A Permutation Test for Two-Sample Means and Signal Identification of High-dimensional Data

Supplement

Efang Kong[a], Lengyang Wang[b], Yingcun Xia[b,a] and Jin Liu[b]

[a] *University of Electronic Science and Technology of China*
[b] *National University of Singapore, Singapore*

This Supplement gives proofs of theorems in the paper and other related results. Subsection 0.1 includes a brief introduction to the key concept of high-dimensional central limit theorem and some propositions used in the proofs. Subsection 0.2-0.5 contain proofs for Theorems 1, 2, 3 and 4 respectively. Subsection 0.6 contains some additional results for the analysis of the WTCCC dataset.

0.1. High dimensional CLT and other prerequisites

Let $\mathcal{A}^{re}$ be the collection of *all* hyperrectangles in R^p:

$$\{w \in R^p : a_j \leq w_j \leq b_j,\ j = 1, \cdots, p\}.$$

Suppose $X_1, \cdots, X_m$ are independent random vectors in R^p, each with zero mean and finite variances. Write $\Sigma_m^X = m^{-1} \sum_i Cov(X_i)$. Then under certain regularity conditions

$$r_m = \sup_{A \in \mathcal{A}^{re}} |Pr(m^{1/2}\bar{X}_m \in A) - Pr(N(0, \Sigma_m^X) \in A)| \to 0,\ m \to \infty, \quad (1)$$

where $N(.,.)$ stands for the $p-$variate Gaussian; see Proposition 2.1 of Chernozhukov et al. (2017). For ease of exposition, write this as $\bar{X}_m \stackrel{d}{\to} N(0, \Sigma_m^X)$. Similarly for $Y_1, \cdots, Y_n$, each with mean zero and finite variance,

$$s_n = \sup_{A \in \mathcal{A}^{re}} |Pr(n^{1/2}\bar{Y}_n \in A) - Pr(N(0, \Sigma_n^Y) \in A)| \to 0,\ n \to \infty. \quad (2)$$

Chernozhukov et al. (2017) also gave upper bounds concerning other classes of sets in R^p, e.g., the simple convex sets and the sparse convex set. These results are also referred to as normal approximations in high dimension settings, and are crucial to the asymptotic study concerning both the null distribution and the permutation distribution. Also because this paper deals with a two-sample problem, these results need to be adapted to fit our purposes. More details could be found in the proof of the Lemmas in the next section.

The following three lemmas are used in the proofs of the theorems.

Proposition 1. *Let $(g_1, \cdots, g_p)$ be centered Gaussian random vectors in R^p, with covariance matrix Σ and $\lambda_{min}(\Sigma^{-1}) \geq c_0$ for some constant c_0. Then there exists some constant $C > 0$ such that*

$$\sup_t P\Big(|p^{-1}\sum_{i=1}^{p}|g_i| - t| \leq \delta\Big) \leq C\delta$$

where $C > 0$ depends only on c_0.

Proof of Proposition 1. Without loss of generality, suppose $\delta < 1$, for otherwise just take $C = 1$. Write $S = \{(s_1, s_2, \cdots, s_p) : s_i = 1 \text{ or } -1, i = 1, \cdots, p\}$. For any $s = s_1, s_2, \cdots, s_p) \in S$, let R_s be the subset of R^p defined as

$$R_s = \{(x_1, \cdots, x_p) : \sum_i x_i s_i \in [pt, p(t+\delta)],\ x_i s_i \geq 0,\ i = 1, \cdots, p\}.$$

Therefore $R_s : s \in S$ are disjoint and

$$\cup_{s\in S} R_s = \{(x_1, \cdots, x_p) : \sum_{i=1}^{p} |x_i| \in [pt, p(t+\delta)]\} \stackrel{def}{=} A_{t,\delta}.$$

Let $|\Sigma|$ denote the determinant of Σ, and $C_p = (2\pi)^{-p/2}|\Sigma|^{-1/2} \leq (2\pi c_0)^{-p/2}$. Since $Z^\top \Sigma^{-1} Z \geq c_0 Z^\top Z \geq c_0 p |Z|_1^2$, we have

$$P(Z \in A_{t,\delta}) = C_p \int_{A_{t,\delta}} \exp\Big\{ - \frac{Z^\top \Sigma^{-1} Z}{2} \Big\} dZ = C_p \sum_{s\in S} \int_{R_s} \exp\Big\{ - \frac{Z^\top \Sigma^{-1} Z}{2} \Big\} dZ,$$
$$\leq C_p \exp(-c_0 p t^2/2) \sum_{s\in S} \int_{R_s} dZ = \exp(-pt^2/2)\{(t+\delta)^{p+1} - t^{p+1}\} C_p 2^p/p!, \qquad (3)$$

where the last inequality follows from the fact that the 'volume' of each of R_s is $\{(t+\delta)^{p+1} - t^{p+1}\}/p!$. Write $h(t|p,\delta) = \exp(-pt^2/2)\{(t+\delta)^{p+1} - t^{p+1}\}$. With p and δ fixed, we have

$$\lim_{t\to 0} h(t|p,\delta) = \delta^{p+1}; \quad \lim_{t\to\infty} h(t|p,\delta) \approx \delta t^p \exp(-pt^2/2) = o(\delta);$$
$$h'(t|p,\delta) = \exp(-pt^2/2)\Big[(p+1)\{(t+\delta)^p - t^p\} - pt\{(t+\delta)^{p+1} - t^{p+1}\}\Big]$$

with $h'(t|p,\delta)=0$ at some $t^*<1$. Thus we could confine t in $(0,1)$. Since $C_p 2^p/p!$ is bounded,

$$P(Z\in A_{t,\delta})\leq \delta(\pi c_0/8)^{-p/2}\exp(-pt^2/2)(p+1)/p! = O(\delta),$$

where the last inequality follows from Sterling's Approximation. □

Proposition 2. *Consider p centred normal random variables g_i, $1\leq i\leq p$, not necessarily independent. Suppose $Eg_i^2=\sigma_i^2\leq\sigma^2$, $1\leq i\leq p$. Then*

$$E(\sum_{i=1}^p |g_i|) = \sqrt{2/\pi}\sum_i \sigma_i, \quad E\max_{1\leq i\leq p}|g_i|\leq \sigma\sqrt{2\log(2p)}.$$

Proof of Proposition 2. The first equation is trivial. The proof of the second is similar to that of Proposition 1.1.3 of Talagrand (2003). Note that for a centered normal random variable g, and any β,

$$E(\exp\beta\,|g|) = 2\Phi(\sigma\beta)\exp\{E(g^2)\beta^2/2\}\leq 2\exp\{E(g^2)\beta^2/2\}. \tag{4}$$

Using Jensen's inequality and the concavity of the logarithm, we have

$$\beta E\max_{1\leq i\leq p}|g_i|\leq E\log(\sum_{i=1}^p\exp(\beta\,|g_i|))\leq \frac{\beta^2\sigma^2}{2}+\log(2p).$$

