Eudes Barbosa[1,2] / Richard Röttger[1] / Anne-Christin Hauschild[1] / Siomar de Castro Soares[2,4] / Sebastian Böcker[3] / Vasco Azevedo[2] / Jan Baumbach[1]

LifeStyle-Specific-Islands (LiSSI): Integrated Bioinformatics Platform for Genomic Island Analysis

[1] University of Southern Denmark, Department of Mathematics and Computer Science, Odense, Denmark, E-mail: jan.baumbach@imada.sdu.dk
[2] Federal University of Minas Gerais, Institute of Biological Sciences, Belo Horizonte, Brazil
[3] Friedrich-Schiller-Universität Jena, Faculty of Mathematics and Computer Science, Jena, Germany
[4] Federal University of Triângulo Mineiro, Department of Immunology, Microbiology and Parasitology, Uberaba, Brazil

Abstract:
Distinct bacteria are able to cope with highly diverse lifestyles; for instance, they can be free living or host-associated. Thus, these organisms must possess a large and varied genomic arsenal to withstand different environmental conditions. To facilitate the identification of genomic features that might influence bacterial adaptation to a specific niche, we introduce LifeStyle-Specific-Islands (LiSSI). LiSSI combines evolutionary sequence analysis with statistical learning (Random Forest with feature selection, model tuning and robustness analysis). In summary, our strategy aims to identify conserved consecutive homology sequences (islands) in genomes and to identify the most discriminant islands for each lifestyle.

Keywords: Bacteria, Lifestyle, Machine Learning, Island, Homologous genes

DOI: 10.1515/jib-2017-0010

Received: March 13, 2017; **Revised:** April 10, 2017; **Accepted:** April 19, 2017

1 Introduction

Bacterial genomes are relatively small, and vary in size by more than one order of magnitude, ranging from approximately 150 kilobases [1] to 13 megabases [2]. Due to processes such as rearrangements, gene duplication or loss, and horizontal gene transfer, bacterial genomes are extremely variable in terms of gene repertoires. Further, bacterial chromosome architecture is subject to a balance between genetic novelty and stability of the gene arrangement in the chromosome. While genetic novelties have great influence in adaptation, the introduction of new genes tends to disrupt the chromosome organization. The trade-off between these two processes depends on bacterial niche and lifestyle [3]. Furthermore, gene order conservation usually involves two categories of genes: namely, rare and persistent genes, where the mechanisms that led to each kind are not identical. In summary, conservation cannot be explained in all instances by operons and lateral gene transfer [4].

Throughout the years, several models were developed to explain gene order conservation [5]. The latest models are the Co-regulation Model (CM) and the Selfish Operon Model (SOM) [6], [7]. CM is based on the observation that genes that are found close together on the chromosome can be regulated efficiently. Therefore, genes involved in the same metabolic pathway or the same protein complex would constitute selective advantages when being clustered. This model leads to the conclusion that operons are the origin of the cluster organization in bacterial chromosomes. The main problem with CM is that it fails to explain the selective advantages of gene proximity while co-transcription still is not possible. The second model, SOM, is based on lateral gene transfer. The model states that if a set of genes provides equivalent fitness (independent of their position), physical proximity provides an advantage to the genes themselves. In this case, clustered genes are favored over spread ones while being transferred. Therefore, genes can be gradually moved close together even before co-transcription is possible [7], [8].

Different environments, habitats, energy sources, and niches ("lifestyles" for short) require particular characteristics from bacterial species to survive, reproduce and proliferate. Hence, one can observe various genome-sizes and mobile DNA elements associated with different lifestyles [9], [10]. It is reasonable to expect that these

Jan Baumbach is the corresponding author.

organisms possess a large and varied genomic arsenal to withstand different environmental conditions. To facilitate the identification of genomic features that might influence bacterial adaptation to a specific niche, we introduce LiSSI (LifeStyle-Specific-Islands).

2 The LiSSI Approach

LiSSI is organized into four modules. Subsequent to data acquisition (GenBank or locally), a standard run consists of: defining groups of putative homologous genes (evolutionary sequence analysis), followed by island detection and identification of the most discriminant islands for a given lifestyle (statistical learning). Further, functional classification can be used to search for protein domains in the selected genes/islands. Optionally, the tool can be used without island detection. In this case, it will report putative homologous genes that are mainly associated with a given lifestyle. LiSSI's analysis pipeline is shown in Figure 1.

