Modeling the mosaic structure of bacterial genomes to infer their evolutionary history

Michael Sheinman, Peter F. Arndt, Florian Massip

January 26, 2024

Supplementary Information

A. Definitions. Here we summarize and define all the quantities used in the article.

μ_i^A, μ^A — effective mutation rate of a genomic segment of species A. The segment index i is omitted when appropriate.

μ_i^B, μ^B — effective mutation rate of a genomic segment of species B. The segment index i is omitted when appropriate.

$\mu_i = (\mu_i^A + \mu_i^B)/2$, $\mu = (\mu^A + \mu^B)/2$ — average effective mutation rate of a genomic segment i of species A and B. The segment index i is omitted when appropriate.

μ_c — average effective mutation rate of the most conserved genomic segment of species A and B.

μ_s — average effective mutation rate of the least conserved genomic segment of species A and B.

$p(.)$ — distribution of some variable.

$\widetilde{p}(.) = \mathscr{L}\{p\}(.)$ – Laplace transform of $p(.)$.

r — length of exact sequence match.

K_i, K — segment length (the segment index i is omitted when appropriate).

θ_i, θ — segment divergence (the segment index i is omitted when appropriate). Assuming long segments, throughout the article we do not distinguish between the expected divergence (mutation rate times the divergence time) and the realized one.

τ — taxa divergence time, twice the time to the most recent common ancestor of taxa A and B.

τ' — divergence time of the locus in the horizontal part, twice the time to the most recent common ancestor of homologous loci which transferred horizontally.

$m(r|\mu,\tau)$ — number of exact sequence matches of length r for an aligned segment with mutation rate μ and divergence τ.

$m(r|\tau)$ — number of exact sequence matches of length r for all aligned segments of two taxa with divergence τ.

ρ — HGT rate per bp per year for a given taxa pair.

L_0 — total length of all homologous sequences between two taxa.

δ — divergence threshold of the alignment software. Any sequence with a divergence larger than δ will not be detected by the alignment software.

$\mu_a = \min\left(\frac{\delta}{\tau}, \mu_s\right)$ — mutation rate of the least conserved detectable segment

L_v — length of the detected conserved part of the alignment, also referred to as the "vertical part".

L_h — length of the detected alignment due to HGT, also referred to as the "horizontal part".

$L = L_v + L_h$ — length of the alignment, sum of all alignable segments.

θ_v — average divergence along the vertical part of the alignment.

θ_h — average divergence along the horizontal part of the alignment.

$m_v(r)$ — number of exact matches of length r along the vertical part of the alignment.

$m_h(r)$ — number of exact matches of length r along the horizontal part of the alignment.

$m(r)$ — number of exact matches of length r along the alignment.

Throughout this article the time (τ) units are years, length (L) units are bp, while the rates (mutation μ and horizontal transfer ρ) are in units of yrs^{-1} bp^{-1}.

B. Derivation of Eq. 1 and impact of correlation between mutation rates along orthologous loci of two species. Here we discuss in detail the distribution of the mean mutation rate $p(\mu_i)$ of two lineages A and B and the derivation of Eq. 1. In addition, we discuss in more detail how correlation between mutation rates μ_i^A and μ_i^B may affect $p(\mu_i)$ and, therefore, match length distributions $m(r)$.

Consider, similarly to Section **Mutation rate distribution** in the main text, loci labeled with index i in two bacterial taxa A and B with $\tau/2$ time to their last common ancestor. Loci mutate with rate μ_i^A in taxon A and μ_i^B in taxon B. The divergence between the orthologous locus i of taxa A and B, $\theta_i = \mu_i\tau$ depends on the average mutation rate $\mu_i = (\mu_i^A + \mu_i^B)/2$. In the following we omit index i and investigate the distribution of μ for different correlation strength of μ^A and μ^B. We assume that the marginal distribution of μ^A and μ^B, $p^A(\mu^A)$ and $p^B(\mu^B)$, are the same. The joint distribution $p^{AB}(\mu^A,\mu^B)$ is unknown, but we use a single parameter to characterize it—correlation coefficient between μ^A and μ^B.

We use the following assumptions about the marginal distribution of mutation rate μ^A (and, therefore, μ^B). First, it is bounded between two values, μ_c and μ_s. Second, inside this range the mutation rate distribution does not change much, such that it can be approximated by a constant:

$$p^A(\mu^A) = \begin{cases} \frac{1}{\mu_s-\mu_c} & \mu_c \le \mu^A \le \mu_s \\ 0 & \text{otherwise} \end{cases} \text{ and } p^B(\mu^B) = \begin{cases} \frac{1}{\mu_s-\mu_c} & \mu_c \le \mu^B \le \mu_s \\ 0 & \text{otherwise} \end{cases}. \quad \text{[S.1]}$$

Below we also analyse what happens for another generic distribution, i.e. the exponential distribution and show that our conclusions are robust.

The MLD of genomic comparison is shaped by the distribution of the mean mutation rate (see Eq. 3)

$$\mu = (\mu^A + \mu^B)/2, \quad \text{[S.2]}$$

which depends on the joint distribution $p^{AB}(\mu^A,\mu^B)$ via (here δ is Dirac delta function)

$$p(\mu) = \int_0^\infty d\mu^A \int_0^\infty d\mu^B p^{AB}(\mu^A,\mu^B)\delta\left(\mu - \frac{\mu^A+\mu^B}{2}\right). \quad \text{[S.3]}$$

Clearly the bounds of μ are the same as those of μ^A and μ^B, namely μ_c and μ_s. However, within this interval the shape of $p(\mu)$ depends on the correlation between μ^A and μ^B. It is instructive to consider first two extreme cases: B.1 independent, uncorrelated μ^A and μ^B and B.2 perfectly correlated (equal) μ^A and μ^B.

B.1. Uncorrelated μ^A and μ^B. If the mutation rates are independent in the two taxa, and if $p(\mu^A) = p(\mu^B) \sim$ const between μ_c and μ_s, the joint distribution is given by $p^{AB}(\mu^A, \mu^B) = p^A(\mu^A)p^B(\mu^B) \sim$ const, so that within this interval

$$p(\mu) = \int_0^\infty d\mu^A \int_0^\infty d\mu^B p^{AB}(\mu^A, \mu^B)\delta\left(\mu - \frac{\mu^A + \mu^B}{2}\right) \sim \int_0^\infty d\mu^A I\left(\mu > \frac{\mu^A}{2}\right) \sim \mu. \qquad \text{[S.4]}$$

where I is the indicator function giving one if the condition is fulfilled and zero otherwise. In sum, the distribution of the average mutation rate μ is bounded between μ_c and μ_s and scales linearly within this interval, following (after proper normalization of the distribution density) Eq. **1**. This is also shown numerically in Fig. S12(a) and empirically in Figs. 1(b) and S13 (see Section **Empirical distribution of mutation rates**). Therefore, we can support Eq. **1** by empirical evidence, numeric and analytic consideration, based on the assumption of Eq. **S.1**.

The MLD can be derived from $p(\mu)$ using $m(r) = \int p(\mu)(\mu\tau)^2 e^{-\mu\tau r} d\mu$ (Eq. **3**). In this case one obtains a -4 power-law tail for the MLD which flattens for small values of $r \ll \frac{1}{\mu_s \tau}$ and exponentially drops for $r \gg \frac{1}{\mu_c \tau}$ (see Fig. S12(b))

$$m(r) \propto \frac{e^{-\mu_c \tau r}\{6 + r\mu_c\tau[6 + r\mu_c\tau(3 + r\mu_c\tau)]\} - e^{-\mu_s \tau r}\{6 + r\mu_s\tau[6 + r\mu_s\tau(3 + r\mu_s\tau)]\}}{\tau^2 \mu_s^2 r^4}. \qquad \text{[S.5]}$$

Here, for simplicity we ignored finite sensitivity of the alignment software, HGT events, and assumed $\mu_s \gg \mu_c$ to show the qualitative trends. For the general case see Eq. **4** or Eq. **S.9** for the vertical part and Eq. **10** for the full MLD, including the horizontal part.

We observe these predicted trends for most bacterial pairs (see Fig. 2 and supplementary file `SI-ExtendedFigures.pdf`), suggesting that the correlation between mutation rates of orthologous loci is negligible for relatively distant taxa pairs. Pairs of eukaryotes exhibit similar behavior, as shown in Fig. S11 and the simulations (see Fig. S4).

Interestingly, we obtain very similar conclusions if μ^A and μ^B are exponentially distributed and larger than μ_c (Fig. S12(c-d))

$$p^A(\mu^A) = \frac{1}{\mu_s} e^{-(\mu^A - \mu_c)/\mu_s} I(\mu > \mu_c) \text{ and } p^B(\mu^B) = \frac{1}{\mu_s} e^{-(\mu^B - \mu_c)/\mu_s} I(\mu > \mu_c). \qquad \text{[S.6]}$$

We turn now to study how correlation between μ^A and μ^B can affect our results. We start by studying the extreme case where μ^A and μ^B are perfectly correlated.

B.2. Perfectly correlated (equal) μ^A and μ^B. If μ^A and μ^B are perfectly correlated, $\mu^B = \mu^A$, their average μ is distributed as μ^A (or, equivalently μ^B):

$$p(\mu) = \int_0^\infty p^A\left(\mu^A\right)\delta\left(\mu - \frac{\mu^A + \mu^B}{2}\right) d\mu^A = p^A(\mu) = p^B(\mu). \qquad \text{[S.7]}$$

In this case, if $p^A(\mu) = p^B(\mu) \sim$ const, then $p(\mu) \sim$ const, as shown in Fig. S12(a). As before, we can calculate the MLD of the two taxa using $m(r) = \int p(\mu)(\mu\tau)^2 e^{-\mu\tau r} d\mu$ (Eq. **3**) and the obtained MLD is distributed as $m(r) \sim r^{-3}$, flattens for small values of $r \ll \frac{1}{\mu_s \tau}$ and exponentially drops for $r \gg \frac{1}{\mu_c \tau}$ (see Fig. S12(b))

$$m(r) \propto \frac{e^{-\mu_c \tau r}[2 + \mu_c r\tau(2 + \mu_c r\tau)] - e^{-\mu_s \tau r}[2 + \mu_s r\tau(2 + \mu_s r\tau)]}{\tau \mu_s r^3}. \qquad \text{[S.8]}$$

We observe such -3 power-law regime for close taxa pairs (see Fig. S7(a)), suggesting that, when analysing closely related taxa, the correlation between mutation rates of orthologous loci should be taken into account. We obtain very similar results if μ^A and μ^B follow either a uniform (Fig. S12(a-b)) or a truncated exponential (Fig. S12(c-d)) distribution.

B.3. Correlated μ^A and μ^B. If the correlation coefficient between $p^A(\mu^A)$ and $p^B(\mu^B)$ is not 0 nor 1, the power-law regimes of $p(\mu)$ and $m(r)$ are in between of the described extreme cases (see Fig. S12). In this case, the correlation between the mutation rates of orthologous loci may affect their MLDs and account for some discrepancies between our theory (which assumes no correlation) and empirical results.

C. Empirical distribution of mutation rates. In the main text of the article we demonstrate that distribution of exact sequence matches is related to the Laplace transform of the mutation rate distribution. Here we analyze the mutation rate more directly, segmenting the pairwise genome-wide mosaic alignments to segments of constant mutations rate (see Fig. S2 for an illustrative example) using our `segmut` `R` package (1).

After the segmentation, we collect the genetic distances calculated for each segment i—θ_i. Along the vertical part of the alignment, the mutation rate is calculated using $\mu_i = \theta_i/\tau$, where τ is the evolutionary time distance between the two taxa, calculated using our method (fitting the MLD). This relation between θ_i and μ_i is valid only for the vertical part of the alignment, because in the horizontal part the divergence time is not fixed to τ. We exemplify the resulting distribution of the mutation rate along the genomes in the article in Fig. 1(b). Here we provide a few more examples and discuss the results in more detail.

In Fig. S13 a comparison between our main assumption Eq. **1** and the derived distribution of the mutation rate from the empirical data is shown for 6 pairs of taxa. For taxa with small divergences (Fig. S13(a-d)), the sensitivity of the aligner does not play a role and the empirical distribution of μ_i agrees well with Eq. **1**. For the pair *Enterobacter asburiae vs. Enterobacter hormaechei* (marked by digit 2 in Fig. 3(a-c)), we obtained unrealistic divergence time estimate $\tau = 1.03 \cdot 10^8$ due to very high level of HGT between the two species (see Fig. 3(b) and Fig. S7(b)). To get a good fit in Fig. S13(c), we manually adjusted τ. The massive uptick of the θ_i distribution for small θ_i in Fig. S13(c) is another signature of the very large HGT rate between these two pairs.

