Supplementary Data: Biomarker Discovery for Heterogeneous Diseases

Garrick Wallstrom, Karen S. Anderson and Joshua LaBaer

Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, 1001 S. McAllister Ave. Arizona State University, Tempe, Arizona, 85287-5001

*Corresponding author:

Garrick Wallstrom, Ph.D.

Center for Personalized Diagnostics, Biodesign Institute, ASU, 1001 S. McAllister Ave

Tempe, AZ 85287-6401

Telephone: 480-727-7482; Email: garrick.wallstrom@asu.edu

Supplementary Methods

Statistical Computation

We conducted all simulations at the ASU Advanced Computing Center, which has more than 5,000 processor cores and 40 teraFLOPS of computation power. We used R(37) for all simulations and statistical analysis. We used the SNOW package(38) for parallel computation. We used the limma package(39) for the empirical Bayes analysis.

Simulation

Our stochastic model of the single-stage biomarker screening process is as follows. Let N_1 and N_2 denote the number of controls and cases, respectively. We let X_{1ij} denote the response (normalized and transformed as appropriate) for control i and candidate biomarker j, and X_{2ij} denote the response for case i and candidate biomarker j. In both the homogeneous and heterogeneous models, all response variables are mutually independent. In the homogeneous model,

$$X_{1ij} \sim N(0,1), \text{ for } i = 1, \ldots, N_1 \text{ and } j = 1, \ldots 10000,$$

$$X_{2ij} \sim N(0.80,1), \text{ for } i = 1, \ldots, N_2 \text{ and } j = 1, \ldots 50, \text{ and}$$

$$X_{2ij} \sim N(0,1), \text{ for } i = 1, \ldots, N_2 \text{ and } j = 51, \ldots 10000.$$

For the heterogeneous model we introduce random variables denoted by W_{ij} which indicate for each case i and each true biomarker j whether case i belongs to the responding subtype. Formally, the heterogeneous model is as follows.

$$X_{1ij} \sim N(0,1), \text{ for } i = 1, \ldots, N_1 \text{ and } j = 1, \ldots 10000,$$

$W_{ij} \sim Bernoulli(0.20)$, for $i = 1, \ldots, N_1$ and $j = 1, \ldots 50$,

$X_{2ij} | W_{ij} \sim N(2.49 \times W_{ij}, 1)$, for $i = 1, \ldots, N_2$ and $j = 1, \ldots 50$, and

$X_{2ij} \sim N(0,1)$, for $i = 1, \ldots, N_2$ and $j = 51, \ldots 10000$.

Selection of Biomarkers

Prior to formal analysis, intensity data are typically normalized and transformed. Normalization is used to remove systematic variation and facilitate unbiased comparisons both across and within arrays. Common sources of variation that may be addressed with normalization include pin effects and print order effects. While there are many approaches to normalization(40, 41), in our study we will assume that appropriate normalization has been performed on the intensity data. Transformations of intensity data are also common to stabilize variances prior to statistical analysis. Logarithmic and arcsinh transformations are common choices for variance stabilization. Again, we will assume that the intensity data have been appropriately transformed.

A vast number and variety of statistical methods have been used for the identification of potential biomarkers. We made no attempt here to be comprehensive and compare the performance of a broad spectrum of biomarker selection methods. Instead, we concentrated on eight methods that were selected to highlight the role that heterogeneity plays in sample size determination for biomarker discovery studies. For the single-stage design and the second stage of a two-stage design, we used the methods to produce p-values and then used the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at 10% for each method(42). For the first-stage of a two-stage design, we used the methods to produce scores which were then used to rank the candidates and select the top 750.

1. Ordinary t-test(43). We used a one-sided ordinary two-sample t-test to evaluate the hypothesis that the mean response for cases is greater than for controls. For computational efficiency, we computed test statistics using lmFit from the limma package as a by-product of the empirical Bayes analysis. First-stage scores for this method were simply the t-test statistics.
2. Welch's t-test(44). We used a one-sided Welch's t-test to evaluate the hypothesis that the mean response for cases is greater than controls. This adaptation of the ordinary t-test is more appropriate when the populations have unequal variances. We used the t.test function in R to compute this test. Scores for this method were the t-test statistics.
3. Empirical Bayes moderated t-test(45). This test uses empirical Bayes methodology to improve estimates of standard deviations for a t-test to evaluate the hypothesis that the mean response for cases is greater than for controls(45). We used the ebayes and lmFit functions in the limma package to compute this test(39). Scores for this method were the moderated t-statistics.
4. Kolmogorov-Smirnov test(46, 47). We used a one-sided Kolmogorov-Smirnov test to test the hypothesis that the distribution of case responses is stochastically greater than that of control responses. We used the ks.test function in R to compute this test. Scores for this method were associated test statistics.
5. Mann-Whitney U test(48, 49). We used a one-sided Mann-Whitney U test to evaluate the hypothesis that the distribution of case responses is stochastically greater than that of control responses. We used the wilcox.test function in R to compute this test. Scores for this method were the associated test statistics.
6. Area under the ROC curve (AUC). The receiver operator characteristic (ROC) curve is a popular approach for summarizing the tradeoff between the sensitivity and specificity of a diagnostic test across a range of classification thresholds(50). Formally, the ROC

curve is defined as a set of points,

$$ROC(\cdot) = \{(1 - \text{specificity}(h), \text{sensitivity}(h)), h \in (-\infty, +\infty)\},$$

where *h* is the classification threshold. The area under the ROC curve (AUC) is one measure of the overall discrimination ability of a diagnostic test. The AUC has an intuitive interpretation as the probability that a randomly selected case has higher intensity than a randomly selected control. It is also equivalent to the Mann-Whitney U test statistic. Here we estimated the AUC using a non-parametric estimator(51). We tested the hypothesis that the AUC is greater than 0.5, which is the expected AUC for a non-discriminatory diagnostic test. We computed p-values using a short-cutting permutation test with up to 50,000 permutations for each p-value. First-stage scores were the estimated AUC values, and are equivalent to the scores from the Mann-Whitney U test.

7. Partial area under the ROC curve (PAUC). One drawback of using AUC is that it is a weighted average of sensitivity across all levels of specificity. However, in practice only a range of specificities may be clinically relevant. For example, if the consequences of false positives are significant, the sensitivity of a potential biomarker at 50% specificity is of little interest. This argument has led some researchers to use the PAUC, which is the area under the ROC curve but only over a range of specificities of interest(51, 52). We find this targeted averaging compelling in the context of disease heterogeneity. Here we considered the PAUC in the region where the specificity is greater than 95% and estimate its value using a non-parametric estimator(51). We tested the hypothesis that the PAUC is greater than 0.00125, which is the expected PAUC for a non-discriminatory diagnostic test. We computed p-values using a short-cutting permutation test with up to 50,000 permutations for each p-value. First-stage scores were the estimated PAUC values.