Taking $\beta=\sqrt{2\log 2p}/\sigma$, and the proof is complete. □

Proposition 3 (Theorem 1 (c) of Cai et al. (2014))**.** *Consider p normal random variables g_i, $1\leq i\leq p$, with zero mean and a covariance matrix Σ, such that $c_0\leq\lambda_{min}(\Sigma)\leq\lambda_{max}(\Sigma)\leq c_1$ for some constants $c_0>0$, $c_1>0$.*

Let $\sigma_{k,k}$ be the diagonal entries of Σ. We have, for any $x \in R$,

$$P\Big(\max_{1\le k\le p} g_k^2/\sigma_{k,k} - 2\log p + \log\{\log p\} \le x\Big) \to \exp\Big(-\pi^{-1/2}\exp(-x/2)\Big),\ p \to \infty.$$

Proposition 4. *Consider p centered normal random variables g_i, $1 \le i \le p$, not necessarily independent with $Eg_i^2 = \sigma_i^2 > 0$. Let $\underline{\sigma} = \min_i \sigma_i$, $\bar{\sigma} = \max_i \sigma_i$. We have, for any $\epsilon > 0$,*

$$\sup_t P\Big(|\max_i |g_i| - t| \le \epsilon\Big) \le C\epsilon\sqrt{1 \vee \log(2p/\epsilon)},$$

where $C > 0$ depends only on $\underline{\sigma}$ and $\bar{\sigma}$.

Proposition 4 is a direct consequence of Corollary 1 of Chernozhukov et al. (2015).

0.2. Proof of Theorem 1

Proof of Theorem 1 is based on the following two lemmas: the first concerns the null distribution of the two statistics, while the second lemma concerns their permutation distributions. As these two distributions are the same, thus the theorem follows.

Lemma 1. *Suppose H_0 is true. Then under conditions (C1)-(C5) of Section 6,*

$$\sup_{t\in R}\Big|P\Big(H^\infty(Z^N) \le t\Big) - P\Big(|N(0, I_p)|_\infty \le t\Big)\Big| \to 0; \qquad (5)$$

while under conditions (C1)-(C2),(C3′),(C4)-(C5) of Section 6,

$$\sup_{t\in R}\Big|P\Big(H^1(Z^N)\le t\Big)-P\Big(|N(0,I_p)|_1\le t\Big)\Big|\to 0. \tag{6}$$

Lemma 2. *Under conditions (C1)-(C5) of Section 6, we have*

$$\sup_{t\in R}\Big|\frac{1}{N!}\sum_{\pi\in G_N} I\{H^\infty(Z_\pi^N)<t\}-P\Big(|N(0,I_p)|_\infty<t\Big)\Big|\xrightarrow{p}0. \tag{7}$$

Under conditions (C1)-(C2),(C3′),(C4), and (C5) of Section 6,

$$\sup_{t\in R}\Big|\frac{1}{N!}\sum_{\pi\in G_N} I\{H^1(Z_\pi^N)<t\}-P\Big(|N(0,I_p)|_1<t\Big)\Big|\xrightarrow{p}0. \tag{8}$$

Lemma 1 could been seen as an adaptation of results on high-dimension normal approximation of Chernozhukov et al. (2017) to the current two-sample problem.

Proof of Lemma 1. The proof is a combination of high-dimension normal approximation with the coupling procedure similar to that devised in Chung and Romano (2013). We begin with the normal approximation to $\delta_N = N^{1/2}(\bar X_m-\bar Y_n)$. Note that $\delta_N = N^{1/2}\{\bar X_m-\mu^X-(\bar Y_n-\mu^Y)\}$, so under H_0, we may take X_i and Y_j as already centered. Let s_i, $i=1,\cdots,N$ be independent random variables such that $s_i=1$ with probability c, $s_i=-1$ with probability $1-c$. If $s_i=1$, $Z_i\sim X_i/c$, and $Z_i\sim Y_i/(1-c)$ otherwise. Let $m=\sum_i I(s_i=1)$, then $c_N=m/N-c=O_p(N^{-1/2})$ and

$$\frac{1}{N}\sum_{i=1}^N Z_is_i=\frac{m}{N}\frac{\bar X_m}{c}-\frac{n}{N}\frac{\bar Y_n}{1-c},\quad (\bar X_m-\bar Y_n)=\frac{1}{N}\sum_{i=1}^N Z_is_i+c_N\Big(\frac{\bar X_m}{c}-\frac{\bar Y_n}{1-c}\Big).$$

The proof of (5) consists of three steps.

Step 1. We first show that

$$\sup_{t\in R}\Big|P\Big(|\delta_N|_\infty < t\Big) - P\Big(|N(0,\tilde{\Sigma})|_\infty < t\Big)\Big| \to 0. \tag{9}$$

It is easy to see that for any $\epsilon > 0$,

$$P\Big(|\delta_N|_\infty < t\Big) \le P\Big(N^{1/2}|\bar{X}_m/c - \bar{Y}_n/(1-c)|_\infty > \epsilon/c_N\Big) + P\Big(N^{-1/2}|\sum_{i=1}^N Z_i s_i|_\infty < t+\epsilon\Big) = I + II, \tag{10}$$

$$P\Big(|\delta_N|_\infty < t\Big) \ge -P\Big(N^{1/2}|\bar{X}_m/c - \bar{Y}_n/(1-c)|_\infty > \epsilon/c_N\Big) + P\Big(N^{-1/2}|\sum_{i=1}^N Z_i s_i|_\infty < t-\epsilon\Big) = -I + III. \tag{11}$$

For I, due to (1) and (2), it is easy to see that

$$\begin{aligned} I &\le P\Big(|N^{1/2}\bar{X}_m|_\infty > \epsilon/(2c_N)\Big) + P\Big(|N^{1/2}\bar{Y}_n|_\infty > \epsilon/(2c_N)\Big) \\ &\le r_m + s_n + P\Big(|N(0,\Sigma^X)|_\infty > \sqrt{c}\epsilon/(2c_N)\Big) + P\Big(|N(0,\Sigma^Y)|_\infty > \epsilon\sqrt{1-c}/(2c_N)\Big). \end{aligned}$$

Since $c_N = O_p(N^{-1/2})$, so for any $\delta > 0$, there exists a $C > 0$ such that $P(c_N \le Cn^{-1/2}) \ge 1-\delta$. Thus

$$\begin{aligned} &P\Big(|N(0,\Sigma^X)|_\infty > \sqrt{c}\epsilon/(2c_N)\Big) \le P\Big(|N(0,\Sigma^X)|_\infty > \sqrt{c}C^{-1}N^{1/2}\epsilon/2\Big) + \delta \\ &\le E(|N(0,\Sigma^X)|_\infty)Cc^{-1}N^{-1/2}\epsilon^{-1} + \delta \le C(\log p/N)^{1/2}\epsilon^{-1} + \delta, \end{aligned} \tag{12}$$

where the first inequality follows from Chebyshev inequality, while the second is due to Proposition 2. Since $\log p = O(n^\alpha)$, $0 < \alpha < 1/7$, let

$\delta \to 0$ and $\epsilon \to 0$ (e.g. $\epsilon = CN^{(\alpha-1)/2}$) be properly chosen. It follows that $I \le r_m + s_n + o(1)$.