For taxa with higher divergences, segments with high effective mutation rates are not detected by the alignment. This effect can be seen in Fig. S13(e-f): the distribution of θ_i reaches its maximum at the alignment sensitivity threshold δ, so that the distribution peak of the empirical μ_i distribution is smaller than the theoretical expectation μ_s.

C.1. Segments with very small θ—putative HGT events. In the main text, we make the hypothesis that segments with very low divergence θ_i result from recent horizontal gene transfers. Here we conduct additional analyses to demonstrate this point.

In Ref. (2) we showed that long identical matches between distant bacterial taxa are enriched with integrative, conjugative elements and resistance genes. This suggests that such long matches are present due to relatively recent HGT events. Here we investigate biological content of segments (obtained using `segmut` `R` package (1)) with very low divergence, as represented in the upticks of distributions of μ below μ_c in Figs. 1(b) and S13) and show that such segments should be indeed interpreted as recent HGT and not extremely conserved sequences. To this end we aligned segments obtained from alignments of *E. coli* and *S. enterica* against the database of resistance genes, transposable elements and integrons using BacAnt software (3). As shown in Fig. S14(a), segments with very low divergences $\theta < 10^{-3}$ are enriched with resistance genes, transposable elements and integrons relative to the rest of segments with $\theta \geq 10^{-3}$. This functional enrichment supports our interpretation of segments with $\theta \ll \mu_c \tau$ as horizontally transferred ones.

Another useful measure of a genomic sequence in the context of HGT inference is its GC content and how it corresponds to the GC content of the whole genome (4). In Fig. S14(b) one can see that GC content of the segments from the from alignments of *E. coli* and *S. enterica* with very low divergence $\theta < 10^{-3}$ has significantly different average GC content and its distribution exhibits peaks at values very different from the mean GC content of *E. coli* and *S. enterica*, supporting the evidence for their origin from HGTs.

D. Consistency of genome segmentation and MLD. To demonstrate the validity of our approach on empirical data, we computed the alignment of one strain of *E. coli vs.* one strain of *S. enterica*. Ignoring the mosaic structure of genomes, assuming that the mutation rates is constant along the two genomes (such that the genome-wide density of mutations θ is uniform along the genome), the MLD would simply follow $m(r) = L\theta^2 e^{-\theta r}$, which is very different from the empirical observation. To make sure that this discrepancy is due to the existence of loci with different effective mutation rates along the genome, we used the segmut R package (see Materials and Methods) to reconstruct the mosaic structure of the genomes. This way we can compute the empirical distribution of genomic divergences $p(\theta_i)$. Assuming that the mutations density is constant along each locus, but differs from one locus to another, the MLD of locus i is given by $m(r|\theta_i) \simeq \theta_i^2 e^{-\theta_i r}$. Using this we can compute a genome-wide pseudotheoretical MLD using $m(r) = \sum_i p(\theta_i) m(r|\theta_i)$, which very closely mimics the empirical distribution, demonstrating the validity of our model (see Fig. S3(a)). The disagreement between the naive model — which assume a single genome-wide mutation rate, and therefore a homogeneous density of mutations along the whole genome — and our mosaic model is even more evident when one considers the comparisons of many pairs of strains as shown for the analysis of all *vs.* all alignments of *E. coli vs. S. enterica* (see Fig. S3(b)).

E. Analytical results. Here we derive the results from Section **Analytical solution** in more detail.

E.1. Vertical part. The MLD from the δ-detectable vertical part of the genome (homologous loci with $\mu < \mu_a$, such that their divergences are smaller than δ) with time divergence τ, using Eq. **1**, is given by

$$m_v(r) = \int_{\mu_c}^{\mu_a} m(r|\mu,\tau)p(\mu)d\mu = L_0 \int_{\mu_c}^{\mu_a} (\mu\tau)^2 e^{-\mu\tau r} p(\mu) d\mu = L_0 \frac{\partial^2}{\partial r^2} \begin{cases} \frac{2}{r^2\tau^2} \frac{e^{-\mu_c \tau r}(\mu_c r\tau+1) - e^{-\mu_s \tau r}(\mu_s r \tau + 1)}{\mu_s^2 - \mu_c^2} & \tau \leq \frac{\delta}{\mu_s} \\ \frac{2}{r^2\tau^2} \frac{e^{-\mu_c \tau r}(\mu_c r\tau+1) - e^{-\delta r}(\delta r + 1)}{\mu_s^2 - \mu_c^2} & \tau > \frac{\delta}{\mu_s} \end{cases}. \qquad \text{[S.9]}$$

One can see that the tail of the MLD from the vertical part scales as r^{-4}, as previously observed in eukaryotes (5, 6). In our previous studies (Refs. (5, 6), we focused on the value of the power-law $m(r) \sim r^\alpha$ exponent $\alpha = -4$ and ignored the (exponential) cutoff of the power-law tail of $m(r)$. Then, to analyse this power-law asymptotic behaviour we analysed the behaviour of $p(\mu)$ close to zero. In that work one can see in Fig. 1C that the power-law drops above a certain threshold. In the current work we investigated this drop more systematically, taking other pairs of mammals (see Fig. S11) and bacteria pairs. We show that this drop of $m(r)$ can be explained by a lower cutoff of $p(\mu)$ at $\mu_c > 0$. In sum, the $m \sim r^{-4}$ power-law behaviour can be understood using the similar arguments to the ones in Refs. (5, 6), but to explain the exponential cutoff of the power-law we had to introduce a lower cutoff of $p(\mu)$ at $\mu_c > 0$.

The total length of the δ-detectable homologous vertical part of the genome decreases with increasing time divergence τ beyond δ/μ_s and is given by

$$L_v = L_0 \int_{\mu_c}^{\mu_a} p(\mu)d\mu = L_0 \frac{\mu_a^2 - \mu_c^2}{\mu_s^2 - \mu_c^2} = L_0 \begin{cases} 1 & \tau \leq \frac{\delta}{\mu_s} \\ \frac{(\frac{\delta}{\tau})^2 - \mu_c^2}{\mu_s^2 - \mu_c^2} & \tau > \frac{\delta}{\mu_s} \end{cases} \qquad \text{[S.10]}$$

Along the vertical region with this length the average divergence is given by

$$\theta_v = \frac{L_0}{L_v} \int_{\mu_c}^{\mu_a} \mu\tau p(\mu)d\mu = \frac{2}{3} \frac{\mu_a^3 - \mu_c^3}{\mu_a^2 - \mu_c^2} \tau = \frac{2}{3}\tau \begin{cases} \frac{\mu_s^3 - \mu_c^3}{\mu_s^2 - \mu_c^2} & \tau \leq \frac{\delta}{\mu_s} \\ \frac{(\frac{\delta}{\tau})^3 - \mu_c^3}{(\frac{\delta}{\tau})^2 - \mu_c^2} & \tau > \frac{\delta}{\mu_s} \end{cases} \qquad \text{[S.11]}$$

See SI Section E.3 for representative plots of these functions.

E.2. Horizontally transferred part. We assume that HGT is a random Poisson process with rate ρ. Then the distribution of the time divergences τ' in the horizontal part is exponential $\propto e^{-\frac{\rho\tau'}{2}}$ but has an upper bound τ (the taxa time divergence). Assuming that only a small fraction of the genome has been transferred (*i.e.* $\frac{\rho\tau}{2} \ll 1$), the MLD from the δ-detectable horizontal part of the genome can be written as:

$$m_h(r) = \int_0^\tau m_v(r|\tau') \frac{\rho}{2} e^{-\frac{\rho}{2}\tau'} d\tau' \simeq \int_0^\tau m_v(r|\tau') \frac{\rho}{2} d\tau' = \frac{\rho L_0}{\mu_c + \mu_s} \frac{\partial^2}{\partial r^2} \begin{cases} \frac{1}{r} - \frac{e^{-\mu_c\tau r} - e^{-\mu_s \tau r}}{\tau r^2 (\mu_s - \mu_c)} & \tau \leq \frac{\delta}{\mu_s} \\ \frac{1}{r} - \frac{e^{-\mu_c\tau r} - e^{-\delta r}(\delta r - \mu_s r\tau + 1)}{\tau r^2 (\mu_s - \mu_c)} & \tau > \frac{\delta}{\mu_s} \end{cases}. \qquad \text{[S.12]}$$

One can see that the tail of the MLD from the horizontally transferred part scales as r^{-3}, as was also derived and shown empirically in Ref. (2).

In the same regime, the total length of the δ-detectable homologous part of the genome due to HGT is given by

$$L_h = \int_0^\tau L_v(\tau') \frac{\rho}{2} e^{-\frac{\rho}{2}\tau'} d\tau' \simeq \int_0^\tau L_v(\tau') \frac{\rho}{2} d\tau' = L_0 \begin{cases} \frac{\rho}{2}\tau & \tau \leq \frac{\delta}{\mu_s} \\ \frac{\delta\rho}{2\mu_s} + \frac{\rho}{2\mu_s} \frac{\tau\mu_c^2(\frac{\delta}{\tau} - \mu_s) - \delta\mu_s(\frac{\delta}{\tau} - \mu_s)}{\mu_s^2 - \mu_c^2} & \tau > \frac{\delta}{\mu_s} \end{cases}. \qquad \text{[S.13]}$$

Along the horizontally transferred region with this length the average divergence is given by

$$\theta_h = \frac{\int_0^\tau \theta_v(\tau') L_v(\tau') \frac{\rho}{2} e^{-\frac{\rho}{2}\tau'} d\tau'}{L_h} \simeq \frac{\int_0^\tau \theta_v(\tau') L_v(\tau') \frac{\rho}{2} d\tau'}{L_h} = \begin{cases} \frac{2}{3} \frac{\mu_s^3 - \mu_c^3}{\mu_s^2 - \mu_c^2} \frac{\tau}{2} & \tau \leq \frac{\delta}{\mu_s} \\ \frac{\delta}{2} \frac{2}{3} \frac{2\delta^2 + \frac{1}{\delta}(\mu_c\tau)^3 - 3\delta\mu_s\tau}{\delta^2 + (\mu_c\tau)^2 - 2\delta\mu_s\tau} & \tau > \frac{\delta}{\mu_s} \end{cases} \qquad \text{[S.14]}$$

E.3. Summary and illustration of analytical results. Shortly after speciation, when the divergence is still low, two bacterial genomes can be well aligned and the total alignment length of the vertical transmitted part (L_v) of their genomes spans their entire genome (L_0), see Fig. S15(a). As evolution progresses mutations will be accumulated and the divergence in the vertical transmitted part θ_v will increase first linearly as $\theta_v \propto \tau$, see Fig. S15(b). However at evolutionary distance $\tau = \delta/\mu_s$ faster evolving loci will accumulated so many mutations such that no alignment algorithm can align these regions. At this point the total alignment length L_v will shrink until only very well conserved loci can be aligned at evolutionary distance $\tau = \delta/\mu_c$. During this time the average divergence of the still alignable segments will stay just below δ. Beyond this point even short segments of the two bacteria can be aligned unless there has been an event of HGT. The corresponding MLDs for different time points are presented in Fig. S15(c).

F. Fitting MLDs.

F.1. Fitting Objective. To fit our theoretical MLD to the empirical results obtained from genomic comparisons, we first bin the data as described in the method section. We then minimize the mean square relative differences between the observed and estimated data points. In principle we could also maximize the Poisson likelihood of the observed counts. However, observed data often contains biological noise due to codon structure or other constraints on short length scales that we do not account for in our modeling. We found that in the presence of such noise on short length scales, minimizing the mean square relative differences results in better overall results. To demonstrate this behavior we generated an exemplary dataset of counts according to our model and a slightly distorted dataset where we moved half of the counts for lengths $r = 1$ and 2 to the count for $r = 3$ mimicking a potential influence of codon structure on the count data, see Fig. S16. One can see that minimizing mean square relative differences one obtains better estimates of ρ, compared to the maximum likelihood approach.

