HERO: Visualizing genome-wide patterns of recombination in microbial species and populations

Cooper J. Park[1#], Pekka Marttinen[2] and Cheryl P. Andam[1,3]

[1]University of New Hampshire, Department of Molecular, Cellular and Biomedical Sciences, Durham, New Hampshire 03824, USA

[2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland

[3]University at Albany, State University of New York, Department of Biological Sciences, Albany, New York 12222, USA

[#]Address correspondence to

Cooper Park, therealcooperpark@gmail.com

Abstract

Background

Homologous recombination is known to influence a myriad of evolutionary and population processes in many microbial species. However, growing evidence suggests that the frequency and distribution of recombination events can be influenced by genetic and ecological barriers between strains within a species. Despite the growing number of tools available to predict recombination events, no software provides the means to characterize donor-recipient relationships and other metrics of genome-wide recombination heterogeneity within a population or species.

Results

We present HERO, a Python tool which uses the output of the recombination detection tool fastGEAR to identify donors and recipients in recombination events. HERO also maps recombination events to user-defined metadata categories to help elucidate potential drivers of bias in recombination patterns and visualizes the results in publication-ready figures using Circos networks. It also reports and visualizes the variation in recombined DNA fragment size and events per gene as additional measures of variation.

Conclusions

HERO is a freely available Python tool for measuring and visualizing heterogeneity within a bacterial population's history of recombination. The code and documentation are available for download from https://github.com/therealcooperpark/hero. A real-world example of using the program can be found at https://github.com/therealcooperpark/hero_example.

Keywords: Recombination, population, genome

Background

Genetic recombination allows a microbial cell to rapidly acquire novel traits through incorporation of DNA fragments from other strains or species into its own genome [1]. It often involves the non-reciprocal unidirectional transfer of a homologous or highly similar segment of DNA from a donor to a recipient [1]. The consequences of genetic recombination are vast. Homologous recombination is known to influence a myriad of evolutionary and population processes, including levels of standing diversity, niche expansion, spread of resistance and virulence determinants, and rapid adaptive changes in response to new or fluctuating environmental conditions [2,3]. It can generate vaccine escape variants and the rapid diversification of surface antigens, allowing immune evasion [4]. Recombination of large DNA segments can also result to the emergence of novel genetic variants or hybrids with unique phenotypes such as multidrug resistance, hyper-virulence and increased transmissibility [3,5,6].

Although many studies have generated crucial insights into the nature and frequencies of recombination between bacterial species [7–9], it is often assumed that all strains recombine at a uniform frequency and randomly across the entire species. Recombination rates between strains of the same species can vary along a continuum spanning several orders of magnitude. Some strains also donate or receive DNA more often than others [10,11], while some strains tend to preferentially recombine with specific partners [12,13]. Such a pair of strains or lineages exchanging DNA more often between them than with others is said to be linked by a highway of recombination (or biased recombination). Highways likely represent specific lineages that function as hubs of gene flow, facilitating the rapid spread of genes associated with antibiotic resistance, host adaptation and immune interactions [12]. Within-species differences in recombination also suggest that lineages respond to selective pressures in different ways. Such

variation also implies that recombination itself can evolve in response to natural selection [14,15] and can occur quickly on an evolutionary timescale [16,17]. Hence, the idea of a single effective recombination rate for a species does not provide a biologically realistic representation of microbial evolution. Equally problematic is when studies attempt to fit the data to evolutionary and population genetic models that assume a constant species-wide rate of recombination.

Rapid recombination detection programs have been developed that can be used to identify recombined DNA fragments in large-scale whole genome datasets. ClonalOrigin generates a clonal phylogenetic tree and considers recombination events as regions of DNA that create localized discrepancies to the clonal phylogeny [18]. BratNextGen clusters regions in a genome that may be more distinct from other taxa than expected by normal mutation-driven evolution and creates a proportion of shared ancestry tree to group genomes that have a greater proportion of shared DNA clusters [19]. Gubbins identifies recombined sequences by iteratively scanning a sequence alignment and examining for elevated densities of nucleotide substitutions, and hence is more appropriate in investigations at the subspecies level [20]. FastGEAR uses a Hidden Markov Model to compare every nucleotide site in the target sequence to all remaining lineages and asks whether it is more similar to something else compared to other strains in the same lineage [21]. However, these programs do not provide a means to characterize population-wide patterns of donor-recipient relationships using the predicted recombination events. Here, we introduce the program HERO (Highways Enumerated by Recombination Observations), which uses the output of fastGEAR to identify donors and recipients in recombination events. We rely on fastGEAR to predict recombination events due to its ability to identify intra-species donors and its prediction accuracy at the species level. HERO also maps recombination events to user-defined metadata categories to help elucidate potential drivers of biases in recombination

