SUPPLEMENTARY MATERIAL

Statistical Properties of *RN* and *SRN* measures

The relative nonsynonymous divergence ($RN = \left| \frac{d_{N12} - d_{N13}}{d_{N23}} \right|$) is a modification of the Canberra distance: $d(x,y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$ where *x* and *y* are *p*-dimensional vector (Johnson and Wichern 2002, p. 671). *RN* is one dimension and uses $x_i + y_i - r_i$ (r_i>0: a random variable) instead of $x_i + y_i$ because we can consider d_{N23} as $d_{N12} + d_{N13} - r$ (r >0: a random variable). The Canberra distance examines the sum of series of a fraction differences between coordinates of a pair of vectors. Each term of fraction difference has value between 0 and 1 (*RN* can be greater than 1 since d_{N23} is occasionally less than $|d_{N12} - d_{N13}|$. Conversely, *SRN* be less than -1 or more than 1, see below). Note that if both coordinate are zeros, it needs to be defined as $\frac{0}{0} = 0$. The Signed *RN* (*SRN* $= \left(\frac{d_{N12} - d_{N13}}{d_{N23}} \right)$) preserves the properties of the Canberra distance since *RN*=|*SRN*| and gives information of asymmetry between a pair. *SRN* is approximately normally distributed in our data set (Fig S1.A).

Taylor series provide a natural method to approximate variance through polynomials (Casella and Berger 1990, p.328). To approximate the variance of *RN* and *SRN*, we will use the first-order Taylor series. First, we will approximate the variance of a

differentiable function, $f(x,y,z)=\frac{x-y}{z}$. Let x,y,z be random variables with means x_0, y_0, z_0 where *x* and *y* are independent, $z=x+y-r>0$ with a random variable *r* (>0), which is independent of x and y, $E(z)=E(x)+E(y)-E(r)$ and $\mathrm{var}(z)=\mathrm{var}(x)+\mathrm{var}(y)+\mathrm{var}(r)$. Then the estimate of the variance of $f(x,y,z)$ is as follows:

$$\mathrm{var}(f(x,y,z)) \approx a\,\mathrm{var}(x)+b\,\mathrm{var}(y)+c\,\mathrm{var}(z)$$

where $a=\frac{(2y_0-2x_0+z_0)(y_0-x_0+z_0)}{z_0^4}$, $b=\frac{(2y_0-2x_0-z_0)(y_0-x_0-z_0)}{z_0^4}$,

$c=\frac{(y_0-x_0)^2}{z_0^4}$. Therefore, we can now approximate the variance of $SRN=\left(\frac{d_{N12}-d_{N13}}{d_{N23}}\right)$ by

$$\mathrm{var}(SRN)\approx\frac{(2d_{N13}-2d_{N12}+d_{N23})(d_{N13}-d_{N12}+d_{N23})}{{d_{N23}}^4}\mathrm{var}(d_{N12})$$

$$-\frac{(2d_{N13}-2d_{N12}-d_{N23})(d_{N13}-d_{N12}-d_{N23})}{{d_{N23}}^4}\mathrm{var}(d_{N13})+\frac{(d_{N13}-d_{N12})^2}{{d_{N23}}^4}\mathrm{var}(d_{N23})\ .$$

We calculated the standard error of *SRN* by using the expression above (Fig S1.B). Since *E(RN)* and *E(SRN)* is approximately *RN* and *SRN*, the estimated variance of *RN* is the same as that of *SRN*.

