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*.

Master's Thesis Process Summary

In this master's thesis in Bioinformatics I started out with the aim to replicate an investigation in the relationship between recombination and GC-content inspired by Lassalle et al. (2015). In order to improve on this study, I wanted to use a new method: *mcorr* (Lin & Kussell, 2019). Unfortunately, this method proved unfit on isolated genes, which is the basis for the experimental setup. This led us to use the same methods as in the original study. So, instead of changing the method, I changed the data set: I received a data set on the coding sequences of the core and accessory genes of 196 samples from 5 genospecies of *Rhizobium leguminosarum* (Cavassim et al., 2020). From this data set I extracted the core genes, and used mainly two software packages, around which I built bioinformatic pipelines, to infer 1) the presence of recombination (Bruen et al., 2006) and 2) the rate of recombination *ClonalFrameML* (Didelot & Wilson, 2015), respectively. Firstly, I compared the inferred parameters from the opposed algorithms. Then I investigated the evidence of a relationship between recombination and synonymous GC-content, and investigated whether this relationship was confounded by population structure. Lastly, I looked for the physical stratification of recombination and synonymous GC-content in the chromosome.

In order to increase the reproducibility of the analyses taken on in this study, I automated most analyses by writing bioinformatic pipelines which can be executed on a high performance computing cluster. The main pipeline that extracts the core genes and infers recombination activity with *PHI* and *ClonalFrameML* is parallelized with *gwf* (grid workflow) and can be accessed at `github.com/cmkobel/gBGC/tree/master/workflow`. All statistics and plotting were performed with *R-tidyverse*.

Introduction

Homologous recombination changes the state of linkage between allelic DNA variants. As selection and drift increase linkage disequilibrium, recombination decreases it. This is an important, and possibly an adaptive process, because it allows more efficient selection for alleles that are linked to less favorable ones (Hill & Robertson, 1966). One phenomenon caused by recombination; gene conversion, has the special effect of transferring variants unidirectionally from one genetic sequence to another. Evidence has come forward that gene conversion is GC-biased in mammals, yeast and bacteria (Duret and Galtier 2009; Galtier et al. 2009; Lassalle et al. 2015). Unfortunately, the underlying mechanism explaining this GC-bias is not yet known, but some are put forward. The hypothesis with the most support is that the donor sequence in gene conversion is more often the GC-rich one (Lassalle et al., 2015). Another hypothesis is that the DNA mismatch repair system (MMR) is biased such that a non-Watson-Crick base pair is more often corrected into a GC base pair (Duret & Galtier, 2009; Lassalle et al., 2015). Galtier et al. (2009) have shown that gBGC can lead to the fixation of deleterious alleles in primates. Because it has been proposed that gBGC is universal throughout all phyla (Lassalle et al., 2015), and because it can impart consequences for the adaptation of an organism, we aim to investigate whether it is present in a species that has not yet been exposed to this analysis - *Rhizobium leguminosarum*.

Rhizobium leguminosarum is a gram-negative soil bacteria in the alphaproteobacteria class. It is characterized by its endosymbiotic relationship with legumes, wherein it fixates atmospheric nitrogen in exchange for carbohydrates provided by the plant. Because the *Rhizobium* genus plays a key role in the nitrogen cycle, we want to explore how it works, also in relation to gBGC.

Mechanism of Gene Conversion

Gene conversion can be initiated by a double strand break (DSB). This DSB might be initiated actively by the *SPO11* protein (Duret & Galtier, 2009), or passively by a chemical agent (free radicals, ionizing radiation etc.). The DSB is then actively repaired either by double strand break repair (DSBR) or by synthesis dependent strand annealing (Figure 1).

The DNA mismatch repair in bacteria, especially in species other than E. coli, is not yet fully understood (Fukui, 2010). The consensus revolves around it being initiated when the MutS enzyme recognizes a mismatch (non-Watson-Crick base pair). The phosphate backbone of one strand (of the double stranded DNA) is incised by the MutL enzyme.

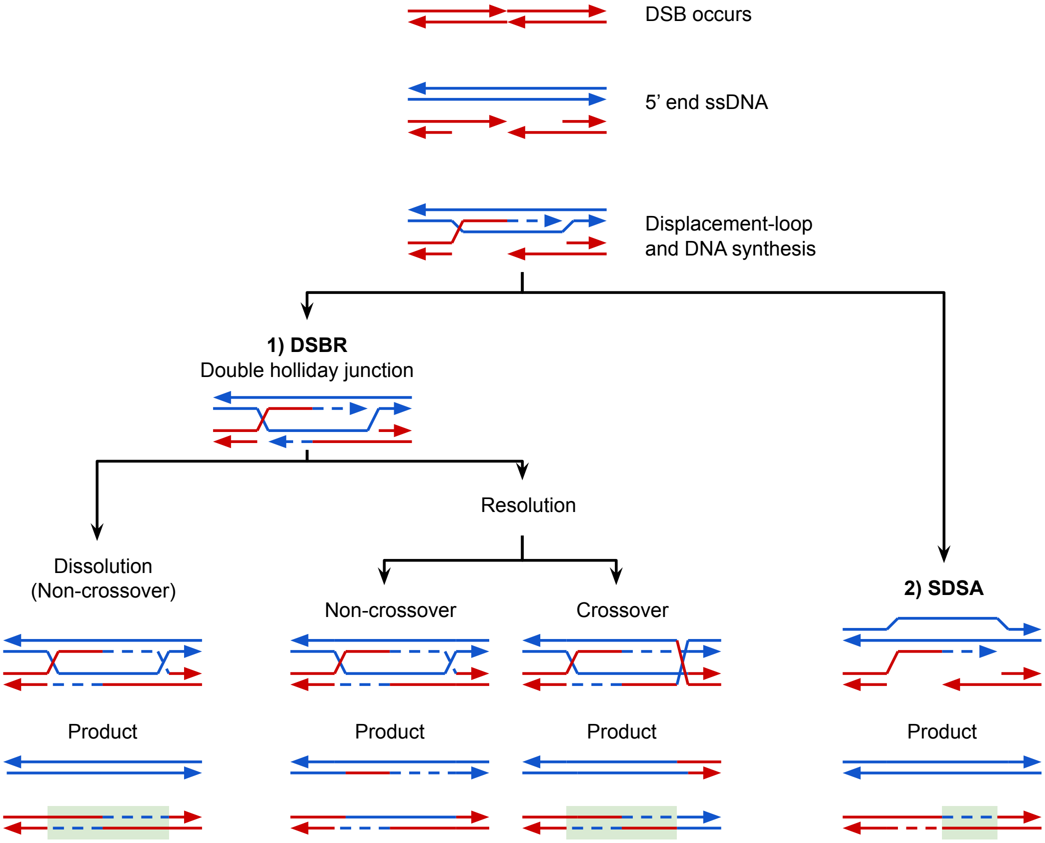

Figure 1: Tree diagram showing the possible ways a double strand break can lead to a gene conversion event. The green boxes in the final products show the areas where heteroduplex DNA is formed. The blue and red DNA depict the donor and recipient sequences, respectively. The arrows show the 3'-ends. Walkthrough from top to bottom: A double strand break (DSB) has taken place in the genomic DNA. The double stranded DNA around the DSB is resected so the 5'-ends are exposed. The red flanking sequence invades the blue in a displacement loop, and DNA is synthesized. Now two possible fates present: 1) Double strand break repair (DSBR) is seeded with the initiation of a double holliday junction where four double stranded DNA sequences are joined twice. The two holliday junctions can be resolved in different ways; they can be dissolved (dissolution), where the donor (blue) strand is left untouched, or resolved (resolution) where either a crossover or non-crossover occurs. This non-crossover product does not transfer unidirectionally, and thus does not lead to gene conversion. DSBR produces two heteroduplex regions distributed on one or two initial sequences. 2) Synthesis dependent strand annealing (SDSA) is carried out, here only one heteroduplex region is produced, hence the amount of gene conversion is half here in relation to DSBR. This figure is adapted from Chen et al. (2007) and Duret & Galtier (2009).