partners. Unlike other tools such as Phandango [22], BratNextGen or fastGEAR whose images depend on phylogenetic trees to visualize population-level dynamics of recombination, HERO's primary figure is a network of donor-recipient pairs. Visualizing recombination this way provides a straightforward approach to elucidating recombination dynamics and interactions in any given microbial genomic dataset.

Implementation

Identifying DNA donors and recipients

HERO is a Python-implemented tool that uses the results of fastGEAR as input to infer donor-recipient pairs in recombination events [21]. Because fastGEAR identifies putative recombination events by predicting the origin of individual nucleotide sites from allelic patterns observed in different lineages, individual fastGEAR-defined lineages (i.e., groups of closely related strains) are reported as the potential donor for each recombination event rather than individual strains. For each putative recombined DNA segment that fastGEAR identifies, a Bayes factor (BF) is also computed based on the density of single nucleotide polymorphisms (SNP) that is compared between the putative recombination event and non-recombinant regions. FastGEAR uses a significance threshold of BF = 1 for recent recombination that represents a middle ground between false positive rate and power to detect recent recombination events.

FastGEAR identifies both recent and ancestral recombination events. HERO considers only the results of recent recombinations inferred by fastGEAR because they are the only events in which the direction of transfer (and thus, the specific donor and recipient for each event) can be identified. Here, a recent recombination event is defined as one which has affected only a subset of the strains found in a fastGEAR-predicted lineage, thus indicating that these

recombination events occurred after the lineages had diverged [21]. HERO first filters predicted recent recombination events by their reported length in base pairs and BF [23]. Recombination detection methods often rely on changes in the density of SNPs between the donor and recipient by scanning differences in SNP density between the putative recombined segment and surrounding non-recombinant genome. However, recombined DNA is often very similar in sequence to the original recipient genome, especially when the event occurs within a species [1]. Therefore, filtering recombination events by the length of the recombined segment is an arbitrary cut-off when the short recombination events predicted by fastGEAR are likely to be only the divergent piece of larger DNA fragments. Therefore, HERO uses the same default thresholds as determined by fastGEAR, a BF of 1 and a minimum fragment length of 0 base pairs, but the values can be adjusted by the user.

HERO accepts associated metadata (e.g., clusters delineated in population structure analysis, environment, specimen source, human or animal host) for each genome in the dataset. HERO identifies a donor-recipient pair between the recipient's metadata group and the most likely donor metadata group. Because fastGEAR identifies a cluster of potential donor strains (i.e., lineage) rather than a single donor genome, HERO uses a simple distance matrix to compare the sequence similarity between the recombined DNA in the recipient to the same region from each genome in the donor lineage. Because shared ecology or genetics often facilitates recombination between closely related strains [24–27], the metadata group containing the genome with the highest similarity to the recipient is considered the donor group for that event. If a single donor-recipient pair cannot be identified for an event (i.e., if donors from multiple different metadata groups tie for the highest similarity to the recipient), the event is discarded from the analysis. Moreover, multiple recombination events with overlapping

nucleotide ranges between the same donor-recipient metadata pair are considered to be a single recombination event.

Identifying highways of recombination

Highways of recombination are donor-recipient pairs which recombine with each other significantly more frequently than they do with other strains in the population [12,13]. HERO identifies a highway of recombination as a pair of metadata groups with a number of recombination events greater than *3*IQR + Q3*, where *IQR* is the inter-quartile range and *Q3* is the third quartile of the distribution of recombination events per donor-recipient pair. Hence, the definition of a highway will vary based on the number of metadata groups and genomes included in the dataset being examined. Furthermore, the direction of a recombination event is considered when determining a highway, making it possible for a recombining pair to be a highway in one direction, but not the other.