F.2. Robustness of time divergence inference to $p(\mu)$ shape. In Section **Numerical validation** in the main text we describe how we validated our inference procedure using simulations of bacterial evolution. Here we analyze specifically the robustness of the inference of τ to the distribution of μ. For simplicity, we ignore HGT events (we take $\rho = 0$) and finite sensitivity of the alignment software (we take $\delta = 1$). In this case the MLD simplifies to Eq. **S.5**. We assume no correlation between mutation rates in taxa A and B and draw both μ^A and μ^B from the same distribution. We study two cases: uniform distribution (Eq. **S.1**) and truncated exponential (Eq. **S.6**). We calculate average mutation rates using $\mu = (\mu^A + \mu^B)/2$ and MLD using $m(r) = \int p(\mu)(\mu\tau)^2 e^{-\mu\tau r} d\mu$ (Eq. **3**) for different values of τ. Then, we fitted the simulated MLD with Eq. **S.5** (minimizing mean squared relative error) with a single free parameter τ (the normalization is taken to ensure the same total length of all matches, $\sum_r m(r)r$). We obtain good fits of the MLDs (Fig. S17) and accurate estimations of the value of τ (inset of Fig. S17). We conclude that our inference procedure is robust for different generic distributions of μ^A and μ^B as long as they are not correlated and the distributions follow assumptions described in Section B: the distributions are negligible outside of the range between two values, μ_c and μ_s and inside this range the distributions do not change much, they can be well approximated by a constant.

F.3. Inferring the parameters from a single pair of genomes. Our results in the main text are mostly based on MLD obtained from many pairwise comparisons of genomes for each taxa pair. A natural question is how the MLD from an individual comparison looks like and how many comparisons one needs to compute to get a good estimate of the parameters. As shown in Figs. S18(a,b), genome-wide properties—mean divergence and total alignment length—do not change dramatically from one pair of *E. coli* and *S. enterica* to another. For other pairs of taxa we refer to Figs. S9,S10. There one can see that, for distant pairs of taxa, different genome comparisons tend to vary because the total alignment length is small and absence/presence of a single gene in the genome can significantly affect the genome-wide alignment.

MLDs of genome pairs of *E. coli* and *S. enterica* are very similar to each other, except genome pairs with rare events of recent HGT which is reflected as very long matches, longer than roughly 10^3, as demonstrated in Fig. S18(c). As shown in Figs. S18(d), estimates of τ using individual pairs and assuming no HGT, $\rho = 0$ results in slightly biased estimate relative to the one obtained from all pairs and fitting ρ as a free parameter (see Fig. 2(e)). For a few pairs with long matches, however, the τ estimate was significantly biased to smaller values. Setting $\rho = 10^{-10}$— value obtained from all the pairs—one gets much smaller overall bias and much less pairs with outlier estimates, as expected (see Figs. S18(d)). In particular, these results suggest that our estimates are not biased significantly due to biased sampling of certain strains in the database. Moreover, to mitigate such sampling biases we removed samples obtained from large multi-isolate projects. In sum, comparing a single pair of genomes one can get a reasonable estimate of τ for most pairs (without recent HGT events). The estimate may be biased to smaller values setting $\rho = 0$ in the fit.

The estimate for ρ cannot be obtained from a single pair of genomes for most cases and the smaller ρ is, the larger number of pairs is needed. One can estimate this number as following. The exponential drop in the vertical part of the MLD occurs at $r \simeq \frac{1}{\mu_c\tau}$ (see Eq. **4** or Eq. **S.5**). The number of matches due to HGT with length longer than this is given by (using Eq. **7**)

$$\int_{\frac{1}{\mu_c\tau}}^{\infty} m_h(r)dr \simeq \frac{\rho L_0(\mu_c\tau)^2}{\mu_s}. \qquad \text{[S.15]}$$

Therefore the number of (independent) pairs of genomes needed to estimate ρ for two taxa with τ divergence has to be much larger than $\frac{\mu_s}{\rho L_0(\mu_c\tau)^2}$. For our taxa pairs we have τ in the $10^7 - 10^9$ range and ρ in the $10^{-13} - 10^{-8}$ range (see Fig. 3). Thus, we need more than 10^4 pairs to estimate the parameters in these ranges. For most taxa pairs this condition is satisfied. In fact, because τ and ρ are negatively correlated, roughly following $\rho \sim 10^6/\tau^2$ (see Fig. 3(c)), more realistic condition is that for the taxa pairs, used in our study, the number of pairs has to be much larger than 1. We made sure that each taxon is represented by at least 20 genomes (see Table 1), such that the number of genome pairs is larger than $20^2 = 400$ for each pair of taxa. This way we managed to detect the $m \sim r^{-3}$ tail from the horizontal part and, therefore, estimate the value of ρ for all taxa pairs (see Supplementary file `SI-ExtendedFigures.pdf`).

1. M. Sheinman, P. F. Arndt, and F. Massip, “segmut,” https://github.com/mishashe/segmut (2023).
2. M. Sheinman, K. Arkhipova, P. F. Arndt, B. E. Dutilh, R. Hermsen, and F. Massip, Elife **10**, e62719 (2021).
3. X. Hua, Q. Liang, M. Deng, J. He, M. Wang, W. Hong, J. Wu, B. Lu, S. Leptihn, Y. Yu, *et al.*, Frontiers in microbiology **12**, 649969 (2021).
4. J. G. Lawrence and H. Ochman, TRENDS in Microbiology **10**, 1 (2002).
5. F. Massip, M. Sheinman, S. Schbath, and P. F. Arndt, Molecular biology and evolution **32**, 524 (2015).
6. F. Massip, M. Sheinman, S. Schbath, and P. F. Arndt, Genetics **204**, 475 (2016).
7. G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin, PLoS computational biology **14**, e1005944 (2018).
8. T. Williams, C. Kelley, C. Bersch, H.-B. Bröker, J. Campbell, R. Cunningham, D. Denholm, G. Elber, R. Fearick, C. Grammes, *et al.*, An interactive plotting program. Available online: http://www.gnuplot. info/docs_5 **2** (2017).
9. L. R. Nassar, G. P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, B. T. Lee, *et al.*, Nucleic Acids Research **51**, D1188 (2023).
10. S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller, Genome research **13**, 103 (2003).

11. R. S. Harris, *Improved pairwise alignment of genomic DNA* (The Pennsylvania State University, 2007).
12. S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprowicz, M. Sanderford, M. Li, G. Stecher, and S. B. Hedges, Molecular Biology and Evolution **39**, msac174 (2022).
13. K. Goldfeld and J. Wujciak-Jens, Journal of Open Source Software **5**, 2763 (2020).

Supplementary figures

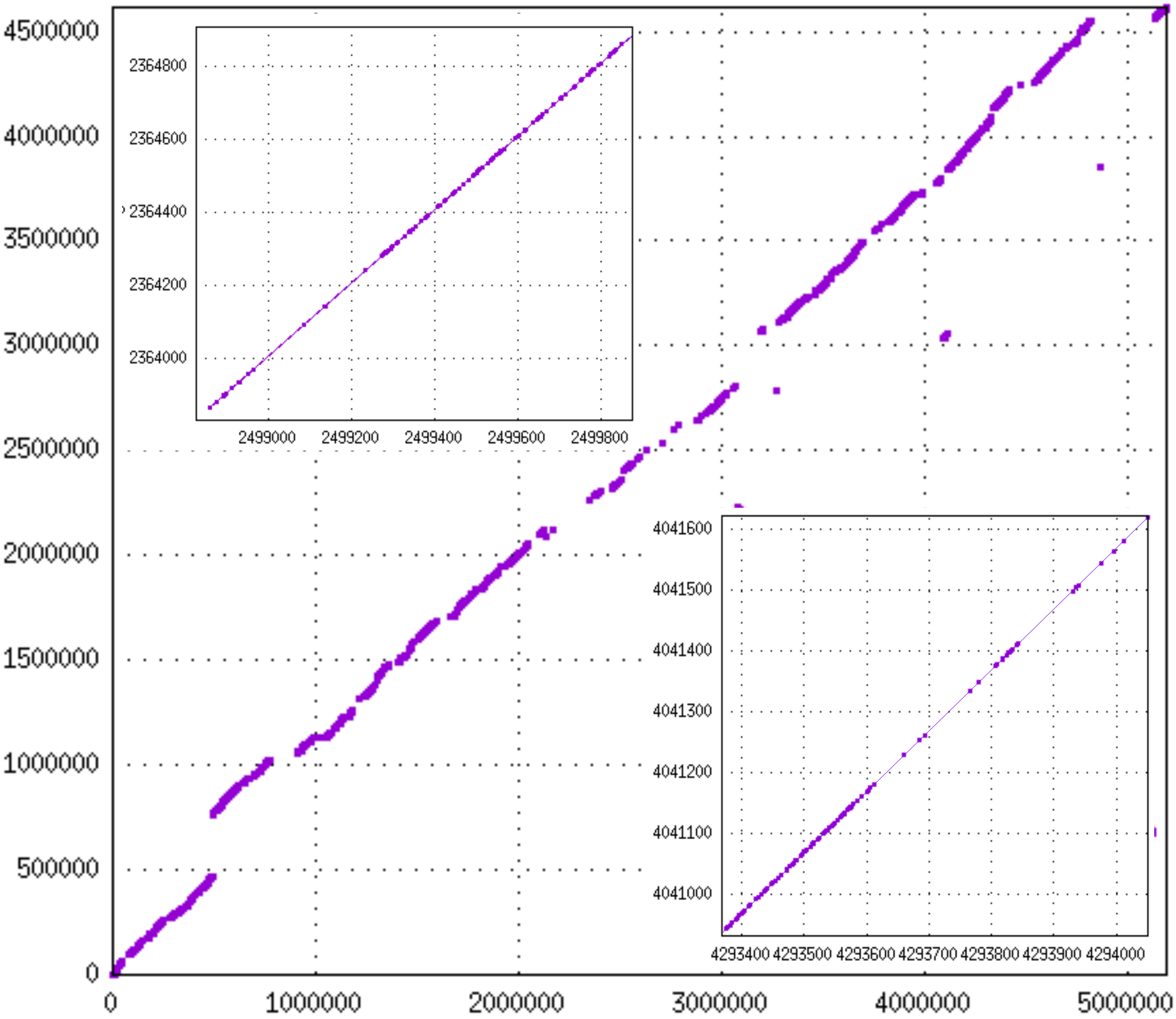

Fig. S1. Dotplot of *E. coli* (NZ_CP068796.1 strain RIVM_C012087) and *S. enterica* (NZ_CP019035.1 str. 9184 isolate ATCC 9184). The alignment was done using `nucmer` (7) with the default settings and `--mum` option and visualized using `gnuplot` (8). Markers denote differences between the two homologous genomic loci.

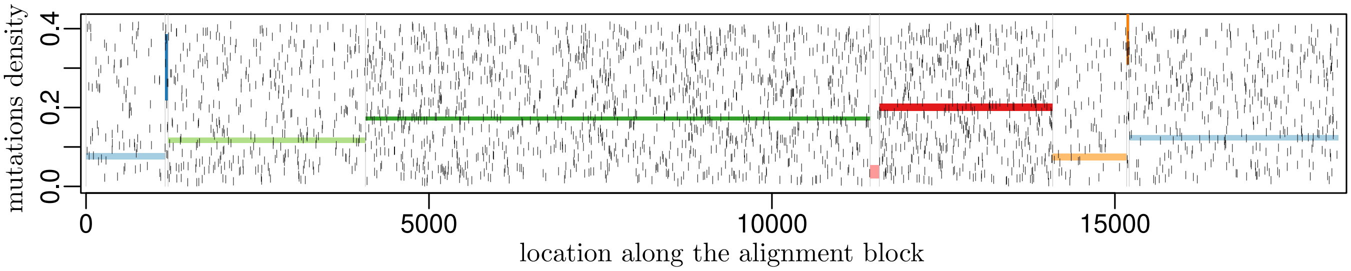

Fig. S2. Mutations along the largest alignment block for *E. coli* strain ATCC 8739 (NZ_CP033020.1) and *S. enterica* strain SE20-72C-2 (NZ_AP026948.1). Each short vertical black line represents a mutation (vertical position is random). Colored rectangles represent the 68% confidence intervals of the mutation density along the segments ($\pm$ one standard deviation, calculated assuming Poisson distribution), obtained using the `segmut R` package (1).

