1 CAI calculation

the first step in the CAI index as described by Sharp et Li Sharp and Li (1987) consists in defining for each synonymous codon of an amino acid a relative adaptiveness score ($w_{c,a}$). In the reference dataset, this value is obtained by calculating the ratio between the frequency of a codon ($f^{\mathrm{ref}}_{c,a}$) and the frequency of its most represented synonymous ($f^{\mathrm{ref}}_{\mathrm{cmax,a}}$):

$$\mathrm{w_{c,a}} = \frac{f^{\mathrm{ref}}_{c,a}}{f^{\mathrm{ref}}_{\mathrm{cmax,a}}}$$

The CAI score is obtained by calculating the geoindexal mean of the relative adaptivenesses multiplied by the occurrences of the related codons found in the query sequence (Sharp and Li, 1987). Here, we name this CAI score CAI$_{59}$, since all 59 synonymous codons have an impact on the CAI score regardless of their relation with their amino acid:

$$\mathrm{CAI}_{59} = \left(\prod_{a \in \mathcal{A}} \prod_{c \in k_a} \mathrm{Occ}^{\mathrm{que}}_{c,a} \times w_{c,a} \right)^{\frac{1}{L}}$$

where $\mathrm{Occ}^{\mathrm{que}}_{c,a}$ is the number of occurrences of codon c in the query and L is the length of the query (total number of amino acids).

In the classical CAI score, the amino acid composition of the query sequence is included in the calculation because all codons contribute equally to the final score. This calculation is analogous to our description of COUSIN$_{59}$, and we therefore refer to it as CAI$_{59}$. For the sake of completeness, we introduce an alternative CAI definition, hereafter named CAI$_{18}$, for which all amino acids contribute equally. The difference between CAI$_{18}$ and CAI$_{59}$ simply lies in the calculation of the geoindexal mean, as follows:

$$\mathrm{CAI}_{18} = \left(\prod_{a \in \mathcal{A}} \prod_{c \in k_a} \frac{\mathrm{Occ}^{\mathrm{que}}_{c,a}}{\mathrm{Occ}^{\mathrm{que}}_{a}} \times w_{c,a} \right)^{\frac{1}{N}}$$

where $\mathrm{Occ}^{\mathrm{que}}_{a}$ is the number of occurrences of the amino acid a in the query, $\mathrm{Occ}^{\mathrm{que}}_{c,a}$ the number of occurrences of codon c in the query and $w_{c,a}$ the relative adaptiveness score (Supplementary Information 1).

Both pairs, COUSIN$_{18}$ and COUSIN$_{59}$ on the one hand and CAI$_{18}$ and CAI$_{59}$ on the other, differ therefore in the way the amino acid composition is accounted for in the calculation. With the "18" indexes, all amino acids contribute equally, independently of their frequency in the protein. The two "18" indexes can be envisioned as the "amino acid by amino acid" CUPrefs of a sequence. With the "59" indexes, all individual codons contribute equally, so that the final contribution of each amino acid is proportional to its frequency in the protein. The two "59" indexes can be envisioned as the "codon by codon" CUPrefs of a sequence. By comparing the "18" and "59" scores of an index, we can estimate the impact of amino acid composition on the observed CUPrefs of a sequence.

2 COUSIN software architecture

The Figure 1 describes the global architecture of the COUSIN software.

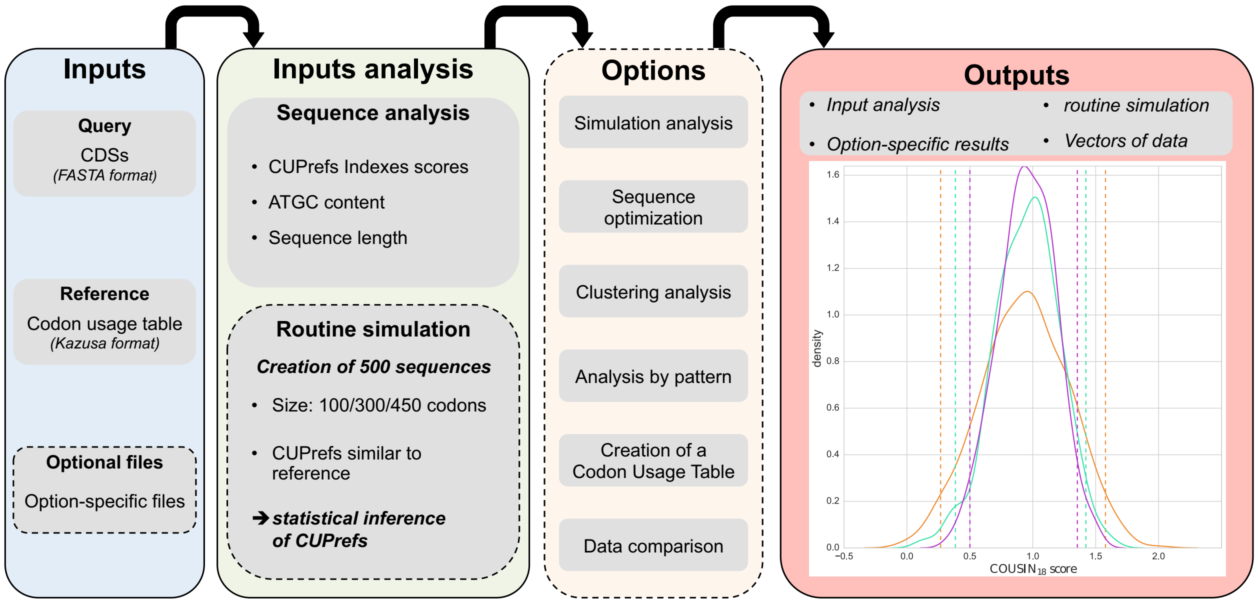

Fig. 1. Architecture of the COUSIN software. The COUSIN software requires input data from the user such as sequences in a FASTA format and a Codon Usage Table in a kazusa-style format (Nakamura *et al.*, 2000). COUSIN performs a CUPrefs analysis on the queries by performing routine tasks. Following user specifications, options can be chosen to deepen the analysis. Graphics and text outputs are given at the end of a COUSIN job. The last part of the figure displays the graphics given by a routine simulation. Here, density curves show COUSIN$_{18}$ range of scores for the generated sequences following a "random-guided" CUPrefs and amino acid composition selection with *E. coli* CUPrefs as reference (Puigbò *et al.*, 2008). Orange, cyan and purple curves refer to sequences with a codon length of 100 (short proteins), 300 (average length of prokaryotic proteins) and 450 (average length of eukaryotic proteins). The longer the generated sequences, the lower the variance and the higher the accuracy of the scores obtained (Comeron and Aguadé, 1998; Roth *et al.*, 2012). Dashed vertical lines indicate the respective 95% interval for orange, cyan and purple curves.

3 Summary data on studied organisms

Table 1 gives detailed informations on the organisms studied. The detailed % of GC content of these same organisms are given in Table 2.

Table 1. Summary statistics of the complete CDSs of the eight organisms included in the analysis. The table shows the species name, reference and accession number in the NCBI database, the number of protein-coding genes kept for the analysis (evaluated by removing isoforms and rejected sequences), the total number of CDSs retrieved (as annotated in genbank files), the ratio between the number of protein-coding genes and the total number of CDSs as well as the global GC3 content found in protein-coding genes.

Species	Reference	Number of protein-coding genes	Total Number of CDSs	Ratio	GC percent (3rd base)
Escherichia coli	K-12 substr. MG1655	3244	4319	0.8	54.9%
Streptomyces coelicolor	A3(2)	6356	8152	0.8	92.3%
Saccharomyces cerevisiae	S288C (assembly R64)	5549	5989	0.9	39.2%
Plasmodium falciparum	3D7 (assembly ASM276v1)	4773	5334	0.9	17.8%
Homo sapiens	Assembly GRCh38.p11	18492	115320	0.1	60.0%
Gallus gallus	Assembly GRCg6a	15751	49767	0.3	60.6%
Mus musculus	Assembly GRCm38.p6	20393	79262	0.3	58.6%
Arabidopsis thaliana	Assembly TAIR10	24774	48148	0.5	42.7%

Table 2. Average % of GC3 content of complete CDSs of *Escherichia coli*, *Streptomyces coelicolor*, *Saccharomyces cerevisiae*, *plasmodium falciparum*, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*. This table describes the average GC3 content found among all CDSs, but also in CDSs with the top 20%, the 60% in the middle and the bottom 20% COUSIN$_{59}$ score.

