Supplementary Method:

Mathematical proof of the empirical null distribution of $\widehat{\mathrm{MR}}_{\mathrm{M\times M}}(i,j)$

Lemma 1:

For the $\mathrm{M}\times\mathrm{M}$ randomly assigned matrix C, $\mathrm{Rank}(i\to j)\perp Rank(j\to i)|\mathrm{C}(i,j)$.

Proof: With given $\mathrm{C}(i,j)$, $\mathrm{Rank}(i\to j)$ only depends on $\mathrm{C}(i,\cdot),\cdot\neq j$ while $Rank(j\to i)$ only depends on $\mathrm{C}(\cdot,j),\ \cdot\neq i$. Since $\mathrm{C}(i,\cdot)\perp\mathrm{C}(\cdot,j)$ for $i,\cdot\neq i,j$ and $\cdot,j!=i,j$, we have $\mathrm{Rank}(i\to j)\perp Rank(j\to i)|\mathrm{C}(i,j)$ □

Proposition 1:

For the $\mathrm{M}\times\mathrm{M}$ matrix randomly assigned from the distribution with pdf f and cdf F, the joint distribution of $Rank(i\to j)\ and\ Rank(j\to i)$ is:

$$\mathrm{P}(Rank(i\to j)=x,Rank(j\to i)=y)=$$
$$\int\binom{M-1}{x-1}((1-F(k))^{x-1}F(k)^{M-x}\cdot\binom{M-1}{y-1}((1-F(k))^{y-1}F(k)^{M-y}\cdot dF(k)$$

Proof:

$$Prob(Rank(j\to i)=x|\mathrm{C}(i,j)=k)=$$
$$Prob\left(\mathrm{C}_{i(j_m)}>k,\mathrm{C}_{i(j_n)}<K,m=1\dots x-1,n=x+1\dots N\right)=$$
$$C_{N-1}^{x-1}(1-F(k))^{x-1}F(k)^{N-x}$$
$$,$$
$$,in\ which\ j_m=1\dots j-1\ and\ j_n=j+1\dots N$$
$$,and\ \mathrm{C}_{i(j)}\ is\ the\ jth\ value\ in\ \mathrm{C}(i,\cdot)\ in\ decreasing\ order$$

By Lemma 1,

$$P(Rank(i\to j),Rank(j\to i)|\mathrm{C}(i,j))=$$
$$P(Rank(i\to j)|\mathrm{C}(i,j))*P(Rank(j\to i)|\mathrm{C}(i,j))$$

$$\mathrm{P}(Rank(i\to j)=x,Rank(j\to i)=y|\mathrm{C}(i,j)=k)=$$
$$(C_{N-1}^{x-1}(1-F(k))^{x-1}F(k)^{N-x})\cdot(C_{N-1}^{y-1}(1-F(k))^{y-1}F(k)^{N-y})$$

Hence

$$\mathrm{P}(Rank(i\to j)=x,Rank(j\to i)=y)=$$
$$\int\binom{M-1}{x-1}((1-F(k))^{x-1}F(k)^{M-x}\cdot\binom{M-1}{y-1}((1-F(k))^{y-1}F(k)^{M-y}\cdot dF(k)$$

□

Proposition 2:

The empirical null distribution of $\widehat{\mathrm{MR}}_{\mathrm{M\times M}}(i,j)$ does not depend on the empirical distribution of $\mathrm{C}(i,j)$

Proof: Since $\mathrm{C}(i,j)$ follows the pdf f, by Proposition 1:

$$\mathrm{P}(Rank(i\to j)=x,Rank(j\to i)=y|\mathrm{C}(i,j)=k)=$$
$$(C_{N-1}^{x-1}(1-F(k))^{x-1}F(k)^{N-x})\cdot(C_{N-1}^{y-1}(1-F(k))^{y-1}F(k)^{N-y})$$

Hence the $\mathrm{MR(i,j)}^2=Rank(i\to j)\cdot Rank(j\to i)$ with given $\mathrm{C}(i,j)=k$ follows a multiplication of two identical binomial distributions with n= N-1 and P=1-F(S_{ij}) ~U(0,1), which does not depend on the empirical distribution of $\mathrm{C}(i,j)$. □

To generate an empirical null distribution of $\widehat{\mathrm{MR}}_{\mathrm{M\times M}}(i,j)$, we can randomly draw $\tilde{\mathrm{p}}$ from

U(0,1) and generate $Rank(i \to j)$ and $Rank(j \to i)$ from two random numbers $X_{i,j,1}\ and\ X_{i,j,2}$ from the binomial distribution $\mathrm{Bin}(\tilde{\mathrm{p}}, \mathrm{n}-1)$ and $\widehat{\mathrm{MR}}(i,j)' = \sqrt{(X_{i,j,1}+1)\cdot(X_{i,j,2}+1)}$. Hence the $\mathrm{P}(Rank(i \to j) = x, Rank(j \to i) = y)$.

