PNAS

# Supporting Information for

## Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast

Elie Marcel Teyssonnière[1,#], Pauline Trébulle[3,#], Julia Muenzner[2,#], Victor Loegler[1], Daniela Ludwig[2,4], Fatma Amari[2,4], Michael Mülleder[4], Anne Friedrich[1], Jing Hou[1], Markus Ralser[2,3,5,*], and Joseph Schacherer[1,6,*]

1. Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France
2. Charité Universitätsmedizin Berlin, Berlin, Germany
3. The Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, UK
4. Core Facility High Throughput Mass Spectrometry, Charité Universitätsmedizin, Berlin, Germany
5. Max Planck Institute for Molecular Genetics, Berlin, Germany
6. Institut Universitaire de France (IUF), Paris, France

# first co-authors

* Corresponding authors

E-mail: ralser359@gmail.com (M.R.) and schacherer@unistra.fr (J.S.)

**This PDF includes:**

Supplementary figures S1-15
Supplementary figure legends
Legends for Datasets S1 to S26

**Other supporting materials for this manuscript include the following:**

Datasets S1 to S26

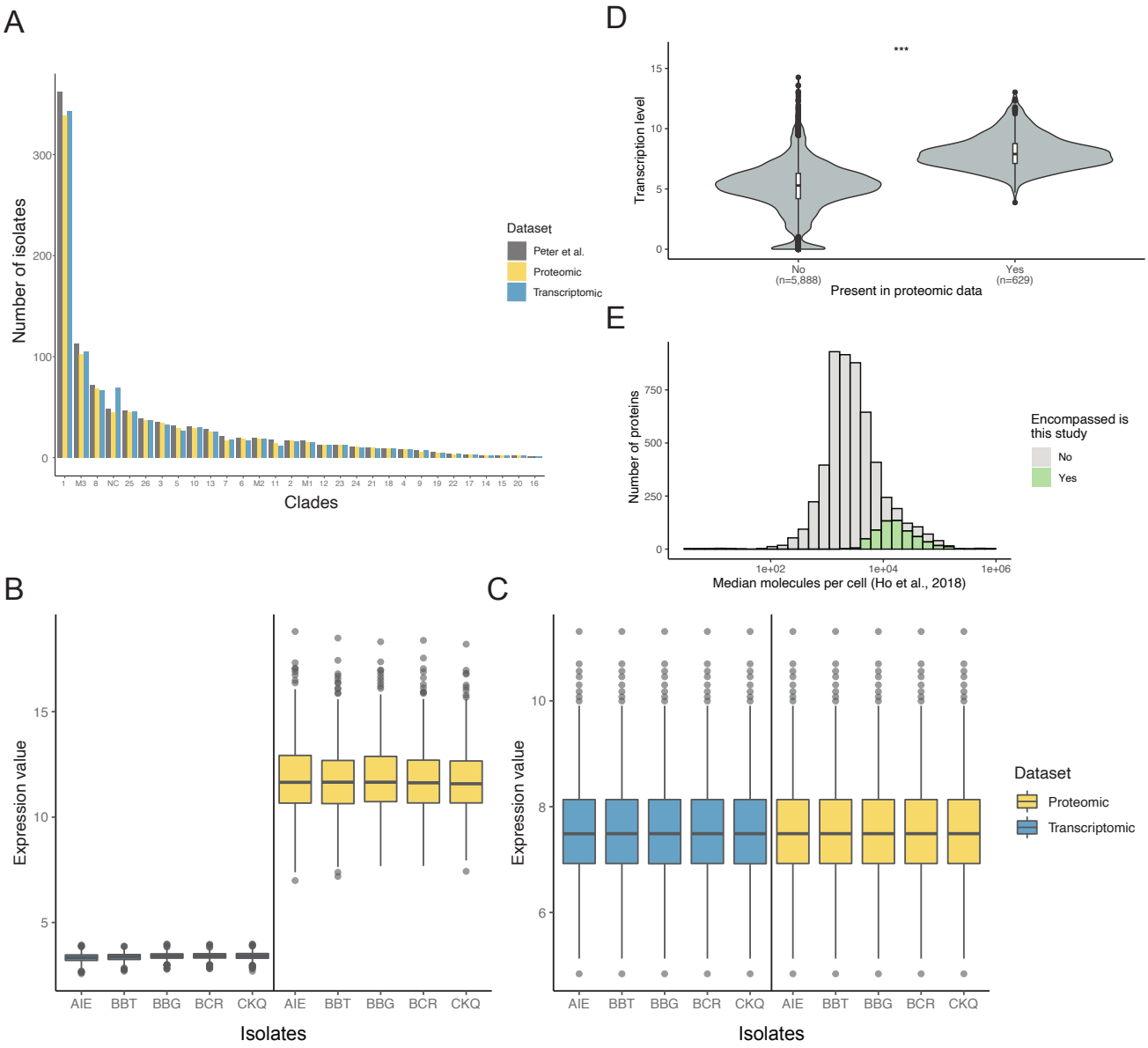


**Figure S1**. **Description of the characteristics and the normalization of the population proteome.**
(A) number of isolates encompassed in the proteomic datasets, in the transcriptomic dataset (1) and in the overall population (2). The x-axis corresponds to the clades (or subpopulations) as defined previously (2). (B, C) Expression values of 5 randomly selected isolates for protein and transcript abundance before (B) and after (C) quantile normalization. (D) mRNA levels of the gene encompassed or not in the proteomic data (*** = p-value $< 2.2 \times 10^{-16}$, Wilcoxon test). (E) Protein levels as defined in (3) of the genes encompassed by our proteomic data.

