Supplementary Materials: Proofs

Proposition 1

Let there be n genes with independent bivariate ranks (R_g^x, R_g^y) and maximum rank statistics $M_g,\ g = 1, \ldots, n$. Define the random variable:

$$W_n(i) = \begin{cases} 2 & \text{if } i \text{ is twice a maximum rank} \\ 1 & \text{if } i \text{ is a unique maximum} \\ 0 & \text{if } i \text{ is not a maximum rank} \end{cases} \tag{1}$$

We calculate $E[W_n(i)]$ because $P(M_g = i) = E[W_n(i)]/n$.

Consider the probability mass function of $W_n(i)$.

$$\begin{aligned} P(W_n(i) = 2) &= P\left(\exists g : (R_g^x, R_g^y) \in \{(l, i) : l < i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(i, l) : l < i\}\right) \\ &= \frac{i-1}{n} \cdot \frac{i-1}{n-1} \\ &= \frac{(i-1)^2}{n(n-1)} \end{aligned} \tag{2}$$

$$\begin{aligned} P(W_n(i) = 1) &= P\left(\exists g : (R_g^x, R_g^y) = (i, i)\right) \\ &\quad + P\left(\exists g : (R_g^x, R_g^y) \in \{(l, i) : l < i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(m, i) : m > i\}\right) \\ &\quad + P\left(\exists g : (R_g^x, R_g^y) \in \{(l, i) : l > i\} \text{ and } \exists h : (R_h^x, R_h^y) \in \{(m, i) : m < i\}\right) \\ &= \frac{1}{n} + \frac{i-1}{n} \cdot \frac{n-i}{n-1} + \frac{i-1}{n} \cdot \frac{n-i}{n-1} \\ &= \frac{-n-1+2(n+1)i-2i^2}{n(n-1)} \end{aligned} \tag{3}$$

$$\begin{aligned} P(Wn(i) = 0) &= 1 - P(W_n(i) = 2) = P(W_n(i) = 1) \\ &= \frac{n^2 + i^2 - 2ni}{n(n-1)} \end{aligned} \tag{4}$$

Then the expectation of $W_n(i)$ can be determined under the null.

$$\begin{aligned} E(W_n(i)) &= 0 \cdot P(W_n(i) = 0) + 1 \cdot P(W_n(i) = 1) + 2 \cdot P(W_n(i) = 2) \\ &= \frac{-n-1+2(n+1)i-2i^2}{n(n-1)} + \frac{2(i-1)^2}{n(n-1)} \\ &= \frac{2i-1}{n} \end{aligned} \tag{5}$$

Thus the marginal probability mass function is known.

$$P(M_g = i) = \frac{1}{n}E(W_n(i)) = \begin{cases} \frac{2i-1}{n^2} & \text{for } i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The marginal pmf for M_h/n is thus

$$f_{n,0}(i/n) = P(M_h = i) = \begin{cases} \frac{2i-1}{n^2} & \text{for } i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Corollary 1

Assume (I1), (I2), and (I3). Further assume $\pi_1 \in (0,1)$ is fixed, and let $j_{\pi_1} = \lfloor n\pi_1 \rfloor$. Then the marginal pmf of M_h/n for an irreproducible gene h is $f_{n,\pi_1}(i/n) = f_{n-j_{\pi_1},0}(i/n - j_{\pi_1})$:

$$f_{n,\pi_1}(i/n) = \begin{cases} \frac{2(i-j_{\pi_1})-1}{(n-j_{\pi_1})^2} & \pi_1 < i/n \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Corollary 2

Let $\pi_1 \in (0,1)$, $x \in (\pi_1, 1)$, and $i_x = \lfloor nx \rfloor$. Then we can derive the marginal cumulative distribution function of M_h/n in the ideal setting:

$$\begin{aligned}
F_{n,\pi_0}(x) &= P(M_h/n \leq x) \\
&= P(\pi_1 < M_h/n \leq x) = \sum_{i=j_{\pi_1}}^{k_x} P(M_h = i) \\
&= \sum_{i=j_{\pi_1}+1}^{k_x} \frac{2(i-j_{\pi_1})-1}{(n-j_{\pi_1})^2} = \frac{1}{(n-j_{\pi_1})^2}\left(2\sum_{i=j_{\pi_1}+1}^{k_x}(i-j_{\pi_1}) - \sum_{i=j_{\pi_1}+1}^{k_x} 1\right) \\
&= \frac{1}{(n-j_{\pi_1})^2}\left(2\frac{(k_x-j_{\pi_1})(k_x-j_{\pi_1}+1)}{2} - (k_x-j_{\pi_1})\right) \\
&= \frac{1}{(n-j_{\pi_1})^2}(k_x-j_{\pi_1})(k_x-j_{\pi_1}+1-1) \\
&= \frac{(k_x-j_{\pi_1})^2}{(n-j_{\pi_1})^2}
\end{aligned}$$