Whether the placement of the incision sites leads to a GC-bias is possible, but unfortunately unknown. Further helicases and exonucleases remove the nicked strand, and DNA Pol III can then synthesize, now possibly GC-biased, double stranded DNA. In a study by Akashi & Yoshikawa, 2013 it was found that a deletion of the mutS enzyme leads to an increase in GC-bias, which indicates the possibility of a link between double strand break repair (DSBR) and genomic GC-bias.

There have been various selectionist views attempting to explain the variation in GC-content across bacterial species. One is that ultraviolet light which catalyzes the formation of thymidine dimers from TpT sites, would induce positive selection on organisms with higher GC content. Another hypothesis was based on the fact that GC base pairs are more thermodynamically stable than AT base pairs. This stability is possibly driven by the

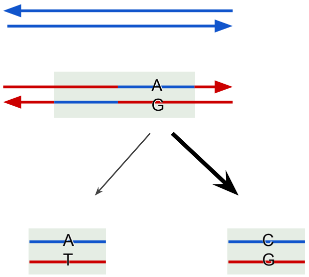

Figure 2: One way gBGC can be biased is that non-Watson-Crick base pairs in the heteroduplex double stranded DNA formed by recombination are corrected into GC base pairs rather than AT base pairs. Figure adapted from Pouyet & Gilbert (2019).

fact that G≡C base pairs are bonded by three hydrogen bonds whereas A=T pairs are only composed of two hydrogen bonds. Yet another attempt to explain GC content variation was based on the observation that bacteria living in aerobic and anaerobic environments would through the use of different metabolic pathways yield different GC-contents, or by enriching the DNA with GC base pairs, the DNA would become more resistant to mutations induced by free oxygen radicals present in an aerobic environments. These aforementioned attempts to explain variation in GC-content from a selectionist point of view have all been refuted (Graur, 2015), and the prevailing view now is that the variation in GC-content stems from a neutral process (Graur, 2015).

In 2009 Duret & Galtier (2009) discussed the presence of gBGC in mammals. They argued that the bias stems from base excision repair (BER) which acts on non-Watson-Crick base pairs which can be introduced by the heteroduplex gene conversion tracts. They attributed the GC-bias as an adaptation to counteract the consequences of the high mutation rate of CpG sites (5-methylcytosine) (Brown & Jiricny, 1987).

Around 2010 Hildebrand et al. (2010) noted that the variation in GC-content was up until that point in time ascribed to different patterns in mutation. They investigated this by examining synonymous genetic variation in a diverse collection of bacterial species. They observed an increased rate of segregation of AT-biased substitutions, especially in GC-rich bacteria. If mutation is AT-biased, then GC-rich bacteria would over time become AT-rich. This is clearly not the case for many GC-rich species, since it would not allow them to exist. Hence they concluded that there must be an opposing mechanism that substitutes reversely (GC-bias), and thereby maintains the equilibrium between the content of the two Watson-Crick base pairs (GC and AT). Because optimal codons often are AT-rich and because GC-bias is observed in data sets without recombination, they concluded that this counteracting mechanism might not be gene conversion, but rather selection.

Bobay & Ochman (2017) investigated the possible drivers of GC-enrichment by measuring the mutation spectrum of polymorphisms in GC4 sites in 91 bacterial species. They found an increased GC->AT substitution rate, both in recombining and non-recombining genes. They suspected that selection plays a role in GC enrichment and observed a significantly higher dN/dS ratio of substitutions incorporated by recombination, compared to that of non-recombinant polymorphisms. On this basis, they concluded that selection is the key driver of GC content. However, taking the literature on the topic into account, it is striking that they fail to measure any relationship between GC-content and recombination. This might be because the mutation signal is many times stronger than the possible gBGC signal. Not taking this into account might lead to the erroneous conclusion that gBGC plays no role in shaping GC-content.

On the other hand, Lassalle et al. (2015), while looking at a diverse set of bacterial species, found no link between optimal codons and selection. In addition, they observed a relationship between synonymous GC-content (GC3) and the presence of recombination. Because of these observations, they argued, opposingly to Hildebrand et al. (2010), that gene conversion might in play a role in GC-content variation. They further showed that an equivalent amount of variation in GC3 can be explained by the crossover rate in human data, which they carried forward to suggest that GC-biased gene conversion may be roughly equivalent in human, further implying that gBGC is an "ancestral feature of cellular organisms", which underlines their allegation that gBGC is a generality across all or most phyla. Lassalle et al. (2015) concluded that it is important to take gBGC into account, because it can confound signals of selection, implying their disagreement with Hildebrand et al. (2010).

Bobay & Ochman, 2017; Yahara et al. (2016) have since criticized Lassalle et al. (2015) by arguing that the classification of genes into being either recombining or non-recombining is too dichotomous. Allegedly because it doesn't take the full spectrum of mutation rates into account. However, we will meet this critique by supplementing with a method that instead measures the recombination rate (*ClonalFrameML*).

Methods and Data

The most prevalent hypothesis explaining the variation in GC-content is tied to gBGC. Gene conversion in itself is computationally infeasible to infer in large data sets. However, because gene conversion is caused by recombination which is more inferable and because the topological pattern of gene conversion is equivalent to recombination (from a phylogenetic perspective), we will use recombination as a proxy directly for gene conversion. In order to validate the inferred recombination parameters; we used two main software packages for inferring recombination activity; so we could compare their results. One is based on the notion of compatibility in the genealogical history between samples and whether recombination is necessary in order to explain this history. The other is based on a maximum-likelihood reconstruction of the ancestral sequence with subsequent expectation-maximization of the recombination parameters. These, we will present in more detail in the following to sections:

Inference of recombination

Pairwise Homoplasy Index (*PHI*)

The pairwise homoplasy index is based on the four gametes test. Imagine two sites with two alleles each. The number of possible allelic combinations is in this case four. If we have sequence data for at least four samples, and we observe all four possible combinations, we know that either recurrent mutation or recombination must have occurred (Figure 3).

If we assume the infinite sites model, which states that no recurrent mutations can occur, and we observe an allelic combination that requires recurrent mutation or recombination, we have evidence that a recombination has indeed occurred.

The pairwise homoplasy index (Bruen et al., 2006) uses this notion of compatibility or, conversely, incompatibility between sites. The distinction is binary such that a site can be either compatible or incompatible. Two sites i and j are compatible if a parsimonious genealogical history exists such that no recurrent mutations or homoplasies are present. Let $|\chi_i|$ and $|\chi_j|$ denote the number of alleles at the sites i and j respectively. If we enumerate all possible trees for our samples then the incompatibility score $l(\chi_i, \chi_j)$ represents the minimum number of mutations that explain the genealogical history. This is equal to the number of colored arrows in Figure 3 II. Because each allele must arise at least once $l()$ has a lower bound of $(|\chi_i| - 1) + (|\chi_j| - 1)$. If we subtract this lower bound from $l()$ we get a more useful function: $i(|\chi_i|, |\chi_j|) = i(|\chi_i|, |\chi_j|) - (|\chi_i| - 1) - (|\chi_j| - 1)$. This function we will refer