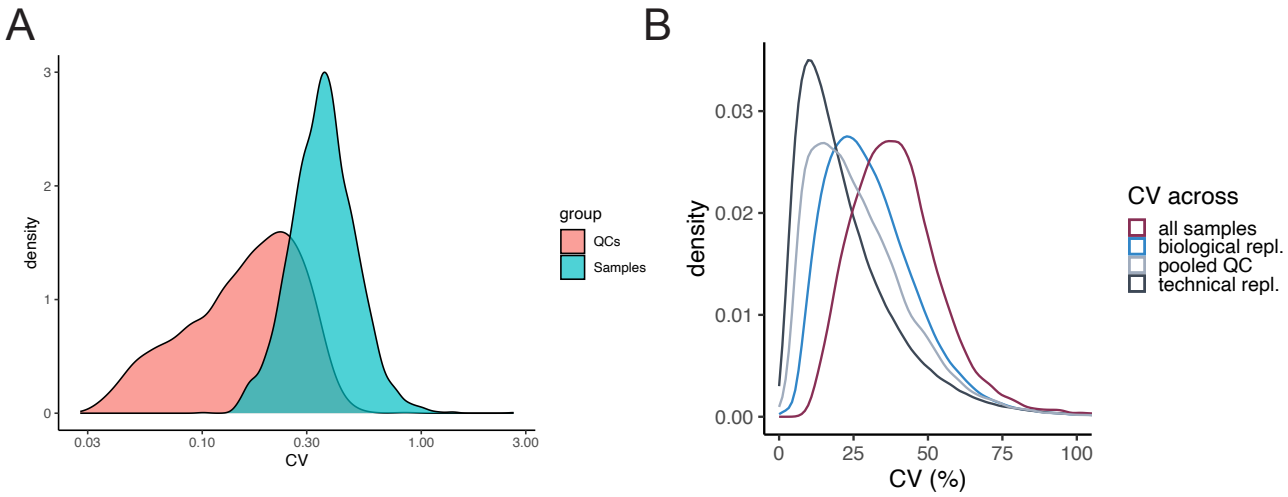


**Figure S2. CV from the QCs and samples precursors.**
(A) The CV was computed using either the QCs set, or the sample set from the population exploration. (B) CV calculated across a set of seven strains cultivated in four biological replicates each, and subsequently measured in technical quadruplicates. CV values are shown between 0-100% to improve the readability of the plot.

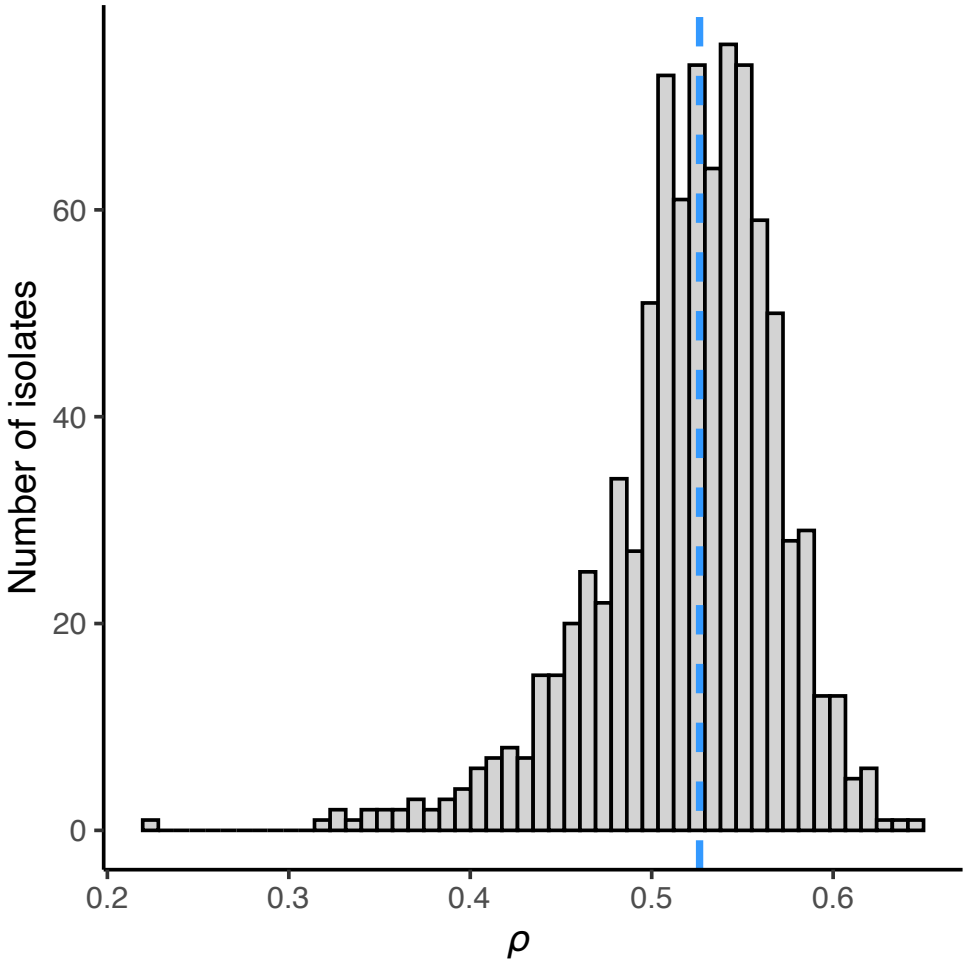


**Figure S3. Across-gene correlation.**
mRNA-protein correlation in each isolate (across-gene correlation). The blue line represents the median (0.53).

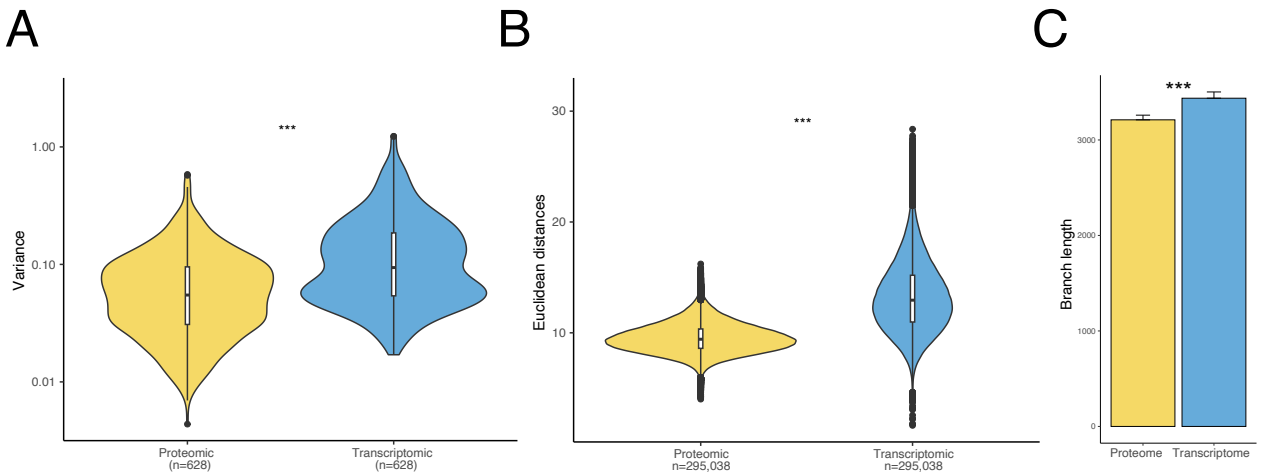


**Figure S4. Detection of the post-transcriptional buffering.**
(A) Comparison between the gene-wise protein and transcript normalized abundance variance (*** = p-value < $2.2x10^{-16}$, Wilcoxon test). (B) Euclidean distances between each isolate using the protein or transcript normalized abundance (*** = p-value < $2.2x10^{-16}$, Wilcoxon test). (C) Branch length difference between the proteome and the transcriptome-based tree. The error bars correspond to 100 bootstrapping steps. We used the bootstrap values to test if the difference in branch length is significant between the two trees (*** = p-value < $2.2x10^{-16}$, Wilcoxon test).