8. Sensitivity. Another method that is similar to the PAUC is to examine the sensitivity at a fixed level of specificity, or vice versa. Here we measured the sensitivity at 95% specificity and tested the hypothesis that the sensitivity exceeds 5%, which would be the sensitivity if there are no differences between the case and control populations. We computed p-values using a short-cutting permutation test with up to 50,000 permutations for each p-value. Scores for this method were the estimated sensitivities.

These selection methods can be organized into three groups. The t-tests (methods 1-3) evaluate whether there is a difference in the means of the two populations and operate under either an assumption that the populations are normal, or that the sample sizes are sufficiently large to induce asymptotic normality via the central limit theorem. The Kolmogorov-Smirnov, Mann-Whitney, and AUC tests (methods 4-6) evaluate whether case responses tend to be larger than control responses. These tests are nonparametric; that is, they do not require distributional assumptions. The PAUC test and the sensitivity test (methods 7-8) are nonparametric methods for evaluating whether there is sufficient mass in the case response distribution in the right tail of the control response distribution. Among all eight methods, most biomarker studies use variants of t-tests, perhaps because of their ability to easily accommodate covariates into an analysis via linear models. The full AUC and Mann-Whitney tests are also commonly used in biomarker studies. While researchers have advocated for the use the partial AUC and sensitivity tests based upon their focus on clinically-relevant portions of the ROC curve(53), their use in biomarker discovery studies remains limited.

First Stage of a Two-stage Screening Study

We conducted a preliminary study to identify the best first stage scoring algorithm (out of 8 considered) for a two-stage screening study. We considered eight scoring algorithms, which were based upon the eight selection methods used in the single-stage study and are described

above, and determined the best scoring algorithm for each first-stage sample size and under both homogeneous and heterogeneous conditions. Using the same normal homogeneous and heterogeneous disease simulation models that we used in the single-stage studies we simulated 20 data sets at each sample size (N), for N between 10 and 100. For each data set and scoring algorithm we measured power to be the proportion of the 50 true biomarkers that were contained in the top 750 candidates. Means and standard errors of power are displayed in Figure S1. Based on these results, for the homogeneous disease model we used empirical Bayes moderated t-statistics when the sample size is at most 40 and the ordinary t-statistic when the sample size is greater than 40. For the heterogeneous disease model we used empirical Bayes moderated t-statistics when N=10 and PAUC otherwise.

Simulation using the Normal Distribution

Single stage results using normal samples are described in the main text and displayed graphically in Figure 2. Tables S1 and S2 provide the numerical results.

Simulation using the T Distribution

In a separate study we investigated the robustness of the normal distribution result by replacing the normal distribution with a t distribution with 3 degrees of freedom (t_3), which has heavier tails than the normal distribution. We simulated case and control responses to non-biomarkers, and control responses to true biomarkers independently from a t_3 distribution. For the homogeneous disease we simulated case responses to true biomarkers independently from a t_3 with location parameter of 1.37. See Figure 1d. For this homogeneous disease model true biomarkers have a sensitivity of 20% at 95% specificity and the AUC is 0.78. For the heterogeneous disease we again simulated responses such that only 20% of cases would have differential response.

Specifically, with 20% probability we simulated the response for an individual case from a t_3 with location parameter of 3.33; otherwise, we simulated the response from a standard t_3. See Figure 1e. Under this heterogeneous mixture model the overall sensitivity is 20% at 95% specificity and the AUC is only .59. The ROC curves for the homogeneous and heterogeneous disease models are shown in Figure 1f.

Power estimates under 10% FDR control using ten simulated data sets at each sample size are given in Tables S3 and S4 and displayed graphically in Figure S2. For the homogeneous disease, the Mann-Whitney, AUC and Kolmogorov-Smirnov tests identified over 90% of the true biomarkers using only 50 cases and 50 controls. By contrast, the PAUC and sensitivity tests identified only approximately 15% and 30%, respectively of the true biomarkers with the same sample size. For the heterogeneous disease, 200 cases and 200 controls were needed before any method achieved 90% power, and only the sensitivity test reached that level of performance. In general we observe that the Mann-Whitney, AUC, and Kolmogorov-Smirnov tests perform the best with t_3 homogeneous disease model, while the PAUC and sensitivity test perform the best with t_3 heterogeneous disease model.

As with normal samples, the optimal statistical methods with t_3 samples differ dramatically depending on whether heterogeneity is present, and the sample sizes required under heterogeneity are much greater than under homogeneity. In fact, the difference between the homogeneous and heterogeneous disease models is more dramatic with t_3 samples than with normal samples. While 90% power can be achieved with 50 cases and 50 controls under the homogeneous disease model, only 5% power can be achieved with the same sample sizes and the heterogeneous disease model.

Performance under 50% FDR Control

Our primary results are based upon controlling the false discovery rate at 10%. In order to assess the impact of changing the false discovery rate, we used the same simulations but controlled the false discovery rate at 50%. Estimated power for a single stage design using normal and t_3 samples are given in Figures S3 and S4, respectively. Estimated power for a two stage design using normal samples with 100 total case and control samples are given in Figure S5. Estimated power with 200 total case and control samples are given in Figure S6.

Actual Breast Cancer Screening Study

Results from applying the eight selection methods to data from an actual breast cancer screening study are described in the main text. Summary statistics for hits found at 5% FDR control are given in Table S5. Table S6 lists the 37 proteins that were highly significant (1% FDR control) using any of the eight methods, along with their q-values (the minimum FDR for which the protein would be significant).

Simulation Study of Empirical Bayes

In order to assess the impact of simulated homoscedasticity of the biomarkers on the performance of the empirical Bayes moderated t-test, we conducted a small simulation study of the single-stage design in which the biomarker variances were simulated according to an inverted gamma distribution with mean 1.0 and standard deviations ranging from 0.1 up to 10. We conducted simulations for both homogeneous and heterogeneous diseases using case and control sample sizes of 25, 50, 100 and 200. For each experimental condition we simulated 20 datasets containing case and control responses for 10,000 candidate biomarkers. Estimated power is shown in Figure S7 for the homogeneous disease and Figure S8 for the heterogeneous disease. For the homogeneous disease, we observe improved power for the empirical Bayes moderated t-test when the sample sizes are 25 and the standard deviation is small (≤ 0.5); however, for larger sample sizes or larger standard deviations, the difference in

power is negligible. For the heterogeneous disease and sample sizes of at least 50 cases and 50 controls, the empirical Bayes method yields greater power relative to the ordinary t-test when the standard deviation is small (say, ≤ 0.2) and equal power for larger standard deviations.

