10% (FDR<=0.1). TAP P-value were calculated at different δ indicated.

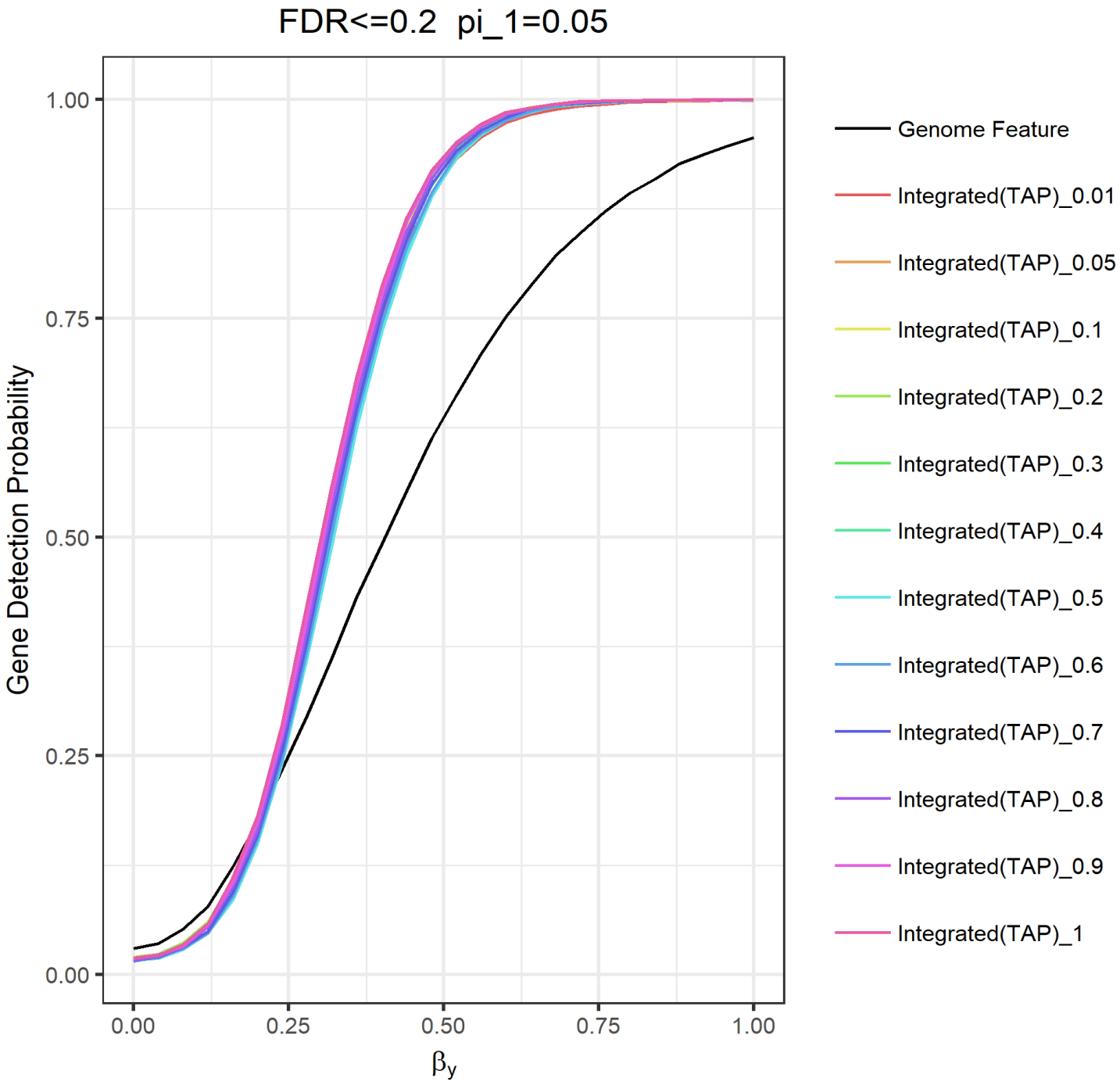

Figure 23. TAP detecting power versus any single feature. (FDR<=0.2, pi_1=0.05)

True positive rate was set to 5% (pi_1=0.05) and the assumed false discovery rate is set at 20% (FDR<=0.2). TAP P-value were calculated at different δ indicated.