Theorem 1

Let $\pi_1 \in (0,1)$, and $x \in (0,1)$ be fixed, and assume (I1), (I2), and (I3). For a fixed n, let $k_x = \lfloor nx \rfloor$, and $j_{\pi_1} = \lfloor n\pi_1 \rfloor$. Then as n tends to infinity:

$$\lim_{n\to\infty} j_{\pi_1}/n \to \pi_1, \text{ and } \lim_{n\to\infty} k_x \to x.$$

Thus the limiting cumulative distribution function can be derived.

$$\begin{aligned}
F_{n,\pi_1}(x) &= \frac{(k_x - j_{\pi_1})^2}{(n - j_{\pi_1})^2} = \frac{(k_x - j_{\pi_1})^2/n^2}{(n - j_{\pi_1})^2/n^2} \\
&= \frac{(k_x/n - j_{\pi_1}/n)^2}{(n/n - j_{\pi_1}/n)^2} \\
&\to \frac{(x - \pi_1)^2}{(1 - \pi_1)^2}
\end{aligned}$$

Theorem 2

Assume π_1 is fixed, $\hat{\pi}_1$, $\hat{S}_n(x)$, $S_n(x)$, and $MSE_n(\lambda)$ are as defined in the manuscript. We define the quantity

$$SS(\lambda) = (1-\lambda)^{-1} \int_\lambda^1 \left(\hat{S}_n(x) - (1-\lambda)S_\lambda(x)\right)^2 dx$$

and show that $\arg\min\{MSE_n(\lambda)\} = \arg\min\{SS(\lambda)\}$. Then the proof outline to show $\hat{\pi}_1 \to \pi_1$ is as follows:

1. Show that the random variables M_h are *absolutely regular.*

2. Use result from Nobel and Dembo (1993) and apply the Glivenko-Cantelli theorem to show that $SS(\pi_1) \to 0$ as $n \to \infty$.

3. Show $SS(\lambda) \nrightarrow 0$ for $\lambda \neq \pi_1$.

4. Use (2) and (3) to prove consistency.

We now proceed with the proof outlined above.

$$
\begin{aligned}
SS(i/n) &= (1-i/n)^{-1}\int_{i/n}^{1}\left(\hat{S}_n(x)-(1-i/n)S_{i/n}(x)\right)^2 dx \\
&\approx (1-i/n)^{-1}\sum_{x=i}^{n}\left(\hat{S}_n(x/n)-(1-i/n)S_{i/n}(x/n)\right)^2 \\
&= n\frac{1}{n-1}\sum_{x=i}^{n}\left(\hat{S}_n(x/n)-(1-i/n)S_{i/n}(x/n)\right)^2 \\
&= n\cdot MSE(i/n) \\
\Rightarrow \arg\min_i \{MSE(i/n)\} &= \arg\min_i \{SS(i/n)\}
\end{aligned}
$$

1. Following Nobel and Dembo (1993), a sequence of random variables $X_1,\ldots,X_N$ is *absolutely regular* if the dependence coefficients $\beta(n),\ n=1,2\ldots$ go to 0, where

   $$\beta(n)=\sup_{A\in\sigma(X_1,\ldots,X_k,X_{n+1+1},\ldots)}\left|P(X)-P_0^k(X)P_{n+k+1}^{\infty}(A)\right|.$$

   Here, P is based on the full joint distribution, and $P_1^{\infty}, P_{-\infty}^{0}$ are joint distributions for $X_{-\infty},\ldots,X_0$ and $X_1\ldots,X_\infty$ respectively and independently. Because the M_h have possible values $1,\ldots,N$, we must let n and N go to ∞ together. Thus set $N=n^2$ and consider the asymptotic behavior of the dependence coefficients below:

   $$
   \begin{aligned}
   \beta(n) &= \sup_{A\in\sigma(M_1,\ldots,M_k,M_{n+1},\ldots,M_N)}\left|P(A)-P_0^k(A)P_{n+k+1}^{N}(A)\right| \\
   &\leq |1-P((A_1,\ldots,M_k,M_{n+1},\ldots,M_N)=(1,2,\ldots,N-n))| \\
   &= \left|1-\sum_{i=1}^{N-n}P(M_h=i)\right| = \left|1-\frac{(N-n)^2}{N^2}\right| \\
   &= \left|1-(1+\frac{n^2-2n}{N^2})\right| \\
   &= \frac{n^2-2n}{N^2} \\
   &\to 0
   \end{aligned}
   $$

   Because $\beta(n)\to 0$, we see that the M_h are absolutely regular.