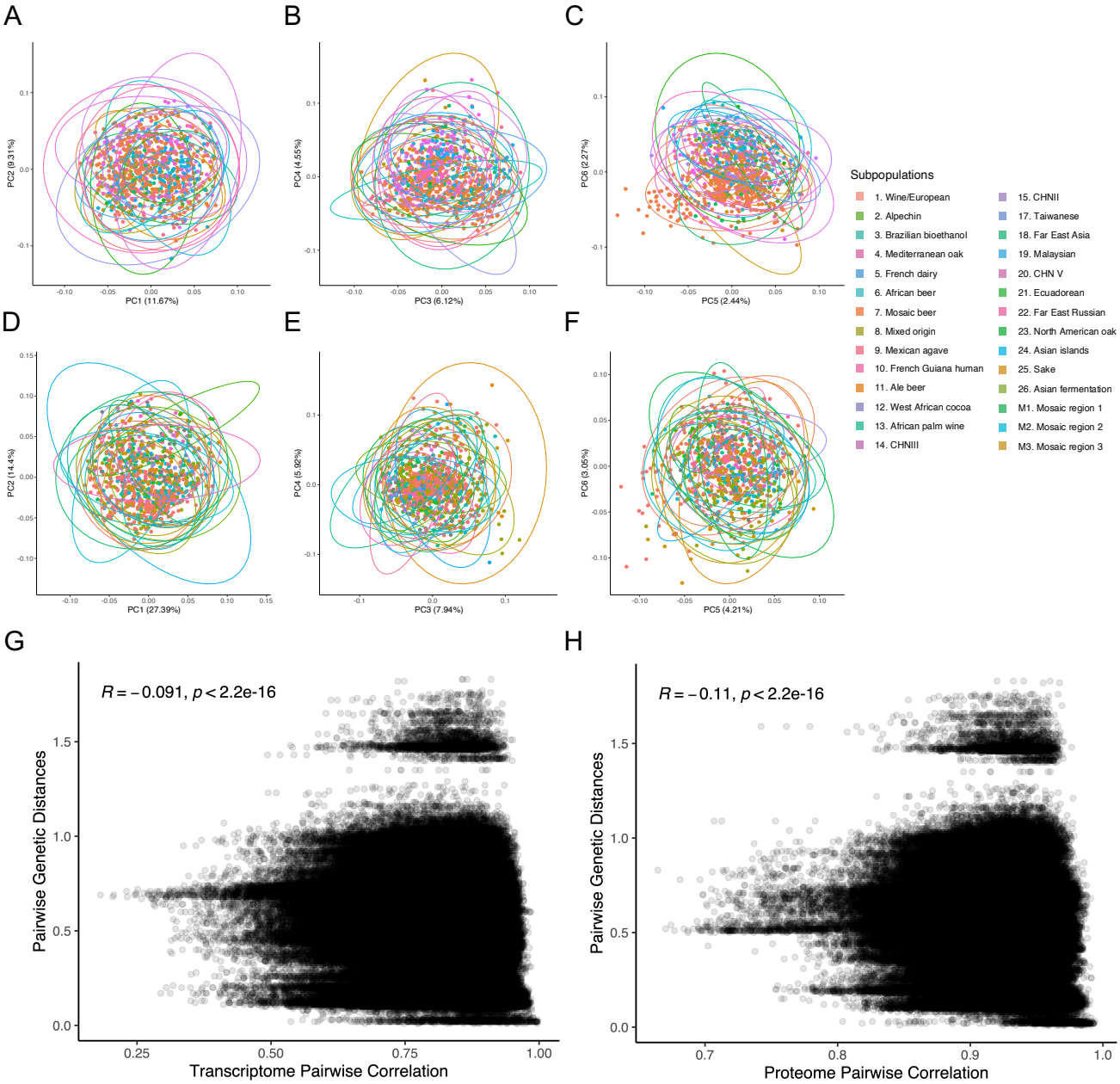


**Figure S5. Population structure is poorly reflected on the proteome and transcriptome.**
PCA using protein (A, B, C) or transcript (D, E, F) abundance. The 6 first PC are plotted together, and the colors correspond to the subpopulations (clades). (G, H) Correlation between the isolate pairwise genetic distance and the isolate pairwise correlation for the transcriptome (G) and the proteome (H).