Now, consider II. As $Var(Z_i s_i) = \tilde{\Sigma}$, apply Proposition 2.1 of Chernozhukov et al. (2017) we get

$$b_N \stackrel{def}{=} \sup_t \Big| P\Big(\Big| N^{-1/2} \sum_{i=1}^{N} Z_i s_i \Big|_\infty < t \Big) - P\Big(|N(0, \tilde{\Sigma})|_\infty < t \Big) \Big| \to 0. \quad (13)$$

On the other hand, for any real z_i, $i = 1, \cdots$, and t, $\epsilon > 0$, $t - \epsilon > 0$,

$$\{t - \epsilon < \max_i |z_i| < t + \epsilon\} \subset \{t - \epsilon < \max_i z_i < t + \epsilon\} \cup \{t - \epsilon < \max_i(-z_i) < t + \epsilon\}.$$

Therefore, with ϵ chosen as above, applying Proposition 4 we have

$$\sup_t P\Big(\Big| |N(0, \tilde{\Sigma})|_\infty - t \Big| \le \epsilon \Big) \le 2C\epsilon \sqrt{1 \vee \log(2p/\epsilon)} = O(N^{\alpha/2 - 1/4}). \quad (14)$$

This together with (13) suggest that for all $t \in R$,

$$P\Big(|\delta_N|_\infty < t \Big) - P\Big(|N(0, \tilde{\Sigma})|_\infty < t \Big) \le b_N + r_m + s_n + O(N^{\alpha/2 - 1/4}). \quad (15)$$

In a similar manner we could obtain, based on (11), an analogue of (15) but in the opposite direction, i.e., for for all $t \in R$,

$$P\Big(|\delta_N|_\infty < t \Big) - P\Big(|N(0, \tilde{\Sigma})|_\infty < t \Big) \ge -b_N - r_m - s_n + O(N^{\alpha/2 - 1/4}).$$

As r_m, s_n and b_N tend to zero as $N \to \infty$, these two inequalities together yield (9).

Step 2. As in Step 1, but with X_i replaced with $\tilde{\Omega}^{1/2} X_i$ and Y_i replaced with $\tilde{\Omega}^{1/2} Y_i$. We have

$$\sup_{t \in R} \Big| P\Big(|\tilde{\Omega}^{1/2}\delta_N|_\infty < t\Big) - P\Big(|N(0, I_p)|_\infty < t\Big)\Big| \to 0. \tag{16}$$

Step 3. We show that

$$\sup_{t \in R} \Big| P\Big(|\tilde{\Omega}_N^{1/2}\delta_N|_\infty \le t\Big) - P\Big(|\tilde{\Omega}^{1/2}\delta_N|_\infty \le t\Big)\Big| \to 0. \tag{17}$$

Note that for any $\epsilon > 0$

$$\begin{aligned} P\Big(|\tilde{\Omega}_N^{1/2}\delta_N|_\infty \le t\Big) &\le P\Big(|\{\tilde{\Omega}_N^{1/2} - \tilde{\Omega}^{1/2}\}\delta_N|_\infty \ge \epsilon\Big) + P\Big(|\tilde{\Omega}^{1/2}\delta_N|_\infty \le t + \epsilon\Big) \\ &= I + II. \end{aligned} \tag{18}$$

To obtain an upper bound of I, note that

$$I \le P\Big(|N^{1/2}\{\tilde{\Omega}_N^{-1/2} - \tilde{\Omega}^{-1/2}\}\bar{X}_m|_\infty \ge \epsilon/2\Big) + P\Big(|N^{1/2}\{\tilde{\Omega}_N^{-1/2} - \tilde{\Omega}^{-1/2}\}\bar{Y}_n|_\infty \ge \epsilon/2\Big)$$

$$P\Big(|N^{1/2}\{\tilde{\Omega}_N^{1/2} - \tilde{\Omega}^{1/2}\}\bar{X}_m|_\infty \ge \epsilon/2\Big) \le P\Big(\sup_{v \in R^p : |v|_1 \le Ca_n} |N^{1/2} v^\top \bar{X}_m|_\infty \ge \epsilon/2\Big)$$

$$= P\Big(Ca_n |N^{1/2}\bar{X}_m|_\infty \ge \epsilon/2\Big) \le P\Big(Ca_n (N/m)^{1/2} |N(0, \Sigma^X)|_\infty \ge \epsilon/2\Big) + r_m, \tag{19}$$

$$= P\Big(|m^{1/2} N(0, \Sigma^X)|_\infty \ge (m/N)^{1/2}\epsilon/(2Ca_n)\Big) + r_m \le C(\log p)^{1/2} a_n/\epsilon + r_m, \tag{20}$$

where $a_n = \|\tilde{\Omega}_N^{1/2} - \tilde{\Omega}^{1/2}\|_{(1,1)}$. Here, (19) is a result of (1), and (20) is due to Chebyshev inequality and Proposition 2.

Regarding II, similar to (14), we have

$$\sup_t \Big| P\Big(|N(0, I_p)|_\infty \le t + \epsilon\Big) - P\Big(|N(0, I_p)|_\infty \le t\Big)\Big| \le C\epsilon(\log p)^{1/2}.$$

Thus, (20) and (a.1) of condition (C5) of Section 6, will both tend to zero with $\epsilon = a_n^{1/2}$. These together with (16) and (18) suggest that

$$P\Big(|\tilde{\Omega}_N^{1/2}\delta_N|_\infty \le t\Big) - P\Big(|N(0, I_p)|_\infty \le t\Big) \le r_m + s_n + o(1).$$

In a similar manner we could establish an analogue of the opposite direction. This completes the proof of (5) for $\gamma = \infty$.

Next, we consider (6). The proof also consists of three steps.

Step 1. We first show that

$$\sup_{t\in R}\Big|P\Big(|N^{1/2}(\bar{X}_m - \bar{Y}_n)|_1 < t\Big) - P\Big(|N(0, \tilde{\Sigma})|_1 < t\Big)\Big| \to 0. \qquad (21)$$

Again it is easy to see that for any $\delta > 0$,

$$\begin{aligned} P\Big(|\delta_N|_1 < t\Big) &\le P\Big(N^{1/2}|\bar{X}_m/c - \bar{Y}_n/(1-c)|_\infty > \epsilon/c_N\Big) \\ &\quad + P\Big(N^{-1/2}|\sum_{i=1}^N Z_i s_i|_1 < t + \epsilon\Big) = I + II, \qquad (22) \\ P\Big(|\delta_N|_1 < t\Big) &\ge -P\Big(N^{1/2}|\bar{X}_m/c - \bar{Y}_n/(1-c)|_\infty > \epsilon/c_N\Big) \\ &\quad + P\Big(N^{-1/2}|\sum_{i=1}^N Z_i s_i|_1 < t - \epsilon\Big) = -I + III. \qquad (23) \end{aligned}$$

It has already been shown in (12) that $I \le r_m + s_n + C(\log p/N)^{1/2}\epsilon^{-1} + o(1)$.