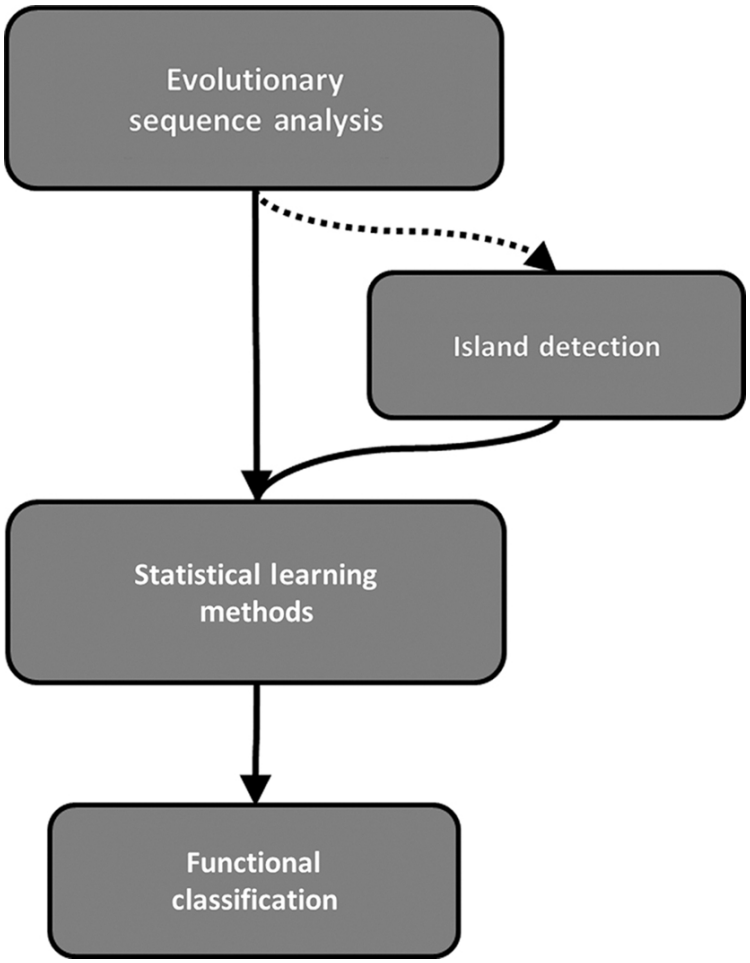

Figure 1: LiSSI pipeline. LiSSI is divided into four modules. After data acquisition (GenBank or locally), a standard run involves: the definition of groups of putative homologous genes (evolutionary sequence analysis), followed by island detection and identification of the most discriminant islands for a given lifestyle (statistical learning methods). Further, functional classification can be used to search for protein domains in the selected genes/islands. Full lines represent mandatory steps, dotted lines represent optional steps.

LiSSI was implemented in Java and R. Java is used to generate the graphical user interface and in file manipulation operations, whereas R is used for the statistical analysis. LiSSI combines evolutionary sequence analysis with statistical learning methods (Random Forest with feature selection, model tuning and robustness analysis); and additional steps for island detection and functional classification of the features. In summary, our strategy aims to identify conserved consecutive homology sequences (islands) in genomes and to identify the most discriminant islands for a given lifestyle.

LiSSI extends the methodology described by Barbosa et al. [11]. Instead of solely analyzing individual genes, we here aim to study the evolution of genome organization. To address island detection, we included Gecko 3 in our pipeline [12]; to address functional classification of the selected features, we relied on a BioJava [13]

module to implement a Pfam search [14]. Pfam stores protein families and is used to identify conserved protein domains. Finally, the software incorporates a module to perform a BLAST search [15] against NCBI. In summary, we now 1) extend our strategy to not just work on single genes but gene islands, and 2) provide a standalone software including a graphical user interface (GUI) guiding the user through all necessary data analysis steps, from importing genomes from NCBI via BLASTing, homology detection, island finding, and classification, to feature selection, decision tree visualization and functional enrichment.

2.1 LiSSI Input Data Processing

LiSSI has an intuitive and straight-forward GUI layout using a wizard dialog guiding the user from selecting and retrieving the genome annotation files to setting the analysis parameters. The result plots and tables are organized in a separate tab.