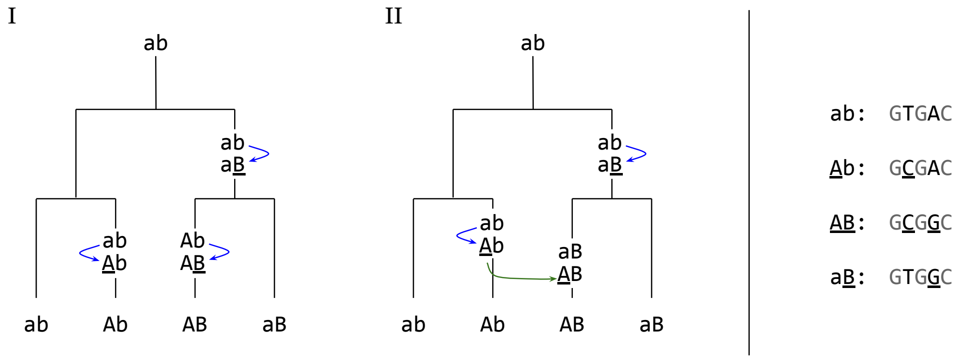

Figure 3: Illustration of two possible genealogical scenarios explaining the same data. Blue arrows denote substitutions, and the green arrow denotes a recombination event (possibly gene conversion). I: Recurrent mutation (homoplasy), assuming no recombination. II: Recombination, assuming the infinite sites model where no recurrent mutations can occur. The right hand side of the figure shows an example of genomic DNA might actually look like. This figure is adapted from Bruen et al. (2006).

to as the refined incompatibility score. $i()$ can be thought of as the non-trivial number of mutations, that is; the number of mutations aside from the mutations necessary to create the alleles in the first place. $i()$ turns out to be equal to the number of homoplasies in the sample-set assuming no recombination. Or conversely; equal to the number of recombination events in the assumption of the infinite sites model. Thus $i()$ is more informative than $j()$ in terms of recurrent mutations or recombination events.

In each equally sized window w on the genome, the proportion of informative sites q together defines $k = wq$. The window size has empirically through simulations been set to $w = 100$, where the recombination state inferred by *PHI* and the known state of the simulated data set are sought to correlate.

By normalizing the refined incompatibility score $i()$ for all neighbouring $i(\chi_i, \chi_{i+1})$ and double neighbouring sites $i(\chi_i, \chi_{i+2})$, we calculate the Pairwise Homoplasy Index as $\Phi_w = c\sum_{j=1}^{k}\sum_{i=1}^{n-j} i(\chi_i, \chi_{i+j})$. Here c is a normalizing factor dependent on k, and n, the last of which represents the number of informative sites in the window.

Because *PHI* doesn't infer the rate of recombination, but rather whether it has occurred or not, we need to test for statistical significance instead. We can calculate a monte carlo p-value of Φ_w by permuting the informative sites on the genome: The informative sites are ordered randomly, and Φ_w is calculated with no regard to the random ordering. This is repeated until a the proportion of values of Φ_w exceeds the significance threshold $\alpha = 1/20$ – Or repeated until a maximum number of iterations has occurred. If this proportion did not exceed α in the maximum number of iterations, we know that the p-value is below α. Thus the occurence of recombination is statistically significant.

ClonalFrameML

ClonalFrameML (Didelot & Wilson, 2015) is based on *ClonalFrame* which was developed by Maiden et al. 1998. Whereas the original version was meant to be used on relatively sparse MLST-data and uses Bayesian inference, *ClonalFrameML* instead uses maximum likelihood on a simpler set of assumptions in the phylogenetic tree that represents the data. This shift is made in order to decrease the running time of the algorithm, such that more data, possibly complete genomes, can be processed.

ClonalFrameML is guided by a phylogenetic tree which is inferred from the core genome of all samples. This we inferred with *RAxML-NG* (Kobert et al., 2014) on the concatenated core genome of each genospecies. This tree is regarded as the best possible representation of the clonal genealogy. On this tree, the maximum likelihood ancestral amino-acid sequence reconstruction on each internal branch is inferred using a sped up exhaustive search algorithm (Pupko et al., 2000). The recombination rate and branch lengths are then estimated using Baum-Welch Expectation-Maximization. For each site, it is inferred whether the variation is from within or outside the species. This is sampled using the Viterbi algorithm. Lastly, variance in the parameters are measured with bootstrapping.

ClonalFrameML assumes that the recombination parameters $R/theta$ (recombination per site per mutation rate per site), delta (mean length of recombination) and nu (relative recombination incidence) are identical for all branches, and also that the length of the branches are proportional to the number of mutations on them. *ClonalFrameML* is implemented as a hidden markov model. On each branch of the clonal genealogy a site can either be affected by recombination (imported) or unaffected by recombination (unimported).

Measuring synonymous GC-content

We seek a way to measure GC-bias under neutrality so it is not confounded by selection. Because bacteria have only 10-15% intergenic regions (Thorpe et al., 2017), we prefer to obtain the measurements from coding genes.

The way we measure synonymous GC-content is by measuring the relative GC content of the third codon. The reason why we can do this is based on the fundamental degeneracy of the genetic code. The degeneracy is implied by the fact that only 20 amino acids are encoded by 64 possible different tri-nucleotide codons (or triplets). On the third codon position, 18 of the 20 amino acids (all except methionine and tryptophan) can be encoded by either; one nucleotide from the A=T base pair, or; one nucleotide from the G≡C base pair. This means that a conceivable GC-bias will have room to present itself here. A more strict measure is based on the fact that 7 of the 20 amino acids can be encoded with any of the four nucleotides on the third position – hence the notion of four-fold degenerate codons, also referred to as GC4 abbreviated from $GC4_{\text{fold}}$.

Deciding whether to use GC3 or GC4 when measuring synonymous GC-content in sequences is a compromise between sensitivity and specificity for GC-bias. GC3 gains sensitivity by taking more sites into account, but does so with the downside of being dependent on the AT→GC substitution to hit the correct strand in order to code synonymously. GC4, on the other hand, takes fewer sites into account which are then more sensitive to GC-bias, but the reduction of the number of sites yields an increased standard error in the measurement.

GC3 and GC4 are used interchangeably in the literature. Lassalle et al. (2015) used GC3 whereas Hildebrand et al. (2010) and Bobay & Ochman (2017) used GC4.

Substitutions or GC-bias in the first and second codon positions are never synonymous. Variation in GC1 and GC2 must therefore be attributed to factors that are linked to the function of the genes the sequences encode i.e. selection.

Data and Processing

The *Rhizobium leguminosarum* DNA data (Cavassim et al., 2020), kindly shared to me by Maria Izabel Cavassim Alves, contains 4068 orthologous genes present in 196 strains (core genes). The strains are divided into five genospecies based on a standard prokaryotic classification measure based on a pairwise average nucleotide identity criteria (ANI threshold > 0.95). Though the data was conveniently shared to be in aligned genomes splitted by coding genes, I will briefly outline how the preprocessing was performed according to the source article:

The data set I received

Paired-end Illumina sequencing was performed on a collection of 196 strains which stem from four geographical locations throughout Western Europe (Denmark, France and the UK). The reads were subjected to *de novo* assembly. Contigs with a low *k*-mer coverage were discarded. By also sequencing and assembling a handful of the strains with long-read technology (PacBio), it was possible to create reference genomes with longer contiguous sequences. On these longer assemblies, a list of syntenic genes was compiled, totalling 3215 orthologs. With the long-read references in mind, scaffolds could then be constructed from the Illumina contigs. Chromosomal contigs were identified by the presence of conserved genes, especially *dnaA*. Plasmids were identified by recognizing the differing *repA* genes. Scaffold sequences were joined using an arbitrary 20 “N” spacer until a complete circular chromid was represented. Chromosomes were started at the *dnaA* gene which normally represents the replication origin. Plasmids were started with the *repA* gene which they carried.

In this data set, the genes are placed into the chromosome and 3 major plasmids. The plasmids sorted by decreasing size are named: pRL12, pRL11 and pRL10 (Young et al., 2006). Because the chromosome contains the majority of genes, and because the order of the genes on the chromosome are more conserved (syntenic), we mainly focused on the genes linked to the chromosome.

Genosp.	#strains	GC	GC3	# core genes (∑inf. sites)				∑genes
				chromosome	pRL12	pRL11	pRL10	
A	32	60.6 %	75.9 %	3’301 (145718)	346	285	137	4069
B	32	60.7 %	76.2 %	3’300 (41606)	346	285	137	4068
C	116	60.6 %	76.0 %	3’300 (178092)	346	285	137	4068
D	5	60.7 %	75.9 %	3’300 (13112)	346	285	137	4068
E	11	60.6 %	75.8 %	3’300 (51992)	346	285	137	4068

Table 1: Overview of the sample size and number of core genes for each genospecies and chromid. Symbols: # means “number of”, ∑means “sum of”. The number of informative sites is only included for the chromosome since we mostly worked with that.