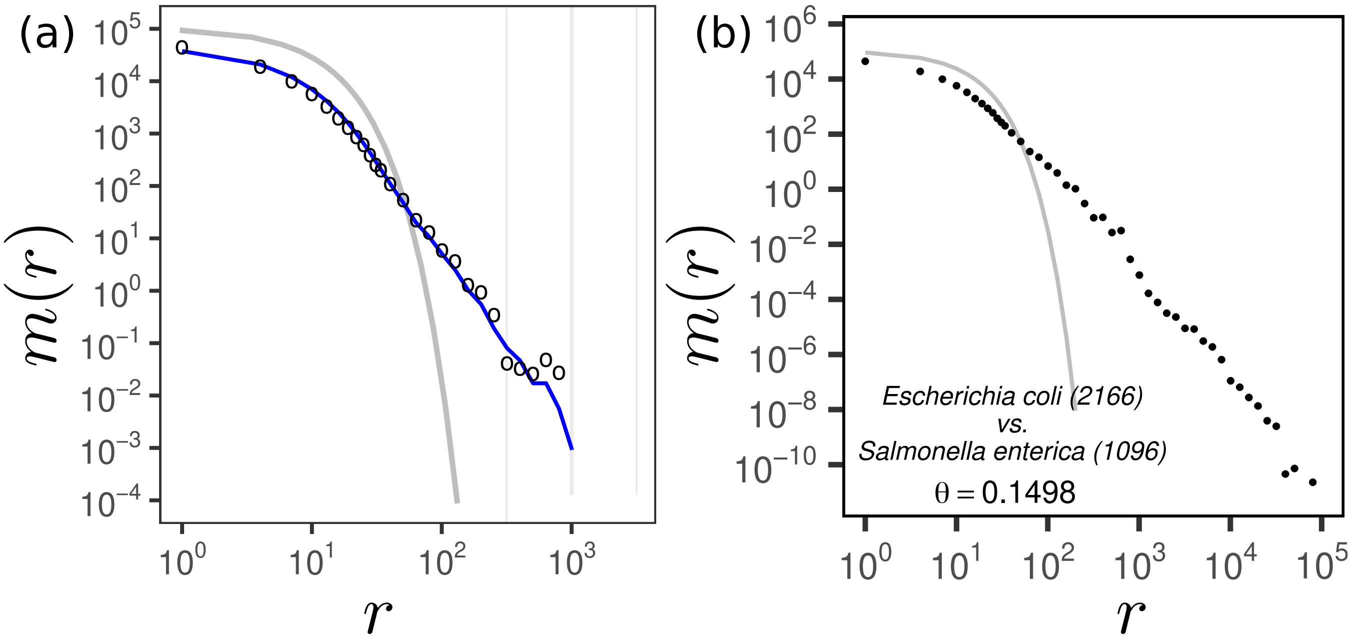

Fig. S3. (a) Empirical *vs.* pseudotheoretical MLD. Circles represent the empirical MLD from an alignment of *E. coli* strain ATCC 8739 (NZ_CP033020.1) and *S. enterica* strain SE20-72C-2 (NZ_AP026948.1). The line is based on the segmentation of the alignment using the `segmut R` package. For each segment i the divergence θ_i and length K_i are calculated and then the pseudotheoretical MLD is calculated as $\sum_i K_i \theta_i^2 e^{-\theta_i r}$. (b) The MLD calculated from all *vs.* all $2,166 \times 1,096$ alignments of *E. coli vs. S. entrerica* strains. The grey solid line in both panels represents theoretical MLD, ignoring the mosaic structure of the genome and assuming that the average genome-wide density of mutations θ is uniform along the genome, $L\theta^2 e^{-\theta r}$ (see Eq. 2).

Taxon	Number
Escherichia coli	2,166
Klebsiella pneumoniae	1,226
Salmonella enterica	1,096
Vibrio	619
Enterobacter hormaechei	279
Citrobacter	192
Serratia	140
Escherichia fergusonii	50
Enterobacter roggenkampii	39
Raoultella	32
Enterobacter asburiae	25
Escherichia albertii	22
Cronobacter	22

Table 1. Number of fully assembled chromosomal genomes used for each taxon.

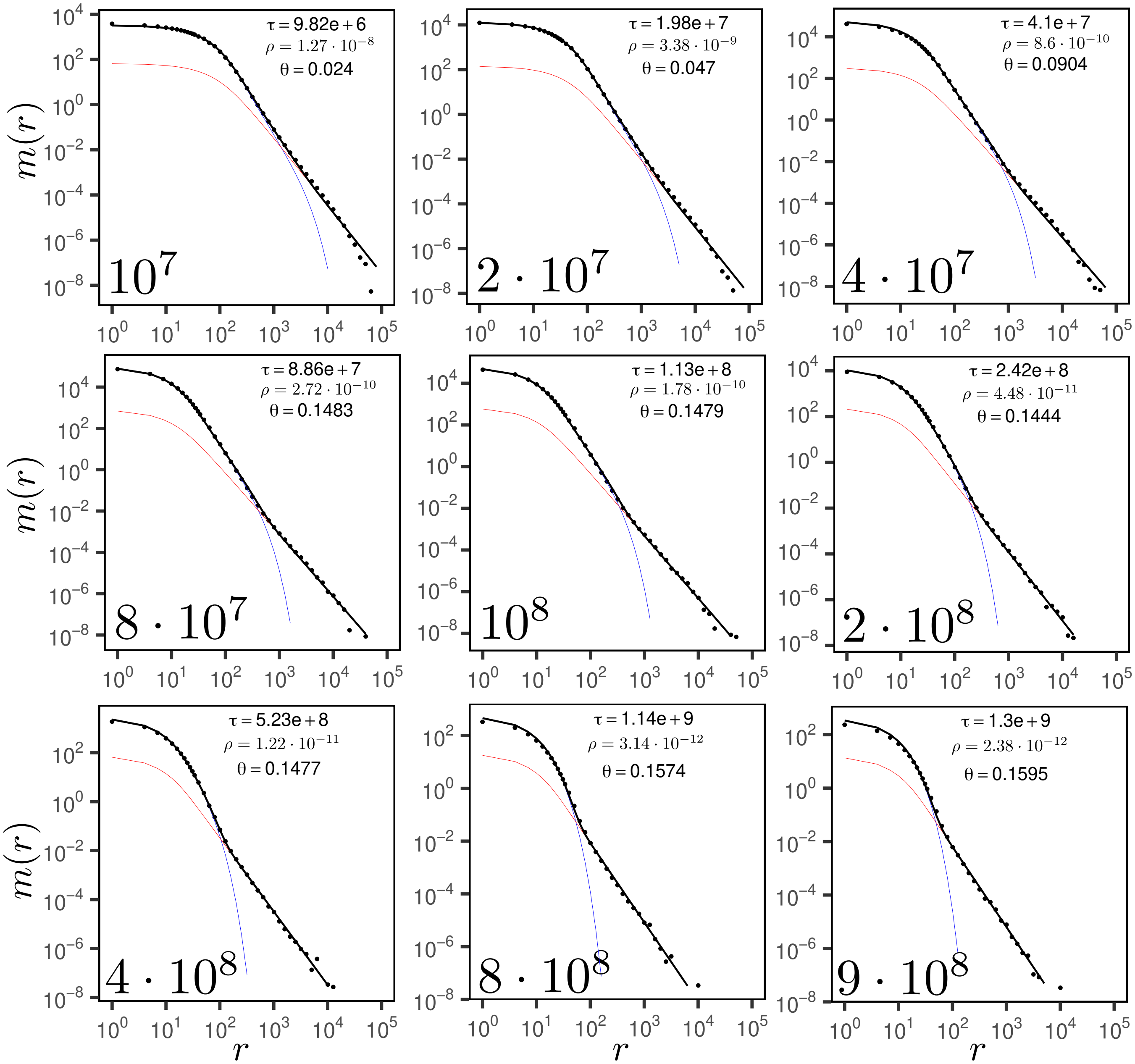

Fig. S4. Match length distributions of simulated pairs of genomes with different divergence times, indicated in the bottom-left corner of the panels. The HGT rate was assumed to be constant in time and taken as $\rho = 10^6/\tau^2$ from the beginning of the simulation. The average divergence is indicated by θ. The simulated data (dots) are fitted using the global parameters $\mu_c = 10^{-10}$ and $\mu_s = 3.64 \cdot 10^{-9}$ and $\delta = 0.25$. The values of τ and ρ are fitted for each pair separately and are shown in the top-right corner.

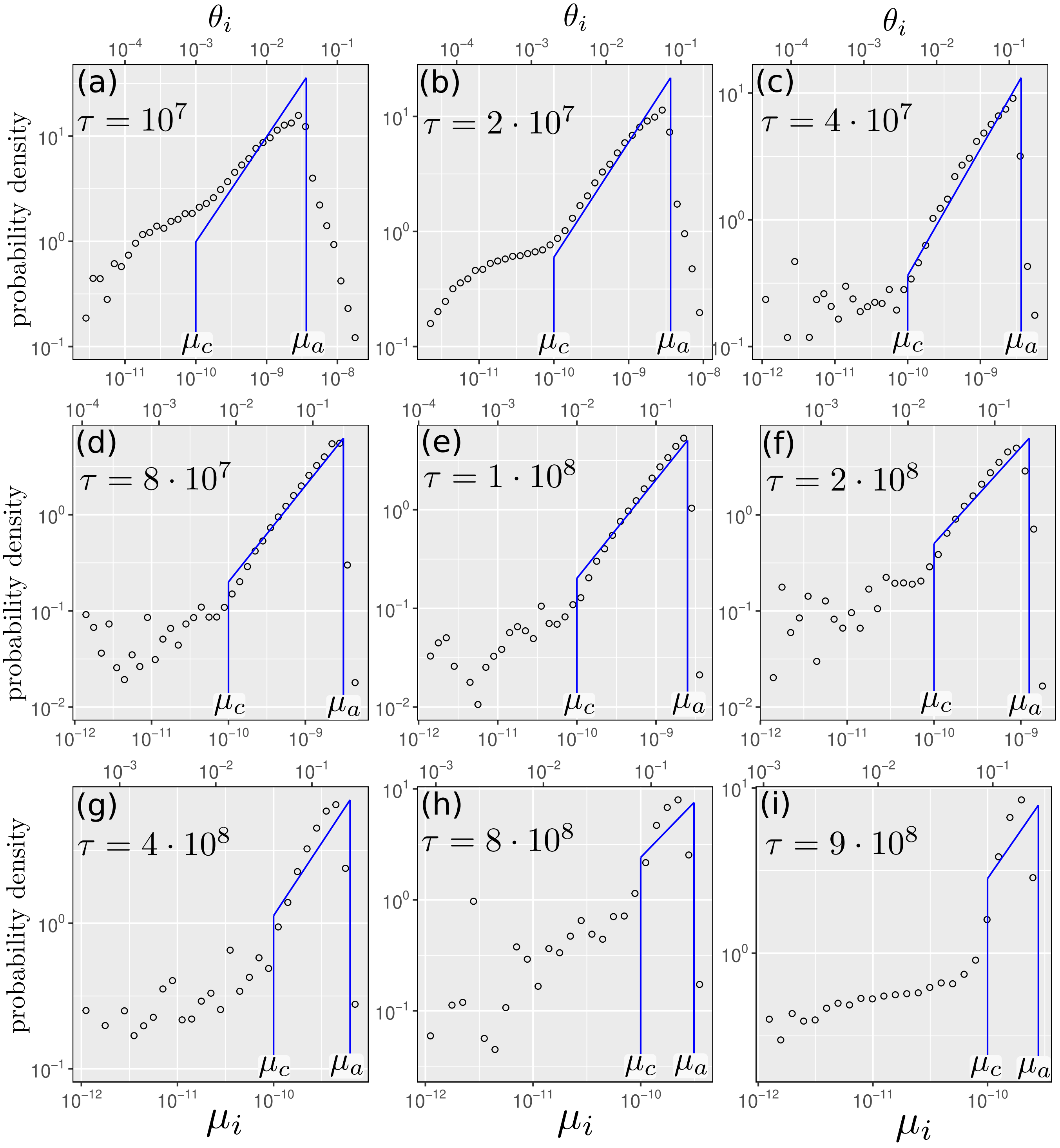

Fig. S5. Distribution of genetic distances θ_i (upper horizontal axes) along segments (longer than 100bp) of alignments of simulated pairs of genomes with different divergence times, indicated in the bottom-left corner of the panels. The HGT rate was assumed to be constant in time and taken as $\rho = 10^6/\tau^2$ from the beginning of the simulation. The segments were obtained using `segmut` R package (1). The distribution of the effective mutation rates μ_i (lower horizontal axes) is obtained using $\mu_i = \theta_i/\tau$. Blue lines correspond to the assumed distribution of the mutation rate, Eq. 1, truncated at $\mu_a = \min(\mu_s, \delta/\tau)$ with $\delta = 0.25$.