References

37. Team RDC. R: A Language and Environment for Statistical Computing. 2.9.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2009.
38. Luke Tierney AJR, Na Li, and H. Sevcikova. Snow: Simple Network of Workstations. 0.3-3 ed.
39. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor: Springer, New York; 2005. p. 397-420.
40. Quackenbush J. Microarray data normalization and transformation. Nat Genet.
41. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265-73.
42. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.
43. STUDENT. THE PROBABLE ERROR OF A MEAN. Biometrika. 1908;6(1):1-25.
44. WELCH BL. THE GENERALIZATION OF ‘STUDENT'S’ PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. Biometrika. 1947;34(1-2):28-35.
45. Smyth G. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical applications in genetics and molecular biology. 2004;3(1).
46. Kolmogorov AN. Sulla determinazione empirica di una legge di distribuzione. Liorn Ist Ital Attuari. 1933;4:83-91.
47. Smirnov NV. Tables for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics. 1948;19(2):279-81.
48. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-3.
49. Whitney HBMaDR. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics. 1947;18(1):50-60.
50. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction: Oxford University Press; 2003.
51. Dodd LE, Pepe MS. Partial AUC Estimation and Regression. Biometrics. 2003;59(3):614-23.
52. McClish DK. Analyzing a Portion of the ROC Curve. Med Decis Making. 1989;9(3):190-5.
53. Pepe M, Longton G, Anderson G, Schummer M. Selecting differentially expressed genes from microarray experiments. Biometrics. 2003;59(1):133 - 42.

Table S1: Normal Samples Homogeneous Disease Summary Statistics

N	Method	Power	FDR	Total Hits	True Hits	False Hits
25	T-test	11.9 (1.7)	8.1 (2.1)	6.5 (0.9)	6.0 (0.8)	0.6 (0.2)
	Welch T	11.0 (1.6)	7.8 (2.3)	6.0 (0.9)	5.5 (0.8)	0.6 (0.2)
	M-W U	5.0 (1.3)	3.2 (1.8)	2.6 (0.7)	2.5 (0.6)	0.1 (0.1)
	AUC	9.9 (1.5)	11.0 (3.2)	5.7 (0.8)	5.0 (0.7)	0.7 (0.2)
	PAUC-95	1.3 (0.4)	18.5 (6.6)	0.9 (0.3)	0.6 (0.2)	0.2 (0.1)
	Sensitivity	1.4 (0.5)	14.8 (7.6)	0.8 (0.2)	0.7 (0.2)	0.1 (0.1)
	EB T	18.9 (2.1)	5.3 (1.4)	10.0 (1.1)	9.4 (1.1)	0.6 (0.1)
	K-S	2.2 (0.6)	2.3 (1.7)	1.1 (0.3)	1.1 (0.3)	0.0 (<0.1)
50	T-test	70.3 (2.0)	9.4 (0.9)	38.9 (1.3)	35.1 (1.0)	3.8 (0.5)
	Welch T	69.8 (2.1)	9.3 (1.0)	38.6 (1.2)	34.9 (1.0)	3.7 (0.4)
	M-W U	65.1 (1.9)	8.5 (1.0)	35.5 (0.9)	32.5 (0.9)	3.0 (0.3)
	AUC	65.0 (1.7)	9.9 (1.0)	36.1 (0.9)	32.5 (0.9)	3.6 (0.4)
	PAUC-95	7.2 (1.1)	13.1 (3.1)	4.2 (0.6)	3.6 (0.6)	0.6 (0.2)
	Sensitivity	9.2 (1.3)	8.1 (3.0)	5.2 (0.7)	4.6 (0.6)	0.6 (0.2)
	EB T	74.5 (1.7)	10.0 (1.2)	41.6 (1.2)	37.2 (0.8)	4.3 (0.6)
	K-S	38.8 (1.3)	4.9 (1.2)	20.4 (0.6)	19.4 (0.7)	1.0 (0.2)
75	T-test	94.6 (0.7)	10.2 (0.9)	52.8 (0.7)	47.3 (0.3)	5.5 (0.5)
	Welch T	94.6 (0.7)	10.1 (0.9)	52.7 (0.7)	47.3 (0.3)	5.4 (0.5)
	M-W U	92.7 (0.9)	8.6 (1.0)	50.9 (0.8)	46.4 (0.5)	4.5 (0.6)
	AUC	92.4 (1.0)	9.3 (0.8)	51.0 (0.7)	46.2 (0.5)	4.8 (0.5)
	PAUC-95	19.3 (1.4)	8.4 (2.1)	10.6 (0.7)	9.7 (0.7)	0.9 (0.2)
	Sensitivity	29.3 (1.8)	4.6 (1.5)	15.3 (0.9)	14.7 (0.9)	0.7 (0.2)
	EB T	95.6 (0.6)	10.2 (1.1)	53.5 (0.9)	47.8 (0.3)	5.7 (0.7)
	K-S	74.6 (1.9)	6.0 (0.9)	39.8 (1.1)	37.3 (0.9)	2.5 (0.4)
100	T-test	98.7 (0.4)	8.8 (1.1)	54.2 (0.7)	49.4 (0.2)	4.9 (0.7)
	Welch T	98.7 (0.4)	8.8 (1.1)	54.2 (0.7)	49.4 (0.2)	4.9 (0.7)
	M-W U	98.1 (0.4)	8.2 (0.8)	53.5 (0.5)	49.0 (0.2)	4.5 (0.5)
	AUC	98.0 (0.5)	9.1 (0.8)	54.0 (0.6)	49.0 (0.2)	5.0 (0.5)
	PAUC-95	33.1 (1.7)	11.8 (1.5)	18.8 (1.0)	16.6 (0.8)	2.2 (0.3)
	Sensitivity	44.4 (2.3)	6.7 (0.9)	23.9 (1.3)	22.2 (1.2)	1.6 (0.3)
	EB T	99.3 (0.3)	8.8 (1.0)	54.5 (0.6)	49.6 (0.1)	4.9 (0.6)
	K-S	91.7 (0.8)	7.9 (0.7)	49.8 (0.5)	45.9 (0.4)	4.0 (0.4)
150	T-test	100.0(<0.1)	10.0 (1.0)	55.7 (0.6)	50.0 (<0.1)	5.7 (0.6)
	Welch T	100.0(<0.1)	10.0 (1.0)	55.7 (0.6)	50.0 (<0.1)	5.7 (0.6)
	M-W U	100.0(<0.1)	10.0 (1.2)	55.8 (0.7)	50.0 (<0.1)	5.8 (0.7)
	AUC	100.0(<0.1)	10.3 (1.0)	55.9 (0.6)	50.0 (<0.1)	5.8 (0.6)
	PAUC-95	67.6 (1.3)	10.4 (1.1)	37.9 (0.9)	33.8 (0.7)	4.0 (0.5)
	Sensitivity	73.8 (1.3)	6.8 (0.9)	39.7 (0.8)	36.9 (0.7)	2.8 (0.4)
	EB T	100.0(<0.1)	10.2 (1.0)	55.8 (0.6)	50.0 (<0.1)	5.8 (0.6)
	K-S	99.7 (0.2)	9.3 (0.9)	55.0 (0.6)	49.9 (0.1)	5.2 (0.6)
200	T-test	100.0(<0.1)	8.9 (1.0)	55.0 (0.6)	50.0 (<0.1)	5.0 (0.6)
	Welch T	100.0(<0.1)	8.9 (1.0)	55.0 (0.6)	50.0 (<0.1)	5.0 (0.6)
	M-W U	100.0(<0.1)	7.8 (1.1)	54.4 (0.6)	50.0 (<0.1)	4.3 (0.6)
	AUC	100.0(<0.1)	8.2 (1.0)	54.6 (0.6)	50.0 (<0.1)	4.6 (0.6)
	PAUC-95	79.4 (1.4)	9.2 (0.9)	43.8 (0.8)	39.7 (0.7)	4.0 (0.4)
	Sensitivity	86.7 (0.9)	7.0 (0.8)	46.6 (0.6)	43.4 (0.4)	3.3 (0.4)
	EB T	100.0(<0.1)	9.5 (0.9)	55.4 (0.6)	50.0 (<0.1)	5.3 (0.6)