On the other hand, similar to (13) we have

$$b_N \stackrel{def}{=} \sup_t \Big|P\Big(\Big|N^{-1/2}\sum_{i=1}^N Z_i s_i\Big|_1 < t\Big) - P\Big(|N(0, \tilde{\Sigma})|_1 < t\Big)\Big| \to 0, \qquad (24)$$

which follows from Proposition 3.1 of Chernozhukov et al. (2017), as ν in condition (C3$'$) is the set of unit vectors that are outward normal to the facets of $\{w \in R^p : \sum |w_j| \leq t\}$, a convex polytope with 2^p facets. According to Proposition 1, there exists some $C > 0$, such that

$$\sup_t P\Big(\Big||N(0,\tilde{\Sigma})|_1 - t\Big| \leq \epsilon\Big) \leq C\epsilon. \tag{25}$$

With $\epsilon = (\log p/N)^{1/4}$, (22), (24) and (25) together imply that for all $t \in R$,

$$P\Big(|\delta_N|_1 < t\Big) - P\Big(|N(0,\tilde{\Sigma})|_1 < t\Big) \leq r_m + s_n + b_N + o(N^{(\alpha-1)/4}). \tag{26}$$

In a similar manner based on (23) we could obtain an analogue of (26) in the opposite direction, i.e., for for all $t \in R$,

$$P\Big(|\delta_N|_1 < t\Big) - P\Big(|N(0,\tilde{\Sigma})|_1 < t\Big) \geq -r_m - s_n - b_N + O(N^{(\alpha-1)/4}).$$

These two inequalities together yield (21).

Step 2. As in Step 1, but with X_i replaced with $\tilde{\Omega}^{-1/2}X_i$ and Y_i replaced with $\tilde{\Omega}^{-1/2}Y_i$, we could obtain

$$\sup_{t\in R}\Big|P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 < t\Big) - P\Big(|N(0,I_p)|_1 < t\Big)\Big| \to 0.$$

Step 3. We show that

$$\sup_{t\in R}\Big|P\Big(|\tilde{\Omega}_N^{1/2}\delta_N|_1 \leq t\Big) - P\Big(|\tilde{\Omega}^{-1/2}\delta_N|_1 \leq t\Big)\Big| \to 0.$$

Note that for any $\epsilon > 0$

$$P\Big(|\tilde{\Omega}_N^{1/2}\delta_N|_1 \leq t\Big) \leq P\Big(|\{\tilde{\Omega}_N^{1/2} - \tilde{\Omega}^{1/2}\}\delta_N|_\infty \geq \epsilon\Big) + P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 \leq t + \epsilon\Big) = I + II.$$

Recall that in (20) it has already been shown that $I \leq C(\log p)^{1/2}a_n/\epsilon + r_m + s_n$. Meanwhile, according to Proposition 1, there exists a constant $C > 0$, such that

$$\sup_t \Big| P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 \leq t + \epsilon\Big) - P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 \leq t\Big)\Big| \leq C\epsilon.$$

Taking $\epsilon = (a_n \log p)^{1/2}$, it follows that $I \to 0$. We have for all $t \in R$,

$$P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 \leq t\Big) \leq P\Big(|\tilde{\Omega}^{1/2}\delta_N|_1 \leq t\Big) + o(1);$$

in a similar manner an analogue of this but in the opposite direction also holds. The proof thus completes. □

Proof of Lemma 2. Derivation of the limit of the permutation distribution is mainly based on the following fact. The permutation distribution based on $Z^N = \{Z_1, \cdots, Z_N\}$ behaves approximately like the permutation distribution based on a sample of N IID observations $\tilde{Z}^N = \{\tilde{Z}_1, \cdots, \tilde{Z}_N\}$, generated from the mixture distribution $\bar{P} = cP_1(.) + (1-c)P_2(.)$.

Let us first introduce a layered coupling construction. Mimicking the coupling procedure in Chung and Romano (2013), we can, except for ordering, construct $\tilde{Z}^N$ too via a two-stage process, so that as many as possible of the original observations are used to make up the Z_is. First draw an index i_1 from $\{0, 1\}$ at random with probabilities of c and $1-c$, respectively; if $i_1 = 0$, then $\tilde{Z}_1 = Z_1$, otherwise $\tilde{Z}_1 = Z_{m+1}$. Next, a second index i_2 is randomly selected from $\{0, 1\}$ (again with probabilities of c and $1-c$, respectively), then $\tilde{Z}_2 = Z_2$ if $i_1 = i_2 = 0$, $\tilde{Z}_2 = Z_{m+2}$ if $i_1 = i_2 = 1$, $\tilde{Z}_2 = Z_1$ if $i_2 = 0$, $i_1 = 1$, and finally $\tilde{Z}_2 = Z_{m+1}$ if $i_1 = 0, i_2 = 1$. We continue in this manner to use the Z_i to fill in the observations $\tilde{Z}_i$ until we reach a point where we have used up all the observations from either 'population'. Suppose this happens with the X_i's group when we are about to fill in the ith coordinate of $\tilde{Z}$; if the newly drawn index is again 0, we simply draw a random $\tilde{Z}_i \sim P_1(.)$. A similar procedure applies if we have used up all the observation from the Y_j's group and an index of 1 is selected. Continue in this manner so that as many as possible of the original Z_i observations are used in the construction of $\tilde{Z}^N$. At the end, Z^N and $\tilde{Z}^N$ have many of the observations in common, and we use D to denote the (random) number of extra observations required to fill up $\tilde{Z}^N$.

Next, we reorder the observations in $\tilde{Z}^N$ by a permutation π_0, so that Z_i and $\tilde{Z}_{\pi_0(i)}$ agree for all i except for the aforementioned random number D. Specifically, let N_1 stands for the number of observations in $\tilde{Z}$, which were filled in when index 0 was drawn. If N_1 is greater than or equal to m, then $\tilde{Z}_{\pi_0(i)}$ for $i = 1, \cdots, n_1$ are filled with the first m of these observations; and put aside the remaining $N_1 - m$ observations. On the other hand, if $N_1 < m$, then fill in all these (originally labelled as X_is) observations and leave the rest of $m - N_1$ slots blank for now. Next, we move onto the observations in $\tilde{Z}^N$, which were filled in when index 1 is drawn, and repeat the above procedure for $m+1, \cdots, N$ spots. The third and also the last step is to complete the observations in $\tilde{Z}^N_{\pi_0}$: simply fill up the empty slots with the remaining observations that have been put aside in the previous two steps, and it does not matter where each of the remaining observations is placed. This arrangement of observations in $\tilde{Z}^N$ corresponds to a permutation π_0, and satisfies $Z_i = \tilde{Z}_{\pi_0(i)}$ for all indices i except for D of them.