2. By Novel and Dembo (Nobel and Dembo, 1993), the Glivenko-Cantelli theorem holds for M_h/n using the marginal distribution. Using the result from Theorem 1 in combination with the

Glivenko-Cantelli theorem the following holds:

$$\begin{aligned}
&\sup_{x\in(\pi_1,1)} |\hat{S}_n(x) - \pi_0 S_{\pi_0}(x)| \overset{a.s.}{\to} 0 \\
\Rightarrow &\int_{\pi_1}^{1} (\hat{S}_n(x) - \pi_0 S_{\pi_0}(x))^2\,dx \overset{a.s.}{\to} 0 \\
\Rightarrow &SS(\pi_1) \to 0.
\end{aligned}$$

3. For an ideal setting where the proportion of reproducible signals is $\pi_1^* \neq \pi_1$,

$$\hat{S}_n(x) \to S_{\pi_1^*}(x)$$

and $\exists x \in (\pi_1, 1)$ such that $S_{\pi_1}(x) \neq S_{\pi_1^*}(x)$. Because $S_{\pi_1}(x),\ S_{\pi_1^*}(x)$ are continuous,

$$\int_{\pi_1}^{1} (\hat{S}_n(x) - \pi_0 S_{\pi_0}(x))^2\,dx \nrightarrow 0 \ \Rightarrow SS(\pi_1) \neq 0$$

4. Thus by (1), (2), (3) above,

$$\hat{\pi}_1 = \arg\inf_{\lambda\in(0,1)} \{SS(\lambda)\} \to \pi_1$$

Justification of assertion $E[\hat{\pi}_1] \leq \pi_1$

Assume the proportion of reproducible signals is π_1, and that of irreproducible signals is $\pi_0 = 1 - \pi_1$. Additionally assume:

(R1) Reproducible signals *tend* to be ranked higher than irreproducible signals. Thus if gene g is reproducible and gene h is irreproducible,

$$P(R_g^x < R_h^x) > 1/2, \text{ and } P(R_g^y < R_h^y) > 1/2).$$

(I2) The correlation between the ranks of reproducible signals is non-negative. Thus for any reproducible gene g,

$$Cor(R_g^x, R_g^y) \geq 0$$

(I3) The correlation between ranks of irreproducible signals is 0. Thus for any irreproducible gene h,

$$Cor(R_h^x, R_h^y) = 0$$

The justification hinges on the following:

1. If $\lambda_1 \leq \lambda_2$ then $(1-\lambda_1)S_{\lambda_1}(x) \geq (1-\lambda_2)S_{\lambda_2}(x)$ $\forall x \in (\lambda_2, 1)$.

2. $E(\hat{S}_n(x)) \geq (1-\pi_1)S_{\pi_1}(x)$ for $x \in (\pi_1, 1)$.

Based on these two statements, it is clear that $\lambda^* \leq \pi_1$ exists such that

$$E[MSE_n(\lambda^*)] \leq E[MSE_n(\pi_1)],$$

and thus based on the definition of $\hat{\pi_1}$ as the argument that gives the first local minimum of the $MSE_n(\lambda)$, $E[\hat{\pi_1}] \leq \pi_1$.

