Supporting Information

Differences in the path to exit the ribosome across the three domains of life

Khanh Dao Duc[1], Sanjit S. Batra[1], Nicholas Bhattacharya[2], Jamie H. D. Cate[3,4,5], and Yun S. Song[1,6,7,*]

[1] Computer Science Division, University of California, Berkeley, CA 94720
[2] Department of Mathematics, University of California, Berkeley, CA 94720
[3] Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720
[4] Department of Chemistry, University of California, Berkeley, CA 94720
[5] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[6] Department of Statistics, University of California, Berkeley, CA 94720
[7] Chan Zuckerberg Biohub, San Francisco, CA 94158

* To whom correspondence should be addressed: yss@berkeley.edu

February 6, 2019

This document contains:

- Detailed methods on the processing of the tunnel coordinates, the phylogenetic analysis of 16S/18S rRNA sequences (Figure S2A), and alternative method to visualize the tunnel (Figure S9).
- A study of the robustness of the tunnel geometry clustering, with regards to replicates, parameters of the tunnel geometric distance and computational method of tunnel extraction.
- Supplementary Table S1.
- Supplementary Figures S1 to S8.

1 Supplementary methods

1.1 Tunnel data processing

We detail here the processing of the exit tunnel coordinates and radius. Coordinates and radius of the tunnel were first extracted using Pymol and Python custom scripts. The tunnel centerline was parameterized by its arc length and the tunnel radius plot along the arc length was smoothened using the spline `fit` function in Matlab (smoothing parameter 0.07). The fitting of the tunnel centerline to a line was done using Matlab. To compute the distance of a given point M to the tunnel and locate where to locate it with respect to the tunnel (as in Figure 6), we first computed, for any given tunnel centerline point C, the Euclidean distance between M and C and then subtracted the tunnel radius associated with C. The distance and position of M to the tunnel was then obtained by minimizing this value over all the tunnel centerline. All codes and scripts used are available upon request.

1.2 Alternative method for tunnel extraction

In addition to the main method used to extract the tunnel geometry (see Methods section), we used a "sphere-filling" method called *Hollow* for visualizing cavities of atomic structures [1]. We applied *Hollow* with default values (specifying as input parameters a "cylinder" of 40 Å diameter, starting from the PTC and ending at the exit port). The coordinates of the output spheres (see Figure S9**a**) were exported and analyzed with Matlab. The tunnel structure was refined by compiling the spheres connected to the tunnel obtained from the main method (see Figure S9**b**). The associated volume (Figure S9**c**) was computed by using the α-shape function in Matlab, with default parameters.

1.3 Phylogenetic analysis of 16S/18S rRNA sequences

We describe here the methods used to carry out the phylogenetic tree inference given in Figure S2a. We first obtained the 16S rRNA sequences for bacteria and archaea and 18S rRNA sequences for eukarya from NCBI, by searching for the relevant sequence in the "Nucleotide" category. These sequences were aligned with MAFFT [2] with default parameters. The Neighbor-Joining method available within MEGA7 [3] was then used to infer the phylogenetic tree based on the aligned 16S/18S rRNA sequences. The evolutionary distances were computed using the Maximum Composite Likelihood method in the units of the number of base substitutions per site. Codon positions included were 1st + 2nd + 3rd + Noncoding.

2 Robustness of the clustering

We studied the robustness of the tunnel geometric analysis with respect to 1) the computational method to extract and represent the tunnel geometry 2) replicates of same species from different cryo EM and X-ray datasets, and 3) parameters of the tunnel geometric distance.

2.1 Robustness to computational method

To assess how our geometric representation of the tunnel (using 3D coordinates of the centerline plus radius) was robust to more detailed representations, we applied an alternative method that provides high resolution visualization of the tunnel (see Supplementary Methods). More precisely, we used the visualizing data of the empty space in the neighborhood of the tunnel (Figure S9**a**) to extract additional surface points of the tunnel (Figure S9**b**). Whereas our original representation tends to

underestimate the true volume of the tunnel, this new representation is likely to overestimate the volume accessible to the nascent polypeptide chain, since it accounts for small non convex local regions (in addition of being sensitive to the noise and resolution of the EM structure). For five species representative of our dataset (2 eukaryotes, 1 archaea and 2 prokaryotes), we compared the associated volume with the volume found with the original method (Figure S9**c**). We found a very high correlation between the two volumes (Pearson's $r > 0.99$, p-value $< 10^{-5}$), leading to the same ordering of species. The additional volume constituted in average $5.85 \pm 0.6\%$ of the total. Upon approximating the tunnel by a cylinder of constant radius, this volume represented an average radius increase of 0.17 ± 0.01Å. Overall, we concluded that our quantitative analysis of the geometric features and the subsequent clustering analysis was robust with respect to the computational method to extract and represent the tunnel geometry (see also Discussion).

2.2 Robustness to replicates

In addition to the main structures of our dataset, we performed the same clustering analysis as in Figure S2c, with replicate structures from *E.coli*, *H. sapiens* and *T. thermophilus* (see Table 1). Such structures, which present a slightly higher resolution than the main ones, were also chosen to come from different labs. As shown in Figure S1a, we obtained a clustering of the main and replicate structures for *T. thermophilus* and *H. sapiens*, while the *E. Coli* structures formed a cluster with *B. subtilis*. We directly compared the tunnel radius plots of these structures in Figure S1b, showing very similar profiles that explain their clustering. Interestingly, most of the difference between the *E. coli* main and replicate tunnel radius variations can be located in the upper part of the tunnel. Since the replicate structure of *E. coli* originally contained the ErmCL nascent chain, which is 19 amino acids long and also occupies the upper part of the tunnel [4], this suggests that the nascent chain and arrest sequences in particular, can potentially affect the geometry of the tunnel. Compared to the hierarchical clustering obtained without replicates (Figure S2c), we also noticed in Figure S1a only minor differences, with the human mitoribosome and *P. falciparum* moving to their neighboring branch. Overall, we conclude that the clustering obtained from comparing the tunnel radius variation plots is robust to replicate structures coming from different experimental datasets.