3.2.4 Application to real data

In a collaborative study with hematology departement at St Jude Children's Research Hospital, we applied the preocdure to analyze effect of multiple genomic variants on Acute Lymphoid Leukemia (ALL) drug resistance at gene level (Autry, et al., 2020). The association of each feature with prednisolone sensitivity was tested with appropriate test. P-values of features within 50 KB upstream or downstream of the gene coding region were integrated by TAP method with $\delta = 0.05$ to obtain a gene-level TAP statistic with associated P-value. We identified 903 out of 19725 genes that are significantly associated with ALL drug resistance. Those genes are then combined with genes discovered by other 2 traditional methods (polygenomic and CRISPER). A total 15 genes were found to have significant association in all three methods. Then one of the top candidate – CELSR2 is selected and validated by another cohort. It's role in ALL drug resistance was further confirmed by gene knockout in cell line. The significant association CELSR2 with drug resistance is found to be mainly driven by highly significant gene expression probes from gene expression array (Fig 24)

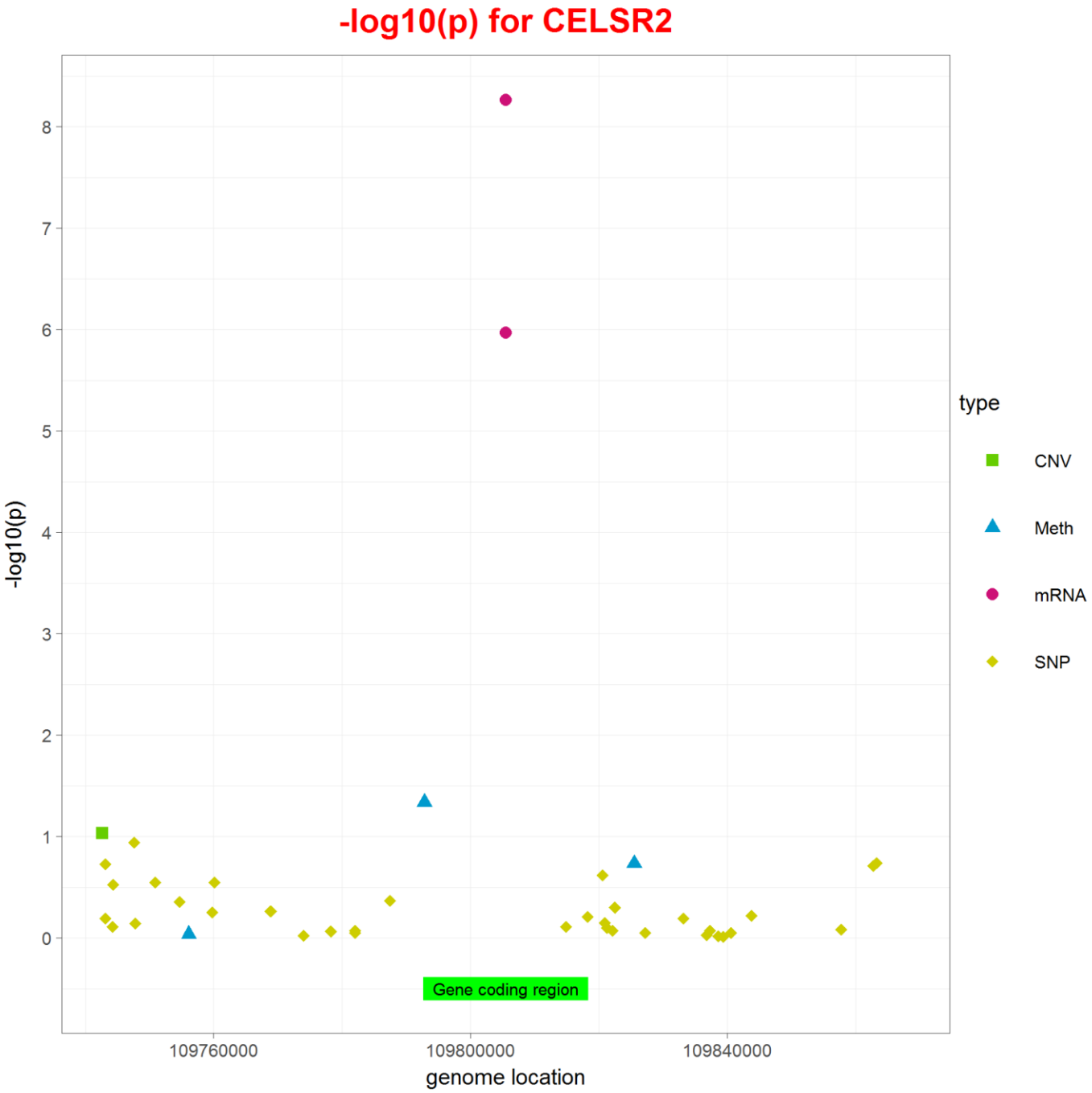

Figure 24. Time Square plot of CELSR2.

-log10 transformed single feature P-values are plotted along their genome location. Feature types are indicated.

3.3 GWAS bootstrap/permutation