To prove (1) assume $\lambda_1 < \lambda_2$. Then consider the quantity

$$\begin{aligned}
(1-\lambda_1)f_{\lambda_1}(x) - (1-\lambda_2)f_{\lambda_2(x)} &= \frac{2(x-\lambda_1)}{1-\lambda_1} - \frac{2(x-\lambda_2)}{1-\lambda_2} \\
&= \frac{2}{(1-\lambda_1)(1-\lambda_2)}((x-\lambda_1)(1-\lambda_2) - (x-\lambda_2)(1-\lambda_1)) \\
&= \frac{2}{(1-\lambda_1)(1-\lambda_2)} \cdot (\lambda_1 - \lambda_2) \cdot (x-1) \\
&= (+) \cdot (-) \cdot (-) \\
&> 0
\end{aligned}$$

Thus

$$\begin{aligned}
(1-\lambda_1)S_{\lambda_1}(x) - (1-\lambda_2)S_{\lambda_2}(x) &= \int_x^1 (1-\lambda_1)f_{\lambda_1}(x) - (1-\lambda_2)f_{\lambda_2(x)}\, dx \\
&> \int_x^1 0\, dx \\
&= 0
\end{aligned}$$

Therefore

$$(1-\lambda_1)S_{\lambda_1}(x) > (1-\lambda_2)S_{\lambda_2}(x)$$

To justify (2):

Note that we use the term 'justify' in place of 'prove'. This substitution is intentional, as a formal proof would require further assumptions about the distribution of reproducible signals. In this justification we outline why the claim that $E(\hat{S}_n(x)) \geq (1-\pi_1)S_{\pi_1}(x)$ for $x \in (\pi_1, 1)$ is reasonable. In the discussion below, let $S(x)$ be the underlying survival function for some mixture distribution

$\pi_1 g(x) + (1-\pi_1) f(x)$, where $g,\ f$ give the marginal distributions of M_g/n for reproducible and irreproducible signals respectively.

Consider the appropriately weighted theoretical survival function dependent on π_1, evaluated at $\pi_1 :\ (1-\pi_1)S_{\pi_1}(\pi_1) = 1-\pi_1$. In the idealistic setting, the survival function must go through the point $(\pi_1, 1-\pi_1)$. In the realistic setting, however, $P(M_g/n > \pi_1) > 0$. Thus $S(\pi_1) > (1-\pi_1)S_{\pi_1}(\pi_1)$. This fact places the expected empirical survival curve above the theoretical survival curve at the key input π_1. Further, as the effect size decreases, $P(M_g/n > \pi_1)$ increases, in turn increasing the difference between $S(\pi_1)$ and $(1-\pi_1)S_{\pi_1}(\pi_1)$.

The same reasoning holds for $x > \pi_1$. For any effect size $P(M_g/n > x) > 0$, inflating $S(x)$ above $(1-\pi_1)S_{\pi_1}(x)$. For small effect sizes this is particularly noticeable. For this reason, survival curves from data sets with smaller effect sizes more closely resemble theoretical curves calculated using a smaller reproducible component.

Outline of derivation for higher-order procedure based on Maximum of three ranks

Assume three independently ranked experiments for the same n signals. Further assume that all signals are independent (completely irreproducible). Thus for each signal g, there are three ranks: $M_g^x,\ M_g^y,\ M_g^z$. Define the maximum rank as

$$M_g^{(3)} = \max\{M_g^x, M_g^y, M_g^z\}.$$

Define the random variable $W_n(i)^{(3)}$ in similar fashion to $W_n(i)$ in Proposition 1:

$$W_n^{(3)}(i) = \begin{cases} 3 & \text{if } i \text{ is thrice a maximum rank} \\ 2 & \text{if } i \text{ is twice a maximum rank} \\ 1 & \text{if } i \text{ is a unique maximum} \\ 0 & \text{if } i \text{ is not a maximum rank} \end{cases} \tag{8}$$

Derivation of a three-dimensional MaRR procedure using the maximum proceeds by the following steps:

1. Determine the marginal probability mass function of $W_n^{(3)}(i)$.

2. Use the fact that $P(M_g^{(3)} = i) = E(W_n^{(3)})/n$ to define a marginal pmf for $M_g^{(3)}$: $f_{n,0}^{(3)}(i/n)$, similar to the marginal pmf $f_{n,0}(i/n)$ in Proposition 1.

3. Assume ideal conditions

   (I1) Reproducible signals are always ranked higher than irreproducible signals for all three experiments.

   (I2) Correlation between ranks of reproducible signals is non-negative.