	E. coli	*S. coelicolor*	*S. cerevisiae*	*P. falciparum*	*H. sapiens*	*G. gallus*	*M. musculus*	*A. thaliana*
GC3 content (all dataset)	54.9	92.3	39.2	17.8	59.8	60.6	58.6	42.7
GC3 content (top 20%)	60.0	96.6	34.0	13.9	79.0	77.9	71.8	37.8
GC3 content (middle 60%)	56.1	93.5	38.0	17.3	61.0	61.2	59.5	42.1
GC3 content (bottom 20%)	46.1	84.7	48.1	23.1	37.7	41.6	42.8	49.4

4 Global analysis of CUPrefs scores against GC content in organisms

The resulting density curves for COUSIN$_{18}$, COUSIN$_{59}$, CAI$_{18}$ and CAI$_{59}$ are presented in Figure 2.

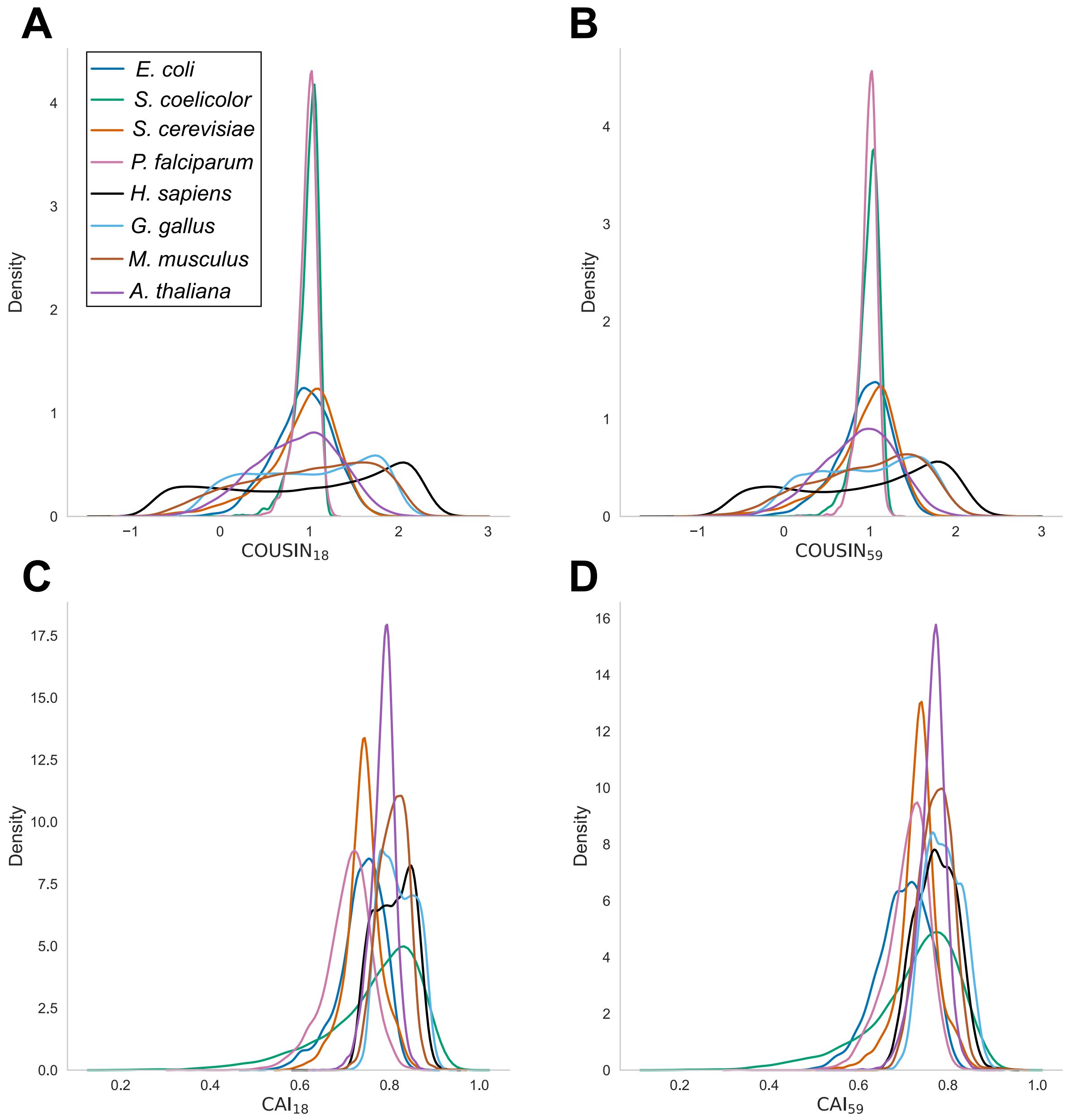

Fig. 2. Density curves for COUSIN$_{18}$ (A) and COUSIN$_{59}$ (B), CAI$_{18}$ (C) and CAI$_{59}$ (D) indices for the complete CDSs of the eight organisms studied (see color legend).. For each CDS, the values for each score were calculated against the average codon usage reference table of the corresponding genome. The COUSIN$_{18}$ and COUSIN$_{59}$ normalisation renders curves centered around 1 allowing for rapid identification of differential dispersion in the leptokurtic curves for organisms with strong nucleotide compositional biases (*e.g. S. coelicolor*, in green) compared to those more platykurtic for organisms with weaker compositional biases (*e.g. E. coli* in blue). Notice the bimodal distributions for *H. Sapiens* (black) and *G. gallus* (light blue) in panel A and B.

In addition to the first analysis putting on sight the dispersion of COUSIN and CAI scores among organisms CDSs, we also draw up the scores of these same CDSs with their GC3 content to bring light on the relation between these two variables. Figures 3, 4 (COUSIN$_{59}$), 5 and 6 (CAI$_{59}$) show individual scatterplots between one of the two index score and the GC3 content along with Pearson correlation tests.