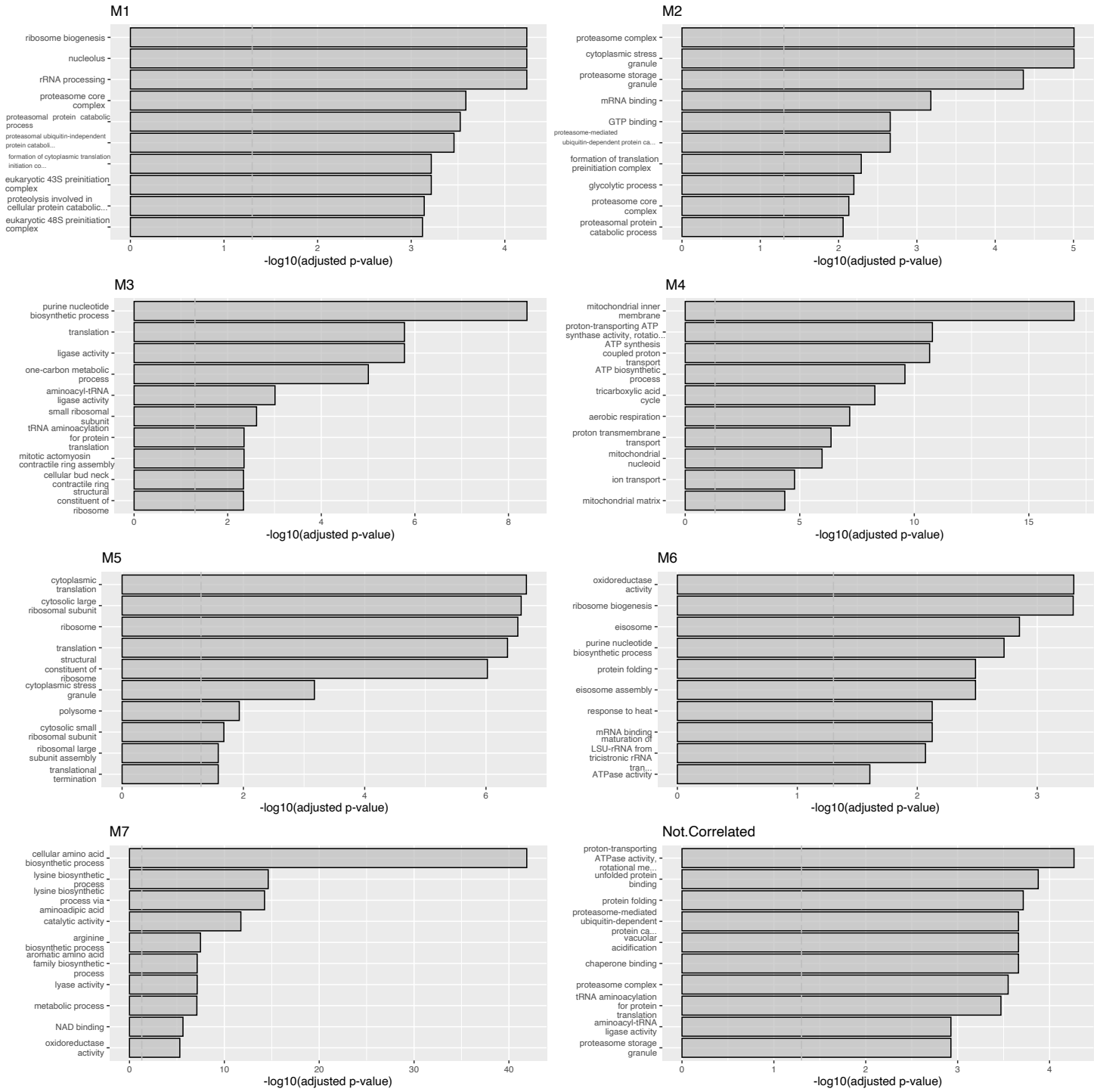


**Figure S6. Functional exploration of the proteome WGCNA modules.**
Functional enrichment of each co-expression module detected using WGCNA on protein abundance data. The enrichment was performed using the CEMiTool package. The dotted lines on each graph represent the significance threshold.

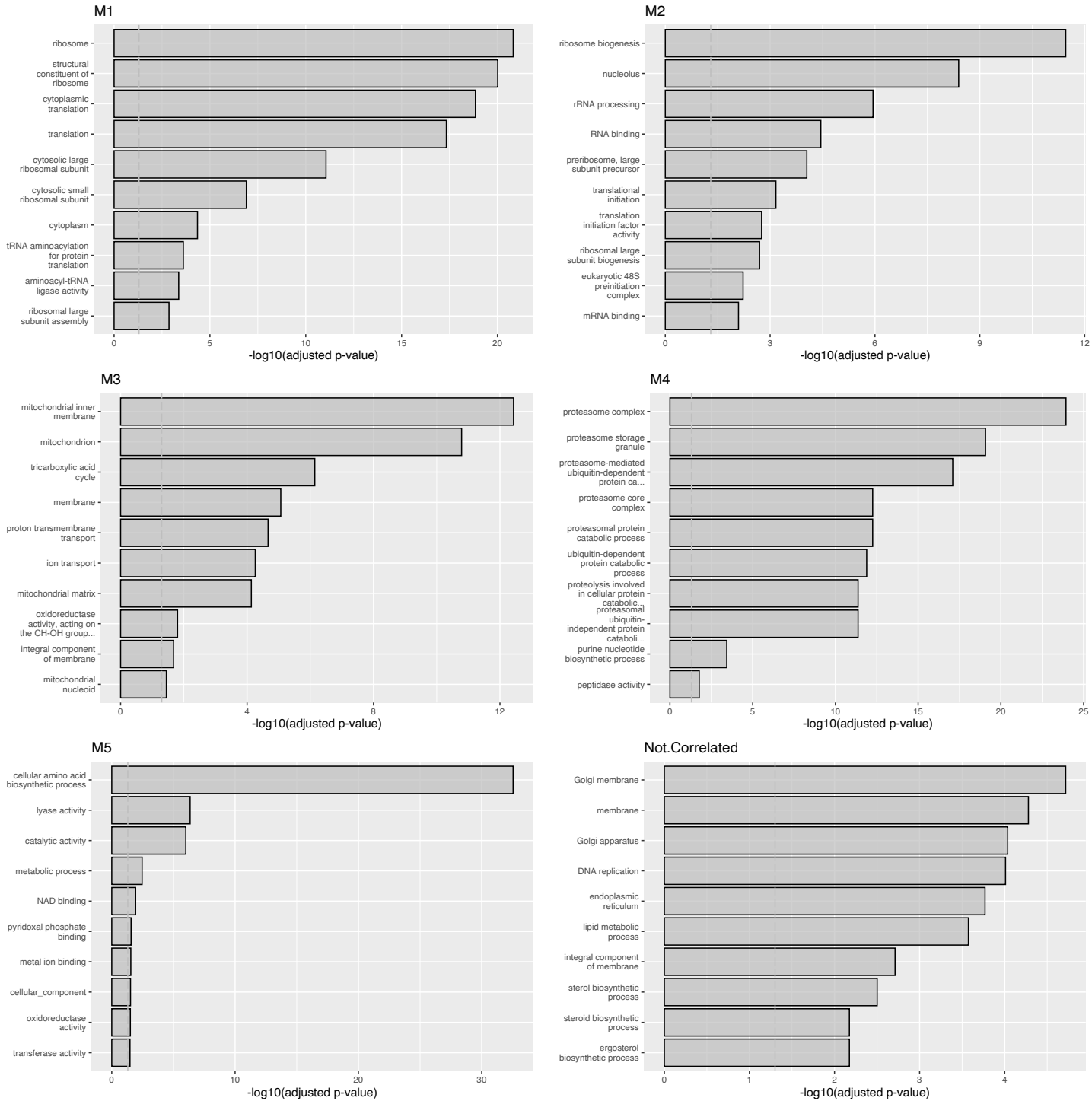


**Figure S7**. **Functional exploration of the transcriptome WGCNA modules.**
Functional enrichment of each co-expression module detected using WGCNA on transcript abundance data. The enrichment was performed using the CEMiTool package. The dotted lines on each graph represent the significance threshold.