Visualizing results

The primary output of HERO is a pair of network images generated using Circos [28] . We first define the phylogenetic relationships and membership of the strains to metadata groups (Fig. 1a). The colors of the fragments on the outer ring of the phylogeny represent distinct metadata groups in which the strains belong to (Fig. 1a). In this example, we used sequence clusters (SC) defined by a Bayesian hierarchical clustering method implemented in BAPS [29] as our metadata group (Fig. 1a). We then create a network that shows the pairs of recombining strains between SCs (Fig. 1b). This figure is generated by HERO with a file name of circos.svg. Here, the length of the fragments in the outer ring is proportional to the number of recombination

events involving the group. The intertwining ribbons between groups represent donor-recipient pairs of recombination where the ribbon is colored to match the donor and the donor edge of the ribbon is indented towards the center of the circle. The width of the ribbon is proportional to the number of recombination events between a pair of genomes. Because the direction of a recombination event is considered when visualizing these pairs, it is possible for two ribbons to exist between the same pair of metadata groups. There is an option to highlight highways of recombination as seen in the output "highway_circos.svg" (Fig. 1c). In addition to the circos networks, HERO generates frequency histograms showing the lengths of recombined DNA sequences, the number of recombination receipts per genome and the number of recombination events per gene. HERO also provides supporting text files of the data for all figures.

Results and Discussion

We next demonstrate the utility of HERO with the same collection of 616 whole-genome *Streptococcus pneumoniae* isolates sampled in Massachusetts, USA [30] that was previously used to demonstrate the effectiveness of fastGEAR to detect recombination [21]. The methods we used to prepare the dataset have been described in detail in Additional File 1 and Accession IDs for all genomes can be found in Additional File 2.

Exploring recombination dynamics in *S. pneumoniae*

We first used Roary [31] to characterize the pan-genome of the entire *S. pneumoniae* population. We identified 1,161 core genes (i.e., genes present in ≥ 99% of strains) and 6,133 shared accessory genes (i.e., genes present in at least 2 genomes, but less than 99% of the population) out of a total of 7,511 genes in the pan-genome. We then filtered out 2,779 gene

alignments with numerous gaps (i.e., the number of gaps in any individual sequence was ≥ 5% of its total gene length). From the remaining 4,732 gene alignments, we identified 262 genes that exhibit evidence of recombination and a total of 938 recent recombination events using fastGEAR. We then used HERO to identify the distribution of these recombination events across the 16 SCs in which each genome was assigned to in its original publication [30] (Fig. 1a). Out of the 256 possible unidirectional pairs of SCs, 183 of them showed evidence for recombination having 1-130 recombination events in any one pair (mean = 5.1 events) (Fig. 1b). Using HERO's definition of a highway of recombination, we found 12 SC pairs that met the definition of a highway (i.e., pairs with ≥ 13 unique recombination events) (Fig. 1c). Highways of recombination accounted for 397 of the total 938 (42.3%) recombination events inferred within the population.

All highways of recombination involved the only multiphyletic cluster SC16 as either a donor or recipient. Based on the phylogenetic tree for the population, SC16 is likely composed of multiple individual clusters too small to be detected independently by the BAPS clustering software [29]. To improve the resolution of our recombination analysis, we used HERO to re-calculate the distribution of recombination events, but this time breaking SC16 into eight smaller SCs (labeled as SC16a-h) where each new SC is separated by at least one monophyletic SC (Fig. 2a). Using these newly assigned clusters, the number of possible unidirectional SC pairs in the population increased to 529. We found 339 of these pairs to have evidence of recombination, having 1-30 events in any one pair (mean = 3.2 events; Fig 2b). These pairs shared a total of 1,087 recombination events across 253 different genes. In this new clustering scheme, the threshold for a highway of recombination decreased to 12 events per pair, yet only seven pairs

(2.1% of all recombining pairs) were identified as highways (Fig. 2c). These highways accounted for 126 (11.6%) out of the total 1,087 recombination events. While 12 of these highways involved SC16c as either a donor or recipient, the remaining six highways were scattered between pairs involving SC16b, SC12 and SC6.