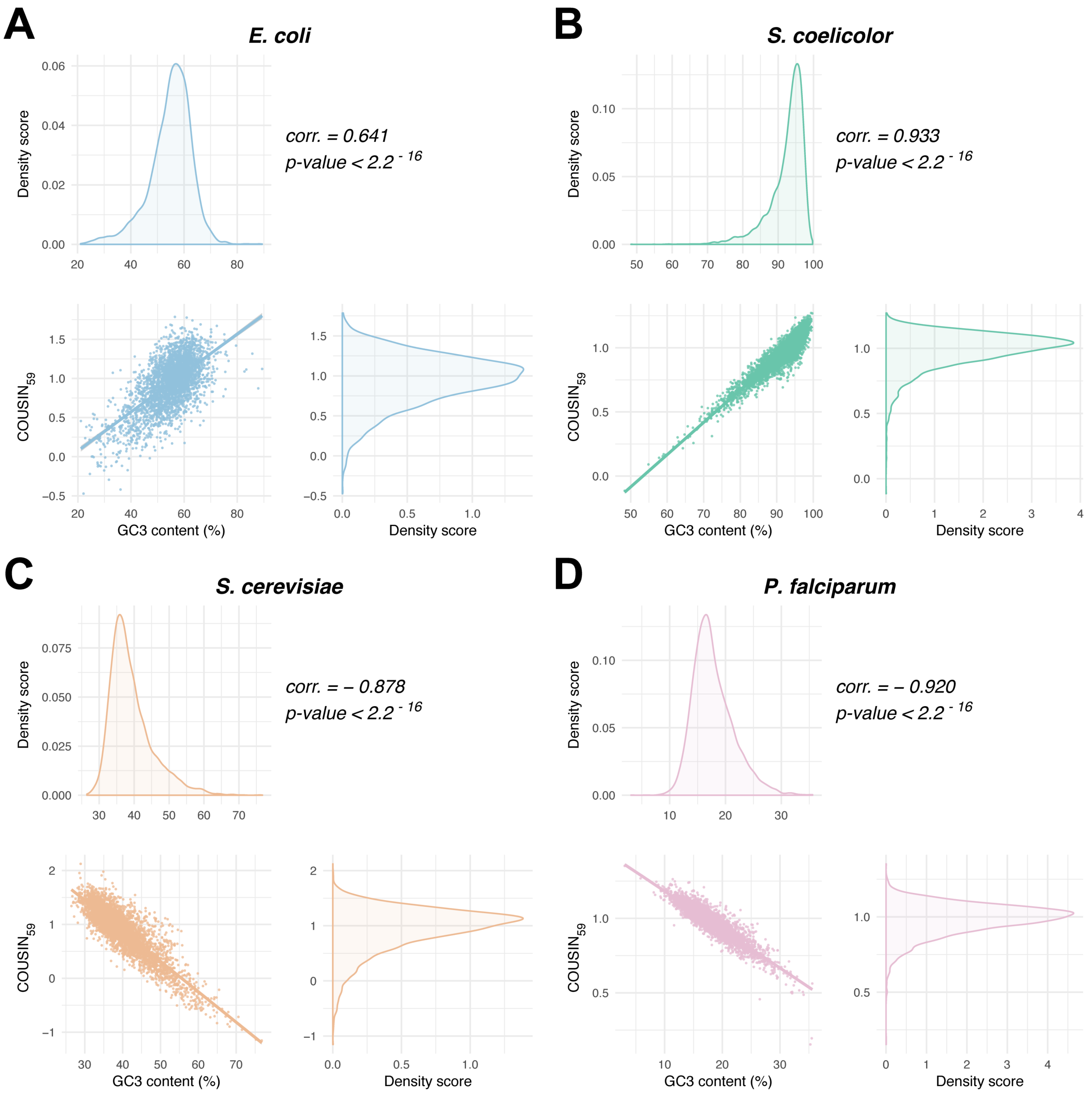

Fig. 3. Scatterplot of COUSIN$_{59}$ (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *E. coli* (A), *S. coelicolor* (B), *S. cerevisiae* (C) and *P. falciparum* (D). Each scatterplot is accompanied by two density curves: COUSIN$_{59}$ (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between COUSIN$_{59}$ scores and GC3 content is given.

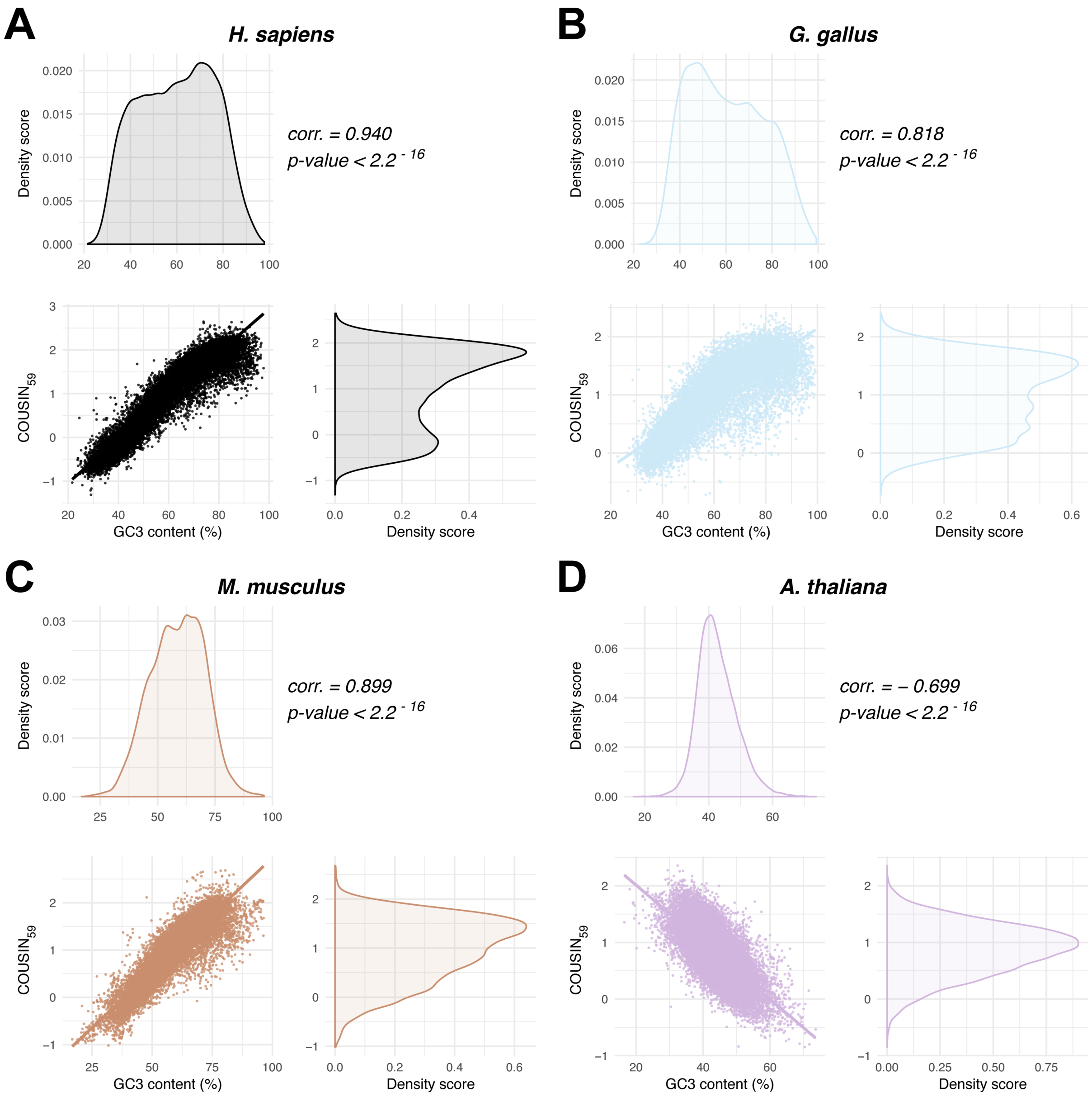

Fig. 4. Scatterplot of COUSIN$_{59}$ (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *H. sapiens* (A), *G. gallus* (B), *Mus musculus* (C) and *A. thaliana* (D). Each scatterplot is accompanied by two density curves: COUSIN$_{59}$ (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between COUSIN$_{59}$ scores and GC3 content is given.

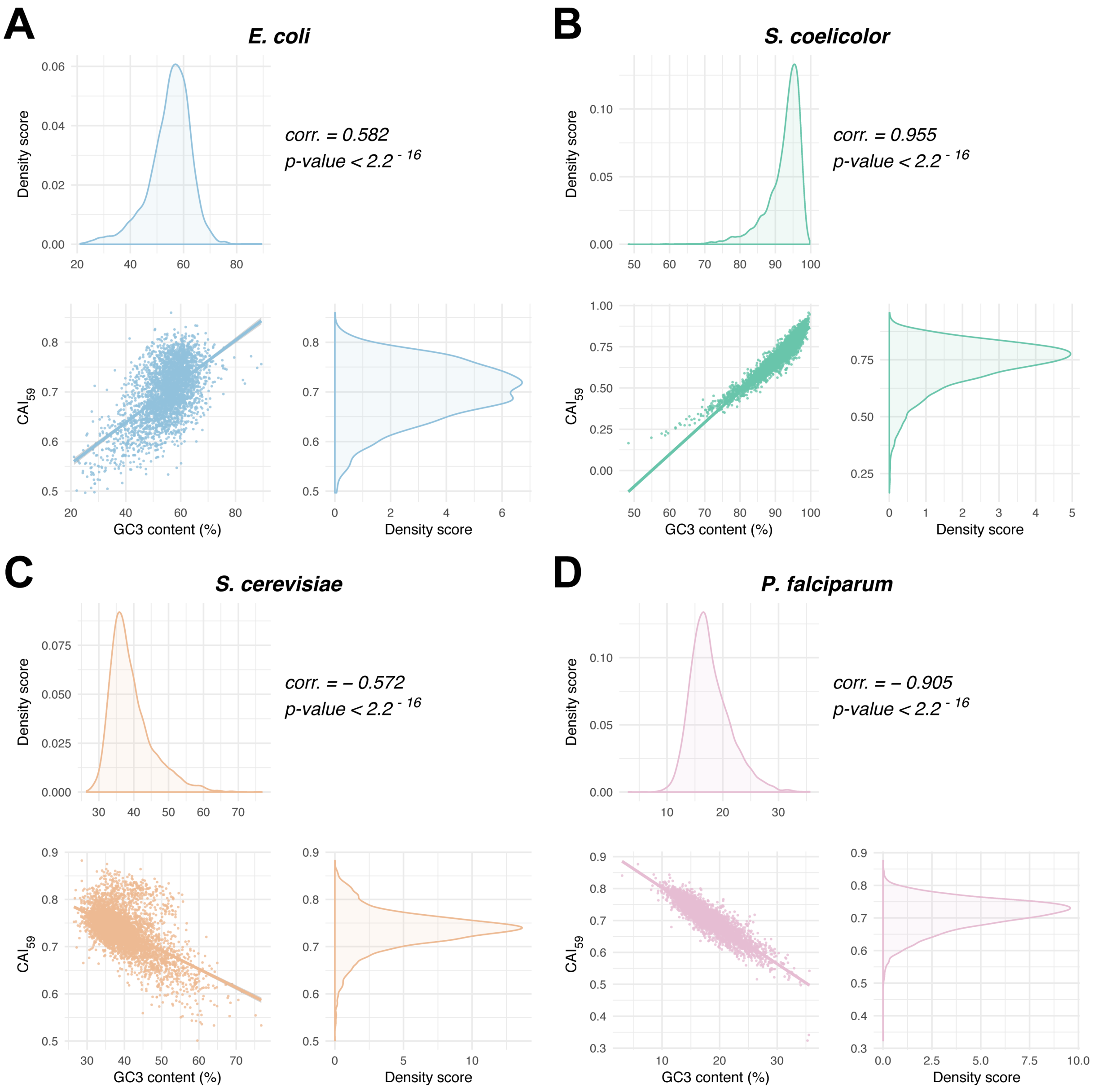

Fig. 5. Scatterplot of CAI_{59} (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *E. coli* (A), *S. coelicolor* (B), *S. cerevisiae* (C) and *P. falciparum* (D). Each scatterplot is accompanied by two density curves: CAI_{59} (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between CAI_{59} scores and GC3 content is given.