2.3 Robustness to the choice of tunnel comparison metric

After evaluating the robustness of the clustering to replicate structures, we did the same with regards to the parameters of the distance introduced to compare the radius plots. For two tunnels T_1 and T_2 parametrized by $T_i = (\bar{S}^{(i)}, \bar{R}^{(i)})$, where $i = 1, 2$, $\bar{S}$ is an arc length parametrization of the tunnel centerline 3D coordinates and $\bar{R}$ is the associated radius, we recall that we define the distance between T_1 and T_2 as

$$D(T_1, T_2) = \min_{|\delta| \leq \ell} d_\delta(T_1, T_2), \tag{1}$$

where ℓ is the maximum shift length and

$$d_\delta(T_1, T_2) = \frac{\displaystyle\int_{\bar{S}_\delta^{(1)} \cap \bar{S}^{(2)}} \left[\bar{R}_\delta^{(1)}(s) - \bar{R}^{(2)}(s)\right]^p ds}{\displaystyle\int_{\bar{S}_\delta^{(1)} \cap \bar{S}^{(2)}} ds} + \varepsilon|\delta|, \tag{2}$$

where $\bar{S}_\delta^{(1)} = \bar{S}^{(1)} - \delta$, and $\forall s \in \bar{S}$, $\bar{R}_\delta(s - \delta) = \bar{R}(s - \delta)$. The parameters associated with the distance D are ℓ (maximum shift), ε (penalty coefficient for the plot shift) and p ($\mathcal{L}^p$ norm coefficient).

To study if our main results are stable with regards to these parameters, we compared the clustering of species obtained in Figure 3b to the ones obtained for different values of parameters (ℓ, ε, p), by computing their Rand index, which gives the probability that two clusterings agree on a randomly chosen pair of elements [5]. The role of the ℓ term is to prevent comparisons of tunnel radius plots over a too small intersecting domain. In practice, this term can also be redundant with the penalty term ε, and we observed that with our default choice of ε (0.01) and upon varying ℓ from 0 to 40 Å, the clustering of Figure 3b was still conserved (so the Rand index is constant, equal to 1). In Table S1, we provide the Rand index obtained for different values of ε and p, showing high similarity to the default clustering. For example, upon varying ε by several orders of magnitude (10^{-3} to 1), we still observed at least 93% agreement of the default pairings, with the variation coming from *P. falciparum* being clustered with trypanosomes, and *E. coli* breaking the bacterial cluster structure into 3 parts (Data not shown). Removing the penalty term ($\varepsilon = 0$) led to 87% agreement, with *M. tuberculosis* clustering apart in the bacterial branch. Similar changes were observed when varying the parameter p, with more than 93% of agreement obtained for $p = 3$ and 4. Using the $\mathcal{L}^1$ norm affects the clustering of *S. cerevisiae*, which clusters with archaeal species, hence decreasing the Rand index to 84%. Therefore, as varying the parameters of the tunnel distance affects the clustering of few species in our dataset, we conclude that the global structure and separation between the different domains of life observed in our main analysis remain well conserved and robust against the tunnel distance parameters.

SUPPORTING TABLE

Table S1. Sensitivity of the clustering to the parameters of the tunnel geometric distance. We compare the clustering obtained in Figure 3b to the one obtained after varying parameters of the tunnel geometric distance by measuring their corresponding Rand index, which gives the probability that two clusterings will agree on a randomly chosen pair of elements (see SI Text). These parameters are ε (penalty term for shifted alignment) and p (coefficient of the $\mathcal{L}^p$ norm) (see Methods and SI text). The upper (lower) table gives the Rand index for different values of ε (p), while other parameters are set to their default values.

