# Supplementary Materials for "Set-Based Tests for Genetic Association Using the Generalized Berk-Jones Statistic"

November 4, 2018

**Supplement A:** Proof of Theorem 1 from Section 3.3.

**Supplement B:** Exact p-value calculation using Equation (5) from Section 3.4.

**Supplement C:** Accuracy of p-value calculation from Section 3.4.

**Supplement D:** Rejection region plots on p-value scale from Section 4.

**Supplement E:** Complete simulation parameters from Section 5.

**Supplement F:** Power benchmarks in structured gene simulations from Section 5.

**Supplement G:** Supplemental power simulations from Section 5.

**Supplement H:** Empirical QQ-plots of CGEMS analysis from Section 6.

**Supplement I:** Evaluation of summary statistic correlation approximation in CGEMS analysis from Section 6.

# Supplement A: Proof of Theorem 1 from Section 3.3

We are interested in the variance of

$$\begin{aligned} S(t) &= \sum_{k=1}^{d} \mathbf{1}\left(|Z_k| \geq t\right), \\ \mathbf{Z} &\sim MVN\left(\mu \cdot \mathbf{J}_d, \mathbf{\Sigma}\right), \end{aligned}$$

where the diagonal elements of $\mathbf{\Sigma}$ are all 1. The variance can be decomposed as

$$\begin{aligned} \mathrm{Var}\left\{S(t)\right\} &= d\lambda(1-\lambda) + 2\sum_{1\leq k<l\leq d}\left\{\Pr\left(|Z_k|, |Z_l| \geq t\right) - \lambda^2\right\}, \\ \lambda &= 1 - \left\{\Phi\left(t-\mu\right) - \Phi(-t-\mu)\right\}. \end{aligned}$$

The summation can be written as

$$\begin{aligned} 2\sum_{1\leq k<l\leq d}\left\{\Pr\left(|Z_k|, |Z_l| \geq t\right) - \lambda^2\right\} &= 2\sum_{1\leq k<l\leq d}\Pr\left(Z_k, Z_l \geq t\right) \\ &\quad +2\sum_{1\leq k<l\leq d}\Pr\left(Z_k, Z_l \leq -t\right) \\ &\quad +2\sum_{1\leq k<l\leq d}\Pr\left(Z_k \geq t, Z_l \leq -t\right) \\ &\quad +2\sum_{1\leq k<l\leq d}\Pr\left(Z_k \leq -t, Z_l \geq t\right) \\ &\quad -d(d-1)\lambda^2. \end{aligned}$$

Each of the four probabilities above can be reexpressed using the standard Mehler kernel for the bivariate normal distribution. For example, the first probability is:

$$\begin{aligned} & 2\sum_{1\leq k<l\leq d}\Pr\left(Z_k, Z_l \geq t\right) \\ = \; & 2\sum_{1\leq k<l\leq d}\int_t^\infty\int_t^\infty \frac{1}{2\pi\sqrt{1-\rho_{k,l}^2}}\exp\left[-\frac{1}{2(1-\rho_{k,l}^2)}\left\{(z_k-\mu)^2 - 2\rho_{k,l}(z_k-\mu)(z_l-\mu) + (z_l-\mu)^2\right\}\right]dz_k dz_l, \\ = \; & 2\sum_{1\leq k<l\leq d}\int_t^\infty\int_t^\infty \phi(z_k-\mu)\phi(z_l-\mu)\sum_{r=0}^{\infty}\frac{\rho_{k,l}^r}{r!}H_r(z_k-\mu)H_r(z_l-\mu)dz_k dz_l, \\ = \; & 2\sum_{1\leq k<l\leq d}\left\{\bar{\Phi}(t-\mu)^2 + \int_t^\infty\int_t^\infty \phi(z_k-\mu)\phi(z_l-\mu)\sum_{r=1}^{\infty}\frac{\rho_{k,l}^r}{r!}H_r(z_k-\mu)H_r(z_l-\mu)dz_k dz_l\right\}, \\ = \; & 2\sum_{1\leq k<l\leq d}\left\{\bar{\Phi}(t-\mu)^2 + \phi(t-\mu)^2\sum_{r=1}^{\infty}\frac{\rho_{k,l}^r}{r!}H_{r-1}(t-\mu)^2\right\}, \end{aligned}$$

$$
\begin{aligned}
&= d(d-1)\bar{\Phi}(t-\mu)^2 + 2\phi(t-\mu)^2 \sum_{r=1}^{\infty} \frac{1}{r!} H_{r-1}(t-\mu)^2 \left( \sum_{1\le k<l\le d} \rho_{k,l}^r \right), \\
&= d(d-1)\bar{\Phi}(t-\mu)^2 + 2\phi(t-\mu)^2 \sum_{r=1}^{\infty} \frac{1}{r!} H_{r-1}(t-\mu)^2 \left( \frac{d(d-1)}{2} \bar{\rho^r} \right), \\
&= d(d-1) \left\{ \bar{\Phi}(t-\mu)^2 + \phi(t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(t-\mu)^2 \right\}, \\
\bar{\rho^r} &= \frac{2}{d(d-1)} \sum_{1\le k<l\le d} \rho_{k,l}^r .
\end{aligned}
$$

Here $\phi(x)$ represents the density function of $N(0,1)$ distribution and $\rho_{k,l}$ is the $(k,l)$ element of $\mathbf{\Sigma}$. We skip the similar derivation for the other three probabilities and give only the final expressions:

$$
\begin{aligned}
2 \sum_{1\le k<l\le d} \Pr\left(Z_k, Z_l \le -t\right) &= d(d-1) \left\{ \Phi(-t-\mu)^2 + \phi(-t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu)^2 \right\}, \\
2 \sum_{1\le k<l\le d} \Pr\left(Z_k \le -t, Z_l \ge t\right) &= d(d-1) \left\{ \Phi(-t-\mu)\bar{\Phi}(t-\mu) \right\} \\
&\quad -d(d-1) \left\{ \phi(-t-\mu)\phi(t-\mu) \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu) H_{r-1}(t-\mu) \right\}, \\
&= 2 \sum_{1\le k<l\le d} \Pr\left(Z_k \ge t, Z_l \le -t\right).
\end{aligned}
$$