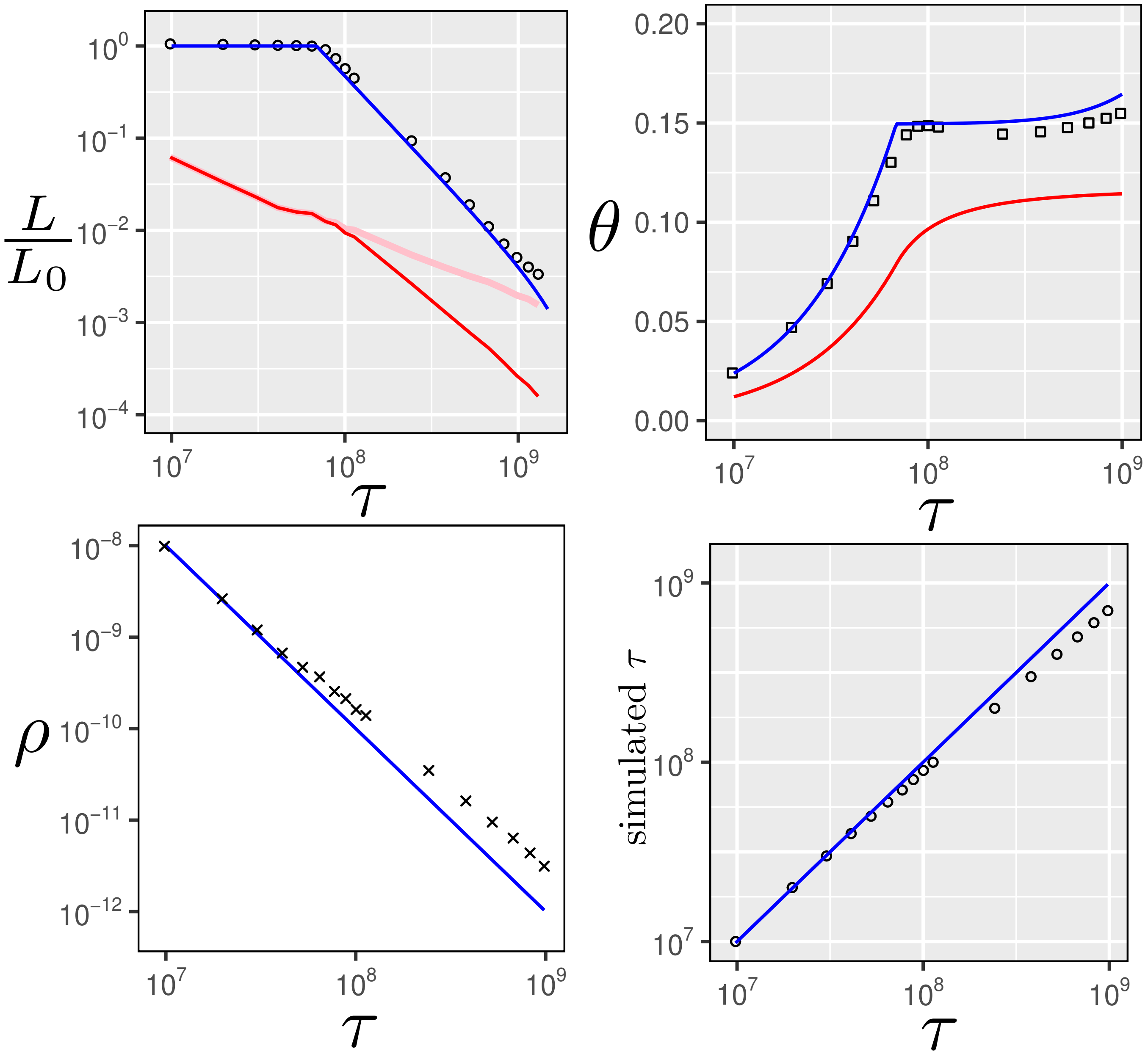

Fig. S6. Analysis of the simulated sequences. (a) Ratio of detectable homologous length and fitted value of the genome length of the common ancestor for all pairs of taxa as a function of fitted time divergence between the taxa. Blue line is L_v/L_0, the predicted length ratio from the vertical part based on Eq. 5, red line is L_h/L_0 the predicted length ratio from the horizontally transferred and detectable part based on Eq. 8. Pink line is the full (detectable and non-detectable) length ratio from the horizontally transferred part: $\rho\tau$. The length ratio L/L_0 from both detectable parts (vertically and horizontally transferred), from Eq. 11 is indistinguishable from L_v—the blue line—on this scale for these data. (b) Empirical divergences for all taxa pairs after Jukes and Cantor distance correction (51) *vs.* fitted time divergence are indicated by squares. Blue line represents the predicted by Eq. 6 divergence along the vertical part, while the red line represents the predicted by Eq. 9 divergence along the horizontally transferred part. The total predicted divergence given by Eq. 12 is indistinguishable from θ_v for these data. (c) Fitted HGT rate as a function of the fitted divergence time for all pairs of taxa. (d) Comparison between the simulated and estimated value of τ. Blue line indicates the identity relation expected for a perfect unbiased estimator.

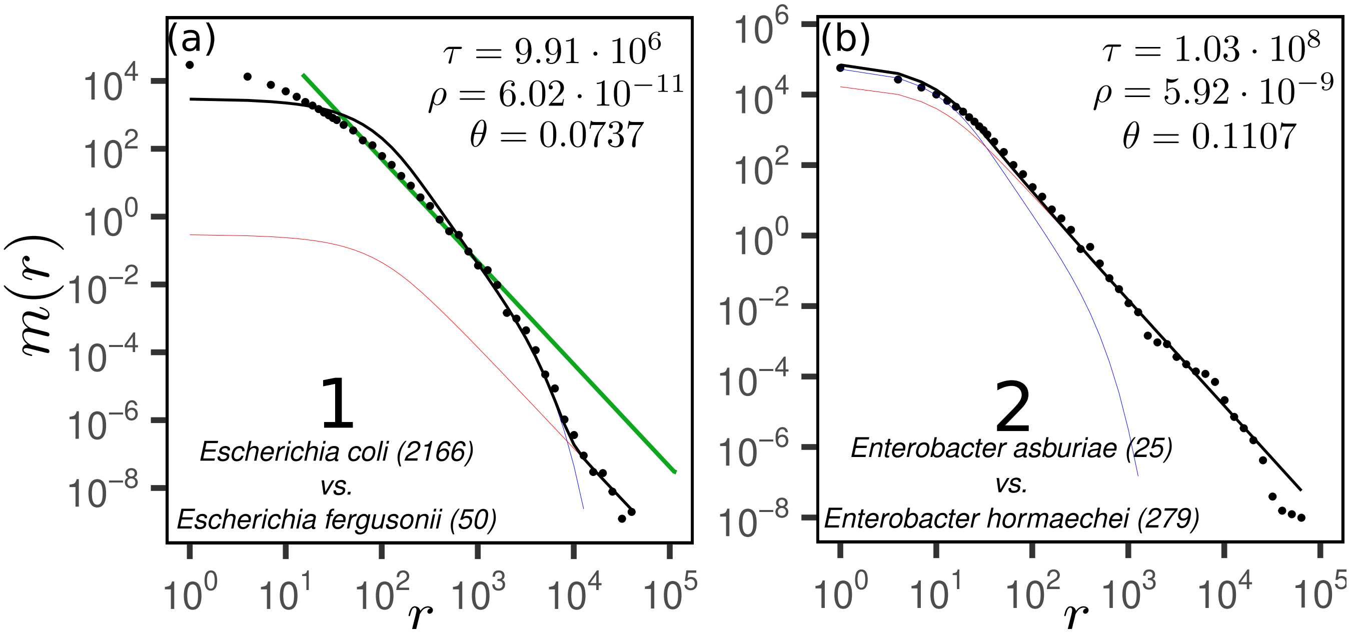

Fig. S7. Match length distributions of 2 pairs of taxa for which the model assumptions are not valid and the time divergence estimation is not accurate. Names of the taxa are indicated in the bottom-left corner of both panels. The numbers in the brackets indicate number of strains. The average divergence is indicated by θ. The empirical data (dots) are fitted with Eqs. 4,7,10 using the global parameters $\mu_s = 3.64 \cdot 10^{-9}$, $\mu_c = 10^{-10}$ and $\delta = 0.25$. The values of τ and ρ are fitted for each pair separately and are shown in the top-right corner. The genome length of the most recent common ancestor of two taxa is assumed to be the minimum of the taxas' genome lengths. The green line in (a) corresponds to $m(r) \sim r^{-3}$ power-law.

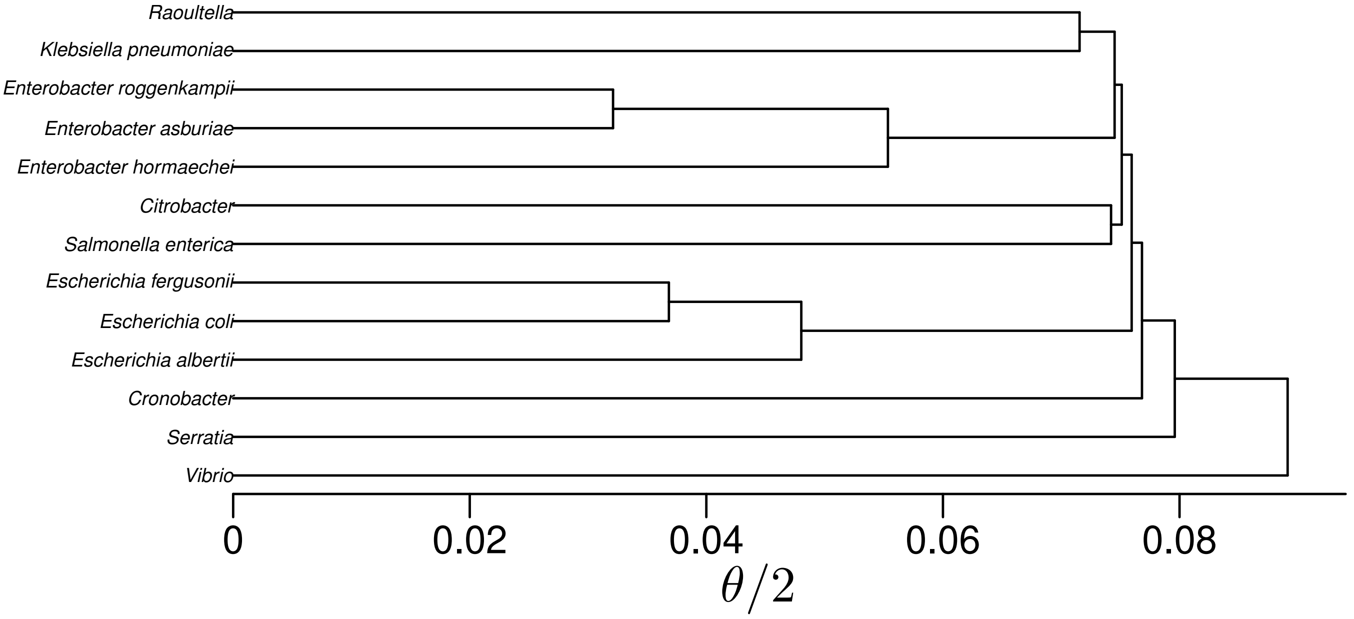

Fig. S8. UPGMA tree using the average pairwise taxa divergences θ.