ε	0	10^{-3}	10^{-2}	10^{-1}	10^{0}
Rand index	0.87	0.96	1	0.93	0.93

p	1	2	3	4
Rand index	0.84	1	0.93	0.96

SUPPORTING FIGURES

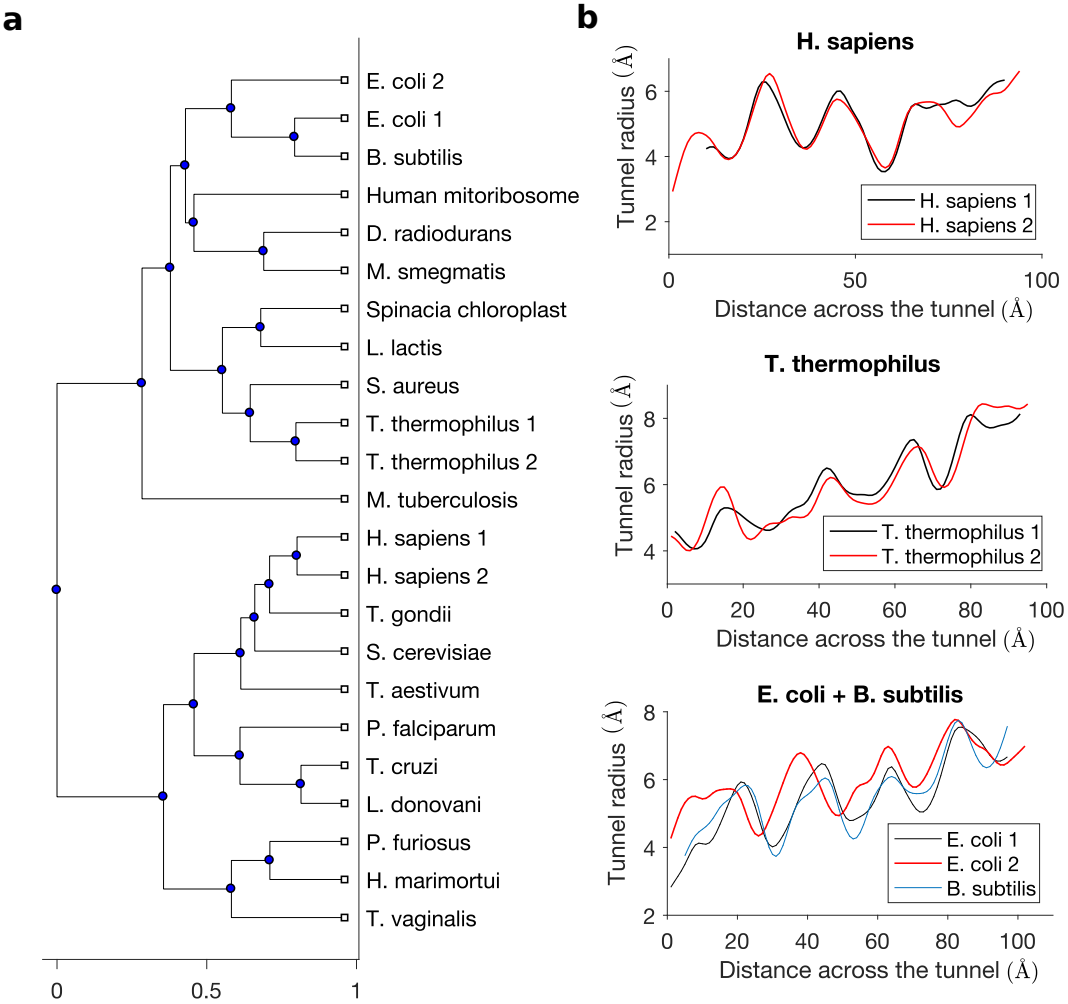

Figure S1. Hierarchical clustering of tunnel geometries with structure replicates from same species. a: We show the clustering tree obtained from the tunnel geometric distance for our main set of species, plus replicate structures (labeled with index number 2) from *H. sapiens*, *T. thermophilus* and *E. coli* (see Table 1 and SI Text). Distance unit (x-axis) is in ångström (see Methods). **b:** We compare the radius variation plots of the structure replicates, aligned after evaluating the tunnel geometric distance (see Methods). For *E. coli* (bottom plot), we include *B. subtilis* radius plot, as it is more similar to the main *E. Coli* than the replicate. Such a discrepancy may be due to the ErmCL nascent arrest chain contained in the replicate structure [4], affecting the geometry of the upper part of the tunnel.

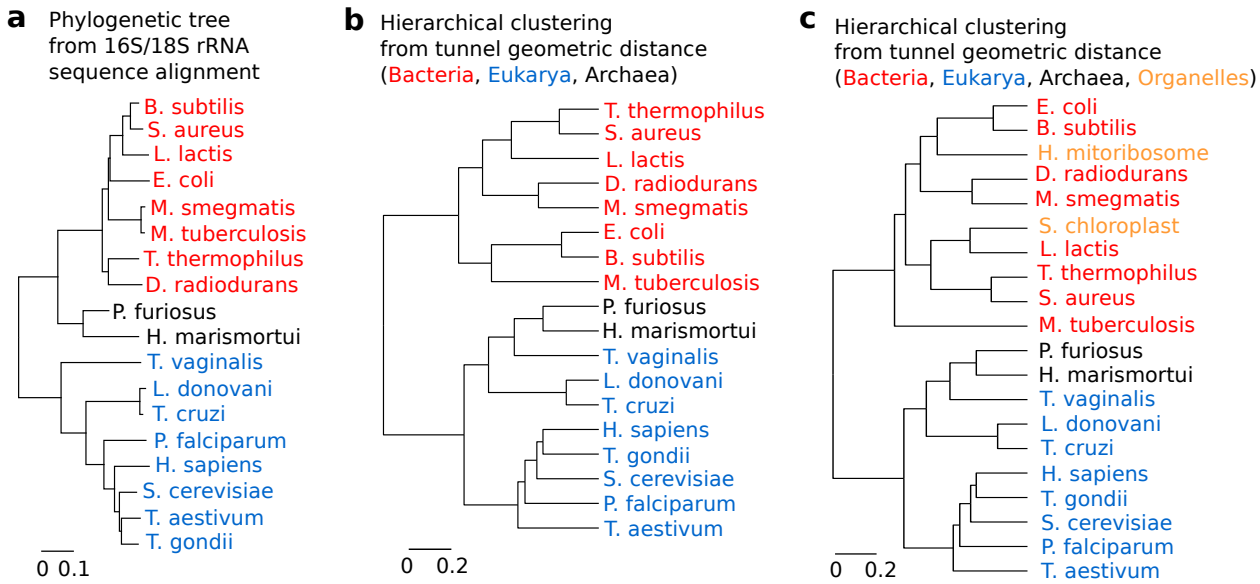

Figure S2. Clustering of species obtained from tunnel geometric distance and RNA sequence analysis. a: We represent the phylogenetic tree inferred from 16S/18S rRNA sequences for bacteria, eukarya and archaea species of our dataset, respectively shown in red, blue and black (scale-bar unit: number of base substitutions per site). **b:** Full clustering tree obtained after applying our tunnel geometric distance metric to the same set of species as in **a** (scale-bar unit: ångström; see Methods). **c:** We show the clustering tree (scale-bar unit: ångström) obtained from the tunnel geometric distance after including the two ribosomes from organelles in our dataset (yellow).

