What do we gain when tolerating loss? The information bottleneck wrings out recombination

Apurva Narechania[1,2]*, Dean Bobo[1,3], Rob DeSalle[1], Barun Mathema[4], Barry Kreiswirth[5], and Paul J. Planet[1,6,7]*

[1]Institute for Comparative Genomics, American Museum of Natural History, New York, NY
[2]Section for Hologenomics, The Globe Institute, University of Copenhagen, Copenhagen, Denmark
[3]Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, USA
[4]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY
[5]Center for Discovery and Innovation, Hackensack Meridien Health, Nutley, NJ
[6] Division of Infectious Diseases, Children's Hospital of Philadelphia, Philadelphia, PA
[7] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

*Correspondence to: Apurva Narechania (anarechania@amnh.org) & Paul J. Planet (planetp@chop.edu)

Word Count
Abstract: 206
Body: 3342

Running Title: Measuring and detecting recombination with the information bottleneck

Abstract

Most microbes have the capacity to acquire genetic material from their environment. Recombination of foreign DNA yields genomes that are, at least in part, incongruent with the vertical history of their species. Dominant approaches for detecting these transfers are phylogenetic, requiring a painstaking series of analyses including alignment and phylogenetic tree reconstruction. These traditional pan-genomic methods do not scale. Here we propose an unsupervised, alignment-free and tree-free technique based on the sequential information bottleneck (SIB), an optimization procedure designed to extract some portion of relevant information from one random variable conditioned on another. In our case, this joint probability distribution tabulates occurrence counts of k-mers against their genomes of origin with the expectation that recombination will create a strong signal that unifies certain sets of co-occuring k-mers. We conceptualize the technique as a rate-distortion problem, measuring distortion in the relevance information as k-mers are compressed into clusters based on their co-occurrence in the source genomes. The result is fast, model-free, lossy compression of k-mers into groups that learns tracts of shared genome sequence differentiating recombined elements from the vertically inherited core. We show that the technique yields a new recombination measure based purely on information, divorced from any biases and limitations inherent to alignment and phylogeny.

Significance

The information bottleneck, a lossy compression technique borrowed from the information theoretic and Natural Langauge Processing literature, is well suited to detecting evolutionary

patterns in sets of co-occuring k-mers. Here we show that we can detect simulated and real recombination events while highlighting a core set of k-mers that comprise the vertically inherited portion of any set of genomes. Moreover, the compressibility of any given set of genomes offers a new way to compare the pangenomes of clades across the microbial tree of life. In our application, the bottleneck is informed by genome origin, our relevance variable, but the technique is general. The information bottleneck can be used for any biological contingency matrix where the goal is to learn groups from unstructured data.

Introduction

Microbial genomes have accumulated at an unprecedented rate[1,2]. Most work on their molecular evolution is grounded in sequence alignment and phylogenetic tree reconstruction, computationally expensive techniques ill-equipped to handle the volume of sequence encountered in a read streaming era[3]. The evolution of microbes is particularly challenging because recombined elements contribute signal unrelated to vertical descent[4]. Current techniques require alignment of reads across a reference genome[5,6], whole genome alignment[7], and/or phylogenomic methods based in orthology[8–10]. Each requires careful curation and deliberate sampling to limit data to reasonable scales. For larger, unbiased datasets that include as much natural variation as possible, these approaches are unsustainable. We need tools that can tolerate information loss without sacrificing knowledge of key evolutionary events.

Lossy compression, where an individual or algorithm makes decisions about which data are important (or relevant) from a large body of information, may offer a solution[11]. To do this

in a principled way, the relevance of a given dataset can be measured as information retained about some other correlated variable. For example, in unsupervised natural language processing (NLP) large corpora of texts are distilled to a few topics that reflect overall themes by comparing patterns of co-occurring words in the source texts. In topic modeling of this sort, the texts themselves are the relevance variable. The goal is to cluster the word distribution with respect to the documents from which they arise. This idea was first described by Tishby, Pereira and Bialek as the information bottleneck (IB)[12] . It was premised on rate distortion, Shannon's original theory of lossy compression which yoked signal distortion to the rate at which that signal can be encoded[13]. The IB's primary innovation was the use of a relevance variable to quantify this distortion and instill it with meaning. Topic modeling was one of this technique's first applications.

Topic modeling has become an important part of the NLP literature with a number of wider applications to unsupervised machine learning. The dominant technique in the field is Latent Dirchilet Allocation (LDA)[14], a probabilistic method, that like the IB, considers each document a mixture of topics. Some groups have applied this idea to whole genomes[15–17], and since the publication of STRUCTURE, LDA has become foundational in the genetics literature[18]. Despite LDA's popularity and success, a number of authors have shown that unbalanced sampling can lead to erroneous or missed population assignments[19]. LDA also makes a number of statistical assumptions including the assignment of hyperparameters and a Dirchilet prior[20]. In contrast, the IB is model free, unaffected by size sample bias.

Genomes are living documents that can be sliced into words of arbitrary size[21,22]. In a genomic context, where words are k-mers and documents are their genomes of origin we

hypothesize that IB derived topics could represent co-occurring groups of k-mers that highlight shared ancestry. K-mers might be arranged in co-linear blocks, distributed across the genome as fragments inherited simultaneously, or may be common to all genomes, providing a simple, operational definition of a genomic core. In all cases, lossy compression of k-mers into topics is guided by how often they co-occur with respect to their genomes of origin. The process requires only two parameters: the k-mer size and the number of clusters expected. In the NLP topic modeling analogy, the core genome of a species could be considered the set of concepts common to every document in a library, while recombined regions are like themes or ideas restricted only to certain shelves.

Here we apply the IB to tens of thousands of microbial genomes. Remarkably, our approach identifies recombination tracts without making any attempt to model evolution, annotate genes, build alignments, or reconstruct trees. Further, squeezing k-mers through the bottleneck yields a kind of distortion we repurpose to measure horizontal signal.

Theory and Implementation

Consider a set of genomes (Y) chopped into k-mers (X). The information bottleneck is designed to compress X under the constraint of Y[12]. The IB extends Shannon's theory of rate distortion by guiding it with an additional, orienting variable. In our case, this variable is genomes of origin. This concept of a relevance variable assigns value to the resulting distortion. The choice of Y defines relevant features in the signal. If X and Y are tabulated as a joint probability distribution, the information that X provides about Y is squeezed through a simpler representation to learn a set of T groups (Figure 1). In a rate distortion sense, T can be considered the channel capacity.

For the technique to work, the two variables in our joint distribution *p(x,y)* must be non-independent, or more precisely, must have positive mutual information, *I(X,Y)*:

$$I(X,Y) = \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{p(y)}$$

T is now a meaningful compression of the data, maximizing the mutual information between the clusters and genomes, I(T;Y), while minimizing the mutual information between the k-mers and the clusters, I(T;X). The IB is a classic optimization problem (Figure 1).

Compressing X into T is equivalent to minimizing the following Lagranian with respect to the source genomes, Y:

$$\mathcal{L}[p(t|x)] = I(X;T) - \beta I(X;Y)$$

This formulation balances the compactness of X, with the erosion of information about Y. β is a multiplier that slides through the optimization landscape. As beta approaches 0, k-mers are clumped into fewer and fewer clusters, emphasizing compression. As beta approaches infinity, every k-mer is its own cluster, preserving all relevant information. Of course, collapsing all k-mers into one cluster is overly reductive, and assigning each k-mer to its own cluster is meaningless. The IB negotiates these two extremes. In NLP, the result is a set of clusters that coalesce into topics over a body of literature[23]. In genomics, the process identifies co-occurring and/or spatially co-located k-mers with distinct biological and/or evolutionary meaning.

Minimizing the Lagranian above has an exact, optimal solution. The most surprising outcome of this solution is that the relative entropy, or Kullback Liebler divergence[24], emerges as the distortion measure for the information bottleneck. The relative entropy is a fundamental quantity in information theory, and in our IB context, it measures the distortion of compressed k-mers:

$$D_{KL} = \sum_{y} p(y|x) \log \frac{P(y|x)}{P(y|t)}$$

We set the number of clusters and implement a sequential clustering procedure. From an initial random distribution of all k-mers across our clusters we draw one k-mer out, and represent it as a singleton. Using greedy optimization, we merge this singleton into one of the existing bulk clusters. Slonim's sequential information bottleneck (SIB)[25] employs the Jensen-Shannon divergence[26] in the cost of merging a k-mer, x, into a cluster, t:

$$d(x,t) = (p(x) + p(t)) * D_{JS}(p(y|x), p(y|t))$$

A k-mer will join a new cluster only if its new address reduces the total distortion. Otherwise, it remains in its existing cluster. With respect to our initial random conditions, this algorithm is guaranteed to converge to a local optimum. We mitigate the risk of getting trapped in local optima by testing several random initializations, and we reduce the overall computation by implementing a k-mer sketch[27,28].

157 Once the clusters stabilize, we quantify the information captured by calculating the
158 normalized mutual information, *NMI = I(T;X) / I(X;Y).* Trivially, the NMI is one when each k-mer
159 occupies its own cluster and is zero when all k-mers are mixed into one cluster. The curve
160 traced between T = 1 and T = *x* is called the relevance compression curve[29]. The choice of T is
161 analogous to the optimization of β in the Lagranian above. Both modulate the Shannon channel
162 capacity, but we vary T for the deterministic case involving hard clustering. As with β, the shape
163 of this curve describes the compressibility of the data.
164 The most important aspect of the SIB, and the reason we chose it for this work, is that
165 its distortion, as measured by NMI, is a proxy for horizontal signal.
166
167 **Results and Discussion**
168 *The bottleneck in test: simulated recombination events*
169 We simulated 50 recombination events across a set of 100 1 Mbase genomes in SimBac[30].
170 These events appear in black on the innermost track if the circos plot in Figure 2. We sketched
171 19-mers from all 100 genomes and applied the information bottleneck, modeling 50 clusters.
172 The outermost track is a frequency plot of k-mers from the core cluster aligned to arbitrary
173 genome from the set. The core emerges as a dense block of shared genome sequence
174 interrupted by our simulated recombination events. K-mers that would otherwise occupy these
175 gaps are sorted into other clusters. Because the compression is driven by genome of origin, if a
176 single ancestor sustains multiple transfer events, all k-mers from these events merge into a
177 single cluster shared by the same subset of descendants. Plots of the bottleneck-defined core

178 function almost as a photographic negative, highlighting the blank spaces as regions scrambled
179 by horizontal signal.

180 Meaning emerges from this spatial organization of k-mers and corresponds to regions of
181 shared ancestry. In the examples shown, the pattern of events is evident by eye, but with
182 increasing input genomes and increased modeling (more clusters), automated inspection is key.
183 We use a signal processing method based in change point detection to programmatically detect
184 changes in k-mer frequency. We specifically employ the PELT algorithm[31] to model probabilities
185 of change along the genome backbone. As shown in the track of red hashes, changepoint
186 detection largely delimits the boundaries of our simulated recombination events. PELT's
187 detection sensitivity is modulated by a penalty parameter that can vastly change the analysis. A
188 sweep through values for this parameter is shown in Supplementary Figure 1. While this
189 parameter can alter the changepoints derived from our primary frequency data, the frequency
190 data itself almost always shows gaps in the core's continuity at recombination events. In the
191 example shown in Figure 2, 48 of the 50 recombination events have k-mer counts significantly
192 lower than the rest of the core's background (Wilcoxon, $p < 0.05$).

193 We compare our approach to ClonalFrameML[6], the leading recombination detection
194 program in microbial genomics. Unlike the IB, ClonalFrameML requires a reference, alignment
195 of all remaining 99 genomes to that reference, tree reconstruction from the resulting character
196 matrix, and recombination detection using a likelihood-based sliding window approach.
197 ClonalFrameML calls are shown in the track of light blue hashes. The program successfully
198 identifies regions of recombination, even picking up smaller regions that IB fails to detect at the

199 shown change point sensitivity. But the program seems to subdivide events into smaller ones,
200 sometimes up to 20 smaller events in one large expanse.