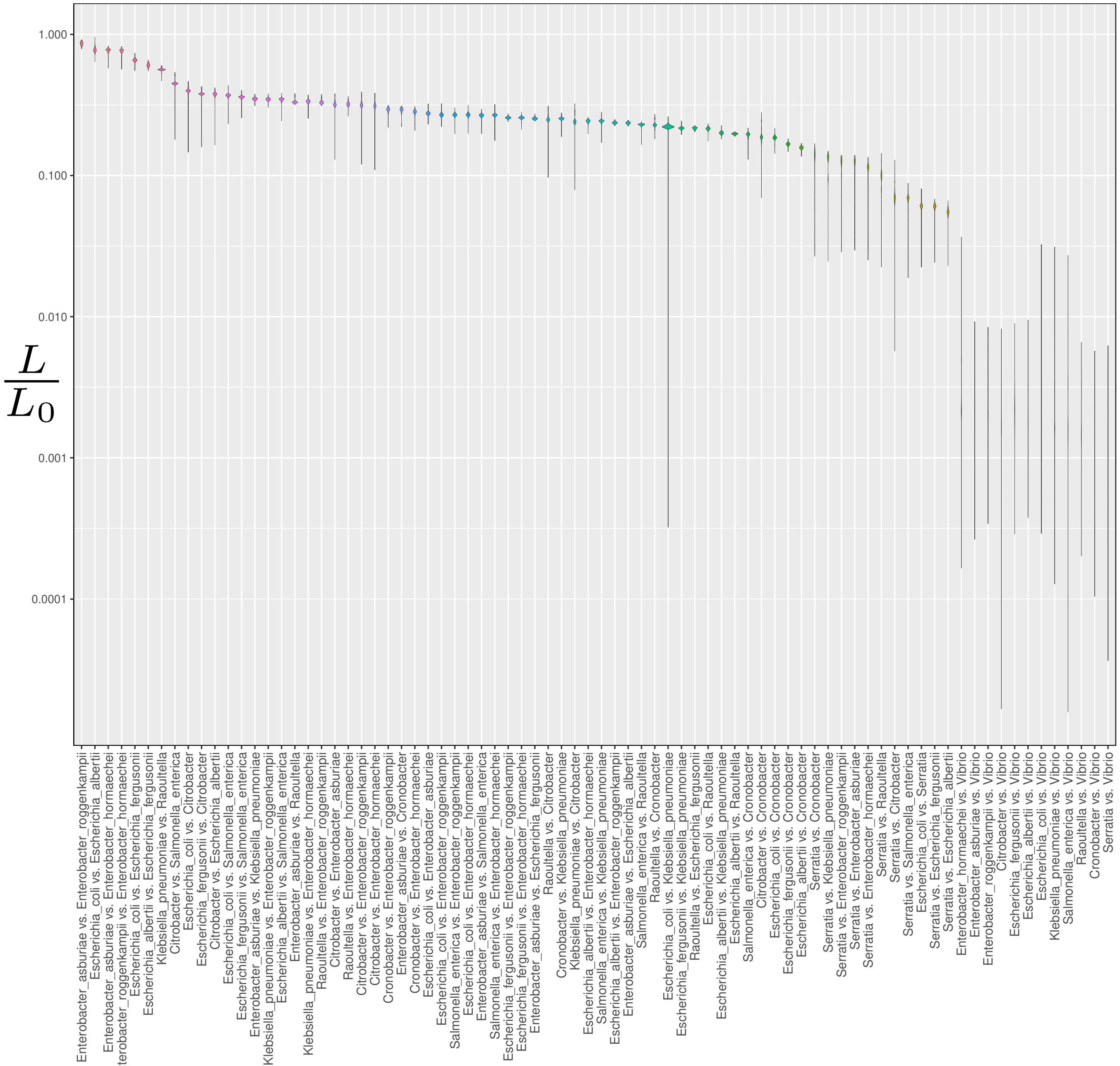

Fig. S9. Violin plot of homologous fraction of the genome, detected by the aligner for all pairwise alignments ordered by the median value.

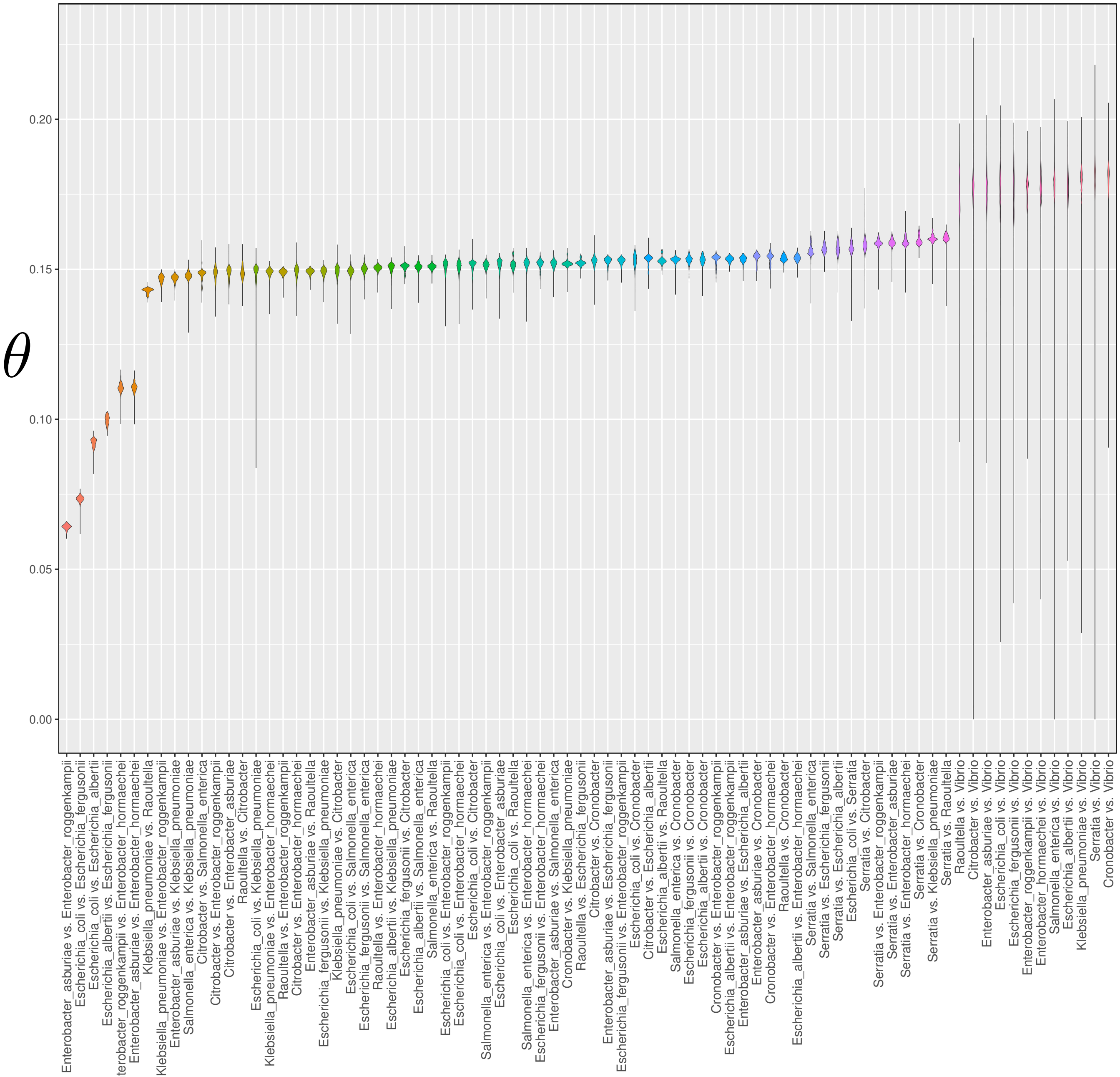

Fig. S10. Violin plot of average divergences θ for all pairwise alignments ordered by the median value.

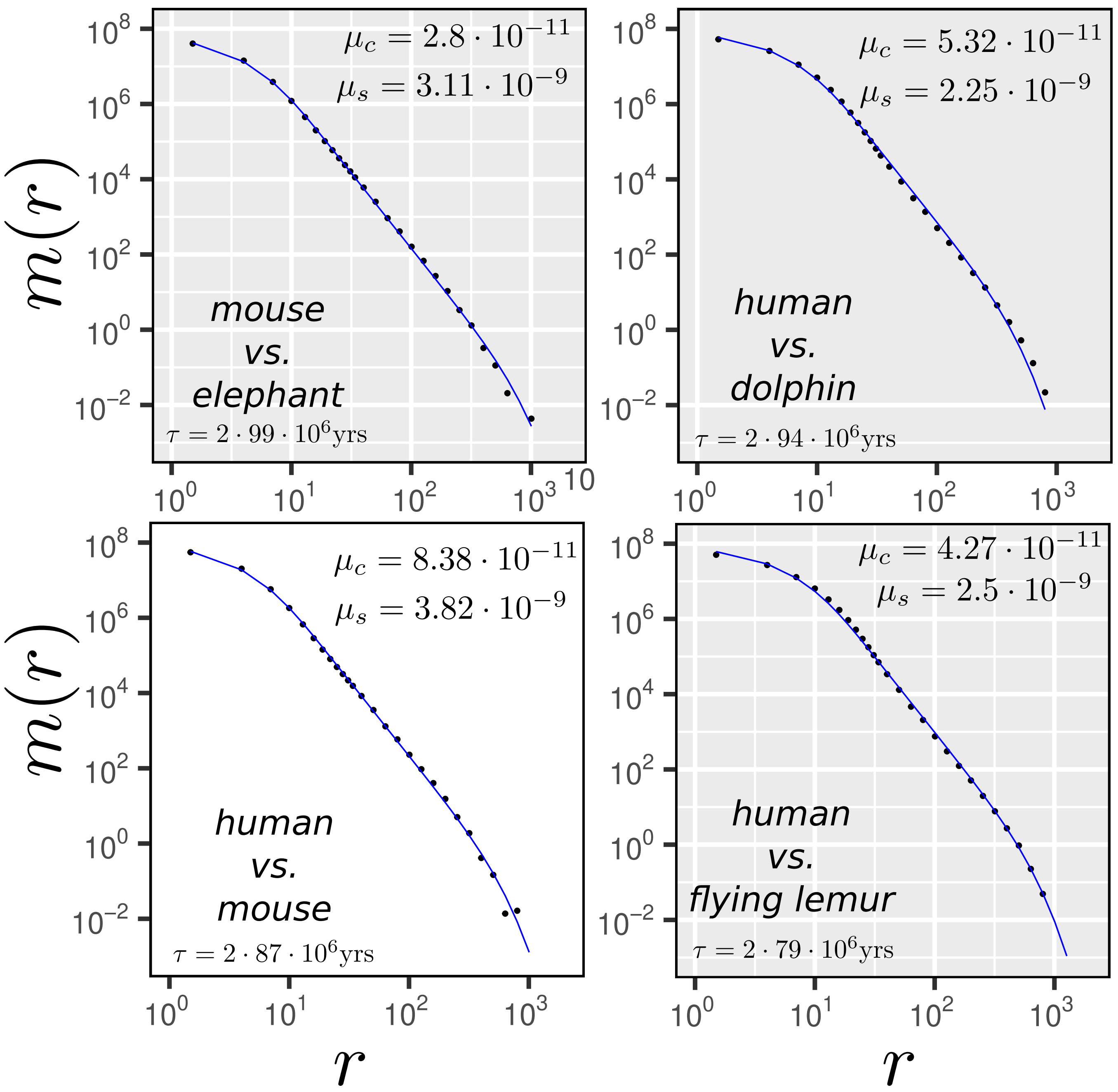

Fig. S11. MLD of animal pairs alignments, downloaded from UCSC (9) (note that here the alignment was done not by `nucmer`, but using chained `lastz` aligner (10, 11)). The analytical fit of μ_c and μ_s (see upper-right corner) was done using Eq. 4—the vertical part of the MLD (we assume that here there is no horizontal transfer). The divergence times τ are taken from Ref. (12) (see bottom-left corners).

