Comparative Genomics Reveals Forces Driving the Evolution of Highly Iterated Palindrome-1 (HIP1) in Cyanobacteria.

Minli Xu

CMU-CB-15-104

March, 2015

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

Thesis Committee:
Dr. Dannie Durand, Chair
Dr. Luisa Hiller
Dr. Jeffrey Lawrence
Dr. Daniel Barker

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported in part by NSF grant DBI1262593. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: Repetitive Sequence, Small Repeat, Cyanobacteria, Highly Iterated Palindrome-1, HIP1, Horizontal Gene Transfer, HGT, Tree Reconciliation, Notung, Temporal Feasibility.

Acknowledgments

Foremost, I would like to express my warmest thanks to my advisor, Dr. Dannie Durand, for her continuous guidance, inspiration, and selfless support of my PhD study and research, for her patience, motivation, and enthusiasm. Her guidance and encouragement kept me going all through the time of research and writing of this thesis.

I also wish to express my sincerest gratitude to other members of my thesis committee: Dr. Jeffrey Lawrence, Dr. Daniel Barker, and Dr. N. Luisa Hiller, for all the valuable suggestions, ideas and feedback on my thesis, and for taking time out from their busy schedule to serve on my committee. Their advice, guidance, and insights profoundly influenced this thesis. Without them, this thesis would not exist.

I would also like to thank all the Durand lab members, for creating such a nice and friendly environment to study and work in, for their support and companionship. Through the years, I learned a lot from them.

Finally I would like to thank all the friends and classmates here at CMU and PITT, with whom the six years of PhD study is such an exciting and memorable journey.

Abstract

The Highly Iterative Palindrome-1 (HIP1) is a highly abundant octamer palindrome motif (5-GCGATCGC-3) found in a wide range of cyanobacterial genomes from various habitats. In the most extreme genome, HIP1 frequency is as high as one occurrence per 350 nucleotides. This is rather astonishing considering that at this frequency, on average, every gene will be associated with more than one HIP1 motif. This high level abundance is particularly intriguing, considering the important roles other repetitive motifs play in the regulation, maintenance, and evolution of prokaryotic genomes. However, although first identified in the early 1990s, HIP1s functional and molecular roles remain a mystery.

Here I present a comparative genomics investigation of the forces that maintain HIP1 abundance in 40 cyanobacterial genomes. My genome-scale survey of HIP1 enrichment, taking into account the background tri-nucleotide frequency in the genome, shows that HIP1 frequencies are up to 300 times higher than expected. Further analysis reveals that in alignments of divergent genomes, HIP1 motifs are more conserved than other octamer palindromes with the same GC content, used as a control. This conservation is not a byproduct of codon usage, since codons in HIP1 motifs are more conserved than the same codons found outside HIP1 motifs. HIP1 is also conserved on a broader scale. I predicted orthologs using the Notung software platform and compared enrichment of HIP1 motifs with control motifs across orthologous gene pairs. The similarity of HIP1enrichment in orthologs is significantly higher than the control. Taken together, my results provide the first evidence for the mechanism driving HIP1 prevalence. The observed conservation is consistent with selection acting to maintain HIP1 prevalence and rejects the hypothesis that HIP1 abundance is due to a neutral process, such as DNA repair. The evidence of selection thus suggests a functional role for HIP1. My analysis of the genome-wide spatial distribution of HIP1 suggests that the motif lacks periodicity, voting against a role in supercoiling. The spatial distribution of HIP1 motifs in mRNA transcript data from Synechococcus sp. PCC 7942 reveals a significant 3 bias, which is suggestive of regulatory functions such as transcription termination and inhibition of exonucleolytic degradation. I conclude by discussing my findings in the context of cyanobacterial evolution and propose testable hypotheses for future work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Repetitive sequences play important roles in prokaryotic genome architecture and evolution (Delihas, 2011; Treangen et al., 2009). Mobile elements contribute to genome plasticity by transposition, moving sequences around the genome. The presence of similar sequences at dispersed locations enable illegitimate recombination, which also contributes to genome plasticity and evolution. Some repetitive sequences, such as USSs and CRISPRs, are gatekeepers of genome integrity, facilitating or inhibiting the acquisition of foreign DNA. Other repetitive sequences are related to prokaryotic chromosome maintenance, such as AIMS and Chi sequence. Bacterial repetitive motifs can also participate in transcriptional regulation by controlling supercoiling or transcription termination (Treangen et al., 2009).

The Highly Iterated Palindrome-1 (HIP1), a palindromic motif (5'-GCGATCGC-3'), is highly abundant in a wide range of cyanobacterial genomes from various habitats. A survey conducted in 2011 (Delaye et al., 2011b) reveals that the high HIP1 abundance is only observed in cyanobacterial genomes, among all the completely sequenced prokaryotic genomes available at the time the study was conducted. Although discovered twenty years ago (Robinson et al., 1995, 1997), the functional and molecular roles of HIP1 remain a mystery. No mechanism or biological system has been identified, to my current knowledge, that explains the observed high level of prevalence. It is not even known whether HIP1 is under selection, or whether HIP1 abundance is an artifact of some neutral process.

In this thesis, I present a comparative genomic analysis of the taxonomic distribution, enrichment, and conservation of the HIP1 motif, with particular attention to evidence for selection acting on HIP1 motifs, which could suggest a functional role.

1.1 Cyanobacteria

Cyanobacteria is one of the oldest lineages on the planet. The first fossil record of cyanobacteria, in the form of stromatolites, dates back to 2.8 billion years ago (Olson, 2006). Cyanobacteria are believed to have played important roles in shaping the atmosphere of the earth during the Great Oxygenation Event (also known as the Great Oxidation Event) 2.4 billion years ago (Kump, 2008; Sessions et al., 2009). It is also believed that plant chloroplasts originated from cyanobacteria according to the endosymbiosis hypothesis (reviewed in Brinkman et al. (2002)). With 2.8 billion years of evolutionary history, currently existing cyanobacterial genomes display a high level of diversity (Whitton, 2012; Whitton and Potts, 2000). Many cyanobacterial species are capable of photosynthesis and nitrogen fixation. Interestingly, cyanobacteria are among those few prokaryotes that display multicellularity, including cell type differentiation and cell-cell communication. Cell type is one important basis for cyanobacterial taxonomy (Rippka et al., 1979). The current commonly used taxonomy classifies the cyanobacterial phylum into five subsections, based on cell morphology and development (Rippka et al., 1979; Shih et al., 2013). Subsections I and II are unicellular cyanobacteria. Subsection I (Chroococcales) are strains that undergo binary fission for reproduction, while Subsection II (Pleurocapsales) are strains that undergo multiple fissions. Subsections III, IV, and V are multicellular cyanobacteria capable of forming filaments, long chains of individual cells. Subsections IV (Nostocales) and V (Stigonematales) are strains capable of differentiating into specific cell types, such as heterocysts. Although the original classification has been revised based on molecular data, the taxonomy based on the five subsections remains. Another important feature of this group of species is a circadian clock, which allows some cyanobacterial strains to undergo circadian regulation of gene expression (Dvornyk et al., 2003; Loza-Correa et al., 2010). It is worth mentioning that while many cyanobacteria are free-living, some form symbiotic relationships with other organisms, including fungi, corals, and angiosperm plants. Cyanobacterial toxins are among the most powerful known poisons in nature, posing potential health risks. Furthermore, cyanobacteria have been used as tools in biotechnology applications, such as biofuel synthesis, food production, pharmaceuticals, and bioremediation (Abed et al., 2009).

1.1.1 Ecology and lifestyles

Over the course of evolution, species in the current cyanobacterial lineage have developed various kinds of sophisticated systems to adapt to their living environment. Some of these (like the circadian clock) are considered signature features that are rarely observed in prokaryotes outside of the cyanobacterial lineage. Considering that the hyper-abundance of HIP1 motif is only observed in cyanobacteria among prokaryotes (Delaye et al., 2011b), it is a valuable practice to study those features that are unique to cyanobacteria or that are only found in a few other groups. Here, I briefly review the most salient cyanobacterial features.

Photosynthesis

Photosynthesis is the process of converting light energy into chemical energy that is usable for various cellular activities. The name of cyanobacteria is actually related to photosynthesis, as cyan means blue in Greek. Some of the earliest described cyanobacteria are blue in color, because of the bluish pigment phycocyanin, which cyanobacteria utilize to capture the light for photosynthesis. Interestingly, it is believed that photosynthesis in eukaryotes (algae and plants) is descended from cyanobacterial photosynthesis, according to the endosymbiotic theory, which is supported by both morphological and molecular evidence (Baum, 2013; Keeling, 2010; McFadden and van Dooren, 2004).

In most known cyanobacteria, Photosystems I and II are the major components for photosynthesis. However, the photosynthetic machinery varies across cyanobacterial species. Genomic data reveal that not all species have the same set of Photosystem I and II genes (Mulkidjanian et al., 2006). In addition, a range of photosynthetic pigments are used to capture light at different wave lengths in different species. The basal cyanobacterium *Gloeobacter violaceus* is unique among cyanobacteria in that it does not have thylakoid membranes, the subcellular location where the light-dependent reaction of photosynthesis occurs in most cyanobacteria. In contrast, photosynthesis occurs in the plasma membrane in *Gloeobacter* (Nakamura et al., 2003). In addition, many genes for Photosystem I and II are absent from this genome.

Nitrogen fixation

Many cyanobacteria are capable of nitrogen fixation, in which inorganic nitrogen gas (N_2) from the atmosphere is converted into ammonia, nitrites or nitrates. Nitrogenase, the key enzyme for this process, has been identified in many cyanobacterial genomes. However, not all cyanobacteria are nitrogen-fixing, and diazotrophs are found in many other prokaryotic phyla.

Circadian clock

Oxidative photosynthesis and nitrogen fixation are incompatible processes, as photosynthesis creates a highly oxidative local enviroment, and nitrogenase, the key enzyme for nitrogen fixation, is inactivated by oxygen. Species that perform both functions have acquired various mechanisms for isolating theses environments from each other. One strategy is temporal separation of the two incompatible processes. It has been observed for many cyanobacteria, that there exists a circadian rhythm of gene expression (Dvornyk et al., 2003), in which different sets of genes are expressed during different temporal intervals over the course of a day. One early study by Liu et al. (1995) suggested that almost all promoters are rhythmically regulated. Three key genes have been identified that control this cyanobacterial circadian clock system, namely KaiA, KaiB, and KaiC (Ishiura et al., 1998). Genome-wide scanning has shown that these three genes are present in current completely sequenced cyanobacterial genomes, except that KaiA is missing in some marine pico-cyanobacteria, KaiA and KaiB are missing in cyanobacterium UCYN-A, and none of the three were found in *Gloeobacter* (Axmann et al., 2014). Further, KaiA orthologs are only found in cyanobacterial genomes. These observations suggest that the circadian system evolved within the cyanobacterial lineage. Numerous additional genes involved in this circadian system have been identified in the last decades (Katayama et al., 1999; Schmitz et al., 2000). The cycling of the phosphorylation status of KaiC protein, which is the output of the KaiABC central oscillator, was reconstructed in vitro (Nakajima et al., 2005). However, it is still unknown how the central oscillator, which is composed of KaiABC, controls global gene expression (Dvornyk et al., 2003; Loza-Correa et al., 2010). It has been speculated that this global gene expression oscillation is regulated at a chromosomal level, and the rhythmic changes of the topology of the entire chromosome (e.g., via super-coiling) could explain the global circadian rhythm (Smith and Williams, 2006; Vijayan et al., 2009; Woelfle et al., 2007).

In addition to the core oscillator genes (KaiABC), other genes have been identified participating in the input (PexA, ldpA, and CikA) and output (CikA, SasA, LabA, and RpaA) pathways to the central circadian oscillator (recently reviewed in Axmann et al. (2014)). Among them, CikA has a dual role in both input and output pathways. CikA is destabilized by oxidized quinones, and affects the KaiC phosphorylation via an unknown mechanism. On the other hand, in the output pathway, CikA interacts with phosphorylated KaiBC complexes and inhibits RpaA, which has a DNA-binding domain. RpaA is currently the final component of the output pathway, and its target(s) and impact on chromosome topology have not been identified. Therefore, how RpaA affects the chromosome topology and regulates the global gene expression is currently the missing link in the cyanobacterial circadian system. CikA, LabA, and SasA participate in the output pathway by interacting with KaiBC complexes and activate or inhibit RpaA. Among these input and output genes, LabA and CikA are missing in marine pico-cyanobacteria, PexA, LdpA, and LabA are missing in cyanobacterium UCYN-A, and Pex, CikA, and SasA are missing in *Gleoebacter* (reviewed in Axmann et al. (2014)). It is not clear if exist circadian systems in the genomes of *Gleoebacter*, cyanobacterium UCYN-A, and marine pico-cyanobacteria.

Cell types

Another strategy for separating nitrogen fixation and photosynthesis is differentiation into distinct cell types, with different functional roles. The various cell types allow spatial separation of the two incompatible processes. For example, in some strains from family Nostocaceae, two types of cells can be observed: the common vegetative cells for photosynthesis, and thick-walled heterocysts where nitrogen fixation occurs (reviewed in Wolk et al. (2004)).

Cyanobacteria are also capable of other forms of cell type differentiation. When the living environment becomes harsh, some species can form a spore-like dormant cell called an akinete. Thanks to thick cells wall and food reserves, akinetes have enhanced survivability. Cyanobacterial cells of Subsections III, IV, V can form filaments, also called trichomes, long chains of individual cells. Filaments can break into shorter reproductive filaments with gliding motility. Filament fragments released from an immotile parental filament are called Hormogonia (reviewed in Rippka et al. (1979)).

1.2 Bacterial small repeat sequences

Various types of small repeat sequences exist in bacterial genomes, and they have different impacts on bacterial genome evolution. Many repetitive sequences described in the literature are related to or are themselves mobile elements. The Repetitive Extragenic Palindromic sequences (REPs) have imperfect palindromic core sequences, are 20-60 bp long and occur hundreds of times in a wide range of bacterial genomes (Stern et al., 1984; Versalovic et al., 1991). Some REP sequences are specific targets for insertion elements. The relationship between REPs and mobile elements has been systematically studied in Tobes and Pareja (2006), and it was found that some Insertion Sequence (IS) elements interact specifically within REPs. The REP sequence has been proposed to replicate itself by an RNA-mediated mechanism of gene conversion that maintains its prevalence within the genomes (Higgins et al., 1988). As one of the earliest discovered repetitive sequence, numerous functional roles have been proposed for REPs. It has been suggested that REP sequences are involved in transcription termination based on the observation that most REP motifs are located near the 3-terminus of genes (Gilson et al., 1986; Manzanera et al., 2001). This hypotheses was later tested experimentally by Aranda-Olmedo et al. (2002), who observed no terminator activity associated with REP sequences experiments with plasmids possessing introduced REP elements in *Pseudomonas putida*. This hypothesis has not been tested in other species, to my knowledge. REPs are also proposed to be related to mRNA stabilization, control of translation and genomic rearrangements (reviewed in Treangen et al. (2009)). Various studies also suggested that REPs are the binding sites for DNA polymerase I, DNA gyrase, and integration host factor (reviewed in Treangen et al. (2009)), which could potentially be related to DNA physiology in bacteria. It was observed that REP can fold into small stem loops (Higgins et al., 1982; Stern et al., 1984), suggesting possible secondary structures at the DNA or mRNA level, with potential functional importance. Interestingly, the genome-wide distribution of REPs can be either dispersed or clustered. In the latter case, clusters of REP sequences are called Bacterial Interspersed Mosaic Elements (BIMEs) (Bachellier et al., 1994). BIMEs typically consist of tandemly repeated doublets of two closely spaced REP elements (Gilson et al., 1991). In *E. coli*, important proteins were reported to interact wtih REPs or BIMEs, including integration host factor, DNA gyrase and DNA polymerase I (reviewed in Nunvar et al. (2013)). Interestingly, BIMEs have also been shown to protect mRNA from 3' exonucleolytic degradation by exonuclease III (McLaren et al., 1991; Newbury et al., 1987; Py et al., 1996).

Miniature Inverted-repeat Transposable Elements (MITEs), also known as class III transposons, are present in many bacterial and eukaryotic genomes (Fattash et al., 2013). Bacterial MITEs are usually between 100 and 400 bp in length, and have a relatively complex structure. Each MITE contains a core sequence, flanked by terminal inverted repeats, which can carry open reading frames (reviewed in Delihas (2011)). MITEs are incapable of self-transposition. MITEs are generally believed to be derived from IS (Jiang et al., 2004). A number of short repeats that have been reported in specific studies, including the RUP motif in *Streptococcus*, the ERIC motif in *Enterobacteriaceae*, and Correia in *Neisseria*, have MITE-like properties and could be considered members of this class (Delihas, 2011).

Shorter repeated units are also common, and are observed both in tandem arrays and dispersed around the genome. A number of known bacterial short repeats are of functional importance. Some of them have functions that are strongly supported by experimental evidence, while others are speculated to be associated to certain biological functions. Architecture IMparting Sequences (AIMS) are non-palindromic octamers responsible for maintaining chromosome architecture, allowing for orderly replication and segregation (Hendrickson and Lawrence, 2006). AIMS are strand-biased, and are overrepresented on one DNA strand. Further, AIMS on leading strands are increasingly abundant toward replication termini, allowing proteins with functions related to chromosome replication and segregation to find the replication termini. It is reported that AIMS have been identified in nearly all bacterial genomes (Hendrickson and Lawrence, 2006).

Another non-palindromic octamer, the Chi site, is associated with recombination hotspots, and found to be involved in recombinational repair of DNA (reviewed in (Smith, 2012)). Similar to AIMS, the distribution of Chi sequence is found to be strand-biased in bacterial genomes, which higher abundance in the leading strand (El Karoui et al., 1999; Uno et al., 2000)

Highly Repetitive Motifs (HRMs) found in *Lactococcus latis* are 13 bp long non-palindromic dispersed motifs (Mrazek et al., 2002). More than 2000 copies have been identified in that genome. An analysis by Mrazek et al. (2002) shows that HRMs are more frequently present in close proximity to the 3'-end of genes, as suggested by the distributions of distances to the starts and ends of genes. Based on this observation, it has been hypothesized that HRMs might be related to transcription termination. The distribution of the spacing between two neighboring HRMs (either on the same strand or not) reveals a strong periodic pattern of 10 bp at the whole genome level. The spacing analysis also showed that HRM motif and the inverted complement of HRM often form close dyads, separated by $\leq$20 bp. In two smaller genomic regions (380 bp and 455 bp), r-scan analysis demonstrated that HRM has a periodicity of 59 bp.

Many bacterial short repeats are related to controlling the influx of foreign genetic material. Some of these promote the uptake of foreign genetic material. For example, DNA Uptake Short Sequences (USSs) are $\sim$10 bp long motifs, first found in *Haemophilus influenzae* (Smith et al., 1999), that allow the species to be naturally transformable. It has also been suggested that USSs could be related to transcription termination given that they are frequently located within transcription terminators (Smith et al., 1999). This hypothesis has never been tested systematically to my knowledge.

In contrast, Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) are direct repeats of palindromic sequences $\sim$30 bp in length, with spacer DNA separating the repeats. Approximately 40% of sequenced Eubacterial and 90% of sequenced Archaeal genomes contain at least one CRISPR locus, according to CRISPRdb (Grissa et al., 2007). Recent studies suggest that CRISPRs act as a prokaryotic immune system that provides resistance to alien genetic material (Barrangou et al., 2007). Interestingly, the CRISPR system has been utilized as an effective bioengineering tool for genome editing and gene regulation in both prokaryotes and eukaryotes (Mali et al., 2013).

Many of these various types of repeats are also observed in Cyanobacterial genomes. Short Dispersed Repeats (SDRs) are found in a wide range of cyanobacterial genomes, with lengths 16-40 bp (Elhai et al., 2008). Their primary sequences are non-symmetrical and non-palindromic. However, SDRs are predicted to form similar secondary structures (Elhai et al., 2008). Alignments of homologous regions revealed SDR insertions within genomes, suggesting they are mobile elements. Elhai et al. (2008) identified eight classes of SDRs (SDR1-8). Among them, all classes of SDRs are present in the three tested genomes from the heterocyst-forming Nostocaceae clade, with at least 10 occurrences per genome. Outside of the Nostocaceae clade, SDR1 and SDR7 were found in four cyanobacterial genomes with lower frequency (less than 10 per genome). No SDR has been detected in marine pico-cyanobacteria. Most interestingly for this thesis, SDR5 has a specific insertion target site, which is the HIP1 motif. However, no SDR5-like sequences were detected in HIP1-rich cyanobacterial genomes outside of Nostocaceae clade (Elhai et al., 2008), suggesting it is unlikely that HIP1 acts as an insertion site for SDR, as its primary function.

Aside from SDR repeats, various types of small repeats have been documented in cyanobacterial genomes. Katayama et al. (2002) reported tandem repeats, with length 7-14 basepairs, in five cyanobacterial genomes (*Nostoc sp.* PCC 7120, *Synechocystis sp.* PCC6803, *Thermosynechococcus elongatus* BP-1, *Synechococcus sp.* WH8102 and *Prochlorococcus marinus* CCMP1986) with abundance ranging from 29 in *Prochlorococcus marinus* CCMP1986 to 294 in *Nostoc sp.* PCC 7120. Several thousand MITE occurrences have been identified in 17 cyanobacterial genomes by Lin et al. (2011). Kaneko et al. (2007) identified eight groups of putative MITE sequences in the cyanobacterium *Microcystis aeruginosa.* MITEs have been found inserted into microcystin genes in an Anabaena strain isolated from the Baltic Sea (Fewer et al., 2011), leading to the inactivation of these genes. Interestingly, in a study by Treangen et al. (2009), the cyanobacterial genome *Microcystis aeruginosa* NIES-843 is found to be among the top ten genomes with highest repeat coverage, out of 659 bacterial genome analyzed. Repetitive regions make up 20.4% of its 5.8 Mb genome. In the same study, the cyanobacterial genome *Prochlococcus marinus* MIT 9312 was found to be among the ten genomes with the lowest repeat coverage. Other marine pico-cyanobacteria have very low repeat coverage as well.

In summary, many functions are ascribed to bacterial repetitive sequences. Sequences such as REP and BIME contribute to genome plasticity by transposition. Similar repetitive elements at dispersed location may enable illegitimate recombination, which also contributes to genome plasticity and evolution (e.g. the Chi sequence). Some repetitive sequences, such as USSs and CRISPRs, are gatekeepers of genome integrity, facilitating or inhibiting the acquisition of foreign DNA. Repetitive sequences contribute to prokaryotic chromosome maintenance, such as the Chi sequence. Repetitive sequences have also been found to be involved in chromosome replication and segregation. Further, bacterial repetitive motifs can also potentially be involved in transcriptional regulation by controlling supercoiling, and transcription termination (e.g. REP, HRM).

1.3 Highly Iterated Palindrome-1 motifs

Highly Iterated Palindrome-1 (HIP1) is a highly abundant DNA motif uniquely found in cyanobacteria. First identified in early 1990s, this octamer motif (5'-GCGATCGC-3') is over-represented in a wide range of cyanobacterial genomes from various habitats (Robinson et al., 1995, 1997). HIP1 abundance can be as high as one occurrence per 500 nucleotides on average, which is rather astonishing considering that at this frequency, on average, every gene in that genome will be associated with more than one HIP1 motif.

The functional and molecular roles of HIP1 have remained a mystery for more than 20 years. Robinson et al. (1997) proposed and tested the hypothesis that HIP1 might function as a protein binding site. They tested this hypothesis using Electrophoretic Motility Shift Assays (EMSA), which failed to unambiguously identify any protein that binds specifically to HIP1 motifs in *Synechococcus sp.* PCC 7942. Given that their results reflected only one cyanobacterial strain and were based on the technology that was available at the time, the role of HIP1 as a potential binding site for protein-DNA interaction remains unresolved.

Akiyama et al. (1998) hypothesized that HIP1 may be involved in site-specific recombination of plasmids in *Synechococcus sp.* PCC7002, but this hypothesis has never been tested. Elhai et al. (2008) proposed, based on the scanning of short mobile elements in 16 cyanobacterial genomes, that HIP1 could be an insertion site for mobile element Short Disperse Repeat-5 (SDR5) in the Nostocaceae lineage. However, no SDR or SDR-like repeat was found in any HIP1-enriched cyanobacterial genome outside of the Nostocaceae lineage. Though SDR may have a functional association to HIP1 in Nostocaceae, it is very unlikely that SDR5 can entirely explain the biological functional of HIP1, given the much wider phylogenetic distribution of the HIP1 motif.

A more recent analysis, based on 40 cyanobacterial genomes, used a phylogenetic profiling approach (Pellegrini et al., 1999) to infer HIP1 function (Delaye et al., 2011b). One goal of the study was to establish a functional linkage for HIP1, by association analysis between HIP1-enriched and non-HIP1-enriched cyanobacterial genomes. The authors first searched for a single PFAM protein domain whose phylogenetic distribution matched the presence and absence of HIP1 hyper-abundance across the 40 cyanobacterial genomes analyzed. Their phylogenetic profiling was unable to identify a single PFAM domain that has the exact same phylogenetic distribution as HIP1 hyper-abundance across the 40 genomes. However, when protein domain architectures were used instead of protein domains for the phylogenetic profiling analysis, Delaye et al. (2011b) identified a candidate gene family, the glucose 6-phosphate dehydrogenase assembly protein (OpcA). OpcA is present in all the cyanobacterial genomes studied in Delaye et al. (2011b). The gene encoding OpcA was found to have one PFAM domain in genomes lacking HIP1, while in hyper-abundant HIP1 genomes, OpcA has two PFAM domains: the OpcA_G6PD_assem (glucose-6-phosphate dehydrogenase subunit) domain and PG_binding_1 (gutative peptidoglycan binding) domain. The authors concluded that HIP1 may be functionally linked to the opcA protein.

This result is intriguing because OpcA is cyanobacteria-specific gene with circadian transcriptional regulation (Min and Golden, 2000). However, this functional correlation analysis was performed without accounting for common ancestry. The distribution of HIP1 among the 40 genomes is not patchy: All cyanobacterial genomes in that study possess abundant HIP1 motifs, except *Gloeobacter* and the species in the marine pico-cyanobacterial clade. Thus, much of the observed correlation between gene content and HIP1 content could be due to inheritance from a common ancestor. A correlation model that takes phylogenetic structure into account (Pagel et al., 2004) would be particularly important for an analysis of HIP1 function. A second weakness of that analysis is that phylogenetic profiling is highly sensitive to the accurate prediction of the trait in question, in this case HIP1 hyper-abundance. In this study, HIP1 enrichment was assessed based on estimates of the expected number of HIP1 motifs derived from nucleotide frequencies, only. Higher order oligomer frequencies were not taken into account, possibly resulting in errors in the assessment of HIP1 hyper-abundance. In addition, in their study, cyanobacterium UCYN-A was included in the set of genomes with HIP1 hyper-abundance, while the two Yellowstone strains were excluded. Last but not least, phylogenetic profiling analysis based on domain presence/absence across the genomes is highly sensitive to accurate domain annotation. Because of these difficulties, the conclusion in Delaye et al. (2011b) is not convincing and the correlation between HIP1 enrichment and genetic function remains an open question.

It is instructive to compare the characteristics of HIP1 with the properties of the repetitive sequences described in the previous section. HIP1 has a very different motif length and structure from the BIMEs and MITEs. HIP1 resembles REP in that both are palindromic. However, HIP1 is perfectly conserved and much shorter than REP sequences. It is reasonable to suggest that HIP1 might be a subtype of USS because of their similarity in motif lengths and their quite comparable genome-wide distribution. However, this hypothesis is not supported by the fact that USS conveys competency to the host genome, while a naturally incompetent cyanobacterial strain *Synechococcus sp.* PCC6714 (Vioque, 2007) is HIP1-rich. HIP1 differs from CRISPRs in that CRISPRs consist of clustered tandem short repeats, rather than being dispersed throughout the genome like the HIP1 motif. Thus, it is unlikely that HIP1 prevalence is related to a CRISPR-like function. HIP1 differs from AIMS and Chi sequences in that HIP1 is palindromic and thus shows no strand bias. In fact, although HIP1 has properties that are similar to those of various other types of bacterial repeat sequences, it represents a unique combination of those characteristics: it is short, palindromic, non-tandem and very conserved.

1.4 Thesis Overview

A review of repetitive sequences reveals their profound importance to prokaryotic genome evolution, chromosome physiology and genetic regulation. In light of this, it is surprising how little is known about the HIP1 motif. This lack of knowledge underscores the importance of gaining a better understanding of the genomic behavior and biological roles of the HIP1 motifs.

The major focus of my thesis includes characterizing the taxonomic distribution of HIP1 abundance and enrichment, the spatial distribution of HIP1 motifs in genomes and within transcriptional units, and the conservation of HIP1 motifs in species diverging from a common ancestor. These studies are directed towards understanding how HIP1 abundance is maintained in cyanobacterial genomes, as well as developing more focused hypotheses concerning HIP1 functional roles.

1.4.1 Cyanobacterial genome data used in this thesis

The majority of the analyses reported in this thesis are based on two cyanobacterial datasets, summarized in Table 1.1.

The primary dataset, referred to as the NCBI dataset, consists of the 40 cyanobacterial genomes that were completely sequenced and assembled when I started my thesis research in December, 2011. This dataset was used for the analyses of HIP1 phylogenetic distribution, enrichment and conservation in Chapter 3 and the intra-genome HIP1 spatial analysis in Chapter 4. This dataset, which I obtained from NCBI's FTP site[1], includes the complete DNA sequences of all chromosomes and plasmids in each genome, as well as an annotation table that specifies the coordinates of protein coding and RNA genes within the genome. The details about the creation of this dataset is explained in Section 3.8.

A second dataset, based on 47 cyanobacterial genomes, was provided by Dr. Daniel Barker (School of Biology, University of St Andrews, Scotland). A total of 39 genomes are common to both datasets. This dataset, referred to as the Barker dataset, contains phylogenies of 13,852 gene families from 65 species. Of these, 49 are cyanobacteria, and 16 are proteobacteria, used as an outgroup. This dataset was created from protein sequences from the Integr8 database (Kersey et al., 2005) for the 65 species. Gene families were predicted based on OrthoMCL 2.0 (Li et al., 2003). This dataset also contains a species phylogeny of the 65 species. The species phylogeny was based on 147 universal, single-copy gene families. This Barker dataset was used for ortholog prediction and calculating the K_S values in Chapter 3. The details about the creation of this dataset are given in Section 3.8.

The genomes in Table 1.1 are primarily unicellular species from Subsection I (binary fission). There are also a number of multicellular species from Subsection IV, and one species from Subsection III. Subsections II and V are not represented in this dataset. Despite

[1] URL: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/

Dataset[1]	ID[2]	Genome name	Genome size	GC%	NCBI accession ID[3]	Taxonomy
N,B	ama	Acaryochloris marina MBIC11017	8.36Mb	46%	NC_009925	I
B	amx	Arthrospira maxima CS-328	5.99Mb	45%		III
N,B	ana	Nostoc sp. PCC 7120	7.21Mb	41%	NC_003272	IV
N,B	ava	Anabaena variabilis ATCC29413	7.11Mb	41%	NC_007413	IV
B	cwa	Crocosphaera watsonni WH 8501	6.24Mb	37%		I
N,B	cya	Cyanothece sp. PCC 7424	6.55Mb	38%	NC_011729	I
N,B	cyb	Cyanothece sp. ATCC51142	5.46Mb	37%	NC_010546	I
B	cyc	Cyanothece sp. CCY0110	5.86Mb	37%		I
N,B	cyd	Cyanothece sp. PCC7425	5.79Mb	50%	NC_011884	I
N,B	cye	Cyanothece sp. PCC 7822	7.84Mb	39%	NC_014501	I
N,B	cyf	Cyanothece sp. PCC8801	4.79Mb	39%	NC_011726	I
N,B	cyg	Cyanothece sp. PCC8802	4.80Mb	39%	NC_013161	I
N,B	gvi	Gloeobacter violaceus PCC7421	4.66Mb	61%	NC_005125	I
B	lyn	Lyngbya sp. PCC 8106	7.03Mb	41%		III
N,B	mae	Microcystis aeruginosa NIES843	5.84Mb	42%	NC_010296	I
N,B	naz	Nostoc azollae 0708	5.49Mb	38%	NC_014248	IV
N,B	npu	Nostoc punctiforme PCC73102	9.06Mb	41%	NC_010628	IV
B	nsp	Nodularia spumigena CCY9414	5.30Mb	41%		IV
N,B	pma	Prochlorococcus marinus AS9601	1.67Mb	31%	NC_008816	I
N,B	pmb	Prochlorococcus marinus MIT9211	1.69Mb	38%	NC_009976	I
N,B	pmc	Prochlorococcus marinus MIT9215	1.74Mb	31%	NC_009840	I
N,B	pmd	Prochlorococcus marinus MIT9301	1.64Mb	31%	NC_009091	I
N,B	pme	Prochlorococcus marinus MIT9303	2.68Mb	50%	NC_008820	I
N,B	pmf	Prochlorococcus marinus MIT9312	1.71Mb	31%	NC_007577	I
N,B	pmg	Prochlorococcus marinus MIT9313	2.41Mb	50%	NC_005071	I
N,B	pmh	Prochlorococcus marinus MIT9312	1.70Mb	30%	NC_008817	I
N,B	pmi	Prochlorococcus marinus NATL1A	1.86Mb	34%	NC_008819	I
N,B	pmj	Prochlorococcus marinus NATL2A	1.84Mb	35%	NC_007335	I
N,B	pmk	Prochlorococcus marinus CCMP1986	1.66Mb	30%	NC_005072	I
N,B	pml	Prochlorococcus marinus CCMP1375	1.75Mb	36%	NC_005042	I
N,B	sel	Synechococcus elongatus PCC7942	2.74Mb	55%	NC_007604	I
N,B	sya	Synechococcus elongatus PCC6301	2.70Mb	55%	NC_006576	I
N,B	syb	Synechococcus sp. PCC7002	3.41Mb	49%	NC_010475	I
B	syc	Synechococcus sp. BL107	2.29Mb	54%		I
N,B	syd	Synechococcus sp. CC9311	2.61Mb	52%	NC_008319	I
N,B	sye	Synechococcus sp. CC9605	2.51Mb	59%	NC_007516	I
N,B	syf	Synechococcus sp. CC9902	2.23Mb	54%	NC_007513	I
N,B	syg	Synechococcus sp. JA-2-3B'a(2-13)	3.05Mb	58%	NC_007776	I
N,B	syh	Synechococcus sp. JA-3-3Ab	2.93Mb	60%	NC_007775	I
N,B	syi	Synechococcus sp. RCC307	2.22Mb	60%	NC_009482	I
B	syj	Synechococcus sp. RS9916	2.66Mb	60%		I
B	syk	Synechococcus sp. RS9917	2.59Mb	64%		I
B	syl	Synechococcus sp. WH 5701	3.12Mb	65%		I
N,B	sym	Synechococcus sp. WH7803	2.37Mb	60%	NC_009481	I
B	syn	Synechococcus sp. WH7805	2.62Mb	58%		I
N,B	syo	Synechococcus sp. WH8102	2.43Mb	59%	NC_005070	I
N,B	syp	Synechocystis sp. PCC6803	3.95Mb	47%	NC_000911	I
N,B	syq	Thermosynechococcus elongatus BP-1	2.59Mb	53%	NC_004113	I
N,B	syr	Trichodesmium erythraeum IMS101	7.75Mb	34%	NC_008312	III
N	uca	cyanobacterium UCYN-A	1.44Mb	31%	NC_013771	I

Table 1.1: A summary of the cyanobacterial genomes used for HIP1 analysis. Genomes are ordered alphabetically. [1]N: NCBI dataset, B: Barker dataset. [2] The three-letter genome ID used in this thesis. [3] Genomes in the Barker dataset were obtained from Inter8 and hence do not have NCBI accession IDs.

the taxonomic bias among the available whole genome sequences, these datasets represent substantial genomic and ecological variety.

Both datasets contain *Nostoc azollae*, which forms a symbiotic relationship with a plant, the water-fern *Azolla filiculoides.* An estimated that 31.2% of the *Nostoc azollae* genome is made up of pseudogenes, suggesting that it may be currently undergoing genome collapse (Ran et al., 2010). The NCBI dataset also includes cyanobacterium UCYN-A [2] , a symbiont of the single-celled alga *Rhopalodia gibba. Cyanothece sp.* UCYN-A also appears to be undergoing genome reduction. Its genome is 1.44 Mb in size, compared with other Cyanothece genomes which range from 4.5 Mb to 7.8 Mb. During the genome reduction process, it lost several key pathways, including carbon fixation and Photosystem II (Kneip et al., 2008; Zehr et al., 2008), which is essential for water splitting and oxygen evolution.