Theorem 1:

Empirical null distribution of $\mathrm{MR(i,j)}$ can be simulated by

$$MR_{i,j} = \sqrt{(X_{i,j,1}+1)\cdot(X_{i,j,2}+1)}$$

$$X_{i,j,1} \sim Binom\left(n-2, \mathrm{p_{ij}}\right), X_{i,j,2} \sim Binom\left(n-2, \mathrm{p_{ij}}\right)$$

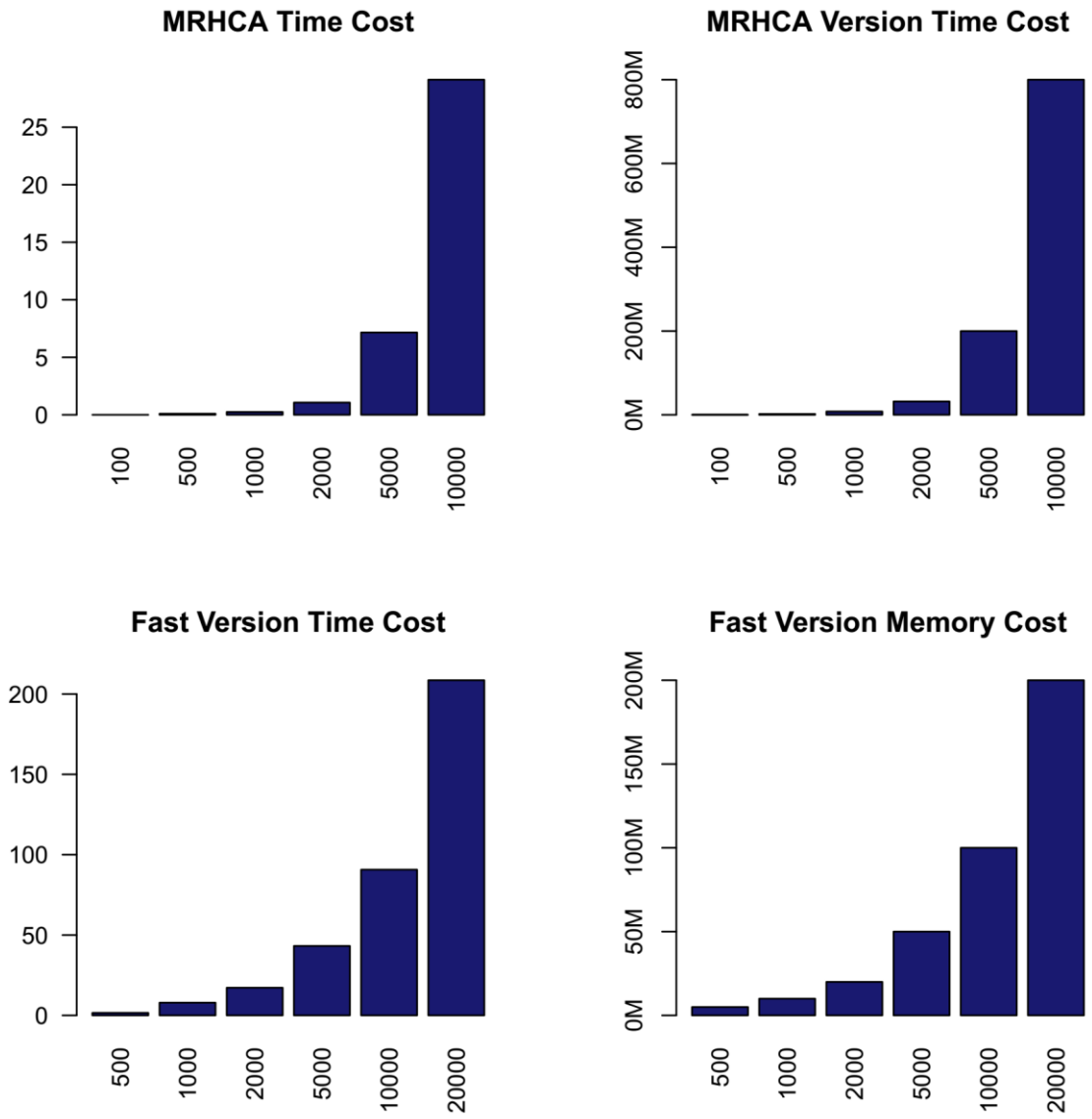
$$\mathrm{p_{ij}} \sim U(0,1)$$

Proof:

By Proposition 1 and 2, and the discussion above, the random numbers simulated as described in the theorem form empirical null distribution of $\widehat{\mathrm{MR}}_{\mathrm{M\times M}}(i,j)$. □

Memory and Computational cost analysis:

We have analyzed the memory and computational cost of MRHCA and the fast version over randomized gene expression profile of 100, 500, 1000, 2000, 5000, 10000 and 20000 genes with 300 samples. The analyses were conducted by R x64 3.4.0 on a desktop with Intel Core i7-7700K CPU and 16.0GB RAM. The memory and algorithm running time are shown in the figure below. The loops in this fast version of the Rank_Index generation was not accelerated by a C library. Hence the time cost of the fast version is more than the MRHCA algorithm. See more details on https://github.com/zy26/mrct.



The Figures show the time (s) and memory (M) costs (y-axis) of the MRHCA and the fast version on simulated data with number of genes shown on x-axis.