The Generalized Higher Criticism for Testing SNP-set Effects in Genetic Association Studies: Supplementary Materials

1 Proof of Theorem 1

We first calculate

$$
\begin{aligned}
&Cov\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}}, \sum_{k=1}^{p} I_{\{|Z_k|>t_j\}}\right) \\
&= E\left((\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}})(\sum_{k=1}^{p} I_{\{|Z_k|>t_j\}})\right) - E\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_i\}}\right)E\left(\sum_{k=1}^{p} I_{\{|Z_k|>t_j\}}\right) \\
&= p[2\bar{\Phi}(max\{t_i,t_j\}) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)] + \sum_{k\neq l}[P(|Z_k|>t_i,|Z_l|>t_j) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)]
\end{aligned}
$$

So it is sufficient to show that

$$
\sum_{k\neq l}[P(|Z_k|>t_i,|Z_l|>t_j) - 4\bar{\Phi}(t_i)\bar{\Phi}(t_j)] = 4p(p-1)\phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{\mathcal{H}_{2n-1}(t_i)\mathcal{H}_{2n-1}(t_j)\overline{r^{2n}}}{(2n)!}
$$

Letting $r_{k,l} = Cov(Z_k, Z_l)$, Schwartzman and Lin (2011) showed that

$$
P(Z_k > t_i, Z_l > t_j) = \bar{\Phi}(t_i)\bar{\Phi}(t_j) + \phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{r_{k,l}^n}{n!}\mathcal{H}_{n-1}(t_i)\mathcal{H}_{n-1}(t_j)
$$

Because Z_k and Z_l are bivariate normal we can rewrite $P(|Z_k| > t_i, |Z_l| > t_j)$ as:

$$
P(|Z_k|>t_i,|Z_l|>t_j) = 2(\bar{\Phi}(t_i) - P(Z_k>t_i, Z_l>-t_j) + P(Z_k>t_i,Z_l>t_j))
$$

Plugging back in yields:

$$\sum_{k\neq l}[P(|Z_k|>t_i,|Z_l|>t_j)-4\bar{\Phi}(t_i)\bar{\Phi}(t_j)]$$

$$=\sum_{k\neq l}2\phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{r_{k,l}^n}{n!}\mathcal{H}_{n-1}(t_i)(\mathcal{H}_{n-1}(t_j)-\mathcal{H}_{n-1}(-t_j))$$

$$=2\phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{\mathcal{H}_{n-1}(t_i)(\mathcal{H}_{n-1}(t_j)-\mathcal{H}_{n-1}(-t_j))}{n!}\sum_{k\neq l}r_{k,l}^n$$

$$=4p(p-1)\phi(t_i)\phi(t_j)\sum_{n=1}^{\infty}\frac{\mathcal{H}_{2n-1}(t_i)\mathcal{H}_{2n-1}(t_j)\overline{r^{2n}}}{(2n)!}$$

2 Proof of the GHC p-value calculation

Using the results in the main text, we have

$$\begin{aligned} pr\,(GHC\geq h) &= 1-pr\left(\bigcap_{t>0}\left\{S(t)<h\sqrt{\widehat{Var}(S(t))}+2p\bar{\Phi}(t)\right\}\right) \\ &= 1-pr\left(\bigcap_{k=1}^{p}\{S(t_k)<p-k+1\}\right) \end{aligned}$$

where the t_k are defined in equation (5) of the main text. We are able to write the intersection over all $t>0$ as an intersection of p events due to the monotone nature of $h\sqrt{\widehat{Var}(S(t))}+2p\bar{\Phi}(t)$ combined with the fact that $S(t)$ can only take on the values $\{0,1,...,p\}$. Applying the chain rule of conditioning leads to:

$$\begin{aligned} pr\,(GHC\geq h) &= 1-pr\left(\bigcap_{k=1}^{p}\{S(t_k)<p-k+1\}\right) \\ &= 1-\prod_{k=1}^{p}pr\left(S(t_k)\leq p-k\,\middle|\,\bigcap_{l=1}^{k-1}\{S(t_l)\leq p-l\}\right)=1-\prod_{k=1}^{p}\sum_{a=0}^{p-k}q_{k,a} \end{aligned}$$