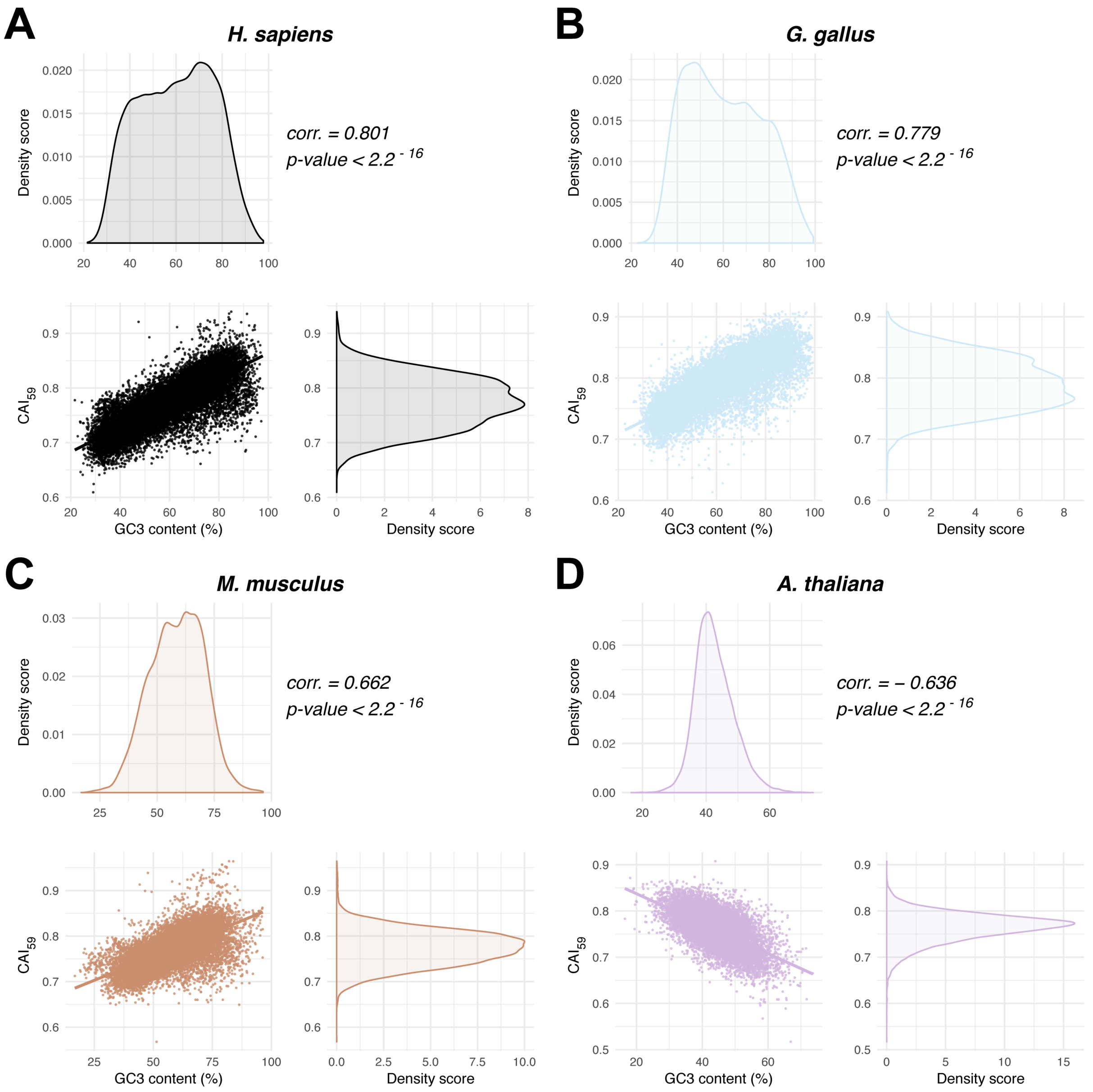

Fig. 6. Scatterplot of CAI$_{59}$ (x-axis) and GC3 content (y-axis) scores for CDSs belonging to *H. sapiens* (A), *G. gallus* (B), *Mus musculus* (C) and *A. thaliana* (D). Each scatterplot is accompanied by two density curves: CAI$_{59}$ (right of the scatterplot) and GC3 content (top of the scatterplot). On the top-right of the scatterplots, statistics of Pearson correlation tests between CAI$_{59}$ scores and GC3 content of CDSs is given.

5 Statistics on studied organisms

The mean values, Huber-M estimator values and Median of Absolute Deviations (MAD) scores related to the CAI and COUSIN analysis on the studied organisms are given in Table 3 (**?**).

Table 3. Mean value, Huber-M estimator value and MAD scores of COUSIN$_{18}$, COUSIN$_{59}$, CAI$_{18}$ and CAI$_{59}$ scores on the studied organisms.

	E. coli	***S. coelicolor***	***S. cerevisiae***	***P. falciparum***	***H. sapiens***	***G. gallus***	***M. musculus***	***A. thaliana***
Mean (COUSIN$_{18}$)	0.93	0.98	0.91	0.98	0.98	0.99	0.98	0.85
Huber-M estimator (COUSIN$_{18}$)	0.94	0.99	0.95	0.98	0.98	0.99	0.99	0.86
MAD (COUSIN$_{18}$)(+/-)	0.33	0.10	0.33	0.10	1.23	0.85	0.81	0.50
Mean (COUSIN$_{59}$)	0.94	0.98	0.93	0.98	0.95	0.97	0.97	0.87
Huber-M estimator (COUSIN$_{59}$)	0.96	1.00	0.97	0.99	0.95	0.97	0.98	0.88
MAD (COUSIN$_{59}$)(+/-)	0.29	0.11	0.32	0.00	1.03	0.74	0.67	0.45
Mean (CAI$_{18}$)	0.74	0.77	0.74	0.71	0.81	0.82	0.81	0.79
Huber-M estimator (CAI$_{18}$)	0.74	0.78	0.74	0.71	0.81	0.23	0.81	0.79
MAD (CAI$_{18}$)(+/-)	0.05	0.09	0.03	0.05	0.05	0.05	0.04	0.02
Mean (CAI$_{59}$)	0.70	0.73	0.73	0.71	0.77	0.79	0.77	0.76
Huber-M estimator (CAI$_{59}$)	0.70	0.74	0.74	0.71	0.77	0.79	0.77	0.77
MAD (CAI$_{59}$)(+/-)	0.06	0.09	0.03	0.04	0.05	0.05	0.04	0.03

6 Global analysis of correlation between COUSIN and CAI indexes

Scatter-plots along with Pearson correlation tests between COUSIN and CAI indexes are given in Figures 7 (*E. coli*), 8 (*S. coelicolor*), 9 (*S. cerevisiae*), 10 (*P. falciparum*), 11 (*H. sapiens*), 12 (*G. gallus*), 13 (*M. musculus*), 14 (*A. thaliana*).