In order to make the workflow more reproducible and manageable, I wrote a pipeline based on *gwf* which is a *python* workflow module that interfaces with the slurm workload manager used on the GenomeDK HPC (`github.com/gwforg/gwf`). The workflow backend computes a dependency graph between individual jobs, and enqueues them to the workload manager in the optimal order. My main workflow takes an aligned core genome in extended multi-fasta format, and a phylogeny which is computed using *RAxML-NG* (version 0.6.0) (Kobert et al., 2014) using a general time reversible substitution model. The workflow isolates all the core genes. For each gene, the recombination activity is then inferred with *PHI* (Bruen et al., 2006) and *ClonalFrameML* (Bruen et al., 2006; Didelot & Wilson, 2015) and the GC content is measured with a custom *python* script (`github.com/cmkobel/gBGC/blob/master/workflow/script/xmfa_gc.py`). When all parameters have been inferred, the results are collated into a single file, which can be imported in an R-environment (Wickham et al., 2019) for data processing. The workflow is repeated independently for each genospecies.

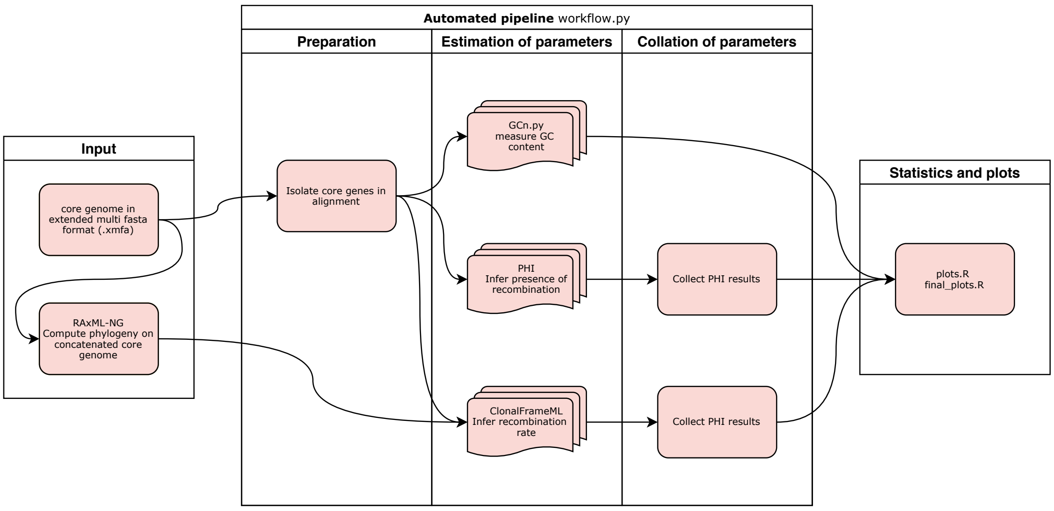
Automated pipeline workflow.py
Preparation
Estimation of parameters
Collation of parameters
Input
core genome in extended multi fasta format (.xmfa)
RAxML-NG Compute phylogeny on concatenated core genome
Isolate core genes in alignment
GCn.py measure GC content
PHI Infer presence of recombination
ClonalFrameML Infer recombination rate
Collect PHI results
Collect PHI results
Statistics and plots
plots.R final_plots.R

Figure 4: Outline of the workflow that computes the results for this study.

The linked github repository contains all necessary scripts to set up the analysis from scratch: `https://github.com/cmkobel/gBGC/tree/master/workflow`. However, the code is not portable in the sense that it will not work immediately on a different system setup than the one used in this study (GenomeDK).

Results & Discussion

Distribution of GC content

To explore the links between GC-content and recombination, we first investigated its distribution. The distribution of GC-content on each codon position (GCn) is equivalent in each genospecies. For that reason we will discuss it on a general above-genospecies level. Because of the structure of the genetic code, we expect to see different distributions of GC-content on each codon position (GC1, GC2 and GC3). We found that the distribution is reminiscent of what is observed in E. coli (Lawrence et al. 1997); which is that the GC2 is lower (median) than GC1 and GC3, and also that GC3 is negatively skewed (Figure 5). Our measurements also closely follow the distribution in another study on *Rhizobium* species (Kogay et al., n.d.), where the median falls within one quartile of what we observe. In Kogay et al. (n.d.), they also find a negative skew in GC3.

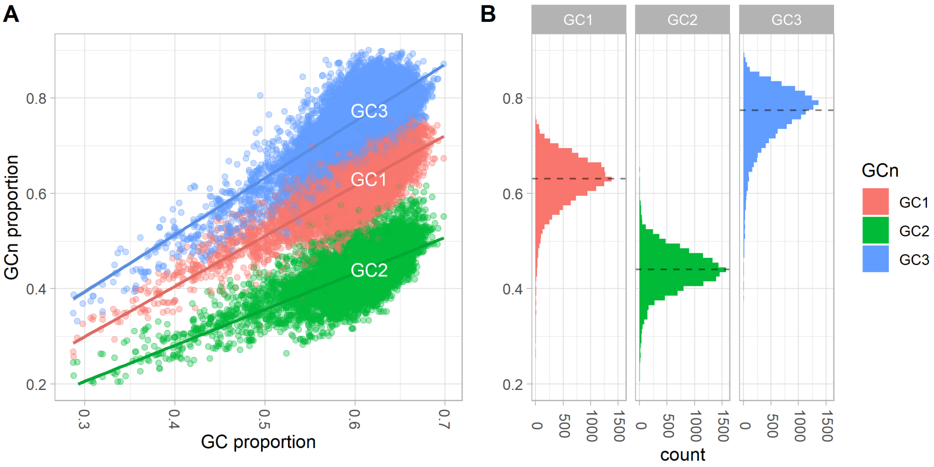

Figure 5: Overall proportion of GC and proportion of GC at each codon position (GCn) in the 5 genospecies pooled together. Each point represents a gene from one genospecies. A: Relationship between overall GC and codon-positional GC (GCn): Here the correlation is emphasized with linear model fits. The R^2-values for these fits are GC1~GC: 0.62, GC2~GC: 0.46 and GC3~GC: 0.57. B: Distribution of GCn shown with a histogram where binwidth = 0.01. The medians are GC1: 0.63, GC2: 0.44, GC3: 0.77. The overall median GC content is 0.62. The horizontal dashed line denotes the median value of the GC-content in each codon position.

The distribution of GC3 might be skewed because of newly horizontally transferred genes which have not yet been ameliorated to match the background GC-content of the sampled genome (Lawrence & Ochman, 1997).

Qualitatively, we note that the distributions of GC2 and GC3 are effectively non-overlapping (Figure 5 pane B). Also that the GC1, GC2 and GC3-content is strongly correlated with the overall GC content. (Figure 5 pane A).

Comparison of the algorithms

We measure recombination as a proxy for gene conversion. Recombination yields an incongruence in the genealogical history which is easier to infer than a gene conversion tract specifically. The variance of the estimates of recombination are already a major bottleneck in this application. This indicates that the inference of gene conversion will most likely yield an even noisier set of parameters. Hence, we use measures of recombination along the genome as a proxy for the expected intensity of gene conversion. For that, we opted to use two fundamentally different packages: *PHI* and *ClonalFrameML*. In order to validate the results of inference, we compared the results of these two: *PHI* measures the statistical significance of recombination (p-value), whereas *ClonalFrameML* measures the rate of recombination relative to the rate of mutation (*R*/*theta*). We will refer to both of these parameters generally as recombination activity throughout this text.