The marine pico-cyanobacteria are over-represented in these datasets, reflecting great research interest in this group among the community. Marine pico-cyanobacteria are some of the most well studied cyanobacteria, consisting of oceanic *Prochlococcus* and *Synechococcus* species. They are believed to be the most abundant photosynthetic organisms on Earth (Scanlan et al., 2009). Marine pico-cyanobacteria have reduced genomes, owing to genome streamlining to adapt to nutrient-rich ocean environments. As the names suggests, marine pico-cyanobacteria are small in sizes (< 3 μm in diameter), and hence have greater surface area to volume ratios, which make them very efficient at nutrient uptake. Both *Synechococcus* and *Prochlococcus* display unique oceanic distribution as well as a wide range of pigmentation (Scanlan et al., 2009).

Both datasets contain *Gloeobacter violaceus* PCC 742. The genus *Gloeobacter* is a sister group to all other known cyanobacteria (Nakamura et al., 2003). *Gloeobacter violaceus* PCC 742 is the only complete genome in this genus. *Gloeobacter* is believed to be the earliest branching cyanobacterial species or a transition form, as it lacks some features found in all other cyanobacteria.

Two thermophilic *Synechococcus* strains (*Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*)), isolated from a hotspring in Yellowstone National park

[2] This genome is currently undergoing name changing. It is currently named as *Candidatus Atelocyanobacterium thalassa*, and previously as *Cyanothece sp.* UCYN-A or cyanobacterium UCYN-A. In this thesis, I refer this genome as cyanobacterium UCYN-A.

Literature	Taxa	Dataset
Gupta 2009	34 cyanobacteria	45 protein families
Bandyopadhyay 2011	61 cyanobacteria	226 protein families
Larsson 2011	58 cyanobacteria	285 protein proteins
Criscuolo 2011	61 cyanobacteria , and 22 primary photosynthetic eukaryotes	191 protein families
Latysheva 2012	49 cyanobacteria and 16 proteobacteria	147 protein families
Schirrmeister 2011	58 cyanobacteria	16S rRNA
Wang 2011	58 cyanobacteria	16S-23S rRNA

Table 1.2: Recent molecular phylogenies of cyanobacteria.

(Bhaya et al., 2007), are present in both datasets. Though named *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) and *Synechococcus sp.* JA-3-3Ab (*syh*), these strains are very different from the marine pico *Synechococcus* in niche and ecology, as well as in genomic structure. *Synechococcus elongatus* PCC7942 (*sel*), a genome common to both datasets, is a model organism used to study the prokaryotic circadian clock. *Microcystis aeruginosa* NIES843 (*mae*), also a genome common to both datasets, harbors a large number of repetitive sequences in its genome, comprising up to ∼20% of its genome (Treangen et al., 2009). This species has been the target of intensive study because it is responsible for toxic blooms, posing serious public health risks.

The species tree constructed by Latysheva et al. (2012) reveals the phylogenetic relationships of genomes in the Barker dataset, reproduced here in Figure 1.1. This tree was constructed from the concatenated sequences of 147 single copy gene families, as described in Latysheva et al. (2012) and Section 3.8.6 on page 117. The genomes in the NCBI dataset are labeled in red in Figure 1.1. Note that cyanobacterium UCYN-A (*uca*), which is in the NCBI dataset, but not in the Barker dataset, is not represented in this tree.

This phylogeny is generally in agreement with other recently published molecular phylogenies for the cyanobacteria, summarized in Table 1.2. In all trees, *Gloeobacter* is the basal species. The marine pico-cyanobacteria are monophyletic in all trees. In addition, all phylogenies place Nostocaceae strains in one clade. Similarly, all *Cyanothece* genomes, except *Cyanothece sp.* PCC7425 (*cyd*), form a clade. Interestingly, *Cyanothece sp.* PCC7425 (*cyd*) is frequently placed together with *Thermosynechococcus elongatus* BP-1 (*syq*) and *Acaryochloris marina* MBIC11017 (*ama*) (Criscuolo and Gribaldo, 2011; Larsson et al., 2011; Latysheva

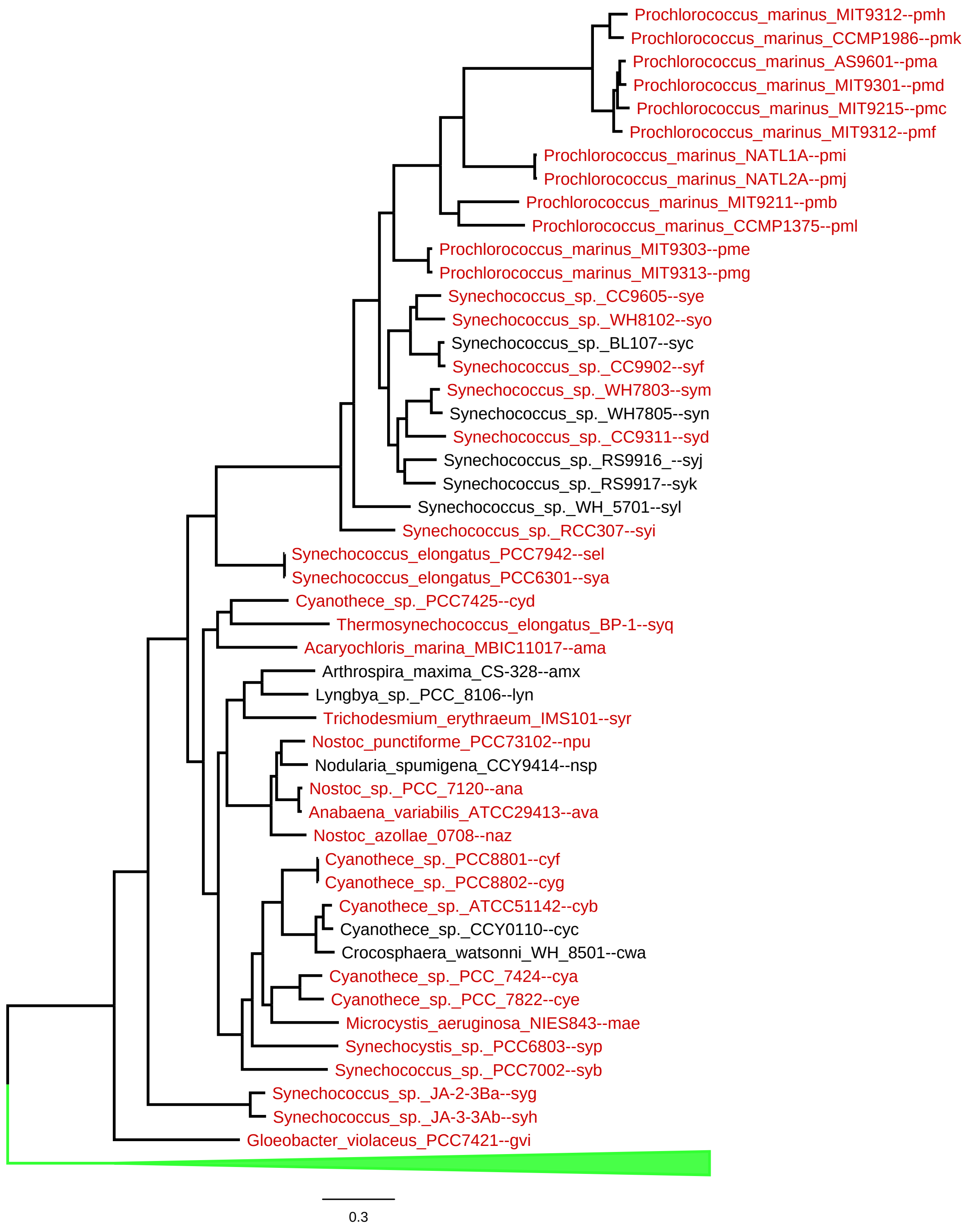

Prochlorococcus_marinus_MIT9312--pmh
Prochlorococcus_marinus_CCMP1986--pmk
Prochlorococcus_marinus_AS9601--pma
Prochlorococcus_marinus_MIT9301--pmd
Prochlorococcus_marinus_MIT9215--pmc
Prochlorococcus_marinus_MIT9312--pmf
Prochlorococcus_marinus_NATL1A--pmi
Prochlorococcus_marinus_NATL2A--pmj
Prochlorococcus_marinus_MIT9211--pmb
Prochlorococcus_marinus_CCMP1375--pml
Prochlorococcus_marinus_MIT9303--pme
Prochlorococcus_marinus_MIT9313--pmg
Synechococcus_sp._CC9605--sye
Synechococcus_sp._WH8102--syo
Synechococcus_sp._BL107--syc
Synechococcus_sp._CC9902--syf
Synechococcus_sp._WH7803--sym
Synechococcus_sp._WH7805--syn
Synechococcus_sp._CC9311--syd
Synechococcus_sp._RS9916_--syj
Synechococcus_sp._RS9917--syk
Synechococcus_sp._WH_5701--syl
Synechococcus_sp._RCC307--syi
Synechococcus_elongatus_PCC7942--sel
Synechococcus_elongatus_PCC6301--sya
Cyanothece_sp._PCC7425--cyd
Thermosynechococcus_elongatus_BP-1--syq
Acaryochloris_marina_MBIC11017--ama
Arthrospira_maxima_CS-328--amx
Lyngbya_sp._PCC_8106--lyn
Trichodesmium_erythraeum_IMS101--syr
Nostoc_punctiforme_PCC73102--npu
Nodularia_spumigena_CCY9414--nsp
Nostoc_sp._PCC_7120--ana
Anabaena_variabilis_ATCC29413--ava
Nostoc_azollae_0708--naz
Cyanothece_sp._PCC8801--cyf
Cyanothece_sp._PCC8802--cyg
Cyanothece_sp._ATCC51142--cyb
Cyanothece_sp._CCY0110--cyc
Crocosphaera_watsonni_WH_8501--cwa
Cyanothece_sp._PCC_7424--cya
Cyanothece_sp._PCC_7822--cye
Microcystis_aeruginosa_NIES843--mae
Synechocystis_sp._PCC6803--syp
Synechococcus_sp._PCC7002--syb
Synechococcus_sp._JA-2-3Ba--syg
Synechococcus_sp._JA-3-3Ab--syh
Gloeobacter_violaceus_PCC7421--gvi
0.3

Figure 1.1: Phylogeny of the cyanobacterial genomes from the Barker dataset, rooted with 16 protobacterial outgroup species (Latysheva et al., 2012). Species in the primary (NCBI) dataset are shown in red. The three letter genome IDs are beside the genome full names, separated by '- -'. Branch length unit: substitution per amino acid site.

et al., 2012), posing an intriguing evolutionary scenario for further study.

On the other hand, there is no consensus on the phylogenetic positions of the closely related strains, *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*). Most studies in Table 1.2 are in agreement with Figure 1.1 regarding the position of *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*). However, the analysis of Gupta (2009) places *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*) as sister group to the clade containing *Cyanothece* and Nostocaceae. The relationship of the Yellowstone strains *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) and *Synechococcus sp.* JA-3-3Ab (*syh*), relative to other cyanobacteria, has also been difficult to resolve. In most protein family-based phylogenies (Bandyopadhyay et al., 2011; Criscuolo and Gribaldo, 2011; Larsson et al., 2011; Latysheva et al., 2012), *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) are sister taxa to all other cyanobacterial genomes except *Gloeobacter*, as shown in Figure 1.1. In contrast in rRNA-based phylogenies (Gupta, 2009; Schirrmeister et al., 2011; Wang et al., 2011), *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) and *Synechococcus sp.* JA-3-3Ab (*syh*) are placed in a clade with *Gloeobacter*, so that *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*), *Synechococcus sp.* JA-3-3Ab (*syh*), and *Gloeobacter violaceus* PCC7421 (*gvi*) together constitute the deepest branching clade in the tree.

1.4.2 Roadmap to rest of the thesis

The rest of this thesis is organized as follows:

In Chapter 2, I describe my contributions to horizontal gene transfer inference using gene tree-species tree reconciliation. First, I present an algorithm that ensures that inferred gene family histories with transfers make temporal sense; that is, in a valid gene family history, it must be possible to order all the inferred events in a manner that is consistent with the forward progression in time. Second, I present case studies that probe the challenges of phylogenetic transfers inference, including model choice, degeneracy of solutions, and temporal inconsistency. I discuss the results of this study for developing best practices and design guidelines for new methods.

Chapter 3 focuses on elucidating the forces that shape HIP1 abundance in cyanobacterial genomes. My analysis of HIP1 enrichment, with a suitable correction for background sequence composition, confirms that HIP1 is a fundamental feature of cyanobacterial genomes. I describe the discovery of a novel HIP1 variant in two thermophillic strains that were sampled from hotsprings in Yellowstone National Park, and that were previously believed to lack HIP1. Further, I present evidence of HIP1 conservation in genome pairs at appropriate evolutionary distances that is consistent with selection acting to maintain HIP1, and allows us to reject the hypothesis that a neutral process underlies HIP1 prevalence. The evidence, following correction for overall selection on coding regions, suggests that HIP1 is not preferentially located in coding or intergenic regions. Further, within coding regions, there is no demonstrable preference for reading frame. Taken together, these results suggest that it is unlikely that selection is acting on HIP1 at the codon or amino acid level.

The evidence for selection in Chapter 3 suggests a functional role for HIP1. I explore this further in Chapter 4, focusing on the spatial distribution of HIP1. The spatial distribution of HIP1 motifs in high-resolution transcript data shows a marked 3' preference. This distribution is consistent with a regulatory role on the mRNA level, such as transcription termination, inhibition of exonucleases, or formation of stable secondary structures. It could also represent selection acting to eliminate HIP1 motifs in relatively AT-rich 5' promoter regions. Efforts to establish an association between 3' HIP1 motif enrichment and features related to transcription (e.g., mRNA abundance, GC content, codon usage) proved inconclusive.

Chapter 2

Phylogenetic Reconciliation with Transfers

Reconciliation is a procedure wherein a gene family tree is compared with a species tree in order to reveal events that occurred during the history of the gene family. Reconciliation is the most robust approach to identifying orthologs (Bourgon et al., 2004; Searls, 2003). In addition, reconciliation is used to estimate gene age and identify gene events that correlate with the emergence of novel functions. Event parsimony provides a basis for rooting an unrooted tree and for species-tree aware methods for correcting gene tree error. In a phylogenomic context, reconciliation is used in the construction of databases of annotated molecular phylogenies.

The earliest work on reconciliation focused only on gene duplication (D) and loss (L). Reconciliation under a DL-model has been studied for more than 30 years, and is a well understood problem. Reconciliation with horizontal gene transfers (HGTs), however, is a more complex and difficult problem. When the event model includes transfers, there may be more than one, and possibly many, optimal event histories. Software is required that can generate all optimal histories efficiently and then present this information to the user in a way that is not overwhelming. Another challenge is that it is possible to generate mathematically optimal histories that are biologically unacceptable because they imply a temporal ordering of events that cannot be realized without the aid of a time travel machine. These conflicts must be recognized and conveyed to the user. Software packages to support reconciliation with transfers are relatively new. There is no agreement on how to meet these challenges. Community consensus on best practices for the application of such tools has yet to be established.

As a doctoral student, I participated in a collaborative project to develop Notung 2.7[1], reconciliation software that supports an event model with transfers. In contrast to published algorithms which do not include losses in the optimization criterion (Berglund-Sonnhammer et al., 2006; Ma et al., 2000; Tofigh et al., 2011; Zmasek and Eddy, 2001), we developed algorithms for inferring tranfers together with duplications and losses. Unlike other algorithms, Notung reports all optimal solutions that are temporally feasible. Further, our algorithm is the first to infer transfers when reconciling a binary gene tree to a non-binary species tree. In addition, it includes a heuristic to distinguish between incongruence arising from either uncertainty or Incomplete Lineage Sorting (ILS) and incongruence arising from gene duplications, and transfers. Notung 2.7 was released on September 2012, and was described in a paper published in Bioinformatics (Stolzer et al., 2012).

Development of Notung 2.7 was a team effort and I participated in many aspects of the project in a collaborative way. Two aspects of the project are uniquely mine: first, I developed an algorithm for testing whether a candidate optimal solution is temporally feasible and implemented this algorithm in Notung; second, I carried out an empirical investigation of algorithm performance on two biological datasets, with particular attention to how the event model and the choice of event costs influences the number and types of events inferred, the frequency of degenerate solutions, and the frequency of temporally infeasible solutions. I also investigated the potential for overestimating events when the branching order of the species tree is poorly resolved. The lessons learned from these case studies provided us with a deeper understanding of the challenges posed by reconciliation with transfers and had a significant impact on downstream design decisions. The results of both contributions were reported in Stolzer et al. (2012).

In Section 2.1, I give a general review of methods for HGT inference, with a focus

[1] URL:http://lampetra.compbio.cs.cmu.edu/Notung/index27.html

on transfer in prokaryotic genomes. In 2.2, I review phylogenetic reconciliation, paying particular attention to the challenges that arise when transfers are included in the event model. My specific contributions are described in Section 2.4 and Section 2.5.

2.1 Inference Methods for Horizontal Gene Transfer

Horizontal Gene Transfer (HGT) refers to the exchange of genetic material between species that are not vertically related. In eukaryotes, HGT events are considered to be rare, while in prokaryotes, HGT occurs much more frequently (Ochman et al., 2000). For example, Medigue et al. (1991) suggested that approximately 15% percent of the *E.coli* genome is subject to HGT activities. HGT is an important aspect of bacterial genome evolution, allowing for rapid acquisition of new systems for adaptation. This is seen in bacterial pathogens, which can acquire antibiotic resistance genes horizontally (Courvalin, 1994).

Elucidating HGT is very important when studying gene functions in bacteria. Currently, the ortholog-based approach still plays a major role in bacterial gene annotation and functional prediction. The underlying assumption is that orthologous genes perform similar biological functions. When a horizontally transferred gene is mistaken for an ortholog, a function might be assigned to that gene erroneously.

In addition, the prevalence of horizontally transferred regions in bacterial genomes makes prokaryotic phylogeny reconstruction a very complicated problem, as different parts of the genome may be of different evolutionary origins (Gogarten et al., 2002). A common question for bacterial evolution is whether it is meaningful to use a species phylogeny to represent bacterial evolution (Gogarten et al., 2002; Kunin et al., 2005).

Therefore, detecting HGT activity and identifying possible horizontally transferred genes is critically important in the study of bacterial genome evolution and the biological function of genes.

Currently, there are two major strategies for inferring HGT activities: (1) parametric methods, and (2) phylogenetic methods. Parametric methods infer HGT largely based on sequence features, such as GC content. Phylogenetic methods, on the other hand, utilize phylogenetic signals for detecting HGTs. Though fundamentally different in technique, parametric and phylogenetic methods are complementary approaches that have their own respective advantages and disadvantages. In practice, it is recommended that both types of approaches be used for studing potential HGT in the biological scenario of interest (Fitzpatrick, 2012).

2.1.1 Parametric Methods

Most parametric methods focus on one genome of interest, and detect HGT activity in that particular genome by identifying regions with sequence composition that strongly deviates from the rest of the genome. In addition to sequence signatures such as atypical sequence composition, genomic context such as surrounding transposase genes can also be used as effective predictors of HGT activity (Hacker et al., 1997).

Commonly used genomic features for HGT detection include nucleotide composition, oligonucleotide spectrum, and codon usage. Genomic composition can be used for inferring HGT activity, because bacterial nucleotide composition varies widely. For example, even within the relatively closely related marine pico-cyanobacterial lineage, the genomic GC content can vary from 30% in *Prochlorococcus marinus* CCMP1986 to 60% in *Synechococcus sp.* WH7803 (see Table 1.1). Thus, GC content can be used as a genome-specific signature for identifying recent HGT, when the genomic signatures from donor and recipient differ significantly. In practice, the GC content at the first and third codon positions are often used for inferring transferred genes (Lawrence and Ochman, 1998).

GC content is a special case of the k-mer spectrum method. Instead of using one single k-mer motif as a genomic signature, the k-mer spectrum method assesses the frequency of all possible k-mers in a particular genome. For example, dinucleotide biases were used as a genomic signature to discriminate between sequences from different genomes (Karlin, 1998; Karlin et al., 1995). The k-mer spectrum tends to vary more between genomes than within genomes, suggesting a fairly good signal-to-noise ratio for use as a genomic signature for detecting HGT activity (Reviewed in Ravenhall et al. (2015)). The choice of the value of k controls the predictive power, as well as the complexity of the calculation. Currently, di-, tri-, and tetra nucleotide frequencies have all been frequently used for inferring HGT activity (Delaye et al., 2011a; Dufraigne et al., 2005).

Codon usage bias is also a very powerful measure for predicting HGT, as each genome has a characteristic preference for certain synonymous codons (Karlin, 1998; Karlin et al., 1999). When using this feature, a gene is predicted to be foreign, if the codon usage in that gene is significantly different from the genome-wide codon usage.

For parametric methods for detecting HGT, genes are often used as the units for prediction. A sliding window of fixed length is also frequently used. A longer window size can better tolerate the within-genome variability; however, it will be worse at the detection of HGT in smaller segments (Ravenhall et al., 2015) Many studies have focused on detecting larger alien genomic regions, termed genomic islands, with typical lengths 10-200kb. For example, Chatterjee et al. (2008) proposed a method for detecting genomic islands by comparing tetra-nucleotide frequencies within a sliding window against that in randomly sampled genomic regions of the same size.

More recently, clustering-based methods have been used to infer sets of genes within a particular genome that are likely to be of foreign origin. These methods are based on the idea that alien genes are likely to possess properties that are similar to each other, and hence form a cluster in sequence space (reviewed in Azad and Lawrence (2012)).

Genomic context is also frequently used as a predictor of HGT activity. The presence of surrounding repetitive sequences, such as transposase, integrase or tmRNA genes, may indicate a horizontally acquired genomic region. Furthermore, regions introducing disruption in gene order, when compared to a closely related sister genome, are indicators of potential HGT activity. For example, a gene forming part of a non-native operon, when all the other genes in that operon are native as evidenced by sister genomes, could be considered to be subject to HGT activity (reviewed in Ravenhall et al. (2015)).

One notable limitation of parametric methods is that the transferred segments need to display a significant level of difference in terms of sequence signal, compared to the host genome. Such difference in signal can be hard to achieve when (1) the transfer donor and recipient of the transfer have similar levels of the sequence signal of interest, and more commonly (2) the transfer event is ancient and the transferred regions have adopted sequence features of the host genome. This occurs because transferred genes are subject to the directional mutation pressures of the recipient genome, a process called amelioration (Lawrence and Ochman, 1997). In addition, the fact that a bacterial genome may have regions with atypical composition that are not due to HGT, but rather resulted from functional constraints such as codon selection, may also introduce noise that can lead to false positive predictions.

In general, parametric methods are particularly useful when closely related genomes are not available, hence comparative genomics-based phylogenetic methods are not possible.

2.1.2 Phylogenetic Methods

Phylogenetic methods use information about species relationships to infer horizontal gene transfer. Phylogenetic methods can be loosely further classified into two categories: the implicit and explicit methods.

Implicit phylogenetic methods

The implicit methods examine evolutionary distance, sequence similarity, or patterns of presence/absence across species, without explicitly modeling horizontal transfer events. Most implicit methods do not require species or gene phylogenies, and rely on extracted signals that can be used as evidence for potential HGT activities.

A simple implicit approach is to detect highly similar sequences in distantly related genomes. For example, Nelson et al. (1999) used BLAST to search for foreign genes in the eubacterial species, *Thermotoga maritima*. For certain *T. maritima* query sequences, the top BLAST hits were archaeal genes, rather than genes from more closely related eubacteria, suggesting that the query genes are of archaeal origin. The discrepancy between gene and species divergence can also be used for inferring HGT without explicitly comparing the trees (Novichkov et al., 2004). According to the molecular clock hypothesis, orthologous genes should evolve in a consistent, clock-wise fashion across species. Under this assumption, the divergence between orthologous genes should be proportional to the divergence between their corresponding species. A discrepancy between gene and species divergence times may indicate potential HGT events. One application of this approach is to look for outliers in the distribution of similarity scores between homologous genes. Statistical tests, such as the Spearman rank correlation-based test used by Lawrence and Hartl (1992) and the likelihood ratio test used by Dessimoz et al. (2008), have been applied to identify transferred genes with similarity scores that significantly differ from genes in the same family.

Phylogenetic profiles, i.e., the pattern of presence and absence of gene family members across species, can also be used to identify horizontally transferred genes (Pellegrini et al., 1999). In this case, a gene that lacks homologs in closely related genomes, but possesses homologs in more distant genomes, may have been horizontally transfered. The presence/ absence pattern can be extended to include counts of gene family members that can be used to reconstruct the evolutionary scenario along the species tree (Pagel, 1999).

Explicit phylogenetic methods

Explicit phylogenetic methods are based on comparing the gene family tree with the corresponding species tree. Incongruence between the two trees is interpreted to be due to an evolutionary event or phenomenon that interrupted standard vertical descent. HGT activity is one source of incongruence between the species tree and the gene tree.

Subtree Pruning and Regrafting (SPR) is one explicit technique to detect HGT (MacLeod et al., 2005). When a SPR operation is performed, an internal branch of the gene tree is selected and cut (pruned) and then regrafted onto another branch, internal or leaf. The edit distance between a gene tree and species tree, which is based on the number of SPR operations required to transform the gene tree topology into that of the species tree, is a clue to potential HGT events. However, finding the minimum number of SPR operations between a gene tree and species tree is NP-hard (Bordewich et al., 2004).

One SPR-based software for inferring HGT events is HorizStory (MacLeod et al., 2005). HorizStory first collapses subtrees that are identical between the species tree and the gene tree, in order to reduce the problem size. It then recursively performs SPR operations until the gene tree topology agrees with the species tree.

Another explicit phylogenetic strategy for detecting HGT activity is to decompose a gene tree into substructures. For example, a gene tree can be decomposed into quartets, or trees containing only four leaves, sub-sampled from the original tree. HGT events can be inferred when the topologies of quartets do not agree with the species tree topology (Bansal et al., 2011; Zhaxybayeva et al., 2006). Such quartet decomposition methods are computationally efficient in handling large-scale phylogenomic analysis.

Tree reconciliation is another explicit approach, that infers an event history, pinpointing exactly where in the gene phylogeny putative HGT events occurred. Briefly, given a rooted gene tree, a rooted species tree, and a mapping from present day genes to present day species, tree reconciliation finds the ancestral association between genes and species and the set of events that best explains incongruence between the gene tree and the species tree. The set of inferred events are postulated events on the gene tree, including speciation, gene duplication, gene loss, and horizontal gene transfer.

Limitations of explicit phylogenetic methods include the requirement of accurate phylogenies for the species and genes of interest, their inability to infer recent HGT between sister taxa, the difficulty of distinguishing between HGT and other sources of incongruence, and the need to deal with inferred transfer events that conflict, resulting in histories that are temporally inconsistent.

2.2 Inferring Transfers with Reconciliation

Reconciliation of gene and species trees is used to investigate many aspects of gene family evolution. Two important distinguishing properties of a reconciliation algorithm are the optimization criterion used for inference and the event model, which is the set of allowable events. Mathematically, reconciliation is based on the rationale that the event history that optimizes a given objective criterion is the best explanation of the observed incongruence.

Commonly used reconciliation criteria include maximum likelihood and maximum parsimony. Maximum likelihood methods assume that gene family events occur according to a neutral, stochastic model. In this model, the process of reconciliation infers the rates of events, as well as the event history that maximizes the likelihood of the observed gene tree given the species tree. Probabilistic approaches to reconciliation have the advantage that they capture uncertainty and provide a general framework that allows for incorporating sequence evolution in the analysis. However, they are also computation intensive, require sufficient data to infer the rates, and may incur errors due to over-fitting. In contrast, parsimony models are based on the assumption that the history with the fewest events is the best explanation of the observed incongruence. A disadvantage of parsimony models is that they do not provide a model of uncertainty and cannot be used to infer rates. On the other hand, parsimony models have the advantage of computational tractability and avoid the risk of over-interpretation due to over-fitting. For the remainder of this thesis, only parsimony models will be considered.

The earliest reconciliation algorithms considered gene duplications and losses (a DL-event model) (Goodman et al., 1998; Page and Charleston, 1997) or duplications alone (Zmasek and Eddy, 2001). Under the DL-event model, the most parsimonious reconciliation is unique and can be found in polynomial time with a greedy algorithm. Given a rooted gene tree and a rooted species tree, the greedy algorithm traverses the gene tree, starting with the leaves, and labels each gene tree node with the associated species node. Once this association has been established, the labels are used to infer duplication and loss events. If the gene tree is unrooted, the root can be inferred using duplication-loss parsimony (Chen et al., 2000). Each branch in the gene tree is assigned a score corresponding to the weighted sum of the events in the most parsimonious reconciliation obtained when the tree is rooted on that branch. The tree is then rooted on the minimum cost branch. Note that there may be more than one optimal root.

The increased awareness of the importance of HGT in bacterial evolution led to the development of reconciliation algorithms based on parsimonious event models with transfers. Most event inference algorithms consider either gene duplication or HGT (reviewed in Doyon et al. (2011); Nakhleh (2011); Nakhleh and Ruths (2009)), but not both. Exact algorithms with exponential time complexity have been presented for the duplication-transfer (DT) (Tofigh et al., 2011) and duplication-transfer-loss (DTL) models (David and Alm, 2011; Stolzer et al., 2012), under a parsimony criterion. These algorithms use dynamic programming to find the event history that minimizes the weighted sum of inferred events,

$$\pi \;=\; \delta \cdot N_D + \lambda \cdot N_L + \tau \cdot N_T, \tag{2.1}$$

where N_D, N_L, and N_T are the number of inferred events, and δ, λ, and τ are costs of duplications (D), transfers (T), and losses (L), respectively. In practice, the costs are often chosen based on empirical results. For example, when performing HGT prediction among prokaryotes where high HGT activity are expected, lower costs of λ are used. The choice of costs is an active research topic (Libeskind-Hadas et al., 2014), and is currently being systematically studied using simulated genome evolution in our group.

In contrast to the DL-model, when the event model includes transfers, the optimal reconciliation is no longer unique, and there can be more than one most parsimonious event history. An example of a reconciliation problem with multiple optimal histories is given in Figure 2.1. Reconciling the gene tree with the species tree, under the DTL model with event costs $\delta = 1$, $\tau = 3$, and $\lambda = 1$, results in three optimal solutions with the same total event cost. The first two reconciliations in Figure 2.1 both have two transfers and one loss; i.e., they have the same total cost and the same number of events of each type. The only difference is the direction of transfer *t2*, and the species in which the loss occurred. This is a very common source of degeneracy, when different sets of transfers lead to the same total event cost. The third reconciliation (Figure 2.1(c)) has one duplication, one transfer and three losses. This is another source of degeneracy, when a transfer can be traded for a duplication and one or more losses. Although all three optimal solutions in this example have the same total event cost, the underlying event histories are fundamentally different. In practice, it is important to consider all multiple optimal solutions when inferring the event history of a gene family, because different solutions could lead to different biological interpretations.

A second challenge associated with reconciliation with transfers is temporal feasibility. Inferred transfer events introduce temporal constraints because the donor and recipient species of each transfer must have co-existed. In a reconciliation with two or more transfers, the constraints may be mutually incompatible, resulting in a gene family history that cannot be realized without traveling backwards in time. Figure 2.2 shows a simple example of temporal

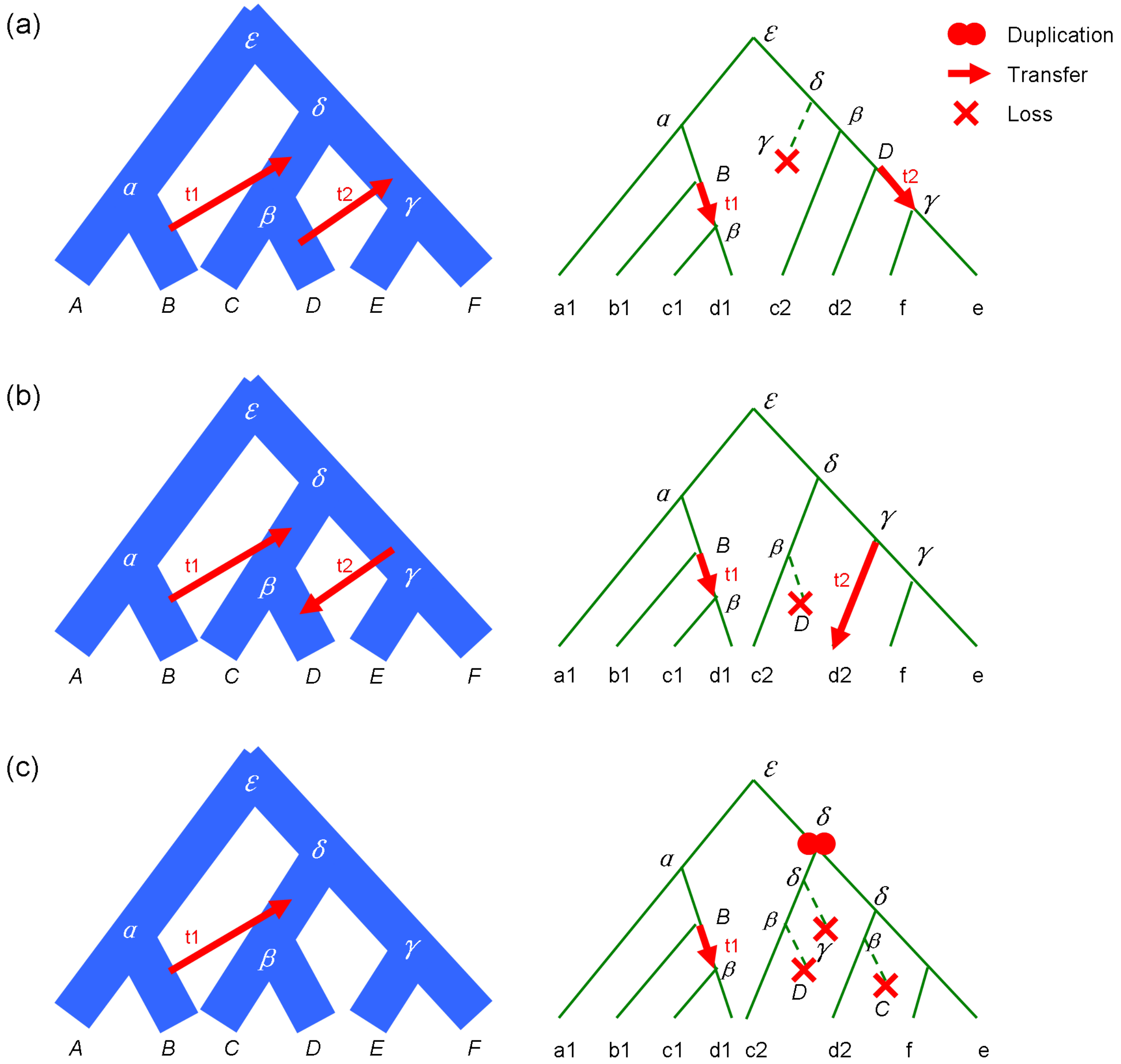


Figure 2.1: Multiple optimal solutions: Hypothetical species and gene trees with three optimal reconciliations under the DTL-model when $\delta = 1$, $\tau = 3$, and $\lambda = 1$. All three solutions have a total cost of seven. For each solution, the inferred events are shown in red on the gene tree (right). Each internal gene tree node is labeled with its associated species. The species tree (left) is annotated with inferred transfer events only.

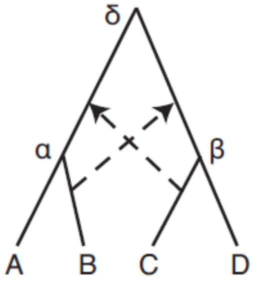


Figure 2.2: A simple example showing a temporal infeasibility caused by two transfers. Here a species phylogeny is shown with two transfers (dashed across).

infeasibility caused by two transfers. If ancestral species α acquired a gene from C, α and C must have been co-existed at the same time. The ancestor of C could not possibly acquire the gene from a descendant of α, at least, not without time travel. Such histories are called temporally infeasible.

In order to be biologically meaningful, an event history must be temporally feasible. Unfortunately, finding the optimal, temporally feasible reconciliation is an NP-complete problem (Hallett et al., 2004; Tofigh et al., 2011). The practical consequence of the NP-completeness of reconciliation with transfers is that there is no known way to find the optimal, temporally feasible event history without considering all possible event histories. There are two approaches to this problem. The first approach requires a species tree with branch lengths, where the branch lengths are proportional to time. In this restricted model, it is possible to determine which species pairs are contemporaneous. The great advantage of this approach is that event inference with transfers can be solved in polynomial time. However, algorithms for this restricted model may fail to recognize transfers if they involve a taxon that is missing from the dataset (Huson and Scornavacca, 2011; Nakhleh, 2011). More importantly, this model (reviewed in Doyon et al. (2011); Huson and Scornavacca (2011)) requires estimates of speciation times, which are frequently not known.