201 Lossy compression drops data. Forcing all the signal in our k-mer occurrence matrix
202 through a narrow 50 cluster channel garbles portions of the original message. This is equivalent
203 to channel noise. It recalls Shannon's original formulation of the rate distortion problem[13]. In
204 this case, modeling more clusters increases the rate of transmission, and reduces the distortion
205 of the message received. With respect to the information bottleneck, we quantify this effect
206 using a relevance-compression curve[32]. In Supplemental Figure 2, we show data from 100
207 simulated sets each of 4, 10, 100 and 1000 one megabase genomes. Regardless of the number
208 of genomes in each experiment, the NMI increases with increased modeling (one example for
209 each shown in Panel A). As expected, this increase is rapid and stark in the smaller datasets;
210 slow and somewhat halting in the larger ones. Complex scenarios with more genomes and
211 more evolutionary paths require a broader channel (more clusters) to transmit signal. The
212 theoretical extremes for the relevance-compression curves are intuitive. At the origin, all the
213 relevant information is destroyed. With too many clusters, we retain more information than we
214 can interpret. The curve traced between these two extremes is a fingerprint of the data. A
215 convex shape suggests natural structure easily modeled with just a few clusters. Data that
216 resists compression flattens this curve. Theoretically, the space above the curve is unachievable
217 by any process, forming an upper bound. The relevance-compression curve therefore defines
218 absolute limits on the quantity and quality of evolutionary information communicated as we
219 sweep through a dilating channel. In Supplemental Figures 2B and 2C, we show how
220 information loss in more complex compression problems is driven by the narrowing gap

221 between k-mer frequency outside versus inside the simulated recombination events. With four
222 genomes, discrimination is easy. With 1000, noise limits our resolution.

223 The relevance-compression curve's responsiveness to evolutionary complexity opens
224 the door to reconceptualizing recombination. Given a sequence set, most researchers would
225 calculate R/theta[33], the ratio of the recombination rate over the mutation rate, to quantify a
226 species' tendency to recombine. But R/theta's calculation is grueling, requiring alignment
227 (either reference based or across whole genomes), tree reconstruction, and algorithms to
228 isolate the clonal frame[5,6]. The information bottleneck clearly pinpoints recombination events
229 in space without resorting to any of these traditional procedures.

230 NMI is a corollary to this compression, but can NMI itself be an informative
231 recombination metric? In Figure 3 we show that for sets of 100 simulated genomes, NMI
232 decreases with increased recombination rate. We also show that even very high background
233 mutation rates have a negligible effect on NMI for all sets except those subject to vanishingly
234 low recombination. Recombination seems to increase distortion indicating that in our context
235 evolutionary complexity is tied to information loss. At high rates of recombination, every base
236 of a 1 Mb genome is likely scrambled. Under such flux, some sites recombine several times,
237 eroding the core genome itself. This erosion is reflected in the steep decline in NMI with
238 increasing recombination rate. The NMI's responsiveness to both recombination rates and
239 mutation rates makes it an attractive alternative to arduous – and in some cases – impossible
240 R/theta calculations. As expected, we capture more information if we allow more clusters, but
241 over-modeling in simpler evolutionary scenarios erodes the information captured
242 (Supplemental Figure 3).

243

244 *The bottleneck in action: real world recombination*

245 We used genomes from ST239 *Staphylococcus aureus* to illustrate that our method can
246 corroborate known, large scale recombination events found in nature. The ST239 strain is a
247 hybrid: a segment from a CC30 (clonal complex 30) donor replaced nearly 20% of the
248 homologous region in a CC8 strain[34]. The evolutionary histories of genes across these segments
249 are incongruent. Previous studies compared the histories of thousands of genes to reach this
250 conclusion[35]. Here, we attempt to localize this same phenomenon using the co-occurrence
251 pattern of k-mers alone. We chose 10 genomes (GCA_000146385.1, GCA_000012045.1,
252 GCA_000011505.1, GCA_000011265.1, GCA_000013425.1, GCA_000204665.1,
253 GCA_000159535.2, GCA_000027045.1, GCA_000017085.1, and SA21300), sampled from both
254 the donor clade (CC30), the recipient clade (CC8), and genomes outside of the evolutionary
255 event.

256 Figure 4 highlights two of these 10 genomes, and three of the 60 clusters we modeled
257 for this analysis. Both *S. aureus* COL (CC8) and *S. aureus* T0131 (ST239) share a large, congruent
258 core. The gap in this core characterizes the dimensions of the recombination event, whose k-
259 mers are split into two other clusters, shown here as the second and third tracks. The
260 bottleneck learns the structural evolution of the clade as tracts of co-occurring sequence. The
261 clusters themselves comprise an evolutionary model for the structural event and highlight the
262 core. Genome origin guides k-mer compression by forming the basis of the distortion measure.
263 We lose information in a controlled and quantitative way, and short circuit the intensive
264 phylogenomic analyses required[35] to identify the tract.

265 Our approach introduces a new type of comparative genomics based on compression.
266 Figure 5 shows relevance compression curves for the ST239 genomes alongside 100 genomes of
267 *Mycobacterium tuberculosis* and *Helicobacter pylori*. The convex shape of the *M tuberculosis*
268 curve reflects its clonality. Data from a more open pangenome that resists compression flattens
269 the curve. Highly recombinogenic species like *H. pylori* suffer this sort of steep information loss.
270 In Supplementary Figure 4 we show relevance compression curves for the 28 most well
271 studied microbial genomes in RefSeq. In all cases, increased modeling captures more
272 information, but at varying rates. We also demonstrate that we can reach a steady-state NMI
273 with a small number of randomly selected genomes regardless of clade (Supplementary Figure
274 5). We interpret the shape of the relevance compression curve as a proxy for evolutionary
275 strategy and show that 50 randomly selected genomes can reflect the nature of the species.
276
277 *The bottleneck as a new measure of recombination*
278 We have shown that in simulation, given a set number of clusters, increased recombination
279 results in decreased NMI. To test whether this pattern holds for real data, we analyzed the top
280 13 most sequenced bacterial species in RefSeq with at least 100 complete, contiguous genome
281 assemblies. For each of these 13 species, we randomly selected one complete genome as a
282 reference, and 100 random genomes as queries. We calculated R/theta for these selections and
283 repeated the experiment 100 times for all 13 species. Figure 6 shows R/theta (A) and r/m (B)
284 plotted against the corresponding five cluster NMI. The data indicate that NMI is an inverse
285 mirror of R/theta: higher levels of recombination tend to lower compressibility. NMI clearly
286 offers a conceptual alternative to R/theta calculable in a fraction of the time (Figure 7). A simple

287 distortion metric produced as sequence compresses through an information bottleneck
288 provides a ready alternative to currently accepted approaches.
289 While time savings is an important feature of the IB, the wide range in R/theta values for
290 *H. pylori* is emblematic of the primary weakness of most dominant recombination methods:
291 reference bias. In Supplementary Figure 6, we show the range of R/theta values calculated for
292 10 randomly chosen genomes of *S aureus*, *M tuberculosis,* and *H pylori* against all possible
293 contiguous, complete references known for these genomes. While *S aureus* and *M tuberculosis*
294 show relatively tight ranges, *H. pylori* varies over one full unit depending on the reference
295 chosen. Even vetted resources like RefSeq can be misleading. While compiling the data shown
296 in Figure 6, we observed that long-branch effects from a single misclassified sequence can ruin
297 an entire replicate. Omitting these misclassified sequences with a MASH distance-based screen
298 clarified the relationship between NMI and R/theta (Supplemental Figure 7). We removed
299 replicates that contained outliers in a set of all pairwise MASH distances and settled on 4.5
300 standard deviations from the mean to remove the heterogenous set of points near the origin
301 while retaining the bulk of the replicates for each species. This uncertainty in RefSeq coupled
302 with the instability inherent to reference-based analyses argues for reference free approaches
303 like the IB. The information bottleneck is not saddled with the references required of many
304 bioinformatic techniques including most of the dominant ones employed to detect and
305 measure recombination.
306 The capacity of the bottleneck – in the Shannon sense – can be used to modulate
307 detection. In the case of ST239 *S. aureus*, asking for just two clusters – a very narrow channel –
308 captures more than 40% of the relevant information (Figure 5). Remarkably, these two clusters

separate the core from the recombined region. Even the simplest model learns the most prominent evolutionary process. Further along the curve, fifteen clusters capture 80% of the information. To put this in perspective, the IB condenses the outsized, uninterpretable dimensions of our original data – 40 million k-mers – into just fifteen buckets while sacrificing only 20% of the information originally present in the relevance variable. Because the technique is inherently lossy, the SIB will never recover all the information originally encoded but aims to extract the most salient bits. Beyond fifteen clusters, the ST239 curve elbows, and modeling gains are slight. In this way, the relevance-compression curve defines the optimal number of clusters,[29,32] and it is the NMI at this asymptote that best reflects the traditional R/theta.

In the light of evolution, the bend in our compression curves may have a deeper meaning. For *S. aureus*, fifteen clusters are enough to adequately capture the complete set of k-mer aggregation patterns across our chosen genomes. This point of diminishing returns may signify an opportunity for interpretive balance: not so many clusters that we drown dominant evolutionary events in noise, and not so few that we neglect to model subtle k-mer co-occurrence patterns. This particular use of the elbow method in our information theoretic context puts a crude limit on the dominant evolutionary paths taken by the elements that comprise our pangenome. And the NMI emitted at this asymptote offers a new way to describe the lability of the species.

Software

NECK (https://github.com/narechan/neck)

332 **Figure Legends**

333 Figure 1. The information bottleneck. In the information bottleneck a distribution, *X*, is
334 compressed into *T* while retaining as much information as possible about a correlated relevance
335 variable, *Y*. The joint distribution, *p(x,y)*, has positive mutual information and the goal of the
336 information bottleneck is to capture as much of that information as possible at interpretive
337 scale. The technique is a classic optimization problem wherein the mutual information between
338 *T* and *X* is minimized, while the mutual information between *T* and *Y* is maximized. At
339 optimality, *T* is presumed to be a lossy but adequate model of *X*.

340

341 Figure 2. Simulations. We simulated 50 recombination events across 100 one megabase
342 genomes. The recombination events are shown as black blocks on the innermost track. IB based
343 changepoint calls for breakpoints are shown in red, and ClonalFrameML based intervals are
344 shown in blue.

345

346 Figure 3. Simulation across various rates of recombination and background mutation. As the
347 recombination rate increases, the NMI decreases. The effect of background mutation rate is
348 swamped by recombination, becoming a factor only when recombination rates are very low.

349

350 Figure 4. Modelling ST239's hybridization event. We selected 10 *S. aureus* genomes to track the
351 ST239 hybridization event with the information bottleneck. We chose COL to represent the
352 CC30 donor strain, and T0131 the CC8 acceptor. Of the 60 clusters we calculated, we show the
353 three that capture the hybridization event. The innermost track is a frequency plot of k-mers

354 that define the core. The second and third tracks are flipsides of the HGT event that created
355 ST239.
356
357 Figure 5. Relevance compression curves across three species. In an information bottleneck
358 experiment, the relevance compression curve traces the increase in normalized mutual
359 information with the number of clusters modeled. The curves quantify the amount of
360 information lost at a given modeling threshold. We show how this type of relationship can
361 function as a marker for evolutionary strategy by calculating curves for three very different
362 groups of microbes: *M. tuberculosis*, a species thought to demonstrate little if any HGT; *S.*
363 *aureus*, a species considered largely clonal with occasional HGT; and H. pylori, a species known
364 to employ HGT as an engine for diversity (A).
365
366 Figure 6. NMI tracks R/theta. We show that for 13 key species in Refseq , more recombinogenic
367 species (as described by r/theta in panel A and r/m in panel B) are less easily compressed.
368
369 Figure 7. Computational savings. We calculate NMI, our proxy for genomic lability, with far less
370 computational resource than required by traditional methods.
371
372 **References**
373 1. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509
374 (2021).