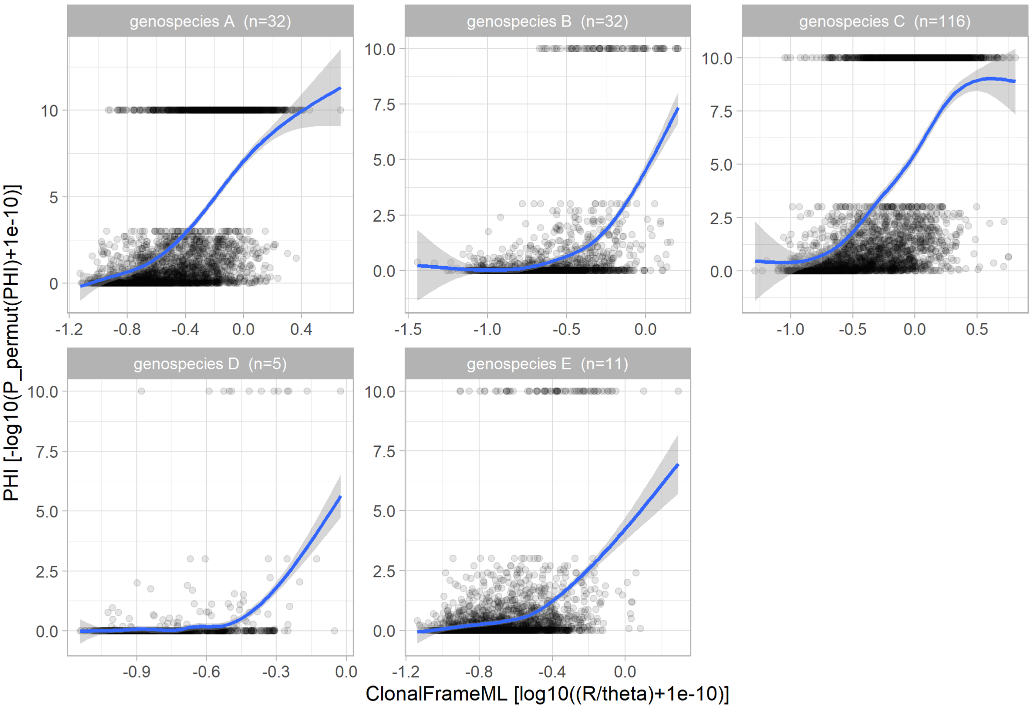

Figure 6: Comparison of the recombination measurements of core genes in the chromosome for all genospecies. Each point represents a core gene. The horizontal axis depicts log transformed *R*/*theta* where *R* is the recombination rate per site, and *theta* is the mutation rate per site. The vertical axis depicts the minus log transformed p-value from a permutation test on the significance of recombination in each gene. 1e-10 has been added to both measures before performing the log-transformation. This is done in order to not exclude zero-values. The lines fitted are cubic splines.

The distribution of p-values for the null hypothesis (H_0) of no recombination calculated in *PHI*, which are computed using a permutation test, are highly polarized in the sense that the distribution is bimodal. This means that most values fall in either a p-value close to zero or close to one, where no values fall in-between. This, combined with the more uniformly distributed recombination rates inferred with *ClonalFrameML* yields a logistic relationship, which can be observed in Figure 6. The fact that the relationship between the recombination activity inferred with the two measures tends to be positive, shows that the algorithms are measuring the same signal, or at least something that is confounded on it.

Evidence for GC-biased gene conversion

From both *PHI* and *ClonalFrameML* we found evidence that variation in recombination partly explains variation in synonymous GC-content. With *PHI* we binned the synonymous GC-content variation into 20 equally sized bins and counted the number of significantly recombining genes. In order to adjust for multiple testing, we divided the significance threshold α=0.05 by the number of genes in each genospecies.

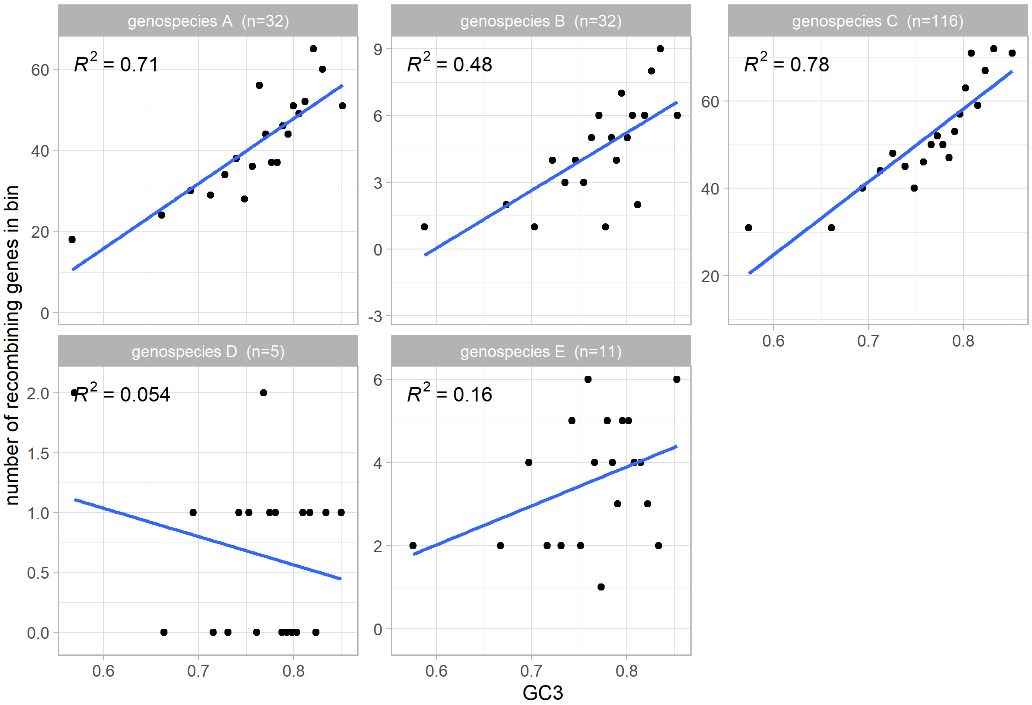

Figure 7: Number of genes with a statistically significant signal that they are recombining for each 20 bins with roughly 203 genes each. The binning is performed on the GC3 feature. Each point represents a bin. For each gene from the chromosome of the five genospecies, the presence of recombination was inferred with *PHI*. A linear model was fitted and the R² was calculated on the shown data points. *n* in each genospecies-pane shows the number of sampled strains.

According to *PHI*; in four of the five genospecies, the relationship between binned GC3 and the number of recombining genes is positive with an R² of 0.16 to 0.78 (Figure 7). Conversely - In genospecies E, which has a sample size of 5 isolates, the relationship is negative, although with a low degree of explanation (R²=0.073). The R²-values indicate what proportion of the variance in a variable can be explained by another variable. In this case these values indicate how much of the variation in GC3 can be explained by variation in binned recombination activity.

Genospecies C represents the largest sample size (n=116). It comprises 3 structured groups in which strains were isolated from clover samples collected in French conventional fields, Danish conventional fields and Danish organic fields. In order to investigate whether this grouping was confounding the relationship between recombination and synonymous GC-content, we executed a second analysis on these three groups. We re-inferred *PHI* on the core genes separately and the R^2-values were 0.82 (Denmark organic field), 0.21 (Denmark conventional field) and 0.58 (France). The fact that much variation of GC3 can explain variation in recombination in all groups leads us to conclude that the structural grouping is not a confounding factor. The other genospecies comprise multiple geographical groups as well, but because the sample size there is limited, we decided to not delve further into their geographical grouping.

According to *ClonalFrame*, the relationship between binned recombination and GC3 is much the same as with *PHI*. Because the distribution of *R/theta* (output from *ClonalFrameML*) is skewed (negatively), we use the median as the representative for each bin.