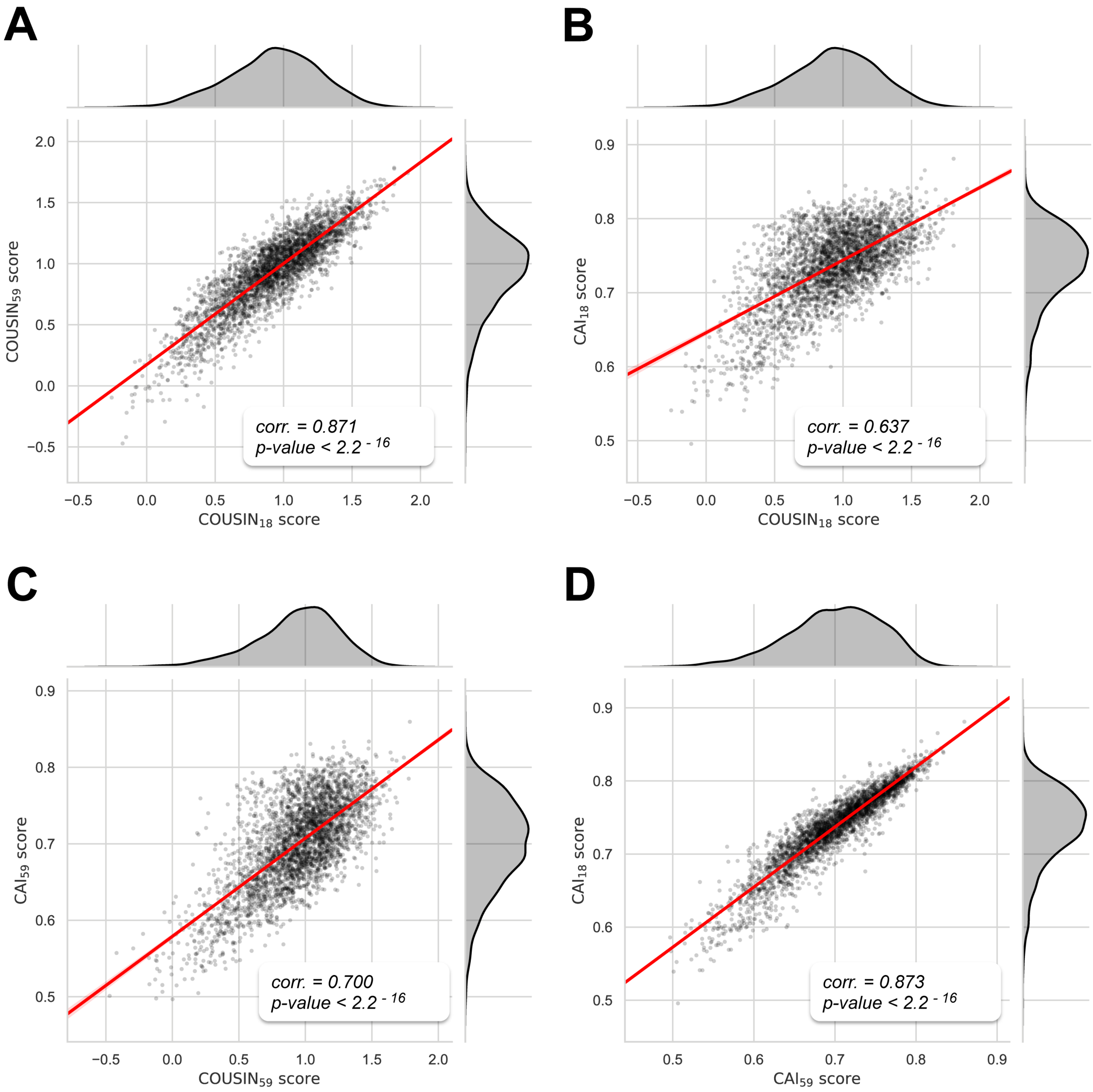

Fig. 7. Dot-plots of *E. coli* CDSs scores between COUSIN metrics declinations, CAI metrics declinations and between both metrics. In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

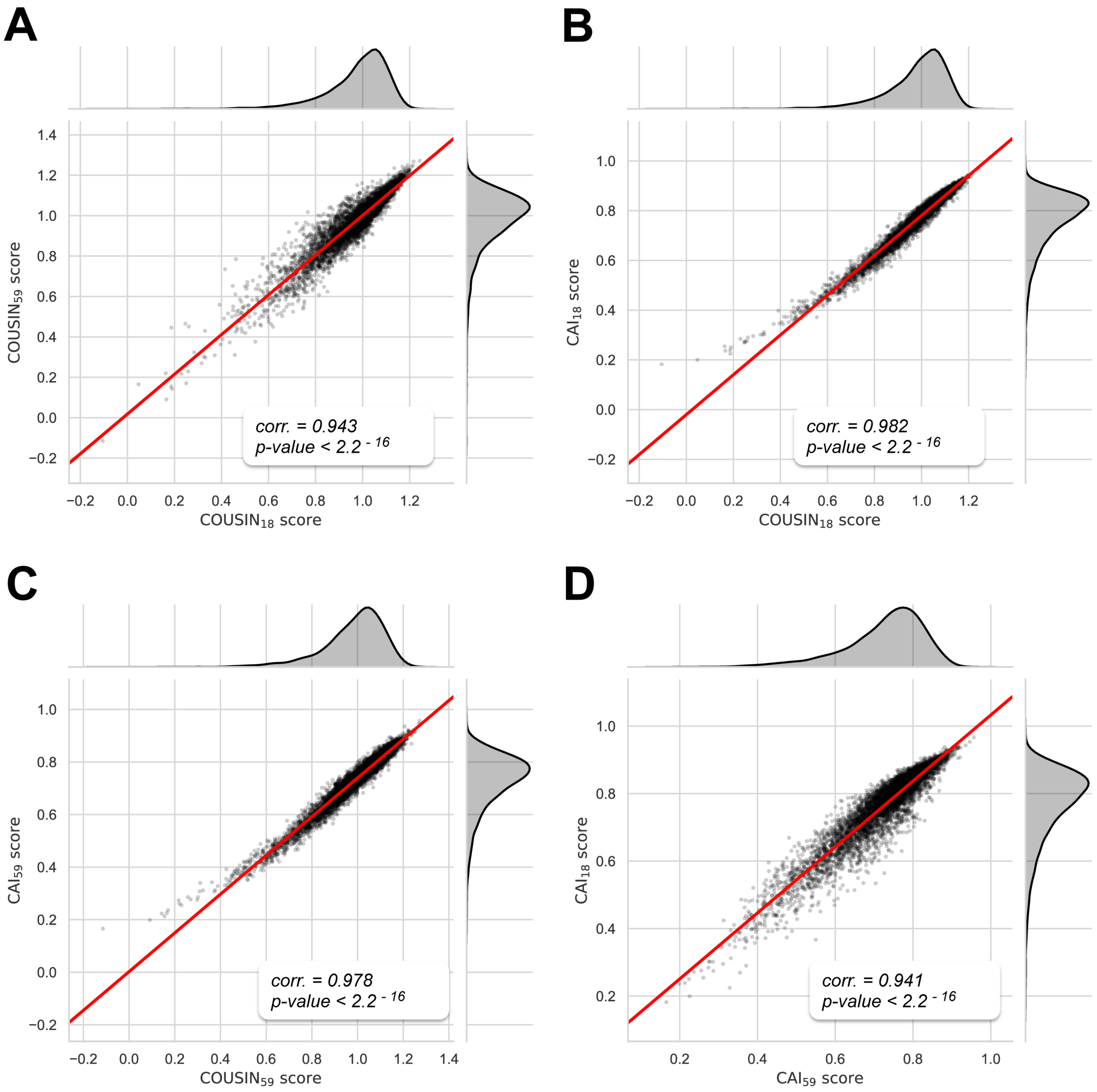

Fig. 8. Dot-plots of *S. coelicolor* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