Step 1 – Load genomes. The first step is to load a set of genome annotations of representatives for each of the lifestyles of interest. With LiSSI, we implemented three options: "Select from local folder", "Download from GenBank" or a combination of both. One may now either select local directories storing files in GenBank format, or select a list of genomes/species for a list of all fully sequenced genomes available at NCBI.

Step 2 – Select genomes. Here, one has to confirm the selection of genomes using some basic information downloaded in the previous step.

Step 3 – Set parameters. The third and final step is displayed in Figure 2. The user is asked to set the parameters 1) for homology detection using BLAST [15] and Transitivity Clustering [16], 2) for island detection using Gecko [17], and 3) for the Random Forest classifier [18], [19]. Note that one may also load intermediate results from previous runs or external analyses here. The LiSSI web site provides detailed tutorials guiding through this process step by step. All preset/standard parameters have been selected based on statistical considerations published in previous work (homology detection, refer to [20] or on our experience with the use case data presented below [island detection and classification]).

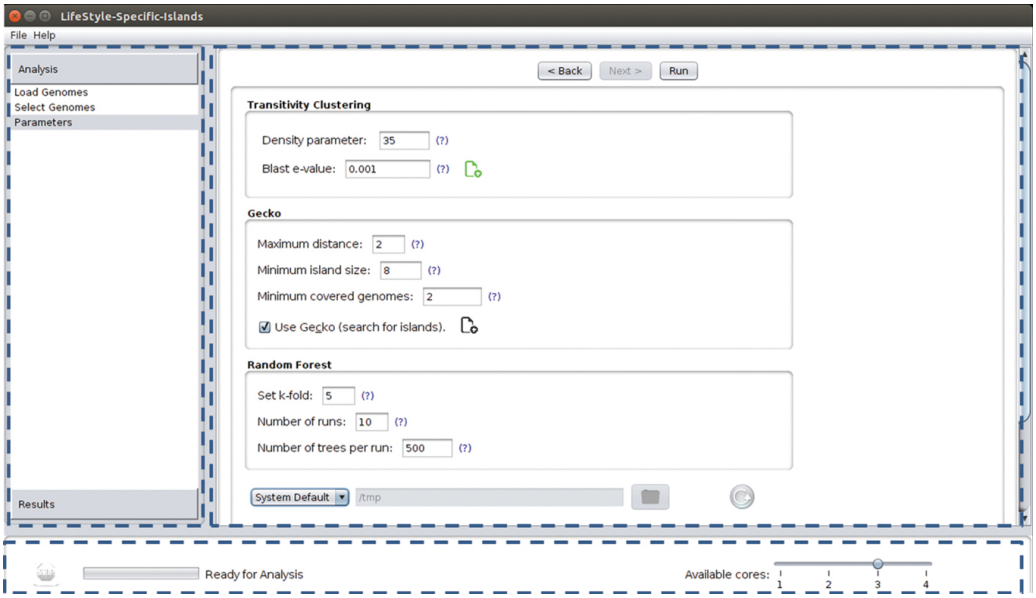

Figure 2: LiSSI software layout. Top-left the selection panel with a tab for Analysis and Results; Top-right the main panel, where all instructions and results will be displayed; and Bottom, the progress panel.

2.2 LiSSI Output Data Analysis

All intermediate results of LiSSI can be inspected individually. We briefly outline them in the following

Homology detection. A corresponding tab summarizes the results found during the homology detection step. It is divided in "Summary" and "Distribution". In the first, one may find basic information about the homology detection process, such as time required for BLAST and Transitivity Clustering. In the latter, one can find a histogram with the cluster size distribution.

Classification. This tab provides all figures created during the classification process. The "Joint Distribution" depicts the distribution of the genetic features (either homologous genes or islands, according to user

selection) among the two lifestyles under investigation. The remaining tabs hold the receiver operating characteristic (ROC) plots depicting the classification performance on three data-sets: 1) the full data set, 2) the data set with a bias towards class "non-pathogenic" (i.e., all features that were mainly found in non-pathogenic organisms), and 3) the data-set with a bias towards class "pathogenic". Each of the plots displays the classification performance using real labels (dark-blue solid line) and using random labels (light-blue dashed line). One would expect a significant drop in performance on random labels, thus a dashed ROC curve being well below the solid one. If this is not the case, one can learn classifier models on randomly labeled data that is as good as the classifier learned from real-label data; rendering it useless. In addition, the distribution of AUC (Area Under Curve) values of the ROC curve for the distinct runs are represented as box-plots. See Figure 3 for an example.