	K-S	100.0(<0.1)	8.8 (0.8)	54.9 (0.5)	50.0 (<0.1)	4.9 (0.5)

Means (standard errors) are given for power, false discovery rate (FDR), total number of hits, number of true hits, and number of false hits. N is the equal number of patients and controls. All numbers are based upon twenty simulations at the specified conditions. The Benjamini-Hochberg procedure was used to control the expected FDR at 10%.

Table S2: Normal Samples Heterogeneous Disease Summary Statistics

N	Method	Power	FDR	Total Hits	True Hits	False Hits
25	T-test	0.1 (0.1)	75.0 (7.9)	0.2 (0.1)	0.0 (<0.1)	0.2 (0.1)
	Welch T	0.1 (0.1)	75.0 (7.9)	0.2 (0.1)	0.0 (<0.1)	0.2 (0.1)
	M-W U	0.0 (<0.1)	100.0(<0.1)	0.1 (0.1)	0.0 (<0.1)	0.1 (0.1)
	AUC	0.1 (0.1)	94.4 (3.0)	0.4 (0.2)	0.0 (<0.1)	0.4 (0.1)
	PAUC-95	0.0 (<0.1)	100.0(<0.1)	0.2 (0.1)	0.0 (<0.1)	0.2 (0.1)
	Sensitivity	0.0 (<0.1)	100.0(<0.1)	0.2 (0.1)	0.0 (<0.1)	0.2 (0.1)
	EB T	2.7 (0.6)	17.4 (5.1)	1.6 (0.3)	1.4 (0.3)	0.2 (0.1)
	K-S	0.0 (<0.1)	100.0(<0.1)	0.2 (0.1)	0.0 (<0.1)	0.2 (0.1)
50	T-test	1.3 (0.4)	26.9 (7.4)	0.9 (0.3)	0.6 (0.2)	0.2 (0.1)
	Welch T	1.2 (0.4)	28.7 (7.6)	0.8 (0.3)	0.6 (0.2)	0.2 (0.1)
	M-W U	0.5 (0.2)	33.3 (9.1)	0.4 (0.2)	0.2 (0.1)	0.2 (0.1)
	AUC	0.6 (0.3)	56.7 (10.4)	0.7 (0.3)	0.3 (0.1)	0.4 (0.2)
	PAUC-95	6.6 (1.1)	13.5 (6.0)	3.7 (0.6)	3.3 (0.5)	0.4 (0.2)
	Sensitivity	1.4 (0.5)	24.2 (9.5)	0.9 (0.3)	0.7 (0.2)	0.2 (0.1)
	EB T	14.9 (1.6)	8.4 (2.6)	8.5 (1.0)	7.5 (0.8)	1.1 (0.4)
	K-S	0.4 (0.2)	20.0 (10.0)	0.2 (0.1)	0.2 (0.1)	0.0 (<0.1)
75	T-test	7.7 (1.3)	2.6 (1.5)	4.1 (0.7)	3.8 (0.6)	0.2 (0.2)
	Welch T	7.7 (1.3)	2.6 (1.5)	4.1 (0.7)	3.8 (0.6)	0.2 (0.2)
	M-W U	1.3 (0.4)	8.3 (5.3)	0.8 (0.2)	0.6 (0.2)	0.1 (0.1)
	AUC	2.2 (0.4)	4.4 (3.8)	1.2 (0.2)	1.1 (0.2)	0.1 (0.1)
	PAUC-95	37.7 (2.4)	10.1 (1.6)	21.2 (1.5)	18.9 (1.2)	2.4 (0.4)
	Sensitivity	11.6 (1.7)	10.2 (3.1)	6.5 (1.0)	5.8 (0.9)	0.7 (0.2)
	EB T	33.1 (1.6)	5.8 (1.5)	17.6 (0.9)	16.6 (0.8)	1.1 (0.3)
	K-S	0.2 (0.1)	33.3 (10.5)	0.2 (0.2)	0.1 (0.1)	0.1 (0.1)
100	T-test	17.9 (2.3)	9.0 (2.3)	10.4 (1.5)	8.9 (1.2)	1.5 (0.4)
	Welch T	17.6 (2.3)	9.1 (2.3)	10.3 (1.6)	8.8 (1.2)	1.5 (0.4)
	M-W U	4.4 (1.0)	9.9 (4.8)	2.7 (0.6)	2.2 (0.5)	0.5 (0.3)
	AUC	5.5 (1.0)	11.7 (3.8)	3.5 (0.7)	2.8 (0.5)	0.8 (0.3)
	PAUC-95	61.6 (2.0)	8.5 (1.2)	33.9 (1.3)	30.8 (1.0)	3.0 (0.5)
	Sensitivity	32.6 (2.2)	6.3 (1.2)	17.6 (1.3)	16.3 (1.1)	1.2 (0.3)
	EB T	47.8 (2.0)	10.0 (1.5)	26.9 (1.4)	23.9 (1.0)	3.0 (0.6)
	K-S	1.3 (0.4)	20.0 (7.8)	0.8 (0.2)	0.6 (0.2)	0.2 (0.1)
150	T-test	52.0 (2.8)	9.7 (1.2)	28.9 (1.6)	26.0 (1.4)	2.9 (0.4)
	Welch T	51.5 (2.8)	9.8 (1.2)	28.6 (1.6)	25.8 (1.4)	2.9 (0.4)
	M-W U	18.4 (1.9)	8.7 (2.2)	10.3 (1.1)	9.2 (1.0)	1.1 (0.3)
	AUC	17.4 (1.8)	11.1 (2.5)	9.9 (1.1)	8.7 (0.9)	1.2 (0.3)
	PAUC-95	93.1 (0.7)	9.4 (0.9)	51.5 (0.6)	46.5 (0.4)	4.9 (0.5)
	Sensitivity	77.1 (1.8)	8.7 (0.8)	42.3 (1.1)	38.5 (0.9)	3.8 (0.4)
	EB T	77.5 (1.7)	10.9 (1.0)	43.6 (1.2)	38.8 (0.8)	4.9 (0.5)
	K-S	8.9 (1.6)	13.1 (3.9)	5.2 (0.9)	4.5 (0.8)	0.7 (0.3)
200	T-test	77.5 (1.6)	11.4 (1.0)	43.9 (1.1)	38.8 (0.8)	5.1 (0.5)
	Welch T	77.4 (1.6)	11.4 (1.0)	43.8 (1.1)	38.7 (0.8)	5.1 (0.5)
	M-W U	36.7 (2.6)	10.8 (1.7)	20.8 (1.5)	18.4 (1.3)	2.4 (0.4)
	AUC	35.3 (2.6)	11.6 (1.8)	20.1 (1.5)	17.6 (1.3)	2.4 (0.4)
	PAUC-95	98.9 (0.3)	9.0 (1.0)	54.5 (0.7)	49.5 (0.2)	5.0 (0.6)
	Sensitivity	93.3 (1.0)	6.3 (0.6)	49.8 (0.5)	46.6 (0.5)	3.1 (0.3)
	EB T	89.9 (1.1)	10.9 (0.9)	50.5 (0.8)	45.0 (0.5)	5.6 (0.5)
	K-S	30.4 (2.7)	9.5 (1.6)	16.9 (1.5)	15.2 (1.3)	1.8 (0.3)