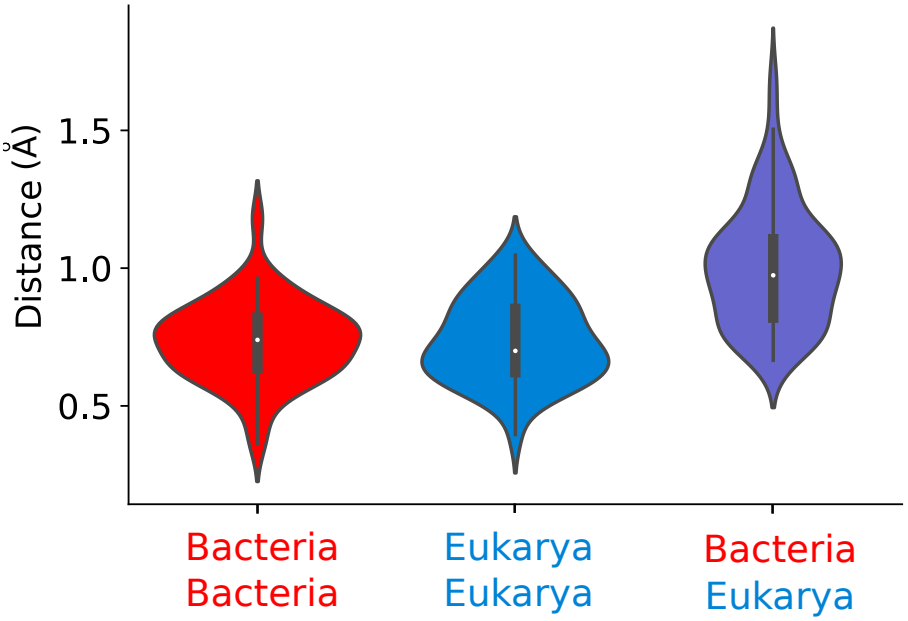

Figure S3. The tunnel geometric distance for intra and inter-domain pairs. We consider the subsets of tunnel geometric distance (see Methods and Figure 3) obtained by considering only 1) couples of bacteria (red), 2) couples of eukarya (blue) and mixed couples of bacteria and eukaryote (violet). The violin plots represent the distribution of pairwise distance for each of the subsets.