So in total we have

$$
\begin{aligned}
2 \sum_{1\le k<l\le d} \left\{ \Pr\left(|Z_k|, |Z_l| \ge t\right) - \lambda^2 \right\} &= d(d-1) \left\{ \bar{\Phi}(t-\mu)^2 + \phi(t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(t-\mu)^2 \right\} \\
&\quad +d(d-1) \left\{ \Phi(-t-\mu)^2 + \phi(-t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu)^2 \right\} \\
&\quad +2d(d-1) \left\{ \Phi(-t-\mu)\bar{\Phi}(t-\mu) \right\} \\
&\quad -2d(d-1) \left\{ \phi(-t-\mu)\phi(t-\mu) \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu) H_{r-1}(t-\mu) \right\} \\
&\quad -d(d-1) \left\{ \bar{\Phi}\left(t-\mu\right) + \Phi(-t-\mu) \right\}^2, \\
&= d(d-1) \left\{ \phi(t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(t-\mu)^2 \right\} \\
&\quad +d(d-1) \left\{ \phi(-t-\mu)^2 \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu)^2 \right\} \\
&\quad -2d(d-1) \left\{ \phi(-t-\mu)\phi(t-\mu) \sum_{r=1}^{\infty} \frac{\bar{\rho^r}}{r!} H_{r-1}(-t-\mu) H_{r-1}(t-\mu) \right\}.
\end{aligned}
$$

Put it all back together for the result given in the theorem.

# Supplement B: Exact p-value calculation using equation (5) from Section 3.4.

We are interested in calculating the probability

$$\Pr\left(G_d \geq g\right) \;=\; 1 - \Pr\left\{\forall j = 1,2,...,d : |Z|_{(j)} \leq b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}.$$

Using the law of total probability, our quantity of interest is

$$\begin{aligned} & \Pr\left\{\forall j = 1,2,...,d : |Z|_{(j)} \leq b_j \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\} \\ = \; & \sum_{\mathbf{a} \in \mathcal{A}} \Pr\left\{\forall j = 1,2,...,d : |Z|_{(j)} \leq b_j, |Z|_{(j)} = |Z_{a_j}| \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}, \end{aligned}$$

where $\mathbf{a} = (a_1, a_2, ..., a_d)$ and $\mathcal{A}$ is the set of all $d!$ possible permutations of the integers from 1 to $d$. Thus the p-value can be expressed as

$$\begin{aligned} & \Pr\left(G_d \geq g\right) \\ = \; & 1 - \sum_{\mathbf{a} \in \mathcal{A}} \Pr\left\{0 \leq |Z_{a_1}| \leq b_1, |Z_{a_1}| \leq |Z_{a_2}| \leq b_2, ...., |Z_{a_{d-1}}| \leq |Z_{a_d}| \leq b_d \middle| \mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})\right\}. \end{aligned}$$

At this point it is apparent that we will need some sort of distribution function for $\mathbf{Y} = (Y_1, Y_2, ..., Y_d) = (|Z_1|, |Z_2|, ..., |Z_d|)$, where $\mathbf{Y}$ is the result of applying the absolute value operator on every element of $\mathbf{Z}$. $\mathbf{Y}$ is also known as the multivariate half-normal distribution. If $\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, then the probability density function of $\mathbf{Y}$ can be written as

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \sum_{\mathbf{s} \in S} (2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}_s|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^T \mathbf{\Sigma}_s^{-1} \mathbf{y}\right\}, \qquad \text{(S.1)} \\ S &= \{(\delta_1, ..., \delta_d) : \delta_j = \pm 1 \forall j = 1,2,...,d\}, \\ \mathbf{\Lambda_s} &= \{\text{diag}\,(\mathbf{s})\}, \\ \mathbf{\Sigma_s} &= \mathbf{\Lambda_s \Sigma \Lambda_s}. \end{aligned}$$

Note that there are $2^d$ elements in $S$. With the use of (S.1), the p-value can be expressed as a $d$-dimensional integral:

$$\Pr\left(G \geq g\right) = 1 - \sum_{\mathbf{a} \in \mathcal{A}} \sum_{\mathbf{s} \in S} \int_0^{b_1} \int_{Y_1}^{b_2} ... \int_{Y_{d-1}}^{b_d} (2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^T \left(\mathbf{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}\right)^{-1} \mathbf{y}\right\} dY_d ... dY_1. \qquad \text{(S.2)}$$

By the use of $\boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}$ we mean the variance matrix that is permuted to account for the ordering $\mathbf{a}$. It can be defined as:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})} &= \boldsymbol{\Lambda}_{\mathbf{s}}\mathbf{P}^{(\mathbf{a})}\boldsymbol{\Sigma}\mathbf{P}^{(\mathbf{a})T}\boldsymbol{\Lambda}_{\mathbf{s}}, \\ \mathbf{P}^{(\mathbf{a})} &= \begin{pmatrix} \mathbf{e}_{a_1}^T \\ \mathbf{e}_{a_2}^T \\ . \\ . \\ \mathbf{e}_{a_d}^T \end{pmatrix},\end{aligned}$$

where $\mathbf{e}_j$ denotes the $d \times 1$ vector with a 1 in the $j$th position and 0 everywhere else. Although equation (S.2) appears to be calculable through many calls to a multivariate normal distribution solver, the lower bounds are functions of variables in the integration, which is not a feature supported by many statistical computing packages. To put the expression into a form more accessible for computation, we can reinterpret the $d$-dimensional integral:

$$\begin{aligned}&\int_0^{b_1}\int_{Y_1}^{b_2}...\int_{Y_{d-1}}^{b_d}(2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(\boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}\right)^{-1}\mathbf{y}\right\}dY_d...dY_1 \\ = \;& \Pr\left\{0 \le Y_1 \le b_1, Y_1 \le Y_2 \le b_2, ..., Y_{d-1} \le Y_d \le b_d \middle| \mathbf{Y} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}\right)\right\}, \\ = \;& \Pr\left\{0 \le Y_1 \le b_1, Y_2 \le b_2, ..., Y_d \le b_d, Y_2 - Y_1 \ge 0, ..., Y_d - Y_{d-1} \ge 0 \middle| \mathbf{Y} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}\right)\right\}. \quad \text{(S.3)}\end{aligned}$$