To further illustrate, we consider an example with $m = n = 3$, $N_1 = 4$; since N_1 as defined above, is the number of observations in $\tilde{Z}$, which were filled in when index 0 was drawn, there are in total 4 observations from $P_1(\cdot)$, and $N_2 = 2$ observations from P_2. In other words, besides the original observations X_1, X_2, X_3, an extra X_4 is drawn from $P_1(\cdot)$ to fill up $\tilde{Z}^N$; and also only Y_1, Y_2 are used to fill up $\tilde{Z}^N$, while Y_3 is unused. Z^N and $\tilde{Z}^N$ have 5 out of 6 observations in common, thus $D = 1$. After the reordering of $\tilde{Z}^N$ through π_0, we have $\tilde{Z}^N_{\pi_0} = (X_1, X_2, X_3, Y_1, Y_2, X_4)$, thus $\tilde{Z}^N_{\pi_0}$ and Z^N differ only at the 6th position.

As proved in Chung and Romano (2013), $E(D/N) \leq N^{-1/2}$. In general, the proof of consistency of the permutation test based on a generic statistic $T_N(\cdot)$ goes as follows. Let $R^T_N(\cdot)$ denote the randomization distribution of

$T_N(\cdot)$, when all the observations are IID generated in the coupling procedure. Suppose π and π' are independent and uniformly distributed over G_N, and they are also independent of Z^N and $\tilde{Z}^N$. The proof begins with showing that

$$(T_N(\tilde{Z}^N_\pi), T_N(\tilde{Z}^N_{\pi'})) \xrightarrow{d} (T, T'), \tag{27}$$

where T and T' are independent random variables with common c.d.f. $R_0(\cdot)$. Equation (27) also implies

$$(T_N(\tilde{Z}^N_{\pi\pi_0}), T_N(\tilde{Z}^N_{\pi'\pi_0})) \xrightarrow{d} (T, T'), \tag{28}$$

where $\pi\pi_0$ stands for π composed with π_0, with π_0 applied first. With the coupling between $\tilde{Z}^N$ and Z^N and the reordering, if we could further show that

$$T_N(\tilde{Z}^N_{\pi\pi_0}) - T_N(Z^N_\pi) \xrightarrow{p} 0, \tag{29}$$

then by the Slutsky's theorem, it follows from (28) and (29) that

$$(T_N(Z^N_\pi), T_N(Z^N_{\pi'})) \xrightarrow{d} (T, T'). \tag{30}$$

Equation (30) is the Hoeffding's condition which enables us to claim that the randomization distribution of $T_N(Z^N)$ to converge in probability to $R_0(\cdot)$. Chung and Romano (2013) proved that this condition is also necessary.

Write $H_0(Z^N) = \{\tilde{\Sigma}/\{c(1-c)\}\}^{-1/2}\delta_N$. Let π and π' be random permutations, independent and uniformly distributed over G_N, and they are also independent of Z^N. The proof consists of the following steps.

- Firstly, based on (a.1) of condition (C5) of Section 6 and through the same arguments used to prove Lemma 5.3 in Chung and Romano (2013), we can show that

$$\left|\{\tilde{\Omega}_N(Z_{\pi(1)},\cdots,Z_{\pi(N)})\}^{1/2} - \sqrt{c(1-c)}\{\Sigma(\bar{P})\}^{-1/2}\right|_{(1,1)} = o_p(1/\log p). \quad (31)$$

- Next, show that

$$(H_0^\gamma(Z_\pi^N), H_0^\gamma(Z_{\pi'}^N)) \xrightarrow{d} (H, H'), \quad (32)$$

  where H and H' are independent and identically distributed as $|N(0, I_p)|_\gamma$.

If both (31) and (32) are true, then

$$|H_0^\gamma(Z_\pi^N) - H_0^\gamma(Z_\pi^N)| \le \left|\left[\{\tilde{\Omega}_N(Z_{\pi(1)},\cdots,Z_{\pi(N)})\}^{1/2} - \{\Sigma(\tilde{P})/\{c(1-c)\}\}^{-1/2}\right]\delta_N\right|_\infty$$
$$\le \sup_{v\in R^p:|v|_1=o_p(1/\log p)} |v^\top\delta_N| \le o_p(1/\log p)|\delta_N|_\infty = o_p(1),$$

where for the last equality we made use of (9). This together with an application of Lemma A.2 of Chung and Romano (2013) lead to the conclusion that the randomization distribution of $H^\gamma(Z_\pi^N)$ converges pointwise (uniformly) in probability to the c.d.f. of $|N(0, I_p)|_\gamma$. □

Proof of (32). Suppose $\tilde{Z}^N = (\tilde{Z}_1, \cdots, \tilde{Z}_N)$, such that $(\tilde{Z}_i, \theta_i)\, i = 1, \cdots, N$, are IID observations generated from the mixture distribution $\bar{P}$. $\tilde{Z}^N$ is also independent of random permutation $\pi \in G_N$. We first show that

$$(H_0^\gamma(\tilde{Z}^N), H_0^\gamma(\tilde{Z}_\pi^N)) \xrightarrow{d} (H, H'), \tag{33}$$

where H, H' are independent and identically distributed.

To this aim, write $\tilde{Z}_l = (\tilde{Z}_{l,1}, \cdots, \tilde{Z}_{l,p})^\top$, $l = 1, \cdots, N$, and $Cov(\tilde{Z}_i) = \Sigma(\bar{P})$. Let $I_1 = \{1, \cdots, m\}$ and $I_2 = \{m+1, \cdots, N\}$. Write $\mu(\bar{P}) = c\mu^X + (1-c)\mu^Y$, and

$$V_1^N = \tfrac{1}{m^{1/2}} \sum_{l=1}^N (\tilde{Z}_l - \mu(\bar{P})) I(l \in I_1),\ V_2^N = \tfrac{1}{n^{1/2}} \sum_{l=1}^N (\tilde{Z}_l - \mu(\bar{P})) I(l \in I_2),$$
$$\tilde{V}_1 = \tfrac{1}{m^{1/2}} \sum_{l=1}^N (\tilde{Z}_l - \mu(\bar{P})) I(\pi(l) \in I_1),\ \tilde{V}_2 = \tfrac{1}{n^{1/2}} \sum_{l=1}^N (\tilde{Z}_l - \mu(\bar{P})) I(\pi(l) \in I_2).$$

For these random vectors, the corresponding central limit theorem in the sense of (1) is such that

$$(V_1^N, V_2^N, \tilde{V}_1^N, \tilde{V}_2^N) \xrightarrow{d} (V_1, V_2, \tilde{V}_1, \tilde{V}_2), \quad m, n \to \infty, \tag{34}$$

where $V_1, V_2, \tilde{V}_1, \tilde{V}_2$ are all $p-$dimensional normal random vectors, with block-wise covariance matrix given as

$$Cov(V_1) = Cov(V_2) = Cov(\tilde{V}_1) = Cov(\tilde{V}_2) = \Sigma(\bar{P}), \tag{35}$$
$$Cov(V_1, V_2) = Cov(\tilde{V}_1, \tilde{V}_2) = \mathbf{0}, \quad Cov(V_1, \tilde{V}_1) = c\,\Sigma(\bar{P}),$$
$$Cov(V_1, \tilde{V}_2) = Cov(\tilde{V}_1, V_2) = \{(1-c)c\}^{1/2}\Sigma(\bar{P}),\ Cov(V_2, \tilde{V}_2) = (1-c)\,\Sigma(\bar{P}),$$

which are deduced based on the simple facts that for i, $j = 1$, 2,

$$\sum_{l=1}^{N} I(l \in I_i, \pi(l) \in I_j)/n_i \xrightarrow{P} p_j, \quad n_1 = m,\, n_2 = n,\, p_1 = c,\, p_2 = 1 - c.$$