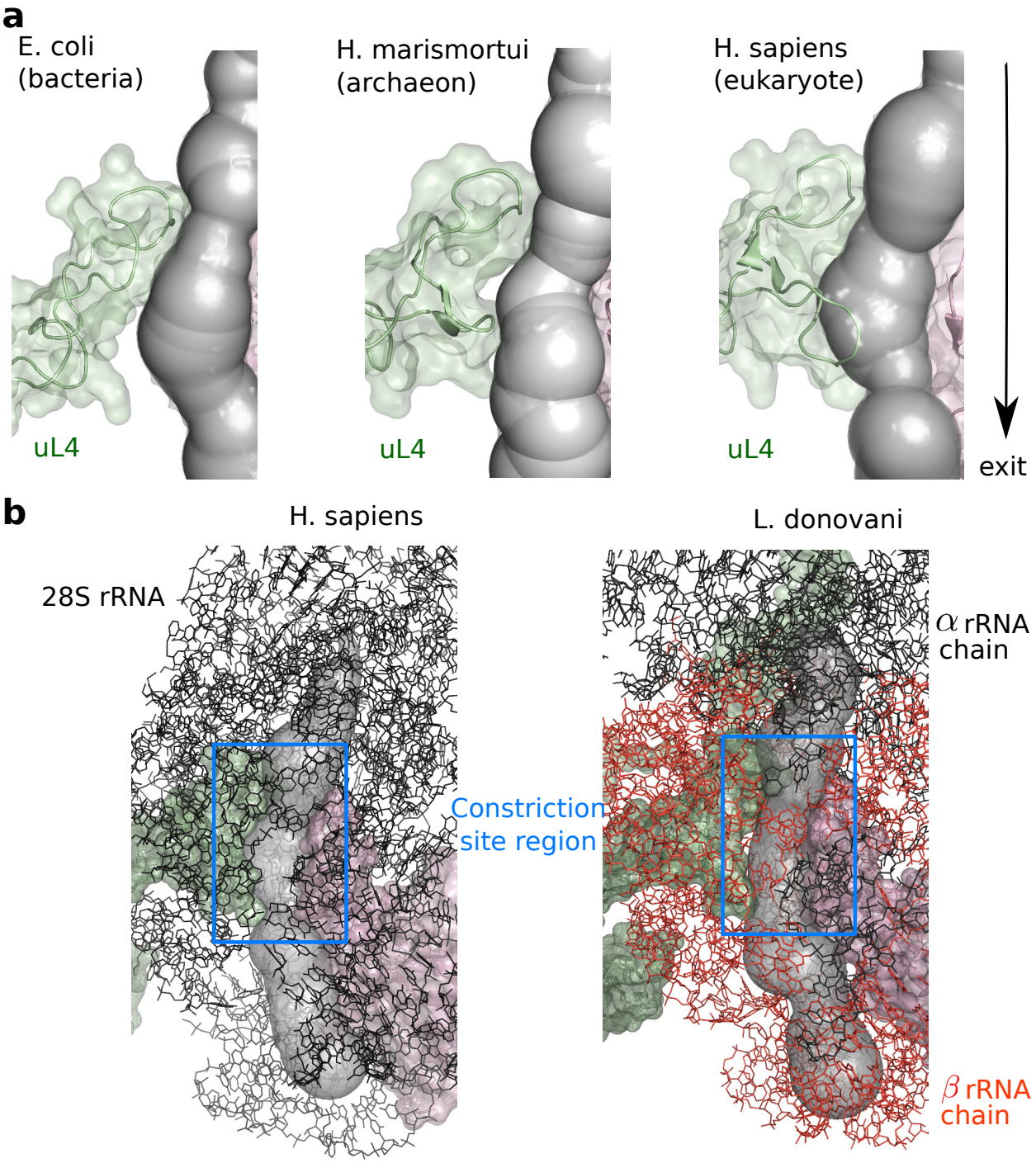

Figure S4. The structure of the ribosome at the constriction site region. a: For *E. coli* (in left, obtained from Fischer *et al.* [6]), *H. marimotui* (in middle, obtained from Gabdulkhakov *et al.* [7]), and *H. sapiens* (in right, obtained from Natchiar *et al.* [8]), we show how the shape of uL4 and the tunnel vary at the constriction site, with the presence of a supplementary arm in *H. sapiens*, which is less prominent in *H. marismortui* and absent in *E. coli*. **b:** We show the structure of the rRNA surrounding the tunnel within a range of 20 Å in *H. sapiens* (left) and *L. donovani* (right). In *L. donovani*, the LSU rRNA breaks from the standard 28S rRNA into six smaller chains, with the two main ones (α and β chains) notably joining around the constriction site.