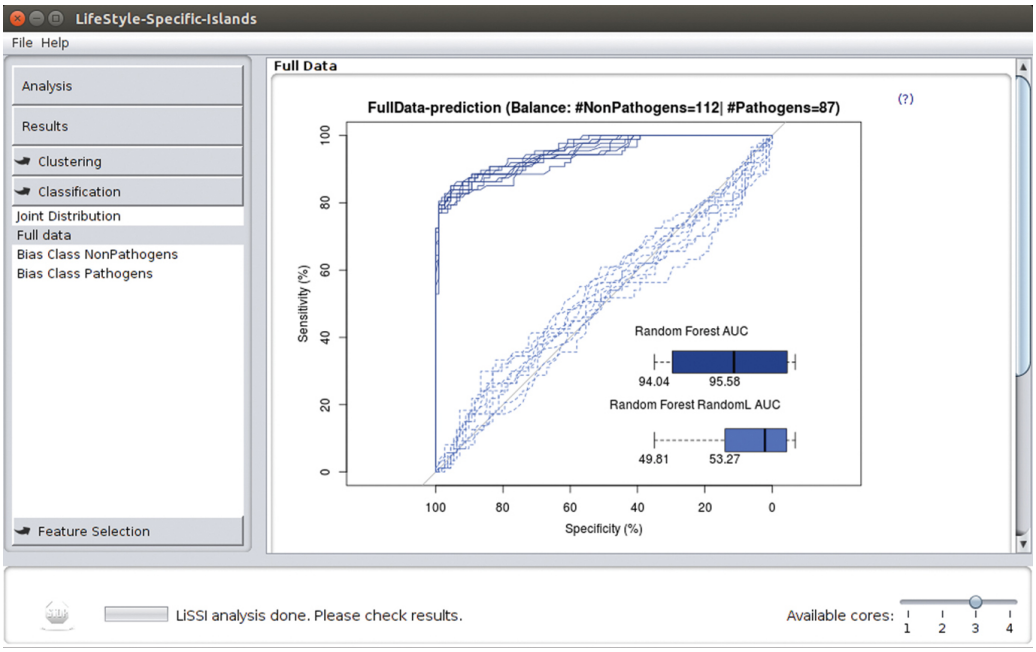

Figure 3: LiSSI: Pathogenicity classification performance results. A Receiver Operating Characteristics (ROC) plots shows the performance of the classification models using genes as features to distinguish pathogenic actinobacteria from non-pathogenic ones. The data was evaluated five times using different 5-fold cross-validation sets to assess the robustness of the classifiers. The real label classifier curves are presented as dark-blue solid lines, while the random label classifiers are depicted as light-blue dashed lines (the ones close to the baseline). The variation of the AUCs (area under curve) in the cross-validation was included in the figure as a box-plot (bottom right). The numbers below each box-plot are the lower and upper quartiles.

Feature Selection. Here, LiSSI presents the decision trees generated subsequent to feature selection. Similarly to the previous tab, it shows figures for the full data set and for each of the two bias directions. By clicking on the nodes one may retrieve more information about the respective genetic feature (homologous genes or islands) or run follow up analysis (Pfam or BLAST). See Figure 4 for an example.