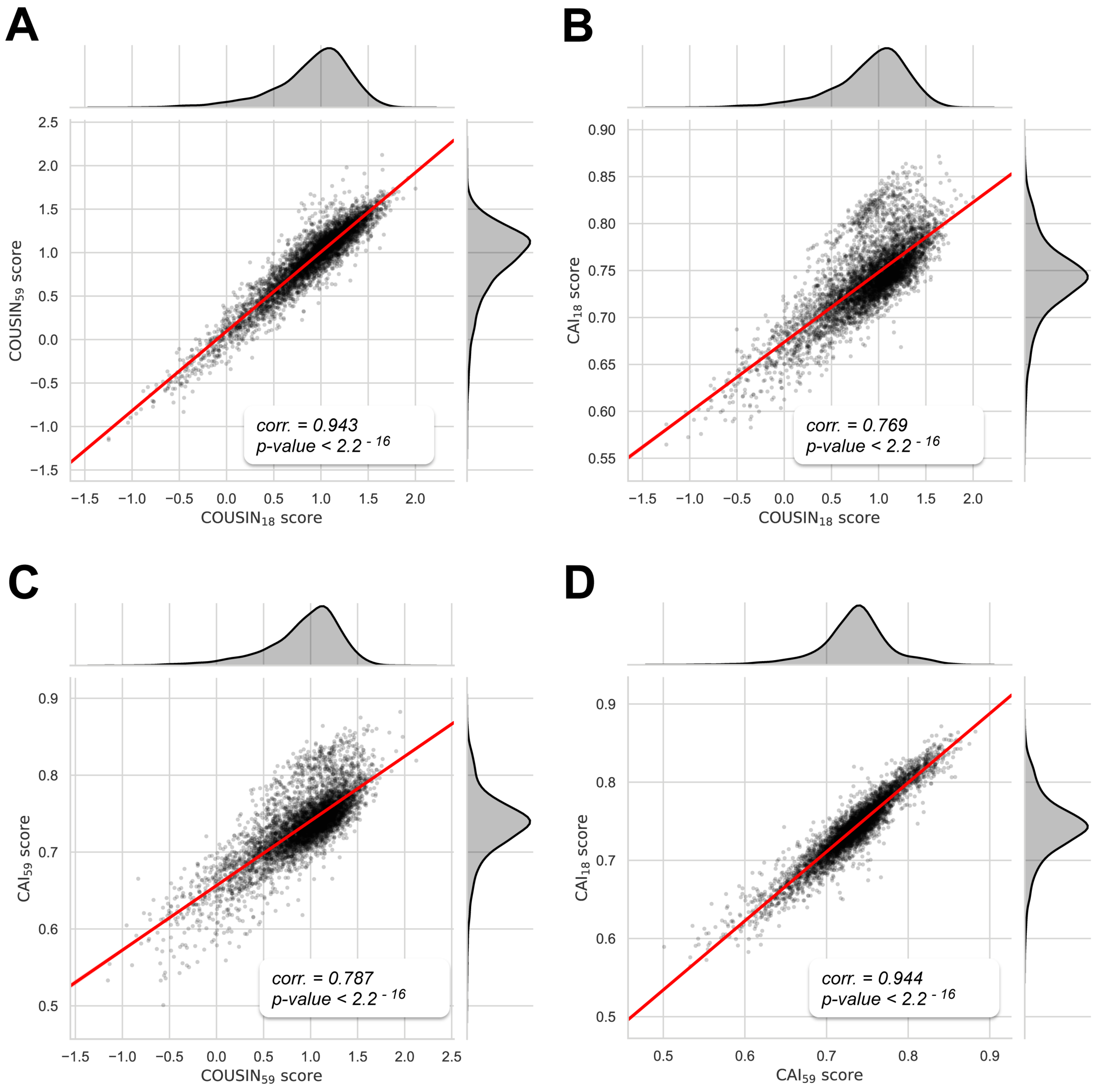

Fig. 9. Dot-plots of *S. cerevisiae* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

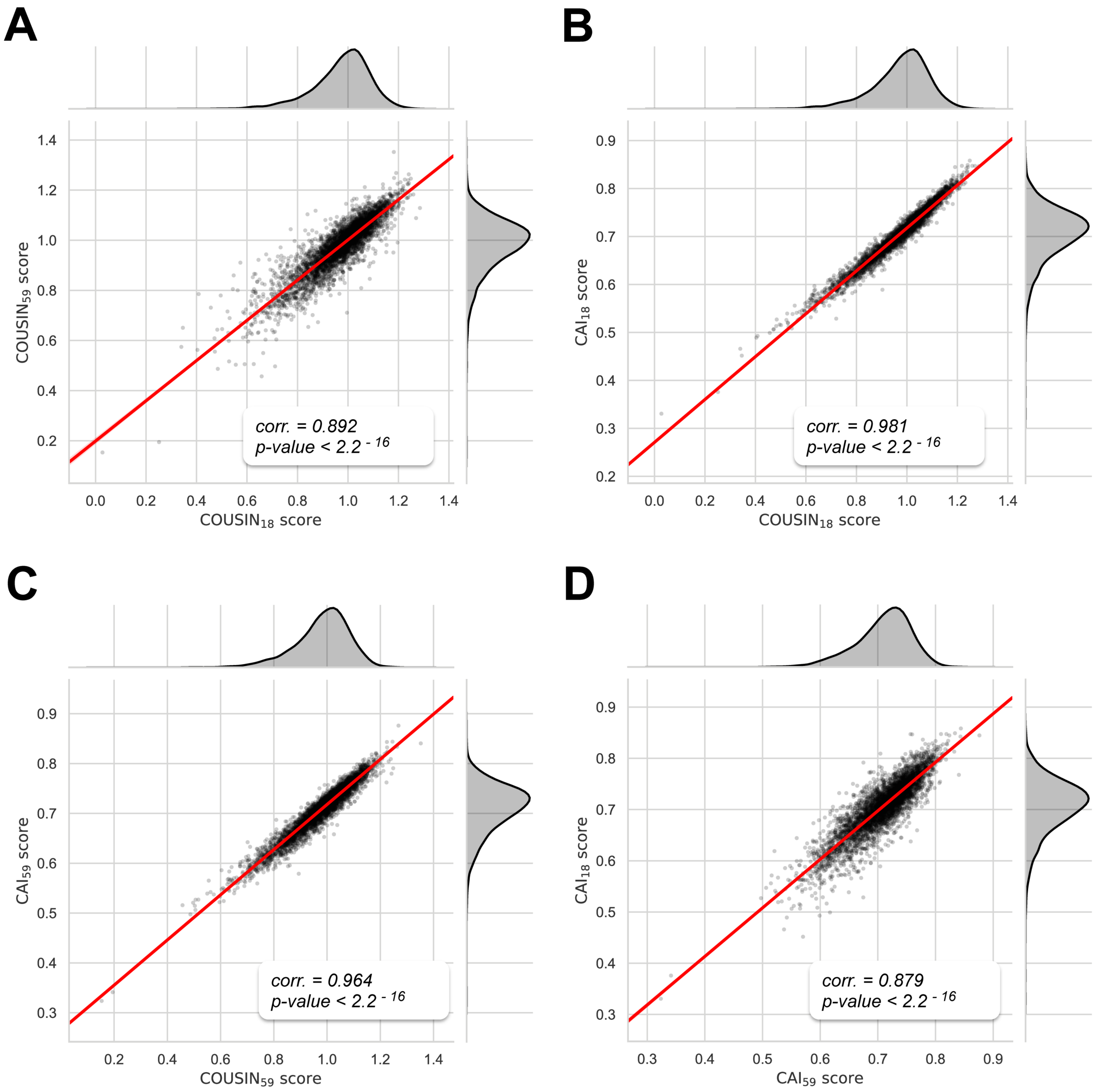

Fig. 10. Dot-plots of *P. falciparum* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