Permutation and bootstrap resampling procedures are computationally intensive and time consuming in the context of big data such as those encountered in genomic application. On a personal computer, more than 15 minutes of computer time are required to complete a genome wide association study (GWAS) for a 500 sample-cohort with multiple genomic array data of two million probes. Hence in the interest of efficiency, the GWAS bootstrap and permutation resampling for those study reported in this thesis were carried out in a parallel high performance computing facility (HPCF). The genome feature data were split into small pieces of about 100 features each. Each piece of 100 features was processed by a Central Processing Unit (CPU). Several CPU's were run in parallel. This strategy ensured that 1000 bootstraps GWAS invoving about 2,500,000 features from 500 subjects could be complete in 24 hrs.

3.4 alignSeg: annotate genomic features to allele

After the association tests were performed for each feature, the important step of summarizing the results from features to designated alleles, which was usually genes, was carried out based on their genome location. The traditional algorithm for implementing this procedure is to search for features that are located within each designated genome region. However this approach is extremely slow because the algorithm requires looping through all designated allele and searching about 2 million features at each loop. An alternative approach is to order the features and genes by their genome location (alignSeg) and then identify the features in the genes based on their location in the ordered data frame. For this purpose, we assume that there are two

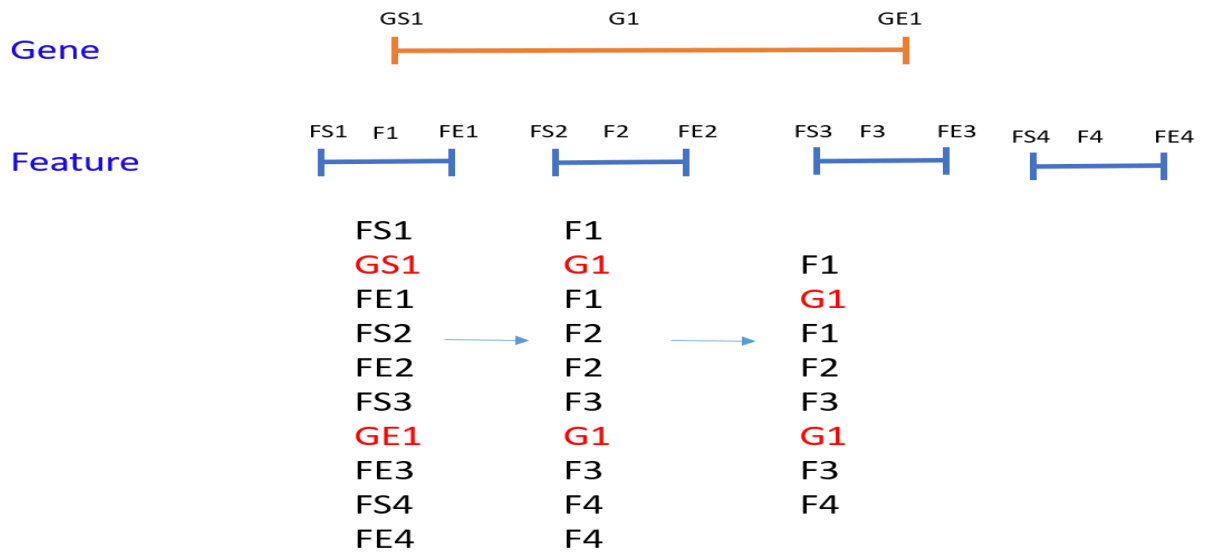

Figure 25. alignSeg algorithm.

