Burton, Corey A. M.S. Variations in recombination rates across *Escherichia coli* populations (2022)
Directed by Dr. Louis Marie Bobay. 17 pp.

The accumulation of bacterial genomic datasets has created a nuanced and difficult challenge for computational analyses. Based on the current trend of genomes being sequenced, it appears that it won't be possible to infer complex parameters such as recombination rates for these entire genomic datasets. We assessed the impact different sampling strategies had on recombination rate estimates, along with the impact of gene content and population structure on recombination rate estimates. Overall, we found that while our novel framework yielded consistent estimates of recombination rates, our sampling strategies, population structure, and gene content did not significantly impact recombination rate estimates.

VARIATIONS IN RECOMBINATION RATES ACROSS *ESCHERICHIA COLI* POPULATIONS

by

Corey Burton

A Thesis

Submitted to

the Faculty of the Graduate School at

The University of North Carolina at Greensboro

in Partial Fulfillment

of the Requirements for the Degree

Master of Science

Greensboro

2022

Approved by

Dr. Louis-Marie Bobay
Committee Chair

APPROVAL PAGE

This thesis written by Corey Burton has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Bobay

Committee Members

Dr. Raymann

Dr. McLean

April 28, 2022
Date of Acceptance by Committee

April 28, 2022
Date of Final Oral Examination

TABLE OF CONTENTS

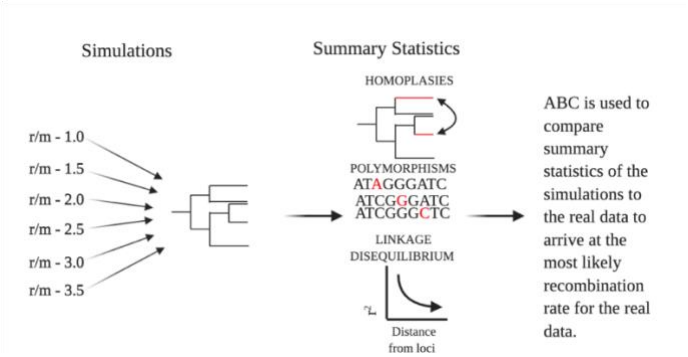
LIST OF FIGURES

CHAPTER I: INTRODUCTION

While bacteria reproduce asexually through binary fission, it is now known that their evolution is also driven by the exchange of various levels of genetic information (Gogarten et al., 2002). In particular, homologous recombination is driving the transfer of short sequences of DNA between chromosomal regions sharing sequence homology and this process is thought to facilitate adaptation (Didelot et al., 2012). However, quantifying the amount of DNA transferred by recombination across species has proven difficult and remains a technical challenge.

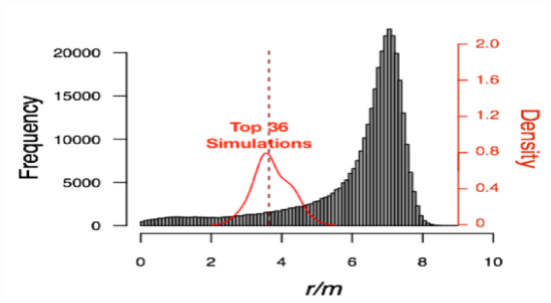
Knowledge on recombination rates is sparse, and while many studies have attempted to quantify this process, these estimates have been derived from various methodologies and datasets (Bobay et al., 2015). As a result, reported estimates of recombination rates have been highly inconsistent across studies, as evidenced by independent studies focusing on the same species. Despite technical biases inherent to the application of different methodologies and different datasets, biological variations may also contribute to these inconsistencies. Bacteria display structured populations and recombination rates can vary within and across populations (Didelot et al., 2012). Ecological factors may also facilitate recombination between populations living in the same niche relative to those living in different niches. Finally, recombination might be beneficial for strains sharing certain traits or phenotypes, and selection might then favor recombination between these populations. Therefore, one key challenge remains to disentangle to what extent these variations of recombination rate estimates are due to technical challenges or biological factors.