Results when we used *RN = $\left|\frac{d_{N12}-d_{N13}}{d_{N12}+d_{N13}}\right|$, *SRN** = $\left(\frac{d_{N12}-d_{N13}}{d_{N12}+d_{N13}}\right)$.**

We can also investigate asymmetric sequence divergence using $d_{N12}+d_{N13}$ as the denominator. Similar to above, we can obtain the variance of $SRN^*=\left(\frac{d_{N12}-d_{N13}}{d_{N12}+d_{N13}}\right)$ by

$\mathrm{var}(SRN^*) \approx \frac{4(d_{N13})^2}{(d_{N12}+d_{N13})^4}\mathrm{var}(d_{N12}) + \frac{4(d_{N12})^2}{(d_{N12}+d_{N13})^4}\mathrm{var}(d_{N13})$. The results from correlation analyses are qualitatively the same and significant. *RN** is significantly correlated with d_N (Kendall's correlation coefficient (τ) = 0.30, P<0.0001, Fig S2.A) and -2loglikelihood ratio (τ = 0.65, P < 0.0001, Fig S2.B). *SRN** is significantly negatively correlated with *SRK* (τ = -0.15, P < 0.001, Fig S2.C), *SRA* (τ = -0.20, P < 0.001, Fig S2.D), while significantly positively correlated with *SRF* (τ = 0.19, P < 0.0001, Fig S2.E). Also, *SRN** is significantly negatively correlated with signed measure of closeness (*SRC*, see below; τ = -0.10, P = 0.0277, Fig S2. F), and the signed measure of betweenness (*SRB*, see below; τ = -0.11, P = 0.0078, Fig S2. G). Again, we performed the Kendall's correlation test with *SRF* and *SRC* where -0.1 < *SRF* < 0.1 (Fig S2. H) and -0.5< *SRC* < 0.5 (Fig S2. I) and found significant correlations (τ = 0.14, -0.09, P = 0.0044, P = 0.0329 respectively).

We examined the covariance of the denominator term ($x_i + x_j$) and the relative measure ($\frac{x_i - x_j}{x_i + x_j}$) and found that there is no linear relationship between these. In other words, the relationship between relative measures is statistically independent from the relationship between the denominators.

REFERENCES

Casella, G. and R. L. Berger. Statistical Inference. (Duxbury Press, Belmont, 1990)

Johnson, R. A. and D. W. Wichern. Applied Multivariate Statistical Analysis. Fifth Ed. (Prentice Hall, New Jersey, 2002)

FIGURE LEGENDS

Figure S1. The distribution of *SRN* and its standard error (SE, the square root of variance divided by the sample number) among the duplicate pairs. A. The bar chart represents the histogram of *SRN* and the solid line indicates the density of *SRN*. The distribution is approximately symmetric. B. The SE of *SRN* for each duplicate pair is indicated by the bar chart. The dashed line represents the SE is 0.5. Almost all of the SE are less than 0.5.

Figure S2. Scatter plots for *RN** and *SRN** (using d_{N12}+ d_{N13} as the denominator). See Supplementary text for description and statistical significance of each plot. The last two scatter plots indicate the relationships between *SRN* and *SRF*, *SRC* when -0.1< *SRF* <0.1 and -0.5< *SRC* <0.5.

Figure S1.