The second approach to the problem of temporal feasibility, which we adopt here, makes no assumption about branch lengths. Instead, we use a heuristic that finds candidate event histories that are mathematically optimal and then test each candidate history for temporal feasibility (Stolzer et al., 2012; Tofigh et al., 2011). The test for temporal infeasibility is one of the contributions of this thesis and is discussed in Section 2.4. This test is based solely on topological considerations. It does not seek to establish the co-existence of donor and recipient of an inferred transfer, but to eliminate histories that could not possibly have occurred. For example, the scenario in Figure 2.2 is intrinsically impossible regardless of the co-existence of any two existing or ancestral species. The species tree topology, together with the two transfers, alone determines infeasibility.

Tofigh et al. (2011), in an extension of their earlier work (Hallett et al., 2004), introduced one of the earliest algorithms for reconciliation with both duplication and transfers. They were also the first to stress the importance of temporal consistency in phylogenetic reconciliation when not using a dated species tree. To address the temporal constraints, Tofigh et al. (2011) proposed two approaches to address temporal constraints. The first uses a dated species tree. The second algorithm performs reconciliation with an undated species tree and then performs a *post hoc* test for temporal feasibility. However, their scheme for testing feasibility is for the less restrictive DT-model, when loss is not considered in the optimization criterion. Under this model, the recipient species of a transfer can be 'lifted' to avoid a potential temporal inconsistency. This will only induce additional losses, but since these losses incur no cost, the solution is till optimal. However, this approach would not work for detecting temporal infeasibility in reconciliations using a DTL-model. Under the DTL-model, such solutions would no longer be guaranteed to be optimal.

2.3 Software for tree reconciliation

Software packages implementing the algorithms cited above are summarized in Table 2.1. Here, I review some of the recent reconciliation models and their corresponding software, in alphabetical order, focusing especially on how they handle temporal constraints and multiple optimal solutions.

AnGST David and Alm (2011) developed an algorithm for reconciliation with the DTL-model. Their method for handling time constraints is to accept a dated species tree as input. Time consistency is hence ensured by forcing that the donor and recipient species of a transfer have intersecting time intervals. This method was implemented in the program

Software/method	Event model	Temporal consistency	Multiple solutions	Interface	Implementation
AnGST	DTL	Dated species tree		Command-line	Python
EUCALYPT	DTL	*post hoc*	Output	Command-line	Java
Mowgli	DTL	Dated species tree	Count only	Command-line	C++
Notung	DTL	*post hoc*	Output	Command-line and GUI	Java
RANGER-DTL	DTL	Dated species tree		Command-line	

Table 2.1: A summary of recent software and methods for phylogenetic reconciliation.

AnGST[2], which is Python-based and has a command-line interface. One particular strength of AnGST is that it can deal with phylogenetic uncertainties in gene trees. It considers alternative topologies in the set of bootstrap subtrees to obtain a gene tree with a minimal cost. However, multiple optimal solutions were not addressed in AnGST.

EUCALYPT EUCALYPT[3] is a program by Donati et al. (2015) that also performs parsimonious tree reconciliation under the DTL-model. Temporal infeasibility is tested using my algorithm, as presented in Stolzer et al. (2012). Unlike the other programs discussed here except Notung, EUCALYPT enumerates all optimal solutions using a polynomial-delay algorithm. EUCALYPT is implemented in Java and has a command-line user interface.

Mowgli Mowgli[4] is a maximum parsimony-based tree reconciliation program for the DTL-model (Doyon et al., 2011). Interestingly, Mowgli can compute the number of optimal solutions, but the option to output all optimal reconciliations is not available. Mowgli, like AnGST, also addresses temporal constraints by using a dated species tree and requiring the transfer donor and recipient species to co-exist within some temporal interval. Mowgli is written in C++ and has a command-line user interface. Like AnGST, Mowgli, as implemented in Mowgli-NNI, can deal with uncertainty in gene trees. It does this by considering edges with weak bootstrap support and performing Nearest-Neighbor Interchanges (NNI) to find the ideal topology with minimal reconciliation cost.

RANGER-DTL RANGER-DTL[5] is another software implemention of the DTL-model for tree reconciliation (Bansal et al., 2012). RANGER-DTL is implemented to efficiently analyze trees with even thousands of taxa, via a preprocessing step of the species tree. RANGER-DTL can take either dated or undated species trees as input. When a dated species tree is used, RANGER-DTL requires that transfer donor and recipient species co-exist. When a undated species tree is used as input, it reports a minimum cost reconciliation but does not check for temporal consistency. In neither case does RANGER-DTL output

[2] URL: http://almlab.mit.edu/angst/

[3] URL: http://eucalypt.gforge.inria.fr/

[4] URL:http://www.atgc-montpellier.fr/Mowgli/

[5] URL: http://http://compbio.mit.edu/ranger-dtl/

multiple solutions. RANGER-DTL has command-line user interface.

2.3.1 Other sources of incongruence

Conceptually, phylogenetic reconciliation is used to explain the incongruence between a species tree and a green tree. In addition to duplication, transfer and loss, gene tree incongruence can arise due to other evolutionary processes, including incomplete lineage sorting and hybridization. In prokaryotes, fractured speciation processes can give rise to gene tree heterogeneity (Retchless and Lawrence, 2010). Incongruence due to other processes can be incorrectly interpreted as duplications or transfers, leading to overestimation of the number of events that occurred. Accurate inference of gene events can also be a problem in a species tree that is not well resolved. If the branching order in the species tree is incorrect, or unknown, gene tree incongruence can similarly be misinterpreted as evidence of duplication or transfer.

To avoid overestimation of gene events, algorithms are needed that can distinguish between gene events and other sources of incongruence. One approach to this problem is to combine explicit models of gene events and of population processes, respectively, in an integrated probabilistic framework. For example, a recently developed algorithm (Rasmussen and Kellis, 2012) for the DL-model uses the multi-species coalescent model (Pamilo and Nei, 1988) to estimate the probability that a given incongruent gene tree node arose through incomplete lineage sorting, and then takes this probability into account when inferring duplications. The strength of this approach is that it includes a specific model of incomplete lineage sorting: the multispecies coalescent explicitly relates the probability of incongruence to the ancestral population size and the time between species divergences. A disadvantage is that estimates of population parameters are only available for a limited set of well studied species. Further, the specificity of the model ceases to be an advantage if the observed incongruence does not reflect the assumptions of the multispecies coalescent, either because the multispecies coalescent is too simple (e.g., it does not capture linkage, migration or selection) or because the incongruence is due to some other process, altogether.

Notung, instead, uses a simple heuristic that is designed to remove noise associated with incongruence arising either from uncertainty, or from incomplete lineage sorting. This heuristic uses a non-binary species tree to distinguish between regions of the species tree where only gene duplication and transfer need be considered (nodes with well resolved branching order) and regions where incomplete lineage sorting (ILS) or uncertainty may be contributing to gene tree incongruence (polytomies).

Typically, a species polytomy can be considered from two perspectives: a "hard" polytomy represents simultaneous divergence of three or more populations; a "soft" polytomy represents a binary branching process in which the branching order is unknown. Notung's heuristic is consistent with both meanings. If a polytomy represents rapid or simultaneous species divergences, then under the multispecies coalescent model all binary resolutions of the polytomy are equally likely. At binary species tree nodes, incongruence is always treated as evidence of gene duplication or transfer. At polytomies, gene duplication and transfer are invoked only if the gene tree does not correspond to one of the binary resolutions of the polytomy. The rationale behind the heuristic is discussed in greater detail in Stolzer et al. (2012), along with a technical description of heuristic algorithm. A soft polytomy can be viewed as a set of possible hypotheses for the time branching order, namely the set of binary resolutions of the polytomy. Our model offers a conservative stance; events are only inferred when the topology of the gene tree does not correspond to any of these hypotheses. Note that in some cases, the hard and soft polytomy models are closely linked: the branching order of species that arose through multiple speciations in rapid successions (Ebersberger et al., 2007; Pollard et al., 2006) is often difficult to resolve.

This model can be invoked for both non-binary species trees and for binary species trees with short branches where ILS is suspected; even when the binary branching order of the species tree is known, the user can collapse edges in the species tree to indicate in which lineages ILS should be considered as an alternate hypothesis.

We add "I" to the name of models, as in DTI and DTLI, to denote such heuristic is being used to model the incongruence, when performing reconciliation with a non-binary species tree.

2.4 Testing for temporal infeasibility with a DTL model

.

A reconciliation is feasible if a temporal ordering of species exists that satisfies the constraints of all inferred transfers. In a reconciliation that has two or more inferred transfer events, the temporal constraints of one transfer may be inconsistent with those of another, and result in a temporally infeasible reconciliation. There are three sources of temporal constraints, illustrated in Figure 2.3:

1. Species ancestor-descendant relationships in the species tree impose a partial temporal ordering on species in the tree.

2. For any inferred transfer, the donor and recipient species must have co-existed.

3. The gene tree imposes a temporal order of individual transfers. When a path from the root to a leaf of the gene tree passes through more than one transfer, then the species associated with the transfer closer to the leaves cannot have occurred before the species associated with the transfer closer to the root.

All three types of temporal constraints must be taken into account when considering a reconciliation. A reconciliation is feasible only if it is possible to assign order to all species in the species tree such that none of these constraints are violated.

To determine whether a reconciliation is temporally feasible, I developed a method that constructs a directed timing graph $G_t = (V_t, E_t)$ that encodes all three types of temporal constraints described above. This graph is designed to have the property that a reconciliation is temporally feasible if and only if the timing graph is acyclic. Prior to describing the construction of the timing graph, I introduce the following notation: Let $P(v)$ be the parent node of v, where v is a node either from V_G or V_S. If (u, v) is an edge either from E_G or E_S, then $P(v) = u$. For node u and v from V_T, where tree T can be either be S or G, $u \geq_T v$ means that u is an ancestor of v in tree T.

Given a gene tree $G = (V_G, E_G)$ and a species tree $S = (V_S, E_S)$, let R_{GS} be a reconciliation of G with S. Let $M(g) = s$ be a function mapping a gene tree node (g) to the corresponding species (s). $\Lambda(R_{GS})$ is the set of transfer edges in R_{GS}.

The vertices in G_t represent species in V_S. However, only species that are the donor, d, or recipient, r, of a transfer edge (g, h) in $\Lambda(R_{GS})$ must be considered. Thus, the vertex set is defined as $V_t = \{v \in V_S | \exists(g, h) \in \Lambda(R_{GS}) \ni v = M(g) \vee v = M(h)\}$.

The edges in E_t represent the three types of temporal constraints mentioned above. These are defined formally as follows:

1. If species s_i is an ancestor of species s_j in S, (i.e., $s_i \geq_S s_j$), then for every (s_i, s_j) in $V_t \times V_t$, add (s_i, s_j) to E_t iff $s_i \geq_S s_j$.

2. Given a transfer $(g, h) \in \Lambda(R_{GS})$, the donor species $M(g)$ and recipient species $M(h)$ must be contemporaneous. Therefore, for every $(s_i, s_j) \in V_t \times V_t$, add (s_i, s_j) to E_t iff s_i and s_j are the donor and recipient species, or vice versa, of some transfer $(g, h) \in \Lambda(R_{GS})$.

3. Let (g, h) and (g', h') be transfers in $\Lambda(R_{GS})$, such that $g \geq_G g'$ (g predates g'). The donor and recipient of (g, h) must have occurred no later than both the donor and the recipient species of (g', h'). In other words, $d = M(g)$ and $r = M(h)$ must have occured no later than both $d' = M(g')$ and $r' = M(h')$. Therefore, $(P(d), d')$, $(P(d), r')$, $(P(r), d')$, and $(P(r), r')$ are added to E_t.

Each candidate reconciliation is then tested for temporal feasibility by verifying that the associated timing graph G_t is acyclic, using a modified topological sorting algorithm in $\Theta(|V_t| + |E_t|)$ (Cormen et al., 2001). Only temporal feasible solutions are output by the Notung software[6].

An example of temporal infeasibility

Figure 2.3(c) shows the construction of the timing graph for the reconciliation in Figure 2.3(a) and (b). There are three inferred transfer events, labeled as $t1$, $t2$, and $t3$. The timing graph is constructed in the following three steps, corresponding to the three temporal constraints described previously. First, vertices were created by placing species that are associated with transfers into the timing graph. Note that species ϵ, A, C, and F are neither donors or

[6] When losses are not considered in the optimization criterion, edges corresponding to Point 2 can be omitted. See Stolzer et al. (2012) for an explanation.

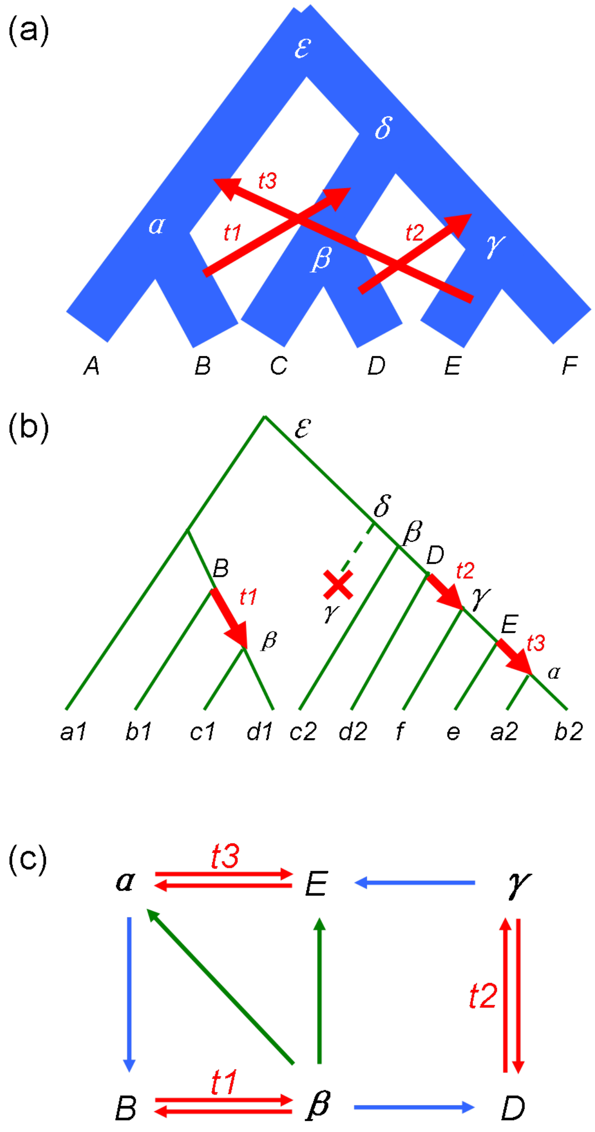


Figure 2.3: An example of a temporally infeasible reconciliation with three transfers. *(a)* A species tree showing the inferred transfers $t1$, $t2$, and $t3$. *(b)* The reconciled gene tree, showing the three inferred transfers. The nodes of each transfer are annotated with the associated donor and recipient species. *(c)* The corresponding timing graph. The colors of edges in the timing graph correspond to the three sources of temporal constraints.

recipients of transfers and do not appear in G_t. Next, edges were added according to the temporal constraints:

1. Uni-directional edges (highlighted in blue) were added according to the vertical descending order in the species tree. For example, G_t contains an edge from β to D because β is an ancestor (and, in fact, the parent) of D in the species tree.

2. Bi-directional edges (shown in red) were then added to the graph by connecting species that are transfer donors and recipients respectively.

3. Finally, unidirectional edges (green) were added according to temporal constraints corresponding to the order of transfers in the gene tree. In this particular example, $t2$ and $t3$ are two inferred transfers that appear on the same path from the root to a leaf node in the gene tree (Figure 2.3(b)). Thus $t2$ must have occurred no later than $t3$. Therefore, we must ensure that the donor and recipient species of $t2$ (D and γ) are no later than the donor and recipient species of $t3$ (E and α). Note that, "a no later than b" implies that either a occurred before b OR a was present at the same time as b. This is equivalent to the assumption that the immediate ancestor of a strictly predated b. As a result, the requirement that D be no later than E and α, is encoded by the green edges from β to E and β to α, respectively.

The constructed timing graph in Figure 2.3(c) contains multiple cyclic paths, e.g. $\alpha \rightarrow B \rightarrow \beta \rightarrow \alpha$. Thus this particular reconciliation is temporally infeasible.

The above described algorithm for testing for temporal infeasibility was implemented in Notung 2.7. For empirical comparison, the test for temporal infeasibility without losses (Tofigh et al., 2011) was also implemented in Notung.

2.5 Reconciliation with transfers: A case study

The use of phylogenetic reconciliation to infer horizontal transfer has many advantages, including the ability to infer transfers between ancestral species, to model the event history, and to infer recipient and donor species. However, HGT inference with reconciliation also poses many challenges:

Event costs: Parsimony-based reconciliation algorithms find the event histories that minimize the weighted sum of the inferred events. Currently, there is no formal framework to guide the user in selecting these costs. This is a cause for concern because the choice of event costs will influence the inference process and, ultimately, the biological interpretation of the results.

Event model: The outcome will also depend on which events are included in the reconciliation model. Many programs (Berglund-Sonnhammer et al., 2006; Ma et al., 2000; Tofigh et al., 2011; Zmasek and Eddy, 2001), do not include losses in the optimization criterion; that is, these algorithms propose candidate solutions that minimize the weighted duplication and transfer cost ($N_D\delta + N_T\tau$), but do not charge for losses.

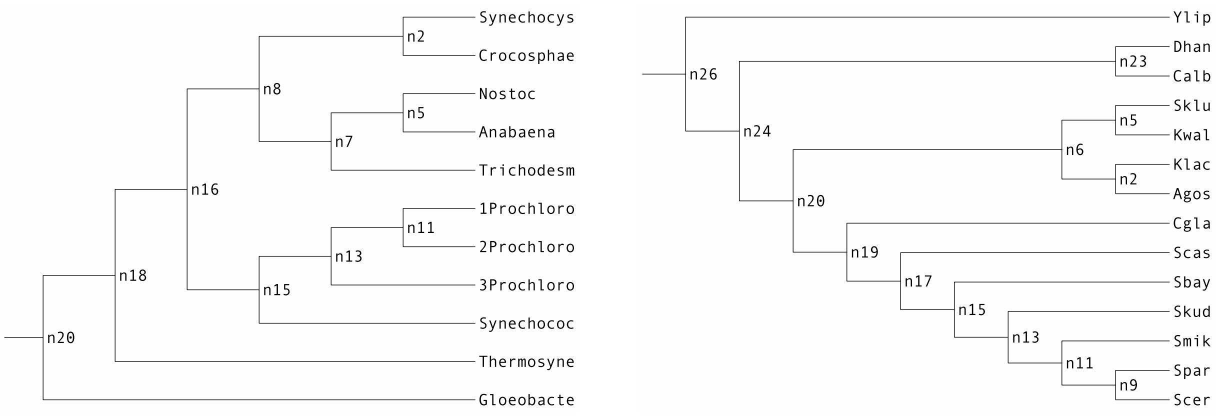
Degeneracy: In a model with transfers, there may be more than one minimum cost solution. When this occurs, how should the information from multiple histories be interpreted?

Temporal infeasibility: There is no known algorithm for efficiently finding a solution that is both optimal and temporally feasible. Notung uses a heuristic that generates candidate solutions and then tests them for feasibility [7]. The heuristic has the property that it can identify whether a candidate solution is feasible, only when the solution is also optimal. However, if all candidate solutions are infeasible, the heuristic fails. We are interested in how often does this occur?

Other sources of incongruence: In addition to duplication, transfer and loss, gene tree incongruence can arise due to uncertainty or to other evolutionary processes (e.g., ILS, hybridization) or due to noise (reconstruction error). Algorithms that do not account for these other sources of incongruence risk overestimation of duplications and transfers. What is the extent of this problem?

To investigate the extent to which users might face these challenges in analyses of real data, I carried out an empirical analysis of two empirical datasets that have been used as

[7] All reconciliation algorithms with transfers that do not assume a species tree with time estimates use some form of this heuristic.



(a) Cyanobacteria

(b) Yeasts

Figure 2.4: The species trees for *(a)* 11 cyanobacterial species and *(b)* 14 yeast species. Only tree topologies, not branch lengths, are shown. Full names were listed in Tables A2 and A3 in the Appendix.

test sets within the community.

Cyanobacterial dataset

This dataset, containing 1128 gene trees from 11 cyanobacterial species, was created by Zhaxybayeva et al. (2006), who applied Quartet Decomposition (QD) method to characterize transfer activity in those species. Gene families were predicted using a bi-directional best hit approach with BLASTP E-value cutoff of $< 10^{-4}$. A multiple sequence alignment for each family was generated using ClustalW for each gene family. Gene family phylogenies were reconstructed using Neighbor-Joining algorithm in PHYLIP with 100 bootstraps. In this dataset, each of the 1128 gene trees has at most one gene copy per species, and each gene tree has at least 7 leaves. The species tree is shown in Figure 2.4(a).

Yeast dataset

This dataset consists of 106 families of single copy orthologs from 14 yeast species that were originally constructed by Rokas and Carroll (2005). The trees used in my study are constructed via maximum likelihood estimation by Herve Philippe and colleagues from Rokas' multiple alignments of these families, as described in Jeffroy et al (2006). This dataset has been used for many studies related to phylogenetic problems and has become a *de facto* gold standard. All 106 gene trees have exactly one copy per species, and thus they all have 14 leaves. The species tree is shown in Figure 2.4(b).

2.5.1 Empirical study – Data Analysis

The trees in both datasets were reconciled using four different event models: DT, DTI, DTL, and DTLI. DT and DTI are models that do not include losses in the optimization criterion. For each model, three cost sets were used, consisting of three different transfer costs with the same fixed duplication and loss costs:

- $\delta = 3$, $\tau = 2.5$, $\lambda = 2$
- $\delta = 3$, $\tau = 6$, $\lambda = 2$
- $\delta = 3$, $\tau = 10$, $\lambda = 2$

The choice of the costs are based on empirical experience. Three different τ values were used, corresponding to different levels of expected horizontal transfer activities. In the case of $\tau = 10$, one transfer is more expensive than one duplication and 2 losses. When $\tau = 6$, one transfer is more expensive than one duplication and 1 loss. When $\tau = 2.5$, one transfer is cheaper than one duplication alone.

For each setting of model and event cost, gene trees were first rooted with Notung's rooting function with corresponding model and parameter settings. Recall that this roots the gene tree on edge that minimizes the total reconciliation cost. The rooted trees were then reconciled and the following information was tabulated: (1) the number of events of each type, (2) the gene and (3) species lineages in which each event occurred, (4) the donor and recipient of each transfer, and (5) the number of temporally infeasible reconciliations. The summarized results are given in Table 2.2 for cyanobacteria and Table 2.3 for yeast. Trees that had no temporally feasible solution for at least one set of parameter values, were eliminated from analysis under all models and values of τ. If a tree had multiple optimal solutions (either multiple optimal roots or multiple reconciliations for a specified root or

Model	τ	n_D	n_T	n_L	Infeasible	Degenerate
DT	2.5	7	1798	1560	84	6
DT	6	1648	191	6096	0	0
DT	10	2066	0	7520	0	0
DTI	2.5	6	1521	1468	3	67
DTI	6	1425	133	5133	0	0
DTI	10	1691	0	5921	0	0
DTL	2.5	0	2121	781	42	13
DTL	6	73	1740	1516	82	50
DTL	10	1324	480	4797	83	40
DTLI	2.5	0	1783	895	92	16
DTLI	6	82	1458	1456	90	109
DTLI	10	1122	405	4093	4	53

Table 2.2: Event counts for the cyanobacteria dataset, with $\delta = 3$ and $\lambda = 2$, based on 814 gene trees. Event counts from 314 gene trees with temporally infeasible or conflicting degenerate solutions in any model were excluded from this analysis. The number of trees not considered for each model and setting is given in the last two columns, respectively.

both), it was only retained if all solutions yielded the same counts for each event type. The total numbers of gene trees eliminated due to temporal infeasibility and degeneracy combined are 314 and 31, for the cyanobacteria and yeast datasets respectively. The detailed criteria for handling conflicting degenerate situation, for inclusion for reporting in Table 2.2 and Table 2.3, was summarized in Table A1 in the Appendix.

2.5.2 Empirical study – Results

Patterns of genetic exchange are presented visually by heatmaps (Figure 2.8- 2.11). The numbers in the heatmaps represent the number of transfer events inferred between a specific pair of existing or ancestral genomes, from the donor to the recipient genome. Therefore, heatmaps are asymmetrical because they describe the directional HGT activity among the genomes. Gene families with no feasible reconciliations were excluded. In the case of multiple optimal solutions, the average number of transfers over all feasible solutions, between each

Model	τ	n_D	n_T	n_L	Infeasible	Degenerate
DT	2.5	1	207	192	3	1
DT	6	192	26	684	0	0
DT	10	245	0	841	0	0
DTI	2.5	8	172	180	4	11
DTI	6	162	25	568	0	0
DTI	10	213	0	720	0	0
DTL	2.5	0	233	138	4	1
DTL	6	6	203	192	3	1
DTL	10	155	53	563	0	11
DTLI	2.5	0	208	115	4	12
DTLI	6	10	172	172	2	13
DTLI	10	138	42	493	1	10

Table 2.3: Event counts for the yeast dataset, with $\delta = 3$ and $\lambda = 2$, based on 175 gene trees. Event counts from 31 gene trees with temporally infeasible or conflicting degenerate solutions in any model were excluded from this analysis. The number of these trees are shown in the last two columns, respectively.

pair of genomes is reported.

My analysis of the result from these two datasets was published in Stolzer et al. (2012). As such, some of the results presented here were adapted from Stolzer et al. (2012). Now I will discuss the results with respect to the challenges described previously in the beginning of this section.

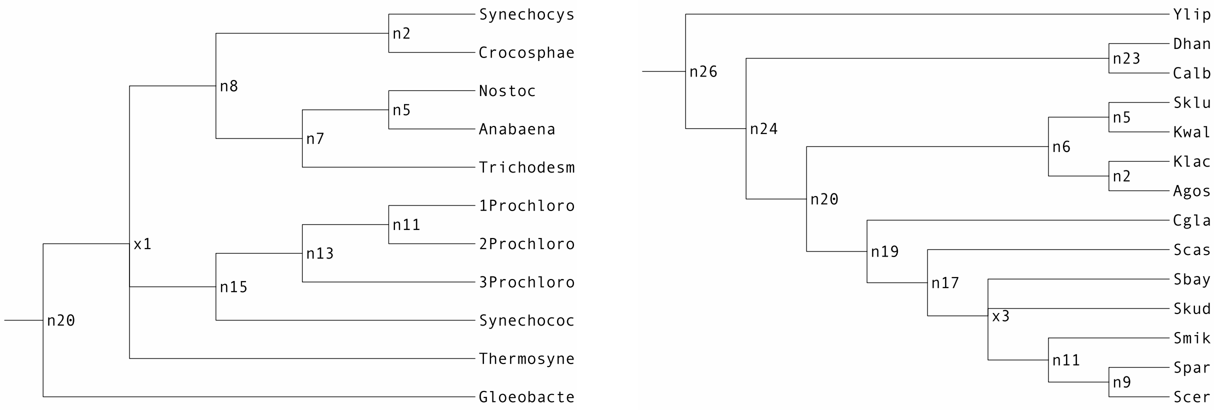
Event costs: The impact of event cost choice is revealed by comparing the event inferred with the three different transfer costs for each of the four event models. The results in Tables 2.2 and 2.3 show that in general, with the increasing cost of transfers (τ), fewer transfers, but more duplication and losses were inferred. This is as expected because transfer events can be exchanged with duplications and losses. Thus, with higher transfer costs, fewer transfer events are inferred and more species-gene tree incongruence is explained by duplications and losses.

Event model: As previously mentioned, several published reconcilation algorithms include duplications and transfers but not losses in the optimization criterion. In order to assess the impact of ignoring losses, we implemented reconciliation with the DT-model in

Notung. When losses are not included in the optimization criterion, candidate event histories that minimize the weighted sum of duplications and transfers are generated. These event histories are evaluated using the algorithm for testing temporal infeasibility without losses proposed by Tofigh et al. (2011). Comparing DT with DTL and DTI with DTLI, we see that when losses are not included in the event model, the inferred histories have more duplications, and more losses, but fewer transfers are inferred. As with increasing transfer cost, this observation can also be explained by the fact that transfer events can be exchanged with duplications and losses. Recall that in Figure 2.1, the same total cost could be obtained by replacing one transfer and one loss (Figure 2.1(b)) with one duplication and three losses (Figure 2.1(c)). As observed in this example, it is not uncommon for histories with duplications to require more losses than histories with transfers. When losses cost nothing, there are frequently more ways to trade duplications for transfers and obtain the same total event cost. In addition, by comparing the corresponding models with and without losses, we tend to observe more infeasible and degenerate cases with event models that include losses. This trend is caused by the fact that more transfer events are inferred with the DTL and DTLI models. Typically as the number of transfer events increases, the number of temporal infeasible solutions increases as well. Similarly, the number of degenerate solutions also increases with the number of transfers. This shows that exclusion of losses in the optimization function will result in substantial changes to the inferred events. Further, excluding losses can hide problems with degeneracy and temporal infeasibility.

Degeneracy: My empirical data suggests that degeneracy is a serious potential concern, even in models that do include losses. In Table 2.2 and Table 2.3, 314 out of 1128 cyanobacterial gene trees and 31 out of 106 yeast gene trees were in the degenerate class. In other words, at least 20% of trees had two or more conflicting optimal solutions in at least one case. This suggests that tree reconciliation based on a single, randomly selected optimal solution, as implemented in some reconciliation software (reviewed by (Doyon et al., 2011; Than et al., 2008)), may result in misleading biological conclusions.

Temporal infeasibility: My results also highlight the prevalence of temporal infeasibility in biological datasets, as approximately 10% of trees were removed because all solutions were temporally infeasible for at least one cost set. Previously, Hallett et al. (2004) reported



(a) Cyanobacteria

(b) Yeasts

Figure 2.5: The species trees for *(a)* 11 cyanobacterial species and *(b)* 14 yeast species, with one edge collapsed on each tree. Only tree topologies, not branch lengths, are shown. Full names were listed in Tables A2 and A3 in the Appendix.

no temporal infeasibility for the application of their DT algorithm to a simulated dataset. Later, it was suggested that temporal infeasibility is rare in biological datasets (Tofigh et al., 2011). Our results reported in Stolzer et al. (2012) contradict these reports, suggesting that infeasible cases may be more prevalent in real data than was previously thought, especially in a model with losses.

Other sources of incongruence: Incongruence can arise from many sources. If all incongruence is attributed to duplication and transfer, the number of inferred events may be unrealistically high. This is particularly the case for species trees where the branching order is unresolved. Notung's DTI and DTLI models are designed to discount incongruence that might arise from incomplete lineage sorting or other sources of noise when the species tree is non-binary. Even when the species tree is binary, if a branch is short or otherwise suspect, the user may replace that branch with a polytomy to invoke the heuristic.

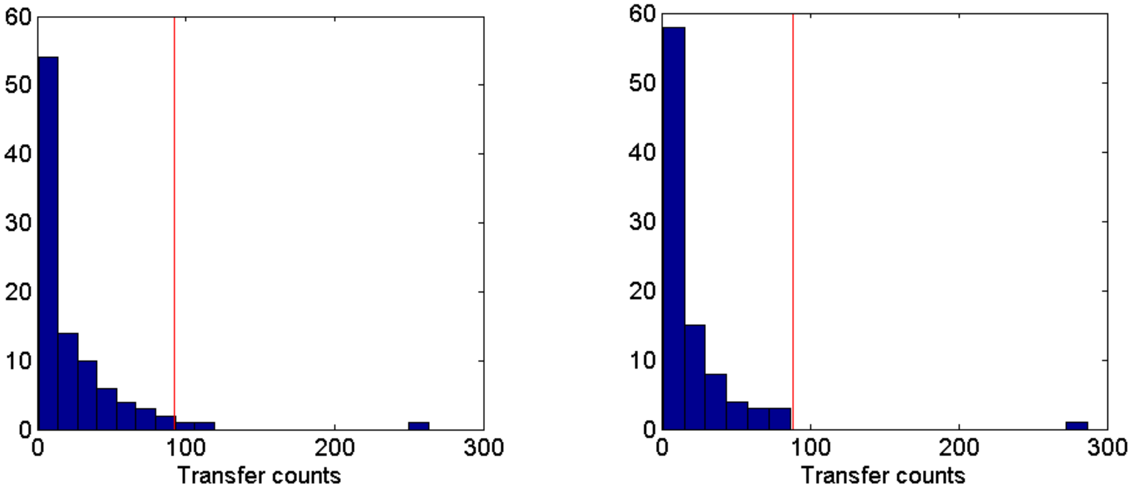
To investigate the impact of other sources of incongruence on the outcome of these analyses, as well as the effectiveness of this heuristic, I generated a non-binary species tree for each of my datasets (Figure 2.5).

For the cyanobacterial tree (Figure 2.5(a)), node n18 was removed because the branch from n20 to n18 is short and it is associated with substantial gene tree incongruence, such that it is an area of the species tree where ILS may be occurring. In yeast (Figure 2.5(b)), the branching order of *S. bayanus* and *S. kudriazevii* has been controversial and one study reported evidence of hybridization in these species (Yu et al., 2013). The edge from n15 to n13 was, therefore, replaced by a polytomy in the yeast species tree (Figure 2.5(b)). The gene trees in both datasets were reconciled with these non-binary species trees to obtain the DTI and DTLI statistics.

When the models with and without ILS are compared, a substantial decrease in the combined number of duplications and transfers was observed, ranging from 15% to 18% in cyanobacteria and from 11% to 14% in yeast. In addition, considerable decreases in the number of losses were observed. In fact, the number of inferred losses with the DTI model, compared to the DT model, decreased by as much as 20%. These differences highlight the extent to which ignoring other sources of incongruence could lead to overestimation of other events.

Comparison with other analyses: I also compared the statistics obtained with Notung to patterns of transfers reported in a study using Quartet Decomposition (QD) (Bansal et al., 2011). Bansal et al. (2011) focused on "highways" of horizontal transfer, that is, pairs of genomes associated with a particularly high degree of genetic exchange (Beiko et al., 2005). Bansal applied the QD method to the same dataset of 1128 cyanobacterial trees and found four HGT highways, shown as dotted lines in Figure 2.7.

In my analysis, hotspots of HGT activity were also apparent in the cyanobacterial data, as shown by the HGT heatmaps (Figure 2.8 and 2.9). To obtain a quantitative definition of an HGT highway, I termed an HGT highway to be a pair of species between which the total number of transfers, in both directions, exceeds two standard deviations above the mean of transfer counts between all pairs of species. The distributions of transfer counts between cyanobacterial genome pairs inferred with the DTL and DTLI models are shown in Figure 2.6. Using my qualitative definition, there are three such HGT highways for DTL model, and 1 case for DTLI model, shown in orange and blue, respectively, in Figure 2.7. The HGT traffic inferred in my analysis with the DTL model is similar to the HGT highways reported by Bansal et al. (2011) (dotted lines in Figure 2.7), for the same dataset. However, when events were inferred with the DTLI model, the elevated transfer counts in the *Gloeobacter*



(a) DTL model

(b) DTLI model

Figure 2.6: Distribution of transfer counts between pairs of genomes under (a) DTL model and (b) DTLI model. The red lines indicate the cutoff at 2 standard deviation above the mean.

group disappeared, resulting a single pair of genomes with high HGT activities (blue line). This further demonstrates the extent to which model choice can influence the biological conclusions of a study. Some of the HGT highways reported by Bansal et al. (2011) may be artifactual. Alternatively, the DTLI heuristic may be eliminating a signal of true horizontal transfer. From a conservative perspective, we could argue that the HGT highway between *Prochlococcus* 3 (MIT) and *Synechococcus* is the most robust of these results.

2.6 Chapter Conclusion

In this chapter, I described my contributions to horizontal gene transfer inference using gene tree-species tree reconciliation. My two major contributions are an algorithm for testing temporal feasibility and case studies that probe the challenges of phylogenetic transfer inference.