   (I3) The three ranks per irreproducible gene are independent.

   and derive marginal mass function $f_{n,\pi_1}^{(3)}(i/n)$ dependent on π_1, the proportion of signals consistent across all three experiments

4. Calculate corresponding marginal cumulative distribution and survival functions, $F_{n,\pi_1}^{(3)}$ and $S_{n,\pi_1}^{(3)}$

5. Derive limiting distributions $F_{\pi_1}^{(3)}$, $S_{\pi_1}^{(3)}$ as $n \to \infty$.

6. Define a loss function dependent on π_1 similar to $SS(\lambda)$:

$$SS_\lambda^{(3)} = \frac{1}{1-\lambda} \int_\lambda \left(\hat{S}_n^{(3)}(x) - (1-\lambda) S_\lambda^{(3)}(x) \right)^2 dx,$$

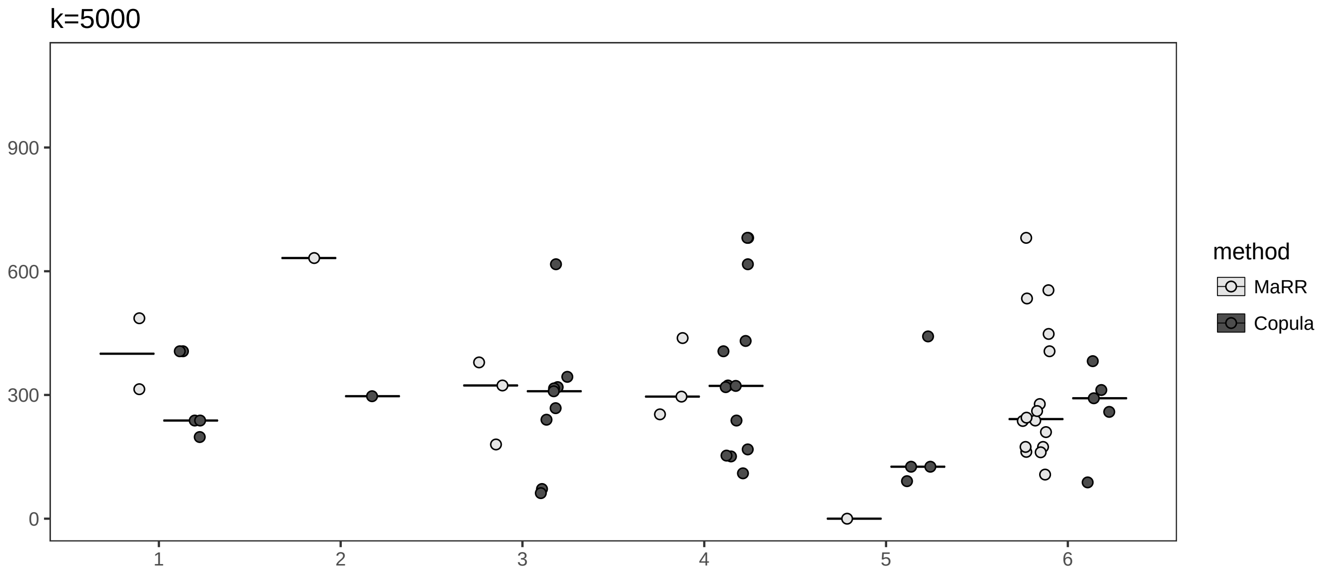
   where $\hat{S}_n^{(3)}(x)$ is the empirical survival function for $M_g^{(3)}$.

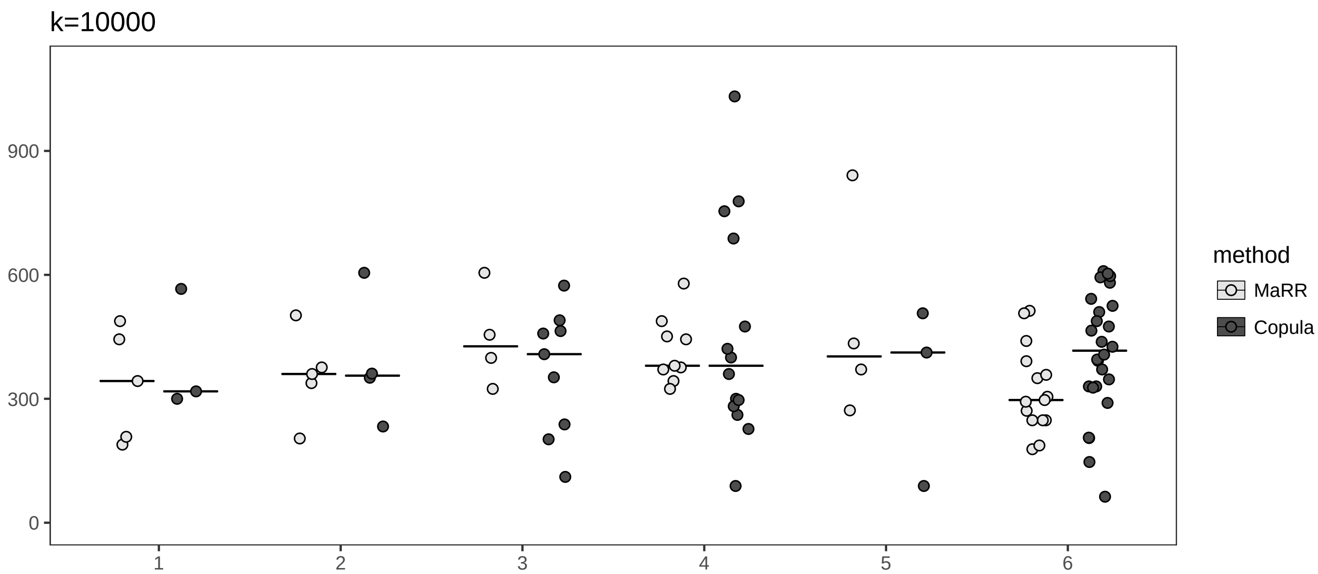
7. Define a new estimate $\hat{k} = \text{argmin}_i \{SS(i/n)\}$.