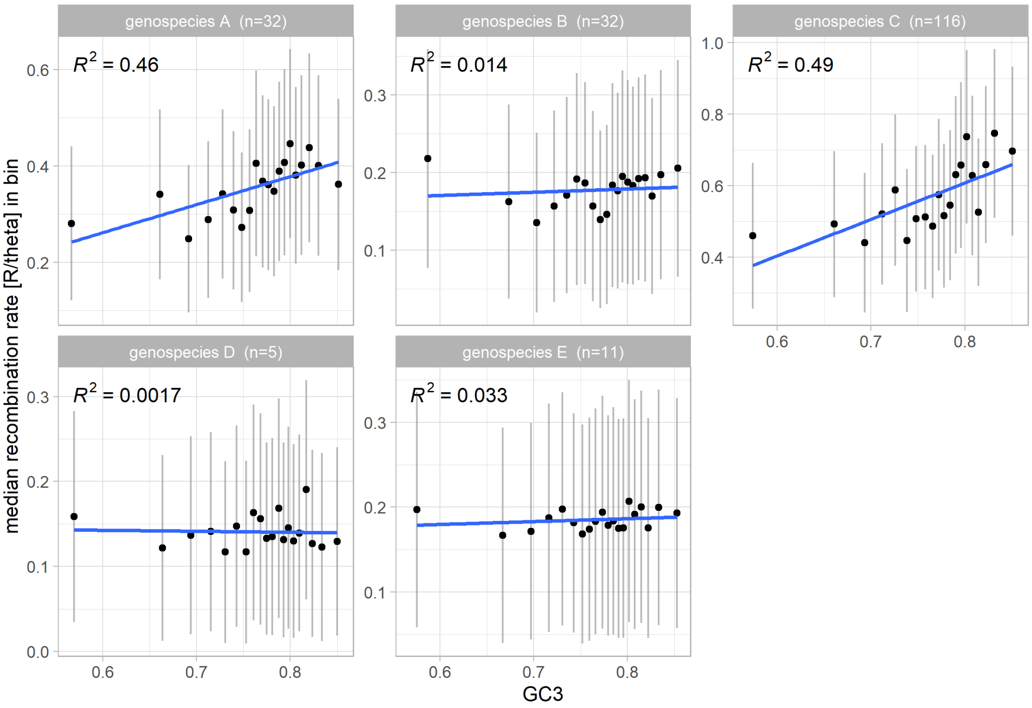

Figure 8: Median recombination rate for each 20 bins with ~203 bins each. Each point represents a bin. The binning is performed on the GC3 feature. For each gene in the chromosome of the five genospecies, the recombination rate was inferred with *ClonalFrameML*. A linear model was fitted and the R^2 was calculated on the shown data. *n* in each genospecies-pane represents the number of sampled isolates. The error bars indicate the median bootstrapped standard deviation in each bin.

The R^2-values follow the same pattern as with *PHI* where they are high for genospecies A, C, less so for B, E and slightly negative for D. This is congruent with the observation that there is a positive relationship between *PHI* and *ClonalFrameML* (Figure 6). The error bars shown in Figure 8 are from bootstrapped replicates of the genes. In each bootstrap replicate a random subset of sites are taken into account so as to measure the standard error in the recombination parameters. The standard deviation is bigger than the difference in signal from the lowest to highest gene bin. This shows that the signal is noisy. The fact that the signal is noisy might indicate that the detection of gBGC may be sensitive to the selection of method. That is, if the method is too sensitive to this noise, the GC3/gBGC relation might be shadowed by it.

The fact that the R^2-values tend to increase with sample-size suggests that the sample-sizes for genospecies besides genospecies C are insufficient, or alternatively that the gBGC activity is less pronounced when less recombination is present.

Adding more samples to the data set doesn't necessarily extend the amount of signal there is to retrieve about recombination. If we add a clone which has a limited amount of additional informative sites than the data set already included in the first place, the sample size is not truly increased but rather; inflated. In order to increase the sample size most efficiently, we should add isolates that have different informative sites – In other words, get a better representation of the population. In Table 1, it is clear that the sum of informative sites is not directly proportional to the number of samples for a given genospecies. For instance, genospecies A and B have the same amount of samples (n=32), but the sum of informative sites is wildly differing: 3.5 times as many in genospecies A (145718) than in genospecies B (41606)

It seems that the the linear relationships (Figure 7 and Figure 8) are linked to the sum of infinite sites in each genospecies. This indicates that genospecies B is relatively more clonal than genospecies A, and might explain why the relationship between GC3 and recombination rate is lacking – because there is not enough signal of recombination in the data set.

Genosp.	# strains	$\sum$inf. sites	R^2 (*PHI*)	R^2 (*ClonalFrameML*)
A	32	145718	0.71	0.46
B	32	41606	0.48	0.01
C	116	178092	0.78	0.49
D	5	13112	0.05	< 0.01
E	11	51992	0.16	0.03

Table 2: Outline of results from linear regression showing the proportion of variance that can be explained by the linear regression on the binned results from each software package. For *PHI* refer to Figure 7, for *ClonalFrameML* refer to Figure 8.

Chromosomal distribution of recombination and synonymous GC-content

In order to investigate how gBGC shapes the GC3-content, we investigated how these parameters vary along the chromosome. If recombination rate and GC3-content follows each other spatially, it might confound the signal for gBGC. By contrasting the spatial signatures of these parameters, we will investigate how they can affect the signals of gBGC.

Because the genome assemblies for the chromosome are divided into three scaffolds, we used a long-read (pacbio-based) reference (SM3). We mapped all genes from the three scaffolds of the chromosome to this reference, and extracted the chromosomal position. This gene mapping was computed with *Exonerate* (Slater & Birney, 2005) using default settings. Because all genes in the data set are syntenic, mapping them to a single reference provides a good representation of where in the genome the genes are located.

On a per gene level, the distribution of recombination and synonymous GC-content throughout the chromosome looks uniform, but when averaging from ~80 genes and upwards, some structuring presents (Figure 9).

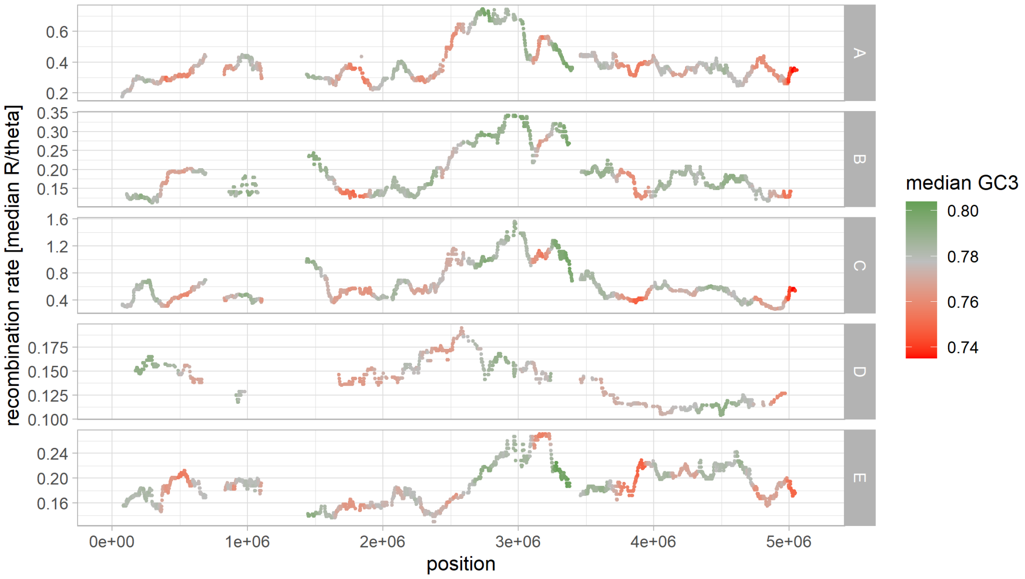

Figure 9: Rolling window of 100 genes on the chromosome in each genospecies. Y-axis: Median recombination rate for each rolling window. Recombination rate is inferred with *ClonalFrameML*. The color shows the mean GC3 for each rolling window scaled such that the grey color corresponds to the median value of the shown data. Because it might be relevant to explore how the structuring changes with different window sizes, we have created an interactive version at `https://cmkobel.shinyapps.io/Rleg_SM3ref`.