375 2. Kyrpides, N. C. *et al.* Genomic Encyclopedia of Type Strains, Phase I: The one thousand
376 microbial genomes (KMG-I) project. *Stand. Genomic Sci.* **9**, 1278–1284 (2014).
377 3. Erlich, Y. A vision for ubiquitous sequencing. *Genome Res.* **25**, 1411–1416 (2015).
378 4. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial
379 speciation. *Science* **315**, 476–480 (2007).
380 5. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial
381 whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15–e15 (2015).
382 6. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole
383 Bacterial Genomes. *PLOS Comput. Biol.* **11**, e1004041 (2015).
384 7. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple Alignment of
385 Conserved Genomic Sequence With Rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
386 8. Chiu, J. C. *et al.* OrthologID: automation of genome-scale ortholog identification within a
387 parsimony framework. *Bioinformatics* **22**, 699–707 (2006).
388 9. Zhao, Y. *et al.* PGAP: pan-genomes analysis pipeline. *Bioinformatics* **28**, 416–418 (2012).
389 10. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*
390 **31**, 3691–3693 (2015).
391 11. Marzen, S. E. & DeDeo, S. The evolution of lossy compression. *J. R. Soc. Interface* **14**,
392 20170166 (2017).
393 12. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method.
394 *arXiv:physics/0004057* (2000).
395 13. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423
396 (1948).

397 14. Blei, D. M. Latent Dirichlet Allocation.

398 15. La Rosa, M., Fiannaca, A., Rizzo, R. & Urso, A. Probabilistic topic modeling for the
399 analysis and classification of genomic sequences. *BMC Bioinformatics* **16**, S2 (2015).

400 16. Chen, X., Hu, X., Shen, X. & Rosen, G. Probabilistic topic modeling for genomic data
401 interpretation. in *2010 IEEE International Conference on Bioinformatics and Biomedicine*
402 *(BIBM)* 149–152 (2010). doi:10.1109/BIBM.2010.5706554.

403 17. Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. An overview of topic modeling and its
404 current applications in bioinformatics. *SpringerPlus* **5**, 1608 (2016).

405 18. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using
406 Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).

407 19. Wang, J. The computer program structure for assigning individuals to populations: easy
408 to use but easier to misuse. *Mol. Ecol. Resour.* **17**, 981–990 (2017).

409 20. Wallach, H. M., Mimno, D. & McCallum, A. Rethinking LDA: why priors matter. in
410 *Proceedings of the 22nd International Conference on Neural Information Processing Systems*
411 1973–1981 (Curran Associates Inc., 2009).

412 21. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with
413 feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci.* **106**, 2677–
414 2682 (2009).

415 22. Cong, Y., Chan, Y. & Ragan, M. A. A novel alignment-free method for detection of lateral
416 genetic transfer based on TF-IDF. *Sci. Rep.* **6**, 30308 (2016).

23. Pereira, F., Tishby, N. & Lee, L. Distributional Clustering of English Words. in *31st Annual Meeting of the Association for Computational Linguistics* 183–190 (Association for Computational Linguistics, 1993). doi:10.3115/981574.981598.

24. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).

25. Slonim, N., Friedman, N. & Tishby, N. Unsupervised Document Classification Using Sequential Information Maximization. in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 129–136 (ACM, 2002). doi:10.1145/564376.564401.

26. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).

27. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

28. Narechania, A. *et al. Hill numbers at the edge of a pandemic: rapid SARS-COV2 surveillance using clinical, pooled, or wastewater sequence as a sensor for population change*. http://medrxiv.org/lookup/doi/10.1101/2022.06.23.22276807 (2022) doi:10.1101/2022.06.23.22276807.

29. Still, S. & Bialek, W. How Many Clusters? An Information-Theoretic Perspective. *Neural Comput* **16**, 2483–2506 (2004).

30. Brown, T., Didelot, X., Wilson, D. J. & Maio, N. D. SimBac: simulation of whole bacterial genomes with homologous recombination. *Microb. Genomics* **2**, (2016).

438 31. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal detection of changepoints with a linear
439 computational cost. *J. Am. Stat. Assoc.* **107**, 1590–1598 (2012).
440 32. Slonim, N. The Information Bottleneck: Theory and Applications. *Dr. Diss. Hebr. Univ.*
441 *Jerus. Isr. 2003* 157.
442 33. Guttman, D. S. & Dykhuizen, D. E. Clonal Divergence in Escherichia coli as a Result of
443 Recombination, Not Mutation. *Science* **266**, 1380–1383 (1994).
444 34. Robinson, D. A. & Enright, M. C. Evolution of Staphylococcus aureus by Large
445 Chromosomal Replacements. *J. Bacteriol.* **186**, 1060–1064 (2004).
446 35. Narechania, A. *et al.* Clusterflock: a flocking algorithm for isolating congruent
447 phylogenomic datasets. *GigaScience* **5**, (2016).
448
449 **Supplemental Methods**
450 *Information Bottleneck with neck.*
451 Neck is the program we use to execute the information bottleneck in our evolutionary context.
452 Neck accepts two key parameters from the user: k-mer size and the number of clusters to
453 model. The program performs a k-mer sketch of its query genomes by sliding a length *k* window
454 across the sequence. No reference is required. We use a 64-bit hash function on canonical k-
455 mers to generate a bottom-k sketch at the desired depth. The learning process inherent to the
456 bottleneck procedure iterates over these k-mers, minimizing cluster distortion on each
457 successive loop through the sketched strings. This learning process is called the sequential
458 information bottleneck. Neck will support both assembled and unassembled sequence, and in
459 both cases, outputs clusters of unlocalized strings. To localize strings on assembled sequence,

460 we provide an auxiliary program called neck_paint (see below) A typical run of neck looks like

461 this:

462 *neck -i genomes/ -k 19 -m 1 -c 50 -n 4 -s 10 -o out_neck_c50/*

463 The query genomes are found in the genomes directory *(-i genomes/*), one genome per file. In

464 this run 19-mers are hashed at a rate of 10% *(-s 10*). In a bottom-k context this means that the

465 lowest tenth of the hash values and their corresponding k-mers are sketched into the set of

466 strings used in the bottleneck procedure. For this analysis, we model 50 clusters *(-c 50*) and

467 compress the clusters over 4 iterations of the optimization (-*n 4*). Output files, including the

468 clusters of k-mers found and the NMI value are redirected to out_neck_c50 (-*o out_neck_c50/*).

469

470 *Visualization with neck_paint*

471 Though not strictly required, assembled genomes give a user the opportunity to contextualize

472 clusters of k-mers generated in neck. The script neck_paint aligns k-mers from the clusters to

473 query genomes using bowtie2 and calculates the depth of the resulting alignments with

474 samtools. Using these depth statistics, neck_paint will calculate changepoints between regions

475 of high and low k-mer frequency. On the core cluster, these regions typically correspond to the

476 core genome and the recombination events. Recombination sites almost always exhibit

477 depleted alignment in the core. De novo recombination intervals independent of any other

478 annotations are inferred from these changepoints. If recombination regions are known,

479 neck_paint can calculate Wilcoxon statistics on the alignment frequency differences between

480 the known regions and the rest of the genome. We use Circos to visualize every cluster/genome

481 pair, their associated changepoints, and any known recombination events. A typical run of
482 neck_paint looks like this:
483 *neck_paint -i genomes/ -c clusters/ -l 200 -r changepoint.R -p 128 -m 3 -v 1000 -x*
484 *log.external -o out_neck_c50_paint/*
485 In this case, neck_paint will perform alignments, perform changepoint analysis, and generate a
486 circos visualization for every genome *(-i genomes/*) cluster (-*c clusters/)* pair. The script
487 changepoint.R (-*r changepoint.R*) delimits the changepoints using the PELT algorithm given the
488 penalty parameter (-*v 1000*). The higher the penalty, the stronger the signal needed to mark a
489 changepoint. A low penalty parameter therefore often suffers from false positives while a high
490 penalty parameter might lack sensitivity. For the circos visualization, alignment frequency is
491 binned into 200 basepair intervals (-*l 200*) and shown as a plot with a maximum of 3 units (-*m*
492 *3*). If available, recombination events can also be shown as a track (-*x log.external*) and the user
493 can generate Wilcoxon statistics for non-parametric difference between frequencies within
494 these intervals and frequencies outside of them. Because there can be many cluster/genome
495 pairs, the process can be parallelized (-*p 128*) and the output collected into a single directory (-*o*
496 *out_neck_c50_paint*).
497
498 **Supplemental Figures**
499 Supplemental Figure 1. Parameter sweep in changepoint detection. Calling changepoints in the
500 mapped k-mer frequency data is sensitive to PELT's penalty parameter. If set too low, the
501 analysis is mired in false postives. If set too high, we fail to call true recombination events.
502

503 Supplemental Figure 2. Compression efficiency with large datasets. We simulated
504 recombination in 100 sets of 4, 10, 100, and 1000 one megabase genomes. In A we show the
505 rate of compression efficiency with increased modeling in the example relevance-compression
506 curves. More complex datasets require more clusters to capture comparable amounts of
507 information. In B we show that compression noise in more complex datasets is due to the
508 narrowing gap between core cluster k-mers captured outside versus inside simulated
509 recombination events. In C, we show examples of k-mers from the core IB cluster mapped to
510 one of the genomes. Simulated events are displayed in black, and changepoints are shown as
511 red tick marks delimiting significant changes in the frequency of the k-mer distribution.
512
513 Supplemental Figure 2. Simulation across various rates of recombination and levels of
514 modeling. NMI decreases with increasing recombination at approximately the same rate across
515 different levels of modeling. Modeling more clusters, captures more information. For lower
516 levels of recombination, over-modeling may cause a decrease in useful information captured.
517
518 Supplemental Figure 3. Relevance compression curves across 28 species. We show how
519 increasing the number of clusters modeled, increases the amount of information captured
520 across the 28 most sequenced microbial species in GenBank.
521
522 Supplemental Figure 4. Species-level optimal compression. Regardless of species, the NMI
523 stabilizes at about 50 randomly selected genomes. Increased genomic information does not
524 seem to add greater complexity beyond that point.

525

526 Supplemental Figure 5. Reference bias in traditional methods. We show that for *H. pylori*,

527 R/theta varies over a large range depending on the reference chosen.

528

529 Supplemental Figure 6. The effect of increased stringency in outlier removal. RefSeq contains

530 misannotated sequences that cause severe long branch effects. We screened any replicates

531 with these sequences using an all against all MASH distance analysis. We show the change in

532 the replicates kept as the stringency for removing experiments is increased from three standard

533 deviations of the mean to 4.5 standard deviations. In the panel with all replicates, we show that

534 the points removed cluster close to the origin.

535

536 Supplemental Figure 7. MASH screen of long branches. Misclassified sequences in RefSeq can

537 cause estimates for r/m to vary widely. We screened replicates that contained outliers in the

538 set of all pairwise MASH distances. Shown are variants of Figure 6B with successively more

539 replicates removed as the threshold for removal is increased. At 4.5 standard deviations from

540 the mean, the cluster of heterogenous points near the origin disappears while retaining most of

541 the replicates for each species.