To be clear, equation (S.3) is meant to show how the integral of equation (S.2) can be viewed as a simpler probability if we reinterpret $\mathbf{Y}$ as possessing a multivariate normal distribution instead of its true multivariate half-normal distribution. Equation (S.3) is simpler because the bounds are all constants, which is a form more amenable to most statistical software. Our final step to simplify the quantity for computation is to introduce the vector $\mathbf{T} = (T_1, T_2, ..., T_{2d-1})$ so that

$$\begin{aligned}&\Pr\left\{0 \le Y_1 \le b_1, Y_2 \le b_2, ..., Y_d \le b_d, Y_2 - Y_1 \ge 0, ..., Y_d - Y_{d-1} \ge 0 \middle| \mathbf{Y} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_s^{(a)})\right\} \\ = \;& \Pr\left\{0 \le T_1 \le b_1, T_2 \le b_2, ..., T_d \le b_d, T_{d+1} \ge 0, ..., T_{2d-1} \ge 0 \middle| \mathbf{T} \sim MVN\left(\mathbf{0}_{(2d-1)\times 1}, \boldsymbol{\Delta}_d\boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})}\boldsymbol{\Delta}_d^T\right)\right\}, \\ \boldsymbol{\Delta}_d \;=\; & \begin{pmatrix} \mathbf{I}_{d\times d} \\ \mathbf{D} \end{pmatrix}_{(2d-1)\times d},\end{aligned}$$

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & . & . & \\ & & . & . & \\ & & & -1 & 1 \end{pmatrix}_{(d-1)\times d}.$$

The final p-value is given by

$$\begin{aligned} \Pr\left(G_d \geq g\right) &= 1 - \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{s}\in S} \Pr\left(\mathbf{L} \leq \mathbf{T}_{\mathbf{a},\mathbf{s}} \leq \mathbf{U}\right), \qquad \text{(S.4)} \\ \mathbf{T}_{\mathbf{a},\mathbf{s}} &\sim MVN\left(\mathbf{0}_{(2d-1)\times 1}, \boldsymbol{\Delta}_d \boldsymbol{\Sigma}_{\mathbf{s}}^{(\mathbf{a})} \boldsymbol{\Delta}_d^T\right), \\ \mathbf{L} &= (0, \underbrace{-\infty, ..., -\infty}_{d-1}, \underbrace{0, ..., 0}_{d-1}), \\ \mathbf{U} &= (b_1, b_2, ..., b_d, \underbrace{\infty, ..., \infty}_{d-1}). \end{aligned}$$

Equation (S.4) gives us the integral bounds as constants, at a cost of increasing the dimension of the multivariate normal distribution of interest from $d$ to $2d-1$. This final expression can be used in any number of computing packages to produce the desired probability.

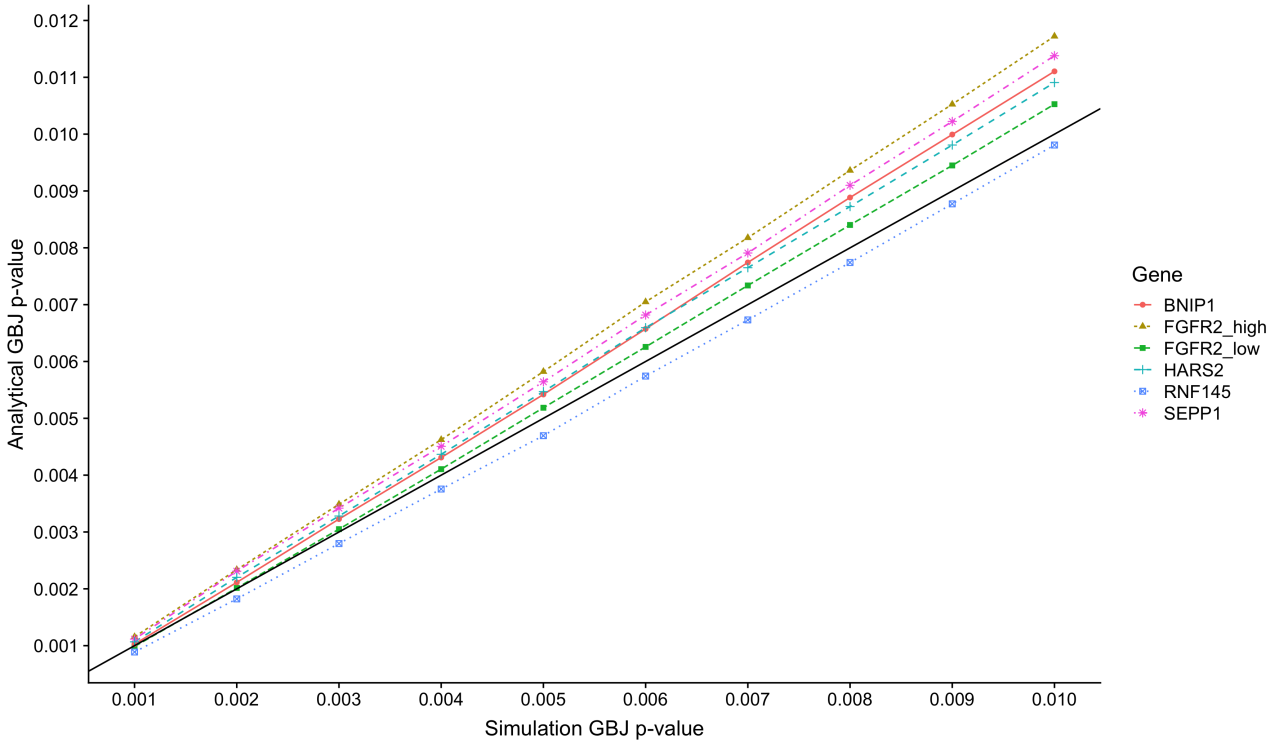
# Supplement C: Accuracy of p-value calculation from Section 3.4.