One of the major challenges associated with an event model that includes transfers is degeneracy. When the event model includes transfers, the minimum cost event history is

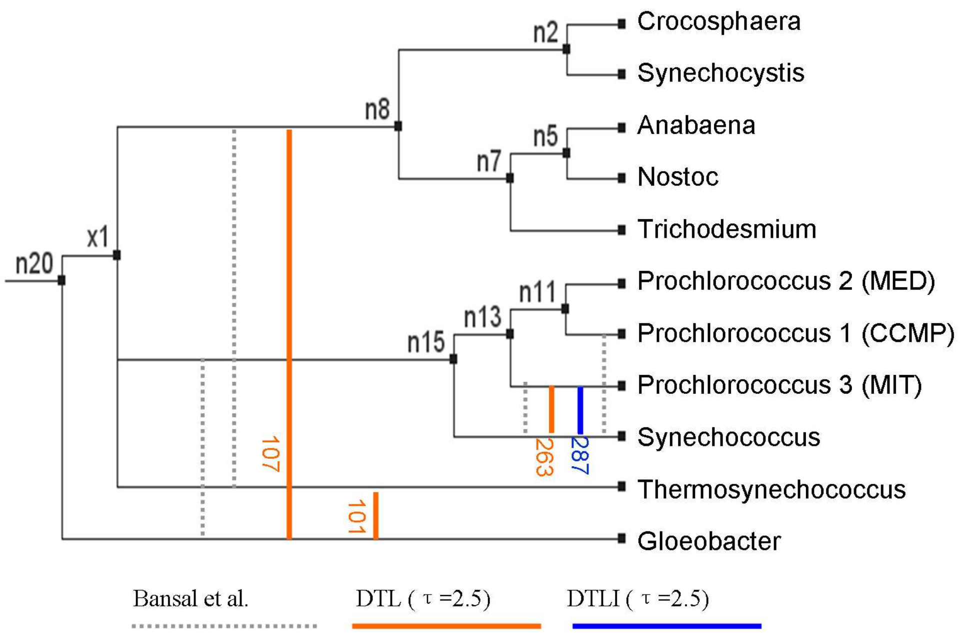
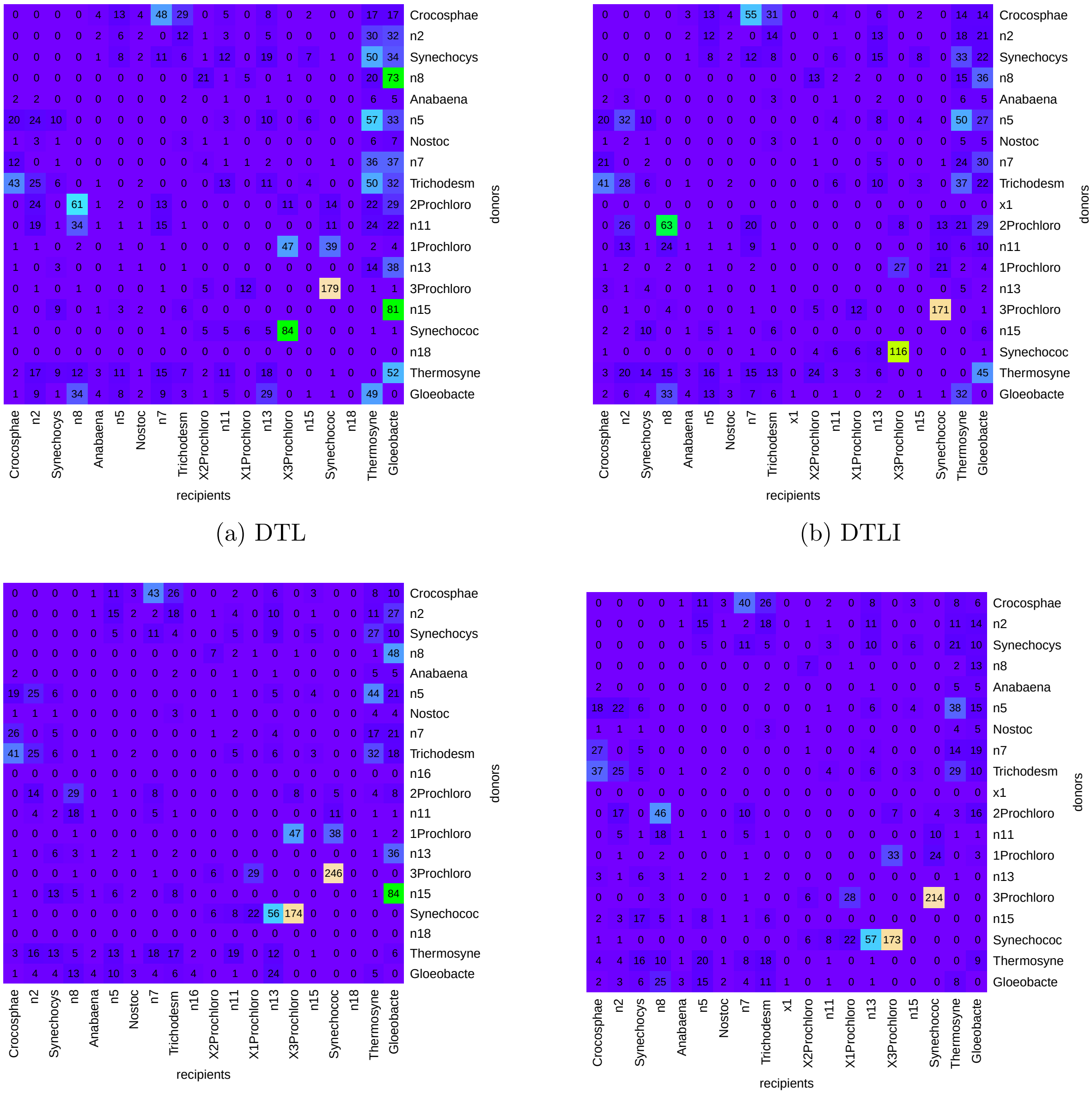


Figure 2.7: High HGT activities between pairs of genomes were detected, using the DTL and DTLI models with $\delta = 3$, $\tau = 2.5$ and $\lambda = 2$. The internal edge n16-n18 was collapsed for the DTLI model. Genome pairs with transfer counts greater than 2 standard deviations above the mean are shown, with the total number of transfers labeled. HGT Highways predicted by Bansal et al. (2011) are shown as dashed lines.



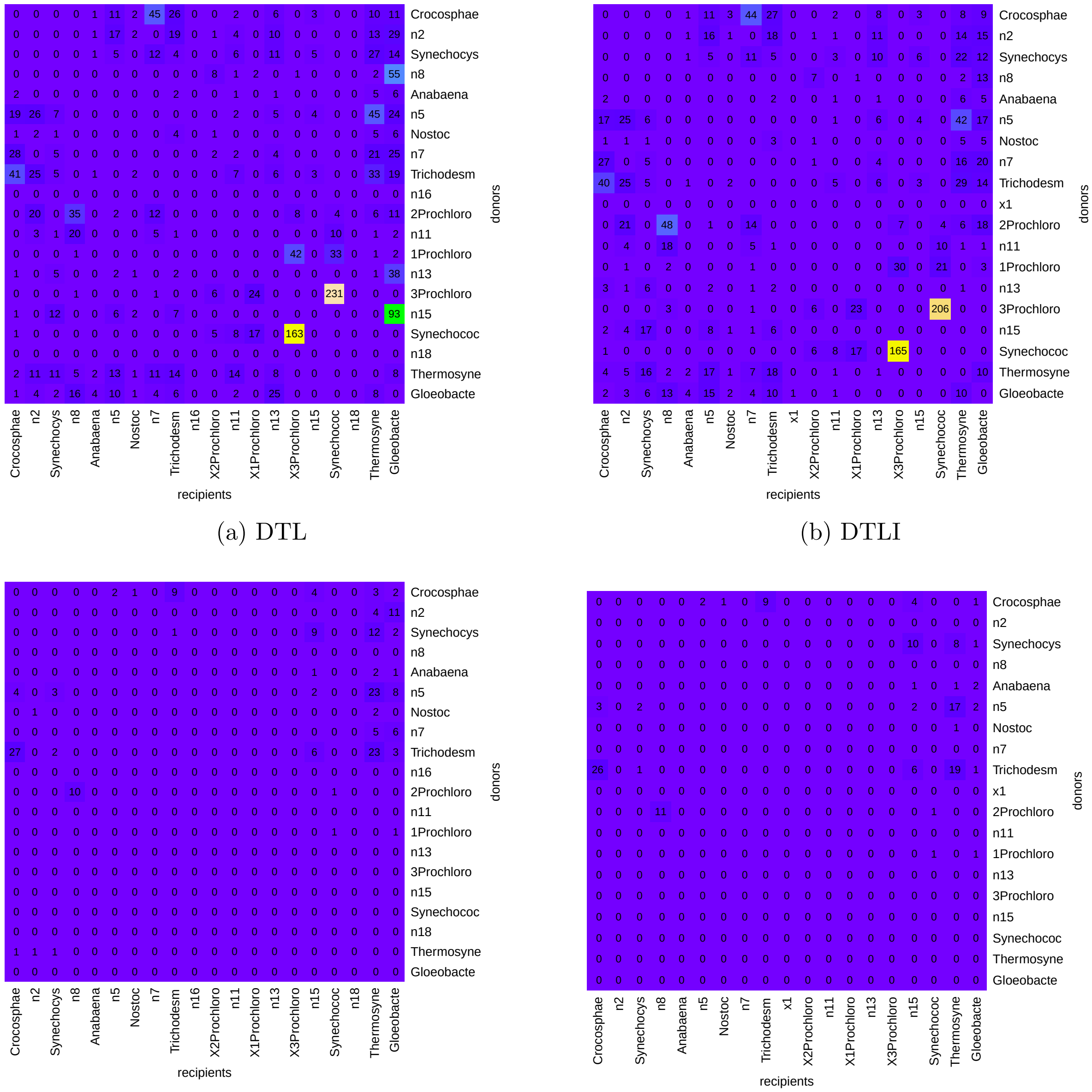
(a) DTL

(b) DTLI

(c) DT

(d) DTI

Figure 2.8: Transfers in cyanobacteria, inferred with $\delta = 3$, $\lambda = 2$ and $\tau = 2.5$ under *(a)* the DTL-model, *(b)* DTLI-model, *(c)* the DT-model *(d)* DTI-model.



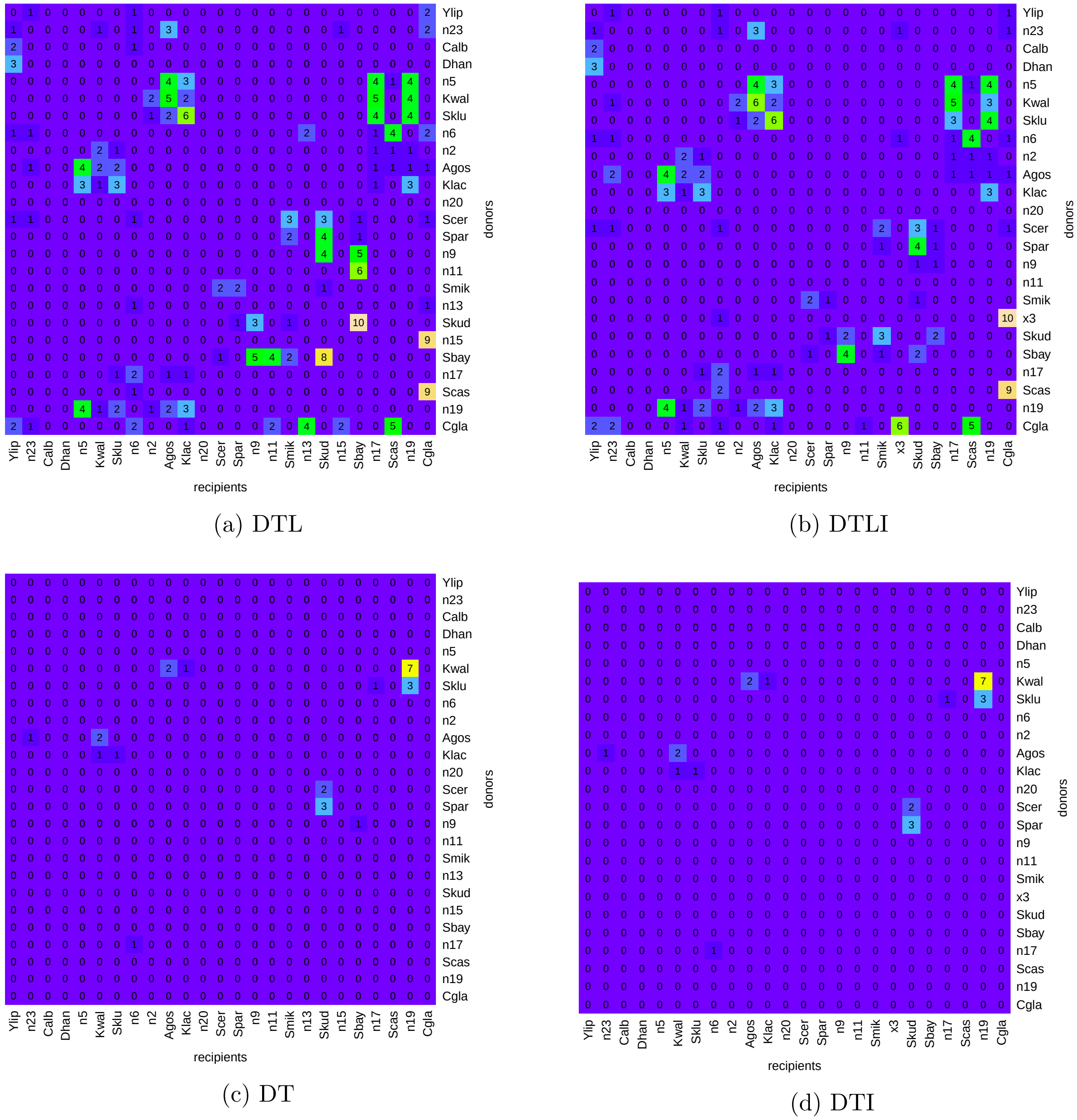
(a) DTL

(b) DTLI

(c) DT

(d) DTI

Figure 2.9: Transfers in cyanobacteria, inferred with $\delta = 3$, $\lambda = 2$ and $\tau = 6$ under *(a)* the DTL-model, *(b)* DTLI-model, *(c)* the DT-model *(d)* DTI-model.



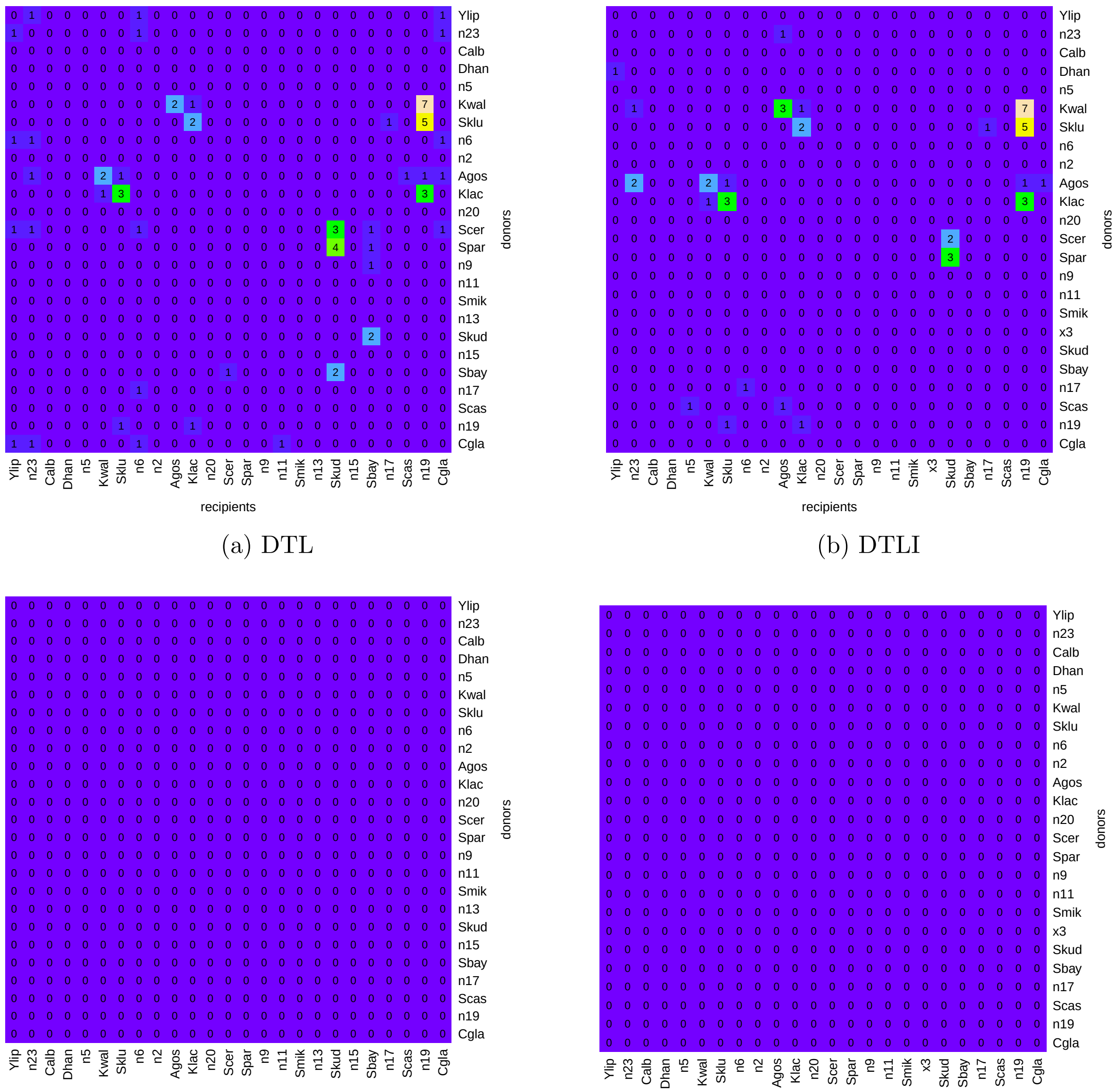
(a) DTL

(b) DTLI

(c) DT

(d) DTI

Figure 2.10: Transfers in yeast, inferred with $\delta = 3$, $\lambda = 2$ and $\tau = 6$ under *(a)* the DTL-model, *(b)* DTLI-model, *(c)* the DT-model *(d)* DTI-model.



(a) DTL

(b) DTLI

(c) DT

(d) DTI

Figure 2.11: Transfers in yeast, inferred with $\delta = 3$, $\lambda = 2$ and $\tau = 6$ under *(a)* the DTL-model, *(b)* DTLI-model, *(c)* the DT-model *(d)* DTI-model.

not, in general, unique. However, many of the algorithms and software packages that have been developed for reconciliation with transfers ignore the problem of degeneracy. With the exception of Notung and EUCALYPT, the algorithms and software cited in Table 2.1, report only one of possibly many optimal solutions.

Yet my case studies suggest that multiple optimal solutions are a frequent occurrence, especially in datasets where transfer is the dominant process. In the analysis reported here, 20% of 1128 cyanobacterial trees had multiple optimal solutions with inconsistent event histories. In other words, for one in five trees, the arbitrary selection of a single optimal solution could lead to conclusions that might not be supported by other optimal solutions.

A second major challenge is temporal infeasibility. One approach to this problem has been to restrict the problem to dated species trees, in which temporally compatible donor-recipient pairs can be readily identified. This requires accurate inference of species trees with branch lengths that are proportional to time.

If the molecular clock hypothesis holds (i.e., if the rate of substitutions is the same in all lineages), then branch lengths in substitutions per site can be used. However, this hypothesis does not hold for most data sets. The development of relaxed molecular clock methods is an active research area (Drummond et al., 2006). These require a significant amount of computation time, rely on model assumptions, and result in very large confidence intervals.

An alternate approach is to estimate dates based on the fossil record. However, fossil dating can be very inaccurate (Drummond et al., 2006; Yang and Rannala, 2006) and only works in cases where fossil evidence is available. This is particularly a problem with prokaryotic data.

Some methods simply ignore temporal infeasibility (Libeskind-Hadas et al., 2014; Wu et al., 2013) based on the argument that it rarely arises in practice. Previous empirical studies have reported that temporal infeasibility is rare (Tofigh et al., 2011) or non-existent (Hallett et al., 2004). My results contradict this assumption as well. Approximately 10% of the trees in my study had no temporally feasible solution.

To address this problem, I developed criteria for recognizing temporally infeasible histories under the DTL-event model. My approach is similar to the infeasibility checking scheme proposed by Tofigh et al. (2011) for the less restrictive DT model, in which losses are not

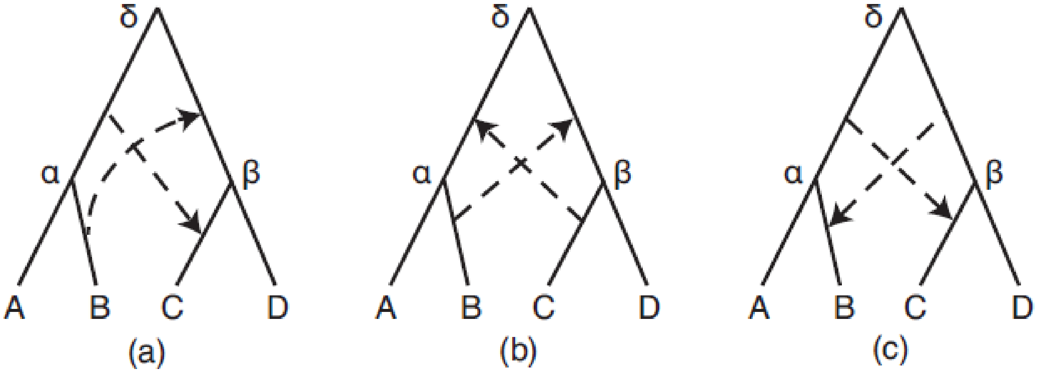


Figure 2.12: (a-c) Three examples of temporally infeasible transfer pairs on a hypothetical species tree with four leaves. Dashed arrows correspond to inferred transfers

penalized, but imposes additional constraints. For example, Figure 2.12 shows three scenarios that are temporally infeasible under the DTL model. All three violate the constraints introduced in Section 2.4. However, only Figure 2.12(b) violates the constraints proposed by Tofigh et al. (2011).

Tofigh's constraints are appropriate for applications where only the number of duplications and transfers is inferred, but the specific events are not of interest. Under the DT model, there exist certain event histories that are infeasible, but for which a temporally feasible reconciliation can be identified that has the same number of duplications and transfers, but more losses. Since losses incur no cost under the DT model, this feasible reconciliation has the same cost as the original, infeasible history. For example, the infeasible histories in Figures 2.12(a) and (c) can be converted into feasible histories by lifting the transfer recipient on (b, C) so that it enters the edge (d, b), above the other transfer. This operation generates a new history that is temporal feasibility, but incurs an additional loss in species D. For the purposes of inferring only the optimal reconciliation cost, it is not necessary to construct the feasible history, but simply to verify that it exists.

Therefore, the constraints proposed by Tofigh et al. (2011) do not rule out infeasible histories for which there exists a feasible history with the same DT cost. This is sufficient if only the cost is of interest, but insufficient for applications where the goal is to infer the donor and recipient species of specific transfers. Similarly, Tofigh's constraints are not appropriate for counting the number of optimal solutions, because the set of optimal solutions reported under these constraints will include infeasible solutions like those in Figures 2.12(a) and (c), as well as their feasible counterparts. This will lead to an overestimate of the number of valid, optimal reconciliations. My approach does not fall prey to these shortcomings.

The literature on methods for reconciliation with transfers has approached these challenges from various perspectives and there is no consensus on the importance of these problems.

Reconciliation algorithms for inferring horizontal transfer could profitably be generalized in several ways.

First, most algorithms and software use a single event cost for all inferred events of each event type. For instance, in an inferred reconciliation scenario, the loss of a 16S rRNA gene in *E.coli* has the same cost as the loss of a gene for cellular motility. However, the potential fitness impact of losing an essential gene like 16S rRNA is much higher. One direction for future work is to address the variation in fitness impact by assigning different event costs to different gene families. Another possibility is a transfer cost that is a function of evolutionary divergence, niche similarity, or the similarity of the genomic nucleotide composition between donor-recipient species to model the difficulty of such inferred transfers.

Second, in current reconciliation algorithms, events are inferred independently. However, a region of the chromosome containing multiple of genes may be transferred in a single event. A challenge in identifying these cases is that foreign genes that originated through a single, large-scale transfer event are not always contiguous. This can occur, for example, when a large chunk of foreign DNA is digested into smaller pieces that are integrated into different parts of the chromosome. Reconciliation models that consider multiple gene trees simultaneously are needed to handle large-scale transfers. Expanded event models are also needed. For example, horizontal transfer may be a single homologous replacement event, instead of distinct loss and transfer events as modeled in current DTL-reconciliation algorithms.

Finally, phylogenetic noise may be present due to the errors in the gene and species phylogenies, leading to erroneous inference of the event history. Currently in Notung, we partially address this by the gene tree rearrangement procedure (Chen et al., 2000). If a gene tree branch is weakly supported, as defined by branch support, Notung can rearrange the topology at that branch to better match the topology of the species tree. Similarly, Mowgli-NNI rearranges edges with weak bootstraps. It does this with NNI operations. The AnGST software deals with phylogenetic uncertainties in gene tree by selecting bootstrap subtrees that minimize costs (David and Alm, 2011). However, both procedures assume that the species tree is error-free, and cannot address phylogenetic noise from the species tree. Appropriate handling of phylogenetic noise is an open question for future research.

Chapter 3

Highly Iterated Palindrome-1 (HIP1) motifs

In this chapter, I systematically characterize the abundance, enrichment and conservation of the HIP1 motif in 40 cyanobacterial genomes in the NCBI dataset. My focuses include: (1) What is the taxonomic distribution of HIP1 among cyanobacterial genomes? (2) What is the abundance when compared to genomic background composition? (3) Is HIP1 much more abundant than other motifs?

In 20 of those genomes, I establish that the HIP1 motif is more abundant than expected by chance. Using a comparative genomic approach, I investigate whether HIP1 prevalence is maintained by a neutral process or by selection acting on the motif. I further consider whether selection acts on specific HIP1 positions or on HIP1 content in a local region of the genome.

3.1 HIP1 Frequency and phylogenetic distribution

In order to characterize the distribution of the HIP1 motif in the 40 genomes in my study (see Table 1.1), and to investigate its frequency relative to other palindromes of length 8, I

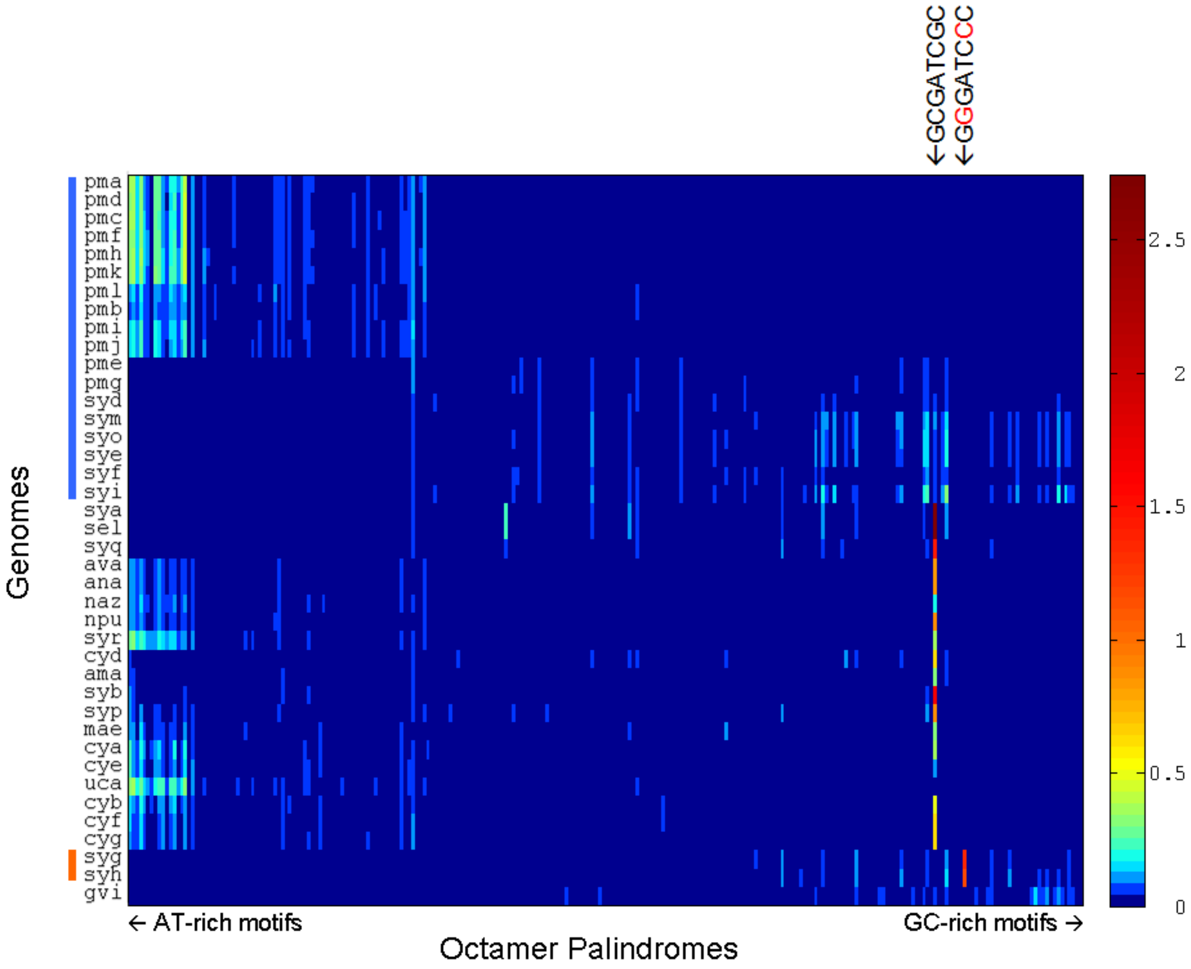


Figure 3.1: The frequency (count/kbp) of all octamer palindromes in 40 cyanobacterial genomes. Color map is based on a log transformation of the motif frequencies. The blue vertical bar beside the genome name indicates the marine pico-cyanobacteria lineage. The orange vertical bar indicates the two Yellowstone strains. The color bands corresponding to the 5'-GCGATCGC-3' (HIP1) and 5'-GGGATCCC-3' motifs are indicated above the heat map.

searched all chromosomal sequences in the NCBI dataset for all octamer palindrome motifs, using in-house Perl scripts. For each genome, the frequency in motifs per kilo basepair, of each of the 64 possible octamer palindromes was calculated. These are summarized by a heatmap in Figure 3.1. The HIP1 motif frequencies in the 40 genomes are shown in Table 3.1.

As shown in Figure 3.1, the canonical HIP1 motif (5'-GCGATCGC-3') has moderate to very high abundance in all genomes, with some notable exceptions including *Gloeobacter violaceus* PCC7421 (*gvi*), cyanobacterium UCYN-A (*uca*), the marine pico-cyanobacteria, and the Yellowstone strains *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*). Comparing this heatmap with the genome-wide GC content (Table 1.1), abundance of AT-rich motifs can be observed in genomes with low GC content, such as most *Prochlococcus* (GC% < 40), many *Cyanothece* (GC% < 45), and many Nostocaceae species (GC% < 45). Interestingly, many of the non-pico AT-rich genomes, such as *Cyanothece* and Nostocaceae genomes, also have high HIP1 frequency. AT-rich motifs are generally low frequency in GC-rich genomes, including pico and non-pico *Synechoccocus*, *Gloeobacter violaceus* PCC7421 (*gvi*), and *Acaryochloris marina* MBIC11017 (*ama*). In addition, a few moderately AT-rich motifs have somewhat elevated frequency in most genomes. Some of these AT-rich motifs may be associated with the Pribnow boxes (5'-TATAAT-3' or variants), that are commonly found in bacterial promoter regions in bacteria. Similarly, GC-rich motifs are moderately elevated in GC-rich genomes, such as some *Synechococcus* strains and *Gloeobacter.* Interestingly, we found that the most abundant octamer palindrome in *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*), two closely related, thermophilic cyanobacterial strains isolated from a hot spring in Yellowstone National Park (Steunou et al., 2006), is 5'-GGGATCCC-3'. This motif highly resembles the canonical HIP1 motif. In fact, this motif and canonical HIP1 are the only ocatmer palindromes with frequency exceeding 0.7 in the 40 cyanobacterial genomes. On the other hand, the frequency of the canonical motif (5-GCGATCGC-3) was zero in those two genomes. This raised the question whether motif 5'-GGGATCCC-3' could be a variant form of the HIP1 motif, playing the same role in the Yellowstone strains that the canonical HIP1 motif (5'-GCGATCGC-3') plays in cyanobacterial species in which the canonical HIP1 motif is highly abundant.

Genome	HIP1 count	HIP1 frequency	HIP1 Enrichment
pma	2	0	1.23
pmd	2	0	1.29
pmc	4	0	2.40
pmf	2	0	1.32
pmh	2	0	1.39
pmk	1	0	0.78
pml	3	0	0.85
pmb	1	0	0.24
pmi	3	0	0.87
pmj	4	0	1.13
pme	104	0.04	1.59
pmg	94	0.04	1.47
syd	165	0.06	1.78
sym	256	0.11	1.49
syo	157	0.06	1.33
sye	151	0.06	1.09
syf	144	0.06	1.44
syi	189	0.08	1.11
sya	7356	2.73	26.36
sel	7402	2.75	26.46
syq	3681	1.42	53.10
ava	5239	0.82	98.48
ana	5260	0.82	87.54
naz	1105	0.21	63.32
npu	7151	0.87	80.29
syr	2908	0.38	302.50
cyd	3397	0.63	39.06
ama	2147	0.33	25.21
syb	5084	1.69	59.50
syp	3160	0.88	133.05
mae	1821	0.31	24.32
cya	2252	0.38	76.78
cye	647	0.11	13.17
uca	37	0.03	28.46
cyb	2392	0.48	152.65
cyf	2959	0.63	97.66
cyg	2980	0.64	97.48
syg	57	0.02	0.69
syh	67	0.02	0.70
gvi	318	0.07	0.58
syg	4099	1.35	18.60
syh	3401	1.16	16.75

Table 3.1: Genome-wide HIP1 motif count, frequency (motifs per kbp), and enrichment in the 40 genomes. The statistics reported in the last two rows are based on the motif 5'-GGGATCCC-3'.

3.2 HIP1 Enrichment

The high abundance of HIP1 motifs in cyanobacterial genomes is intriguing, but could result from the underlying sequence composition. The observed high frequency of HIP1 motifs might not be significantly different from the expected frequency resulting from the background sequence composition. In order to rule out this possibility, I calculated HIP1 motif enrichment, which is the ratio between the observed and expected number of motifs[1].

The expected number of HIP1 motifs was calculated using a second order Markov model of sequence composition, which accounts for the background tri-nucleotide frequency (Karlin and Brendel, 1992). Let $W = w_1 w_2 ... w_n$ be a motif of length n, and let $E(W)$ be the expected number of instances in a region of length L. Then $E(W)$ can be approximated using a second order Markov model as follows:

$$\begin{aligned} E(W) &= P(W)L \\ &\approx P(w_1 w_2)P(w_3|w_1 w_2)P(w_4|w_2 w_3)...P(w_n|w_{n-2}w_{n-1})L, \end{aligned} \tag{3.1}$$

where $P(w_i w_{i+1} ... w_{i_R})$ is the frequency of string $w_i w_{i+1} ... w_{i_R}$ in the genome and $P(w_{i+2}|w_i w_{i+1})$ is the conditional probability of observing w_{i+2} given that the previous two bases were w_i and w_{i+1}. This conditional probability can be calculated using di- and tri-nucleotide counts:

$$P(w_{i+2}|w_i w_{i+1}) = \frac{N(w_i w_{i+1} w_{i+2})}{N(w_i w_{i+1})}, \tag{3.2}$$

where $N(s)$ is the number of instances of string s in the region of length L. Substituting the right hand side of Equation (3.2) for each of the conditional probabilities in Equation (3.1), we obtain

$$E(W) \approx \frac{N(w_1 w_2 w_3)N(w_2 w_3 w_4)...N(w_{n-2}w_{n-1}w_n)}{N(w_2 w_3)N(w_3 w_4)...N(w_{n-2}w_{n-1})}. \tag{3.3}$$

[1] $\frac{O}{E}$ is used, instead of $\frac{(O-E)^2}{E}$, as the motif enrichment. This is because $\frac{(O-E)^2}{E}$ is directional and can not distinguish over-representation from under-representation of motifs.