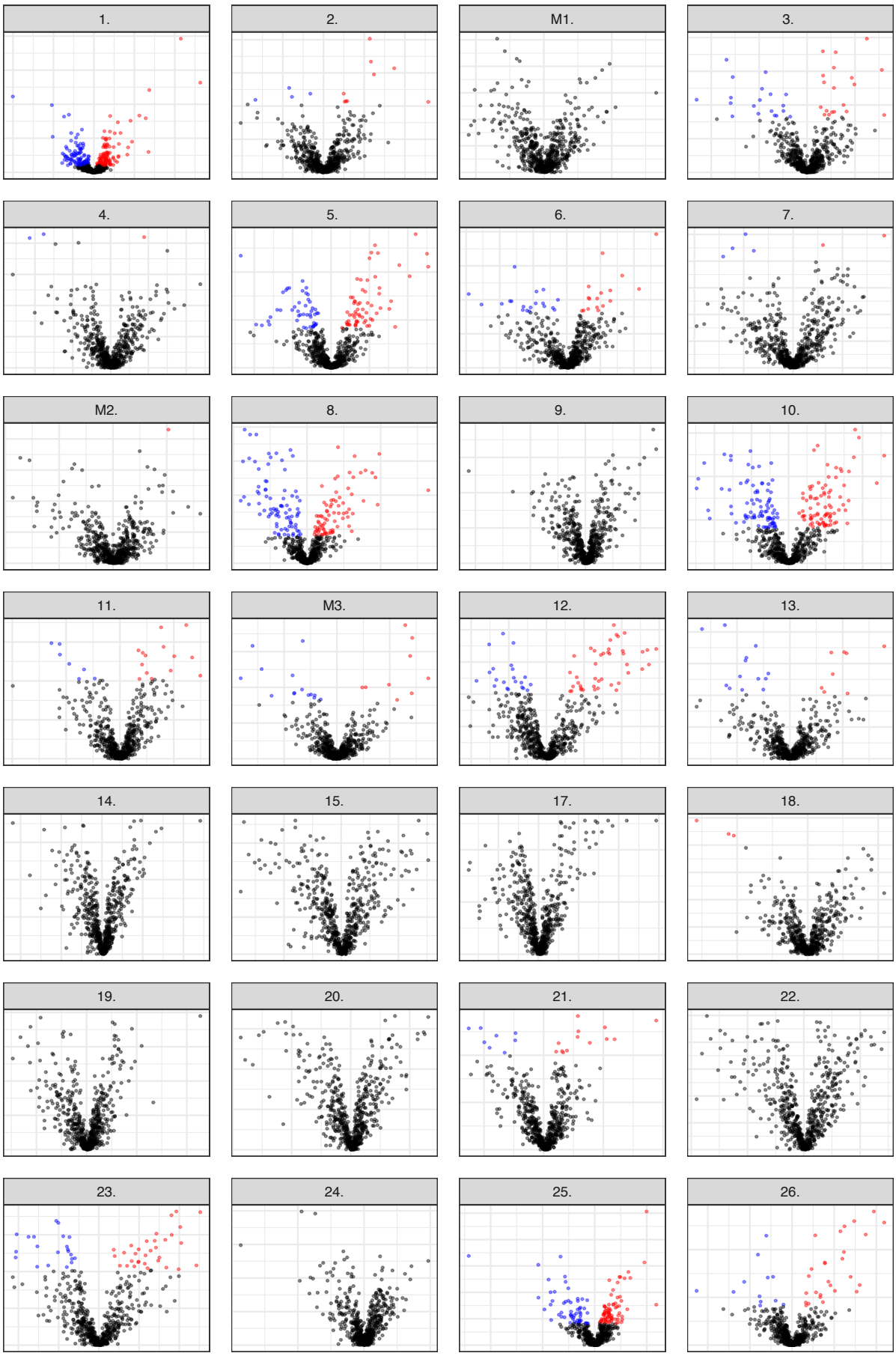


**Figure S8. DEPs detected in each subpopulation.**
Volcano plots for each subpopulation highlighting the DEPs. The blue points correspond to under-expressed gene in a subpopulation while the red points correspond to over-expressed genes.

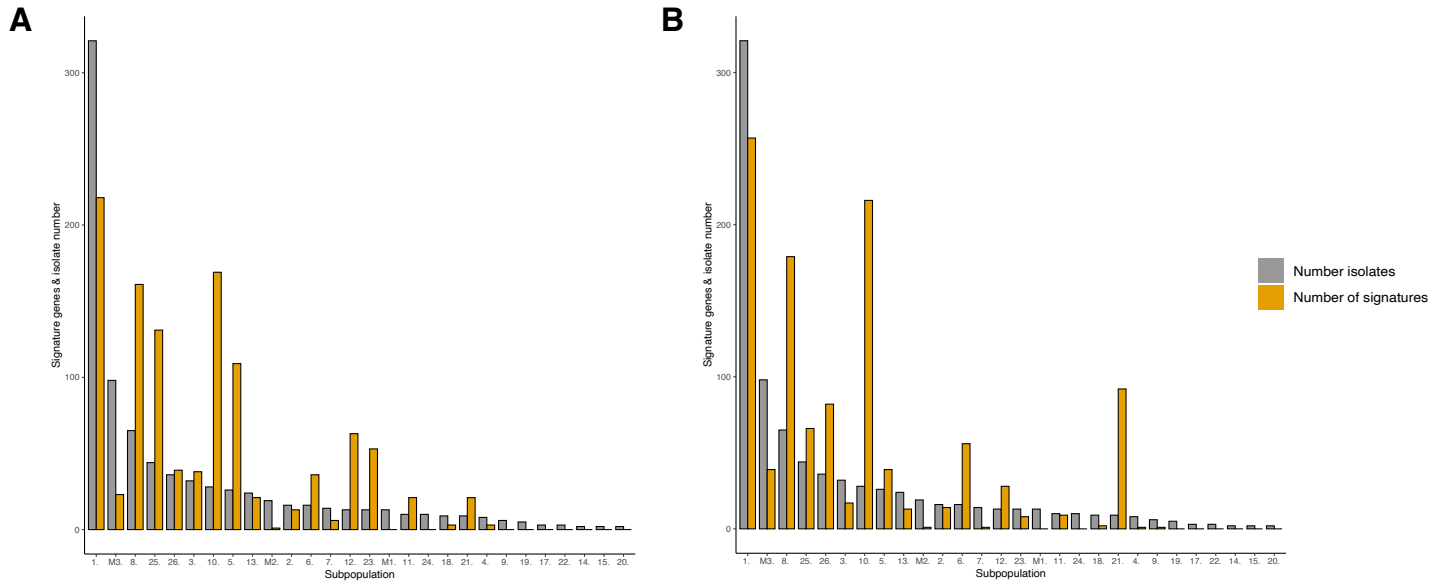


**Figure S9. Number of DEP and differentially expressed transcripts.**
(A, B) Number of proteome (A) and transcriptome (B) DEPs (or differentially expressed transcripts for the transcriptome) in each subpopulation together with the number of isolates in each subpopulation.