Here we examine the accuracy of the p-value calculation given in Section 3.4. In particular, the following results are designed to complement and extend the Type I error simulation reported in Table 1. Table 1 showed that the analytical GBJ p-value calculation protects the size of the test under the null, and in this analysis we further demonstrate that the analytical p-values are very close to p-values obtained through simulation.

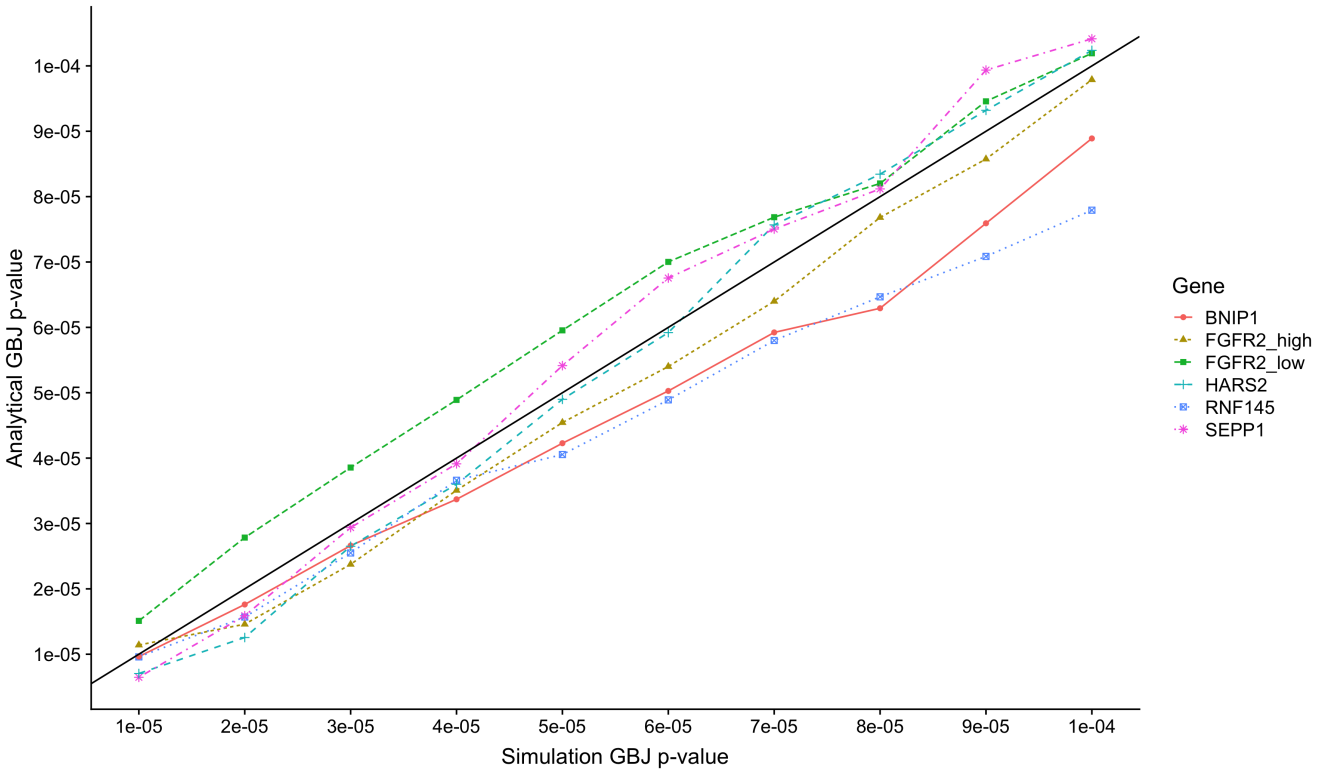
We focus on the low LD FGFR2 region, the high LD FGFR2 region, and four genes of varying size chosen at random from chromosome 5. For each of these regions separately, we first carry out $10^6$ simulations under the null (with the parameters used to construct Table 1) and find the observed value of the GBJ statistic that would correspond to the ten simulated p-values (0.001, 0.002,..., 0.01). We then calculate the analytical p-value of each GBJ

value taken from the above procedure and compare the difference between the analytical and simulated p-values. For example, an observed GBJ statistic of 6.934065 corresponds to the simulated p-value of 0.001 for the high LD FGFR2 region. Calculating the analytical p-value of observing 6.934065 for this region produces $p = 0.00099$.

We then move further into the tail and repeat the analysis for the ten smaller p-values ($1 \cdot 10^{-5}$, $2 \cdot 10^{-5}$,..., $1 \cdot 10^{-4}$). Full results are shown in Figures S1 and S2 below. We see that generally the analytical p-values are very close to the simulated p-values, with some differences possibly attributable to Monte Carlo error. The analytical p-values trend slightly more conservative in Figure S1, which matches the results of Table 1, where the analytical p-value calculation is slightly conservative at nominal levels $\alpha = 0.01$ and $\alpha = 0.001$.