542

FIG 1

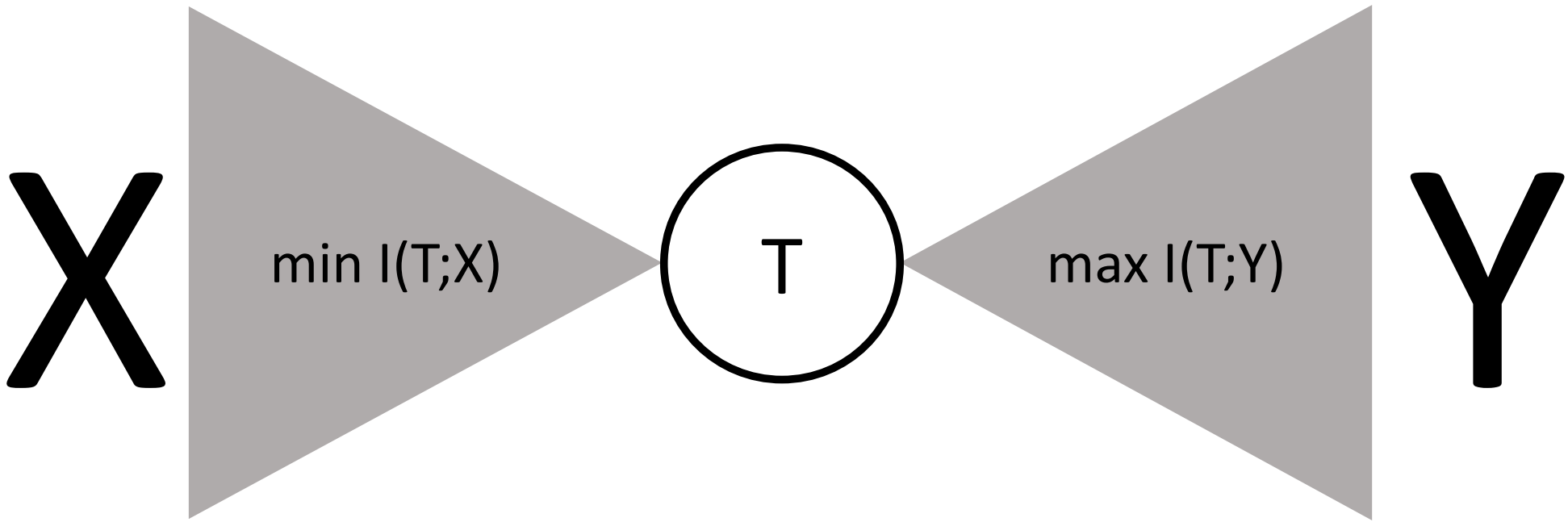
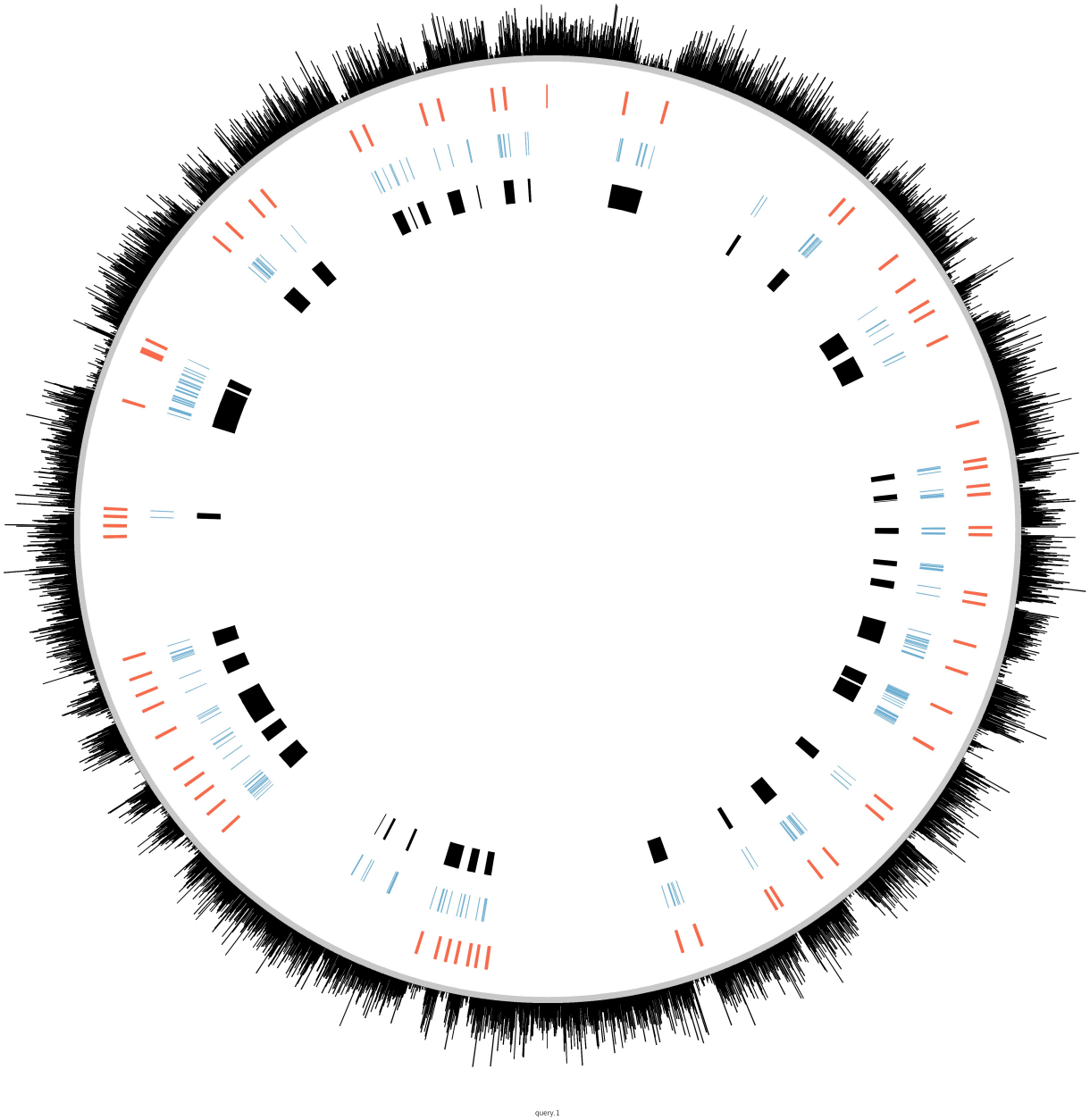


Figure 1

FIG 2



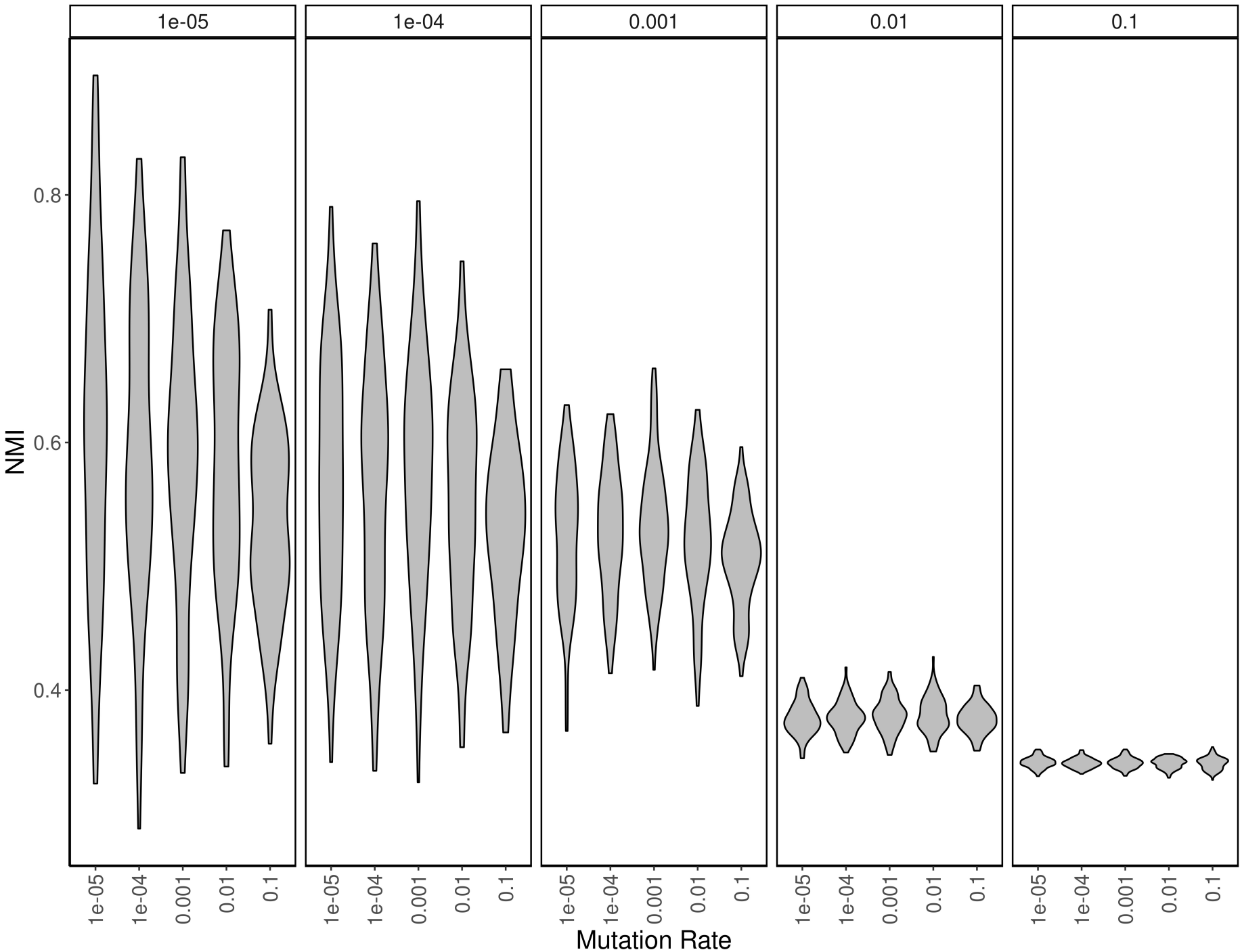
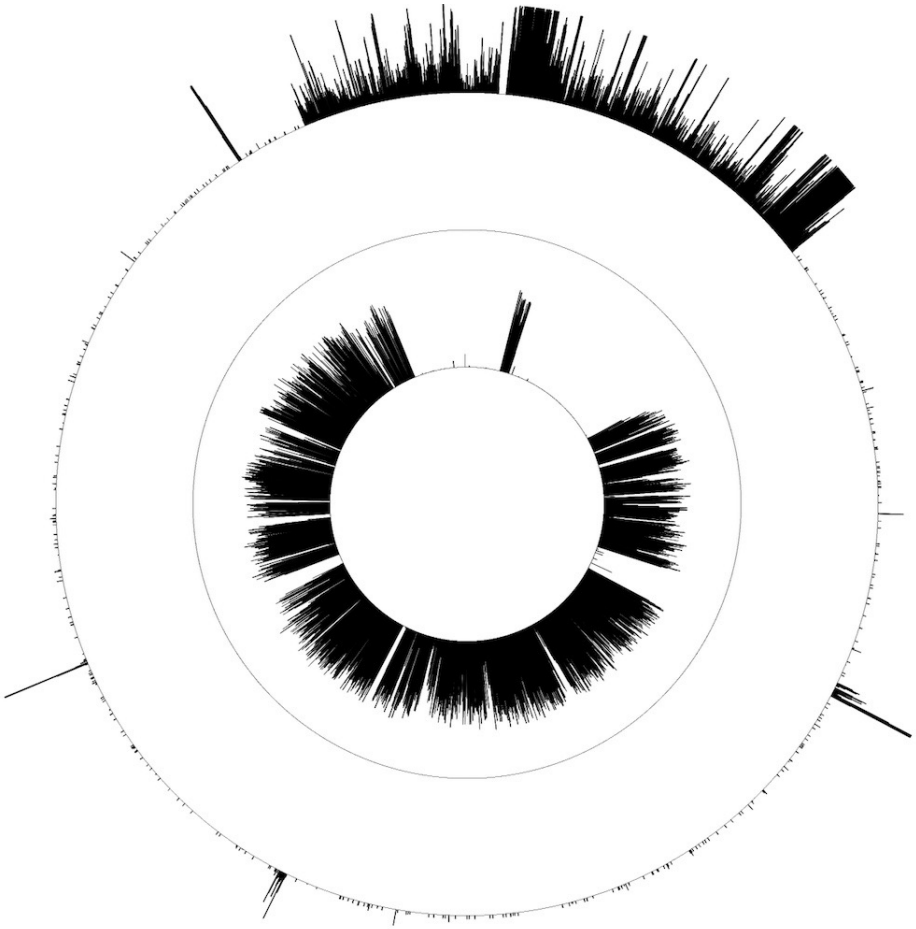
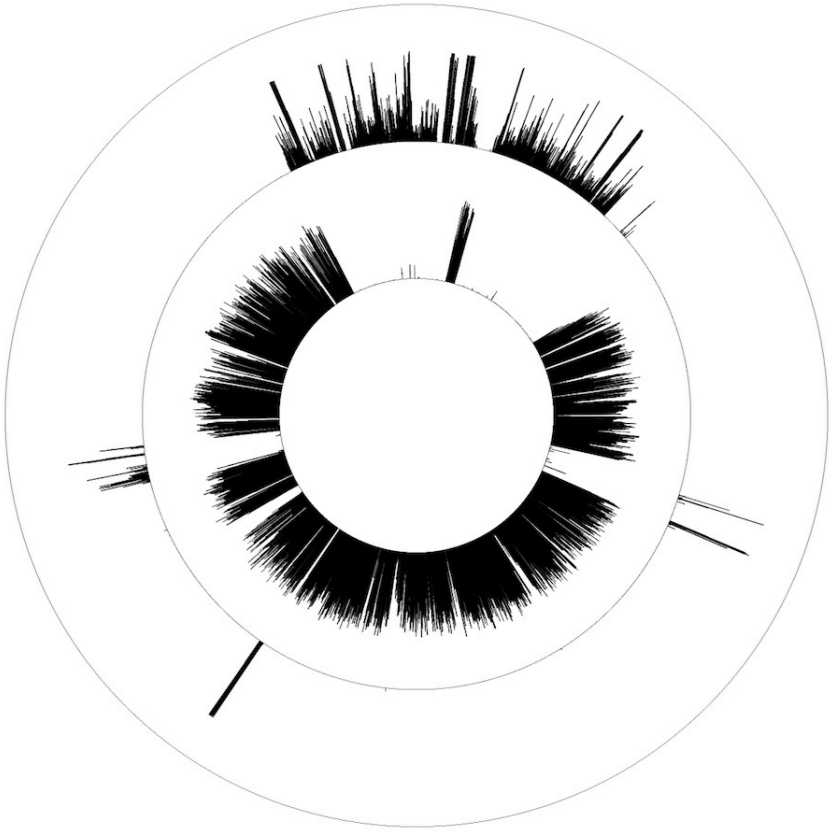