The number of recombination events per genome varied, ranging from 5 and 36 recombination events in a single genome (mean = 13.6 events; median = 13) (Fig. 3a). The detected fragment size of recombination events varied from 2 - 4,446 bp (mean = 209.6 bp; median = 94) (Fig. 3b). Lastly, we detected variation in the number of recombination events per gene, ranging from 1 and 34 events per gene (mean = 4.3 events; median = 2) (Fig. 3c).

Characterizing the properties of a recombination pair

Intra-species variation in recombination has been found to exist within multiple bacterial species and across broad ecological settings [12,13,26]. However, the extent to which genetic and ecological factors drive this variation remains poorly understood. By combining results generated by HERO with other common measures of population diversity we sought to identify trends within the *S. pneumoniae* population that could be extrapolated to other species and populations.

One of the most significant challenges to predicting recombination pairs is the effect of sampling bias on donor identification. Under-sampling a population risks missing genomes with unique gene repertoires that could potentially be the source of a recombination event. In contrast, having one or a few well-sampled subpopulations may exaggerate the credit these larger groups

get as a donor by being the dominant source of variation to which putative recombination events are compared against. Using the 616 *S. pneumococcus* dataset, we tested the effect of sampling in our population. We first compared the number of genomes in each SC to the number of recombination events involving the SC and found a weak but insignificant correlation between the two (p-value = 0.14, R^2 = 0.06) (Fig. 4a). We also compared the number of shared genes within an SC to its number of recombination events and found a significant positive correlation (p-value = 0.001, R^2 = 0.37) (Fig. 4b). Lastly, we calculated the Average Nucleotide Identity (ANI) for each SC using fastANI v1.0 [32]. ANI estimates the average nucleotide identity of all orthologous genes shared between any two genomes [32]. We calculated the SC-wide ANI, which refers to the mean of all possible pairwise ANI values between any two genomes in the SC. For each SC, we compared the number of recombination events with its SC-wide ANI and found a negative but insignificant correlation (p-value = 0.97, R^2 = 0) (Fig. 4c). While the size of an SC can influence the amount of diversity within it, it is not clear that sample size alone is significantly influencing the distribution of predicted recombination events among the SCs. Instead, the number of potential genes which can be measured for recombination appear to be the primary driver of sampling bias. Therefore, the primary limitations to HERO stem from the assigned metadata groups. Multiphyletic clades, such as SC16 from this *S. pneumoniae* population, are likely to distort findings from SCs derived from sequence data as the cumulative genetic diversity from multiple clades will contribute many more opportunities to find recombination than from within a single monophyletic clade. However, if multiphyletic clades are expected (e.g., in ecologically derived clusters), sufficient representation for each cluster will be crucial to accurately attributing recombination events to a donor cluster.

Conclusions

In summary, we present HERO, a user-friendly python program that uses the output from the popular recombination detection tool fastGEAR to identify donor-recipient pairs in recombination events. We propose a definition of a highway of recombination which can capture unique trends in recombination frequencies within a population. The simplicity of HERO's usage combined with its informative visualization output provide a detailed look into recombination dynamics within a population or species.

Figure Legends

Figure 1. HERO-derived recombination pairs compared to SC positions in a phylogeny. a) Core genome phylogeny of 616 *S. pneumoniae* genomes. The phylogeny was reconstructed using the concatenated alignment of 1,161 core genes. The scale bar represents substitutions per site. Outer ring indicates SC membership identified using BAPS. b) A recombination network generated by HERO. Outer ring fragments are individual BAPS-derived SCs. Length of each fragment is proportional to the number of recombination events affecting the SC. Ribbons connect clusters that share recombination events where the thickness of the ribbon is proportional to the number of shared events, the color of the ribbon matches the color of the donor cluster in panel a, and the recombination donors are indicated by edges of the ribbon indented towards the middle of the circle. c) A recombination network (identical to panel b) highlighting the highways of recombination. Non-highway recombinations are colored gray.