2.1 Proof of Theorem 2

Let $\sigma_a(t)=\sqrt{Var(S(t))}$ and $\sigma_s(t)=\sqrt{2p\bar{\Phi}(t)(1-2\bar{\Phi}(t))}$, and then let $HC(t)=\{S(t)-2p\bar{\Phi}(t)\}/\sigma_s(t)$ and $GHC(t)=\{S(t)-2p\bar{\Phi}(t)\}/\sigma_a(t)$. Noting that $GHC(t)$ is a mean 0 variance 1 random variable,

$$\begin{aligned} pr_{H_0}(GHC > c) &\leq \sum_{t \in [s, \sqrt{5 \log p}] \cap \mathbb{N}} pr_{H_0}(GHC(t) > c) \\ &\leq \sum_{t \in [s, \sqrt{5 \log p}] \cap \mathbb{N}} 1/c^2 \qquad \text{by Chebyshev's Inequality} \\ &= \frac{O(\sqrt{\log p})}{c^2} \end{aligned}$$

Hence for $c = O(\log p)$ we have that $pr_{H_0}(GHC > c) \to 0$. Without loss of generality take $c = \log p$.

Now we study the behavior of GHC under the alternative. By Arias-Castro et al. (2011) we have that if $\max_j |\beta_j| \geq \sqrt{6 \log p}$, then

$$HC(\sqrt{5 \log p}) \geq p^{3/4} \tag{S.1}$$

with probability greater than $1 - o(1/\sqrt{p})$. For the rest of the alternatives satisfying $A \leq \max_j |\beta_j| \leq \sqrt{6 \log p}$, it suffices to show that there exists a $t \in [\sqrt{2 \min(1, 4c^*(\alpha)) \log p}, \sqrt{5 \log p}] \cap \mathbb{N}$ such that $E_{H_1}(GHC(t)) \gg \log p$ and $\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} \to \infty$.

Letting $HC(t) = GHC(t)\frac{\sigma_a(t)}{\sigma_s(t)}$, we have that

$$\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} = \frac{E_{H_1}(HC(t))}{\sqrt{Var_{H_1}(HC(t))}}.$$

In Arias-Castro et al. (2011), proof of theorem 3, they show that for $t = \sqrt{2 \min(1, 4\gamma) \log p}$, $\frac{E_{H_1}(HC(t))}{\sqrt{Var_{H_1}(HC(t))}} \to \infty$. Hence, for the same t, $\frac{E_{H_1}(GHC(t))}{\sqrt{Var_{H_1}(GHC(t))}} \to \infty$.

We will show that for that same t, $E_{H_1}(GHC(t)) = \frac{\sigma_s(t)}{\sigma_a(t)} E_{H_1}(HC(t)) \gg \log p$. For the same t, Arias-Castro et al. (2011) show that $E_{H_1}(HC(t)) \gg (\log p)^2 \sqrt{\Delta}$. This implies that $E_{H_1}(GHC(t)) \gg \frac{\sigma_s(t)}{\sigma_a(t)} (\log p)^2 \sqrt{\Delta}$.

Arias-Castro et al. (2011) showed that $Var_{H_0}(HC(t')) \leq c'(\log p)^2 \Delta$ for some constant $c' > 0$. Combine this inequality with the fact that $Var_{H_0}(HC(t')) = \frac{\sigma_a^2(t')}{\sigma_s^2(t')}$, we have that $\frac{\sigma_s(t)}{\sigma_a(t)} \leq \frac{1}{\sqrt{c'}\log p\sqrt{\Delta}}$. Hence,

$$E_{H_1}(GHC(t)) \gg \frac{1}{\sqrt{c'}\log p\sqrt{\Delta}}(\log p)^2\sqrt{\Delta} = O(\log p)$$

Therefore $E_{H_1}(GHC(t)) \gg \log p$ as required. Using equation (S.1) we evaluate the case where $t = \sqrt{5\log p}$ as

$$GHC(\sqrt{5\log p}) \quad = \quad HC(\sqrt{5\log p})\frac{\sigma_s(\sqrt{5\log p})}{\sigma_a(\sqrt{5\log p})} \gg p^{3/4}\frac{1}{\log p\sqrt{\Delta}} \gg \log p.$$

References

Arias-Castro, E., Candès, E., and Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics* **39,** 2533–2556.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98,** 199–214.