Now consider the following linear combinations of $V_1^N, V_2^N, \tilde{V}_1^N$ and $\tilde{V}_2^N$:

$$W_i^N = V_i^N - (n_i/N)^{1/2}\{c^{1/2}V_1^N + (1-c)^{1/2}V_2^N\}, \tag{36}$$

$$\tilde{W}_i^N = \tilde{V}_i^N - (n_i/N)^{1/2}\{c^{1/2}\tilde{V}_1^N + (1-c)^{1/2}\tilde{V}_2^N\},\ i = 1, 2,\, n_1 = m,\, n_2 = n.$$

In view of (34), $(W_1^N, W_2^N, \tilde{W}_1^N, \tilde{W}_2^N)$ is also asymptotically normal

$$(W_1^N, W_2^N, \tilde{W}_1^N, \tilde{W}_2^N) \xrightarrow{d} (W_1, W_2, \tilde{W}_1, \tilde{W}_2), \quad m, n \to \infty;$$

as $Cov(W_1, \tilde{W}_1) = Cov(W_1, \tilde{W}_2) = Cov(W_2, \tilde{W}_1) = Cov(W_2, \tilde{W}_2) = \mathbf{0}$, (W_1^N, W_2^N) and $(\tilde{W}_1^N, \tilde{W}_2^N)$ are also asymptotically independent. As $H_0^\gamma(\tilde{Z}^N)$ only depends on W_1^N and W_2^N, $H_0^\gamma(\tilde{Z}_\pi^N)$ only depends on $\tilde{W}_1^N$ and $\tilde{W}_2^N$, $H_0^\gamma(\tilde{Z}^N)$ and $H_0^\gamma(\tilde{Z}_\pi^N)$ are also asymptotically independent. Moreover, as (V_1^N, V_2^N) and $(\tilde{V}_1^N, \tilde{V}_2^N)$ are identically distributed, so are $H_0^\gamma(\tilde{Z}^N)$ and $H_0^\gamma(\tilde{Z}_\pi^N)$. In particular, since

$$W_1^N = (1-c)V_1^N - \{c(1-c)\}^{1/2}V_2^N = m^{1/2}n/N(\bar{X}_m - \bar{Y}_n) = (m/N)^{1/2}n/N\delta_N,$$

and $Cov(W_1^N) = (1-c)\Sigma(\bar{P})$, according to Proposition 2.1 (for $\gamma = \infty$) and Propostion 3.2 (for $\gamma = 1$) of Chernozhukov et al. (2017),

$$\sup_{t \in R} \left| P(H_0^\gamma(\tilde{Z}^N) \le t) - P(|N(0, I_p)|_\gamma \le t) \right| \to 0.$$

Next we show that with a random permutation $\pi \in G_N$ independent of Z^N,

$$H_0^{\gamma}(Z_{\pi}^N) - H_0^{\gamma}(\tilde{Z}_{\pi\pi_0}^N) \xrightarrow{p} 0. \tag{37}$$

Let $I_1 \subset \{1, \cdots, m\}$ and $I_2 \subset \{m+1, \cdots, N\}$ are two collection of indices such that $I_1 \cup I_2$ stand for the set of indices where Z_{π}^N and $\tilde{Z}_{\pi\pi_0}^N$ differ. Therefore, $D = \#(I_1) + \#(I_2) = O_p(N^{1/2})$ as we have seen in the proof of Lemma 2.For any $i \in I_1 \cup I_2$, write $\tilde{Z}_{\pi\pi_0(i)}^N = (\tilde{Z}_{i1}, \cdots, \tilde{Z}_{ip})$ and $Z_{\pi(i)}^N = (Z_{i1}, \cdots, Z_{ip})$. Thus for $\gamma = 1$ or $\gamma = \infty$

$$\begin{aligned} |H_0^{\gamma}(Z_{\pi}^N) - H_0^{\gamma}(\tilde{Z}_{\pi\pi_0}^N)| &\leq N^{1/2}\frac{m+n}{mn}\Big|\tilde{\Sigma}^{-1/2}\{\sum_{i\in I_1}(\tilde{Z}_i - Z_i) - \sum_{i\in I_2}(\tilde{Z}_i - Z_i)\}\Big|_{\gamma} \\ &\leq N^{1/2}\frac{|m-n|}{mn}|\tilde{\Sigma}^{-1/2}|_{(1,1)}\Big|\sum_{i\in I_1}(\tilde{Z}_i - Z_i) - \sum_{i\in I_2}(\tilde{Z}_i - Z_i)\Big|_{\infty}. \end{aligned} \tag{38}$$

As $|\tilde{\Sigma}^{-1/2}|_{(1,1)}$ is finite, there exists some constant $C > 0$, that for any $\epsilon > 0$,

$$\begin{aligned} P\Big(|H_0^{\gamma}(Z_{\pi}^N) - H_0^{\gamma}(\tilde{Z}_{\pi\pi_0}^N)| > \epsilon\Big) &\leq P\Big(|\sum_{i\in I_1}(\tilde{Z}_i - Z_i)|_{\infty} > CN^{1/2}\epsilon\Big) \\ &+P\Big(|\sum_{i\in I_2}(\tilde{Z}_i - Z_i)|_{\infty} > CN^{1/2}\epsilon\Big). \end{aligned} \tag{39}$$

If D is finite, then a simple application of Chebyshev inequality and Proposition 2,

$$P\Big(|\sum_{i\in I_1}(\tilde{Z}_i - Z_i)|_{\infty} > CN^{1/2}\epsilon\Big) \leq C(\log p/N)^{1/2}\epsilon^{-1} \to 0,$$

under condition that $\log p = o(n^{\alpha})$ with $\alpha \leq 1/6$. If, $\#(I_1)$ is infinite, we could still apply Proposition 2.1 of Chernozhukov et al. (2017), to show that

$$\sup_t \Big| P\Big(\#(I_1)^{-1/2} |\sum_{i\in I_1}(\tilde{Z}_i - Z_i)|_\infty > t\Big) - P\Big(|\Xi|_\infty > t\Big)\Big| \to 0, \tag{40}$$

where Ξ is zero-mean p-dim normal distribution with finite variance (matrix), so that $E|\Xi|_\infty = O((\log p)^{1/2})$. Since $\#(I_1) = O_p(N^{1/2})$,

$$\begin{aligned} P\Big(|\sum_{i\in I_1}(\tilde{Z}_i - Z_i)|_\infty > CN^{1/2}\epsilon\Big) &\leq P\Big(|\Xi|_\infty > CN^{1/4}\epsilon\Big) + o(1) \\ &\leq C(\log p/N^{1/2})^{1/2}\epsilon^{-1} \to 0. \end{aligned} \tag{41}$$