I estimated HIP1 enrichment using the approximation for the expected number of motifs given in Equation 3.2. Since background frequencies in coding and non-coding regions can differ substantially, I calculated the expected number of motifs in coding and non-coding regions separately.[2] When estimating the expected number of motifs in coding regions, the expected number of HIP1 motifs in each reading frame was calculated separately using the background frequency from that reading frame. For example, the expected number of HIP1 motifs in Reading Frame 0 ($RF0$), using the tri-nucleotide background frequency model (2nd-order Markov model), can be estimated by:

$$E_0(W) \approx \frac{N_0(w_1 w_2 w_3) N_1(w_2 w_3 w_4) \ldots N_f(w_{n-2} w_{n-1} w_n)}{N_1(w_2 w_3) N_2(w_3 w_4) \ldots N_f(w_{n-2} w_{n-1})}, \tag{3.4}$$

where $N_f(w_i w_{i+1} w_{i+2})$ is the number of instances of the tri-nucleotide $w_i w_{i+1} w_{i+2}$, when w_i is in codon position f ($f = 0, 1, 2$). Similar expression can be derived for the expected number of motifs in Reading Frame 1 ($RF1$) and Reading Frame 2 ($RF2$). The expected number of motifs in coding regions within a given genome is the sum of the expected number of motifs in each of the three reading frames:

$$E_{CDS}(W) = E_0(W) + E_1(W) + E_2(W). \tag{3.5}$$

The expected number of motifs genome-wide is the sum of expected number of motifs from coding and non-coding regions:

$$\begin{aligned} E_{GW}(W) &= E_{CDS}(W) + E_{NC}(W) \\ &= E_0(W) + E_1(W) + E_2(W) + E_{NC}(W), \end{aligned} \tag{3.6}$$

where E_{CDS} and E_{NC} denote the expected number of motifs in Open Reading Frames (ORFs) and in intergenic regions, respectively. Intergenic regions are defined to be regions between ORFs that do not contain other annotated elements, such as RNA genes or transposons,

[2] For the definition of coding and non-coding regions, please refer to the Methods section at the end of this chapter.

and that do not exceed given length threshold. The upper limit on the length of intergenic regions was imposed to rule out unannotated elements such as transposons and long repetitive sequences. I experimented with several criteria for the inclusion of the intergenic regions, as described in Section 3.8.3. While specific numbers obtained with the different criteria varied, the overall trends were unaffected.

The enrichment, $\mathcal{E}(W)$, is the ratio of observed number of motifs to expected number of motifs:

$$\mathcal{E}(W) = \frac{O(W)}{E(W)}, \tag{3.7}$$

where $O(W)$ is the number of observed instances of W in the region of interest.

The enrichment of each octamer palindromes in 40 cyanobacterial genomes was calculated genome-wide and for coding and intergenic regions. Table 3.2 summarizes the enrichment results. Overall, the enrichment clearly indicates that HIP1 motif abundance is not due to genome composition. The enrichment is less than 16 in all marine pico-cyanobacteria genomes and less than 5 in most of them. In *Gloeobacter violaceus* PCC7421 (*gvi*), the number of observed motifs is slightly lower than expected. This is consistent with the conclusion that HIP1 is absent from the marine pico-cyanobacteria and *Gloeobacter violaceus* PCC7421 (*gvi*). This is also consistent with the observations from Delaye et al. (2011b). Therefore, these 19 genomes will not be considered further. Interestingly, it is observed that the enrichment in intergenic regions is often higher than in coding regions. This can be partially explained by the fact that bacterial intergenic regions are more AT-rich in general, as GC content in intergenic regions are observed to be 5%-10% lower than in coding regions in bacterial genomes (Brocchieri, 2014).

Figure 3.2 shows the abundance and enrichment of all octamer palindromes that are enriched by a factor of 5 or greater and occur at least 10 times in one or more genomes[3]. Of 40 × 64 = 2560 motif-genome pairs, only 42 meet both criteria. Of the 42 cases that meet these criteria, almost half are instances of canonical HIP1 motif. Further, both the enrichment and abundance of HIP1 dwarfs most other motif-genome pairs in Figure 3.2. The exceptions are 5'-TAGTACTA-3' in *Synechococcus sp.* WH8102 (*syo*) and 5'-GGGATCCC-3'

[3] To efficiently compute the enrichment for all the 8-mer motifs, expected number of motifs were estimated using the genome-wide tri-nucleotide frequencies.

	Observed				Expected				O/E	
id	intergenic	RF0	RF1	RF2	intergenic	RF0	RF1	RF2	intergenic	coding
pma	0	1	1	0	0.12	0.51	0.50	0.50	0.00	1.32
pmd	1	0	1	0	0.09	0.45	0.50	0.51	10.86	0.69
pmc	2	1	1	0	0.13	0.52	0.51	0.51	15.62	1.30
pmf	0	1	1	0	0.12	0.46	0.45	0.48	0.00	1.44
pmh	0	2	0	0	0.11	0.42	0.45	0.46	0.00	1.51
pmk	1	0	0	0	0.10	0.37	0.47	0.35	10.27	0.00
pml	0	0	2	1	0.29	1.16	1.05	1.01	0.00	0.93
pmb	0	1	0	0	0.27	1.26	1.30	1.27	0.00	0.26
pmi	1	2	0	0	0.26	1.15	0.97	1.06	3.84	0.63
pmj	1	2	1	0	0.27	1.03	1.14	1.11	3.77	0.91
pme	7	45	43	9	3.60	20.96	21.03	19.99	1.94	1.56
pmg	7	40	38	9	4.70	19.32	19.97	20.07	1.49	1.47
syd	5	90	60	10	4.30	30.86	29.67	27.90	1.16	1.81
sym	5	132	106	13	5.23	56.61	56.26	53.18	0.96	1.51
syo	8	74	64	11	5.54	55.51	50.99	6.32	1.44	1.32
sye	7	83	58	3	7.64	42.91	43.81	43.55	0.92	1.11
syf	6	76	53	9	4.53	32.59	32.89	29.88	1.33	1.45
syi	5	73	107	4	3.60	56.24	56.54	53.74	1.39	1.10
sya	1008	2679	2484	1185	15.75	90.53	86.08	86.75	64.00	24.10
sel	907	2755	2539	1201	12.81	90.17	89.08	87.65	70.82	24.34
syq	227	1192	1769	493	2.76	22.75	21.49	22.32	82.29	51.89
ava	474	2444	1575	746	2.97	16.59	16.92	16.72	159.41	94.86
ana	457	2412	1622	769	3.11	28.33	25.77	2.88	147.08	84.29
naz	135	538	304	128	3.32	7.27	6.13	0.73	40.69	68.61
npu	1045	3136	1985	985	8.57	27.12	26.21	27.16	121.92	75.86
syr	978	1039	664	226	1.98	2.58	2.76	2.29	494.08	252.75
cyd	243	1827	874	453	5.16	27.39	28.12	26.29	47.05	38.56
ama	97	1019	750	281	4.99	26.59	26.05	27.52	19.45	25.57
syb	527	2707	1423	427	3.73	29.83	25.96	25.93	141.24	55.77
syp	280	1978	745	157	1.53	7.25	7.11	7.86	182.97	129.64
mae	247	981	454	139	6.21	23.23	22.87	22.57	39.77	22.92
cya	238	1212	597	205	2.21	8.90	9.11	9.11	107.84	74.26
cye	44	412	153	38	2.76	26.99	17.49	1.89	15.93	13.00
uca	2	25	8	2	0.08	0.39	0.40	0.43	25.39	28.67
cyb	122	1094	854	322	0.70	5.06	5.09	4.82	173.69	151.60
cyf	256	1255	1029	419	1.61	10.09	9.45	9.15	158.80	94.20
cyg	241	1283	1040	416	1.49	9.88	9.80	9.40	161.64	94.18
syg	803	1036	875	1385	29.88	36.92	87.29	66.34	26.87	17.30
syh	623	879	700	1199	29.67	32.87	81.35	59.17	21.00	16.02
gvi	12	191	107	8	15.39	175.07	174.19	182.83	0.78	0.58

Table 3.2: HIP1 motif enrichment. The HIP1 statistics reported for genome *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) are based on the HIP1 variant 5'-GGGATCCC-3'.

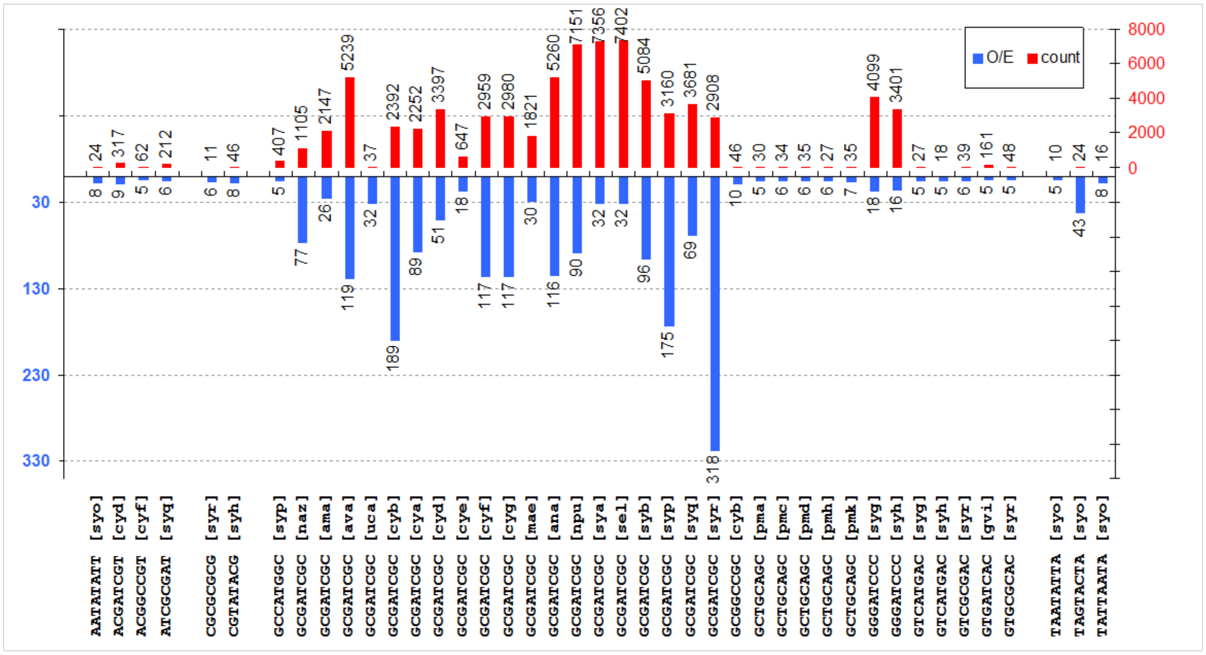


Figure 3.2: Enrichment (blue) and abundance (red) for all instances of an 8-mer palindrome that is enriched by a factor of 5 or more and has at least 10 copies in one or more genomes.

in the Yellowstone strains. In all other genomes, HIP1 is by far the most abundant motif and it is also more abundant than expected by a factor of at least 20.

There are 24 instances of the AT-rich motif 5'-TAGTACTA-3' in *Synechococcus sp.* WH8102 (*syo*); it has an enrichment of 43. However, an in-depth investigation of the motif 5'-TAGTACTA-3' in the *Synechococcus sp.* WH8102 (*syo*) genome revealed that 19 out of 24 motif occurrences are clustered in a 10 kbp window within a single annotated ORF. This particular genomic region is likely a localized genome-specific repetitive region, and thus the intra-genome distribution of 5'-TAGTACTA-3' is very different from that of the HIP1 motif. As a result, I did not consider this particular AT-rich octamer pattern further.

The Yellowstone strains have 4099 and 3401 instances of the motif 5'-GGGATCCC-3', respectively. This is comparable to the abundance of the canonical HIP1 motif in *Thermosynechococcus elongatus* BP-1 (*syq*) and *Cyanothece sp.* PCC7425 (*cyd*). Similarly, the enrichments of 5'-GGGATCCC-3' in *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) are 18 and 16, respectively, comparable to that of *Cyanothece sp.* PCC7425 (*cyd*). A closer look at the intra-genomic motif distribution also revealed that

5'-GGGATCCC-3' is broadly distributed across the genomes of *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*). Taken together, these observations support the hypothesis that 5'-GGGATCCC-3' is a variant of the HIP1 motif in the two Yellowstone strains. Subsequent analyses will be applied to this variant in *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) and *Synechococcus sp.* JA-3-3Ab (*syh*), and to the canonical form in other genomes. I will use the term "HIP1" to refer to 5'-GGGATCCC-3' in *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) and *Synechococcus sp.* JA-3-3Ab (*syh*), unless otherwise noted.

Comparison of HIP1 enrichment and frequency for all 40 genomes (Figure 3.3) shows that the genomes with the highest HIP1 frequency (*Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*)) do not have the highest enrichment, suggesting a substantial number of HIP1 motifs in those genomes are contributed by background composition. On the other hand, *Trichodesmium erythraeum* IMS101 (*syr*), a genome with moderate HIP1 frequency, shows the highest HIP1 enrichment. In addition, two genomes (*Cyanothece sp.* PCC 7822 (*cye*) and cyanobacterium UCYN-A (*uca*)) with very low HIP1 frequency did not appear to be HIP1-rich. However, HIP1 enrichment in these genomes is 13 and 28, respectively. Note that there are only 37 instances of the HIP1 motif in the cyanobacterium UCYN-A (*uca*) genome, a reduced genome lacking many biological pathways. As it has been shown that binding sites rapidly degrade following the loss of the corresponding transcription factor (Moses et al., 2006), it is possible that the mechanism responsible for HIP1 prevalence is no longer present in cyanobacterium UCYN-A (*uca*) genome, and that HIP1 motifs in cyanobacterium UCYN-A (*uca*) are undergoing a degeneration process. However, even though there are only a few HIP1 motifs in the cyanobacterium UCYN-A (*uca*) genome, the number is much greater than the expected number, suggesting those HIP1 motifs are not a chance occurrence. Given that 19 out of 24 5'-TAGTACTA-3' motifs in *Synechococcus sp.* WH8102 (*syo*) are clustered in a 10 kbp window, I investigated the spatial distribution of the 37 HIP1 motifs within the cyanobacterium UCYN-A (*uca*) genome to determine whether those 37 motif instances are located within a single genomic region. That would suggest that those HIP1 motifs were transferred into cyanobacterium UCYN-A (*uca*) from a HIP1-rich cyanobacterial genome. Interestingly, the spatial distribution reveals that those 37 HIP1 motifs are relatively evenly distributed along the cyanobacterium UCYN-A (*uca*) chromo-

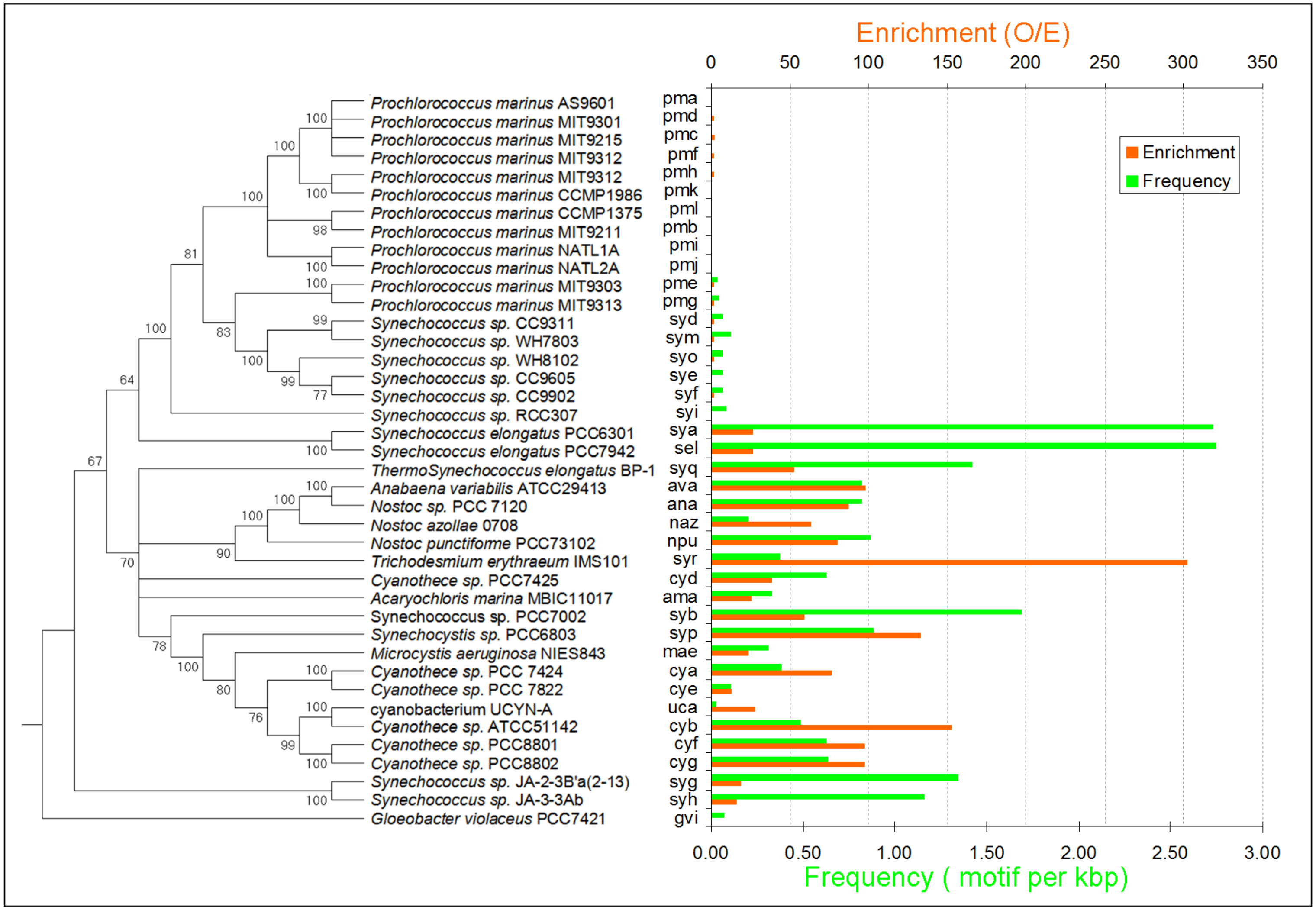


Figure 3.3: HIP1 enrichment compared with HIP1 frequency across the 40 cyanobacterial genomes in the NCBI dataset. The species phylogeny shown on the left is a maximum likelihood tree based on the concatenation of the DNA alignments of 16S and 23S rRNA genes. The tree is rooted at *Gloeobacter violaceus* PCC7421 (*gvi*), and branch supports are based on 100 bootstraps. Weakly supported edges (with < 60 bootstrap value) are collapsed.

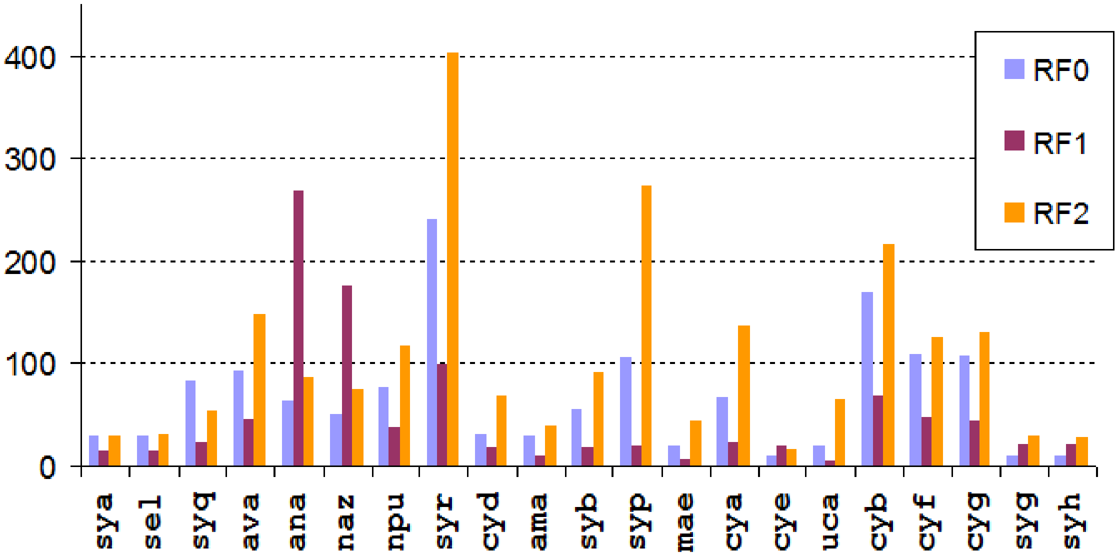


Figure 3.4: HIP1 enrichment among the three reading frames in HIP1-rich genomes.

some (data not shown), suggesting that it is unlikely that those HIP1 occurrences are all of foreign origin. To summarize, these results shown in Figure 3.2 and Table 3.1 indeed suggest that the high observed HIP1 frequency cannot be explained by the background sequence composition.

The dependence of HIP1 enrichment on reading frame was further investigated. When considering HIP1 enrichment in individual reading frames, it was observed that the HIP1 distributions among the three reading frames vary within a relatively large range (Figure 3.4). Overall, the distribution of reading frames shows a marked preference for RF2. In 17 out of 21 genomes, RF2 is most enriched. In the remaining four genomes, RF1 is enriched in three genomes (*Cyanothece sp.* PCC 7822 (*cye*) and the Nostoc strains *Nostoc sp.* PCC 7120 (*ana*) and *Nostoc azollae* 0708 (*naz*)) and RF0 is enriched in one (*Thermosynechococcus elongatus* BP-1 (*syq*)). In most genomes (16/21), RF0 ranks second for enrichment. A G-test, based on contingency tables with a phylogenetic correction, shows that this distribution deviates significantly from a uniform distribution (see Table 3.3). Details on the phylogenetic correction used to address the phylogenetic dependency among the genomes are described in the Methods section on page 119. As *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) have the variant form of HIP1, they were excluded from the G-

	RF0	RF1	RF2
Observed	1.88	2.12	14.01
Expected	6	6	6

(a) G-test p-value: 5.59×10^{-4}

	RF0	RF1	RF2
Observed	1.17	2.48	15.35
Expected	6.33	6.33	6.33

(b) G-test p-value: 9.22×10^{-5}

Table 3.3: Contingency tables showing the number of genomes in which each reading fram is most enriched, corrected for the phylogenetic dependency. (a) Phylogenetic correction based on the species phylogeny from the Barker dataset (Figure 1.1). cyanobacterium UCYN-A (*uca*) is not represented in the Barker tree and is excluded from this table. (b) Phylogenetic correction based on a maximum likelihood tree inferred using the concatenation of alignment of 16S and 23S rRNA genes. (Figure 3.3).

tests.

Since HIP1 motifs within the coding region will be translated into amino acid sequences, the observed bias among the three reading frames may reflect genome-specific codon usage, and amino acid preference in proteins. However, given the variation across the genomes, it seems unlikely that HIP1 is enriched due to selection on a single encoded peptide sequence. If a specific sequence of amino acid sequences was under selection across these HIP1-rich genomes, I would expect to see that HIP1 is universally much more enriched in one of the three reading frames than in the other two. However, in almost all genomes where RF2 is most enriched (Figure 3.4), the number of HIP1 motifs in RF0 and RF1 still exceeds the expected number by a substantial margin. Further, enrichment in RF2 is never more than 3 times higher than that in the second most enriched reading frame. As shown in Figure 3.4, in all genomes, enrichment varies between reading frames, but in no case is a single reading frame enriched to the exclusion of the other two.

The number of contexts in which HIP1 can be observed varies greatly between the three reading frames. The canonical motif in RF0 corresponds to four sequences of three codons each (GCG ATC GC*), all of which encode the hydrophobic triplet, AIA. In RF1, the canonical motif also corresponds to four sequences of four codons (*GC GAT CGC), each of which encodes a different peptide. In contrast, there are 255 ways[4] to embed the canonical motif in RF2 (**G CGA TCG C**). Further, Thirteen different amino acids are encoded by codons ending in G and five different amino acids are encoded by codons starting with C, resulting in 65 peptides of length four. The elevated degeneracy associated with the encoding

[4] One of the 16 nucleotide triplets ending in G, one is a STOP codon.

G_A	G_V	sub / site	HIP motifs	MAUVE blocks	K_S
sya	sel	0.0016	GCGATCGC	4	<0.01
cyf	cyg	0.0031	GCGATCGC	5	0.02
ana	ava	0.0277	GCGATCGC	263	0.18
syh	syg	0.1288	GGGATCCC	404	0.59

Table 3.4: Divergence of genome pairs selected for conservation analyses.

of HIP1 in RF2 may contribute to the observed preference for this reading frame.

3.3 HIP1 motif conservation

The observed enrichment shows that the high abundance of HIP1 is not solely due to the underlying genomic sequence composition, suggesting that high HIP1 abundance must be maintained by some genomic mechanism. We expect to see conservation of HIP1 sites in related genomes if HIP1 is maintained by selection, but not if HIP1 abundance is maintained by neutral processes. If HIP1 motifs are a byproduct of some neutral process, the motif, once inserted, will decay due to random stochastic mutation. In this case, HIP1 abundance would remain high because HIP1 motifs are constantly replenished, not because existing HIP1s are maintained. To rule out the possibility that a neutral process is driving HIP1 abundance, I assessed the conservation of HIP1 motifs in four pairs of genomes (*sel*-*sya*, *cyf*-*cyg*, *ana*-*ava*, and *syh*-*syg*), selected at varying evolutionary distances (Table 3.4). Evolutionary distance between a pair of genomes was assessed using K_S (synonymous substitutions per synonymous site), as described in Methods (3.8.4).

The four selected pairs of genomes, highlighted on the phylogeny in Figure 3.5, also represent a broad range of ecological strategies and taxonomic groups. Among them, *Synechococcus elongatus* PCC7942 (*sel*), a model organism used to study the circadian clock in cyanobacteria, is very closely related to *Synechococcus elongatus* PCC6301 (*sya*). Both *sya* and *sel* are freshwater obligate photoautotrophs. The two *Cyanothece* strains (*cyf* and *cyg*) were isolated from rice fields in Taiwan during springtime. Both genomes produce phycoerythrin, a pigment that absorbs light at frequencies that are not utilized by many other photosynthetic organisms. The *Nostocaceae* strains, *ana* and *ava*, are filamentous

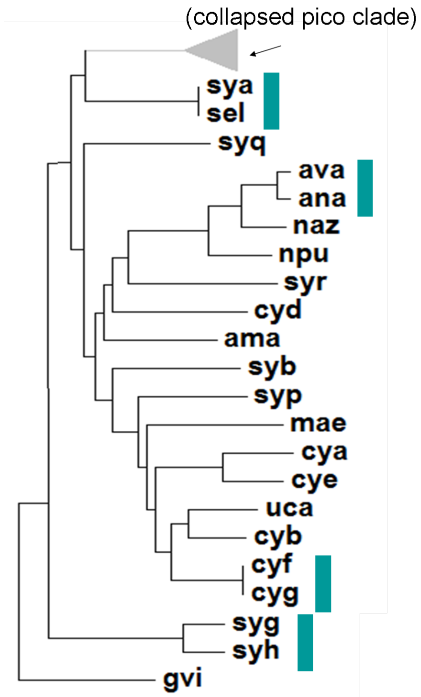


Figure 3.5: Phylogenetic relationships of genome pairs selected for conservation analyses.

heterocyst-forming diazotrophs with larger genome sizes. The fourth pair, *syh* and *syg*, are the two thermophilic strains with the alternative form of HIP1.

In order to quantify positional motif conservation, I first generated pairwise whole genome alignments using MAUVE ver. 2.3.1 (Darling et al., 2004, 2010) for the four selected pairs of genomes, as described in Methods (Section 3.8.5). Conservation was assessed over all aligned blocks in each pairwise alignment.

Motif conservation was quantified using two different conservation scores: the C score and the S score. S is a symmetrical measure of the conservation between a pair of genomes G_a and G_b. S is a specific application of the Jaccard score, which represents the fraction of elements in the union of two sets that also appear in their intersection. This symmetrical conservation score can be expressed as

$$S = \frac{n_{ab}}{n_a + n_b - n_{ab}}, \tag{3.8}$$

where n_a and n_b are the numbers of motif sites in G_a and G_b, respectively, that occur in the alignable blocks, and n_{ab} is the number of sites where the motifs in G_a and G_b are perfectly aligned[5].

The C score is an asymmetric measure of motif conservation between a pair of genomes, where one genome is treated as the reference genome (e.g., genome G_a) and the other as the comparison genome (e.g., genome G_b). C represents the fraction of sites in the reference genome that are also found in the comparison genome, and can be expressed as

$$C = \frac{n_{ab}}{n_a}. \tag{3.9}$$

The mean and variance of the number of conserved motif sites can be estimated under the assumption that conserved sites are binomially distributed. Given a motif in G_a, let p be the probability that it is perfectly aligned with the same motif in G_b, and let X be the random variable representing the number of sites that are conserved. Then, $X \sim B(n_a, p)$ with expectation

$$E(X) = n_a \cdot \hat{p}, \tag{3.10}$$

and variance

$$Var(X) = n_a \cdot \hat{p}(1 - \hat{p}), \tag{3.11}$$

where $\hat{p} = C$ is the maximum likelihood estimator of p. These statistics can be used to estimate bounds on the C score. Using the normal approximation, we obtain the 95% confidence interval of p:

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{3.12}$$

A p-value can be estimated directly from $B(n_a, \hat{p})$.

A more conservative approximation for p can be obtained by estimating the likelihood

[5] In all four genome pairs, the aligned blocks do not cover the entire length of either genome. Therefore, the motif counts in the conservation analysis (3.5 and 3.6) are lower than the total motifs counts given in Table 3.2.

interval. The log likelihood of p is

$$\mathcal{L}(p) = \ln\left(\binom{n_a}{n_{ab}} p^{n_{ab}}(1-p)^{n_a - n_{ab}}\right). \tag{3.13}$$

The likelihood interval for $\hat{p}$ is computed by searching for the lower bound ($\hat{p}_l$) and upper bound ($\hat{p}_u$) such that

$$\mathcal{L}(\hat{p}_l) = \mathcal{L}(\hat{p}_u) = \max \mathcal{L}(p) - 2 \tag{3.14}$$

and

$$\hat{p}_u > \hat{p}_l.$$

The upper bound $\hat{p}_u$ can be substituted for $\hat{p}$ in Equation 3.12 to obtain a more conservative estimate of the error bounds of C.

The conservation of HIP1 sites was assessed using both S and C measures. As a control, I also calculated the conservation of motifs that have properties similar to HIP1, but are not enriched. For example, an octamer palindrome with the same GC content as HIP1 can be used as a control motif. To have a large set of control motifs, I used the combined set of all octamer palindromes with 75% GC content as the control motif set, as described in Methods. When using this heterogeneous set of motifs, n_a refers to the number of sites in the aligned regions of G_a where any one of the palindromes in the control set appears; n_b is defined similarly. For control motifs, n_{ab} is the number of sites at which a control motif in G_a is perfectly aligned with the same palindrome in G_b. The conservation of HIP1, S_{HIP1}, can be assessed in comparison to the conservation of the control, S_{Ctrl}, to determine whether HIP1 motifs are more conserved than other palindromes with the same GC content. The asymmetric conservation scores are similarly compared.

Genome-wide motif conservation is reported in Table 3.5. Confidence intervals were obtained by estimating the upper bound, p_u, as described above, and substituting p_u into Equation 3.12. The p-values were estimated directly from the binomial distribution, $B(n_a, \hat{p_u})$. They indicate that the distinction between HIP1 conservation and control motif conservation is significant ($p < 10^{-14}$). For all four pairs of genomes, the HIP1 conservation lies outside

	G_a	G_b	K_S	n_a	n_b	n_{ab}	S	c	$\hat{p}_u^1$	95% C.I.^{2}	p-value^3
HIP1	sya	sel	0.00	7356	7399	7348	0.992	0.999			4.68×10^{-14}
	cyf	cyg	0.02	2947	2972	2825	0.913	0.959			1.30×10^{-25}
	ana	ava	0.18	4814	4837	4211	0.774	0.875			0
	syh	syg	0.59	3181	3767	2409	0.531	0.757			0
Control	sya	sel	0.00	2825	2808	2796	0.986	0.990	0.993	0.990-0.996	
	cyf	cyg	0.02	602	623	535	0.775	0.889	0.909	0.886-0.932	
	ana	ava	0.18	532	576	236	0.271	0.444	0.480	0.438-0.522	
	syh	syg	0.59	3710	3190	1015	0.172	0.274	0.286	0.271-0.301	

Table 3.5: Genome-wide positional conservation of HIP1 and control motifs. [1] $\hat{p}_u$, the estimate of C calculated based on the upper bound of the likelihoood interval. [2,3] 95% confidence intervals and p-values based on the binomial distribution estimated using the $\hat{p}_u$.

the confidence interval of the control motif conservation.

For the assignment of reference genome and comparison genome when calculating the C score, conservation was calculated twice, exchanging the reference and the comparison genomes, to ensure that the results are not biased by the choice of reference genome. The results, shown in Table A5 in the Appendix, are very similar to those in Table 3.5, suggesting that the direction of the analysis does not change the observations.

3.4 Conservation in coding and non-coding regions

HIP1 motifs in coding regions are more conserved than in intergenic regions. The same observation can be made for control motifs, as shown in Table 3.6. The greater conservation of HIP1 in coding regions could be an indication that HIP1 motifs are of particular importance in coding regions. It could also be due to the fact that coding regions are, in general, more conserved than intergenic regions. If selection were acting specifically to maintain HIP1 motifs in coding regions, we would expect the conservation in coding regions, relative to intergenic regions, to be greater for the HIP1 motif than for the control. Let $\mathcal{R}^{HIP1}$ be the ratio of the coding and intergenic C scores for HIP1 motifs and let $\mathcal{R}^{ctrl}$ be the same ratio for control motifs. Table 3.6 lists all such ratios.

Comparing $\mathcal{R}^{HIP1}$ with $\mathcal{R}^{ctrl}$ for each pair of genomes, the values of $\mathcal{R}^{HIP1}$ and $\mathcal{R}^{ctrl}$ are similar in all four cases. The ratio of $\mathcal{R}^{HIP1}$ to$\mathcal{R}^{ctrl}$ varies from 1.19 (*cyf*-*cyg*) to 0.75 (*syg*-*syh*). In other words, it is relatively close to one, suggesting that the degree of conservation in coding regions, relative to intergenic regions, is not dramatically different for HIP1 than for the control. Further, $\mathcal{R}^{HIP1}$ is neither consistently larger, nor consistently smaller, than $\mathcal{R}^{ctrl}$. However, we have no rigorous basis for testing the hypothesis that HIP1 is more conserved in coding regions than would be expected by chance. Comparing $\mathcal{R}^{ctrl}$ with $\mathcal{R}^{HIP1}$ does not address the variance of $\mathcal{R}^{ctrl}$, nor does it convey the statistical significance of the difference between $\mathcal{R}^{ctrl}$ with $\mathcal{R}^{HIP1}$. Several approaches could be used to develop a formal hypothesis test. A broader set of control motifs (for example, based on 6-mers) could be used to obtain a distribution of control conservation ratios. Alternatively, bootstrapping could be used to

	Genome			Coding regions					Intergenic regions					
	a	b	K_S	n_a	n_b	n_{ab}	C	S	n_a	n_b	n_{ab}	C	S	$\mathcal{R}$
HIP1	sya	sel	0.00	6341	6382	6337	0.99	1.00	1015	1017	1011	0.99	1.00	1.00
	cyf	cyg	0.02	2693	2694	2614	0.94	0.97	249	273	207	0.66	0.83	1.17
	ana	ava	0.18	4437	4280	3946	0.83	0.89	376	556	264	0.40	0.70	1.27
	syh	syg	0.59	2647	3039	2114	0.59	0.80	533	727	295	0.31	0.55	1.44
Control	sya	sel	0.00	2534	2522	2512	0.99	0.99	268	268	266	0.99	0.99	1.00
	cyf	cyg	0.02	507	499	449	0.81	0.89	84	114	76	0.62	0.90	0.98
	ana	ava	0.18	480	481	217	0.29	0.45	45	89	14	0.12	0.31	1.45
	syh	syg	0.59	3320	2787	957	0.19	0.29	357	394	54	0.08	0.15	1.91

Table 3.6: Positional conservation of HIP1 and control motifs in coding and intergenic regions.

estimate the variance, using either the current or an expanded control set. As more whole genome sequences become available, additional genome pairs at appropriate K_S distances promise increased statistical power for both of these approaches.

3.5 Codon conservation in HIP1 motifs

Each bacterial genome has its own codon usage preference. Differences in codon usage can affect the analyses described in Section 3.3, because the observed HIP1 conservation could be merely due to the fact that the nucleotide triplets within HIP1 are preferred codons.

I tested whether HIP1 conservation is caused by codon usage by analyzing codon conservation. If HIP1 conservation is a byproduct of codon usage, then the tri-nucleotides that appear in HIP1 motifs should have the same level of conservation within HIP1 motifs and outside of HIP1 motifs. Alternatively, if those tri-nucleotides are more conserved within HIP1 motifs, then we can conclude that codon usage is not the force driving the observed HIP1 conservation.

For each tri-nucleotide that can occur in HIP1, I calculated the conservation score (both C and S) for instances of the tri-nucleotide found in HIP1 motifs and instances found outside HIP1 motifs. There are six possible tri-nucleotides within the canonical HIP1 motif 5'-GCGATCGC-3': GCG, CGA, GAT, ATC, TCG, and CGC. In the Yellowstone strains, the tri-nucleotides within the variant HIP1 motif are: GGG, GGA, GAT, ATC, TCC, and CCC. This analysis was performed using only the genomic regions in the alignable blocks that are annotated as coding regions in both genomes. This comparison was also carried out separately for each reading frame.