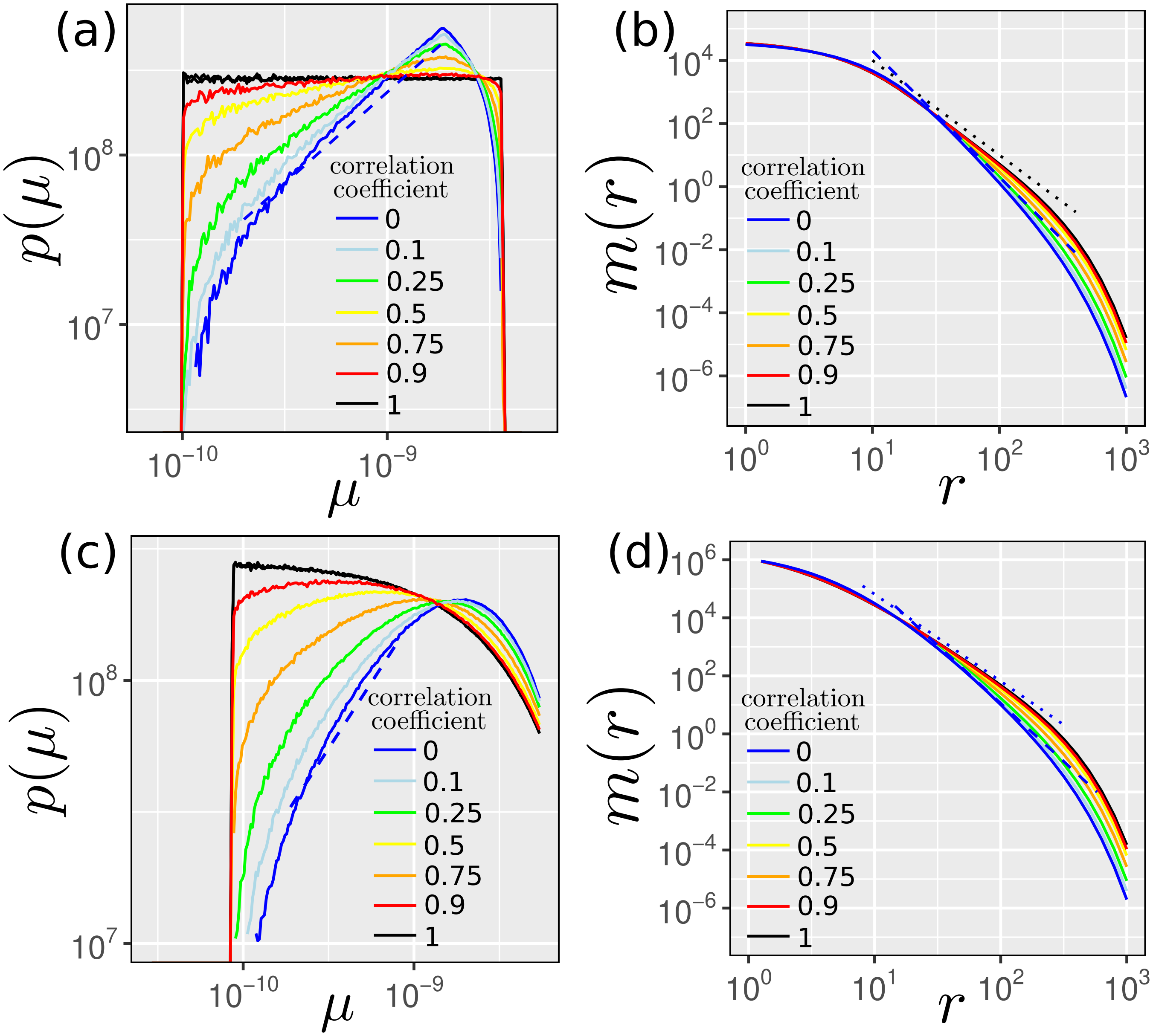

Fig. S12. Impact of correlation between mutation coefficients of two species loci on the match length distribution. (a) Distribution of mean of two correlated random variables $\mu = (\mu^A + \mu^B)/2$ with varying correlation coefficient (see legend). Correlated variables were generated using the `simstudy` R package (13). Both μ^A and μ^B are uniformly distributed between $\mu_c = 10^{-10}$ and $\mu_s = 3.64 \cdot 10^{-9}$ (see Eq. S.1). Dashed line represent linear dependence. (b) MLDs calculated from the mutation rates from panel (a) using $m(r) = \int p(\mu)(\mu\tau)^2 e^{-\mu\tau r} d\mu$ with $\tau = 10^8$. Dashed line corresponds to r^{-4} and dotted line to r^{-3} dependencies. (c-d) The same as upper panels, but both μ^A and μ^B are exponentially distributed with $\mu_s = 3.64 \cdot 10^{-9}$ mean and truncated at μ_c (see Eq. S.6).

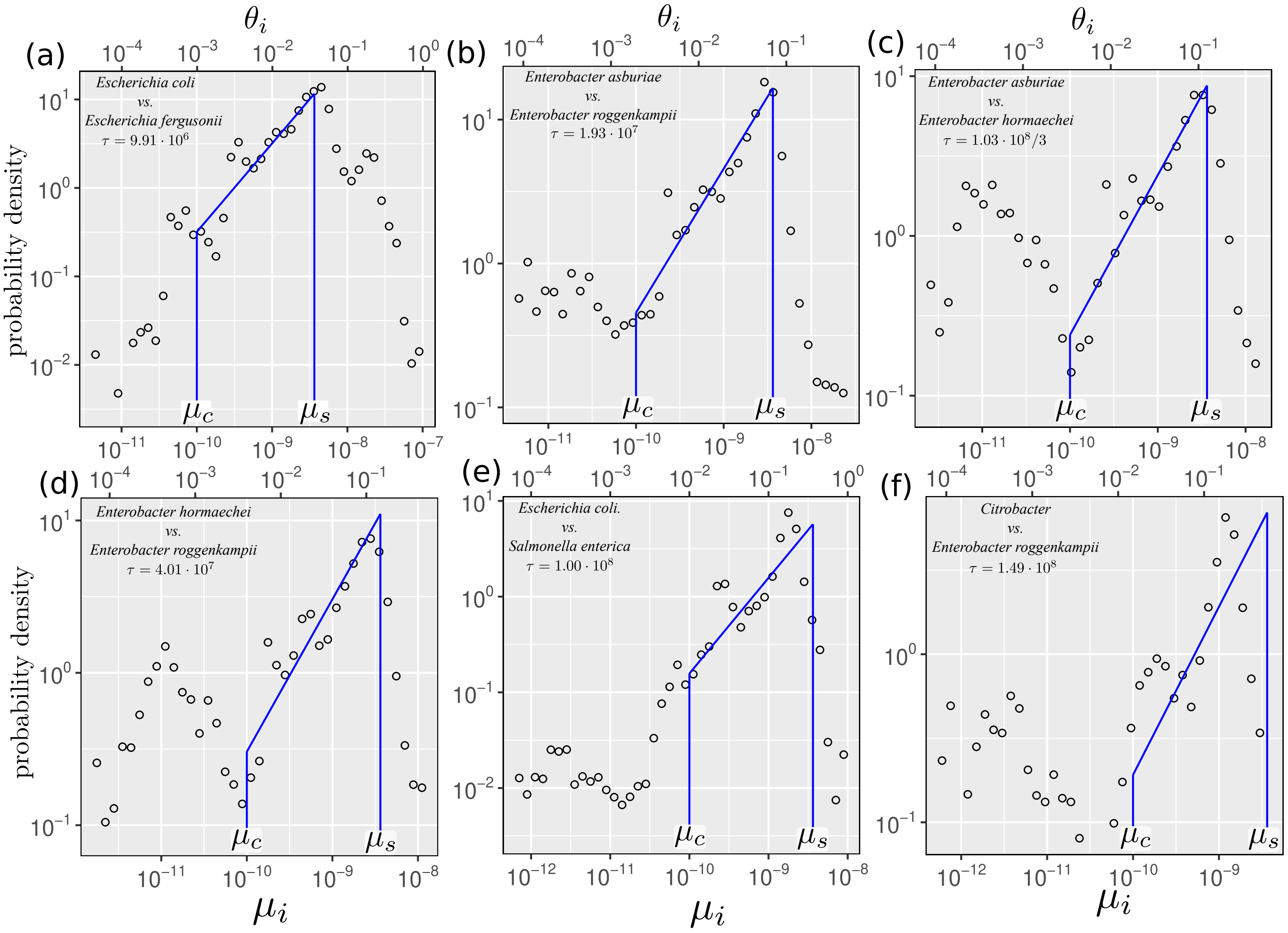

Fig. S13. Distribution of genetic distances θ_i along segments of alignment between different taxa pairs (see upper-left corners). The segments were obtained using `segmut` `R` package (1). The distribution of the effective mutation rates μ_i is obtained using $\mu_i = \theta_i/\tau$. The values of τ are taken from the fitting of the MLD for each taxa pair. For the taxa pair in panel (c) the estimation of τ failed and to get a good fit in panel (c) τ was taken as $3.4 \cdot 10^7$. Blue lines correspond to the assumed distribution of the mutation rate Eq. 1.

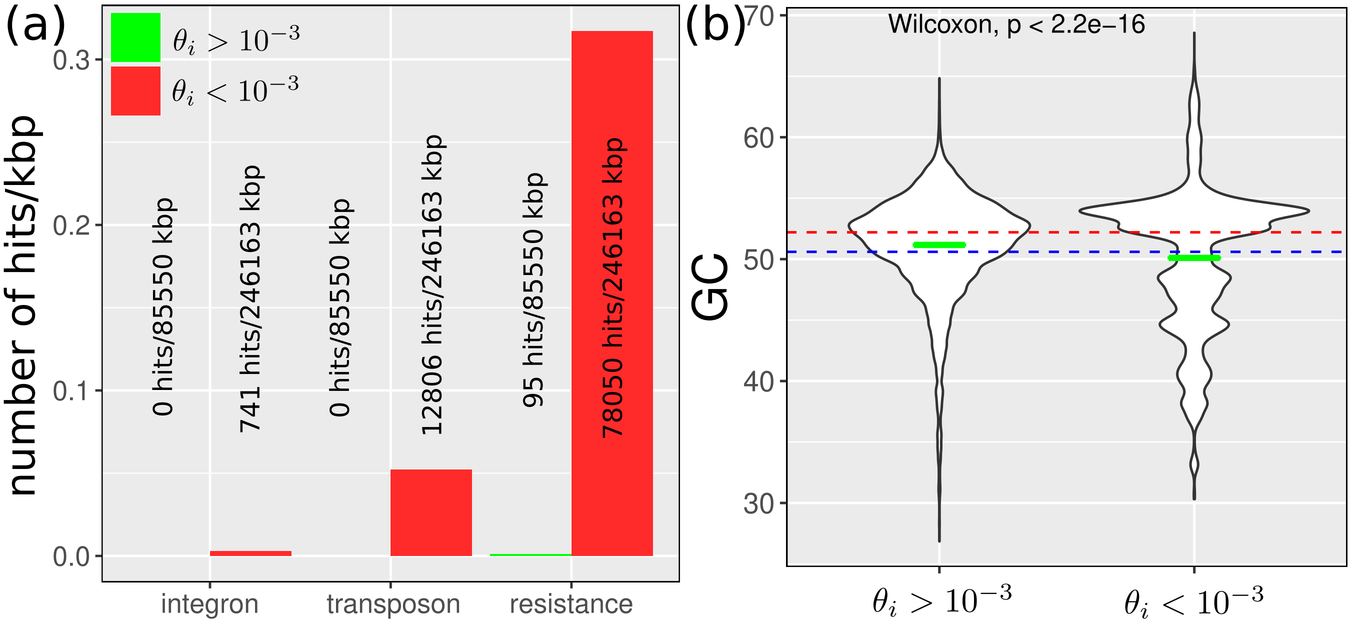

Fig. S14. (a) Functional analysis of randomly sampled 49947 (with $\theta_i > 10^{-3}$) + 58671 (with $\theta_i < 10^{-3}$) = 108,618 segments obtained from `segmut` segmentation of *E. coli vs. S. enterica* alignments. The total length of all $\theta_i < 10^{-3}$ segments is 246,162 kbp while $\theta_i > 10^{-3}$ segments comprise 85,550 kbp. Each segment was searched for homology against the database of resistance genes, transposable elements and integrons using `BacAnt` software (3). In the figure the number of hits and the hits density per kbp are presented for each functional category. (b) Violin plot of the GC content of less (left) and more (right) similar segments of the *E. coli vs. S. enterica* alignments. Green lines indicate the average GC content of the two groups. Blue (red) dashed line correspond to GC content of *E. coli* (*S. enterica*).