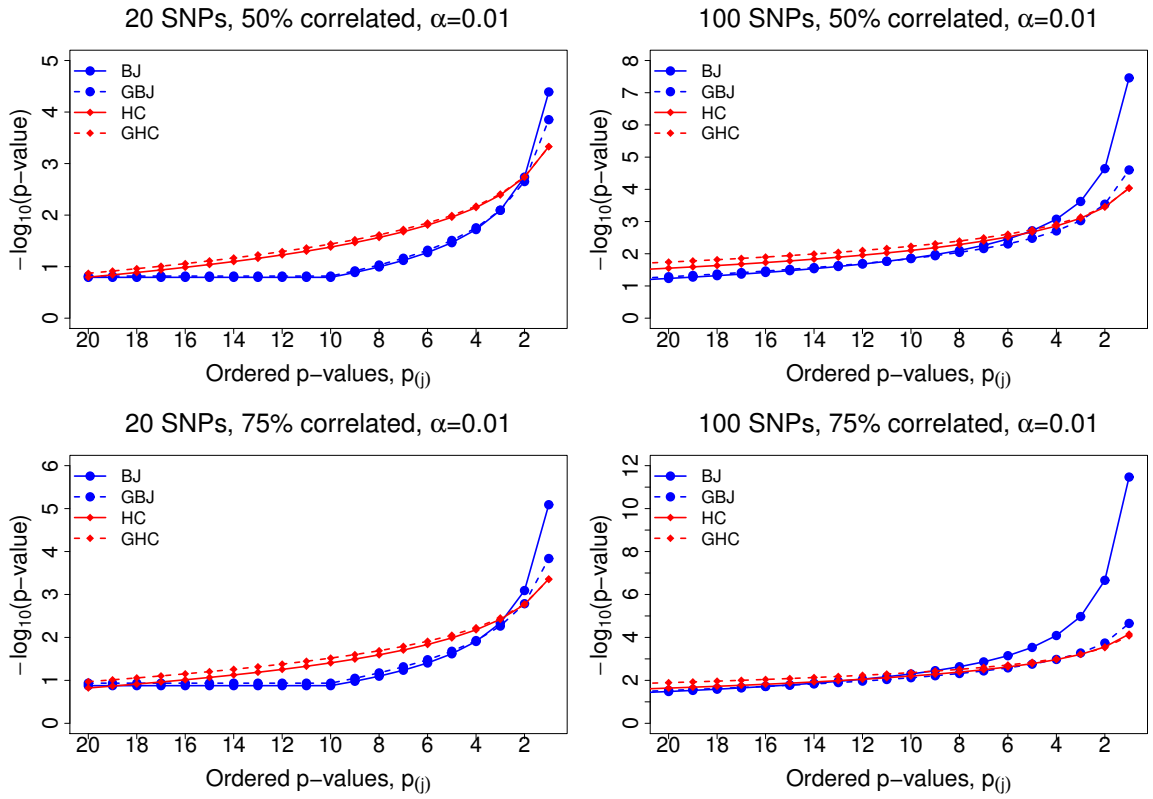

Supplementary Figure 1: Accuracy of analytical p-value calculation. For each region, we find the observed GBJ statistics that would produce an empirical p-value of (0.001, 0.002,..., 0.01). We then calculate the analytical p-values of those observed GBJ statistics and plot them against the simulation p-values. Solid line with no points is the $x = y$ line.




Supplementary Figure 2: Accuracy of analytical p-value calculation. For each region, we find the observed GBJ statistics that would produce an empirical p-value of ($1 \cdot 10^{-5}$, $2 \cdot 10^{-5}$,..., $1 \cdot 10^{-4}$). We then calculate the analytical p-values of those observed GBJ statistics and plot them against the simulation p-values. Solid line with no points is the $x = y$ line.

# Supplement D: Rejection region plots on p-value scale from Section 4.

Here we reproduce the rejection region plots of Figure 1 on the p-value scale. At each integer $j$ on the x-axis, the y-axis gives the rejection boundary in terms of the $j$th smallest p-value. Certain readers may prefer to interpret boundary differences in terms of orders of magnitude on the p-value scale.




Supplementary Figure 3: Rejection region of Berk-Jones, Generalized Berk-Jones, Higher Criticism, and Generalized Higher Criticism tests, plotted according to the order statistics of the p-values. At each point $j$ on the x-axis, if the $j$th smallest p-value is less than the boundary point for a specific test at $j$, then we would reject the null using that test at level $\alpha = 0.01$. The difference between BJ and GBJ becomes much more pronounced as both the size of the set and the amount of correlation increase.

# Supplement E: Complete simulation parameters from Section 5.

Below we give the effect size of the causal SNPs for the simulations using SNPs with pre-determined correlation structures:

Supplementary Table 1: Effect sizes $\beta_j$ for each set of simulations in Figures 2 and 3. For a given number of causal SNPs, all causal SNPs have the same effect size.

| **Correlation** | | | **Number of Causal SNPs** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $\rho_2$ | $\rho_3$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.0 | 0.0 | 0.0 | 0.120 | 0.100 | 0.090 | 0.090 | 0.090 | 0.090 | 0.090 | 0.090 | 0.090 | 0.090 |
| 0.3 | 0.0 | 0.0 | 0.110 | 0.080 | 0.060 | 0.050 | 0.040 | 0.040 | 0.035 | 0.030 | 0.030 | 0.030 |
| 0.3 | 0.0 | 0.3 | 0.110 | 0.090 | 0.060 | 0.050 | 0.050 | 0.040 | 0.030 | 0.030 | 0.030 | 0.030 |
| 0.3 | 0.3 | 0.3 | 0.100 | 0.070 | 0.050 | 0.040 | 0.035 | 0.030 | 0.025 | 0.025 | 0.025 | 0.025 |

Next we give the effect size of the causal SNPs for simulations using HAPGEN2-generated random blocks of 40 SNPs on chromosome 5:

Supplementary Table 2: Effect sizes $\beta_j$ for simulation with HAPGEN2-generated genotypes using random blocks of 40 SNPs on chromosome 5 (Figure 4). For a given number of causal SNPs, all causal SNPs have the same effect size.

| | **Number of Causal SNPs** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $\beta_j$ | 0.15 | 0.14 | 0.12 | 0.11 | 0.11 | 0.10 | 0.10 | 0.10 |

# Supplement F: Power benchmarks from Section 5.

Below we give the number of SNPs necessary to produce a power of at least 80% for a constant effect size of $\beta_j = 0.1$ when testing at $\alpha = 0.01$:

Supplementary Table 3: Minimum number of SNPs necessary to produce a power of at least 80% for a constant effect size of $\beta_j = 0.1$.