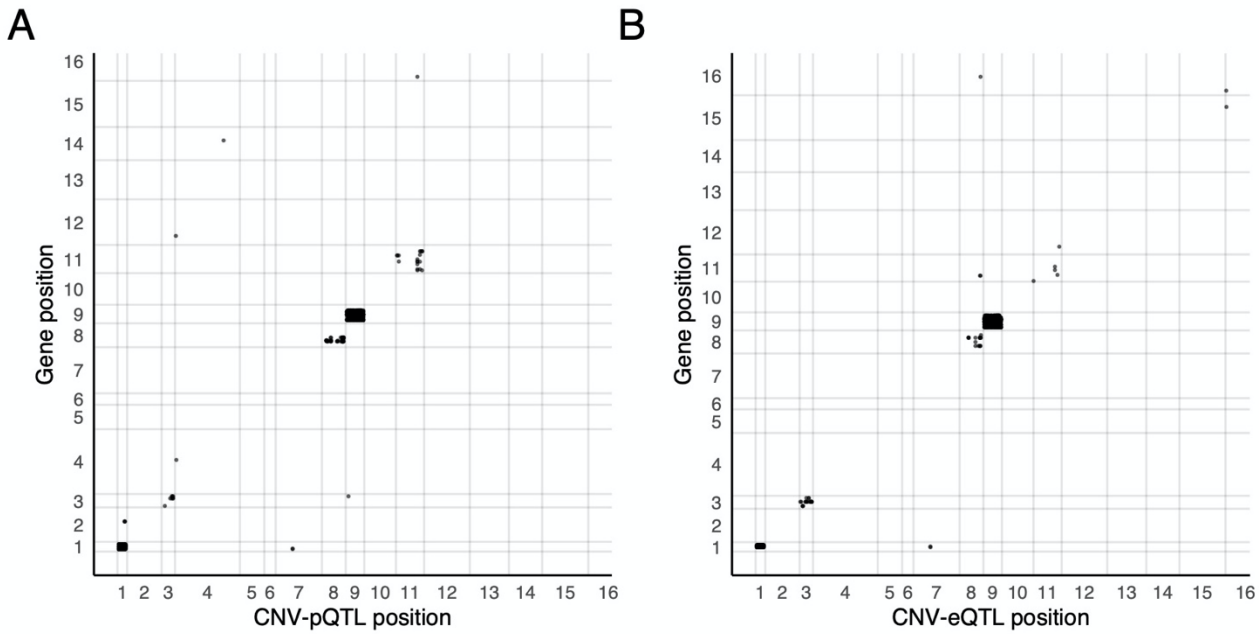


**Figure S10. Genomic location of the CNV-pQTL and -eQTL.**
(A, B) Map of the CNV-pQTL (A) and CNV-eQTL (B) eQTL. The x-axis is the QTL positions on the genome and the y-axis the position of the affected genes on the genome. The x and y-axis numbers represent the 16 chromosomes of *S. cerevisiae*.

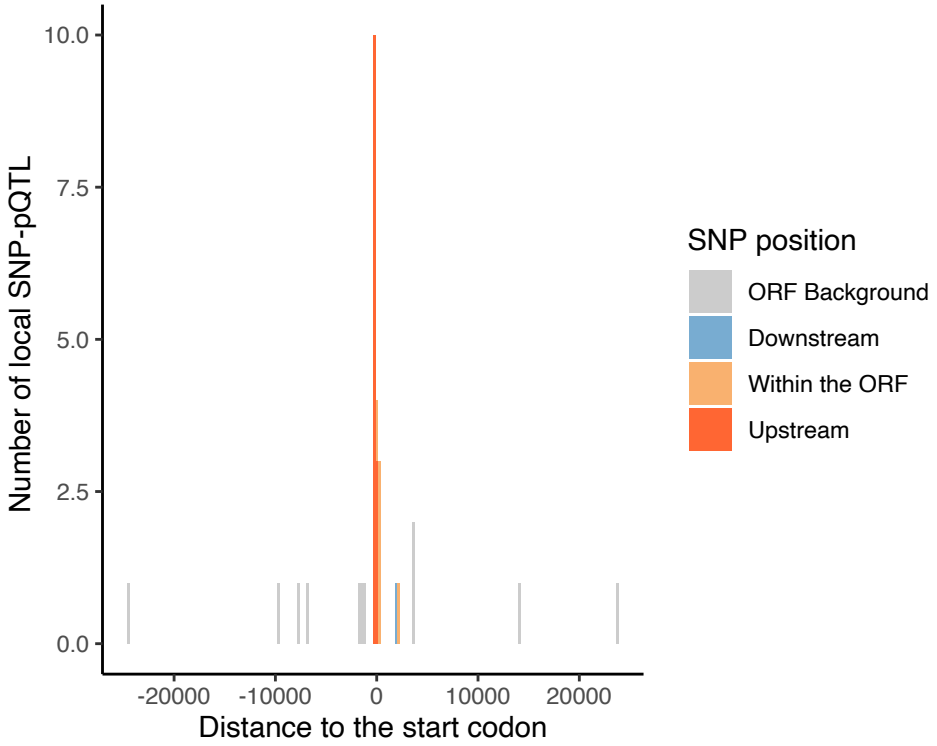


**Figure S11. Location of the local SNP-pQTL.**
Distribution of the local SNP-pQTL around the start codon of their target gene. Downstream pQTL correspond to QTL located between the stop codon and 200 bp after the stop codon, upstream correspond to pQTL located between the start codon and 1,000 bp before the start codon.

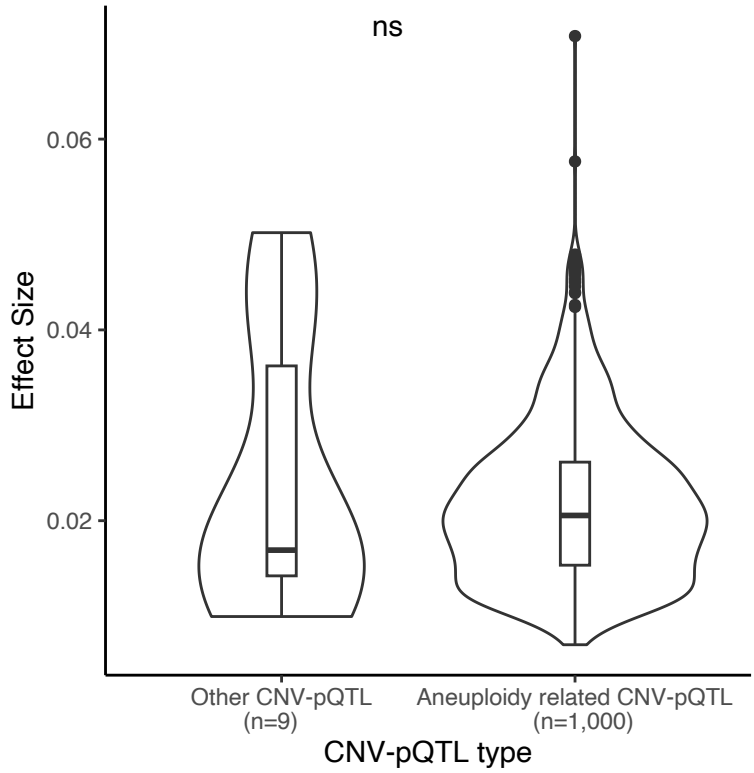


**Figure S12. No difference is observed between the effect size of aneuploidy related CNV-pQTL and the other CNV-pQTL.**
Difference in effect size between the aneuploidy related CNV-pQTL and the other CNV-pQTL (p-value = 0.77, Wilcoxon test).