Table 3.7 summarizes the conservation for each of the six tri-nucleotides. For the three more divergent pairs of genomes, tri-nucleotides are clearly more conserved in HIP1 motifs, than outside HIP1 motifs, as evidenced by both C score and S scores. The significance of the observed result was then determined using the Wilcoxon signed-ranked test, as described in Methods (Section 3.8), which tests the null hypothesis that the overall conservation of tri-nucleotides within HIP1 motifs is statistically indistinguishable from codon conservation

Genome			Within HIP1					Outside HIP1				
a	b	codon	n_a	n_b	n_{ab}	C	S	n_a	n_b	n_{ab}	C	S
sya	sel	GCG	7266	7277	7262	1.00	1.00	56422	56415	56308	1.00	1.00
		CGA	7257	7274	7255	1.00	1.00	50403	50361	50283	1.00	1.00
		GAT	7248	7271	7246	1.00	1.00	50677	50723	50616	1.00	1.00
		ATC	7252	7273	7252	1.00	1.00	50409	50459	50353	1.00	1.00
		TCG	7251	7272	7251	1.00	1.00	50182	50149	50072	1.00	1.00
		CGC	7256	7268	7252	1.00	1.00	56250	56230	56129	1.00	1.00
cyf	cyg	GCG	2825	2838	2730	0.97	0.93	29320	29355	27035	0.92	0.85
		CGA	2826	2840	2735	0.97	0.93	45298	45444	41664	0.92	0.85
		GAT	2833	2853	2757	0.97	0.94	85487	85435	78523	0.92	0.85
		ATC	2837	2856	2764	0.97	0.94	84454	84548	77713	0.92	0.85
		TCG	2817	2843	2732	0.97	0.93	45407	45424	41716	0.92	0.85
		CGC	2813	2841	2727	0.97	0.93	29600	29666	27337	0.92	0.86
ana	ava	GCG	4724	4760	4246	0.90	0.81	47807	47744	31834	0.67	0.50
		CGA	4730	4772	4266	0.90	0.81	43985	44153	28794	0.65	0.49
		GAT	4768	4787	4319	0.91	0.82	92111	92578	67014	0.73	0.57
		ATC	4769	4787	4318	0.91	0.82	92932	93118	67446	0.73	0.57
		TCG	4736	4761	4257	0.90	0.81	44427	44502	29171	0.66	0.49
		CGC	4727	4747	4234	0.90	0.81	48071	48234	32055	0.67	0.50
syh	syg	GGG	3465	3899	2968	0.86	0.68	72607	73466	43266	0.60	0.42
		GGA	3598	4008	3217	0.89	0.73	48127	50164	27410	0.57	0.39
		GAT	3633	4005	3250	0.89	0.74	42967	47750	27629	0.64	0.44
		ATC	3606	3993	3215	0.89	0.73	43142	47822	27719	0.64	0.44
		TCC	3564	3974	3155	0.89	0.72	48198	50813	27706	0.57	0.39
		CCC	3396	3854	2868	0.84	0.65	72338	73578	43458	0.60	0.42

Table 3.7: Codon conservation within and outside of HIP1 motifs.

outside of HIP1, based on the conservation of all six tri-nucleotides and all reading frames. The results are highly significant; $p = 0$ for *cyf-cyg*, *ana-ava*, and *syg-syh*. Taken together, the results indicate that codons are indeed significantly more conserved within HIP1 motifs. This analysis rules out the possibility that codon usage is the primary cause of HIP1 conservation.

3.6 HIP1 content in orthologous genes

In the previous two sections, I showed that HIP1 motifs are positionally conserved between pairs of genomes, and that this conservation is not a byproduct of the codon usage. Specifically, the conservation analysis showed that the positions of HIP1 motifs in aligned regions are conserved. This positional conservation led me to expect that HIP1 content in local regions is also correlated between orthologous regions. To see whether the local HIP1 content can reflect HIP1 positional conservation, relative to the control motif, I investigated the per-gene motif content correlation in orthologous gene pairs.

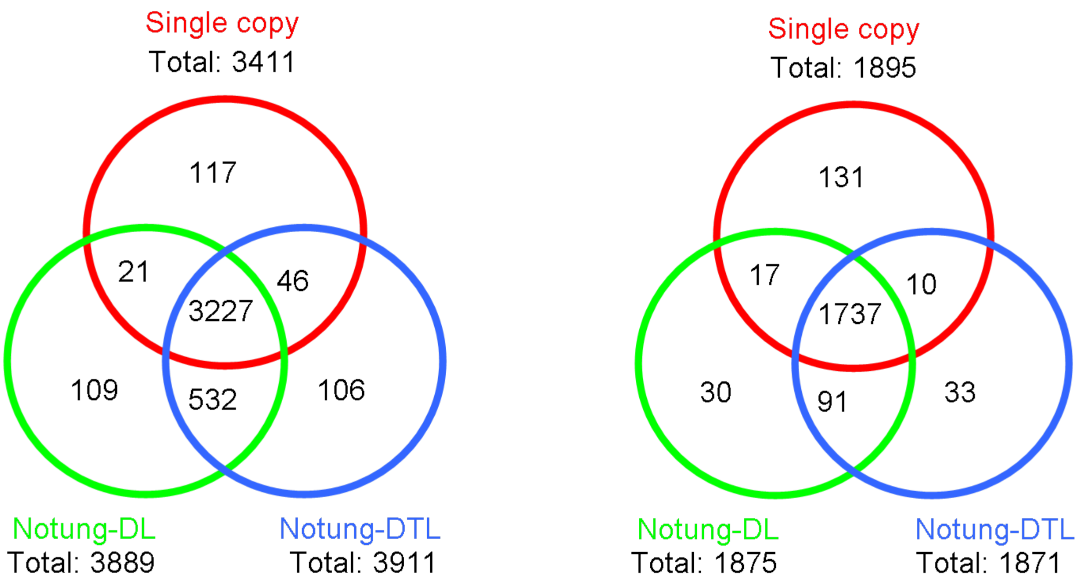
3.6.1 Datasets

HIP1 content analyses were carried out using two genome pairs. These are *ana-ava* and *syh-syg*, the most divergent pairs of the four pairs to test for HIP1 positional conservation. Instead of using MAUVE aligned blocks, I compared motif content in pairs of orthologous genes. The predicted gene families from the Barker dataset were used for ortholog prediction. The orthologs were predicted using three methods:

1. Gene pairs were taken from families with exactly one gene copy in each genome.

2. Orthologs were inferred using phylogenetic reconciliation to identify gene pairs that diverged from a common ancestor via speciation. These were obtained using the ortholog prediction function in Notung-2.8 with a DL-event model.

3. Orthologs were predicted using phylogenetic reconciliation with Notung-2.8 with a



(a) *Nostoc sp.* PCC 7120 (*ana*) and *Anabaena variabilis* ATCC29413 (*ava*)

(b) *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*)

Figure 3.6: The number of shared and unique predicted orthologs obtained using three different methods.

DTL-event model. The details of these three methods are explained in Methods (Section 3.8).

The Venn diagrams in Figure 3.6 summarize the number of shared and unique predicted ortholog pairs, obtained with these three methods. The set of orthologs that were predicted by all three methods were then used for the HIP1 content conservation analysis. These datasets contain 3227 ortholog pairs for *ana*-*ava*, and 1737 pairs for *syh*-*syg*.

3.6.2 Correlation of HIP1 content in orthologous genes

First, I investigated how well the motif content of a gene correlates with the motif content of the corresponding ortholog. Motif content was assessed by motif frequency and by motif enrichment. Scatter plots of HIP1 frequency and enrichment in ortholog pairs are shown in Figures 3.7 and 3.8. For comparison, I also plotted the control motif content in the same ortholog sets for both pairs of genomes. Correlation coefficients were calculated for each plot, and summarized in Table 3.8. The correlation coefficients indicate that motif frequency and enrichment are significantly more correlated for HIP1 than for control motifs ($p < 1 \times 10^{-15}$),

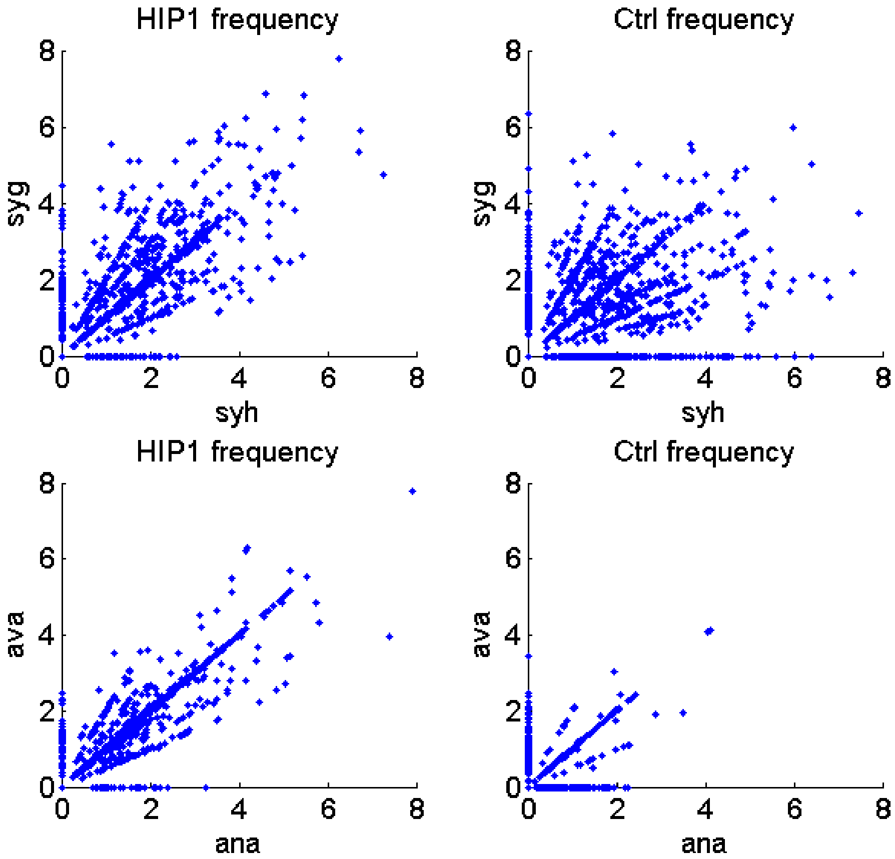


Figure 3.7: Scatter plots showing the motif frequency, in occurrences per kilo basepair, in orthologous gene pairs. Each dot represents an ortholog pair. Only frequencies below 8 motif per kbp are shown to reveal detail. Scatter plots with all the data points can be found in the Appendix (Figure A3).

for both frequency and enrichment. These observations are consistent with the result that HIP1 positions are conserved in aligned blocks, reported in previous sections.

3.6.3 Conservation of HIP1 per-gene content in orthologs

The analysis in Section 3.6.2 assesses motif content similarity across orthologs for the entire set of orthologous pairs. Here, I compare HIP1 content conservation with control motif content conservation, for each pair of orthologs individually. Again, the per-gene HIP1 con-

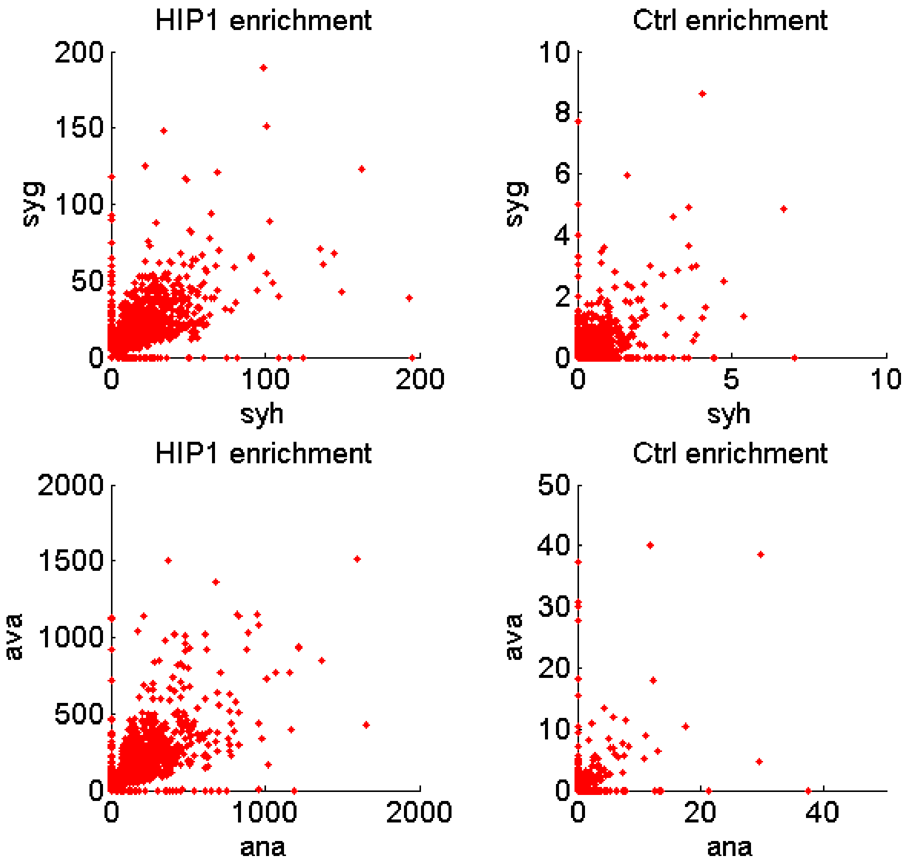


Figure 3.8: Scatter plots showing HIP1 enrichment in orthologous genes. These scatter plots are full views showing all the data points. Each dot represents an ortholog pair.

	Frequency		Enrichment	
Genomes	HIP1	Control	HIP1	Control
syh-syg	0.772	0.310	0.527	0.208
ana-ava	0.939	0.581	0.748	0.481

Table 3.8: Correlation coefficients of HIP1 content and control motif content between pairs of ortholgous genes. Values are based on the plots in Figure 3.7 and Figure 3.8. All correlation coefficients are supported by p-values smaller than 1×10^{-15}.

tent conservation is what would be expected given the positional conservation demonstrated previously.

Since HIP1 and control motif frequencies differ substantially, the following normalization procedures were used to obtain an unbiased comparison. Let g_A and g_B be a pair of orthologous genes from genomes G_A and G_B. To adjust for differences in motif abundance and enrichment across genomes, the values were first standardized and the resulting z-scores were compared. For a given raw score, x, let $\tilde{x}$ denote the standardized score

$$\tilde{x} = \frac{x - \bar{x}}{s}, \tag{3.15}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. For this analysis, the raw scores of interest are the motif frequency and enrichment in orthologous genes.

To test whether HIP1 motif content is more conserved than the control in individual ortholog pairs, I considered two test statistics, the absolute value and the square of the difference of the standardized motif content scores in g_A and g_B:

$$d_a(AB) = |\tilde{x}_A - \tilde{x}_B| \tag{3.16}$$

and

$$d_s(AB) = (\tilde{x}_A - \tilde{x}_B)^2. \tag{3.17}$$

The distributions of these test statistics both HIP1 and control motifs are plotted in Figures 3.9–3.12. The same plots but with full range are shown in Appendix Figures A4-A7. I used the Kolmogorov-Smirnov (KS) test to determine whether the distributions of differences in motif content between orthologous genes was significantly smaller for the HIP1 motif than for the control motif. The test was performed both for the set of all orthologous pairs and for the restricted set of orthologous pairs where the motif content was greater than zero in both genes. The resulting p-values are summarized in Table 3.9, which shows that the per-gene HIP1 content is significantly more conserved than the control motif content. This observation is consistent with the previous results on HIP1 positional conservation.

		Frequency		Enrichment	
Tests	Genome pair	all orthologs	non-zero[1]	all orthologs	non-zero[1]
$d_a[HIP]$ vs. $d_a[Ctr]$	syh - syg	3.36E-191	3.77E-96	3.35E-256	7.71E-215
$d_a[HIP]$ vs. $d_a[Ctrl]$	ana - ava	0	0	0	0
$d_s[HIP]$ vs. $d_s[Ctrl]$	syh - syg	5.10E-68	2.36E-32	2.03E-24	7.56E-09
$d_s[HIP]$ vs. $d_s[Ctrl]$	ana - ava	0	0	0	0

Table 3.9: Comparison of HIP1 and control motif conservation. The p-values are based on one-sided KS tests. [1] Tests results obtained using the partial dataset that contains only the orthologs that have at least 1 HIP1 motif in both genomes.

3.7 Chapter Conclusion

In this chapter, I showed that the HIP1 motif was abundant in most genomes tested with the exception of the marine pico-cyanobacteria, *Gloeobacter*, and two *Cyanothece* strains. The HIP1 motif content was also enriched, relative to the expected content, in all genomes with high HIP1 frequency, suggesting HIP1 abundance is not caused by the background genomic nucleotide composition. The two *Cyanothece* genomes with low abundance were also revealed to be enriched for HIP1 when background was taken into account. A prior study of HIP1 enrichment reported similar results (Delaye et al., 2011b). That study used an overly simple model of underlying sequence composition in estimating the expected number of motifs. My HIP1 enrichment results are based on a model of expected number of motifs in which tri-nucleotide frequencies in intergenic regions and in all three reading frames are taken into account.

In order to determine whether the HIP1 content of HIP1-rich genomes is maintained by selection or by a neutral process, I further investigated the extent to which HIP1 motifs are conserved. My results show that HIP1 positions were more conserved than the positions of control motifs. Analysis of codon conservation suggests that the observed conservation is not a byproduct of codon usage. Rather, my results suggest that the selection acting on HIP1 motifs is driving codon conservation within HIP1 motifs, and not the other way around.

These results, taken together, support the hypothesis that selection is acting to maintain HIP1 motif abundance. The alternative hypothesis, that HIP1 abundance is maintained by a neutral process that continually generates new HIP1 motifs, predicts high abundance and

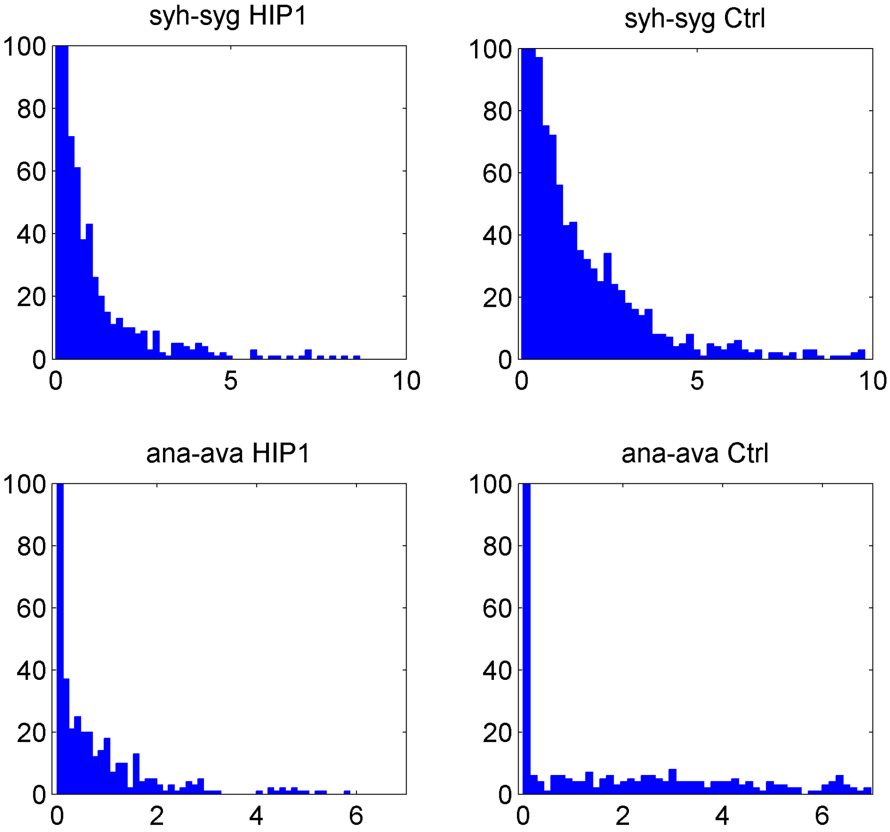


Figure 3.9: Histograms (detailed view) of d_s[HIP1] and d_s[Ctrl] based on motif frequency for the genome pairs *ana-ava* and *syh-syg*. The maximum values on the horizontal and vertical axes are capped. The complete histograms are shown in the Appendix.

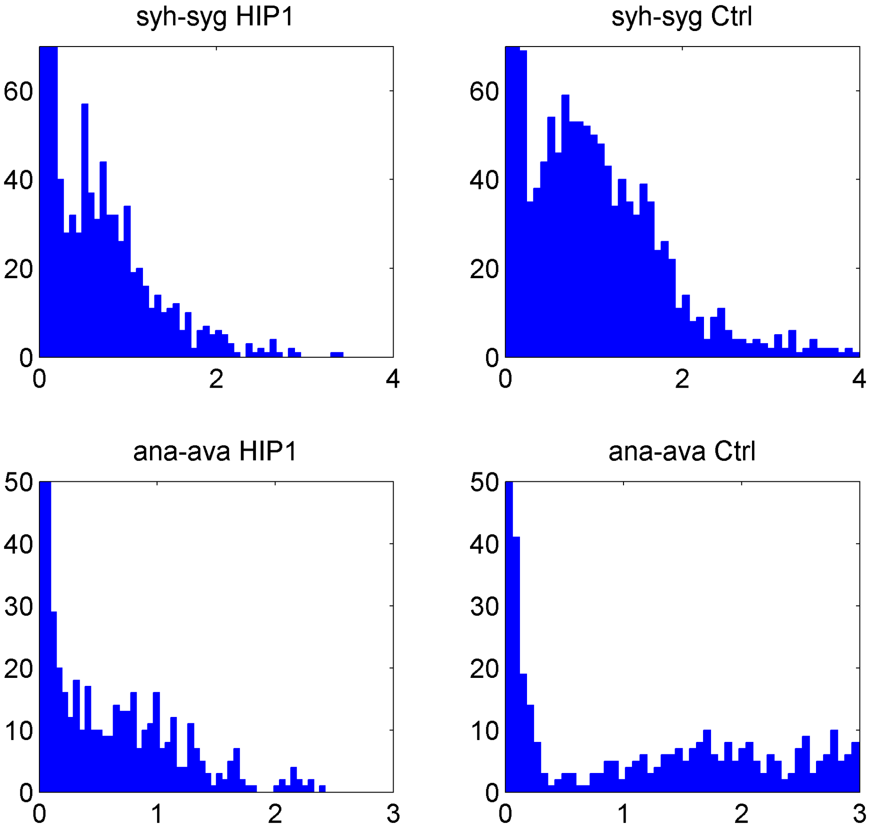


Figure 3.10: Histograms (detailed view) of d_a[HIP1] and d_a[Ctrl] based on motif frequency for the genome pairs *ana*-*ava* and *syh*-*syg*. The maximum values on the horizontal and vertical axes are capped. The complete histograms are shown in the Appendix.

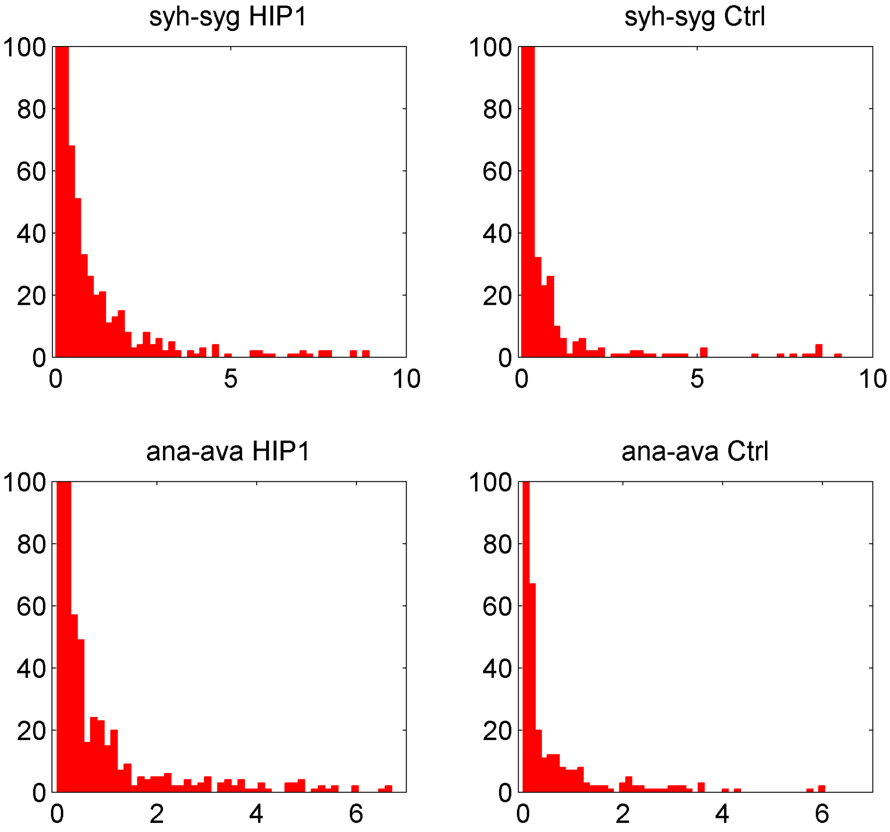


Figure 3.11: Histograms (detailed view) of d_s[HIP1] and d_s[Ctrl] based on motif enrichment for the genome pairs *ana-ava* and *syh-syg*. The maximum values on the horizontal and vertical axes are capped. The complete histograms are shown in the Appendix.

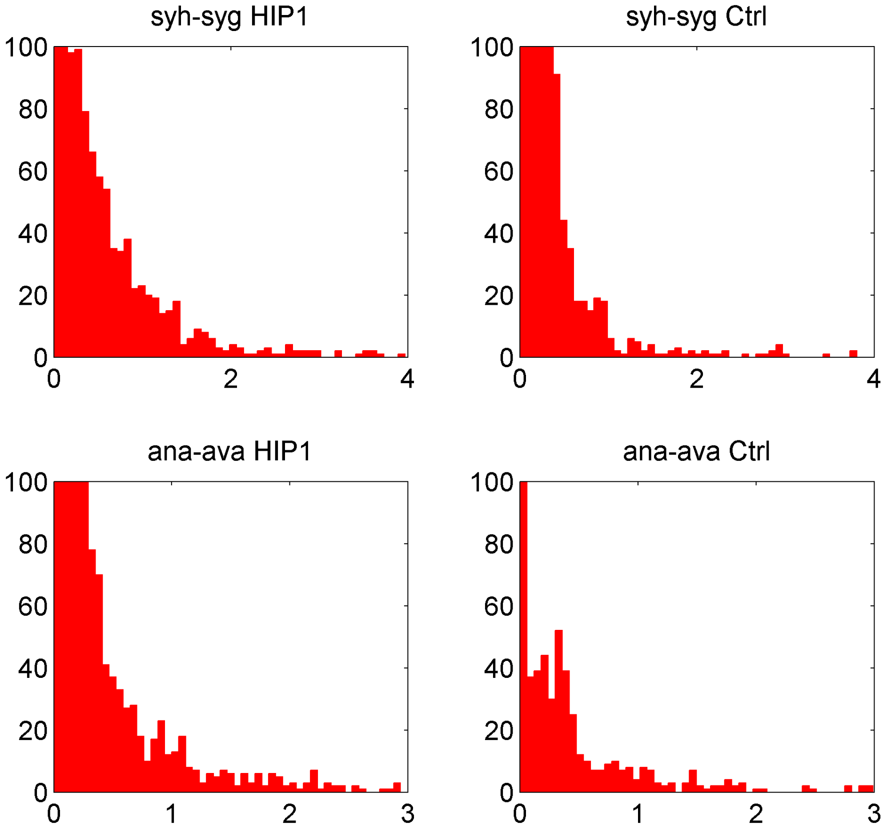


Figure 3.12: Histograms (detailed view) of d_a[HIP1] and d_a[Ctrl] based on motif enrichment for the genome pairs *ana*-*ava* and *syh*-*syg*. The maximum values on the horizontal and vertical axes are capped. The complete histograms are shown in the Appendix.

enrichment, but not elevated motif conservation.

I discovered a variant HIP1 motif (5'-GGGATCCC-3') in the Yellowstone strains, from which the canonical HIP1 motif is absent. I hypothesized that, in these strains, the variant motif might play the same role as the canonical HIP1 in HIP1-rich genomes. The variant HIP1 displayed characteristics similar to those of the canonical motif in all subsequent analyses: The variant motif was not only abundant, but also enriched in the Yellowstone strains. Like the canonical motif in other genomes, it has greater positional conservation than control motifs. Codon conservation analysis demonstrated that this conservation is not caused by codon usage. Further, in the content conservation analysis, the trends that were observed in the *ava*-*ava* pair (canonical HIP1) and the *syg*-*syh* pair (variant HIP1) were similar. These observations are consistent with the hypothesis that 5'-GGGATCCC-3' is a HIP1 variant. This suggests that the Yellowstone strains are HIP1-rich genomes, in contrast to the conclusions of Delaye et al. (2011b). The presence of a variant form could due to the thermophilic ecology of these strains. On the other hand, considering deep branching positions of *syh* and *syg* on the cyanobacterial phylogeny, it is also possible that the variant pattern reflects the ancestral form of this sequence motif.

My investigations on motif content in orthologs show that HIP1 content in orthologous pairs is more correlated than control motif content. Further, when HIP1 content conservation in an orthologous pair is compared to control content conservation in the same pair, the difference in HIP1 frequency tends to be smaller than the difference in control motif frequency. The same is true for enrichment. When considered over the entire set of orthologs, this trend is highly significant, according to KS tests.

The biological implications of HIP1 content conservation are difficult to interpret. The observed conservation of HIP1 content in orthologous pairs could be a byproduct of selection acting on HIP1 positions. Alternatively, selection could be acting to keep HIP1 motifs in a local region, suggesting a functional role that requires the accumulation of HIP1 instances. Such selection would also result in conserved motif sites. Since orthologs tend to inhabit regions with conserved synteny, this selective pressure might also act, indirectly, to promote conservation of HIP1 content in orthologous genes. If selection is acting to maintain HIP1 in local regions, then we would also expect to see elevated HIP1 content conservation when windows from aligned blocks were compared. A third possibility is that the target of selection is not specific regions in the genome, but the orthologs themselves.

Enrichment of HIP1 can be detected among all the genomes surveyed, except for *Gloeobacter* and the marine pico-cyanobacteria. Considering the placement of *Gloeobacter violaceus* PCC7421 (*gvi*), the HIP1 trait could have been acquired after the separation of *Gloeobacter* from the rest of the cyanobacterial lineages, and lost at the last common ancestor to marine pico-cyanobacteria. Alternatively, HIP1 abundance could be present at the last common cyanobacterial ancestor, and lost in *Gloeobacter*. As the deepest branching cyanobacterium, *Gloeobacter* is different from the rest of cyanobacteria in many ways. The thylakoid membrane, together with several key photosynthetic genes, are lacking in *Gloeobacter*. KaiABC, the key genes for cyanobacterial circadian clocks, are also missing in *Gloeobacter*. It is seemingly possible that HIP1 may perform a function that *Gloeobacter* does not require, and consequently was lost in this lineage.

Given the cyanobacterial phylogeny and the phylogenetic distribution of HIP1 enrichment (Figure 3.3), loss of the HIP1 trait in the marine pico-cyanobacteria is more likely than independent gains in the non-pico lineages. Marine pico-cyanobacteria could have lost HIP1 because of changes in underlying genome composition during genome streamlining. It was shown that picos have very low repetitive sequence coverage (Treangen et al., 2009). Therefore HIP1 might have been lost at the same time that other repetitive elements were lost. Alternatively, marine pico-cyanobacteria might have lost the phenotype that requires HIP1.

The phylogenetic distribution of HIP1 abundance and enrichment, as shown in Figure 3.3, provides interesting insights into the evolution of HIP1 enrichment. It seems that within the HIP1-rich genomes, the levels of enrichment are constantly changing during the course of genome evolution. There is no obvious, trend of increasing HIP1 enrichment change along the species tree. For example, even within the relatively closely related *Cyanothece* genomes, the enrichments range from 18 in *Cyanothece sp.* PCC 7822 (*cye*) to 189 in *Cyanothece sp.* ATCC51142 (*cyb*), which is more than a 10-fold change. On the other hand, the variation of HIP1 enrichment among the *Nostocaceae* is relatively small.

3.8 Methods

3.8.1 Control motifs

The combined set of all 63 octamer palindromes with 75% GC content, other than HIP1, was used as a control. The complete list of all possible octamers and their abundance can be found in Table A4 in the Appendix.

3.8.2 Genome dataset acquisition

Forty complete cyanobacterial genomes were retrieved from NCBI's FTP site [6] in Decemeber 2011. For each genome, the following files were acquired:

1. *.fna files containing the primary DNA sequences from all the replicons,
2. *.ptt files, protein annotation tables specifying the coordinates, reference accession id and locus tag for each protein coding gene in the genome,
3. *.rnt files, annotation tables for RNA genes.

3.8.3 Specification of coding and intergenic regions

In comparisons of coding and intergenic regions, coding regions were defined to be all genomic regions annotated as open reading frames in the NCBI protein annotation table. In other words, coding regions are ORFs. Non-coding regions, also referred to as inter-genic regions in this thesis, are the genomic regions that are not coding for proteins or RNAs, according to the annotation tables.

Intergenic regions were defined to be regions that are not included in any annotated ORF, RNA gene, and transposon. When calculating HIP1 motif enrichment and conservation within non-coding regions, long non-coding regions were excluded because they may contain unannotated genetic elements such as unknown protein coding or RNA genes, transposable

[6] URL: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/

elements, tandem repeats, pseudogenes, etc. The upper limit on the length of between-ORF regions was imposed to rule out unannotated elements such as transposons and long repetitive sequences.

I experimented with different criteria for inclusion as non-coding region for the conservation analysis and the results are summarized in Figure A2 in the Appendix. The criteria I experimented with were:

1. No limit (i.e., all non-coding regions out side of annotated features were treated as intergenic regions).

2. Between-ORF regions start 10bp downstream (3') of the stop codon and extend to the next annotated feature or to a position X bp downstream, which ever comes first. Values of X considered were 500 bp, 1000 bp, 2000 bp, and 3000 bp, respectively.

While specific numbers obtained with the different criteria varied, the overall trends were unaffected. For results reported in the main text of this thesis, a cutoff of 2000bp was used.

3.8.4 K_S calculation

K_S, the fraction of synonymous substitutions per synonymous site, was calculated from a concatenated alignment of 147 single copy genes obtained from a previous study (Latysheva et al., 2012). In that study, an amino acid sequence alignment for each family was previously generated, as described in Section 3.8.6. The protein sequence alignments were converted to DNA sequence alignments using the web-based tool PAL2NAL (Suyama et al., 2006), which generates codon-aware DNA alignments from pre-aligned amino acid sequences. K_S values were then calculated using the software KaKs Calculator (Zhang et al., 2006) using the criteria described in Nei and Gojobori (1986). K_S is computationally easier to calculate than the more commonly used d_S (Ota and Nei, 1994; Yang, 1998), and given the range of the genomes I am interested in, K_S is a good approximation of d_S (Li et al., 1985).

3.8.5 Whole genome alignment

Pairwise whole genome alignment was performed in MAUVE ver. 2.3.1 (Darling et al., 2004, 2010) using default parameter settings. Only the genomic regions in the alignable blocks were used for the conservation analyses.

3.8.6 Ortholog prediction

Orthologous gene pairs were prediction between *ana*-*ava*, and between *syh*-*syg*, based on the dataset of gene families and their corresponding phylogenies provided by Dr. Daniel Barker (School of Biology, University of St Andrews).

Gene family prediction and phylogeny reconstruction

The Barker dataset contains 65 species, among which 47 are cyanobacteria. The complete genomes were obtained from Integr8 database (release 108). Gene families were predicted using OrthoMCL 2.0 (Li et al., 2003) with MCL 09-308, on BLAST results with 'm S' masking and an E-value cutoff of 10^{-5}. The inflation parameter 1.6 was chosen.

The Barker dataset contains a species phylogeny of the 65 species. The species phylogeny was based 147 universal, single-copy gene families. Multiple alignment of each gene family was performed in MAFFT in 'E-INS-I' mode with 1000 iterations (Katoh and Toh 2008). LG+Γ was in MODELGENERATOR (Keane et al., 2006) using the using the Bayesian Information criterion (BIC). The phylogeny was constructed using PhyML (Guindon et al., 2010) with 'best' rearrangements.

Phylogenies of 13,852 gene families from 65 species were based on multiple sequence alignments performed in MAFFT. A model was selected for each family using MODELGENERATOR with 4 rate categories and BIC. 13,852 gene trees were constructed in PhyML via 'best' rearrangements. Each gene tree was bootstrapped with 200 replicates.