| **Correlation** | | | **Test** | | | | |
|---|---|---|---|---|---|---|---|
| $\rho_1$ | $\rho_2$ | $\rho_3$ | GBJ | GHC | minP | SKAT | OMNI |
| 0.0 | 0.0 | 0.0 | 7 | 8 | 9 | 7 | 7 |
| 0.3 | 0.0 | 0.0 | 3 | 3 | 3 | 3 | 3 |
| 0.3 | 0.0 | 0.3 | 3 | 3 | 3 | 5 | 3 |
| 0.3 | 0.3 | 0.3 | 3 | 3 | 3 | 3 | 3 |

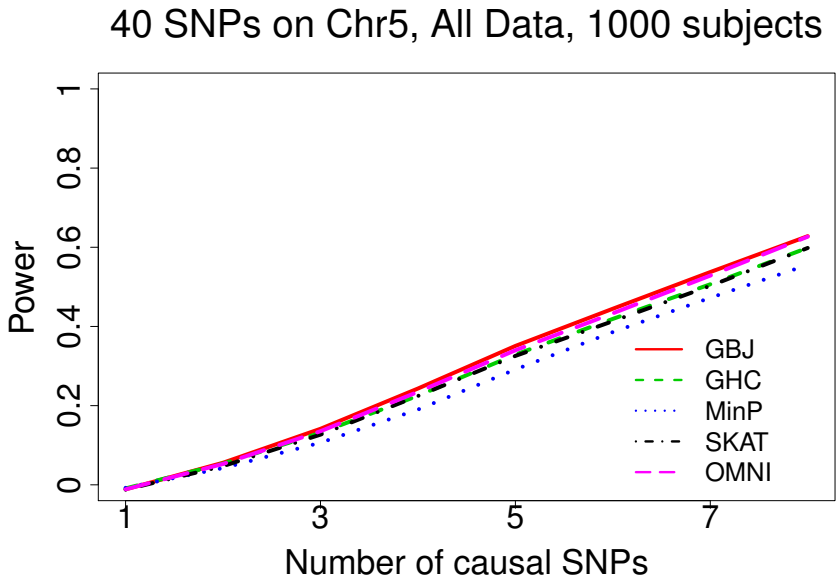
Next we give the number of SNPs necessary to produce a power of at least 80% for a constant effect size of $\beta_j = 0.15$ when testing at $\alpha = 0.01$:

Supplementary Table 4: Minimum number of SNPs necessary to produce a power of at least 80% for a constant effect size of $\beta_j = 0.15$.

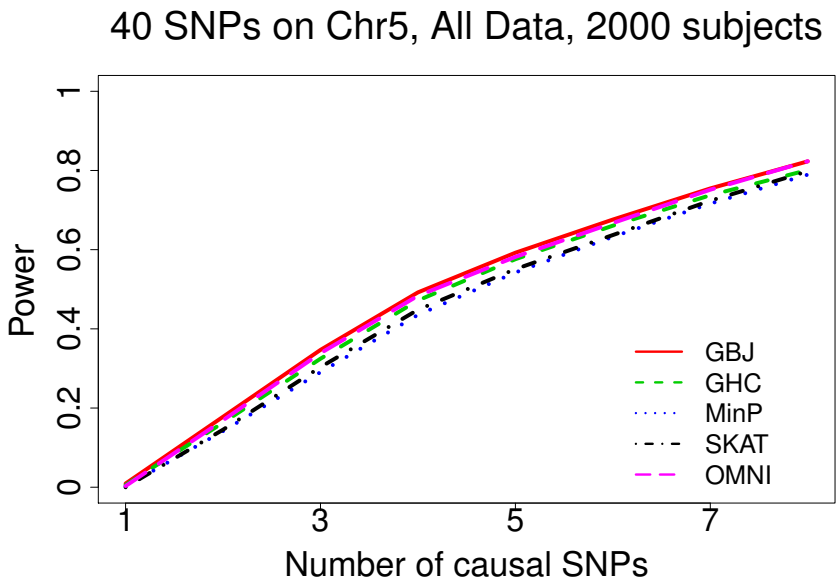
| | **Correlation** | | | **Test** | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho_1$ | $\rho_2$ | $\rho_3$ | GBJ | GHC | minP | SKAT | OMNI |
| 1 | 0.0 | 0.0 | 0.0 | 3 | 2 | 2 | 3 | 2 |
| 2 | 0.3 | 0.0 | 0.0 | 2 | 2 | 2 | 2 | 2 |
| 3 | 0.3 | 0.0 | 0.3 | 2 | 2 | 2 | 3 | 2 |
| 4 | 0.3 | 0.3 | 0.3 | 2 | 2 | 2 | 2 | 2 |

# Supplement G: Supplemental power simulations from Section 6.

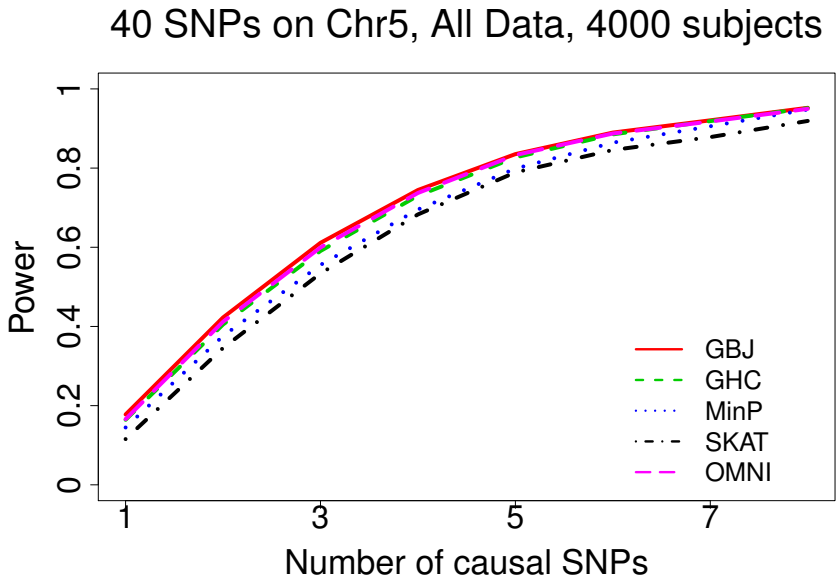
Here we present power simulations similar to Figure 4, except we keep the effect size constant at $\beta_j = 0.1$. We consider three different sample sizes, $n = 1000$, $n = 2000$, and $n = 4000$, and we test at $\alpha = 3.34 \times 10^{-6}$. When there are four causal SNPs, GBJ has power of approximately 23%, 53%, and 74% at $n = 1000, 2000$, and 4000, respectively.