1e-05
1e-04
0.001
0.01
0.1
NMI
0.8
0.6
0.4
1e-05
1e-04
0.001
0.01
0.1
Mutation Rate

FIG 3

FIG 4



SaCOL



SaT0131

FIG 5

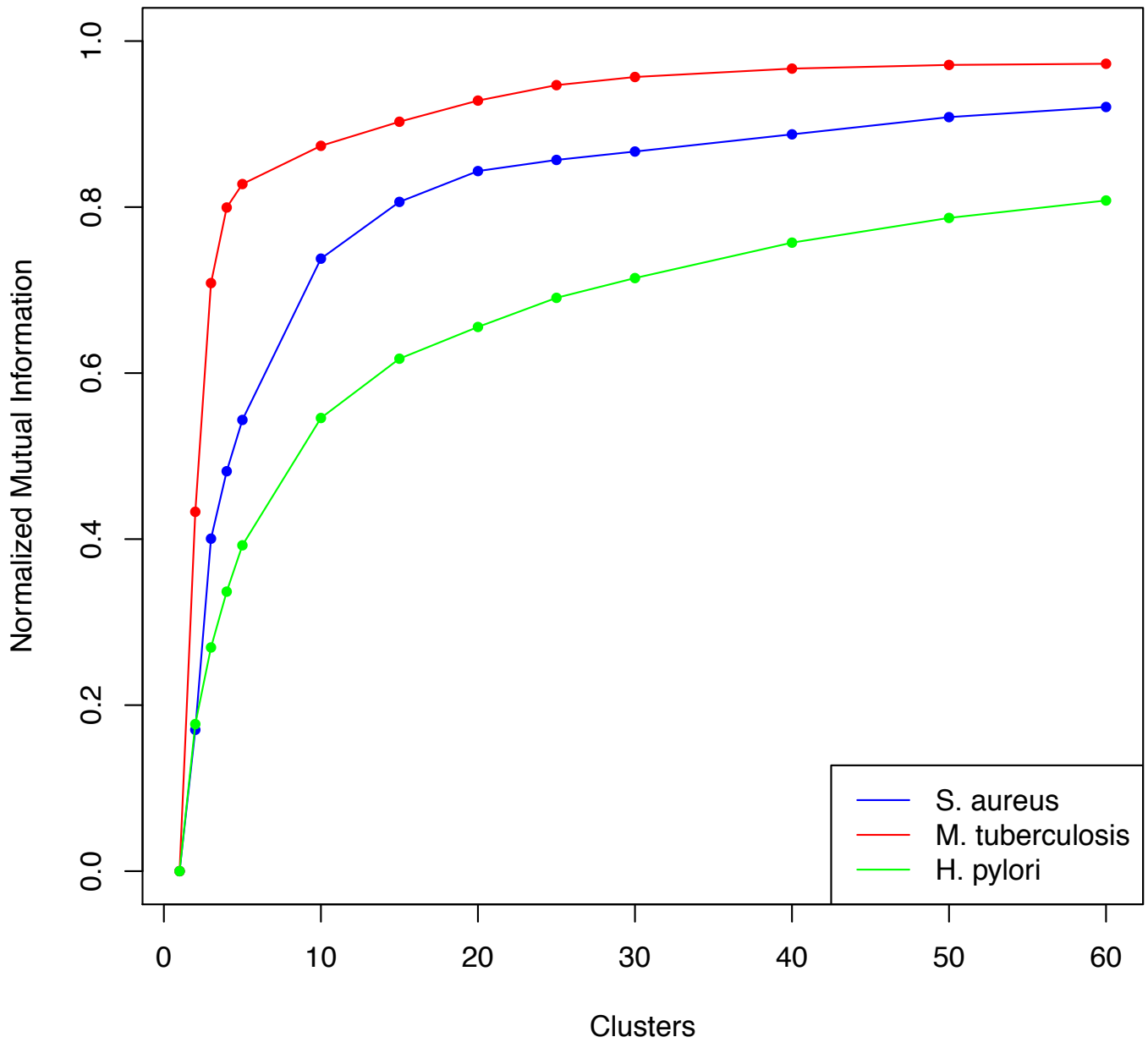

Normalized Mutual Information
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Clusters
S. aureus
M. tuberculosis
H. pylori

FIG 6A

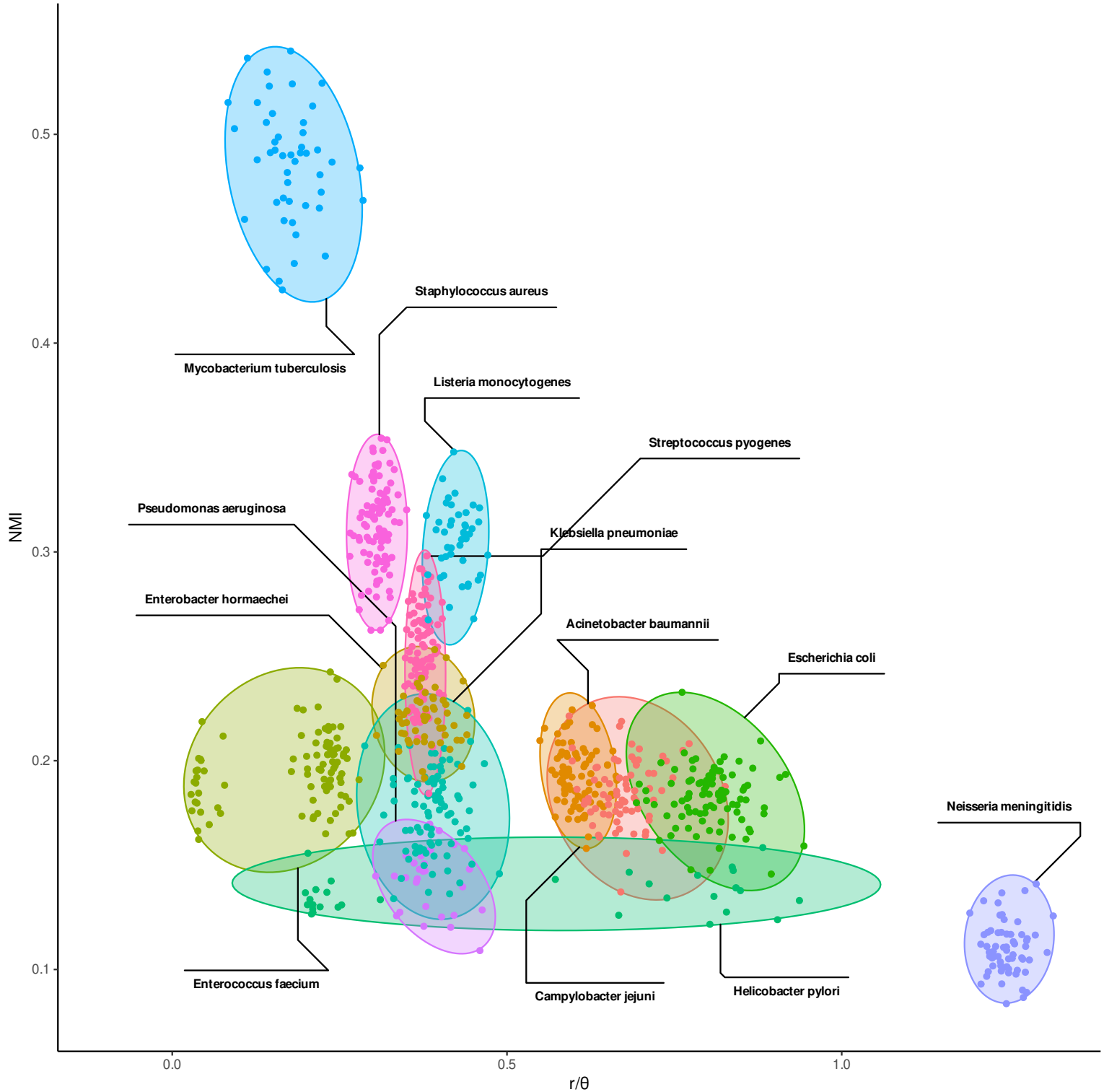

Mycobacterium tuberculosis
Staphylococcus aureus
Listeria monocytogenes
Streptococcus pyogenes
Pseudomonas aeruginosa
Klebsiella pneumoniae
Enterobacter hormaechei
Acinetobacter baumannii
Escherichia coli
Neisseria meningitidis
Enterococcus faecium
Campylobacter jejuni
Helicobacter pylori
NMI
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
r/θ