Orthologs were predicted from these gene families using three different methods. The first method is based on sequence comparison. The other two methods are based on phylogenetic reconciliation, a process that compares a gene tree with a species tree to infer the evolution process that gave rise to each bifurcation in the gene tree. Notung 2.7 (Stolzer et al., 2012) was used to reconcile the gene and species trees, to root the gene trees, and to rearrange weakly supported branches.

Single copy presence in both genomes

Specifically, if genes g_A and g_B, from genome G_A and G_B respectively, are both in gene family F, and no other members of F are found in G_A and G_B, then g_A and g_B are predicted to be orthologs.

Notung prediction with DL event model

Given a species tree and a gene tree, Notung can predict orthologs once the DL-reconciliation is done, by looking at the events history. If the last common ancestor of g_A and g_B, from genomes G_A and G_B respectively, is predicted to be a speciation event, then g_A and g_B are predicted to be orthologs. The gene trees were rooted using duplication-loss (DL) parsimony. To remove noise, the gene trees were rearranged using an edge support threshold of 70%. The reconciliation was performed after a DL-rearrangement and a DL-rerooting process. In all steps of this pipeline, the duplication and loss costs used were $\delta = 3$ and $\lambda = 2$, respectively, as described in Section 2.2 in Chapter 2.

Notung prediction with DTL event model

If the the last common ancestor of g_A and g_B, from genomes G_A and G_B respectively, on the gene tree is predicted to be a speciation event, and there is no transfer predicted along the path between g_A and g_B, then g_A and g_B are predicted to be orthologs (Fitch, 2000). The DTL reconciliation was performed after a DL-rearrangement and a DL-rerooting process. In all steps of this pipeline, the event costs used for DL-model are $\delta = 3$ and $\lambda = 2$. The event costs used for DTL-model are $\delta = 3$, $\tau = 2.5$, and $\lambda = 2$. For gene families with more than one optimal reconciliation, only the orthologs that are supported by all optimal reconciliations are reported.

3.8.7 Correction for phylogenetic dependency

To address the phylogenetic dependency when counting the number of genomes in which HIP1 is most enriched in a particular reading frame, weights are assigned to each genome, reflecting the presence or absence of closely related genomes. Given a species tree with branch lengths, the weight assigned to a species node (A), either an internal or leaf node, is defined to be

$$W_A = \frac{a + S_A}{a + S_A + b + S_B} W_{P(A,B)}, \tag{3.18}$$

where B is the sister group to A, and $P(A, B)$ is the parent of A and B. Here, a and b are the branch lengths from A and B to $P(A, B)$, respectively. S_A is the sum of all branch lengths below A. When A is a leaf node, S_A is zero. The value of W_{toot} is a scaling factor and can be chosen arbitrarily. To avoid underflow due to multiplication of numbers less than one, W_{Root} is frequently set to large, positive integer; e.g., 1000. . Weights for leaf nodes are normalized so that they will sum to 1. Thus, the normalized weight of leaf node X will be:

$$W'_X = \frac{W_X}{\Sigma_{i \in L} W_i}, \tag{3.19}$$

where L is the set of all leaf nodes.

Chapter 4

Intra-Genome Variation of HIP1 Motifs

Previous results from Chapter 3 indicate HIP1 is conserved. The HIP1 positional conservation, as well as the HIP1 content conservation, implies selection, indicating HIP1 might have important functional roles. Such function might be acting on the whole-genome level. If this is the case, the spatial distribution of HIP1 may provide useful insights as many repetitive motifs related to chromosome structure are suggested to have specific spatial patterns, such as periodicity. In addition, the spatial distribution of HIP1 within genetic units can also provide valuable clues about the functional aspects of HIP1, as motifs with regulatory roles tend to be located within or near the promoter regions and terminator regions of the genetic units.

In this chapter, I explore the spatial distribution of the HIP1 motifs on a genome scale, as well as within genetic units. I further look into the spatial distribution with respect to potential functional implications.

4.1 Genome-wide HIP1 spatial variation

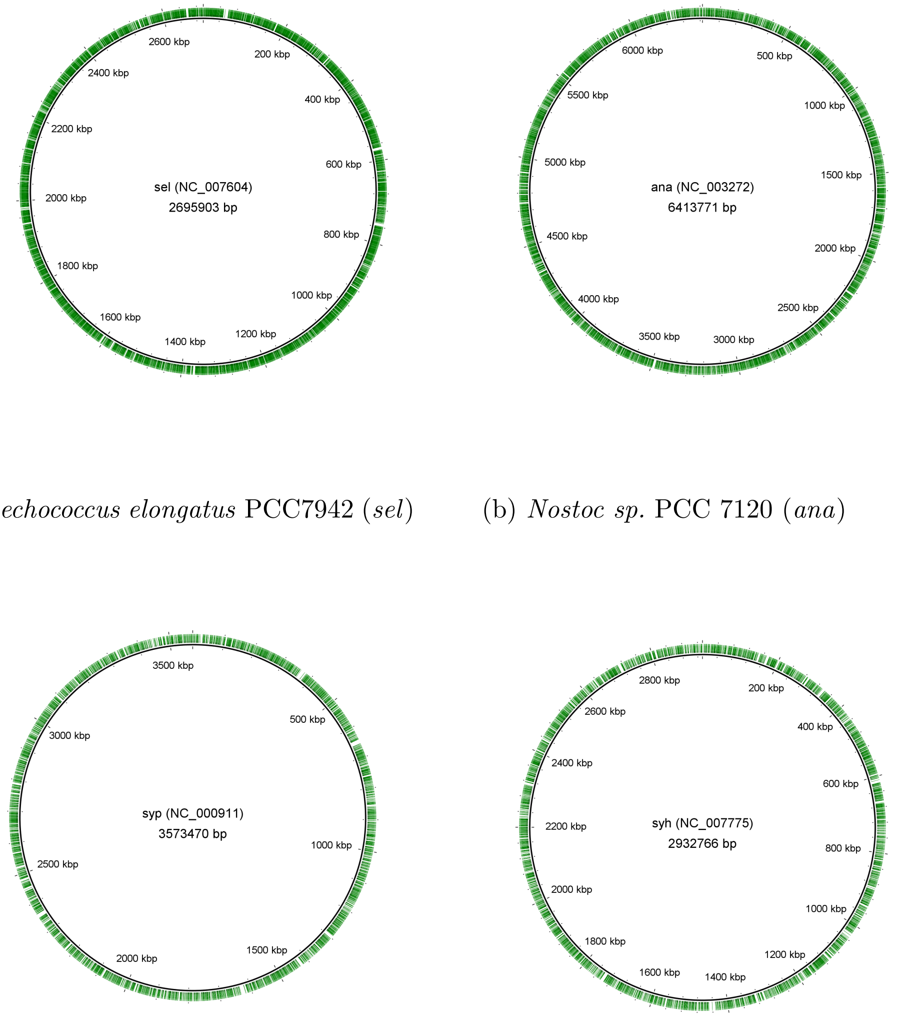
I first consider the spatial organization of HIP1 motifs on a genome-scale. I am interested in whether there exist genome regions with a burst of HIP1 motifs or regions lacking HIP1 motifs. If there exist such regions, how can I detect them and systematically characterize the intra-genome variation of motif abundance?

4.1.1 Genome-wide HIP1 distribution

Four genomes were chosen for visual inspection of genome-wide patterns. The four selected genomes are *Synechococcus elongatus* PCC7942 (*sel*), *Nostoc sp.* PCC 7120 (*ana*), *Synechocystis sp.* PCC6803 (*syp*), and the Yellowstone strain *Synechococcus sp.* JA-3-3Ab (*syh*). Of these, *Synechococcus elongatus* PCC7942 (*sel*), *Nostoc sp.* PCC 7120 (*ana*) and *Synechococcus sp.* JA-3-3Ab (*syh*) are the genomes used in conservation analyses in Chapter 3, and *Synechococcus elongatus* PCC7942 (*sel*), *Nostoc sp.* PCC 7120 (*ana*), *Synechococcus sp.* PCC7002 (*syb*) are the model organisms in cyanobacterial lineage. These genomes exhibit a range of HIP1 frequencies (Table 1.1).

The spatial distribution of HIP1 motifs is presented visually in Figure 4.1. All four strains display variations in HIP1 density along the genome, with gaps corresponding to HIP1-free regions.

Figure 4.2 shows examples of HIP1-free and HIP1-rich regions in the genome of *Synechococcus elongatus* PCC7942 (*sel*), the genome with the highest genome-wide HIP1 frequency. Figure 4.2(a), (b), and (c) show HIP1-free regions containing a protein coding gene, a transposase, and rRNA genes. Visual inspection of HIP1-depauparate regions in HIP1-rich genomes suggests that transposase and rRNA genes are frequently associated with such regions. The transposons are often of foreign origin; these may have originated in genomes lacking HIP1. HIP1 motifs are rarely found in RNA genes, suggesting that secondary structure constraints conflict with the HIP1 motif. Figure 4.2(d) shows a dense cluster of HIP1 motifs in a coding region and, more generally, a pattern of varying HIP1 density in a dozen genes. Again, visual inspection reveals that this variation is typical.



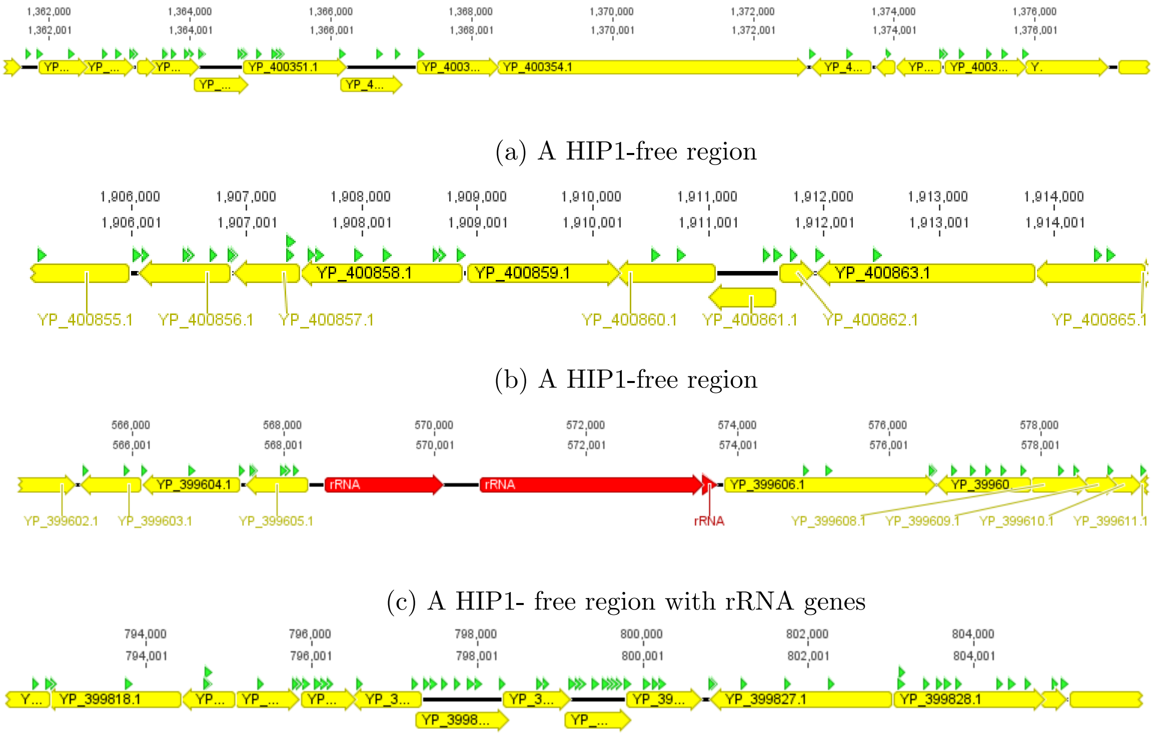
(a) *Synechococcus elongatus* PCC7942 (*sel*)

(b) *Nostoc sp.* PCC 7120 (*ana*)

(c) *Synechocystis sp.* PCC6803 (*syp*)

(d) *Synechococcus sp.* JA-3-3Ab (*syh*)

Figure 4.1: Spatial variation in HIP1 motif content in four HIP1-rich cyanobacterial genomes. Figures produced by the BLAST Ring Image Generator (Alikhan et al., 2011). (a) *Synechococcus elongatus* PCC7942 (*sel*), (b) *Nostoc sp.* PCC 7120 (*ana*), (c) *Synechocystis sp.* PCC6803 (*syp*), (d) *Synechococcus sp.* JA-3-3Ab (*syh*).



(a) A HIP1-free region

(b) A HIP1-free region

(c) A HIP1- free region with rRNA genes

(d) A HIP1-rich region

Figure 4.2: Selected examples of HIP1 intra-genome variation in *Synechococcus sp.* PCC 7942. *(a)* A HIP1-free region containing the integrin alpha chain. *(b)* A HIP1-Hole containing one single gene which is annotated as transposase IS605. *(c)* a HIP1-free rRNA region containing 16S and 23S rRNA genes. Most commonly, rRNA genes are HIP1-free. *(d)* A case of a HIP1-rich region, containing a 16S rRNA methyltransferase (RsmE), and a 4-hydroxythreonine-4-phosphate dehydrogenase. Images generated with the genome browser interface of the Geneious software package (Kearse et al., 2012).

4.2 HIP1 spatial organization within genetic units

The visual inspection of genome-wide HIP1 distribution is intriguing. A natural extension would be quantitative assessment of the motif clustering and clumpiness effect, as well as special motif distribution properties such as periodicity. It is not uncommon to see oligonucleotides that have periodic patterns of occurrence within bacterial genomes (Collings et al., 2010; Mrazek, 2010; Mrazek et al., 2002). However, such analyses are beyond the scope of this thesis.

Next, I considered HIP1's spatial distribution at the sub-genomic level, focusing on the HIP1 distribution within transcripts, operons, and genes. The distribution of the HIP1 motifs within those genetic units may be of functional importance. Usually the 5' end of these genetic units are considered to be the promoter regions with regulatory elements. For example, a potential preferred distribution of HIP1 motif may indicate HIP1 is related to transcriptional regulation. There is evidence that repetitive sequences influence transcription in a variety of ways, including modulating supercoiling, promoting transcription termination, and inhibiting 3' oriented RNA degradation (reviewed in Treangen et al. (2009)). Repetitive sequences may contribute to the formation of RNA secondary structures that influence transcription. HIP1 could act as a DNA binding site for proteins that regulate transcription. Given the palindromic nature and the high abundance of HIP1 in some genomes, it is quite possible that pair motifs interact with HIP1s to form inter-motif secondary structures in a single mRNA molecule. A non-uniform distribution of HIP1 motifs would be an indirect indicator of this type of functional role.

In prokaryotic genomes, two or more neighboring genes in the same orientation may be co-transcribed into a single mRNA and then translated into separate proteins. The concept of an operon was proposed to describe a sequence of contiguous genes that are co-transcribed.

An operon and a transcript may contain the same set of genes, however, the two have very different definitions. A transcript is an rRNA molecule, and the level of a certain type of transcript can be measured in a biological system. An operon is an abstract structural concept. It describes an organization of genes. Operons are predicted and can be verified in biological systems. Transcripts can often have overlaps, as they may have alternative start and stop sites. The genes are typically (but not always) separated by non-coding intergenic regions and the entire sequence of genes may be flanked by 5' and 3' UnTranslated Regions (UTRs).

My analyses in this subsection focus on HIP1 spatial distribution within genes, transcripts, and (predicted) operons. The spatial distribution of HIP1 motif in those genetic units can give useful insight of HIP1 function.

4.2.1 Definitions and Datasets

Here I describe the three genetic units under consideration; genes, transcripts, and operons, and how they are defined in the context of the datasets.

Genes: In this thesis, I use 'gene' mainly to refer to predicted protein coding genes. I also use the term open reading frame (ORF) to refer to the protein coding genes, in the genome annotation files. Note that ORF may have a different meaning in other contexts, such as continuous sequence of DNA triplets that starts with Met codon and ends with STOP codon.

The coordinates for annotated genes were obtained from the protein annotation tables (*.ptt files) downloaded together with the genome sequences from NCBI's ftp site. The actual protein coding sequences were then extracted from the genomic sequence in the *.fna files using in-house scripts.

Transcripts: A transcriptome dataset from is available for *Synechococcus sp.* PCC7942 (Vijayan et al., 2011). This dataset consists of a transcriptome map, generated by combining three high-resolution datasets obtained using RNA sequencing, tiling expression microarrays, and RNA polymerase chromatin immuno-precipitation sequencing (RNA pol ChIP-Seq). The transcriptome map provides the start and end positions in the genome sequence for each transcript. This dataset also provides the expression level of each transcript, measured in RNA read abundance. There are a total of 1415 transcripts in the dataset, some of which overlap. When two neighboring transcripts overlap, I remove the shorter one from the dataset. A set of 1375 non-overlapping transcripts is thus obtained. In the following

distance cutoff (d)	P	TP	precision	sensitivity
100	1455	887	0.610	0.627
125	1345	825	0.613	0.583
150	1265	770	0.609	0.544
175	1210	729	0.602	0.515
200	1169	702	0.601	0.496

Table 4.1: Assessment of operon prediction under various inter ORF distance cutoff d.

analyses, unless otherwise stated I use the term 'transcript dataset' to refer to the set of 1375 non-overlapping transcripts.

Operons: The high-resolution transcriptome dataset from Vijayan et al. (2011) is for one genome, *Synechococcus sp.* PCC 7942. The transcriptomes for other genomes are lacking and experimentally expensive to generate. In order to study transcriptional units in other genomes, I predicted operons in all 20 cyanobacterial genomes in this study.

There are various studies on operon prediction methods for bacterial genomes. Here, I use a relatively simple one: Two neighboring ORFs in the same orientation are placed in the same operon if the inter-ORF distance between them does not exceed a user specified cutoff distance, d. In other words, a predicted operon is a run of consecutive of ORFs with the same orientation that are not too far apart.

To assess the accuracy of this operon prediction method for various cutoff distances, I compared the operons predicted in the *Synechococcus elongatus* PCC7942 (*sel*) genome with the 1415 transcripts in the transcriptome dataset. In this comparison, I considered a prediction to be a true positive (TP) if the start and end positions of the operon coincide exactly with the boundaries of the transcript once the 5' and 3' UTRs have been removed. Predicted operons that are not perfect matches for a transcript are considered false positives (FP), even if they substantially overlap with the transcript. Transcripts that do not perfectly align to a predicted operon are considered to be false negatives (FN). I use

$$precision = \frac{TP}{TP + FP} \tag{4.1}$$

and

$$sensitivity = \frac{TP}{TP + FN} \tag{4.2}$$

to assess the operon prediction performance.

The results of this comparison, summarized in Table 4.1, suggest that a cutoff of $d = 125$ gives the best precision. Note that $d = 100$ is almost as precise as d=125, but has better sensitivity. Because of the nature of analyses, I weighed precision over sensitivity, so that the majority of the predicted operons dataset resembled the experimentally detected transcripts. Based on these results, operons were predicted in all 20 HIP1-rich genomes using a cutoff of $d = 125$. Also note that the value of d is estimated from a single genome (*Synechococcus elongatus* PCC7942 (*sel*)), and then applied to all 20 HIP1 rich genomes. It is possible that bias was introduced because each genome may have its own optimal value of d.

The operon prediction method described above does not take untranslated regions into account. However, most transcripts in the Vijayan dataset have UTR regions; only 175 transcripts lacked a 5' UTR and all but 266 transcripts had a non-zero 3' UTR. The distribution of UTR lengths is summarized in Table 4.2. Statistics were calculated for all UTRs and for the set of UTRs of non-zero length.

To better mimic the behavior of transcripts, I also predicted operons with UTRs by adding an extra 60 bp to the 5' end of the predicted operon and 100 bp to the 3' end of the predicted operon. These values correspond to the mean of the lengths of all non-zero 5'- and 3'-UTRs in the *Synechococcus sp.* PCC 7942 transcript dataset, respectively (Table 4.2).

4.2.2 HIP1 spatial distribution in genes, transcripts, and operons

I used two approaches to quantify the HIP1 spatial distribution within those genetic units: the Fractional Spatial Distribution (FSD) and Binned Statistics (BS).

All	5'-UTR	3'-UTR
mean	50.04	82.99
median	29	50
s.d.	66.96	143.61
Non-zero	5'-UTR	3'-UTR
mean	56.79	101.28
median	33	64
s.d.	68.59	152.702

Table 4.2: A survey of the lengths (in bp) of UTRs in the *Synechococcus sp.* PCC 7942 transcriptome dataset.

Fractional Spatial Distribution (FSD)

FSD is a measure of the motif abundance at a given position in a genetic unit, normalized by its length. Given a motif in a genetic unit of length L, let m be the position, in base pairs, of the first nucleotide in the motif. Then, $f = m/L$ is the position of the motif expressed as a fraction of the length of the genetic unit. A value of f that is closer to 0 indicates a motif located closer to the 5' terminus; if f is closer to 1, then the motif is closer to the 3' end. Given a set of genetic units and a motif of interest, the FSD is the histogram of the fractional positions of all instances of the motif in the data set, normalized by the total number of observed motifs. The benefit of this second normalization is that it is easier to visually compare FSDs.

If the motif exhibits no spatial preference, one would expect the motifs to be uniformly distributed within the transcripts. The observed FSD can be tested against a uniform distribution using a one-sample Kolmogorov-Smirnov (KS) test. The resulting p-value indicates the significance of the deviation of the the FSD from this null hypothesis.

I used the FSD to quantify the spatial distribution of HIP1 motifs in the set of 1375 non-overlapping transcripts (Figure 4.3). For comparison, I plotted the FSD for control motifs in the same set of transcripts. Visual inspection of the HIP1 FSD suggests a preference for the 3' end of transcripts. KS tests show that the HIP1 FSD deviate significantly from a uniform distribution ($p = 2.5 \times 10^{-21}$), in contrast to the control motifs, where the deviation from uniform is much less significant ($p = 2.1 \times 10^{-3}$). A two-sample KS test was applied to

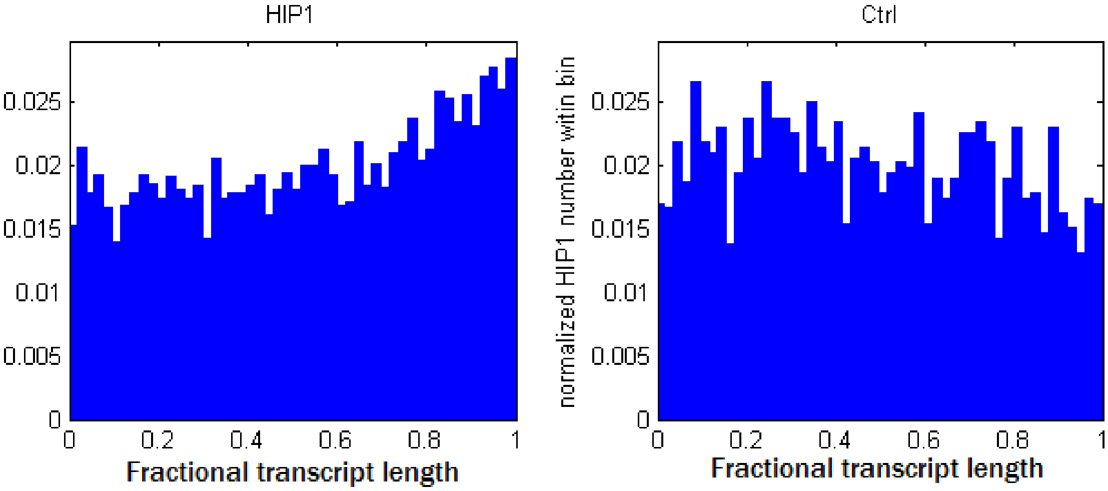


Figure 4.3: Fractional Spatial Distributions of HIP1 motifs (left) and control motifs (right) in 1375 non-overlapping transcripts.

compare the HIP1 with control motif FSDs, suggesting that the spatial distribution of HIP1 motifs is significantly different from the control motif ($p = 3.24 \times 10^{-14}$).

To assess the contribution of the 3' UTR to the observed preference for HIP1 motifs at the 3' terminus, I re-plotted the HIP1 and control FSDs after removing the 5'- and 3'-UTR regions from the transcripts. Again, I used a KS test to determine the statistical significance of the HIP1 FSD relative to the uniform and the control distributions. The resulting p-values are noticeably less significant than those obtained using transcripts with UTR regions (see Table 4.3). This observation suggests that the UTRs make a substantial contribution to the observed 3' preference.

For comparison, the spatial distribution of the HIP1 motif in predicted operons was also studied. FSDs based on the set of 1345 predicted operons (without added UTRs) is shown in Figure 4.4. For comparison, I plotted the FSD for control motifs in the same set of predicted operons. I compared both FSDs to the uniform distribution and to each other, using KS tests. The resulting p-values are shown in Table 4.3. Visually, the trend of 3' HIP1 preference is much weaker, compared to the FSD for the transcipt dataset. The

	p values		
Dataset	HIP1 vs unif	Ctrl vs. unif	HIP1 vs Ctrl
Transcripts	2.46×10^{-21}	2.12×10^{-03}	3.24×10^{-14}
Transcripts (noUTR)	5.67×10^{-10}	5.73×10^{-03}	1.27×10^{-08}
Operon (with UTR)	1.05×10^{-07}	6.57×10^{-05}	5.29×10^{-10}
Operon (noUTR)	1.92×10^{-02}	5.47×10^{-04}	2.83×10^{-05}
Genes	7.54×10^{-02}	1.81×10^{-03}	1.07×10^{-02}

Table 4.3: Comparison of the Fractional Spatial Distributions of the HIP1 and Control motifs with the uniform distribution (one sample Kolmogorov Smirnov test) and with each other (two sample Kolmogorov Smirnov test). All tests are for genome Synechococcus sp. PCC7942 (sel)

deviation from the uniform distributions is barely significant (1.92×10^{-2}). I repeated the above analysis using operons with added predicted UTRs (figures not shown). When UTRs are included in the operon prediction, the deviation of the HIP1 FSD from the uniform distribution is much more significant (1.05×10^{-7} versus 1.92×10^{-2}), suggesting that UTRs substantially contribute to the observed 3' HIP1 preference.

I further repeated the FSD analyses and KS tests for genes. The resulting p-values, shown in Table 4.3, are barely significant. These observations do not support a 3' HIP1 preference in individual genes.

In summary, I observed a strong and significant 3' preference for HIP1 distribution in transcripts. Such spatial preference is significant as revealed by comparing the HIP1 FSD with the uniform distribution and control motif FSD. I also compared the control motif FSD with the uniform distribution using one-sample KS test. The resulting p-values (Table 4.3) suggest that there is only a mild difference between the control motif FSD and the uniform distribution.

The 3' preference of HIP1 was also observed in predicted operons, but to a much weaker extent. The HIP1 distribution revealed no significant 3' spatial preference of in annotated genes. My results suggest that the 3' preference of HIP1 distribution is at least partially contributed by the UTR sequences. The observed 3' preference of HIP1 motif distribution in transcripts is intriguing. However, all the above analyses are based a single genome *Syne-*

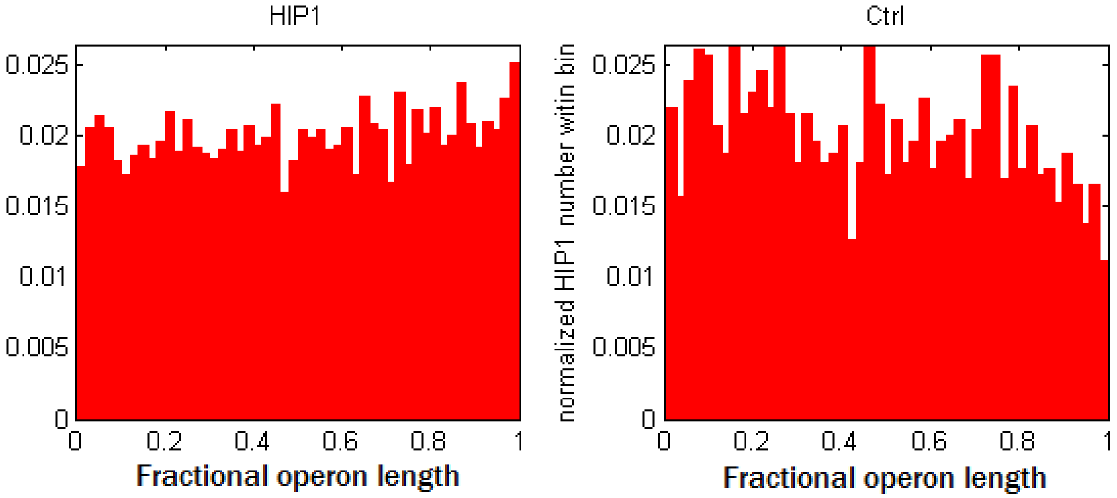


Figure 4.4: Fractional Spatial Distributions of HIP1 motifs (left) and Control motifs (right) in 1345 predicted operons.

chococcus elongatus PCC7942 (*sel*). I then further investigated whether such 3' HIP1 spatial preference was present in other HIP1-rich genomes. The spatial motif distribution in other genomes can only be assessed using predicted operons. Since the UTRs contribute substantially to the magnitude of the observed 3' preference, the assessment of 3' preference in other genomes will likely be influenced by the quality of the UTR predictions. I plotted the HIP1 FSDs for predicted operons with predicted UTRs in Figure 4.5. Visually, among the 20 genomes, only *Synechococcus elongatus* PCC6301 (*sya*), a genome closely to *Synechococcus elongatus* PCC7942 (*sel*), showed a 3' preference similar to that observed in *Synechococcus elongatus* PCC7942 (*sel*). Further, the p-values from KS tests against a uniform distribution suggest that *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*) are the only two genomes that exhibit a significant 3' preference in predicted operons. Given these observations, it is difficult to determine whether the 3' preference of HIP1 is a specific property of *Synechococcus elongatus* PCC6301 (*sya*) and *Synechococcus elongatus* PCC7942 (*sel*), or whether the failure to observe a 3' bias in other genomes is due to the prediction accuracy of operons, and UTRs.

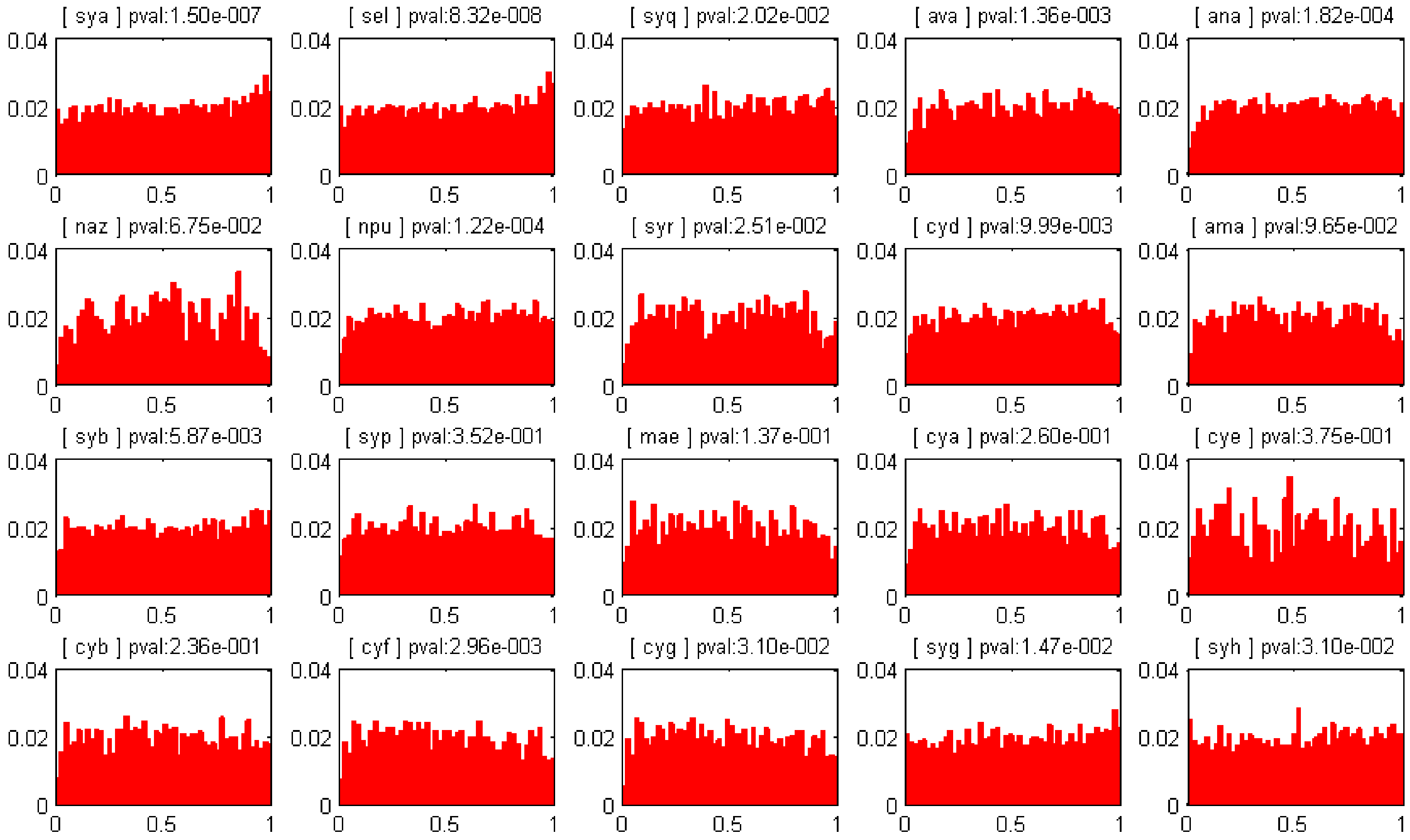


Figure 4.5: Fractional Spatial Distributions of HIP1 motifs in predicted operons in 20 HIP1 rich-genomes. Horizontal axis: Fractional operon length. The p-values from the KS-test against the uniform distribution are displayed above the subfigures.

Binned Statistics

The FSD gives an assessment of the spatial distribution of motif abundance, but cannot be used to assess whether the spatial distribution of motif enrichment deviates from a uniform distribution. In order to assess the spatial distribution of motif enrichment, I used a binned statistic, in which the genetic unit is divided into b bins and the ratio of observed to expected number of motifs is calculated for each bin separately.

For a genetic unit of length L, each bin corresponds to a subsequence of length

$$l = \frac{L}{b}. \tag{4.3}$$

The i^{th} bin is the subsequence of the genetic unit, starting at position $l \cdot (i - 1) + 1$ and ending at position $l \cdot i$. Let n_i^j be the number of motifs observed in the subsequence in bin i in genetic unit j. Given a set of genetic units, the abundance for bin i is simply

$$O_i = \sum_j n_i^j. \tag{4.4}$$

The expected abundance in bin i is calculated by estimating the di- and tri-nucleotide frequencies in the concatenation of the sequences associated with bin i in each of the genetic units. E_i, the expected number of motifs in bin i, can then be estimated from these frequencies using the second order Markov model introduced in Chapter 3 (Equation 3.3). The motif enrichment in bin i is

$$\mathcal{E}_i = \frac{O_i}{E_i}. \tag{4.5}$$

Figure 4.6 shows binned abundance and binned enrichment for the HIP1 and control motifs in the transcript data. Visual inspection of the binned HIP1 abundance again reveals a strong 3' preference. This preference is not observed in the plot of binned control motif abundance. Interestingly, the binned HIP1 enrichment plot shows a preference for both termini: the enrichment in the first and last bin is noticeably elevated, relative to the middle bins. Since the binned abundance in the 5' terminal bin is not elevated, the elevated 5'

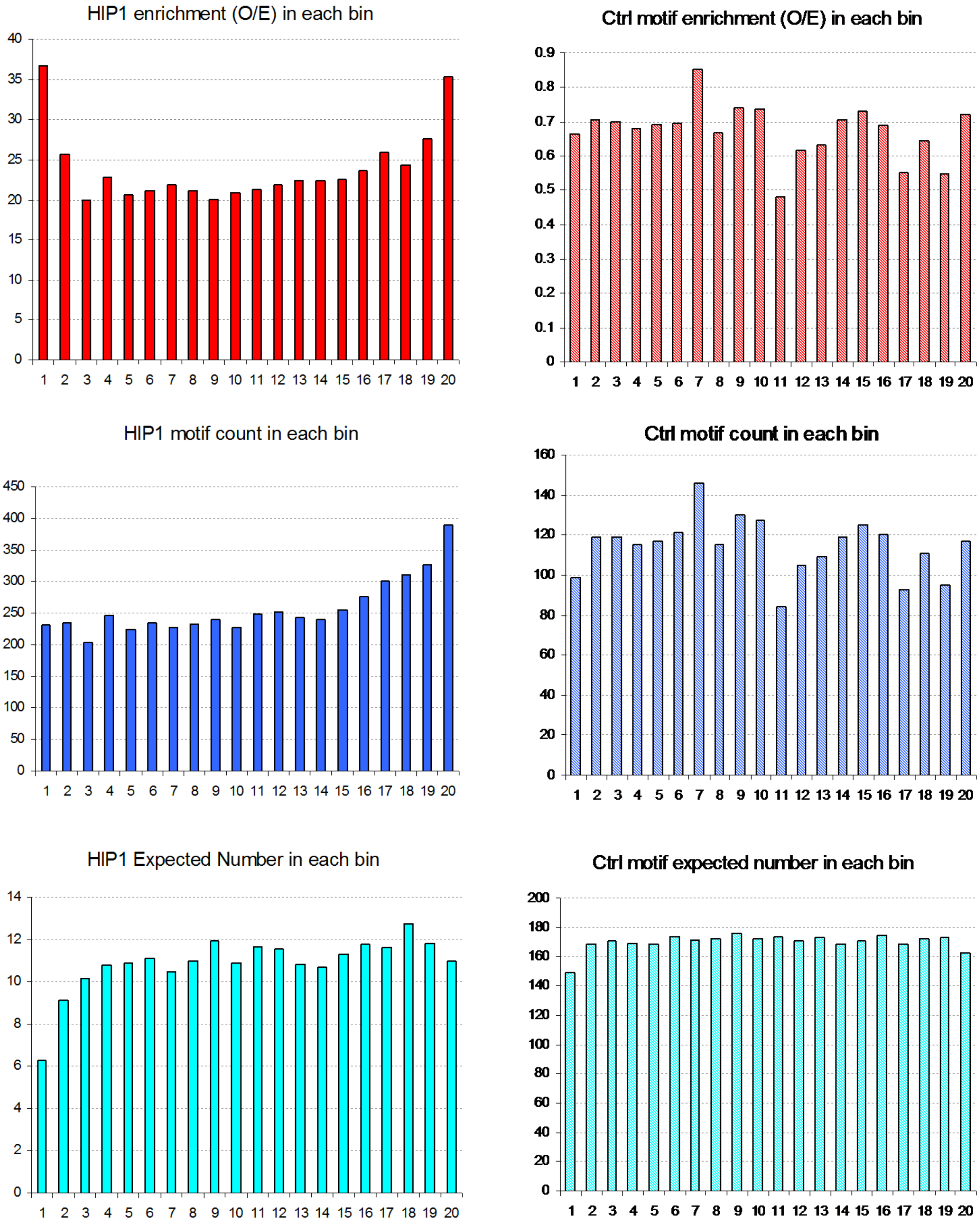


Figure 4.6: Binned statistics: abundance (blue), enrichment (red), and expected number of motifs (green) for HIP1 (left) and the control motif (right) in *Synechococcus elongatus* PCC7942 (*sel*) transcript data.

enrichment is presumably due to reduction in the expected number of motifs in the first bin. This is consistent with the observation that intergenic regions tend to have lower G+C content than coding regions and that promoters, in particular, tend to be relatively AT rich.