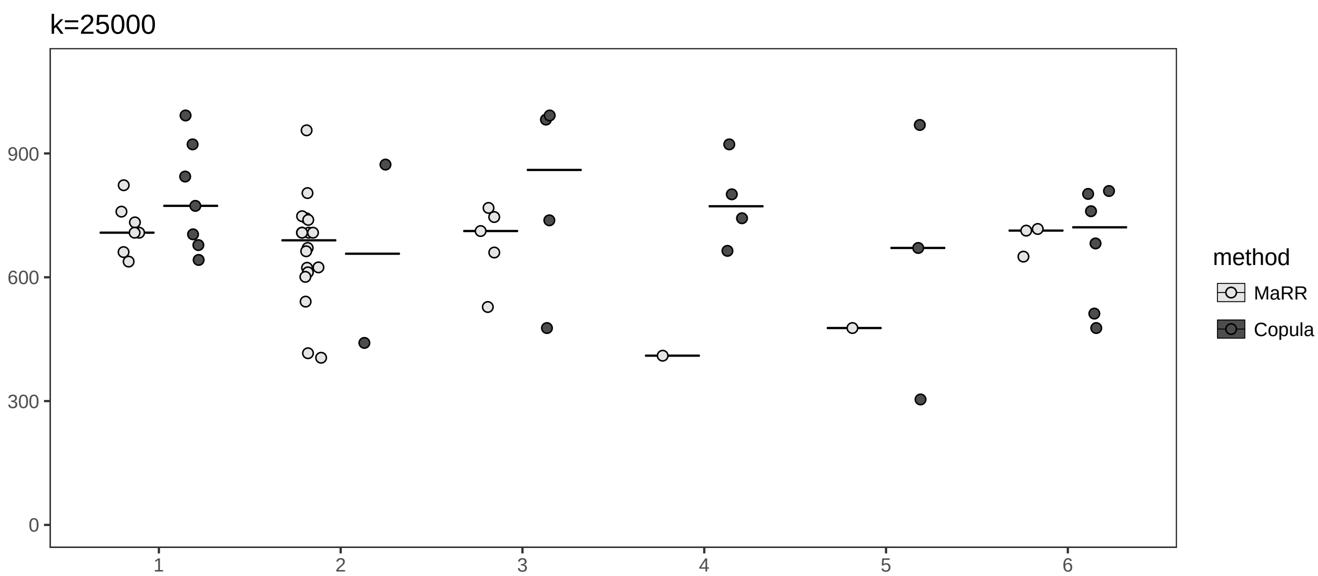
8. Estimate mFDR using for rejection region $f_{n,\hat{k}/n}^{(3)}$.

Figure for top k genes

The below figure shows the rank of method-specific PCR genes in the top k (5000, 10000, and 25000) for only MaRR (light gray) or for only the copula mixture model (dark gray) for comparisons 1–6 of the SEQC data. Horizontal lines indicate median values. PCR genes with lower-valued ranks are more highly expressed.


k=5000
900
600
300
0
1
2
3
4
5
6
method
MaRR
Copula


k=10000
900
600
300
0
1
2
3
4
5
6
method
MaRR
Copula


k=25000
900
600
300
0
1
2
3
4
5
6
method
MaRR
Copula

References

A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 1993.