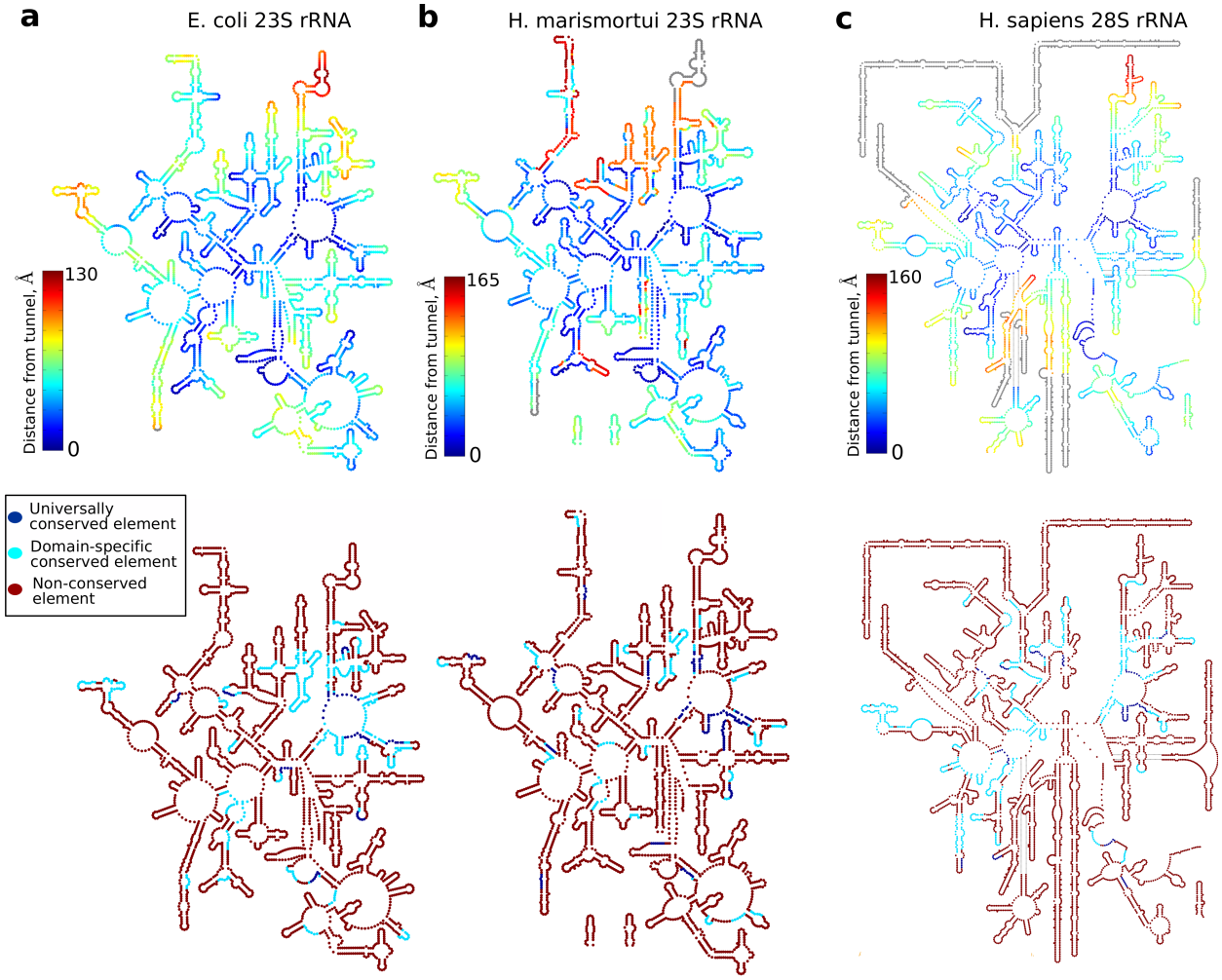

Figure S5. Distance from the tunnel and conservation of the ribosomal large subunit RNA. For *E. coli* (**a**), *H. marismortui* (**b**) and *H. sapiens* (**c**), we show maps of the secondary structure of the LSU RNA, colored by the distance from tunnel (up) and conserved elements (down), as defined by Doris *et al.* [9].

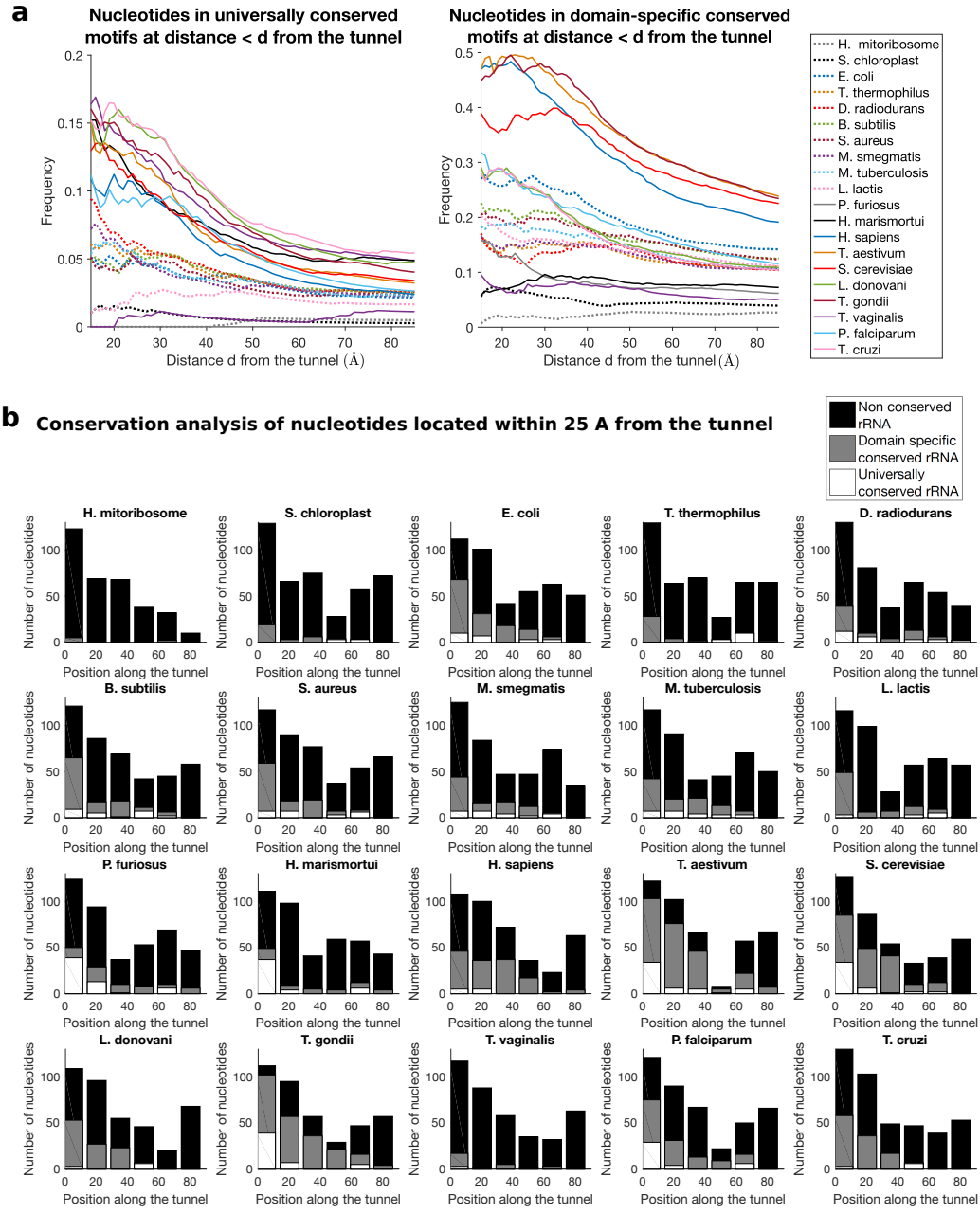

Figure S6. Conservation of rRNA nucleotides at the tunnel. a: As in Figure 6b, we plot for all the species of our dataset, and as a function of distance d, the frequencies of universally (in left) and domain-specific (in right) conserved elements, located within a range d from the tunnel. **b:** For each species of our dataset, we divide the tunnel into different regions along the centerline (each region covering 15 Å), look at the closest rRNA nucleotides located within 25 Å from the tunnel, and plot the associated number of conserved, domain-specific and universally conserved nucleotides, as defined by Doris *et al.* [9].