Means (standard errors) are given for power, false discovery rate (FDR), total number of hits, number of true hits, and number of false hits. N is the equal number of patients and controls. All numbers are based upon twenty simulations at the specified conditions. The Benjamini-Hochberg procedure was used to control the expected FDR at 10%.

Table S3: T_3 Homogeneous Disease Summary Statistics

N	Method	Power	FDR	Total Hits	True Hits	False Hits
25	T-test	33.0 (2.3)	5.4 (1.4)	17.4 (1.1)	16.5 (1.1)	0.9 (0.2)
	Welch T	31.8 (1.9)	4.5 (1.5)	16.6 (0.9)	15.9 (1.0)	0.7 (0.2)
	M-W U	50.8 (2.8)	5.3 (1.0)	26.8 (1.4)	25.4 (1.4)	1.4 (0.3)
	AUC	55.2 (2.3)	8.4 (1.1)	30.1 (1.1)	27.6 (1.2)	2.5 (0.3)
	PAUC-95	5.0 (1.0)	5.6 (3.5)	2.7 (0.5)	2.5 (0.5)	0.2 (0.1)
	Sensitivity	11.8 (2.1)	11.9 (4.4)	6.7 (1.1)	5.9 (1.0)	0.8 (0.3)
	EB T	33.4 (2.4)	4.3 (1.5)	17.5 (1.3)	16.7 (1.2)	0.8 (0.3)
	K-S	38.4 (1.8)	5.3 (1.4)	20.3 (1.0)	19.2 (0.9)	1.1 (0.3)
50	T-test	77.8 (1.7)	8.1 (1.3)	42.4 (1.1)	38.9 (0.8)	3.5 (0.6)
	Welch T	77.8 (1.7)	8.1 (1.3)	42.4 (1.1)	38.9 (0.8)	3.5 (0.6)
	M-W U	96.4 (0.9)	9.4 (1.2)	53.3 (0.8)	48.2 (0.4)	5.1 (0.7)
	AUC	96.4 (1.0)	9.6 (1.3)	53.4 (0.7)	48.2 (0.5)	5.2 (0.8)
	PAUC-95	14.4 (1.9)	8.1 (3.5)	7.9 (1.0)	7.2 (1.0)	0.7 (0.3)
	Sensitivity	29.0 (3.1)	4.6 (1.9)	15.2 (1.7)	14.5 (1.6)	0.7 (0.3)
	EB T	79.2 (1.7)	6.7 (1.5)	42.5 (0.9)	39.6 (0.8)	2.9 (0.7)
	K-S	92.6 (1.0)	5.4 (1.2)	49.0 (0.8)	46.3 (0.5)	2.7 (0.7)
75	T-test	93.6 (1.0)	5.0 (1.0)	49.3 (0.7)	46.8 (0.5)	2.5 (0.5)
	Welch T	93.6 (1.0)	4.6 (1.1)	49.1 (0.7)	46.8 (0.5)	2.3 (0.6)
	M-W U	99.6 (0.3)	8.9 (1.0)	54.7 (0.6)	49.8 (0.1)	4.9 (0.6)
	AUC	99.8 (0.2)	8.8 (1.1)	54.8 (0.6)	49.9 (0.1)	4.9 (0.7)
	PAUC-95	15.6 (1.6)	12.2 (2.0)	8.8 (0.8)	7.8 (0.8)	1.0 (0.1)
	Sensitivity	39.6 (1.3)	6.5 (1.0)	21.2 (0.7)	19.8 (0.7)	1.4 (0.2)
	EB T	94.6 (0.8)	4.7 (1.0)	49.7 (0.6)	47.3 (0.4)	2.4 (0.5)
	K-S	100.0(<0.1)	8.5 (1.2)	54.7 (0.7)	50.0 (<0.1)	4.7 (0.7)
100	T-test	97.8 (0.9)	7.3 (0.8)	52.8 (0.6)	48.9 (0.4)	3.9 (0.5)
	Welch T	97.8 (0.9)	7.2 (0.9)	52.7 (0.6)	48.9 (0.4)	3.8 (0.5)
	M-W U	100.0(<0.1)	11.7 (1.1)	56.7 (0.7)	50.0 (<0.1)	6.7 (0.7)
	AUC	100.0(<0.1)	12.4 (1.4)	57.2 (0.9)	50.0 (<0.1)	7.2 (0.9)
	PAUC-95	19.2 (3.0)	9.0 (3.8)	10.6 (1.6)	9.6 (1.5)	1.0 (0.4)
	Sensitivity	47.2 (3.3)	4.9 (1.2)	24.8 (1.7)	23.6 (1.6)	1.2 (0.3)
	EB T	97.8 (0.9)	7.5 (1.0)	52.9 (0.7)	48.9 (0.4)	4.0 (0.6)
	K-S	100.0(<0.1)	9.0 (0.7)	55.0 (0.4)	50.0 (<0.1)	5.0 (0.4)
150	T-test	99.8 (0.2)	8.4 (1.0)	54.5 (0.6)	49.9 (0.1)	4.6 (0.6)
	Welch T	99.8 (0.2)	8.4 (1.0)	54.5 (0.6)	49.9 (0.1)	4.6 (0.6)
	M-W U	100.0(<0.1)	9.8 (1.2)	55.5 (0.8)	50.0 (<0.1)	5.5 (0.8)
	AUC	100.0(<0.1)	9.6 (1.3)	55.4 (0.8)	50.0 (<0.1)	5.4 (0.8)
	PAUC-95	37.4 (2.8)	9.1 (1.8)	20.6 (1.5)	18.7 (1.4)	1.9 (0.4)
	Sensitivity	66.8 (2.4)	6.0 (1.0)	35.5 (1.1)	33.4 (1.2)	2.1 (0.3)
	EB T	100.0(<0.1)	7.6 (1.0)	54.2 (0.6)	50.0 (<0.1)	4.2 (0.6)
	K-S	100.0(<0.1)	6.7 (0.7)	53.6 (0.4)	50.0 (<0.1)	3.6 (0.4)
200	T-test	99.6 (0.3)	10.6 (1.0)	55.8 (0.6)	49.8 (0.1)	6.0 (0.6)
	Welch T	99.6 (0.3)	10.5 (1.1)	55.7 (0.7)	49.8 (0.1)	5.9 (0.7)
	M-W U	100.0(<0.1)	11.2 (1.2)	56.4 (0.8)	50.0 (<0.1)	6.4 (0.8)
	AUC	100.0(<0.1)	12.5 (1.5)	57.3 (1.0)	50.0 (<0.1)	7.3 (1.0)
	PAUC-95	48.2 (2.8)	9.5 (2.6)	27.0 (2.0)	24.1 (1.4)	2.9 (1.0)