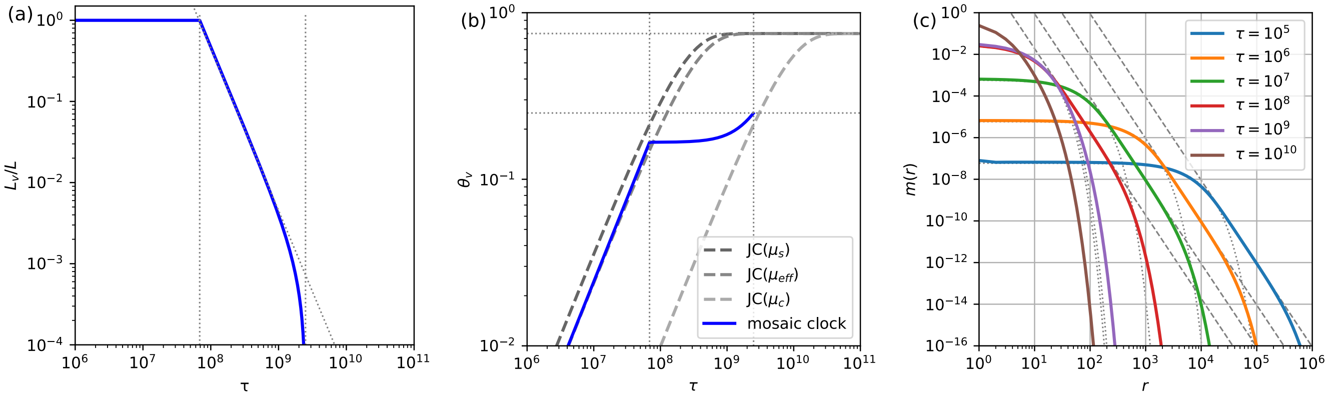

Fig. S15. (a) δ-detectable fraction of a genome. The vertical lines are at δ/μ_s and δ/μ_c. The detectable part decreases proportional to $1/\tau^2$ for $\delta/\mu_s < \tau \ll \delta/\mu_c$ as indicated by the dotted line. (b) The empirical divergence as a function of τ. The dashed lines represent the prediction due to the Jukes-Cantor model with rates μ_s, $\mu_{eff} = (2/3)(\mu_s^3 - \mu_c^3)/(\mu_s^2 - \mu_c^2)$, and μ_c. (c) Match length distributions for several values of τ. The functions for small $\tau < \delta/\mu_s$ show a power-law regime with exponent -4, as indicated by the dashed gray lines. Corresponding exponential distributions with the same mean are shown with gray dotted lines.

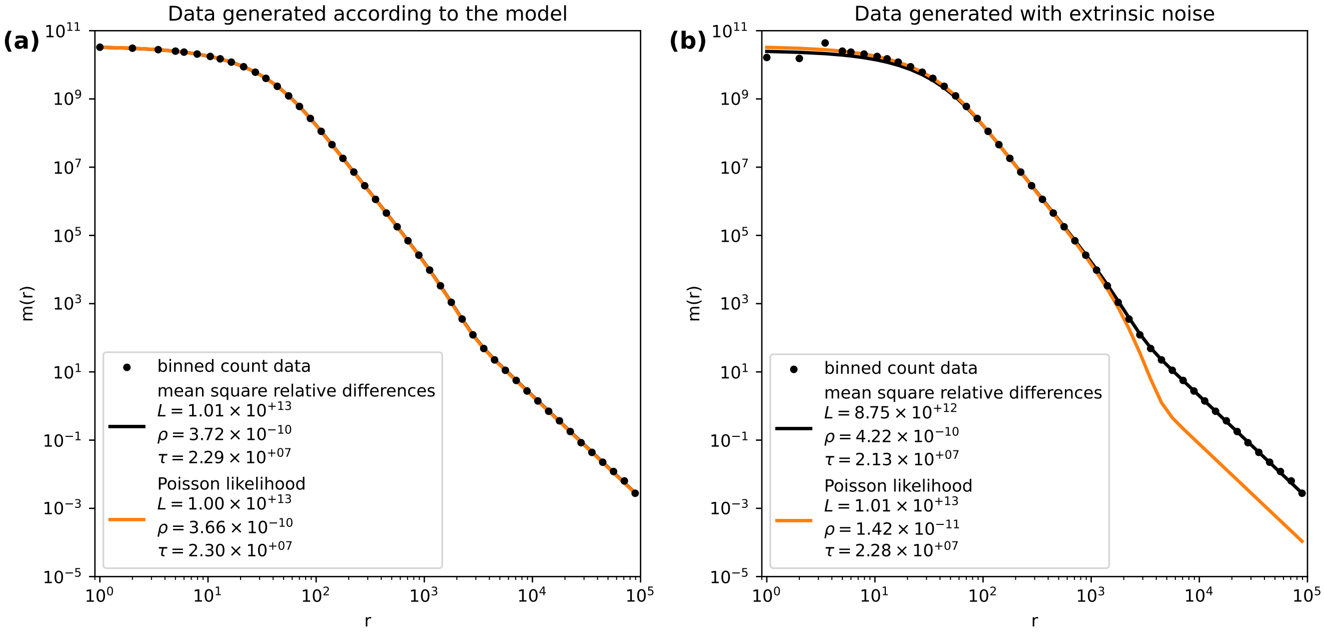

Fig. S16. Comparison of objectives for fitting MLD to binned count data. (a) We generated count data according to our model for $L = 10^{13}$, $\tau = 2.3 \cdot 10^{7}$, and $\rho = 3.66 \cdot 10^{-10}$. After binning the data we fitted MLDs minimizing the mean square relative differences or maximizing the Poisson Likelihood. Both fit the data equally well. (b) To exemplify the robustness of parameter interference in the presence of noise due to biological reasons on small length scales we distorted the data in panel (a) and moved half the counts from $r = 1$ and 2 to $r = 3$. In this scenario the fit minimizing the mean square relative differences yields a better overall fit. Estimated values are given in the legends.

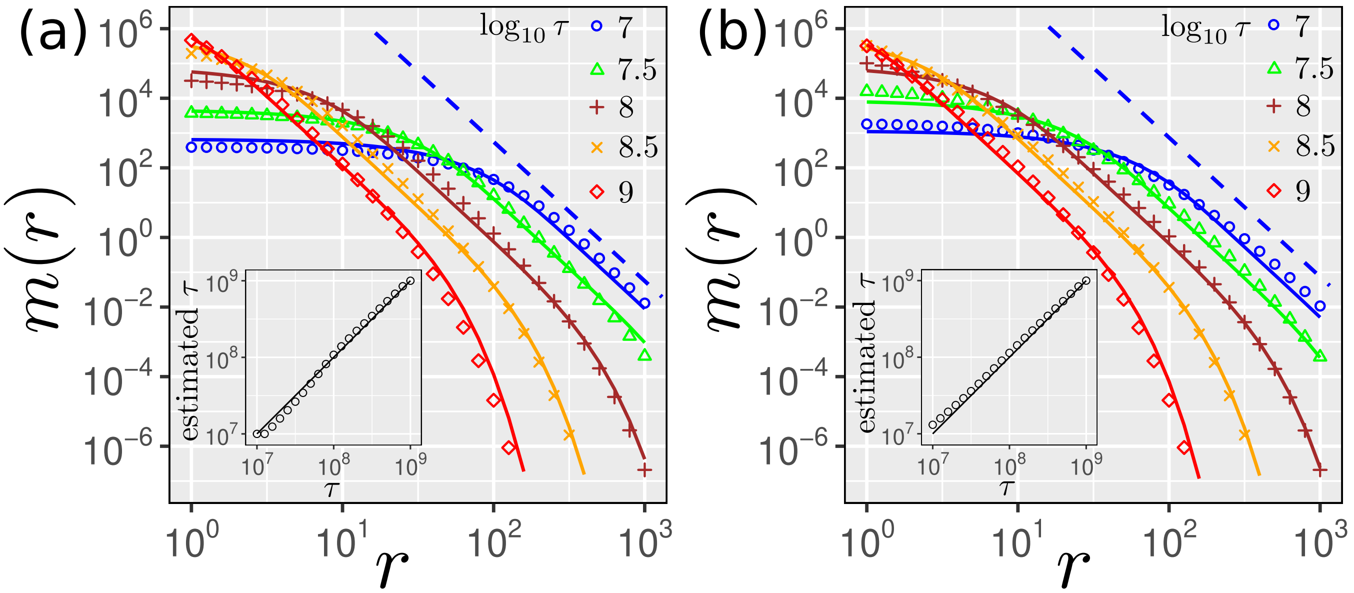

Fig. S17. Robustness of the inference procedure to the distribution of the mutation rate. Here we calculate the MLD for (a) uniform distributions (Eq. S.1) and (b) truncated exponential distributions of $p^A(\mu^A)$ and $p^B(\mu^B)$ (Eq. S.6). We assume no correlation between mutation rates in taxa A and B, ignore HGT events and the finite sensitivity of the alignment software. The results of the simulations for different values of τ are presented as markers while fits (using Eq. S.5, i.e. a specific case of the general formula Eq. 10 with no HGT events, $\rho = 0$, and high sensitivity limit, $\delta = 1$) are represented by the solid lines. Dashed lines represent the $m \sim r^{-4}$ power-law. For a representative set of values of τ we compare the estimated value of τ to the used one in the insets.

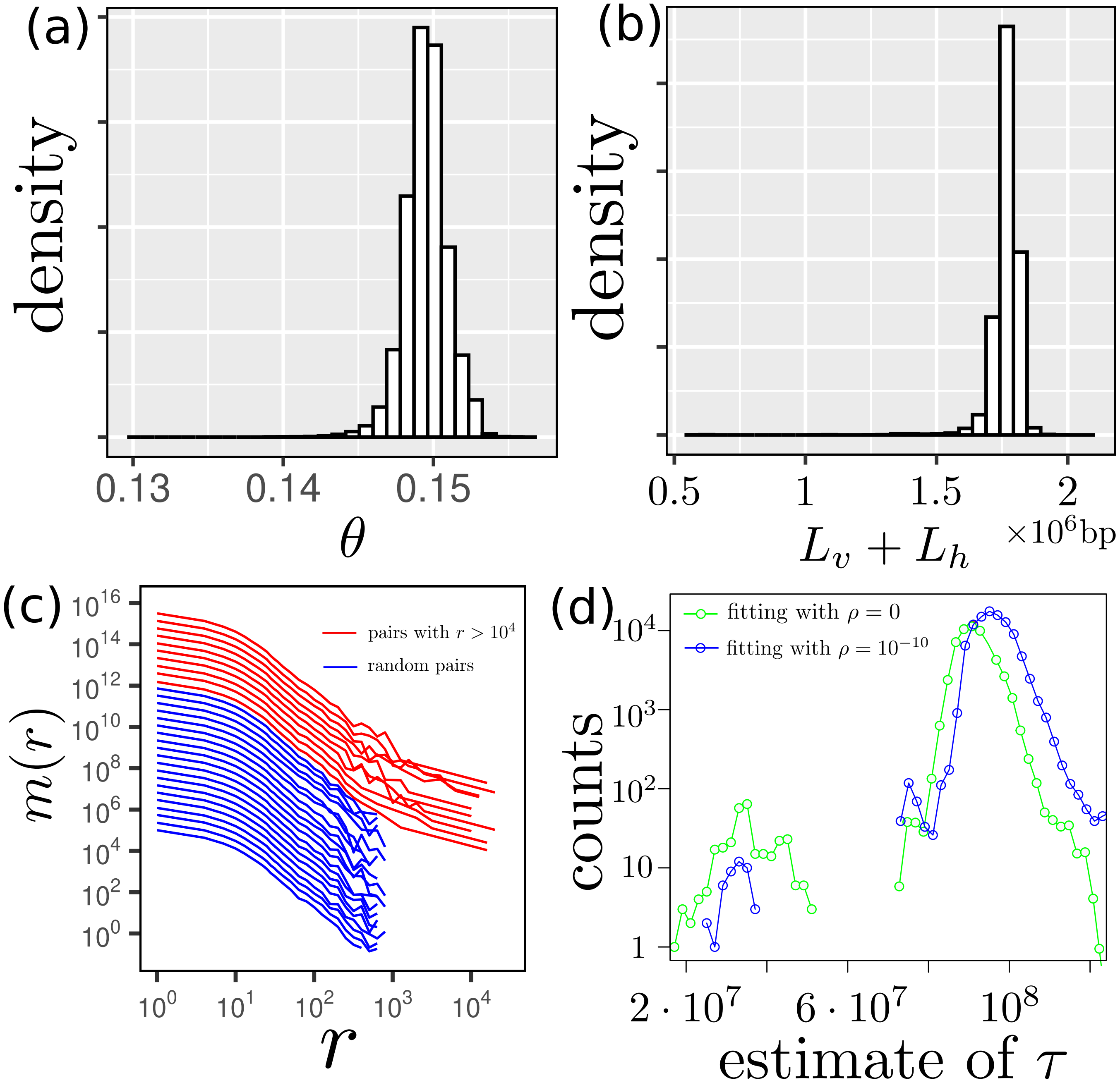

Fig. S18. Analysis of individual *E. coli vs. S. enterica* pairs. (a) Distribution density of divergences between pairs of genomes. (b) Distribution density of length of alignable region of pairs of genomes. (c) A few examples of MLDs using (blue) random pairs and (red) pairs with at least one long ($r > 10^4$) exact match. Each MLD is normalized differently for visibility. (d) Distribution of estimates of time divergence τ between the two taxa using fits of MLDs by Eq. 10 with $\rho = 0$ (red) and $\rho = 10^{-10}$ (blue) HGT rate.