Equation (37) thus follows from (38),(39), (40) and (41). □

0.3. Proof of Theorem 2

The proof consists of two steps in a way similar to that Theorem 1 follows directly from Lemma 1 and Lemma 2. The only minor change is that the condition on the consistency of $\tilde{\Omega}_N$ is replaced with (a.3). The details of proof is thus omitted. □

0.4. Proof of Theorem 3

Derivation of the (limiting) null distribution is done in exactly the same manner as in Lemma 1. For the un-permuted observations, $\bar{X}^* - \bar{Y}^* \equiv \bar{X}_m - \bar{Y}_n$, with or without μ^X in (2.9) replaced with $\bar{X}_m$. The covariance matrix of

δ_n^* is thus given by $(1-c)/c\Sigma^X + \Sigma^Y$. Also simple algebra could verify that under (a.3) of (C6) in Section 6, an analogue for $\{v_{n,k}^*\}^2$ as a consistent estimator of $Var(\delta_{n,k}^*)$ also holds. Therefore, the limiting distribution of $S_1^\gamma(Z^n)$ of (2.12) is the same as given in (2.8) with the covariance matrix given by the correlation matrix associated with $(1-c)/c\Sigma^X+\Sigma^Y$. In the case of $S_1^\gamma(Z^n)$, the corresponding covariance matrix is simply $(1-c)/c\Sigma^X+\Sigma^Y$.

The derivation of the permutation distribution could be done in a way identical to how Lemma 2 is proved after we have clarified the corresponding coupling procedure. We recycle some notations used in the proof of Lemma 2 . We only consider the case where $v = c/(1-c)-K > 0$, for it would only be easier to deal with $v = 0$. The coupling procedure largely follows the routine outlined in Section 0.1, only an extra layer of sampling is involved whenever an index of zero is drawn. When this happens, if the original pseudo observations $\{X_i^*, i = 1, \cdots, n\}$ have not all been used up, then set $\tilde{Z}_i$ to be the next X_i^* in line; otherwise, draw a random value u from the uniform $(0, 1)$. If $u < v$, we first obtain a random sample of $K + 1$ observations of Xs from $P_1(.)$, their sum multiplied by $(1 - c)/c$ gives a new $\tilde{Z}_i$; if $u > v$, then a random sample of K observations from $P_1(.)$ is used, and a new $\tilde{Z}_i$ is then given by the sample sum multiplied by $(1-c)/c$. In the end, $\tilde{Z}^n = (\tilde{Z}_1, \cdots, \tilde{Z}_{2n})$ can be seen as IID observations from the mixture distribution $\tilde{P} = 0.5P_1^*(\cdot,.) + 0.5P_2(\cdot,.) \equiv 0.5vP_{11}(\cdot,.) + 0.5(1 - v)P_{12}(\cdot,\cdot) + 0.5P_2(\cdot,\cdot)$, where $P_{11}(\cdot,\cdot)$ is the sum of $K+1$ i.i.d. observations from $P_1(\cdot,\cdot)$ multiplied by $(1-c)/c$, while $P_{11}(\cdot,\cdot)$ is the sum of K i.i.d. observations from $P_1(\cdot,\cdot)$ again multiplied by $(1-c)/c$. Z^n and $\tilde{Z}^n$ also have many of the observations in common. After a reordering procedure of $\tilde{Z}^n$ by a permutation π_0 similar to that described in the proof of Lemma 2, only now the reordering is applied to three 'populations', $\tilde{Z}^n_{\pi_0}$ should agree with Z^n at many places; regarding D, the number of spots where $\tilde{Z}^n_{\pi_0}$ and Z^n differ, we again have $E(D) \le n^{1/2}$. □

0.5. Proof of Theorem 4

Denote the limit on the RHS of (3.14) by b_0 which could be infinite, and with a slight abuse of notation, write $\delta^*_{n,k} = n^{1/2}(\bar{X}^* - \bar{Y}^* - \delta_0)$ and denote the diagonal entries of $(1-c)/c\Sigma^X + \Sigma^Y$ by a_k^2, $k = 1, \cdots, p$. We have, based on Theorem 3 and Proposition 3,

$$P\Big(\max_{1\le k\le p}\Big|\frac{\delta^*_{n,k}}{a_k}\Big| \le \sqrt{2\ln p - \ln(\ln p) + x}\Big) \to F(x) \qquad (42)$$

as n, $p \to \infty$. Let A be the event that there exists some $k \in s_0$, such that $|\delta_{n,k}/v_{n,k}| < q_{\alpha_n}$. Since $\{|\delta_{n,k}/v_{n,k}| < q_{\alpha_n}\} \subset \{|\delta^*_{n,k}/v_{n,k}| > n^{1/2}|\delta_{0k}/v_{n,k}| - q_{\alpha_n}\}$ and for any $\epsilon > 0$, $\{|\delta^*_{n,k}/v_{n,k}| > t\} \subset \{|\delta^*_{n,k}/v_{n,k}| > t - \epsilon\} \cup \{|(a_k -$

$v_{n,k})\delta^*_{n,k}/a_k| > (s_0/c)^{1/2}\epsilon/2\}$, we have for any $M > 0$, and any $0 < r < 1$,

$$\begin{aligned}
P(A) &\le P\Big(\max_{k\in I_0}|\delta^*_{n,k}/v_{n,k}| > n^{1/2}\min_{k\in I_0}|\delta_{0k}/v_{n,k}| - q_{\alpha_n}\Big)\\
&\le P\Big(\max_{k=1,\cdots,p}|\delta^*_{n,k}/a_k| > r(c/s_1)^{1/2}n^{1/2}\min_{k\in I_0}|\delta_{0k}| - q_{\alpha_n} - \epsilon\Big)\\
&\quad + P\Big(\sup_k|\delta^*_{n,k}/a_k| > M\epsilon\ln p\Big) = I + II,
\end{aligned}$$

where we have used the fact that $s_0/c \le a_k^2 \le s_1/c$ and that $|v_{n,k} - a_k| = o_p(1/\log p)$ for all $k = 1, \cdots, p$. Set $\epsilon = (\ln p)^{-1/2}$, then by (42), $II \to 0$. With this ϵ,

$$I \le P\Big(\max_{k=1,\cdots,p}|\delta^*_{n,k}/a_k| \ge 2(2\ln p)^{1/2} - q_{\alpha_n} - \epsilon\Big) \le \exp(-(\pi\ln p)^{-1/2}) \to 0;$$

if α is chosen so that $q_\alpha/(2\ln p)^{1/2} \to 1$. On the other hand, with the same ϵ above,

$$\begin{aligned}
P\Big(\max_{k\notin I_0}|\delta^*_{n,k}/v^*_{n,k}| \ \ \ge\ \ q_{\alpha_n}\Big) &\le P\Big(\max_{k\notin I_0}|\delta^*_{n,k}/a_k| \ge q_{\alpha_n} - \epsilon\Big)\\
&\quad + P\Big(\sup_k|\delta^*_{n,k}/a_k| > M\epsilon\ln p\Big),
\end{aligned}$$

where, according to (42), both terms on the RHS are zero. This finishes the proof. $\square$

0.6. More analysis results of WTCCC dataset

Using the same quality control as described in Section 6, we have 1,969 cases and 2,992 controls over 304,279 SNPs for BPD, 1,979 cases and 2,992 controls over 306,030 SNPs for HT, and 1,952 cases and 2,992 controls over 307,089 SNPs for RA. We applied the permutation test with S_1^∞ to the data, and the resulting Manhattan plots are shown in Figure 3. The analysis for each disease can be done around 16 minutes on a Windows console with 2.30GHz intel Xeron CPU E5-2697.