This structuring might be due to nucleoid structuring, particularly due to supercoiling (Rocha, 2008). The DNA is protected in supercoiled structures, which may be unwound for transcription, which possibly exposes the DNA to factors such as gene conversion etc. The supercoiled structures are reported to have a size up to around 100kbp. As the average chromosomal gene length of this *R. leguminosarum* data set is close to 1kb (937bp), observing a distinct pattern of structuring in Figure 9 starting around 80-100 genes corresponds roughly to the reported size of these nucleoid supercoil-structures in other species (Rocha, 2008).

Varying the rolling window size (see the Figure 9 description) reveals that the highest rates of recombination are placed around 3Mb. However, we do not observe a link between recombination rate and GC3.

We failed to confirm the location of the origin(s) of replication in the circular chromosome. The replication origins are supposed to be located in the beginning of the assembly. However, according to the genbank *R. leguminosarum* reference assembly, a single gene for the chromosomal replication initiator protein, *dnaA*, is located on the chromosome (genbank gene id: 47499913).

In Figure 10 we bring an alternative representation of the distribution of GC3 and recombination along the chromosome. For both GC3 and recombination rate (*R*/*theta*) we first centered the measurements to a mean equal to zero. Then we took the cumulative score from the first to the last position on the chromosome. It should be noted that the maximum value (amplitude) of this measure depends on the structuring of this centered value (Daubin & Perrière, 2003). The motivation for this visualization is that it shows the cumulative G+C skew on synonymous positions. The skew in GC3 content might be explained by codon usage which is correlated with tRNA abundances (Daubin & Perrière, 2003). Genes in the beginning of the leading strand are more conserved between species, at least in enterobacteria (Daubin & Perrière, 2003). This suggests that these genes that are important early in the replication and hence are made available for transcription earlier in cell division.

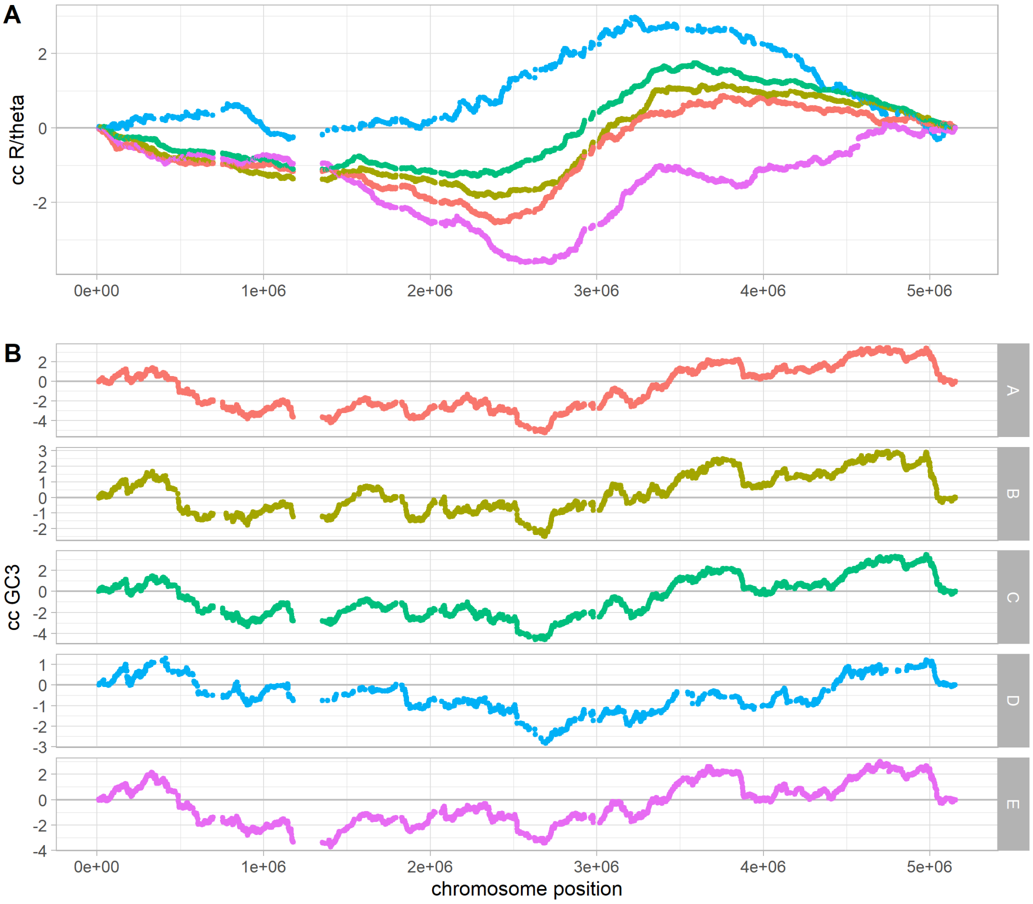

Figure 10: A: Cumulative sum of centered recombination rate (R/theta). B: Cumulative sum of centered GC3, scaled to SD = 1 for clarity. Each point (most are indistinguishable) represents a gene from one genospecies. The x-axis in both panes is the chromosome position in the SM3-reference strain. The same colors represent the five genospecies in both panes (A and B).

This visualization (Figure 10) is unit-less in the sense that it shows structuring of GC-content without the need to choose a window size as in Figure 9. It highlights the regions of the chromosome where all consecutive genes have polarized values of GC3 or recombination rate. In Figure 10 pane A we observe that the cc *R*/*theta*, in varying degrees, has a sinusoidal signature for all genospecies. This is especially the case for genospecies C (green) where we have the best parameter estimations due to a higher sample size. The chromosome can be divided into parts where the rate of recombination is monotonically increasing or decreasing from the mean value. In Figure 10 pane B we observe a shared signature between all genospecies. There is a tendency for the cc GC3 to follow the levels of the sinusoidal cc *R/theta*. Yet in many regions, the two signatures disagree.

Observing that cc GC3 and cc *R/theta* agree on the spatial signature might suggest either that gBGC is dependent on the chromosomal position or contrastingly that gBGC is confounded by the chromosomal position. However, because we don't observe a strong correlation between recombination and GC-content on the chromosome level, we conclude that neither is the case.

We know that the *dnaA* replication origin gene is probably placed in the beginning of the assembled chromosome. In Figure 10 we don't observe a specific pattern in the beginning of the chromosome, other than it being relatively flat as opposed to the peaks at ~2.7Mbp, ~3.7Mbp and ~4.6Mbp which are present in varying degrees between the genospecies. We note that there is a very significant pattern in the cc *R/theta* (Figure 10), where the maximum lies around ~3.4Mb and the minimum lies at ~2.5Mb in all genospecies.

Conclusion

We have shown that the variation in recombination activity can explain a major part of the variation in GC3 (Table 2). Moreover, this observation is robust to how recombination is inferred (Table 2). We find no indication that the spatial structuring of recombination and GC-content along the chromosome can explain or confound such a signal. This indicates that a mechanism linked to recombination, likely gene conversion, shapes the variation of GC-content (gBGC). We also found that the relationship between these two quantities is stronger in the genospecies with a higher amount of informative sites (Table 2). This indicates that in the genospecies with a higher amount of gene conversion activity, the GC-bias is more pronounced. We note that the number of informative sites in the sampled genospecies does not correspond directly to the number of strains sampled. This indicates that some, namely genospecies B, likely is more clonal than the others. The lack of correlation between sample size and the number of informative sites might also be because the data set contains a bad representation of these genospecies.

We conclude that gBGC is present in this *Rhizobium leguminosarum* species complex, and that it is not confounded by the chromosomal structuring of GC-content. This possibly means that gBGC is present in the whole *R. leguminosarum* species which expands the number of clades (Duret & Galtier, 2009; Lassalle et al., 2015), where gBGC has already been found.