Figure 1. ABC Framework



Much effort has been conducted to estimate the recombination rate of the model bacterium *Escherichia coli.* Over several decades of work *E. coli* has been estimated to be non-recombining at all in some studies, to recombining at extremely high rates in other studies (Bobay et al., 2015). More recent studies, however, tended to infer an intermediate rate of recombination for this species. *E. coli* offers an ideal test case to test approaches aiming at estimating recombination rates (Touchon et al., 2009; Didelot et al., 2012). Indeed, this species presents a wealth of genomic data with over 20,000 complete genomes available so far. In addition, *E. coli* has a well characterized population structure and has been divided into seven main phylogroups: A, B1, B2, C, D, E, and F (Tenaillon et al., 2010). Finally, *E. coli*'s strains exhibit various phenotypes and live across differing environments. For instance, it comprises commensal strains, uropathogenic strains, enteropathogenic strains, and *Shigella* strains causing shigellosis (Sims & Kim., 2011).

Figure 2. Bayesian Output



In this study, I aim to apply and test a novel method based on Approximate Bayesian Computation (ABC) (**Figures 1 and 2**) to estimate recombination rates in *E. coli*. Using subsampling approaches, I will determine to what extent this method yields consistent estimates of recombination rates across genomic datasets sampled from the same species. I will further analyze the impact of population structure, bacterial ecology, and genome content on these estimates. Results will establish to what extent this new method is reproducible across samples and how much of the variation of recombination rates are due to biological factors rather than technical biases. Altogether, this work will determine how to best subsample large genomic datasets to derive robust estimates of recombination rates.

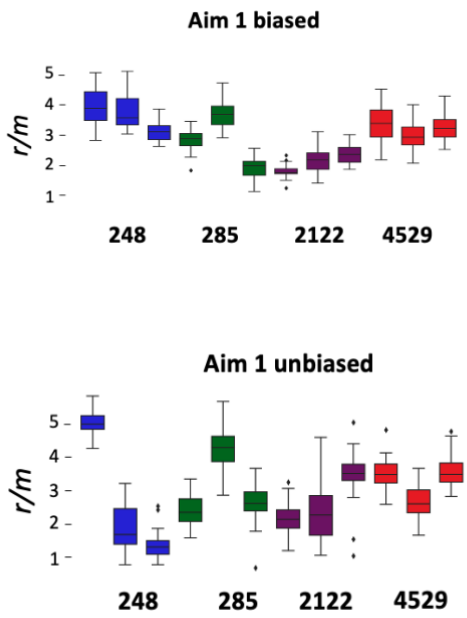
CHAPTER II: RESULTS

The main objective of these analyses is to assess the robustness of recombination rate estimates derived from a new unpublished method based on ABC recently developed in the lab. The overall approach consists of simulating genome evolution with known parameters of recombination rates. The simulated genomes are then compared to the real dataset using three summary statistics which are known to be signatures of recombination. The simulations, and their corresponding recombination rates, which exhibit summary statistics closest to those inferred for the real dataset are then inferred as the most probable recombination rates based on ABC (see Methods). For each sampling, genomes are evolved *in silico* using *CoreSimul*, which is a simulator of genome evolution for prokaryotes, with parameters specific to each sampling of *E. coli* genomes such as nucleotide composition, topology of the phylogenetic tree, substitution rates, and transition/transversion ratio (Bobay., 2020). These parameters are empirically estimated from the dataset. These genomes are evolved *in silico* with diverse rates of recombination. Recombination rates are expressed as the number of alleles exchanged by recombination relative to mutation events (r/m). The three summary statistics are then inferred for each simulation and for the real dataset. By conducting many simulations (n>300,000 for each estimate) of genome evolution with diverse recombination rates, we can compare these summary statistics across thousands of simulated genome datasets to the summary statistics estimated in the real samples of genomes. Based on these summary statistics, our ABC procedure allows us to estimate which simulations are the most similar to the real datasets of *E. coli* genomes, and therefore to infer which recombination rates are the most probable for these genomes.

Aim 1: Impact of sampling biases on recombination rate estimates

The first goal of my analysis was to establish the robustness of recombination rate estimates to basic genome sampling biases. Like most genomic analyses, our ABC approach is unable to process the entirety of genomic datasets (i.e. over thousands of genomes), and estimates are therefore derived from smaller samples. I conducted diverse sub-samplings of the same dataset to measure to what extent these estimates varied from one another. I analyzed a previously published dataset of 400 non redundant *E. coli* genomes assembled in Bobay & Ochman 2018. This dataset represents the set of >1,000 core genes that were aligned and concatenated into a single alignment. Different subsamplings were conducted on this concatenate. First, I aimed to determine how consistent recombination rate estimates were across random unbiased samplings. I also measured how the presence of genomic outliers impacts my estimates by generating sampling biases with uneven genomic divergence. Unbiased and biased subsampling of genomes from four different subtrees of the whole *E. coli* tree were conducted, each with three replicates consisting of fifteen genomes each. Unbiased samplings were conducted by randomly selecting fifteen genomes from the same subtree of *E. coli*. Biased samplings were conducted by randomly selecting fifteen genomes from the same subtree and by including one more divergent genome from a different subtree. Recombination rates were then computed via our ABC framework across these different samples. In this analysis, I am estimating the effective rate of recombination (*r/m*) which represents the number of alleles exchanged by recombination relative to the number of substitutions.