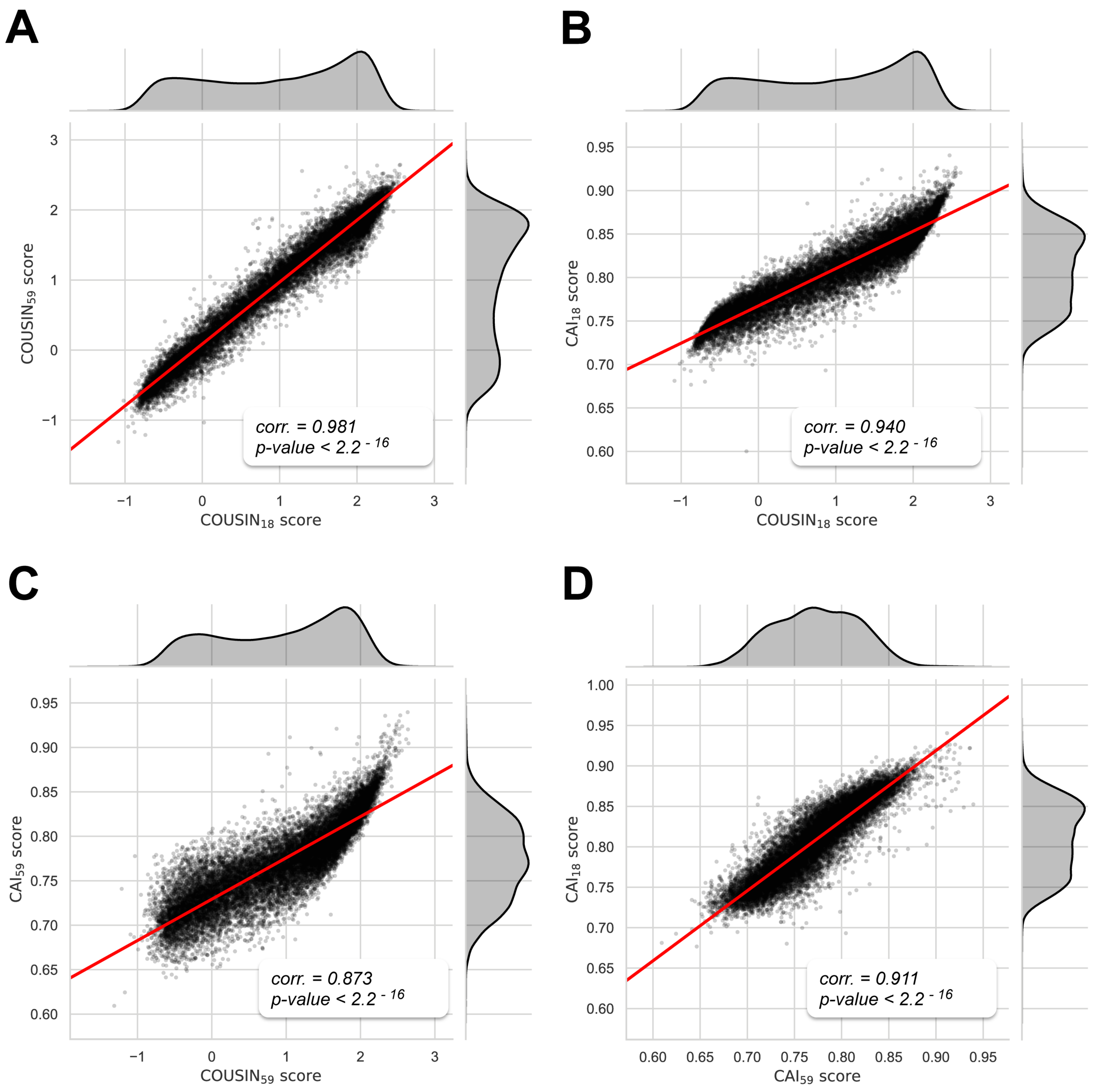

Fig. 11. Dot-plots of *H. sapiens* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

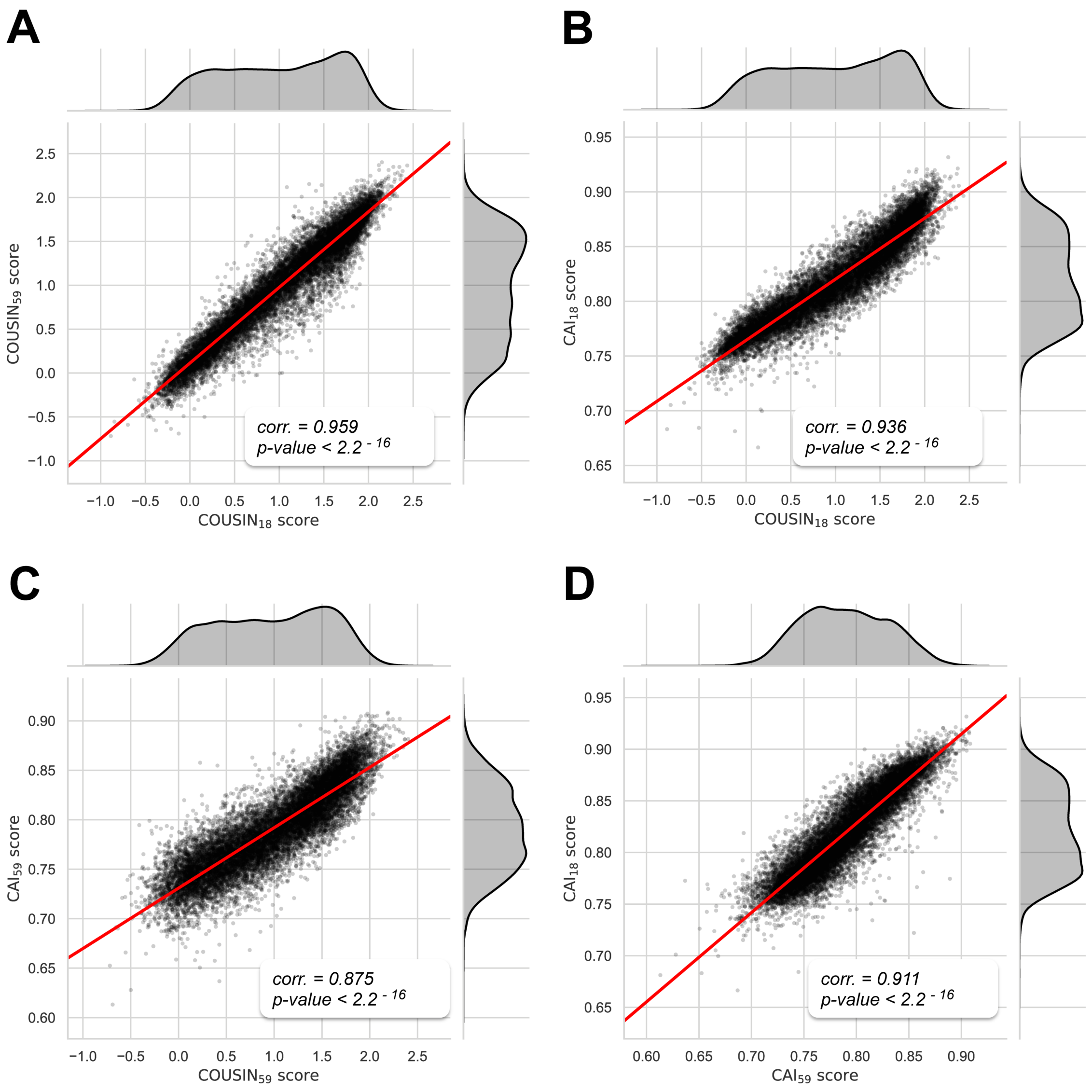

Fig. 12. Dot-plots of *G. gallus* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

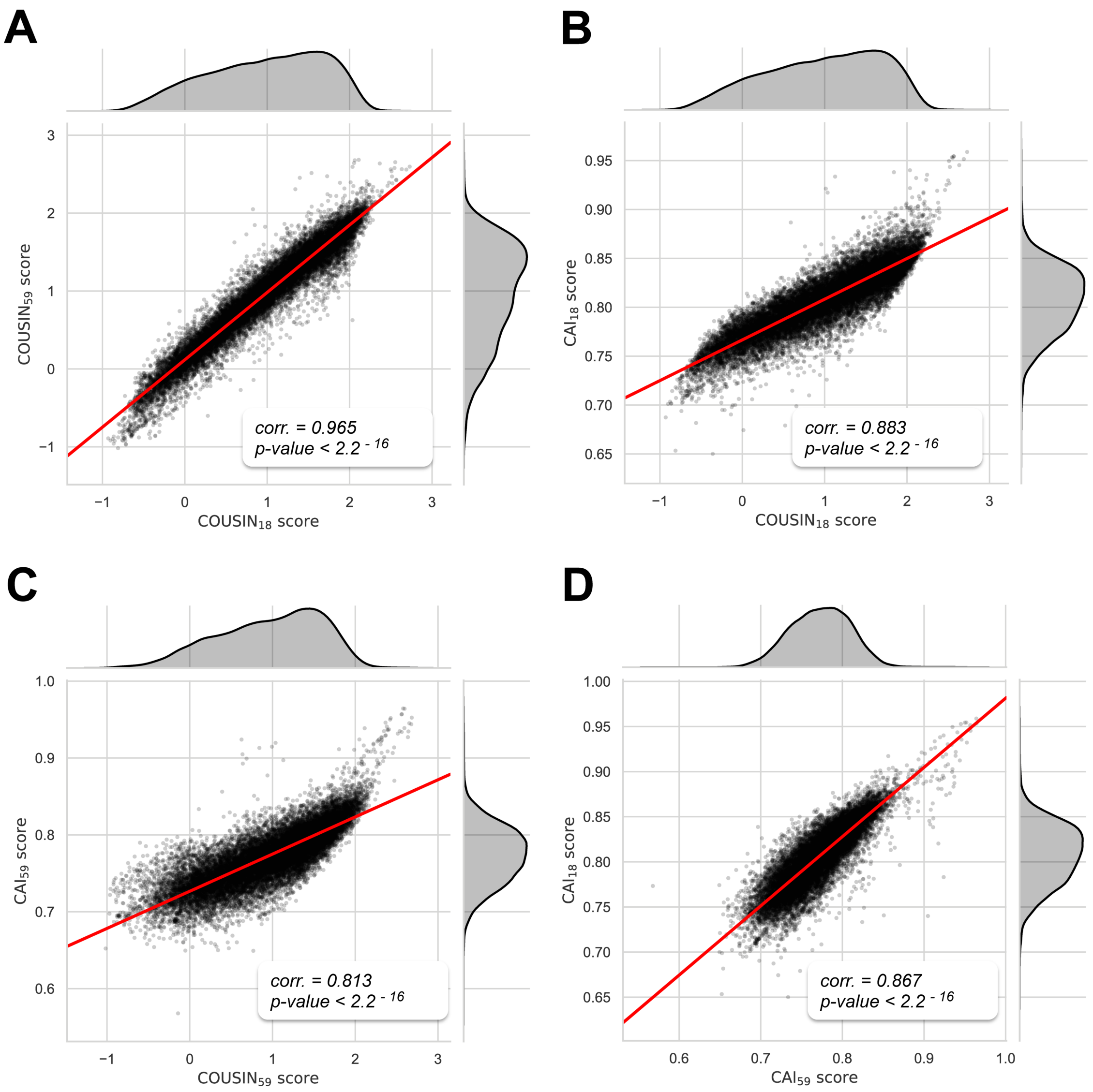

Fig. 13. Dot-plots of *M. musculus* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