FIG 6B

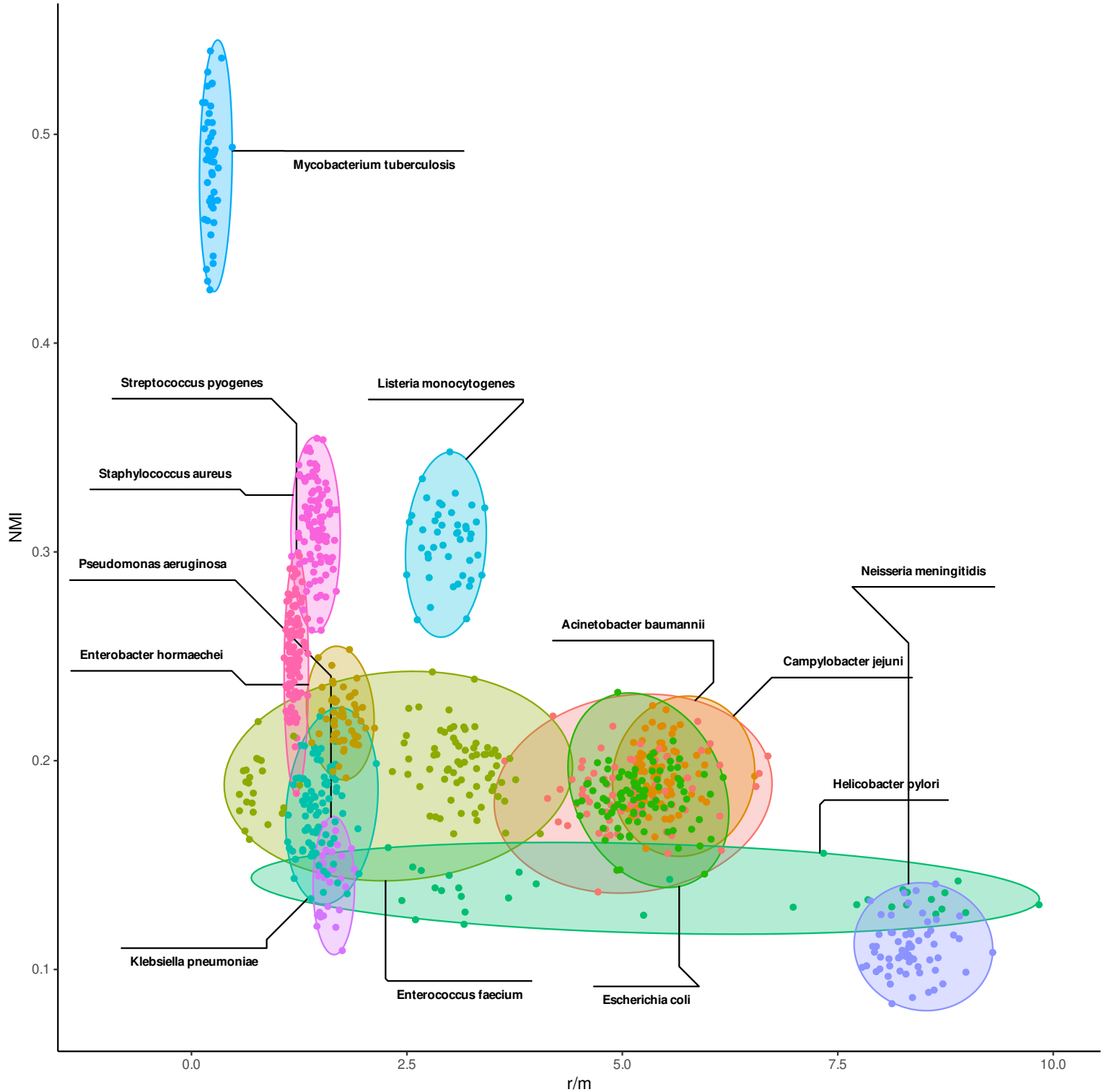
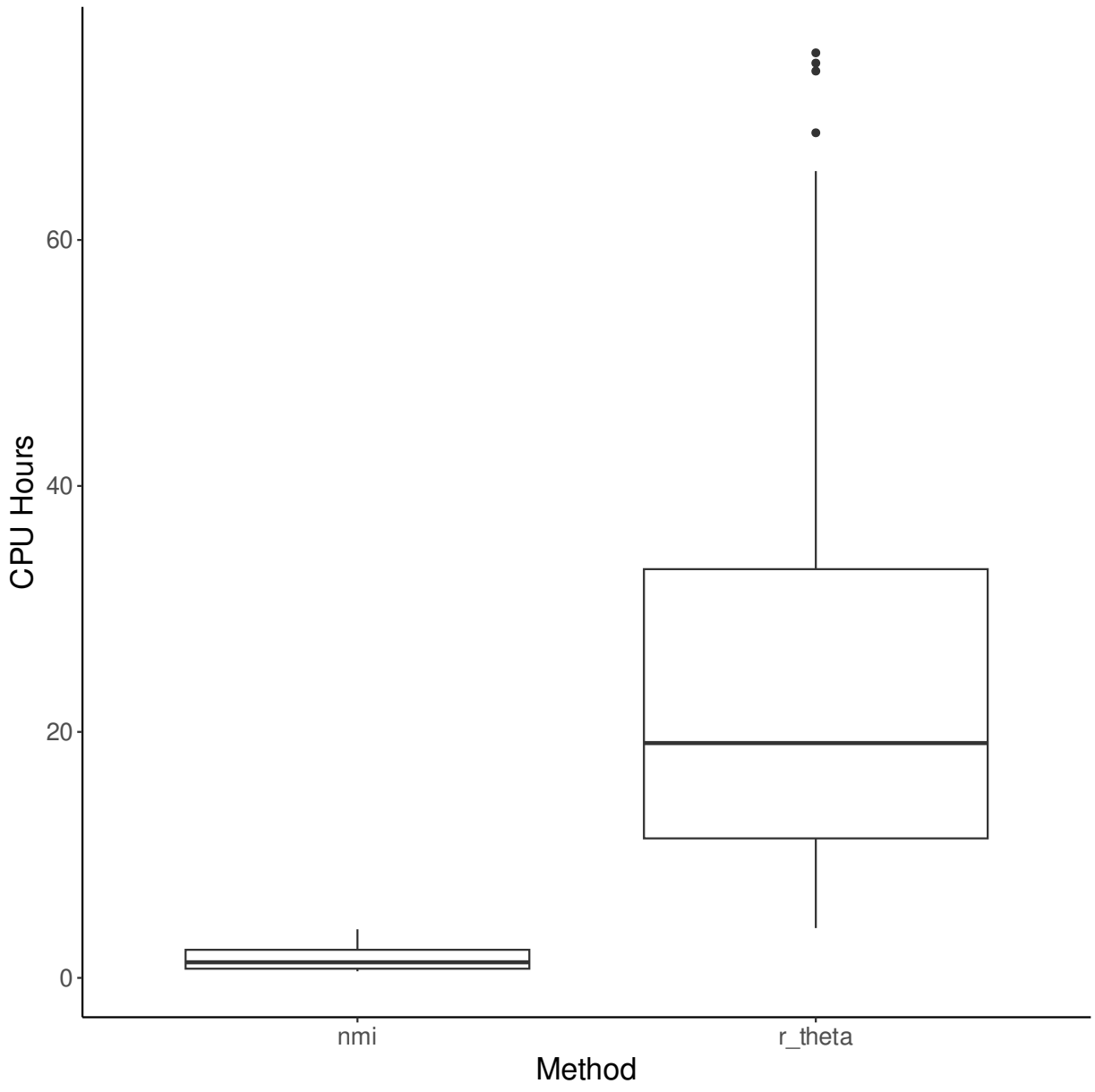
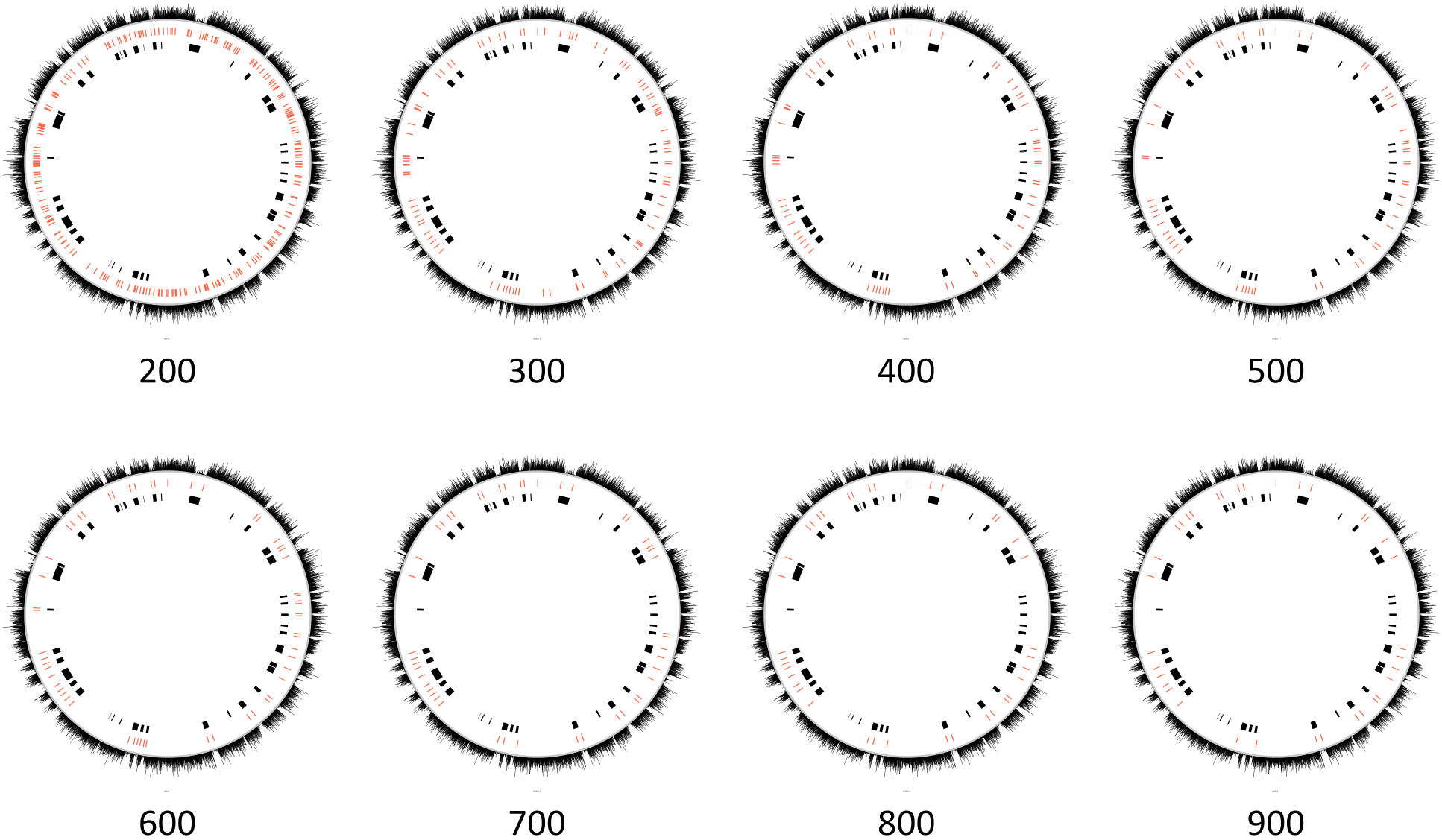
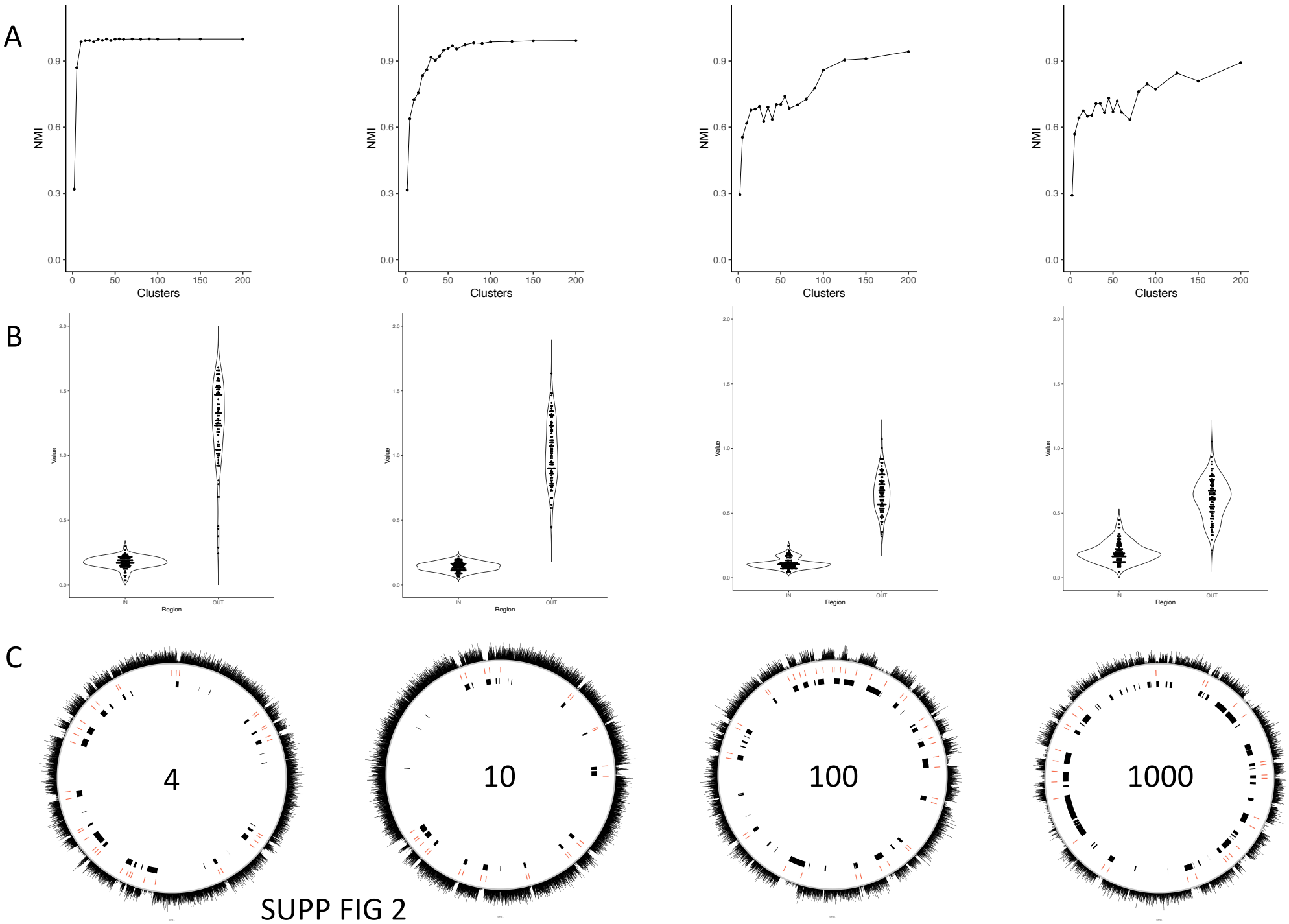