Figure 3. Aim 1 Output



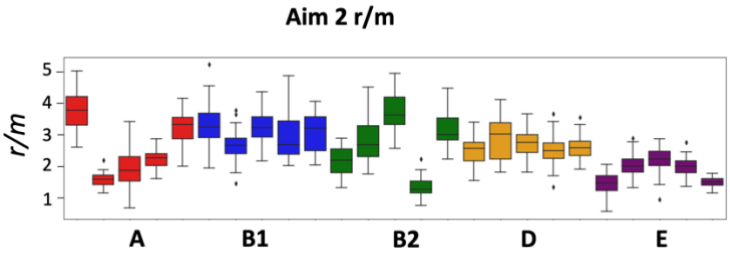
Estimates of the recombination rate varied from $r/m = 0.96$ to $r/m = 4.48$ for the unbiased samplings and $r/m = 1.66$ to $r/m = 3.09$ for the biased samplings. **(Figures 3)**

The introduction of an outlier had a weak impact on *r/m* estimates, ($P < 0.05$, Kruskal-Wallis test) but the origin of the sampled subtree had a more pronounced effect on the estimates ($P < 10^{-15}$, Kruskal-Wallis test). Interestingly, the introduction of a more divergent genome did not systemically decrease *r/m* estimates, indicating that recombination rates are not systematically lower among more distantly related strains. Overall, recombination rate estimates appear to be modestly impacted by sampling biases. Although sampling strategies should be designed to minimize biases by including genomes that are representative of the overall genomic diversity of the species, our results show that the inclusion of outliers did not substantially impact *r/m* estimates. However, significant variations were found based on the origin of the sampling on the tree, which suggests that variations in recombination rates may be driven by population structure.

Aim 2: Impact of population structure on recombination rate estimates

Although many works have attempted to estimate the recombination rate of a given species, it has been suggested that recombination rates may vary across populations within the same species (Didelot & Maiden., 2010). Indeed, populations are strongly structured in bacteria and this might lead to preferential patterns of recombination within and between the population of each species **(**Touchon et al., 2020**).** The previously published phylogenetic tree of the 400 *E. coli* genomes was used to classify each genome into their respective phylogroups using previous classifications (Beghain et al., 2018; Diamant et al., 2004; Sims & Kim, 2011). Most of the genomes in the tree could be assigned to one of the five main phylogroups of *E. coli* (A, B1, B2, D, E) by randomly sampling 15 genomes within each phylogroup, this analysis was repeated five times per phylogroup and each sample of 15 genomes was used to estimate recombination rates with our ABC approach as in Aim 1.

Figure 4. Aim 2 Output



Results show that recombination rate estimates were significantly different across several phylogroups (**Figure 4**). Phylogroups B1 and E presented recombination rate estimates that were systemically significantly different from all other phylogroups and from one another ($P<0.05$, Wilcoxon test with Bonferroni correction). The last three phylogroups A, B2, and D did not display significantly different estimates of recombination rates from one another($P>0.05$,

Wilcoxon test with Bonferroni correction). Thus, this confirms that population structure does play some role in shaping recombination rates. However, variations in recombination rates, albeit significant, were relatively modest when compared to the possible range of recombination rates that were reported across different species (Vos., 2008). Indeed, our estimates varied from a minimum of $r/m = 1.41$ to a maximum of $r/m = 3.76$, while variations across species have been reported to vary from $r/m = 0$ to $r/m = 60$ (Vos., 2008). Moreover, we observed that *r/m* estimates were highly consistent across samplings obtained from phylogroups B1, D, and E. In contrast, *r/m* estimates were much more variable across samplings conducted within phylogroups A and B2.