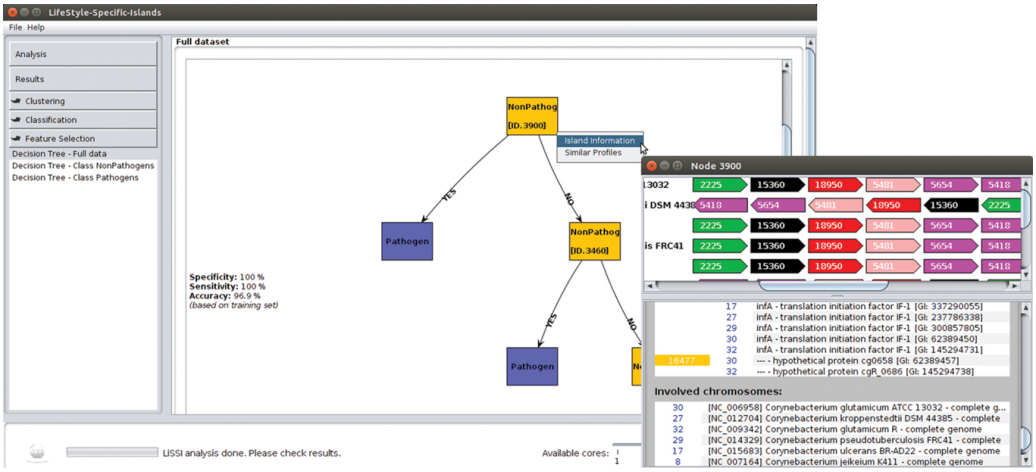

Figure 4: Decision tree created using the most discriminative non-pathogenicity islands. Our classification pipeline (see text) selected the above island as the most representative island for non-pathogens. Nodes containing an identifier represent a genetic feature, in this case, an island. By clicking on the nodes it is possible to visualize the island structure, as well as the gene content and the respective genomes.

3 Application

We selected all 199 completely sequenced Actinobacterial genomes that belonged to at least one of the following lifestyles: aerobes (AE), anaerobes (AN), non-pathogenic (NP), and pathogenic (PA). The annotation for oxygen tolerance was extracted from fusionDB [21], while the pathogenicity annotations were extracted from Barbosa et al. [11]. These labels distribute as follows: 63 AE, 23 AN, 112 NP, and 87 PA. The whole-genome annotation was downloaded from NCBI. For the complete list of species and labels see Supplementary File 1.

We used LiSSI's default parameters for all steps and data sets. The homology threshold of 35 was set to the lower bound of the interval of reasonable values for Actinobacterial species identified in Röttger et al. [20]. We used the following island detection parameters: minimum number of genomes: 2; minimum size: 8; maximum indels: 2. The classification parameters were set to ten runs of 5-fold cross-validation, growing 500 trees per run.

3.1 Pathogenicity

Bacteria can be roughly divided according to pathogenicity classes into non-pathogens and pathogens. This subdivision is an oversimplification of what might actually be a continuum, where the bacteria classification depends not only on intrinsic characteristics but on the host and environmental factors [22].

Pathogens have a higher tendency to gene loss which leads to smaller genomes when compared to non-pathogens, which is commonly explained by the metabolic abundance provided by the host. That eventually leads to the loss of metabolic genes that are no longer under selective pressure [23]. Alternatively, the lack of certain metabolic pathways might actually enhance bacterial virulence [24], [25].

In contrast to pathogens, non-pathogens are exposed to a constantly changing environment, where they need to quickly adapt to extreme changes in salinity or sunlight exposure. Furthermore, their survival depends on the ability to metabolize several sources of nutrients [25], [26], [27], [28]. Therefore, it is expected that these organisms possess a large genetic arsenal ensuring their survival in ever changing environmental conditions.

We applied LiSSI to first find homologous gene sets and genomic islands. We observe homologous genes exclusively found in either pathogens or non-pathogens. We also see islands exclusively found in one of the two classes, PA or NP. We found 375,427 distinct homologous genes, where 317,751 were mainly present in non-pathogens and 57,676 in pathogens. The situation is the opposite for islands. Most of the 465 islands are mainly present in pathogens (386); the remaining 79 are mainly present in non-pathogens. Note that there is no island that is present in more than 35 % of the either, non-pathogens or pathogens.

Afterwards, we applied LiSSI to learn random forest classifiers and decision trees on the best extracted features. The results vary heavily depending on whether we use genes or islands as features. When using genes, the classifiers show good performance for both non-pathogen bias ($\overline{AUC}$ = 94.16 %) and pathogen bias ($\overline{AUC}$ = 93.91 %), as well as for the classifiers using the full data set ($\overline{AUC}$ = 94.81 %). It indicates that we can find gene sets specific for pathogens, as well as gene sets specific for non-pathogens. On the other hand, when using islands as features, the scenario is fairly different. The overall classification performance dropped massively: non-pathogen bias performs poorly ($\overline{AUC}$ = 63.61 %) and pathogen bias as well although not as bad ($\overline{AUC}$ = 88.84 %). See Supplementary File 2 for more details.

