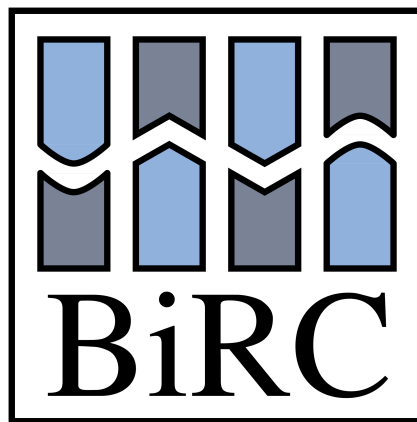


Analysis of the Coevolution of Bacteriophages to a Diverse Bacterial Population



Master's Thesis in Bioinformatics (30 ECTS)

Cornelia Buus and Lotte Hindhede

201509179 | 201406060

Supervisor: Thomas Bataillon

29th May, 2020

Title:

Analysis of the Coevolution of Bacteriophages to a Diverse Bacterial Population

Authors:

Cornelia Buus

Student ID: 201509179

E-mail: corneliabuus@gmail.com

Lotte Aas Hindhede

Student ID: 201406060

E-mail: lottehindhede@gmail.com

Supervisor:

Thomas Bataillon

E-mail: tbata@birc.au.dk

Bioinformatics Research Centre

Aarhus University

C.F. Møllers Allé 8

8000 Aarhus C

This study aims to investigate bacteriophage evolution and their coevolutionary dynamics with bacterial hosts. The data analyzed is obtained from an experiment using the bacteria strain *Streptococcus thermophilus* and its virulent phage 2972. The experiment includes three different experimental setups, which makes it possible to investigate multiple aspects: How bacteria adapt, how bacteriophages adapt, and how diversity affects these dynamics.

The first step was to examine the performance of three different variant callers: FreeBayes, VarDict, and GATK. An initial exploration of the called variants in the context of phage evolution and coevolution with bacteria, was performed. Furthermore, different ways of combining the called variants into genotypes were investigated. When having constructed the genotypes, the evolutionary analysis was conducted. This analysis included the measures; genetic diversity, differentiation, mean pathogen fitness, and mean local adaptation in the bacteriophage populations. Further, these measures were related both to each other and other quantities like bacterial and bacteriophage population size.

From the results we see an excess of mutations arising in regions causing escape from bacterial immunity, i.e. escape mutations, compared to other genomic regions in the bacteriophage. Furthermore, the frequency of these escape mutations tends to increase over time. In regards to diversity and differentiation, it is seen that bacteriophage populations with no initial diversity experience an increase with time. The bacteriophage populations with initial diversity, experience a constant differentiation and a slight decrease in diversity with time. We see an excess of escape mutations in regions targeted by bacteriophage insensitive mutants that experience higher fitness. Furthermore, mean pathogen fitness shows a positive correlation with bacteriophage population size and a negative correlation with bacterial population size. It is also clear that the bacteriophage populations with initial diversity, experience a higher level of local adaptation. The results reveal that the measures of adaptation decrease at the end of the experimental time frame, indicating that the bacterial host have the highest potential to win the coevolutionary arms race.

All these observations support the conclusion that coevolution is occurring between bacteriophage and bacterial populations.

Acknowledgements

Thanks to Thomas Bataillon for supervision and close collaboration. He has been of great help in the discussion of how to best explore the data given and where to focus. Especially thanks for putting us in contact with relevant people providing us with interesting data and algorithms to explore and use in our study.

Thanks to Sylvain Gandon and Enrique Ortega from the research group in Montpellier for the great collaboration in regard to the data analyses, and for many important discussions of these. Moreover, we appreciate that we are given the opportunity to work with data from their experiment.

Thanks to Susan Bailey and Paula Tataru for providing us with a new and interesting algorithm to construct genotypes and estimate their frequencies. This step has been of great help, making it possible to investigate coevolutionary measures, despite the short time frame of our study.

Thanks to Per Kryger for relevant feedback in regard to the content of the thesis and discussion of population genetic aspects.

We are very grateful for the opportunity to work with coevolution to this level of detail in such a short time frame. This could not have been possible without the people mentioned above.

Abbreviations and Nomenclature

nt:	nucleotide(s)
CRISPR:	Clustered Regularly Interspersed Sequentially Palindromic Repeats
Cas:	CRISPR associated proteins
crRNA:	CRISPR RNA
Spacer:	The interspersed segments in the CRISPR locus, complimentary to a protospacer
Protospacer:	30 nt region in bacteriophage genome. A possible target for a spacer
PAM:	Protospacer Adjacent Motif (4 nt long)
Seed:	The first 20 nt in the protospacer closest to the PAM
Escape region:	A 37 nt long region including a PAM and a protospacer. The bacteriophages escape bacterial immunity if variants arise here.
Escape position:	A position between 1 and 37 revealing where a variant is found in the escape region.
BIM:	Bacteriophage Insensitive Mutant
GATK:	Genome Analysis Toolkit
VAF:	Variant allele frequency
PPV:	Positive predictive value
VCF:	Variant Call Format (file format)
CSV:	Comma Separated Values (file format)
AO:	Alternate Allele Observation count
DP:	Read depth

Contents

	Page
1 Introduction	1
1.1 Evolution	1
1.2 Adaptation in Bacteriophages and Bacteria	3
1.3 Genetic Diversity and Differentiation	5
1.4 Measures of Adaptation	7
1.5 From Raw Data to Variant Calling and Genotype Reconstruction	10
1.6 Aim of Study	14
2 Methods	16
2.1 Experimental Setup	16
2.2 Next Generation Sequencing Pipeline	17
2.3 Defining Escape Regions	19
2.4 Inferring Bacteriophage Genotypes and Estimating Frequencies	20
2.5 Measures of Diversity and Differentiation	20
2.6 Jackknife	21
2.7 Modifying Bacterial Genotypes	21
2.8 Measuring Adaptation	22
2.9 Software and Code	22
3 Results and Discussion	23
3.1 Evaluation of NGS Pipeline	23
3.2 Variants Found in Escape Regions	23
3.3 Interpretation of Called Variants	25
3.4 Escape Mutations in Bacteriophage Genotypes	30
3.5 Bacterial Genotypes	32
3.6 Diversity and Differentiation	32
3.7 Adaptation and Coevolution	34

4	General Discussion	39
4.1	Factors Affecting Local Adaptation	39
4.2	The Coexistence of Bacteria and Bacteriophage	40
4.3	How the Experimental Setup Affects the Interpretation of Results Regarding Bacteriophage Evolution	40
4.4	Evaluating Genotype Inference	41
5	Conclusion	43
	Appendices	48
A	Comparison of Variant Callers	49
B	Inferring Phage Genotypes and Estimating Frequencies	59
C	Additional Tile Plots	72
D	Additional Circlize Plots	77
E	Escape and Non-escape tables	81
F	Bacterial Plots	86
G	Population Size	89