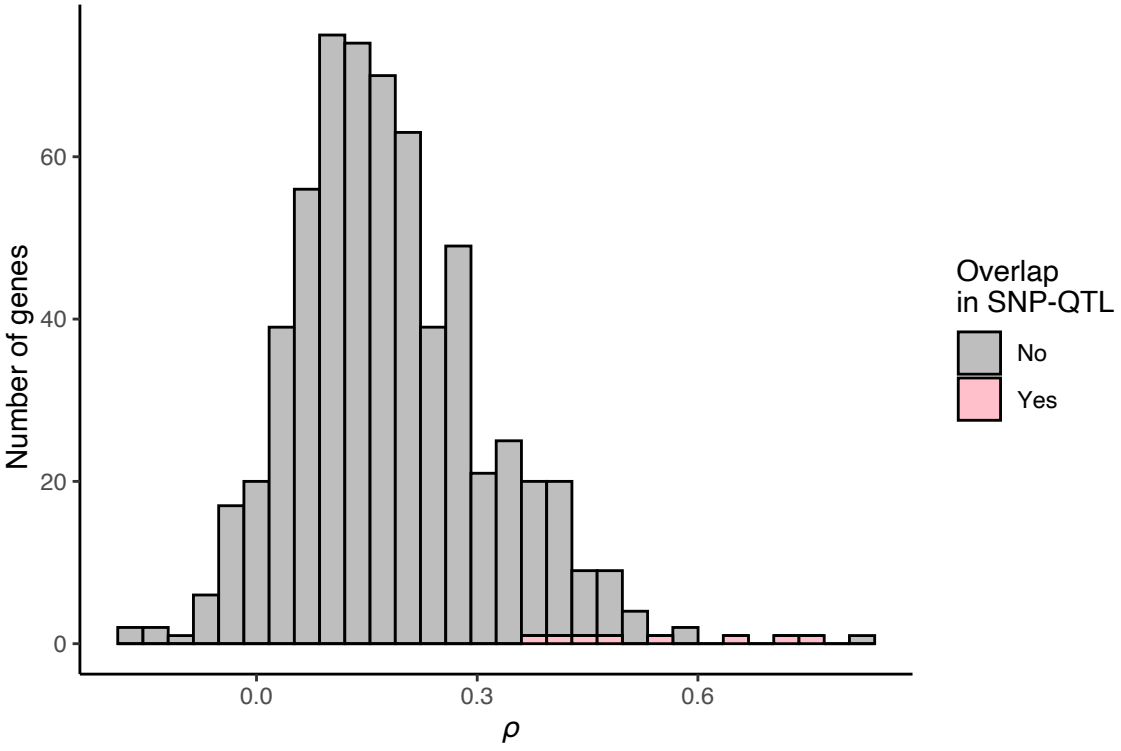


**Figure S13. The genes with an overlapping SNP-QTL tend to have a high within-gene mRNA-protein correlation.**
Within-gene correlation coefficients (Spearman correlation test) between the proteome and the transcriptome. The genes with an overlapping SNP-pQTL and SNP-eQTL are highlighted in pink.

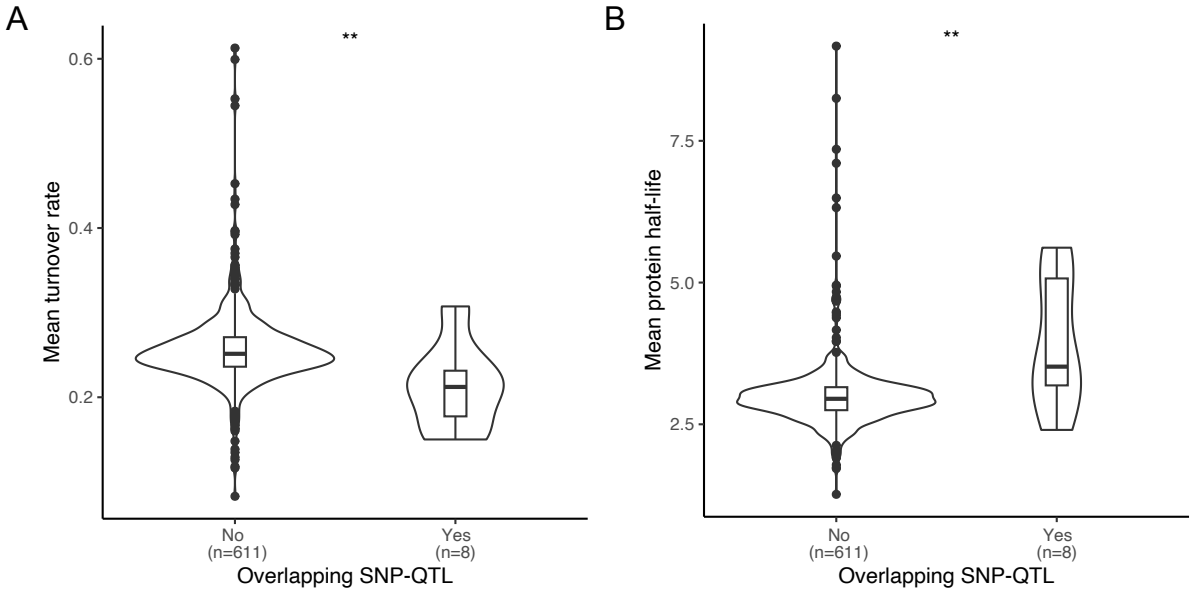


**Figure S14. Turnover rate and half-life of the proteins with or without an overlapping SNP-QTL.** (A, B) The turnover rates (A) and protein half-life values (B) were obtained from (4). The difference was tested using a Wilcoxon test (respective p-values = 0.0057 and 0.0064).