Proposals for further studies

Because there is an ongoing debate in the literature (Bobay & Ochman, 2017; Duret & Galtier, 2009; Hildebrand et al., 2010; Lassalle et al., 2015) on whether the GC-enrichment is due to selection or a neutral process (i.e. gBGC), it would be appropriate to investigate whether selection can explain the GC-enrichment. But because gBGC closely mimics selection by leading to an increased frequency in the population, this is hard to do. (Lassalle et al., 2015) test whether selection of optimal codons can explain the GC-enrichment, which they find that it doesn't. However, because selection of optimal codons is closely linked to the spatial structure along the chromosome due to the leading strand bias (Rocha, 2008), it would be relevant to investigate whether distance to the replication origin can explain the GC-enrichment. In order to do so, we would first have to confirm the correct location of the *dnaA* replication origin genes in the chromosomes.

Another analysis I wanted to explore was to measure the age of the GC-enrichments. Because we already inferred the internal branches of the genealogies for each gene, it should be possible to measure whether the GC-enrichments are distributed evenly or concentrated in the older or newer parts of the phylogeny. However, it should be noted that this will depend on a data set that represents as much genetic variation in a species as possible, because the parameters likely will be highly noisy. Genospecies C would be a good candidate.

A completely different idea which is inspired by the investigations in recombination hotspots in mammals (Duret & Galtier, 2009), would be to sequence the genomic DNA of a cell line over time, and track the gene conversion tracts. This would allow for a specific investigation of gene conversion without regard to other products of recombination which, as mentioned earlier, are possibly confounding the signals we found in this study. This would require the meticulous breeding of bacterial cells with iterative sequencing, followed by the development of algorithms that can track these tracts.

Because recombination in bacteria is not well understood, I think more resources should be prioritized into describing the proteins and mechanisms involved in this process, especially the mechanism that takes care of the correction of non-Watson-Crick base pairs. Recombination is a vital process for all living organisms. Failing to understand how it works ultimately limits our understanding on the limits of the effects of evolution.

A different thing I also want to do is to investigate if the distinct sinusoidal signature of cumulative centered recombination rate (cc $R/theta$) can be reconstructed with different methods and different species/data sets. Either it is a fluke, present only in this data set, or its presence suggests that the distribution of recombination rates is shaped by positional constraints in the chromosome.

Acknowledgements

Thanks Maria. This thesis would not have been the same without your feedback and insightful ideas. I'm sad to learn that the corona-circumstances prohibits me from inviting you to my physical defense.

I hope my processing of your data set has given you at least *something* in return. Nonetheless, I hope we will get another opportunity to work together.

Thomas, thanks for your help. I certainly enjoyed your enthusiasm around my final project at BiRC.

Cheers, Carl.

Bibliography

Akashi, Motohiro and Yoshikawa, Hirofumi (2013). *Relevance of GC content to the conservation of DNA polymerase III/mismatch repair system in Gram-positive bacteria.*

Bobay, Louis-Marie and Ochman, Howard (Oct. 2017). "Impact of Recombination on the Base Composition of Bacteria and Archaea". en. In: *Mol. Biol. Evol.* 34.10, pages 2627–2636.

Brown, Thomas C. and Jiricny, Josef (1987). *A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine.*

Bruen, Trevor C., Philippe, Hervé and Bryant, David (Apr. 2006). "A simple and robust statistical test for detecting the presence of recombination". en. In: *Genetics* 172.4, pages 2665–2681.

Cavassim, Maria Izabel A. et al. (Mar. 2020). "Symbiosis genes show a unique pattern of introgression and selection within a species complex". en. In: *Microb Genom.*

Chen, Jian-Min, Cooper, David N., Chuzhanova, Nadia, Férec, Claude and Patrinos, George P. (2007). *Gene conversion: mechanisms, evolution and human disease.*

Daubin, Vincent and Perrière, Guy (Apr. 2003). "G+C3 structuring along the genome: a common feature in prokaryotes". en. In: *Mol. Biol. Evol.* 20.4, pages 471–483.

Didelot, Xavier and Wilson, Daniel J. (Feb. 2015). "ClonalFrameML: efficient inference of recombination in whole bacterial genomes". en. In: *PLoS Comput. Biol.* 11.2, e1004041.

Duret, Laurent and Galtier, Nicolas (2009). "Biased gene conversion and the evolution of mammalian genomic landscapes". en. In: *Annu. Rev. Genomics Hum. Genet.* 10, pages 285–311.

Fukui, Kenji (July 2010). "DNA mismatch repair in eukaryotes and bacteria". en. In: *J. Nucleic Acids* 2010.

Galtier, Nicolas, Duret, Laurent, Glémin, Sylvain and Ranwez, Vincent (Jan. 2009). "GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates". en. In: *Trends Genet.* 25.1, pages 1–5.

Graur, Dan (Dec. 2015). *Molecular and Genome Evolution.* en.

Hildebrand, Falk, Meyer, Axel and Eyre-Walker, Adam (Sept. 2010). "Evidence of selection upon genomic GC-content in bacteria". en. In: *PLoS Genet.* 6.9, e1001107.

Hill, W. G. and Robertson, A. (Dec. 1966). “The effect of linkage on limits to artificial selection”. en. In: *Genet. Res.* 8.3, pages 269–294.

Kanaar, R., Hoeijmakers, J. H. and Gent, D. C. van (Dec. 1998). “Molecular mechanisms of DNA double strand break repair”. en. In: *Trends Cell Biol.* 8.12, pages 483–489.

Kobert, Kassian, Flouri, Tomáš, Aberer, Andre and Stamatakis, Alexandros (2014). *The Divisible Load Balance Problem and Its Application to Phylogenetic Inference.*

Kogay, Roman, Wolf, Yuri I., Koonin, Eugene V. and Zhaxybayeva, Olga (no date). *Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents.*

Lassalle, Florent et al. (Feb. 2015). “GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands”. en. In: *PLoS Genet.* 11.2, e1004941.

Lawrence, J. G. and Ochman, H. (Apr. 1997). “Amelioration of bacterial genomes: rates of change and exchange”. en. In: *J. Mol. Evol.* 44.4, pages 383–397.

Lin, Mingzhi and Kussell, Edo (Feb. 2019). “Inferring bacterial recombination rates from large-scale sequencing datasets”. en. In: *Nat. Methods* 16.2, pages 199–204.

Pouyet, Fanny and Gilbert, Kimberly J. (Sept. 2019). “Towards an improved understanding of molecular evolution: the relative roles of selection, drift, and everything in between”. In: arXiv: `1909.11490 [q-bio.PE]`.

Pupko, T., Pe’er, I., Shamir, R. and Graur, D. (June 2000). “A fast algorithm for joint reconstruction of ancestral amino acid sequences”. en. In: *Mol. Biol. Evol.* 17.6, pages 890–896.

Rocha, Eduardo P. C. (2008). “The organization of the bacterial genome”. en. In: *Annu. Rev. Genet.* 42, pages 211–233.

Slater, Guy St C. and Birney, Ewan (Feb. 2005). “Automated generation of heuristics for biological sequence comparison”. en. In: *BMC Bioinformatics* 6, page 31.

Thorpe, Harry A., Bayliss, Sion C., Hurst, Laurence D. and Feil, Edward J. (May 2017). “Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species”. en. In: *Genetics* 206.1, pages 363–376.

Wickham, Hadley et al. (2019). *Welcome to the Tidyverse.*

Wilkinson, Leland (2011). *ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H.*

Yahara, Koji et al. (Feb. 2016). “The Landscape of Realized Homologous Recombination in Pathogenic Bacteria”. en. In: *Mol. Biol. Evol.* 33.2, pages 456–471.

Young, J. Peter W. et al. (Apr. 2006). “The genome of Rhizobium leguminosarum has recognizable core and accessory components”. en. In: *Genome Biol.* 7.4, R34.