GC-biased gene conversion in a *Rhizobium leguminosarum* species complex

Master's Thesis in Bioinformatics

GC-partisk gen-konvertering
i et *Rhizobium leguminosarum*
artskompleks

Carl Mathias Kobel

Supervisors:
Thomas Bataillon &
Maria Izabel Cavassim Alves

Bioinformatics Research Centre
Aarhus University

June, 2020

Contents

Abstract **1**

Master's Thesis Process Summary 2

Introduction **3**

Mechanism of Gene Conversion 3

Methods and Data **7**

Inference of recombination 7

Data and Processing 10

Results & Discussion **12**

Distribution of GC content 12

Comparison of the algorithms 13

Evidence for GC-biased gene conversion 14

Chromosomal distribution of recombination and synonymous GC-content 16

Conclusion **20**

Proposals for further studies 20

Acknowledgements **22**

Bibliography **23**

Abstract

Variation along genomes and across species for genomic G≡C base pair content of DNA-based organisms is not fully explained. It varies hugely between phylogenetic clades and even within genomes. In mammals, yeast and bacteria there is an ongoing debate on whether GC-biased gene conversion (gBGC) might be a major cause behind this variation in GC-content. The gBGC hypothesis entails that gene conversion will incidentally locally bias nucleotide composition towards GC. In this study, we explore whether patterns of recombination and GC content suggest that homologous recombination and gene conversion shape GC content in a set of five sympatric bacterial genospecies of *Rhizobium leguminosarum*.

By analyzing the core genes shared by all genospecies, we found co-variation in recombination and synonymous GC-content (GC3). After validating the inferred recombination parameters by contrasting the results from two fundamentally different methods, we observed a varying relationship between the per-gene rate of recombination and the amount of GC3. We found that the strength of this relationship is largely dependent on the amount of genetic variation (number of informative sites) in each of the five genospecies. This implies either that the data set contains a poor representation of the genospecies population, or that the relationship is more pronounced when there is more gene conversion activity, hence more recombination. We also find that the relationship is not confounded by population structure. We concluded that the relationship between recombination and GC3 might be evidence for gBGC in this representation of *R. leguminosarum*.