Figure 2. HERO recombination pairs in a phylogeny with the SC16 split into smaller clusters. a) Core genome phylogeny of the *S. pneumoniae* population. The tree is identical to Fig. 1a, except

that SC16 was divided into eight smaller clusters labeled 16a-h. b) A recombination network generated by HERO. c) Highways of recombination. Network legend is identical to Fig. 1b, c.

Figure 3. Measures of genome-wide variability in recombination. a) Histogram showing the frequency distribution of events per recipient genome. b) Histogram showing the frequency distribution of recombination fragment size (bp). c) Histogram showing the frequency distribution of events per gene.

Figure 4. Characteristics of recombination pairs. a) Relationship between the number of genomes in an SC and the total number of predicted recombination events per SC. b) Relationship between the number of shared genes in an SC and the number of predicted recombination events. c) Relationship between SC-wide ANI and the number of recombination events per SC.

Availability and Requirements

Project name: HERO

Project home page: https://github.com/therealcooperpark/hero

Operating System(s): Linux

Programming language: Python 3.6

Other requirements: BioPython (Python), Pandas (Python), Plotnine (Python), fastGEAR, Circos, GNU Parallel

License: MIT

Any restrictions to use by non-academics: None

List of Abbreviations

ANI – Average nucleotide identity

BF – Bayesian Factor

IQR – Inter-quartile Range

Q3 – Third Quartile

SC – Sequence Cluster

Declarations

Availability of data and materials

Methods for preparing the dataset used here can be found in Additional File 1. Accession IDs for *S. pneumoniae* genomes can be found in Additional File 2. A tutorial of the HERO analysis of the *S. pneumococcus* dataset including intermediate files for each step can be found at https://github.com/therealcooperpark/hero_example

Competing Interests

The authors declare that they have no competing interests.

Funding

The study was supported by the National Science Foundation (grant number 1844430) and start-up funds from UAlbany College of Arts and Sciences to C.P.A**.** The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

C.P.A. conceived of the project. C.J. P. designed the software, wrote documentation and performed bioinformatic analyses. C.P.A. and P. M. guided the work. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the University of New Hampshire Resource Computing Center where all bioinformatics analyses were performed. The authors thank Anthony Westbrook for providing technical and bioinformatics assistance. We also thank Dr. Claire Chewapreecha for discussions on software implementation.

Additional Files

Additional File 1. Details about methods of *S. pneumoniae* analysis.

Additional File 2. Accession IDs and metadata for 616 *S. pneumoniae* genomes used in analysis.

Additional File 3. Supplementary Data for all figures

References

[1] Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol 2010;18:315–22. https://doi.org/10.1016/j.tim.2010.04.002.

[2] Hanage WP. Not so simple after all: Bacteria, their population genetics, and recombination. Cold Spring Harb Perspect Biol 2016;8:a018069. https://doi.org/10.1101/cshperspect.a018069.

[3] Spoor LE, Richardson E, Richards AC, Wilson GJ, Mendonca C, Gupta RK, et al. Recombination-mediated remodelling of host–pathogen interactions during staphylococcus aureus niche adaptation. Microb Genomics 2015;1:1–14.

https://doi.org/10.1099/mgen.0.000036.

[4] Croucher NJ, Campo JJ, Le TQ, Liang X, Bentley SD, Hanage WP, et al. Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. Proc Natl Acad Sci U S A 2017;114:E357–66. https://doi.org/10.1073/pnas.1613937114.

[5] Coyle NM, Bartie KL, Bayliss SC, Bekaert M, Adams A, McMillan S, et al. A Hopeful Sea-Monster: A Very Large Homologous Recombination Event Impacting the Core Genome of the Marine Pathogen Vibrio anguillarum. Front Microbiol 2020;11. https://doi.org/10.3389/fmicb.2020.01430.

[6] Perron GG, Lee AEG, Wang Y, Huang WE, Barraclough TG. Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations. Proc R Soc B Biol Sci 2012;279:1477–84. https://doi.org/10.1098/rspb.2011.1933.

[7] González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. Impact of homologous recombination on the evolution of prokaryotic core genomes. MBio 2019;10. https://doi.org/10.1128/mBio.02494-18.