3.2 Oxygen Consumption

The presence of atmospheric oxygen is a limiting factor for bacterial growth; specifically, oxygen levels cannot exceed those found in a bacterium's native habitat [29]. Above these levels bacteria are subject to decrease in population growth and ultimately death due to the harmful effects of oxidation caused by superoxide and hydrogen peroxide in cellular components [29], [30]. During oxidative stress, lipids are the major target, leading to alterations in membrane fluidity and potentially disrupting membrane-bound proteins. Further, modifications in proteins can lead to conformational changes and consequently loss of function. Another main target of oxidative stress is the DNA, leading to single- or double-strand breaks and, in extreme cases, blocking replication

by cross-linking the DNA to other molecules [31], [32]. Regarding oxygen tolerance, bacteria can be roughly divided into aerobes and anaerobes.

Aerobes are defined as organisms that require atmospheric oxygen conditions (roughly 20 %) to achieve optimal growth. The overhead associated with an oxidative environment is compensated by enabling aerobic respiration, a pathway substantially more efficient than fermentation [33]. Aside from the presence of a metabolic pathway that can use oxygen as the final electron acceptor, other features are ubiquitous among these organisms, such as enzymes that degrade peroxide (catalases and peroxidases) [29], [34]. Other metabolic features are also expected to be found to prevent oxidative agents formation, plus, mechanisms to repair oxidative damage and eliminate damaged molecules [30].

Anaerobic organisms are defined as organisms that can tolerate only low amounts of atmospheric oxygen and are not capable of performing cellular respiration. Organisms of this class lack the mechanisms for cellular respiration and to protect the cellular components against oxidative damage [35]. It is not clear which genes might be either exclusive or essential for this class of organism [36].

Comparing the two classes using LiSSI, we found 335,532 distinct homologous genes, where 198,529 were mainly present in aerobes and 28,974 in anaerobes. Similarly, most of the 181 identified islands were mainly present in aerobes (107); the remaining 74 were mainly present in anaerobes. Again, the classification results differ between the gene-based and the island-based procedure. Using homologous gene sets, the classifiers had good performance for both biases: aerobe bias ($\overline{AUC}$ = 92.48 %) and anaerobe bias ($\overline{AUC}$ = 99.15 %), as well as for the classifier using the full data set ($\overline{AUC}$ = 95.15 %). On the other hand, using islands as features, the scenario is different. The classification performance dropped significantly: aerobe bias achieved an $\overline{AUC}$ of 66.51 %, and the anaerobe bias yielded an $\overline{AUC}$ of 78.26 %. See Supplementary File 2 4 for more details.

4 Conclusion

We introduced LiSSI, a bioinformatics software for identifying signature genes or islands (conserved consecutive sequences of homologous genes) that distinguish bacterial lifestyles. We illustrate its functionality by identifying genetic features for bacterial pathogenicity and tolerance for atmospheric oxygen. While signature genes could always be detected with high accuracy, islands are harder to identify and less well conserved. Although the use of genomic islands as classification features is a highly valuable function, it is also very vulnerable to test set bias – as the classification performance for tolerance for atmospheric oxygen might suggest.

Note that, in contrast to *de novo* approaches (e.g. PIPS and GIPSy [37], [38]) LiSSI is a comparative approach, thus, dedicated (and limited) to detecting genetic features discriminating between two sets of species, i.e. genes/islands appearing in most of the species of one lifestyle but rarely in any species of the other lifestyle. Genetic elements that are not conserved among the species of one of the two sets may remain undetected.

Acknowledgement

EB would like to acknowledge financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) from Brazil (grant no. BEX12954-12-8). JB receives financial support from the SDU2020 initiative and his VILLUM Young Investigator Grant.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] McCutcheon JP, von Dohlen CD. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol. 2011;21:1366–1372.

[2] Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, et al. Complete genome sequence of the myxobacterium sorangium cellulosum. Nat Biotechnol. 2007;25:1281–9.

[3] Rocha EP. Order and disorder in bacterial genomes. Curr Opin Microbiol. 2004;7:519–27.

[4] Fang G, Rocha EP, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics. 2008;9:4.