	Sensitivity	79.4 (1.7)	6.9 (1.5)	42.8 (1.3)	39.7 (0.8)	3.1 (0.8)
	EB T	99.6 (0.3)	10.6 (1.2)	55.8 (0.8)	49.8 (0.1)	6.0 (0.8)
	K-S	100.0(<0.1)	8.7 (1.0)	54.8 (0.6)	50.0 (<0.1)	4.8 (0.6)

Means (standard errors) are given for power, false discovery rate (FDR), total number of hits, number of true hits, and number of false hits. N is the equal number of patients and controls. All numbers are based upon ten simulations at the specified conditions. The Benjamini-Hochberg procedure was used to control the expected FDR at 10%.

Table S4: T_3 Heterogeneous Disease Summary Statistics

N	Method	Power	FDR	Total Hits	True Hits	False Hits
25	T-test	0.0 (<0.1)	0/0	0.0 (<0.1)	0.0 (<0.1)	0.0 (<0.1)
	Welch T	0.0 (<0.1)	0/0	0.0 (<0.1)	0.0 (<0.1)	0.0 (<0.1)
	M-W U	0.0 (<0.1)	0/0	0.0 (<0.1)	0.0 (<0.1)	0.0 (<0.1)
	AUC	0.0 (<0.1)	100.0(<0.1)	0.1 (0.1)	0.0 (<0.1)	0.1 (0.1)
	PAUC-95	0.6 (0.4)	61.1 (11.0)	0.6 (0.3)	0.3 (0.2)	0.3 (0.2)
	Sensitivity	0.0 (<0.1)	0/0	0.0 (<0.1)	0.0 (<0.1)	0.0 (<0.1)
	EB T	0.2 (0.2)	0.0 (<0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (<0.1)
	K-S	0.0 (<0.1)	100.0(<0.1)	0.1 (0.1)	0.0 (<0.1)	0.1 (0.1)
50	T-test	0.4 (0.3)	33.3 (18.3)	0.3 (0.2)	0.2 (0.1)	0.1 (0.1)
	Welch T	0.4 (0.3)	33.3 (18.3)	0.3 (0.2)	0.2 (0.1)	0.1 (0.1)
	M-W U	0.0 (<0.1)	0/0	0.0 (<0.1)	0.0 (<0.1)	0.0 (<0.1)
	AUC	0.4 (0.3)	50.0 (18.3)	0.4 (0.2)	0.2 (0.1)	0.2 (0.1)
	PAUC-95	5.0 (1.5)	17.7 (10.2)	3.0 (0.9)	2.5 (0.7)	0.5 (0.2)
	Sensitivity	2.2 (0.9)	22.2 (12.8)	1.3 (0.5)	1.1 (0.5)	0.2 (0.1)
	EB T	0.8 (0.4)	25.0 (15.8)	0.5 (0.2)	0.4 (0.2)	0.1 (0.1)
	K-S	0.2 (0.2)	0.0 (<0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (<0.1)
75	T-test	1.4 (0.5)	10.0 (7.1)	0.8 (0.3)	0.7 (0.3)	0.1 (0.1)
	Welch T	1.2 (0.4)	10.0 (7.1)	0.7 (0.3)	0.6 (0.2)	0.1 (0.1)
	M-W U	0.8 (0.6)	12.5 (5.6)	0.5 (0.4)	0.4 (0.3)	0.1 (0.1)
	AUC	1.0 (0.4)	25.0 (10.1)	0.8 (0.4)	0.5 (0.2)	0.3 (0.2)
	PAUC-95	15.0 (2.8)	8.3 (3.6)	8.0 (1.4)	7.5 (1.4)	0.5 (0.2)
	Sensitivity	10.0 (1.9)	3.4 (2.3)	5.2 (1.0)	5.0 (1.0)	0.2 (0.1)
	EB T	1.4 (0.5)	10.0 (7.1)	0.8 (0.3)	0.7 (0.3)	0.1 (0.1)
	K-S	0.8 (0.4)	38.9 (11.0)	0.8 (0.5)	0.4 (0.2)	0.4 (0.3)
100	T-test	6.0 (1.3)	7.3 (4.3)	3.4 (0.8)	3.0 (0.6)	0.4 (0.3)
	Welch T	5.2 (1.2)	6.7 (3.9)	2.9 (0.7)	2.6 (0.6)	0.3 (0.2)
	M-W U	2.8 (0.7)	3.1 (2.8)	1.5 (0.4)	1.4 (0.3)	0.1 (0.1)
	AUC	3.4 (0.8)	4.2 (3.7)	1.8 (0.4)	1.7 (0.4)	0.1 (0.1)