Start location and end location of each segment are sorted in increasing order. The unique list of feature identifications between two identical gene identifications are the ones that located within the gene.

datasets – one for features and the other for gene datasets which contain column of genome location: chromosome, start location and end location (Fig 25). First, we create a data subset by chromosomes and following steps will be performed within each chromosome.

1. Split each feature/gene into two rows, one with start location and the other with end location.
2. Create a data-frame with a column named “loc” to hold start/end location and a column with corresponding gene or feature identification.
3. Reorder the data frame by “loc” from smallest to largest.
4. Features located between two identical gene id will be the features in this gene region.

Using this procedure, we only need to performed search and comparison among all features once, which significantly reduced the time to expedite the process.

4. Discussion

GWAS is a well-established procedure for finding association between genes and traits. However, the selection bias and lack of reliable method to summarize feature results over gene substantially makes the procedure inefficient for researchers in the lab. In this study we have proposed TAP statistics to summarize P-values of each features over gene. We have developed a theoretical method for constructing the empirical distribution function of TAP using a template CDF and a Kernel-based smoothing function.

Furthermore, we constructed a scheme to simulate genotype data without disrupting SNP's linkage structure, methylation data, expression data, CNV data and clinical data. The TAP statistic performance was evaluated under null hypothesis ($\beta_y = 0$) and different β_y between (0,1). The ratio of true positive among all features within gene, the P-value threshold (δ), the strength of association between gene and outcome (β_y) and parameters used in FDR procedure were all important in determining the sensitivity and specificity of TAP. Moreover, we evaluated the effect of P-value threshold (δ) and association strength (β_y) on behavior of TAP P-value. To simplify the simulation model, we fixed the number of true positive features (13 features) and the number of all features (51 features). We found that smaller ratio of true positive features and all features reduced the power of TAP when δ is fixed. In our current study model that ratio is large. It would be interesting in future studies to check how the TAP power would change when the ratio is close to 0 such as 13/5100.

Permutation is the most time-consuming steps in TAP procedure. In current setting, 200 permutations are performed. The number of permutations will directly affect the precision of TAP P-value. The more permutation we can perform, the smaller TAP P-value we can go but with high price paid. On the other hand, we would expect that only a small fraction of genes out of ~20,000 genes are associated with outcomes. A small P-value won't be needed for majority genes but a lot of computing power are spent on them. So, an adaptive permutation procedure may be a good choice to reduce the computing burden. It considers the use of permutation to find EDF and P-value as a negative binomial process. Thus, the P-value from permutation would be a random variable follow negative binomial distribution. Therefore, we may not need to perform all permutations to exclude some genes that has low probability to gain a small P-value (Che R, 2014).

Furthermore, instead of using truncated gamma distribution to derive template CDF, we could divide the single feature P-value by selected δ, which may simplify the template CDF. When conditioned on selected δ we may have a more elegant template CDF which may lead to better small TAP P-value performance.

To implement TAP method in real study, two computation challenges need to be addressed. First, a limited number of permutations (~200) should be performed for each feature. Since it takes ~5 hours to finish one GWAS test for a 500-patient cohort with combined ~2,000,000 features in a personal computer. Thus, we propose to develop a pipeline based on high performance computing facility (HPCF) at St Jude Children's Research hospital which can process thousands of jobs concurrently. With this pipeline it will be possible to perform 200

permutations in ~2 hours.