[8] Levin BR, Cornejo OE. The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria. PLoS Genet 2009;5:e1000601. https://doi.org/10.1371/journal.pgen.1000601.

[9] Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J 2009;3:199–208. https://doi.org/10.1038/ismej.2008.93.

[10] Rodríguez-Beltrán J, Tourret J, Tenaillon O, López E, Bourdelier E, Costas C, et al. High recombinant frequency in extraintestinal pathogenic escherichia coli strains. Mol Biol Evol 2015;32:1708–16. https://doi.org/10.1093/molbev/msv072.

[11] Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLoS Genet 2019;15:e1008114. https://doi.org/10.1371/journal.pgen.1008114.

[12] Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 2014;46:305–9. https://doi.org/10.1038/ng.2895.

[13] Park CJ, Andam CP. Distinct but Intertwined Evolutionary Histories of Multiple Salmonella enterica Subspecies. MSystems 2020;5:e00515-19-e00515-19. https://doi.org/10.1128/msystems.00515-19.

[14] Lobkovsky, Alexander E, Wolf, Yuri I, Koonin E V. Gene Frequency Distributions Reject a Neutral Model of Genome Evolution. Genome Biol Evol 2015;5:233–42.

[15] Peñalba J V., Wolf JBW. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet 2020;21:476–92. https://doi.org/10.1038/s41576-020-0240-1.

[16] Cowley LA, Petersen FC, Junges R, Jimson D. Jimenez M, Morrison DA, Hanage WP. Evolution via recombination: Cell-to-cell contact facilitates larger recombination events in Streptococcus pneumoniae. PLOS Genet 2018;14:e1007410. https://doi.org/10.1371/journal.pgen.1007410.

[17] Evans BA, Rozen DE. Significant variation in transformation frequency in Streptococcus pneumoniae. ISME J 2013;7:791–9. https://doi.org/10.1038/ismej.2012.170.

[18] Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 2010;186:1435–49. https://doi.org/10.1534/genetics.110.120121.

[19] Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 2012;40:e6. https://doi.org/10.1093/nar/gkr928.

[20] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 2015;43:e15. https://doi.org/10.1093/nar/gku1196.

[21] Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Mol Biol Evol 2017;34:1167–82. https://doi.org/10.1093/molbev/msx066.

[22] Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: An interactive viewer for bacterial population genomics. Bioinformatics 2018;34:292–3. https://doi.org/10.1093/bioinformatics/btx610.

[23] Bernardo, JM, Smith A. Bayesian Theory | Wiley. IOP Publ 2001. https://www.wiley.com/en-us/Bayesian+Theory-p-9780471494645 (accessed October 29, 2020).

[24] Skippington E, Ragan MA. Phylogeny rather than ecology or lifestyle biases the construction of Escherichia coli-Shigella genetic exchange communities. Open Biol 2012;2:120112. https://doi.org/10.1098/rsob.120112.

[25] Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 2011;21:599–609. https://doi.org/10.1101/gr.115592.110.

[26] Sheppard SK, Cheng L, Méric G, De Haan CPAA, Llarena A-KK, Marttinen P, et al. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol

Ecol 2014;23:2442–51. https://doi.org/10.1111/mec.12742.

[27] Dillon MM, Thakur S, Almeida RND, Wang PW, Weir BS, Guttman DS. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the Pseudomonas syringae species complex 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0603 Evolutionary Biology. Genome Biol 2019;20:1–28. https://doi.org/10.1186/s13059-018-1606-y.

[28] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res 2009;19:1639–45. https://doi.org/10.1101/gr.092759.109.

[29] Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 2008;9:539. https://doi.org/10.1186/1471-2105-9-539.

[30] Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 2013;45:656–63. https://doi.org/10.1038/ng.2625.

[31] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3. https://doi.org/10.1093/bioinformatics/btv421.

[32] Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9:5114. https://doi.org/10.1038/s41467-018-07641-9.

[33] Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria 06 Biological Sciences 0604 Genetics. BMC Evol Biol 2018;18. https://doi.org/10.1186/s12862-018-1272-4.