Mycobacterium tuberculosis
Streptococcus pyogenes
Listeria monocytogenes
Staphylococcus aureus
Pseudomonas aeruginosa
Neisseria meningitidis
Acinetobacter baumannii
Enterobacter hormaechei
Campylobacter jejuni
Helicobacter pylori
Klebsiella pneumoniae
Enterococcus faecium
Escherichia coli
NMI
0.5
0.4
0.3
0.2
0.1
0.0
2.5
5.0
7.5
10.0
r/m

FIG 7



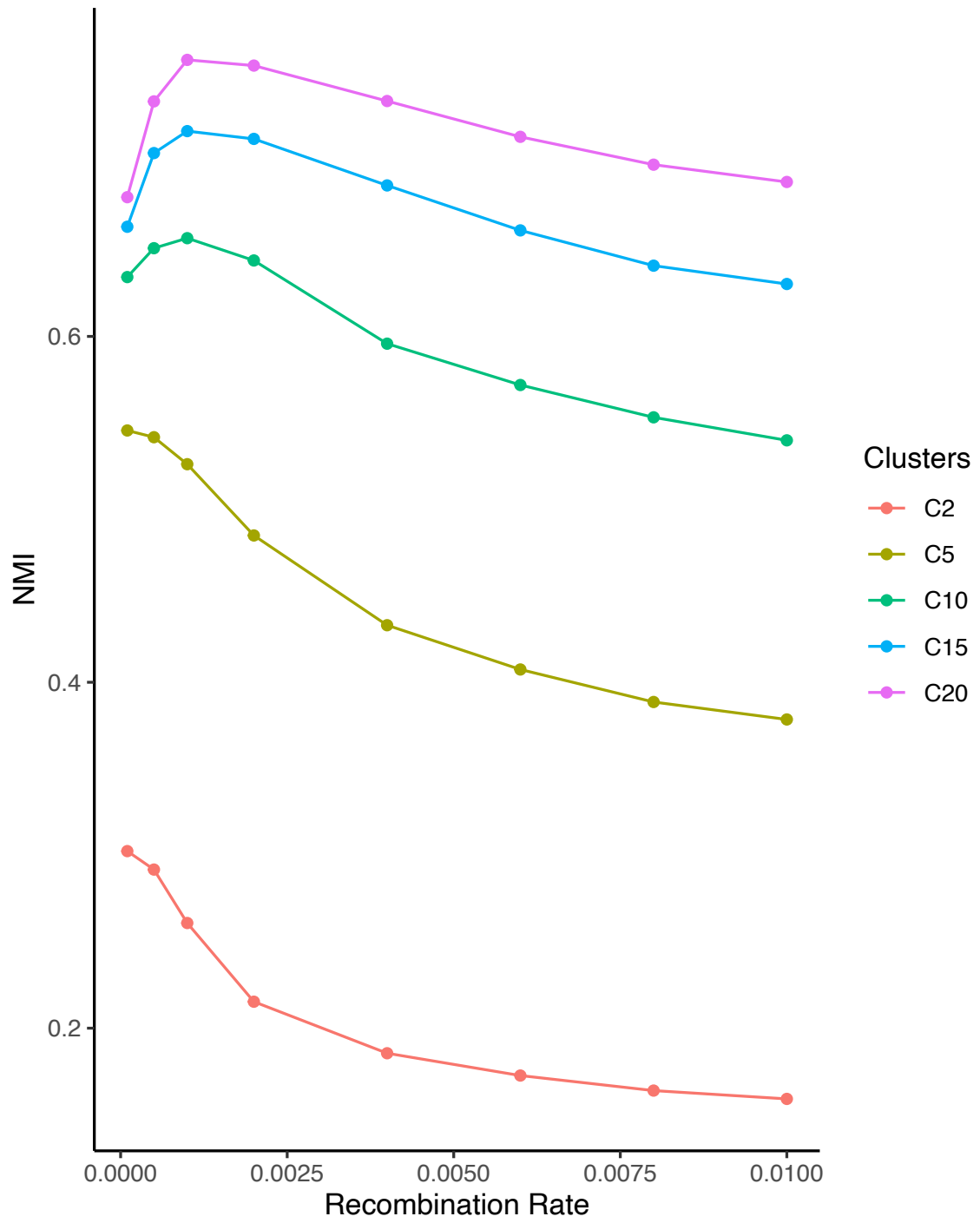


SUPP FIG 1

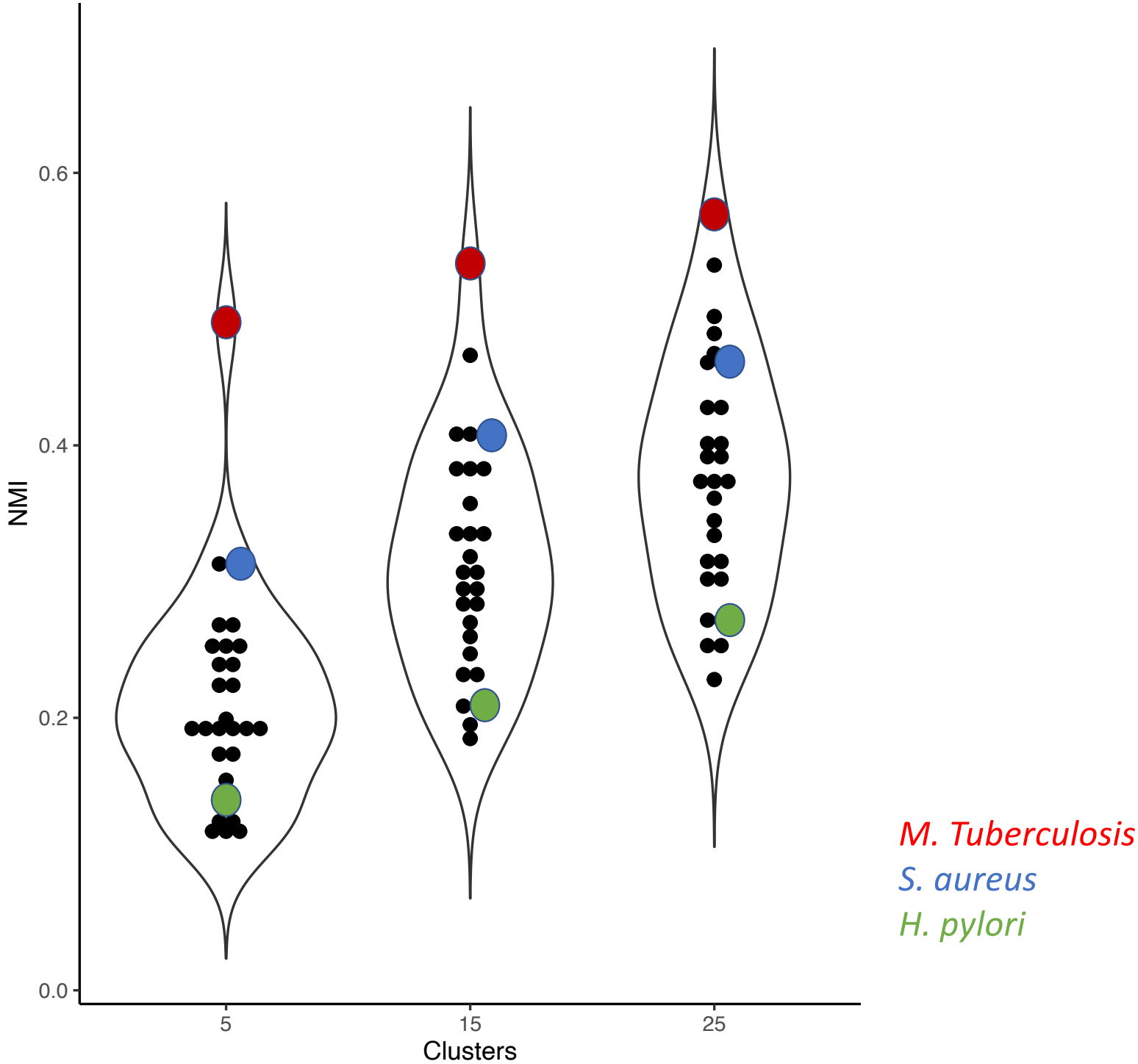


SUPP FIG 2

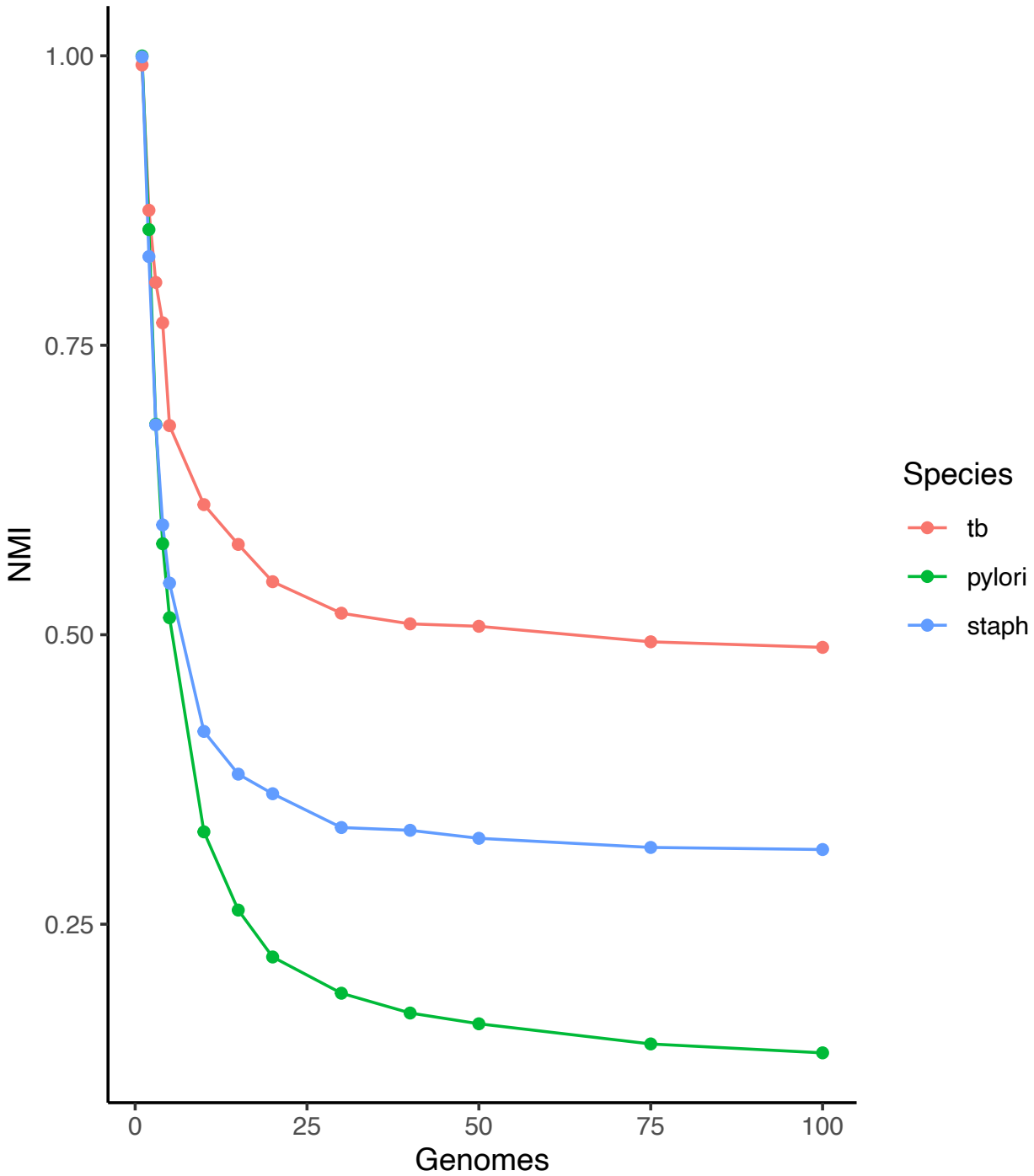
FIG 3A



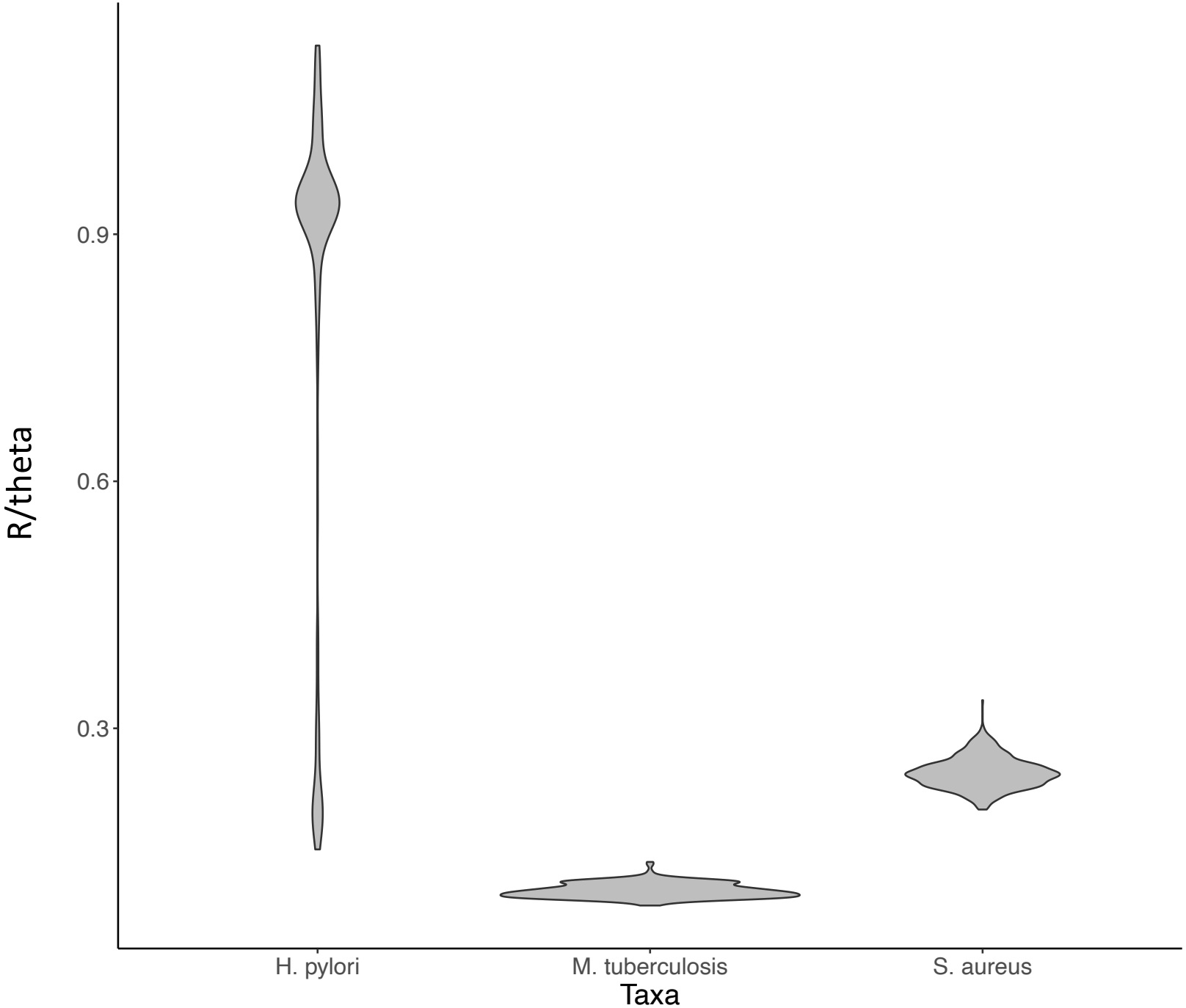
SUPP FIG 3