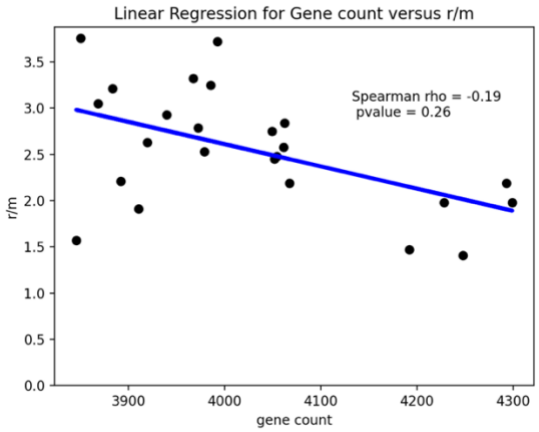
These results indicate that, although recombination rates appear to vary significantly across some *E. coli* phylogroups, they can also vary substantially vary within each phylogroup (**Figure 4**). The fact that *r/m* estimates can vary widely within the same phylogroup suggests that other factors may be driving the variations of recombination rates within *E. coli's* genomes Previous studies have reported that Phylogroup A is primarily composed of commensal strains, and that phylogroup D is mostly composed of enteropathogenic strains (Touchon et al., 2020). It is therefore possible that the observed variations in recombination rate estimates across phylogroups may in fact reflect the heterogeneity of their ecological niches and lifestyles. Although the ecology of a particular strain is difficult to predict, studies have suggested that strains sharing similar ecological niches tend to share similar gene contents (Touchon et al., 2020).

Aim 3: Impact of gene content on recombination rate estimates

I am therefore hypothesizing that strains sharing more similar gene contents are more likely to display higher recombination rates than strains that share less similar gene contents. To

test this hypothesis, I compared the gene content of all the analyzed genomes in my dataset. I defined the set of orthologous genes for each pair of genomes

Figure 5. Aim 3 Output



by best reciprocal hit (see Methods) and from this, I estimated the average number of orthologous genes shared within each sampling. The strains from the different samplings shared from 3827 to 4299 genes on average. Some phylogroups displayed higher levels of gene content similarity; for instance, phylogroup E showed higher average gene content similarity and this group is primarily composed of enteropathogenic strains with similar phenotypes and infectious strategies (Denamur et al,. 2021). Overall, I did not observe any significant correlation between gene content and recombination rate (**Figure 5,** rho=-0.19,P=0.26), indicating that strains with more similar gene content are not likely to engage in recombination more frequently.

CHAPTER III: METHODS

Aim 1: Subsampling approach

Subsampling analyses were conducted from the core genome alignments of 400 non-redundant stains of *E. coli* assembled in Bobay & Ochman 2018. The phylogenetic tree was built with RAxML using a GTR + Gamma model. Several subtrees of the phylogeny were randomly selected to conduct the samplings, which were done via the *random* library in Python. To generate the unbiased sampling, multiple random samplings of 15 genomes were conducted from different subtrees. The biased samplings were generated by randomly selecting 15 genomes from a subtree and adding a randomly selected genome from a different subtree. For each of the generated samplings, the core genome alignments were extracted from the main alignment and a maximum likelihood phylogeny was generated for each alignment using RAxML with a GTR + Gamma model. From the phylogenetic tree and core genome alignments of each sampling, additional statistics were inferred: nucleotide composition, transition/transversion ratio, and polymorphisms across codon positions.

Aim 1: Estimation of recombination rates

Our ABC framework consists of simulating genomes under various recombination rates and to compare several summary statistics that represent signatures of recombination between the simulated and real genomes. The underlying idea is that simulations that were evolved with the right rate of recombination should give rise to genomic signatures that are similar to those observed in the real dataset. For each sampling, recombination rates were estimated with our ABC approach independently. To obtain *r/m* estimates, genomes were evolved 300,000 times using *CoreSimul* with various recombination rates (from rho/theta = 0 to 20) following a Poisson distribution and different average recombination tract length (from delta = 10 to 1,000) following

a geometric distribution. Rho/Theta represents the rate of DNA fragments exchanged by recombination relative to the number of substitutions. In contrast, *r/m* represents the effective rate of recombination which represents the number of alleles exchanged by recombination relative to the number of substitutions. All sites in the genome had equal probabilities of recombination and all branches existing at the same time *t* in the tree had equal chances of recombining. For each recombination event, the amount of transferred alleles *nu* was recorded. For these parameters, the effective rate of recombination was defined as:

$$\frac{r}{m} = \delta \times \nu \times \frac{\rho}{\theta}$$

(Didelot & Wilson., 2015). Each simulation was initiated by generating a random sequence with the length of the sampling alignment and the same nucleotide composition. The sequence was evolved along the tree of its corresponding sampling with a substitution rate estimated from the branches of the tree. Substitution rates followed a K2P model using the transition/transversion ratio empirically inferred for each sampling as explained above. Substitution rates were also varied across the three codon positions, whose relative rate was defined for each sampling as described above. Each simulated alignment was then analyzed to infer three summary statistics that were compared to the three summary statistics inferred from the real genomes of each sampling. Our first summary statistic, the homoplasy ratio (*h/m*) represents the ratio of homoplasic alleles to non-homoplasic alleles. Recombination is known to increase the number of homoplasic alleles i.e. alleles that are not consistent with the vertical inheritance from a single common ancestor. The second summary statistic is Linkage Disequilibrium (LD), which measures the co-inheritance between pairs of alleles located at two loci on the chromosome, recombination decreases LD as a function of the distance between loci. Finally, nucleotide diversity (π), is used as the third summary statistic. Nucleotide diversity is the average number of

nucleotide differences in a sample of genomes. This metric is known to be strongly correlated with recombination rates, since *r/m* measures the effective number of alleles exchanged by recombination. This metric ensures that the simulations have been conducted using realistic mutation and recombination rates. For each sampling, the three summary statistics were compared between the 300,000 simulations and the real datasets with the R package *abc*. The most probable simulations and the median of these values were used as the effective recombination rate for each sampling.

Aim 2: Phylogroup analysis

Using the same dataset as Aim 1, I identified the five main phylogroups, A, B1, B2, D, E using the phylogenetic tree. I used the published datasets of (Beghain, et al., 2018: Sims & Kim., 2011: Diamant et al., 2004) to classify the strains shared with my dataset and to infer the phylogroups in the tree. I then selected 15 genomes from each phylogroup (three times) to conduct the unbiased samplings and I selected 15 genomes from each phylogroup and one genome from another phylogroup to generate biased samplings. I then used the same ABC pipeline to estimate recombination rates for each sampling independently.

Aim 3: Analysis of gene content

The goal of this analysis was to estimate the average number of orthologous genes shared by the genomes selected in each sampling. First, the 400 genomes of *E. coli* were compared against each other using *Blastn*. For each genome pair, orthologous genes were defined as the best reciprocal hits (gene A in genome 1 is most similar to gene B in genome 2 and vice versa) with a nucleotide sequence identity threshold of 95%. For each sampling the number of shared genes was reported with the mean, median, min and max values.

CHAPTER IV: DISCUSSION

Recombination rate is a key parameter in evolution and as such, various methods and approaches have been generated to estimate *r/m* in bacteria. However, estimating accurate recombination rates in bacteria has proven methodologically challenging and estimates of the same species have been largely inconsistent across studies. Here we conducted various samplings to test the robustness of a new approach based on an ABC framework and we further tested how population structure and bacterial gene content may impact or relate to recombination rates.

Our recombination rate estimates of *E. coli* varied from *r/m* = 1.34 to *r/m* = 3.76 across samplings and the average estimates across all the conducted samplings are yielding an estimate of *r/m* = 2.54 for *E. coli* as a species. Although previous studies have yielded highly variable estimates of recombination rates for *E. coli*, most recent studies have inferred a recombination rate around *r/m* = 1. The fact that this estimate is substantially lower than the estimates generated by our ABC approach is not surprising. Indeed, other methods developed to infer recombination rates are unlikely to catch all recombination events since these methods are theoretically incapable of inferring events that don't leave a direct signal of recombination. For instance, some alleles may be exchanged by recombination without being homoplasic (i.e. without being incongruent with the overall phylogeny of the species) and those cannot be inferred as recombinant by these methods. In contrast, because our ABC framework is based on simulations, such events can be accounted for as recombination events in our approach.

Our results revealed that conducting random samplings across 400 genomes of *E. coli* yielded rather consistent recombination rate estimates. We found that biased sampling strategies, where a more distant genome was introduced, did substantially impact recombination rate

estimates. This result shows that recombination rate does not systematically decrease when more divergent genomes are added to a sample. This result has further implications for the sampling strategies, which appear to be robust to biased samplings. Significant variations to our *r/m* estimates were observed across phylogroups, suggesting that population structure may shape recombination rates. However, this result should be contrasted by the fact that some of the most extreme variations in recombination rates were observed within the same phylogroup rather than between phylogroups.