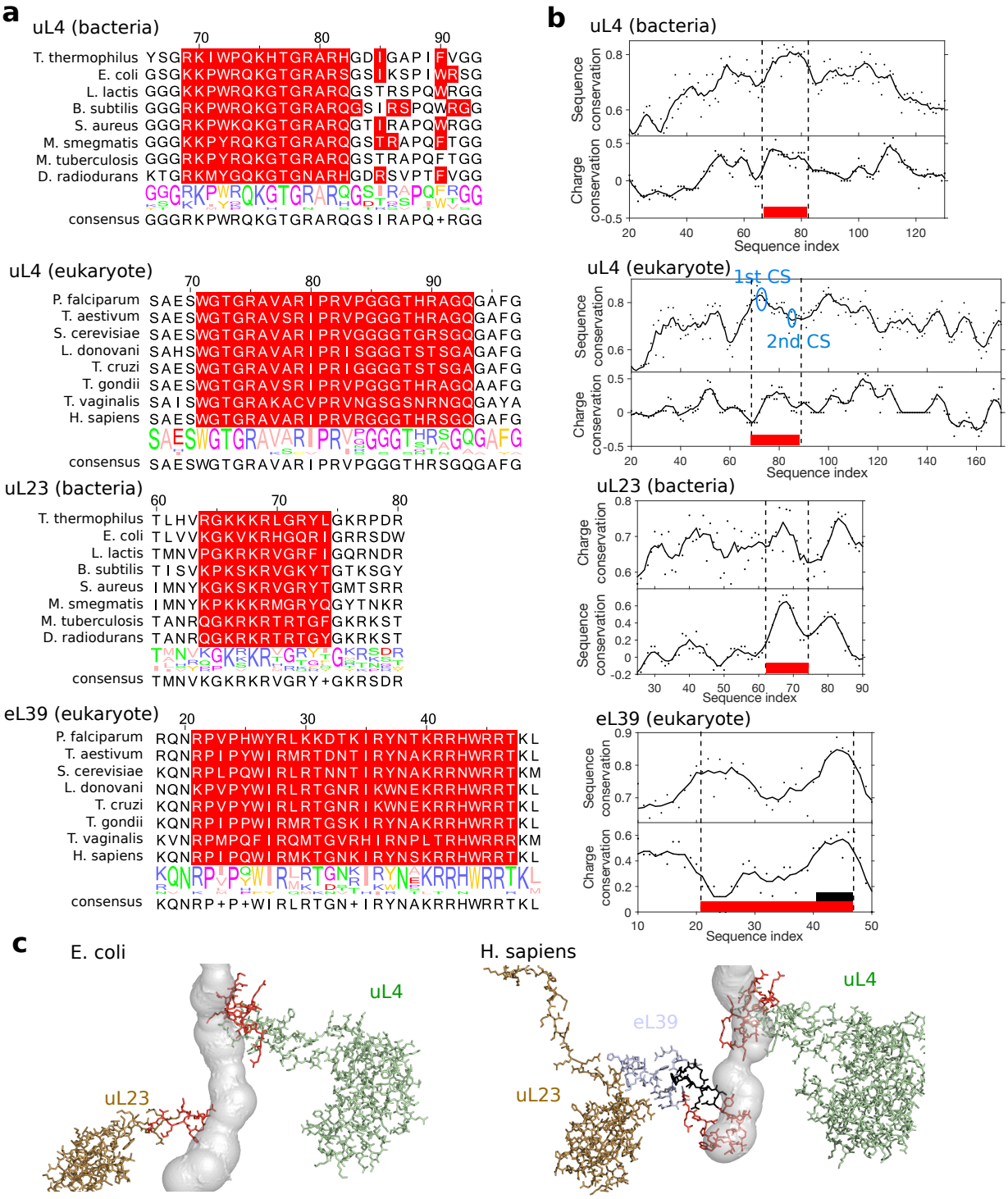

Figure S7. Conservation of sequence and charge of ribosomal proteins. a: Multiple sequence alignments of uL4 and uL23 from bacterial species of our dataset and uL4 and eL39 from eukarya, in the main region close to the tunnel. Residues located less than 10 Å from the tunnel are in red. **b:** The sequence and charge conservation scores; red regions indicate residues as in **a**. Circled positions correspond to the first and second constriction sites. A subregion of eL39 of high charge and sequence conservation is in black. Continuous lines show the signal averaged over 5 sites. **c:** Corresponding structure for *E. coli* (in left) and *H. sapiens* (in right). Regions in red and black correspond to the ones in **b**. Highly conserved region in eL39 matches with uL23 residues at the bacterial tunnel.