	PAUC-95	27.2 (3.7)	11.2 (3.5)	15.4 (2.0)	13.6 (1.9)	1.8 (0.5)
	Sensitivity	35.2 (2.9)	7.1 (1.9)	19.1 (1.7)	17.6 (1.5)	1.5 (0.4)
	EB T	7.2 (1.6)	6.2 (3.9)	4.0 (0.9)	3.6 (0.8)	0.4 (0.3)
	K-S	1.8 (0.6)	0.0 (<0.1)	0.9 (0.3)	0.9 (0.3)	0.0 (<0.1)
150	T-test	26.6 (2.5)	2.4 (1.2)	13.6 (1.2)	13.3 (1.3)	0.3 (0.2)
	Welch T	26.4 (2.5)	2.4 (1.2)	13.5 (1.2)	13.2 (1.3)	0.3 (0.2)
	M-W U	13.6 (2.0)	9.6 (4.6)	7.6 (1.1)	6.8 (1.0)	0.8 (0.4)
	AUC	12.2 (1.8)	13.1 (6.9)	6.9 (1.0)	6.1 (0.9)	0.8 (0.4)
	PAUC-95	61.6 (3.1)	7.4 (1.0)	33.3 (1.7)	30.8 (1.5)	2.5 (0.4)
	Sensitivity	70.2 (2.3)	4.2 (0.5)	36.6 (1.1)	35.1 (1.1)	1.5 (0.2)
	EB T	28.2 (2.3)	2.3 (1.2)	14.4 (1.1)	14.1 (1.1)	0.3 (0.2)
	K-S	8.8 (2.6)	9.0 (3.5)	5.0 (1.5)	4.4 (1.3)	0.6 (0.3)
200	T-test	50.0 (2.8)	9.4 (1.7)	27.7 (1.7)	25.0 (1.4)	2.7 (0.6)
	Welch T	50.0 (2.8)	9.4 (1.7)	27.7 (1.7)	25.0 (1.4)	2.7 (0.6)

	M-W U	31.2 (2.6)	6.1 (2.5)	16.9 (1.7)	15.6 (1.3)	1.3 (0.6)
	AUC	30.2 (2.3)	9.6 (2.8)	17.0 (1.6)	15.1 (1.2)	1.9 (0.7)
	PAUC-95	74.6 (1.5)	9.7 (1.9)	41.5 (1.3)	37.3 (0.8)	4.2 (0.9)
	Sensitivity	89.6 (1.0)	6.2 (1.5)	47.9 (1.1)	44.8 (0.5)	3.1 (0.8)
	EB T	51.4 (2.6)	9.3 (1.6)	28.4 (1.5)	25.7 (1.3)	2.7 (0.6)
	K-S	32.8 (3.6)	6.9 (2.3)	17.9 (2.2)	16.4 (1.8)	1.5 (0.6)

Means (standard errors) are given for power, false discovery rate (FDR), total number of hits, number of true hits, and number of false hits. N is the equal number of patients and controls. All numbers are based upon ten simulations at the specified conditions. The Benjamini-Hochberg procedure was used to control the expected FDR at 10%.

Table S5: Summary Statistics for Significant Hits in Breast Cancer Screening Study

		AUC			Sensitivity at 95% Specificity		
	n	Q1	Median	Q3	Q1	Median	Q3
T-test	77	0.614	0.632	0.647	10.9%	14.9%	20.6%
Welch t-test	78	0.612	0.632	0.647	11.8%	15.7%	20.6%
M-W U	72	0.628	0.637	0.648	10.9%	15.7%	20.8%
AUC	76	0.627	0.636	0.648	10.9%	15.7%	20.6%
PAUC	40	0.584	0.603	0.639	18.3%	20.7%	23.3%
Sensitivity	45	0.578	0.604	0.634	20.6%	22.5%	24.0%
EB t-test	76	0.614	0.632	0.648	11.5%	15.3%	20.6%
K-S	38	0.63	0.64	0.66	10.8%	13.7%	17.6%

First quartile, median, and third quartile of AUC and sensitivity at 95% specificity for significant finds at 5% FDR control using each of the 8 selection methods. The Benjamini-Hochberg procedure was used to control the expected FDR.