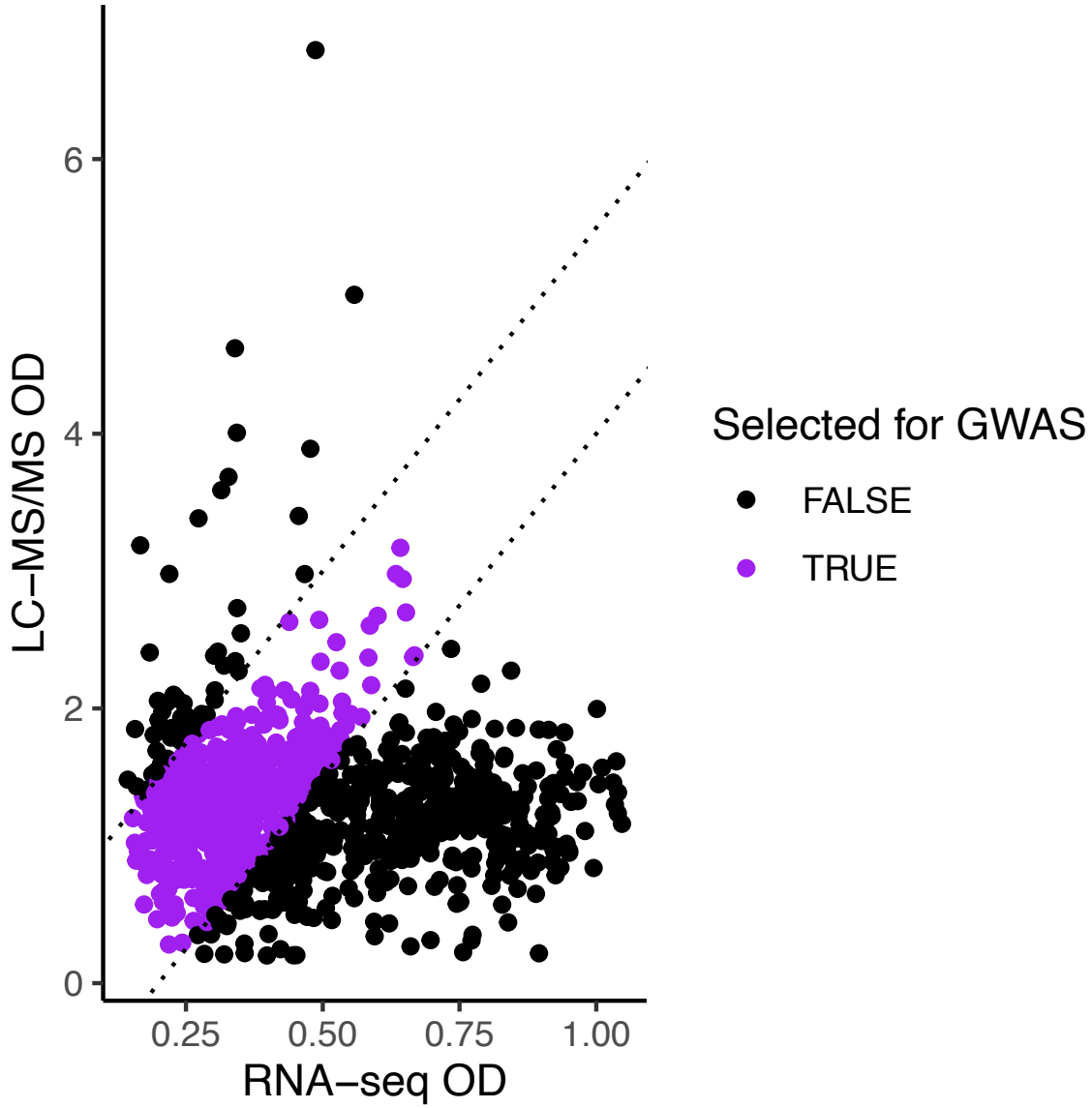


**Figure S15. GWAS were performed using a subset of strain for which harvesting OD were correlated.** Representation of the harvesting OD of the RNA-seq (x-axis) and the LC-MS/MS (y-axis) experiments. The purple points indicate the strains that were selected for both transcriptome and proteome GWAS and that display a high harvesting OD correlation (Pearson correlation coefficient > 0.6).

**Dataset S1.**
Isolates used in this study with their respective OD. Information gathered from (2).

**Dataset S2.**
Proteomic data across the 942 isolates.

**Dataset S3.**
Normalized RNA-seq and proteomic data.

**Dataset S4.**
GO analysis results on the genes encompassed by the proteomic data.

**Dataset S5.**
GSEA results using the raw protein abundance CV.

**Dataset S6.**
Correlation between RNA-seq abundance and protein abundance for each gene.

**Dataset S7.**
GO analysis results on the genes with a significant correlation between RNA-seq and protein abundance.

**Dataset S8.**
GO analysis results on the genes (n=33) with a high correlation between RNA-seq and protein abundance.

**Dataset S9.**
Reduced GO terms used for the branch length exploration.

**Dataset S10.**
Ratio between the proteome and transcriptome tree branch for 101 reduced BP GO term.

**Dataset S11.**
Module detected for each gene in the proteomic data.

**Dataset S12.**
GO enrichment for each module (obtained with CEMiTool) detected using proteomic data.

**Dataset S13.**
Module detected for each gene in the RNA-seq data.

**Dataset S14.**
GO enrichment for each module (obtained with CEMiTool) detected using RNA-seq data.

**Dataset S15.**
Over and underexpressed proteins in each subpopulation (2).

**Dataset S16.**
GSEA analysis on the DEG detected using the normalized proteomic data. If log10pval is missing, it means that the pathway is not significant.

**Dataset S17.**
DEG in the domesticated vs wild comparison for both RNA-seq and proteomic data.

**Dataset S18.**
GSEA analysis on the DEG detected in the domesticated vs wild comparison (using proteomic data).

**Dataset S19.**
GSEA analysis on the DEG detected in the domesticated vs wild comparison (using transcriptomic data).

**Dataset S20.**
SNP GWAS results for the transcriptomic and proteomic GWAS (local: SNP located 25kb before or after start).

**Dataset S21.**
CNV GWAS results for the transcriptomic and proteomic GWAS (local: the causal CNV affects its own gene expression).

**Dataset S22.**
Results for the SNP based proteome GWAS.

**Dataset S23.**
Results for the CNV based proteome GWAS.

**Dataset S24.**
Results for the SNP based transcriptome GWAS.

**Dataset S25.**
Results for the CNV based transcriptome GWAS.

**Dataset S26.**
CVs computed on either the QCs set or the sample set.

**SI References**


1. E. Caudal, *et al.*, Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *BioRxiv* 2023.05.17.541122 (2023). https://doi.org/10.1101/2023.05.17.541122.

2. J. Peter, *et al.*, Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).

3. B. Ho, A. Baryshnikova, G. W. Brown, Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **6**, 192-205.e3 (2018).

4. J. Muenzner, *et al.*, The natural diversity of the yeast proteome reveals chromosome-wide dosage compensation in aneuploids. *BioRxiv* 2022.04.06.487392 (2022). https://doi.org/10.1101/2022.04.06.487392.