With significance level 0.01, we summarize our findings as follows.For BPD, our method S_1^∞ identifies two SNPs from genes *UBR1* and *SLC35F4*, respectively, where *SLC35F4* was reported to be associated with BPD previously (WTCCC, 2007). For RA, S_1^∞ identified 148 SNPs. Among them, 98 SNPs are within 42 gene regions, where 21 genes were reported to be associated with rheumatoid arthritis in previous studies (Raychaudhuri et al., 2008; Hu et al., 2008; Jiang et al., 2015). Note that we are unable to identify significant SNPs associated with HT. The possible reason is that the control population was not screened to remove individuals with hypertension (Doris , 2011).

In the analysis of these three diseases using WTCCC data, we identified many "new" SNPs which were not reported in the original study of

WTCCC (2007), but were detected in later studies. These SNPs and their corresponding studies are listed in Table S1.

Table S1: SNPs identified by both the proposed permutation test and other recent studies, but not by WTCCC (2007).

Assoc. trait	Rs	Base pair position	Chrom.	Gene	studies that detected the same genes: studies	sample size
CAD	rs564398	22029548	9	AL359922.1	Kulminski et al. (2018)	33,431
CAD	rs7865618	22031006	9	AL359922.1 CDKN2B-AS1	Wild et al. (2011)	5,031
CAD	rs4977574	22098575	9	CDKN2B-AS1	C4D Genetics Consortium (2011)	30,472
CAD	rs2891168	22098620	9	CDKN2B-AS1	Nelson et al. (2017)	63,731
CAD	rs4977574	22098575	9	CDKN2B-AS1	van der Harst and Verweij (2018)	547,261
CAD	rs1333042	22103814	9	CDKN2B-AS1	Lee et al. (2013)	5,714
CAD	rs1333048	22125348	9	CDKN2B-AS1	Slavin et al. (2011)	5,000
CD	rs2201841	67228519	1	C1orf141 IL23R	Raelson et al. (2007)	1,146
CD	rs6752107	233252802	2	ATG16L1	de Lange et al. (2017)	40,266
CD	rs3828309	233271764	2	ATG16L1	Barrett et al. (2008)	9,541
CD	rs9292777	40437846	5	AC093277.1	Parkes et al. (2007)	4,686
RA	rs805297	31654829	6	APOM	Hu et al. (2008)	700
RA	rs9268557	32421528	6	BTNL2	Jiang et al. (2015)	2,573
RA	rs6457620	32696222	6	HLA-DQB1	Raychaudhuri et al. (2008)	15,853
RA	rs9784858	32819398	6	AL669918.1	Jiang et al. (2015)	2,573
RA	rs10484565	32827255	6	AL669918.1	Jiang et al. (2015)	2,573
T1D	rs2292239	56088396	12	ERBB3	Cooper et al. (2008)	8,207
T1D	rs9268645	32440750	6	HLA-DRA	Barrett et al. (2009)	16,559
T1D	rs9268853	32461866	6	HLA-DRB9	Kawabata et al. (2019)	676
T2D	rs7901695	112994329	10	TCF7L2	Zeggini et al. (2007)	4,862
T2D	rs8050136	53782363	16	FTO	Zhao et al. (2017)	183,651

References

Barrett, J. C., Hansoul, S. and others (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, **40**, 955.

Barrett, J. C., Clayton, D. G. and others (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, **41**, 703.

Cai, T. T., Liu, W., and Xia, Y. (2014) Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B*, **76**, 349-372.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015) Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, **162**, 47-70.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017) Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45**, 2309-2352.

Chung, E. and Romano, J. P. (2013) Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41**, 484-507.

Coronary Artery Disease (C4D) Genetics Consortium and others (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genetics*, **43**, 339.

Cooper, J. D. and Smyth, D. J. and others (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, **40**, 1399.

de Lange, K. M. and Moutsianas, L. and others (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, **49**, 256.

Doris, Peter (2011) The genetics of blood pressure and hypertension: the role of rare variation. *Cardiovascular therapeutics*, **29**, 37-45.

Hu, H-J and Jin, E-H. and others (2011) Common variants at the promoter region of the APOM confer a risk of rheumatoid arthritis. *Experimental & Molecular Medicine*, **43**, 613.

Jiang, X., Källberg, H. and others (2015) An Immunochip-based interaction study of contrasting interaction effects with smoking in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Rheumatology*, **55**, 149-155.

Kawabata, Y. and Nishida, N. and others (2019) Genome-Wide Association Study Confirming a Strong Effect of HLA and Identifying Variants in CSAD/lnc-ITGB7-1 on Chromosome 12q13. 13 Associated With Susceptibility to Fulminant Type 1 Diabetes. *Diabetes*, **68**, 665–675.

Kosorok, M. and Ma, S. (2007) Marginal asymptotics for the large p, small n paradigm: with applications to microarray data. *The Annals of Statistics*, **35**, 1456-1486.

Kulminski, A. M. and Huang, J. and others (2018) Strong impact of natural-selection–free heterogeneity in genetics of age-related phenotypes. *Aging (Albany NY)*, **10**, 492.

Lee, J.-Y. and Lee, B. and others (2013) A genome-wide association study of a coronary artery disease risk variant. *Journal of Human Genetics*, **58**, 120.

Nelson, C. P. and Goel, A. and others (2017) Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Statistica Sinica*, **49**, 1385.

Parkes, M. and Barrett, J. C. and others (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics*, **39**, 830.

Raelson, J. V. and Little, R. D. and others (2007) Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proceedings of the National Academy of Sciences*, **104**, 14747–14752.

Raychaudhuri, S. and Remmers, E. and others (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nature genetics*, **40**, 1216.

Slavin, T. P., Feng, T. and others (2011) Two-marker association tests yield new disease associations for coronary artery disease and hypertension. *Human Genetics*, **130**, 725–733.

van der Harst, P. and Verweij, N. (2018) Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, **122**, 433–443.

Wild, P. S. and Zeller, T. and others (2011) A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. *Circulation: Cardiovascular Genetics*, **4**, 403–412.

Wellcome Trust Case Control Consortium and others (2011) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661.

Talagrand, M. (2003) *Spin Glasses: A Challenge for Mathematicians.* Springer.

Zeggini, E. and Weedon, M. N. and others (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.

Zhao, W. and Rasheed, A. and others (2017) Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature Genetics*, **49**, 1450.