Table S6: Highly Significant Hits in Breast Cancer Screening Study

	T-test	Welch t-test	M-W U	AUC	PAUC	Sensitivity	EB t-test	K-S
ARF1	**0.0001**	**0.0004**	**0.0003**	**0.0000**	0.1192	0.0744	**0.0001**	**0.0085**
ATP6AP1	0.3400	0.3049	0.1072	0.1299	**0.0000**	0.0171	0.3403	0.1310
BMX	0.4303	0.4163	0.4943	0.5416	**0.0085**	0.0359	0.4306	0.2336
CSNK1E	0.0174	0.0154	0.0733	0.1098	**0.0085**	0.0212	0.0174	0.0568
CTBP1	0.2479	0.2401	0.0928	0.0696	**0.0085**	0.0744	0.2464	0.0894
CYR61	**0.0074**	0.0266	0.0338	0.0420	0.1479	0.2003	0.0071	0.0568
DAPK2	**0.0067**	**0.0077**	0.0166	0.0156	0.0151	0.0181	**0.0068**	0.0504
DBT	**0.0074**	0.0121	0.0202	0.0179	**0.0000**	0.0171	**0.0073**	0.0985
EIF3E	0.0283	0.0318	**0.0092**	**0.0087**	**0.0000**	0.0171	0.0284	0.0349
GJA1	0.0164	0.0137	**0.0092**	**0.0087**	0.0257	0.0209	0.0166	0.0526
GRAP2	**0.0099**	0.0225	0.0273	0.0256	0.2942	0.1306	**0.0099**	0.1322
HOMER2	**0.0099**	0.0137	0.0166	0.0143	0.3342	0.0542	0.0106	0.0349
HOOK1	**0.0035**	**0.0067**	0.0166	0.0182	0.0257	0.0209	**0.0033**	0.0504
HSD17B3	**0.0067**	**0.0087**	**0.0046**	**0.0051**	0.5876	0.7840	**0.0068**	**0.0061**
IFRD2	**0.0001**	**0.0009**	0.0128	**0.0095**	0.5721	0.1940	**0.0001**	0.0341
IGSF11	**0.0067**	**0.0067**	0.0103	0.0143	0.0895	0.0666	**0.0068**	0.0249
ITGB1BP1	**0.0035**	**0.0077**	0.0733	0.0920	0.1147	0.3628	**0.0033**	0.0568
MESP1	**0.0052**	**0.0087**	0.0247	0.0233	0.1782	0.3075	**0.0049**	0.0349
MYCBP	**0.0075**	0.0107	0.0202	0.0182	0.1192	0.1063	**0.0074**	0.0533
NFKB1	**0.0067**	**0.0077**	0.0103	0.0135	0.3258	0.5290	**0.0068**	0.0249
OPTN	**0.0067**	0.0207	0.0245	0.0312	0.2298	0.4878	**0.0068**	0.0341
PAGE1	**0.0067**	0.0109	**0.0092**	0.0106	0.5007	0.5420	**0.0068**	0.0143
PAX8	**0.0067**	0.0107	0.0253	0.0366	0.4132	0.3309	**0.0068**	0.0743
PDCD6IP	0.0521	0.0400	0.0665	0.0920	**0.0030**	0.0300	0.0530	0.1766
PITX1	**0.0067**	**0.0077**	**0.0099**	**0.0087**	0.0367	0.0437	**0.0068**	0.0349
RAB8A	**0.0052**	0.0154	0.1108	0.1822	0.1626	0.4352	**0.0049**	0.1322
RAC3	0.1416	0.1155	0.1846	0.2504	**0.0051**	0.0212	0.1441	0.1228
RNF24	**0.0067**	**0.0077**	0.0166	0.0155	0.0498	0.0209	**0.0068**	0.0625
RPL14	**0.0067**	**0.0077**	**0.0092**	**0.0087**	0.4132	0.2907	**0.0068**	0.0401
RPS26	**0.0067**	**0.0099**	0.0202	0.0190	0.0238	0.0181	**0.0069**	0.0767
SERPINH1	**0.0099**	0.0141	0.0247	0.0190	0.0329	0.0942	0.0109	0.0349
SF3A1	0.4150	0.3759	0.0756	0.0650	**0.0000**	0.0171	0.4116	0.0349
SLC35B1	**0.0067**	**0.0087**	**0.0092**	0.0106	0.2950	0.2964	**0.0068**	0.0311
TAF9B	**0.0052**	**0.0067**	**0.0099**	**0.0051**	0.0510	0.0519	**0.0049**	0.0349
THRA	**0.0004**	**0.0067**	0.0175	0.0150	0.9550	0.7840	**0.0003**	0.0354
TMED1	**0.0067**	**0.0087**	0.0166	0.0182	0.1889	0.2964	**0.0068**	0.0349
UBAP1	0.4281	0.3814	0.4260	0.4622	**0.0097**	0.0432	0.4272	0.2238

Q-values for proteins that were found to be significant with 1% FDR control by any of the eight methods using the breast cancer screening data. The Benjamini-Hochberg procedure was used to control the expected FDR.

Figure Legends

Figure S1. Estimated power for the first stage of a two-stage design with normal samples for a (a) homogeneous disease and (b) heterogeneous disease. The horizontal axis indicates the number of patients and controls in the first stage. The vertical axis indicates power, the proportion of the true biomarkers that are selected by each method. Standard error bars are shown.

Figure S2. Estimated power for single stage design using t_3 samples for a (a) homogeneous disease and a (b) heterogeneous disease. The horizontal axis indicates N, the equal number of cases and controls. The vertical axis indicates power, the proportion of the true biomarkers that are selected by each method. Standard error bars are shown.

Figure S3. Estimated power with FDR controlled at 50% for single stage design using normal samples for a (a) homogeneous disease and a (b) heterogeneous disease. The horizontal axis indicates N, the equal number of cases and controls. The vertical axis indicates power, the proportion of the true biomarkers that are selected by each method. Standard error bars are shown.

Figure S4. Estimated power with FDR controlled at 50% for single stage design using t_3 samples for a (a) homogeneous disease and a (b) heterogeneous disease. The horizontal axis indicates N, the equal number of cases and controls. The vertical axis indicates power, the proportion of the true biomarkers that are selected by each method. Standard error bars are shown.

Figure S5. Estimated power with FDR controlled at 50% for a two-stage design using 100 total patients and 100 total controls with normal samples for a (a) homogeneous disease and a (b) heterogeneous disease. The horizontal axis indicates the allocation of the total number of patients and controls across the two stages. The vertical axis indicates power, the proportion of

the true biomarkers that are selected by each method. The dashed horizontal line indicates the estimated power of the best method in a single-stage design with 50% FDR control and using the same number of total patients and controls. Standard error bars are shown.

Figure S6. Estimated power with FDR controlled at 50% for a two-stage design using 200 total patients and 200 total controls with normal samples for a (a) homogeneous disease and a (b) heterogeneous disease. The horizontal axis indicates the allocation of the total number of patients and controls across the two stages. The vertical axis indicates power, the proportion of the true biomarkers that are selected by each method. The dashed horizontal line indicates the estimated power of the best method in a single-stage design with 50% FDR control and using the same number of total patients and controls. Standard error bars are shown.

Figure S7. Estimated power of the ordinary t-test and empirical Bayes moderated t-test with FDR controlled at 10% for single-stage design using 25, 50, 100, or 200 patients and controls with normal samples for a homogeneous disease. The horizontal axis indicates the standard deviation of the scaled inverted gamma distribution used to generate the variances of the biomarker candidates.

Figure S8. Estimated power of the ordinary t-test and empirical Bayes moderated t-test with FDR controlled at 10% for single-stage design using 25, 50, 100, or 200 patients and controls with normal samples for a heterogeneous disease. The horizontal axis indicates the standard deviation of the scaled inverted gamma distribution used to generate the variances of the biomarker candidates.