Strains from the same phylogroup often display different phenotypes and frequently occupy different niches (Touchon et al., 2020). Strain phenotypes are frequently dictated by the presence of accessory genes that tend to be specific to groups of strains living in the same environment. Because these genes tend to be frequently exchanged via horizontal gene transfer, their distribution is usually not limited to a specific phylogroup. I therefore expected that strains sharing more similar gene content would present higher recombination rates. Results did not reveal a significant correlation between gene content and *r/m*. This pattern could be due to the fact that homologous recombination may not be directly correlated to the frequency of horizontal gene transfers (HGT events do not necessarily rely on homologous recombination). Alternatively, the assumption that genes sharing more similar gene content are more likely to present similar phenotypes and live in the same environment may be an over-simplification. The gain and loss of a single or several accessory gene(s) can be responsible for drastic phenotypic and ecological modifications in bacteria (Iranzo et al., 2019). Conversely, many accessory genes in bacteria are attributable to mobile elements (Bobay MBE., 2013), and these elements are not always associated with a clear phenotype or environmental specialization. Therefore, the link between ecological lifestyle and recombination rate would be better investigated by sampling

and sequencing strains from clearly defined environments. Unfortunately, characterizing basic attributes of the habitat and the niche of a given bacterial strain remains very challenging.

The fast accumulation of bacterial genomic datasets is generating new challenges for computational analyses. For example, *E. coli* currently has >20,000 complete genomes available for analyses, and more genomes are being sequenced every month. Based on this trend, it appears that it won't be possible to infer complex parameters such as recombination rates from entire genomic datasets of bacteria. As a result, subsampling strategies are, or will be, required to generate such parameters for all bacterial species in the near future. Although the accumulation of genomic data has helped to establish some links between population structure and recombination rate, other questions remain difficult to infer from genomic data alone. In particular, the link between bacterial ecology and recombination is a complex question to address due to the scarcity of high-quality data on bacterial ecology. The development of metagenomic approaches coupled with the accurate analysis of the sampled environment may soon provide novel insights into these questions.

REFERENCES

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., & Clermont, O. (2018). ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microbial genomics, 4(7).

Bobay, L.-M., Traverse, C. C., & Ochman, H. (2015). Impermanence of bacterial clones. *Proceedings of the National Academy of Sciences*, *112*(29), 8893–8900. https://doi.org/10.1073/pnas.1501724112

Bobay, LM. CoreSimul: a forward-in-time simulator of genome evolution for prokaryotes modeling homologous recombination. BMC Bioinformatics 21, 264 (2020). https://doi.org/10.1186/s12859-020-03619-x

Didelot, X., Méric, G., Falush, D., & Darling, A. E. (2012). Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics*, *13*(1), 256. https://doi.org/10.1186/1471-2164-13-256

Denamur, E., Clermont, O., Bonacorsi, S., & Gordon, D. (2021). The population genetics of pathogenic Escherichia coli. Nature reviews. Microbiology, 19(1), 37–54. https://doi.org/10.1038/s41579-020-0416-x

Diamant, E., Palti, Y., Gur-Arie, R., Cohen, H., Hallerman, E. M., & Kashi, Y. (2004). Phylogeny and strain typing of Escherichia coli, inferred from variation at mononucleotide repeat loci. Applied and environmental microbiology, 70(4), 2464–2473. https://doi.org/10.1128/AEM.70.4.2464-2473.2004

Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Computational Biology 11(2): e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Didelot, X., & Maiden, M. C. (2010). Impact of recombination on bacterial evolution. *Trends in microbiology*, *18*(7), 315–322. https://doi.org/10.1016/j.tim.2010.04.002

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (2002). Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution*, *19*(12), 2226–2238. https://doi.org/10.1093/oxfordjournals.molbev.a004046

Iranzo, J., Wolf, Y.I., Koonin, E.V. et al. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun 10, 5376 (2019). https://doi.org/10.1038/s41467-019-13429-2

Sims, G. E., & Kim, S.-H. (2011). Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences*, *108*(20), 8329–8334. https://doi.org/10.1073/pnas.1105168108

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, *8*(3), 207–217. https://doi.org/10.1038/nrmicro2298

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., … Denamur, E. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLoS Genetics*, *5*(1), e1000344. https://doi.org/10.1371/journal.pgen.1000344

Touchon, M., Perrin, A., de Sousa, J., Vangchhia, B., Burn, S., O'Brien, C. L., Denamur, E., Gordon, D., & Rocha, E. P. (2020). Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS genetics, 16(6), e1008866. https://doi.org/10.1371/journal.pgen.1008866