0.6
0.4
0.2
0.0
NMI
5
15
25
Clusters
M. Tuberculosis
S. aureus
H. pylori

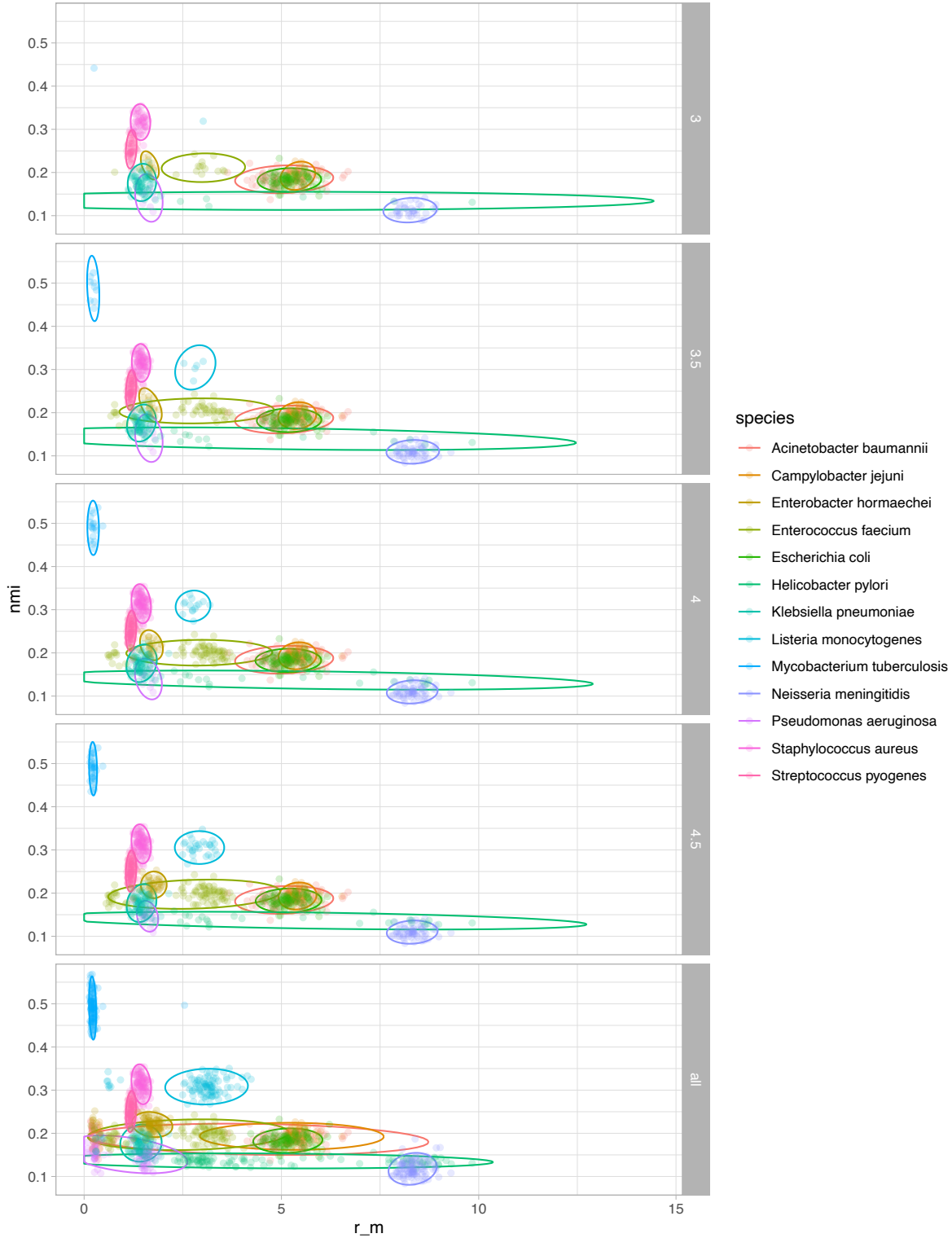
SUPP FIG 4


1.00
0.75
0.50
0.25
NMI
0
25
50
75
100
Genomes
Species
tb
pylori
staph

SUPP FIG 5



SUPP FIG 6


nmi
r_m
3
3.5
4
4.5
all
species
Acinetobacter baumannii
Campylobacter jejuni
Enterobacter hormaechei
Enterococcus faecium
Escherichia coli
Helicobacter pylori
Klebsiella pneumoniae
Listeria monocytogenes
Mycobacterium tuberculosis
Neisseria meningitidis
Pseudomonas aeruginosa
Staphylococcus aureus
Streptococcus pyogenes

SUPP FIG 7