Supplementary Figure 4: Power simulation of Figure 4 except with constant effect size of $\beta = 0.1$ and $n = 1000$.



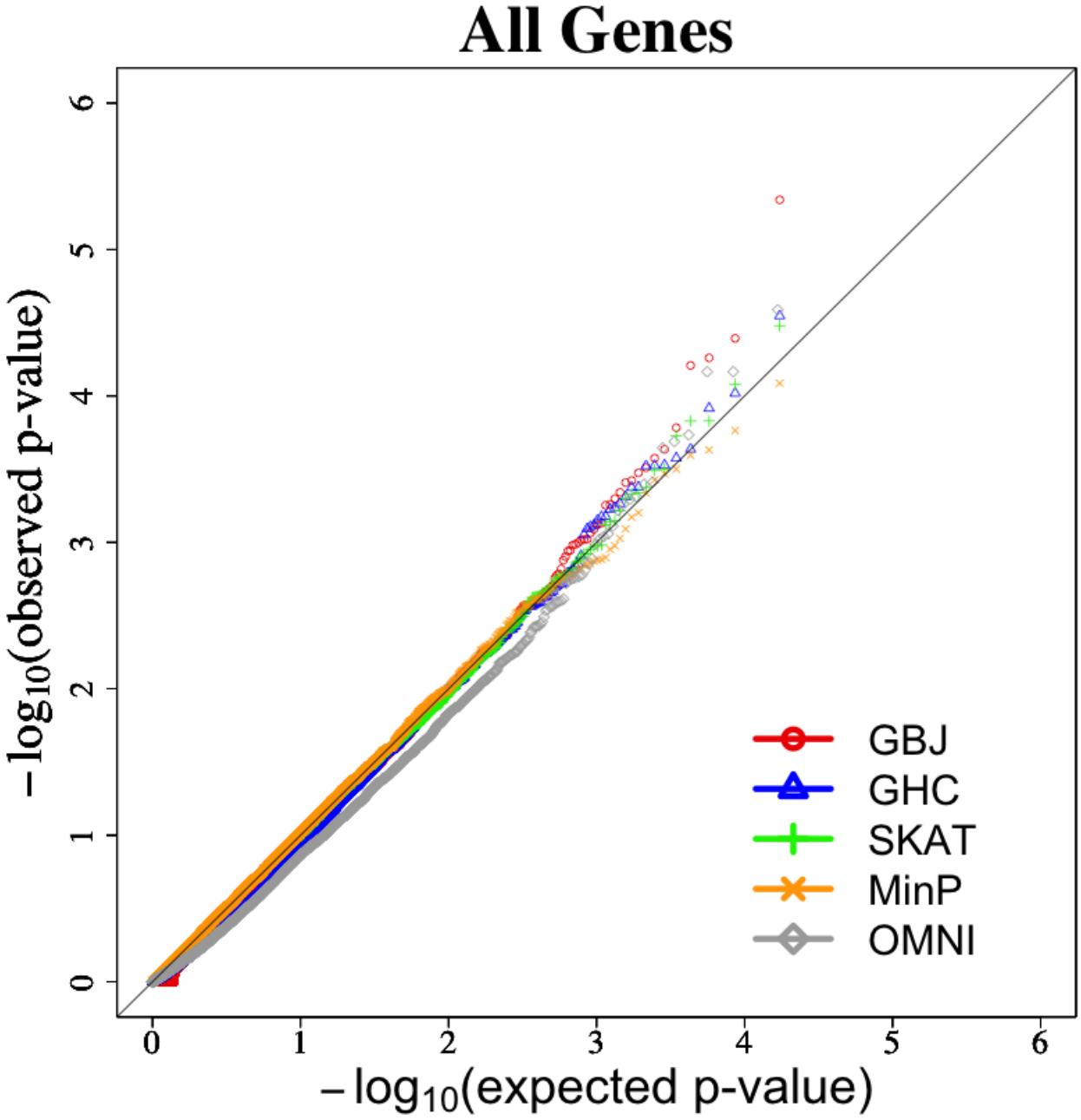

Supplementary Figure 5: Power simulation of Figure 4 except with constant effect size of $\beta = 0.1$ and $n = 2000$.