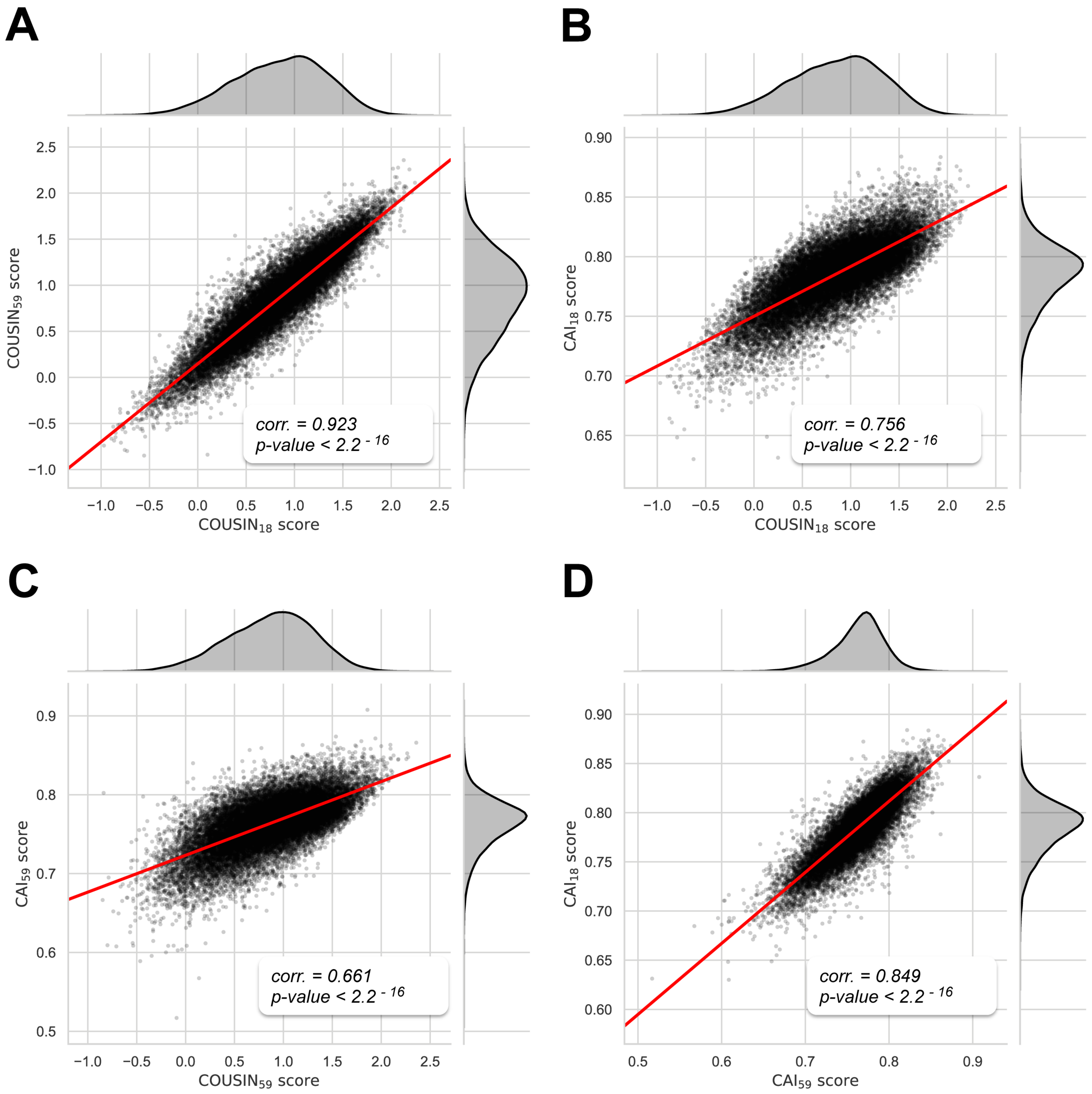

Fig. 14. Dot-plots of *A. thaliana* CDSs scores between COUSIN$_{18}$ and COUSIN$_{59}$ (A),COUSIN$_{18}$ and CAI$_{18}$ (B), COUSIN$_{59}$ and CAI$_{59}$ (C) and between CAI$_{18}$ and CAI$_{59}$ indexes (D). In addition to the dot-plot, a red regression line is given. For each plot, the x-axis and y-axis represent scores obtained for one metric. Results of Pearson's correlation test are indicated on the top-right of plots. Histograms and density plots are given at the opposite of x-axis and y-axis legends. These additional plots indicate the distribution of scores with the related index.

7 Statistics on *G. gallus* by chromosomes

Statistics by chromosomes are given in Tables 4 (*G. gallus*).

Table 4. Size, number of CDSs, Huber-M estimator values and MAD values for GC3 and COUSIN$_{59}$ among *G. gallus* chromosomes

Chromosome	**Size (Mb)**	**CDSs**	**Huber-M estimator (GC3)**	**MAD (+/-)(GC3)**	**Huber-M estimator (COUSIN$_{59}$)**	**MAD (+/-) (COUSIN$_{59}$)**
1	197.6	2025	53.8	14.2	0.8	0.7
2	149.7	1315	51.4	12.5	0.7	0.6
3	110.8	1124	53.1	14.0	0.7	0.7
4	91.3	1094	56.0	16.4	0.8	0.7
5	59.8	920	58.3	17.6	0.9	0.8
6	36.4	520	57.5	16.9	0.9	0.7
7	36.7	470	54.3	15.3	0.8	0.7
8	30.2	491	57.1	19.1	0.9	0.8
9	24.2	415	59.8	19.0	0.9	0.8
10	21.1	400	60.4	19.9	1.0	0.8
11	20.2	356	63.5	22.8	1.1	0.7
12	20.3	340	64.0	22.4	1.0	0.7
13	19.1	354	65.7	20.2	1.1	0.6
14	16.2	392	63.8	18.7	1.1	0.6
15	13.1	347	63.6	18.1	1.1	0.6
16	2.8	133	70.3	10.5	1.3	0.37
17	10.8	284	66.2	16.5	1.2	0.6
18	11.4	306	66.3	19.0	1.2	0.6
19	10.3	324	65.7	16.8	1.2	0.6
20	13.9	335	64.4	18.2	1.2	0.6
21	6.8	236	65.0	17.2	1.2	0.5
22	5.5	186	73.8	13.4	1.4	0.4
23	6.2	248	69.9	17.4	1.3	0.5
24	6.5	184	68.1	15.0	1.4	0.4
25	4.0	259	76.4	12.2	1.5	0.4
26	6.1	268	70.7	14.3	1.4	0.4
27	8.1	296	75.1	13.1	1.5	0.4
28	5.1	308	73.1	14.5	1.4	0.4
30	1.8	79	79.9	6.4	1.3	0.4
31	6.2	213	67.8	4.4	1.3	0.4
32	0.7	55	84.6	5.7	1.4	0.3
33	7.8	463	73.7	11.8	1.4	0.3
W	6.8	37	46.6	13.3	0.4	0.5
Z	82.5	773	52.6	15.8	0.7	0.7

References

Comeron, J. M. and Aguadé, M. (1998). An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution*, **47**(3), 268–274.

Nakamura, Y. *et al.* (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, **28**(1), 292.

Puigbò, P. *et al.* (2008). CAIcal: A combined set of tools to assess codon usage adaptation. *Biology Direct*, **3**, 38.

Roth, A. *et al.* (2012). *Measuring codon usage bias.* ResearchGate.

Sharp, P. M. and Li, W. H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**(3), 1281–1295.