Another computational obstacle we faced was to annotate features to allele/gene based on their genome location. For each permutation, we need to be able to summarize features to genes so that a single value can be computed for each gene. It will take traditional method 20-30 minutes to finish one permutation. While this time was reduced to ~30 seconds with alignSeg algorithm, all 200 permutations can be finished within ~2 hours.

TAP statistics with alignSeg and GWAS implemented with HPCF provides a comprehensive GWAS solution for researchers in lab starting from raw data to comput P-value. This information gathered from these procedures can be used to decide if the gene is worth further investigation.

5. References

Aichin, M. & Gensler, H., 1996. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods.. *Am J Public Health,* 86(5)(May), pp. 726-728.

Autry, R. J., Paugh, S. W., Carter, R. & al, e., 2020. Integrative genomic analyses reveal mechanisms of glucocorticoid resistance in acute lymphoblastic leukemia.. *Nat Cancer,* Volume 1, pp. 329-344.

Benjamini, Y. & Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics,* 29(4), p. 1165–1188.

Bush WS, M. J., 2012. Genome-Wide Association Studies.. *PLoS Comput Biol* , 8(12).

Che R, J. J. M.-R. A. B. C., 2014. An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. *BioData Mining,* Volume 14, pp. 7-9.

Cheng, c. & Parzen, E., 1997. Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference,* Volume 59, pp. 291-307.

Devore, R., 1972. *The Approximation of Continuous Functions by Positive Linear Operators.* New York: Springer-Verlag.

Dudbridge, F. & Koeleman, B., 2003. Rank truncated product of P-values, with application to genomewide association scans.. *Genet Epidemiol.,* 25(4), pp. 360-6.

Edgington, E. S., 1972. An additive method for combining probability values from independent experiments.. *J. Psychol.,* Volume 80, pp. 351-363.

Faye, L., Sun, L., Dimitromanolakis, A. & Bull, S., 2011. Flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design.. *statistical methods,* Volume 30, pp. 1898-1912.

Firth, D., 1993. Bias Reduction of Maximum Likelihood Estimates. *Biometrika,* 80(1), pp. 27-38.

Fisher, R. A., 1932. *Statistical methods for research workers.* 4th ed. London: Oliver and Boyd.

Heard, N. A., 2018. Choosing between methods of combining p-values. *Biometrika,* Volume 105, p. 239–246.

Hedges, L. & Olkin, I., 1985. *Statistical methods for meta-analysis.* San Diego: Academic Press.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), p. 800–802.

Hochberg, Y. & Benjamini, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57(1), pp. 289-300.

Holm, S., 1979. A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics,* Volume 6, pp. 65-70.

Huber, P., 1977. *Robust statistical procedures.*. New York: SIAM Regional Conference Series in Applied Mathematics Society for Industrial and Applied Mathematics.

International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature,* Volume 437, p. 1299–1320.

Lancaster, H., 1961. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics,* 3(1), pp. 20-33.

Lee S, A. G. B. M. L. X., 2014. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet.,* 95(1), pp. 5-23.

Mudholkar, G. & George, E., 1979. *The logit method for combining probabilities.*. New York, J. Rustagi, ed..

Mudholkar, G. & George, E., 1983. On the Convolution of Logistic Random Variables.. *Metrika,* Volume 30, pp. 1-14.

Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M., 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiology,* Volume 32, pp. 381-385.

Pearson, K., 1933. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random.. *Biometrika,* Volume 25, pp. 379-410.

Rosenthal, R., 1978. Combining results of independent studies. *Psychological Bull,* Volume 85, pp. 185-193.

Sarkar, S. K., 1998. Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Ann. Statist,* 26(2), pp. 494-504.

Storey, J. D. & Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.,* 100(16), pp. 9440-9445.

Stouffer, S. et al., 1949. *The American Soldier Adjustment During Army Life.*. Princeton, New Jersey: Princeton University Press.

Tippett, L., 1931. *The Methods of Statistics*.. London: Williams and Norgate..

Yu, K. et al., 2009. Pathway analysis by adaptive combination of P-values.. *Genet Epidemiol.,* 33(8), pp. 700-9.

Zaykin, D., 2002. Truncated product method for combining P-values. *Genetic Epidemiology,* Volume 22, pp. 170-185.