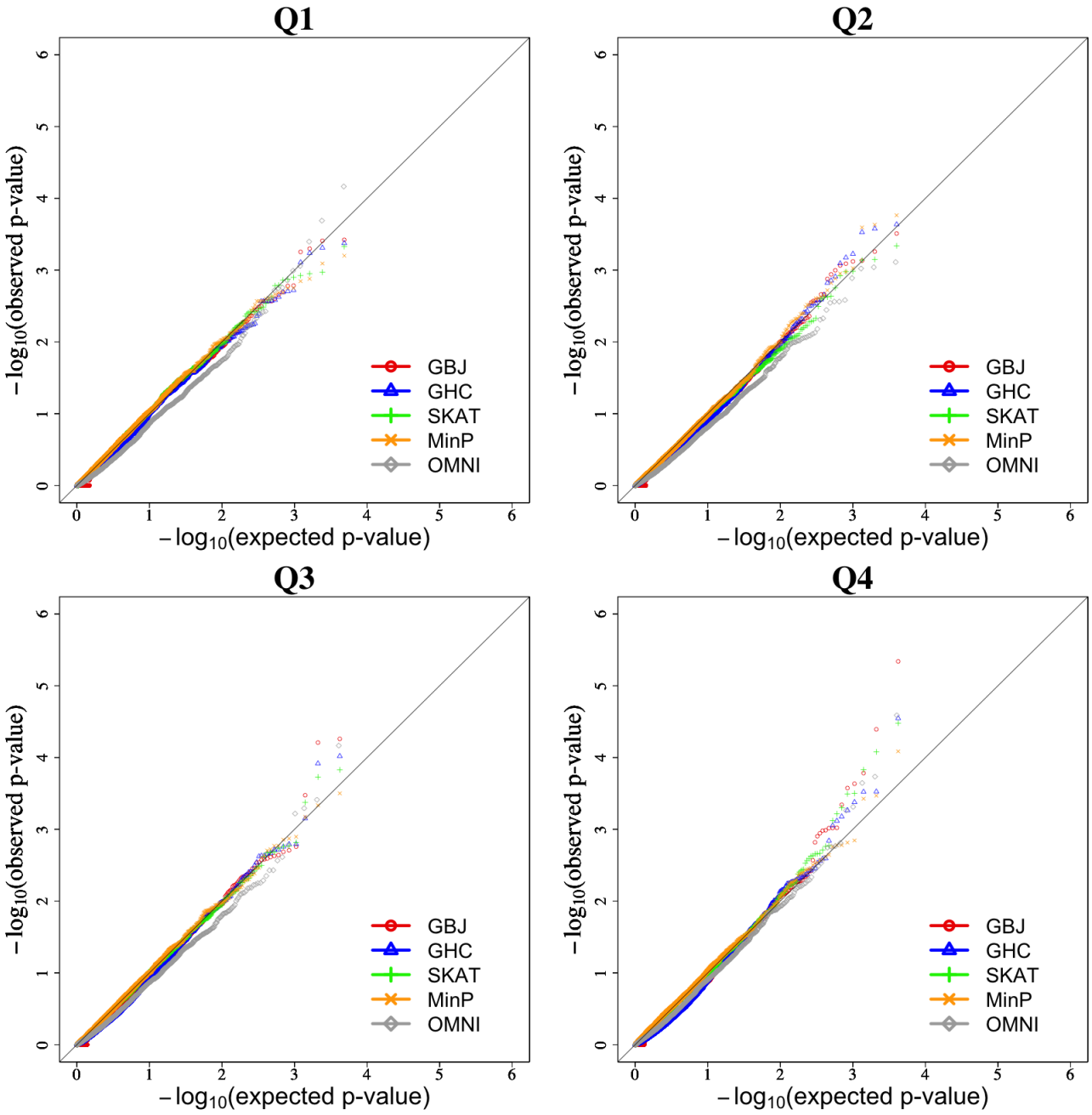

Supplementary Figure 6: Power simulation of Figure 4 except with constant effect size of $\beta = 0.1$ and $n = 4000$.

## Supplement H: Empirical distributions of p-values in CGEMS analysis from Section 6.

Here we use QQ-plots to show the empirical distribution of p-values for each test used in the CGEMS analysis. The vast majority of p-values from all tests lie on the 45-degree line, appearing to indicate that each method is indeed producing valid p-values distributed as Uniform(0,1) random variables under the null. We also stratify the QQ-plots by the four quartiles of gene size. The QQ-plot for the largest quartiles seems to show slightly more signals in the tail across all methods. This behavior might be expected since large genes contain more SNPs and are thus more likely to possess a significant SNP simply by chance.




Supplementary Figure 7: QQ-plots showing empirical distribution of p-values across all genes in CGEMS analysis.




Supplementary Figure 8: QQ-plots showing empirical distribution of p-values stratified by four quartiles of gene size in CGEMS analysis.

# Supplement I: Evaluation of summary statistic correlation approximation in CGEMS analysis from Section 6.

In Section 2.2 we presented an estimate for the correlation structure of precomputed summary statistics. Below we present a small demonstration of the accuracy of this approach using the CGEMS data. We estimate the correlation structure of the top ten genes in the CGEMS analysis using both the individual-level data (as in Section 2.1) and also acting as if the individual-level data were not available (as in Section 2.2). Specifically, in the second approach we mimic a practical analysis of summary data by using reference genotypes from the CEU population of the 1000 Genomes Project as well as the first three principal components calculated from this data. Denote the two estimated correlation matrices by $\hat{\mathbf{\Sigma}}_{CGEMS}$ and $\hat{\mathbf{\Sigma}}_{1000G}$.

For a $d \times d$ matrix $\mathbf{A} = \hat{\mathbf{\Sigma}}_{CGEMS} - \hat{\mathbf{\Sigma}}_{1000G}$ with $(i,j)$th element $a_{ij}$, let the Frobenius norm be given by $||\mathbf{A}||_F = \sqrt{\sum_{i,j} a_{ij}^2}$ and let the matrix $\ell_1$ norm be given by $||\mathbf{A}||_{L_1} = \max_{1 \leq j \leq d} \sum_{i=1}^{d} |a_{ij}|$. We report both norms as well as the median values of $|a_{ij}|$ and $a_{ij}$ below. We also note that p-values calculated using the method of Section 2.2 differed from those calculated using the full data by a median of only $3.15 \times 10^{-5}$ over these ten genes.

Supplementary Table 5: Difference between correlation matrices estimated with original CGEMS data and those estimated using reference data from 1000 Genomes.

| Gene | Frobenius | Matrix $\ell_1$ | Median $\lvert a_{ij} \rvert$ | Median $a_{ij}$ |
|---|---|---|---|---|
| FGFR2 | 2.21 | 3.28 | 0.04 | -0.01 |
| CNGA3 | 1.96 | 3.19 | 0.04 | 0.01 |
| PTCD3 | 0.59 | 0.82 | 0.03 | 0.01 |
| POLR1A | 1.96 | 3.55 | 0.05 | 0.02 |
| ZNF263 | 0.03 | 0.02 | 0.00 | 0.00 |
| VWA3B | 2.13 | 3.04 | 0.03 | 0.00 |
| TBK1 | 1.13 | 1.38 | 0.06 | 0.01 |
| ABCA1 | 3.49 | 3.76 | 0.04 | 0.00 |
| MMRN1 | 0.39 | 0.42 | 0.03 | -0.00 |
| TIGD7 | 0.08 | 0.08 | 0.01 | -0.00 |