[5] Lawrence J. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev. 1999;9:642–8.
[6] Price MN, Huang KH, Arkin AP, Alm EJ. Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res. 2005;15:809–19.
[7] Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics. 1996;143:1843–60.
[8] Pál C, Hurst LD. Evidence against the selfish operon theory. Trends Genet. 2004;20:232–4.
[9] Ochman H, Davalos LM. The nature and dynamics of bacterial genomes. Science. 2006;311:1730–3.
[10] Newton IL, Bordenstein SR. Correlations between bacterial ecology and mobile DNA. Curr Microbiol. 2011;62:198–208.
[11] Barbosa E, Röttger R, Hauschild A-C, Azevedo V, Baumbach J. On the limits of computational functional genomics for bacterial lifestyle prediction. Brief Funct Genomics. 2014;13:398–408.
[12] Winter S, Jahn K, Wehner S, Kuchenbecker L, Marz M, Stoye J, et al. Finding approximate gene clusters with gecko 3. Nucleic Acids Res. 2016;44:9600–10.
[13] Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, Troshin PV, et al. BioJava: an open-source framework for bioinformatics in 2012. Bioinformatics. 2012;28:2693–5.
[14] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
[15] Altschul S, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997:3389–3402. 25.
[16] Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, et al. Partitioning biological data with transitivity clustering. Nat Methods. 2010;7:419–20.
[17] Jahn K. Efficient computation of approximate gene clusters based on reference occurrences. J Comput Biol. 2011;18:1255–74.
[18] Liaw A, Wiener M. Classification and regression by randomforest. R News. 2002;2:18–22.
[19] Diaz-Uriarte R. varSelRF: variable selection using random forests. 2009. Available from: http://ligarto.org/rdiaz/Software/Software.html, R package version 0.7-1.
[20] Röttger R, Kalaghatgi P, Sun P, de Castro Soares S, Azevedo V, Wittkop T, Baumbach J, et al. Density parameter estimation for finding clusters of homologous proteins – tracing actinobacterial pathogenicity lifestyles. Bioinformatics. 2012;29:215–222.
[21] Zhu C, Mahlich Y, Bromberg Y. fusion DB: assessing microbial diversity and environmental preferences via functional similarity networks, 2016. DOI:10.1101/035923.
[22] Ehrlich GD, Hiller NL, Hu FZ. What makes pathogens pathogenic. Genome Biol. 2008;9:1.
[23] Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell. 2002;108:583–6.
[24] Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A. "black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of shigella spp. and enteroinvasive escherichia coli. Proc Natl Acad Sci. 1998;95:3943–8.
[25] Rohmer L, Hocquet D, Miller SI. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. Trends Microbiol. 2011;19:341–8.
[26] Casadevall A. Cards of virulence and the global virulome for humans. Microbe-Am Soc Microbiol. 2006;1:359.
[27] Casadevall A, Pirofski L-a. Accidental virulence, cryptic pathogenesis, martians, lost hosts, and the pathogenicity of environmental microbes. Eukaryotic Cell. 2007;6:2169–74.
[28] Görke B, Stülke J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. Nat Rev Microbiol. 2008;6:613–24.
[29] Imlay JA. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. Nat Rev Microbiol. 2013;11:443–54.
[30] Gutteridge JM. Biological origin of free radicals, and mechanisms of antioxidant protection. Chem Biol Interact. 1994;91:133–40.
[31] Cabiscol E, Tamarit J, Ros J. Oxidative stress in bacteria and protein damage by reactive oxygen species. Int Microbiol. 1999;3:3–8.
[32] Sies H, Menck CF. Singlet oxygen induced DNA damage. Mutat Res. 1992;275:367–75.
[33] Poole RK, Cook GM. Redundancy of aerobic respiratory chains in bacteria? routes, reasons and regulation. Adv Microbial Physiol. 2000;43:165–224.
[34] Pahl HL, Baeuerle PA. Oxygen and the control of gene expression. Bioessays. 1994;16:497–502.
[35] Morris RL, Schmidt TM. Shallow breathing: bacterial life at low O2. Nat Rev Microbiol. 2013;11:205–212.
[36] Müller-Herbst S, Wüstner S, Mühlig A, Eder D, Fuchs TM, Held C, et al. Identification of genes essential for anaerobic growth of listeria monocytogenes. Microbiology. 2014;160:752–65.
[37] Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, Baumbach J. Pips: pathogenicity island prediction software. PLoS One. 2012;e308487.
[38] Soares SC, et al. GIPSy: Genomic island prediction software. Journal of Biotechnology. 2016;232:2–11. DOI:10.1016/j.jbiotec.2015.09.008.

Supplemental Material: The online version of this article offers supplementary material (DOI: https://-doi.org/10.1515/jib-2017-0010).