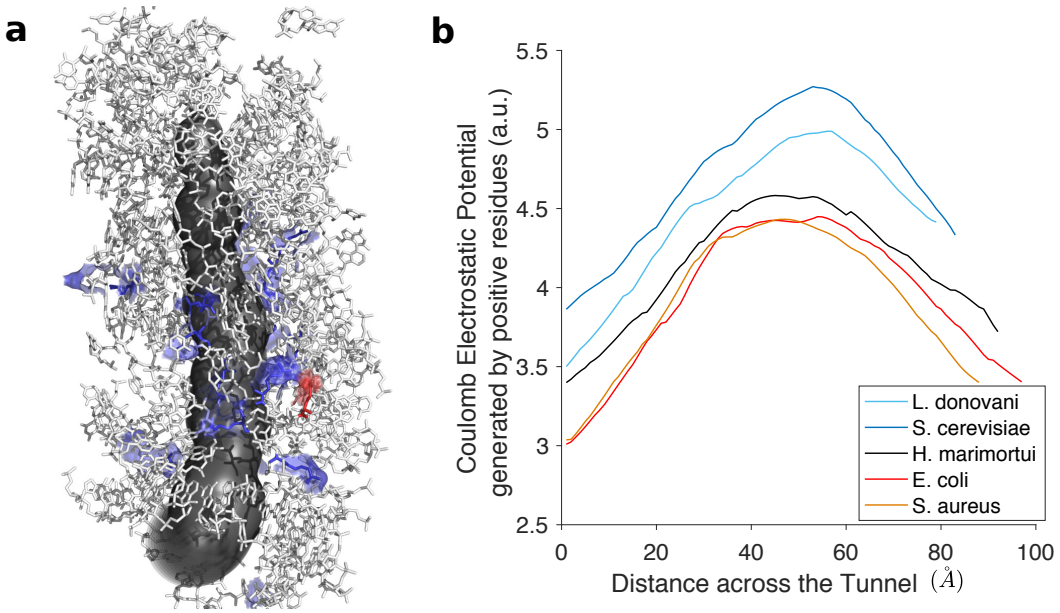

Figure S8. Localization and distribution of the charged residues of ribosomal proteins around the tunnel. a: Visualization of the charged residues (positive in blue, negative in red) in the neighborhood of the ribosome exit tunnel (structure PDB filename: 5NRG) shows a concentration of positive residues around the constriction site. **b:** For five structures representative of our dataset, we compute the electrostatic Coulomb potential generated by charged residues from ribosomal proteins, showing a peak located in the constriction site region.

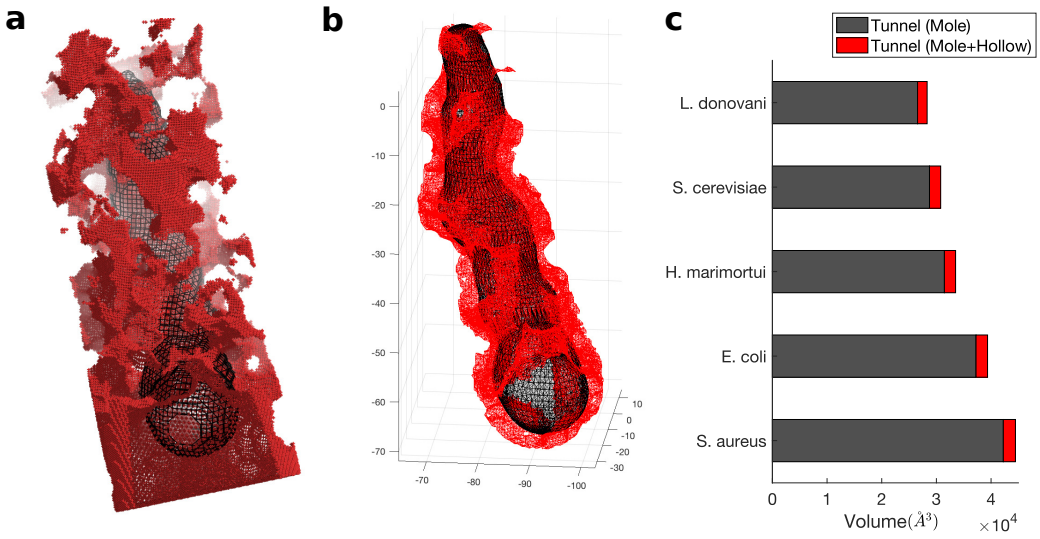

Figure S9. Tunnel volume comparison using alternative method: a: Visualization (see Supplementary methods) of the space (in red) around the exit tunnel computed using our main method (in black) that is not occupied by the ribosome structure (PDB filename of ribosome structure : 5NRG). **b:** The resulting extra surface of the tunnel was processed and represented in red. **c:** Horizontal bar plots of the original volume of the tunnel (grey bar) and additional volume (red) that come from the previous visualization, for five species in our dataset. We found a very high correlation between the original volume and the one obtained by adding the supplementary regions (Pearson's $r > 0.99$, p-value $< 10^{-5}$).

REFERENCES

1. Ho BK, Gruswitz F. HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. BMC structural biology. 2008;8(1):49.

2. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772–780.

3. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution. 2016;33(7):1870–1874.

4. Arenz S, Meydan S, Starosta AL, Berninghausen O, Beckmann R, Vázquez-Laslop N, et al. Drug sensing by the ribosome induces translational arrest via active site perturbation. Molecular cell. 2014;56(3):446–452.

5. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association. 1971;66(336):846–850.

6. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, et al. Structure of the E. coli ribosome–EF-Tu complex at< 3 Å resolution by C s-corrected cryo-EM. Nature. 2015;520(7548):567.

7. Gabdulkhakov A, Nikonov S, Garber M. Revisiting the Haloarcula marismortui 50S ribosomal subunit model. Acta Crystallographica Section D: Biological Crystallography. 2013;69(6):997–1004.

8. Natchiar SK, Myasnikov AG, Kratzat H, Hazemann I, Klaholz BP. Visualization of chemical modifications in the human 80S ribosome structure. Nature. 2017;551(7681):472–477.

9. Doris SM, Smith DR, Beamesderfer JN, Raphael BJ, Nathanson JA, Gerbi SA. Universal and domain-specific sequences in 23S–28S ribosomal RNA identified by computational phylogenetics. RNA. 2015;21(10):1719–1730.