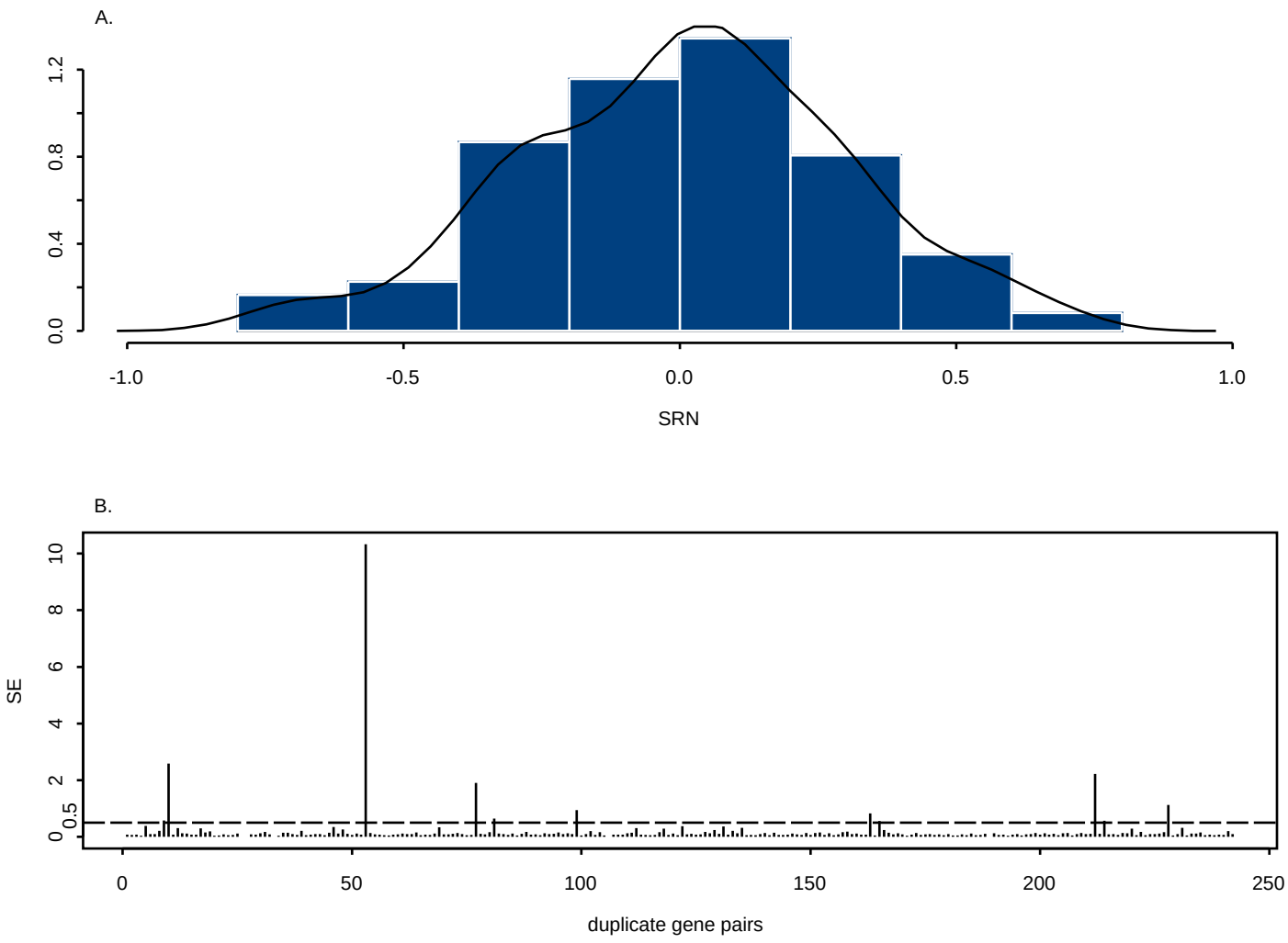
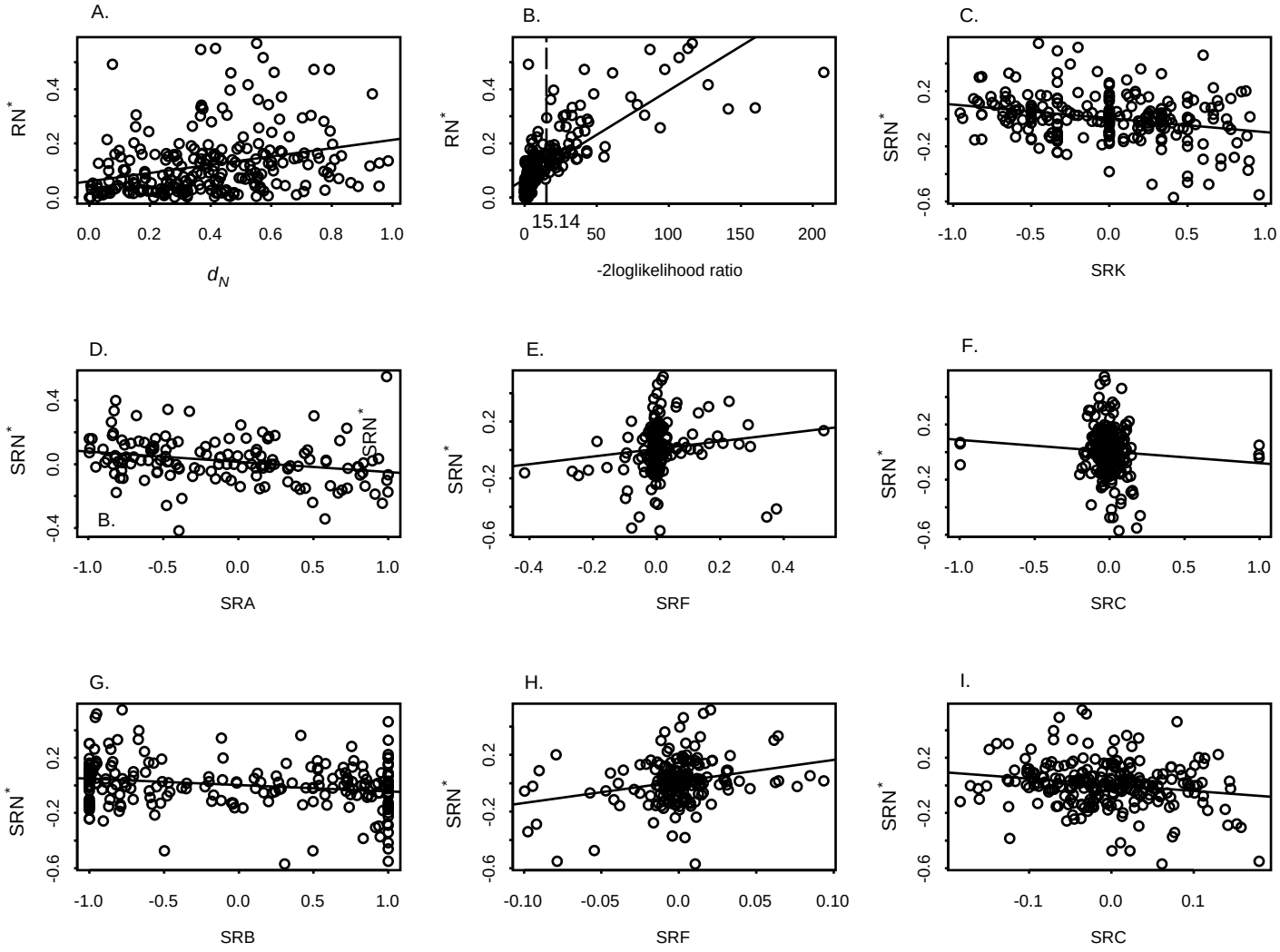

A.
0.0
0.4
0.8
1.2
-1.0
-0.5
0.0
0.5
1.0
SRN
B.
SE
0
0.5
2
4
6
8
10
0
50
100
150
200
250
duplicate gene pairs

Figure S2.


A.
RN*
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
d_N
B.
RN*
0.0
0.2
0.4
15.14
0
50
100
150
200
-2loglikelihood ratio
C.
SRN*
0.2
-0.2
-0.6
-1.0
-0.5
0.0
0.5
1.0
SRK
D.
SRN*
0.4
0.0
-0.4
SRN*
B.
-1.0
-0.5
0.0
0.5
1.0
SRA
E.
SRN*
0.2
-0.2
-0.6
-0.4
-0.2
0.0
0.2
0.4
SRF
F.
SRN*
0.2
-0.2
-0.6
-1.0
-0.5
0.0
0.5
1.0
SRC
G.
SRN*
0.2
-0.2
-0.6
-1.0
-0.5
0.0
0.5
1.0
SRB
H.
SRN*
0.2
-0.2
-0.6
-0.10
-0.05
0.0
0.05
0.10
SRF
I.
SRN*
0.2
-0.2
-0.6
-0.1
0.0
0.1
SRC