Randomized permutation tests can be used to assess the significance of the spatial distribution of a binned statistic. To assess the significance of a 3' bias, the test statistic is defined to be the difference between the average value in the first $b - k$ bins and the average in the last k bins. The distribution of this test statistic under the null hypothesis is simulated by repeatedly permuting the order of the bins. For this analysis, we used a value of $k = 3$.

Unfortunately, there is a limit to the statistical power available for this test. Since the order within the first $b - k$ bins or within the last k bins does not matter, there are only $\binom{b}{k}$ possible permutations. For example, when $k = 3$, the total number of permutations is 19,600. Thus, unless the observed test statistic is more extreme than all permutations, the most significant p value that can be obtained is $5e^{-5}$. The number of permutations, and hence the statistical power, could be increased by increasing k, but this would dilute the signal because the mean motif abundance would decrease. Alternatively, one could increase the number of possible permutations by increasing the number of bins, but then the estimate of the expected number of motifs in any one bin would become less accurate.

Figure 4.7 shows the binned statistics for 1345 predicted operons in *Synechococcus elongatus* PCC7942 (*sel*), with and without UTRs. The significance in each case was estimated by permutation testing. In order to discount 5' effects, the test statistic used is the difference between the average values of the top 3 bins and the middle 44 bins. Once again, 3' bias of HIP1 abundance and enrichment is observed. However, the p-values are only mildly significant for HIP1 enrichment and abundance when UTRs are included, and for HIP1 abundance when UTRs are not included. The p-values for other conditions are not significant (Figure 4.7). The main conclusions to be drawn from this figure are that (1) the distributions of both the binned abundance and the binned enrichment are more significant for the HIP1 motif than for the control motif and (2) that the distributions of BS for HIP1 motifs in operons predicted with UTRs are more significant than in operons lacking UTRs.

The results from BS in other genomes are similar (data not shown). Among the 20 HIP1-rich genomes, only *Synechococcus elongatus* PCC7942 (*sel*), *Synechococcus elonga-*

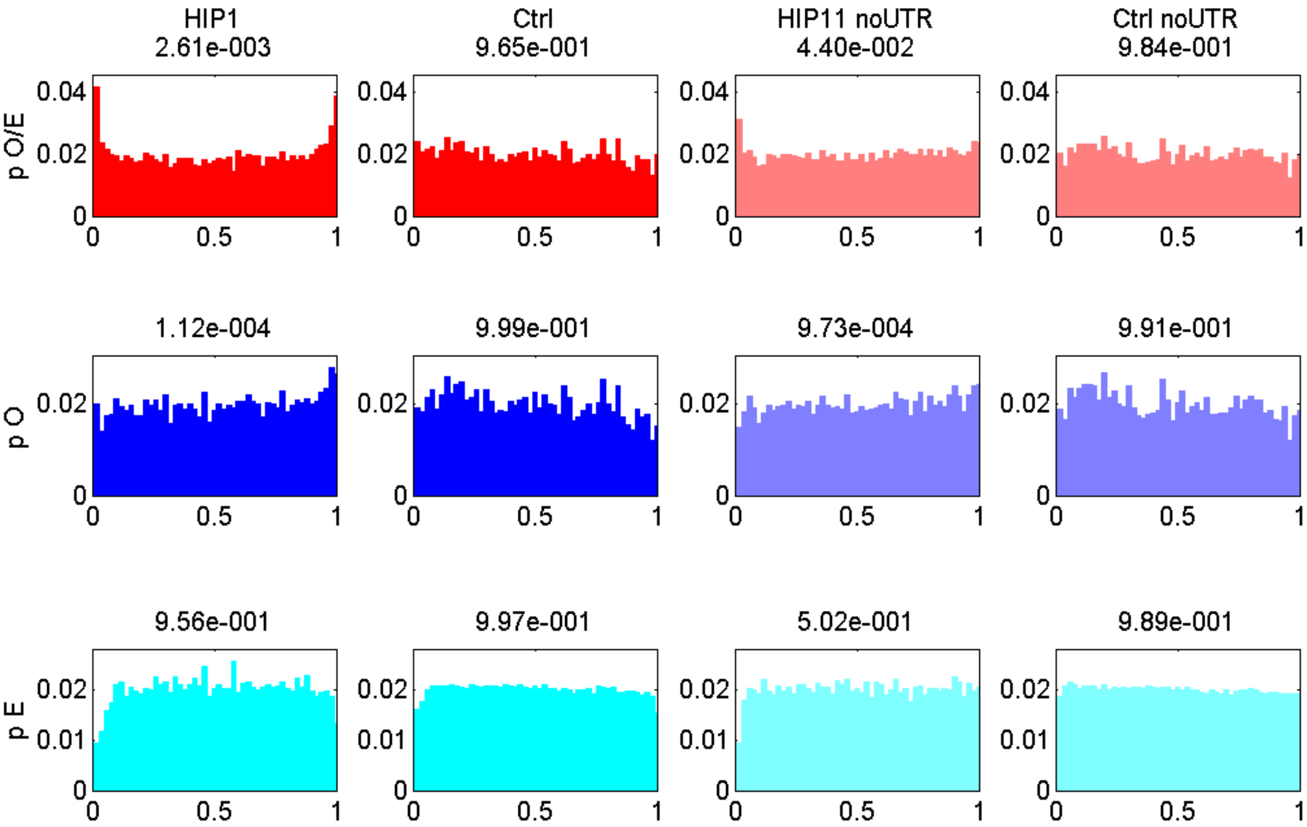


Figure 4.7: Distribution of binned statistics for the observed (blue) and expected (aqua) number of motifs and for motif enrichment (red). p-values represent the significance of the observed distribution, relative to a null distribution generated with permutation testing.

tus PCC6301 (*sya*), *Anabaena variabilis* ATCC29413 (*ava*), *Nostoc sp.* PCC 7120 (*ana*), *Cyanothece sp.* ATCC51142 (*cyb*), and *Synechococcus sp.* PCC7002 (*syb*) have mildly significant p-values for the 3' bias of HIP1 enrichment, with p values 2.61e-3, 2.64e-3, 1.95e-5, 5.29e-5, 2.95e-4 and 1.35e-3, respectively, when the predicted UTRs are included. All other genomes have p-values greater than 0.01. When the predicted UTRs are not included, in all 20 genome, HIP1 enrichment shows no significant 3' bias of HIP1 enrichment ($p > 0.01$). These observations again suggest that UTRs contribute to the 3' bias of HIP1 enrichment.

4.2.3 Signal isolation (via Boys and Girls)

The observation of a 3' bias in the spatial distribution of HIP1 motifs is intriguing because some of the functions that are known to be associated with other types of repeats, act at the

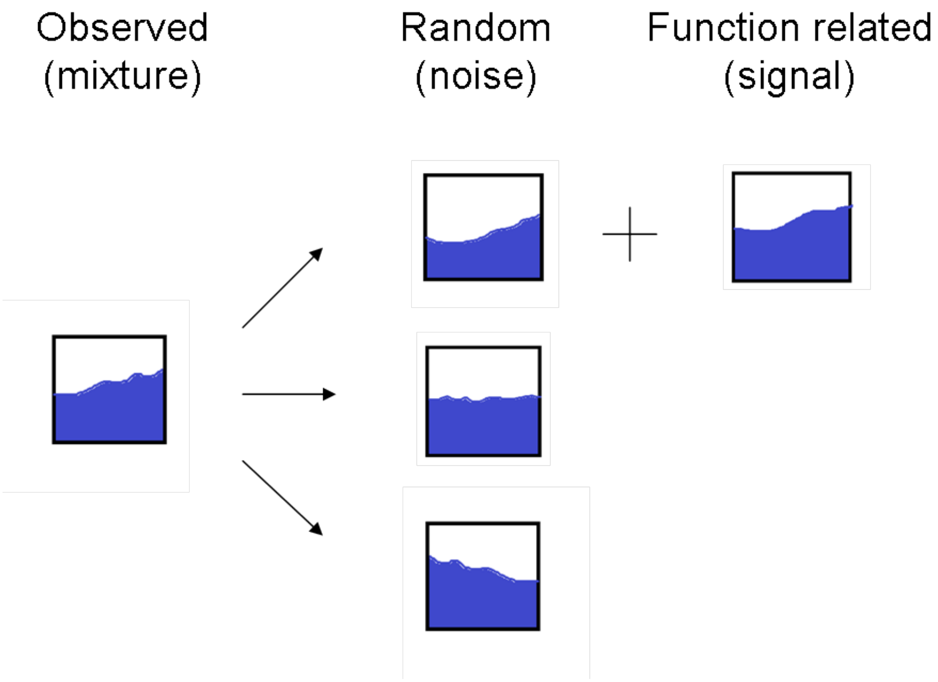


Figure 4.8: An abstract representation of a hypothesized mixture of spatial distributions exhibit a 3' bias when combined.

3' ends or transcripts as transcription termination or inhibition of exonucleolytic degradation.

However, the observed the 3' bias was based on aggregate measures of spatial preference, whereas functions like transcription termination are acting in individual transcripts. How do individual transcripts contribute to this observed aggregate behavior? And if the the 3' bias is, indeed, related to HIP1 function, then in which transcripts is this function operative?

I hypothesized that the set of transcripts represents a mixture of spatial motif distributions, in which some transcripts have a 5' motif bias, some have 3' bias, some have no strong bias, and yet others are HIP1-free. Further, there is a preponderance of transcripts with a 3' bias in this mixture, resulting in the observed 3' preference in the FSD and BS plots. In addition, I hypothesized that HIP1 has a functional role associated with its 3' position in some transcripts and that this accounts for the preponderance of 3' biased transcripts (assuming this preponderance exists). Figure 4.8 shows a schematic representation of these hypotheses.

In order to determine whether the aggregate 3' preference arises because there are more transcripts with a 3' bias than would be expected under a uniform distribution, it is necessary to extract those transcripts that have a preponderance of HIP1 motifs at the 3' end. I used two approaches for this purpose. First, since the 3' bias was observed in the combined set of transcripts, it is possible that the positional bias is too weak to be seen in individual transcripts. With this in mind, I used an iterative heuristic optimization procedure to search for batches of transcripts with a 3' bias. There are two variants of this method: Boys and Girls (BG) and normalized Boys and Girls (nBG). The second approach seeks individual transcripts with a 3' bias, using Ranking by Mean Position (RMP).

Boys and Girls (BG) is an iterative heuristic optimization process that, given a set of N transcripts, seeks the subset of N transcripts to that contribute most to the 3' bias. The goal of the heuristic is to separate the set of transcripts into batches of n transcripts, such that the first batch has the strongest 3' bias and the last batch has the weakest 3' bias. This approach requires an optimization criterion to assess the strength of the 3' bias in a set of transcripts. I quantified the 3' bias of a batch of transcripts by calculating the slope of the FSD, which is calculated by fitting the FSD array to a straight line via least-square fitting in MatLab. Pseudocode describes the heuristic procedure is given in **Algorithm: Boys-and-Girls**.

Ideally, the inner loop of this heuristic will output a batch of n transcripts in which the 3' bias is maximized. However, since this is a hill-climbing procedure that does not consider all possible bi-partitions, the output of the inner loop may not be optimal. As a result, the final list of batches may not be in strictly decreasing order. The accuracy of this heuristic depends on how well the optimization criterion (in this case, the FSD slope) captures 3' bias and how thoroughly the heuristic searches the space of bipartitions. This, in turn, depends on the the maximum number of failed attempts, κ, in the termination condition.

Normalized Boys and Girls (nBG) One potential problem of the BG process is that the FSD slope used in the optimization criterion has a length bias: longer transcripts tend to contribute higher slopes because they contain more HIP1 motifs. To address this problem, I used a variant of the heuristic, in which the optimization criterion is the FSD slope, normalized by the total number of motifs in the FSD.

Algorithm: Boys-and-Girls

```
Input:
    $T$ = a set of $N$ transcripts.

Initialization:
    $Sorted = list()$

Batches:
    While ( $T$ is not empty ) {
        $B$ = a set of $n$ transcripts, selected at random.
        $R = T - B$
        $S = slope(B)$
        Repeat {
            $t_b$ = select a transcript from B uniformly at random.
            $t_r$ = select a transcript from R uniformly at random.
            Move $t_b$ from $B$ to $R$ and $t_r$ from $R$ to $B$          # Swap a pair of transcripts
            $S' = slope(B)$
            if ($S' > S$)
                { continue }                                    # Accept swap
            else
                { move $t_r$ from $B$ to $R$ and $t_b$ in $R$ to $B$ }   # Reverse swap
        } until ($\kappa$ attempts without a successful swap)
        $T = R$
        $Sorted = list(Sorted, B)$                              # Add $B$ to sorted list of batches.
    } # End while

Output:
        $Sorted$: A list of batches of transcripts sorted by slope.
```

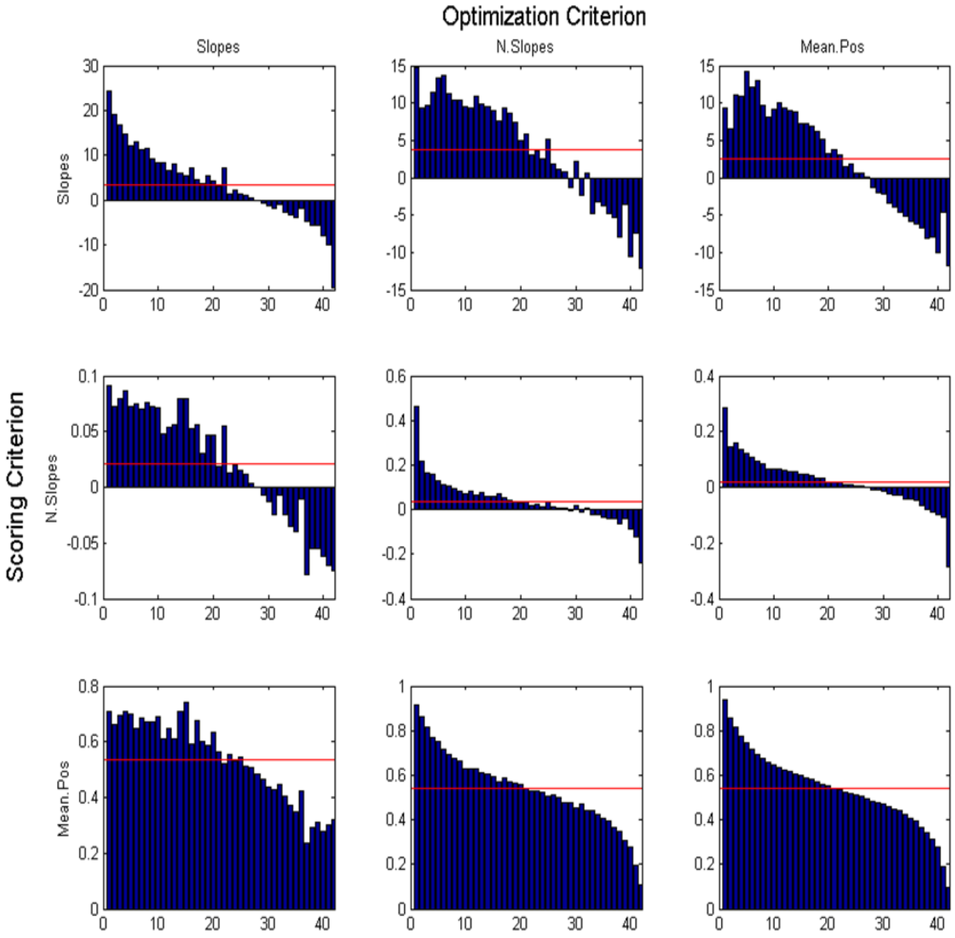


Figure 4.9: Comparison between the three criterion for separation signal from noise.

Ranking by Mean Position (RMP) This approach does not subdivide the transcripts into batches. In this case, the transcripts are ranked according to mean normalized motif position. The normalized position of a single motif is the position of the first nucleotide in the motif, divided by the length of the transcript.

Figure 4.10 shows an example of the output of the BG heuristic, with a batch size of $n = 100$. The first 13 plots show the binned abundance for each batch, where the batch resulting from the first iteration of the outer loop appears in the upper left hand corner. The red barplot in the lower right hand corner shows the slopes of the 13 batches. The barplot shows that there are 8 batches with positive slopes, three with negative sloes and

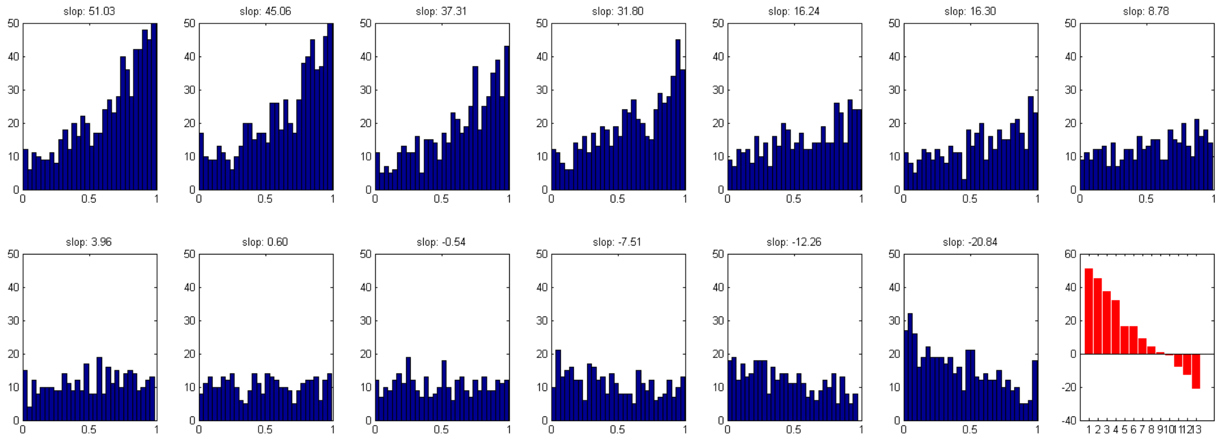


Figure 4.10: A partition of 1375 transcripts into 13 batches with decreasing slopes, obtained with BG process. The last barplot in the bottom row shows the slopes in the resulting batches.

two with slopes close to zero. This is consistent with the hypothesis that the aggregate 3' bias observed in the FSD plots is due to an excess of individual transcripts with a 3' bias.

Figure 4.9 shows a comparison of the three methods. All HIP1-free transcripts were removed from the data set prior to the analysis. The remaining 1056 transcripts were partitioned into 42 batches using the BG and nBG heuristics, with a batch size of $n = 25$ and a termination criterion of $\kappa = 1000$. In addition, each transcript was scored according to its mean normalized motif position. The transcripts were then sorted according to their scores and combined into 42 batches according.

Figure 4.9 shows the resulting three partitions of the same set of transcripts into 42 batches, where the three partitions were obtained with the BG, nBG and RMP methods, respectively. Each column in Figure 4.9 corresponds to one data set, scored using the three different scoring schemes, slope, normalized slope, and RMP. Thus, the plots on the main diagonal show batches obtained with BG, nBG and RMP, respectively, and scored with slope, normalized slope, and RMP. The slopes in the upper left and middle plots are not strictly decreasing. This is because the heuristic does not perfectly partition the data.

The off-diagonal plots represent batches that were generated with one scoring method and then re-scored with a different method. This reveals the extent to which the scoring methods have similar properties. For example, the left and middle plots in the bottom row were generated using BG and nBG, respectively. The resulting batches were then scored by calculating the mean RMP for the 25 transcripts in each batch. Note that the shape of the middle plot is very similar to that of the right-most plot, in which the transcripts were partitioned and scored with the same measure (RMP). This suggests that normalized slope and RMP behave similarly. In contrast, the plot on the left, which was generated with the BG method, lacks monotonicity and looks quite different from the other two plots, suggesting that the slope does not have behavior similar to normalized slope and RMP. The plots in the middle row exhibit a similar trend.

The top 50 transcripts with highest mean HIP1 position are reported in Table A6 and Table A7 in the Appendix. These transcripts contribute most to the observed 3' bias of HIP1 spatial distribution.

4.3 Functional implication of HIP1 spatial variation

One possible hypothesis for the observed 3' bias in sel transcripts is that HIP1 plays a role in transcriptional or translational regulation. To further explore the functional implications of HIP1 spatial variation in transcripts, I investigated the extent to which 3' bias correlates with properties that are directly or indirectly related to expression.

Those properties include the motif frequency within each transcript, transcript expression level, circadian behavior of transcript expression, codon usage bias (GCB, ACE), transcript lengths (in base pairs or in the number of ORFs), and transcript GC content. I also considered motif frequency within each transcript to check if there is a relationship between motif abundance and 3' bias.

Transcript expression level and codon usage bias reflect two different aspects of expression. Transcript expression level is a direct measure of RNA abundance and reflects mRNA expression under a specific set of experimental conditions. My analysis is based on a transcript expression data set that was sampled from cells during exponential growth in constant light, so called "circadian free-run" conditions (Vijayan et al., 2009). Transcripts were sampled at successive time points. Those displaying circadian behavior were annotated "subjective dawn" and "subjective dusk", providing a data set suitable for testing the hypothesis that HIP1 3' bias is related to the circadian regulation of genes. This hypothesis is inspired by the observation that both HIP1 hyper-abundance and the circadian gene regulation are observed exclusively in cyanobacteria among prokaryotes.

Codon usage bias is an indirect measure of expression and reflects selection acting on expression levels under many conditions (reviewed in Plotkin and Kudla (2011)). The rationale for using codon usage bias to assess expression levels is that highly expressed genes are more likely to use preferred codons. The use of preferred codons is also linked to translational accuracy and efficiency.

A correlation between codon usage bias and gene length has been reported in *E. coli* (Eyre-Walker, 1996; Moriyama and Powell, 1998). This suggests a potential link between gene length and gene expression, although codon usage bias in this case may be driven by a need for greater translational accuracy in longer transcripts. GC content is correlated with the length of coding sequences, possibly because stop codons are AT-rich.

The transcriptome dataset used in this chapter (Vijayan et al., 2011) specifies the start and termination coordinates of each experimentally verified transcript. Transcript expression level is a direct measure of RNA abundance and reflects the expression of the gene under a specific set of experimental conditions. The expression level data is in the form of absolute transcript level (mRNA molecule per cell), measured by direct RNA sequencing [1]. Thus, the transcript lengths can be obtained directly from the dataset. Transcript GC content was calculated from the transcript sequence, which was obtained by mapping the transcript coordinates onto the genome. Similarly, the number of ORFs (nORF) within each transcript was obtained by comparing the transcriptome map and gene annotation table. The motif frequency was calculated by dividing number of motifs in the transcript by the transcript length.

A number of methods for quantifying codon usage bias have been proposed. I used ACE (Retchless and Lawrence, 2011) and GCB (Merkl, 2003) for this purpose. These quantities were estimated for each gene using DNAMaster [2]. I estimated codon usage bias for

[1] The number of transcripts per cell is estimated assuming 1,500 mRNAs per cell.

[2] Software URL: http://cobamide2.bio.pitt.edu/

each transcript by averaging the codon usage bias for all the protein coding genes in the transcript. The circadian behavior of genes was from a study by Vijayan et al. (2009), in which whole genome microarrays were used to measure gene expression over a 60-hour period. From that data, each gene in *Synechococcus elongatus* PCC7942 (*sel*) was assigned to one of the three categories for circadian expression behavior: Non-Circadian, peak expression at Dusk, and peak expression at Dawn. To assign the circadian categories of genes to transcripts, three properties were associated to each transcript: the number non-circadian genes, the number of circadian genes which peak at dusk, and the number of circadian genes which peak at dawn, within the transcript.

I used several approaches to investigate a possible association between the HIP1 spatial distribution within transcripts and the properties of the transcripts. Transcripts possessing at least one HIP1 motifs were ranked according to their mean motif position and partitioned into batches of 25 transcripts each, resulting in 42 batches. The same procedure was carried out for the control motif. Because there were fewer transcripts possessing at least one control motif, the control set had only 35 batches.

For each batch resulting from this analysis, I calculated the mean value of each of the numerical properties, averaged over all transcripts in the batch. The fraction of transcripts in each batch that have o ORFs, for $o = 1, 2, 3, 4$ and $o \geq 5$, was determined. I also tabulated, for each batch, the number of genes in each of the three circadian categories. These counts were also normalized by the total number of genes in the batch, to obtain the fraction of genes associated with each of the circadian categories. These values are displayed in Figures 4.16 for the HIP1 motif and 4.17 for the control motif.

Comparing Figure 4.11 and Figure 4.12, no functional property stands out visually. Interestingly, the subplots for the averaged transcript lengths, motif numbers, and motifs frequencies are bell shaped for both HIP1 and control motifs. These observations suggest that there exists a correlation between the transcript length and the 3' bias in spatial motif distribution. The shorter the sequence, the more likely the transcripts are to have extreme bias (both 5' and 3'). The bell shapes on the subplots for motif count and frequency also exhibit this length effect, as the motif abundance is related to the transcript length. On average, the number of motifs associated with a transcript increases with its length. This length effect can distort the association between the 3' bias and the various properties. To control for this confounding effect, I used a stratified analysis in which the transcripts were partitioned into sets with similar lengths. The correlation between 3' bias and the various properties was then considered for each set separately.

4.3.1 Stratified analyses

Before performing the stratified analyses, I first carried out a survey of the length effect. I considered both the transcript length in basepairs, and the number of ORFs within the transcript (nORF). Scatter plots of mean position, the number of motifs, and transcript length (bp) are shown in Figures 4.13 and Figures 4.14. Plots of mean position, the number of motifs, and the number of ORFs per transcript are shown in Figure 4.15 and Figure 4.16. These plots reveal the extent of the dependence between these quantities. As the number of motifs increases, the variance in mean position decreases and approaches 0.5. This observation can be made for both HIP1 and control motifs. In general, the longer the transcript, the more likely it is to have a mean position close to 0.5, because HIP1 abundance and transcript length are strongly correlated. This again shows the importance of stratified analyses, to avoid the potential bias introduced by the transcript length.

Though the number of motifs and the transcript length are strongly correlated, it is not necessarily true that this reflects two sides of the same effects. To be safe, I separated the whole set of transcripts according to both the transcript length in terms of the number of ORFs, as well as the number of motifs in the transcript (nMotif). I assigned each transcript to one of the four categories for nORF (nORF = 1,2,3, and nORF > 3), and 5 categories for nMotif (nMotif = 1,2,3,4, and nMotif > 5). Therefore, the whole dataset was divided into 20 subsets. I then studied the relationship between mean motif positions and various functional properties. Here, I show results from two functional properties: ACEu (Figure 4.17 and 4.18), and expression level (Figure 4.19 and 4.20). In these plots, each dot represents a

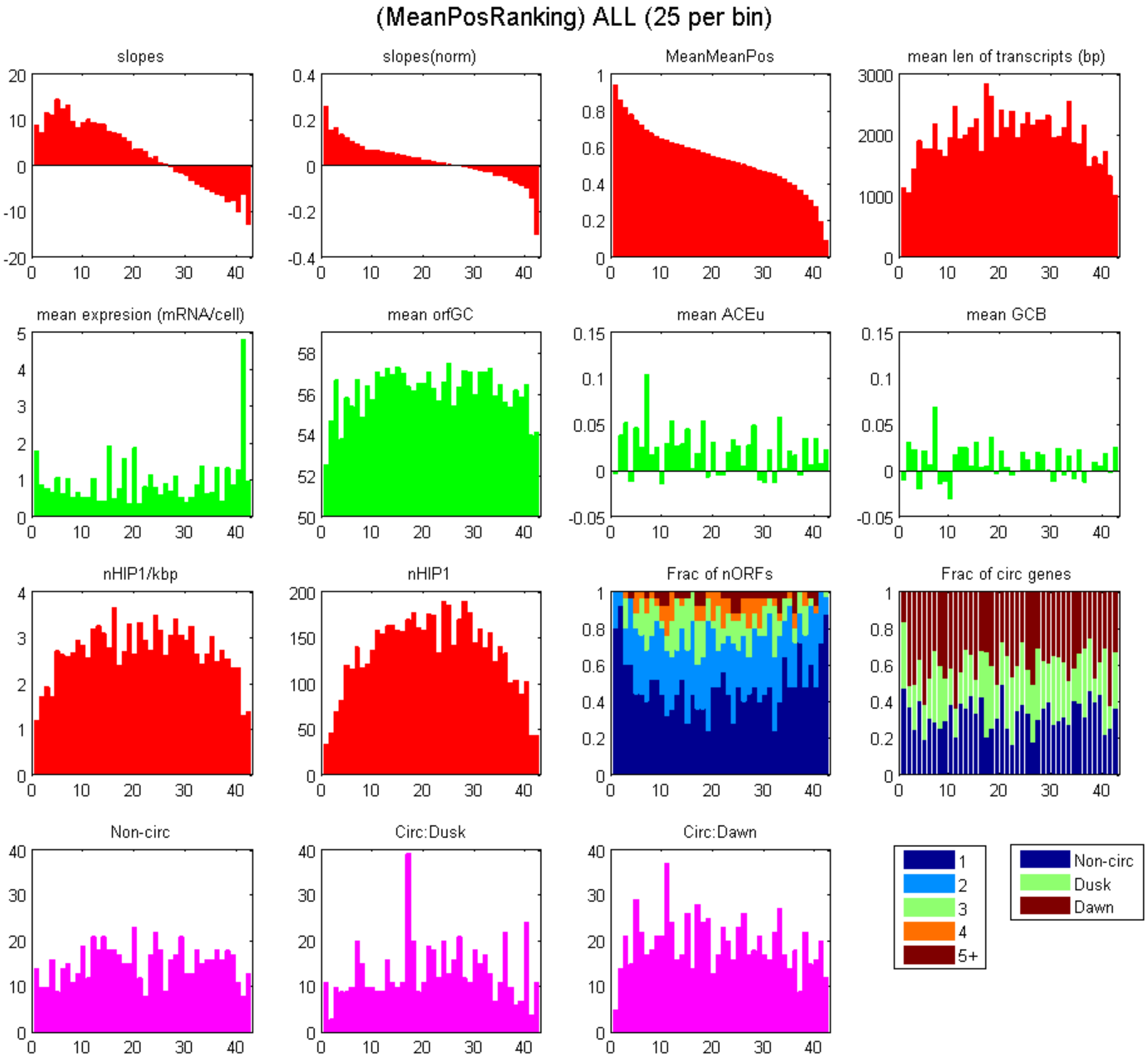


Figure 4.11: Functional properties of interest, in different RMP bins (bin size: 25).

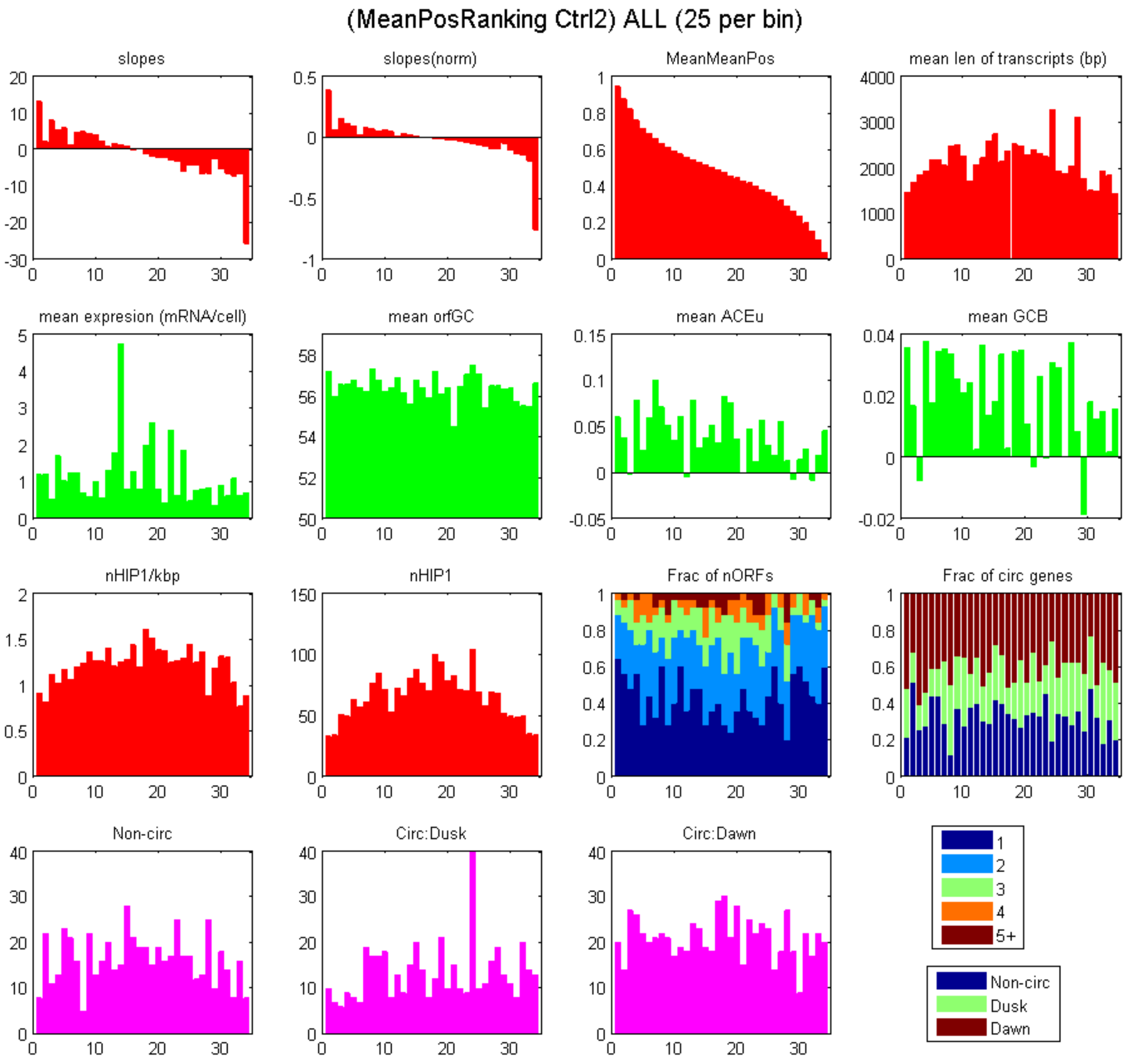


Figure 4.12: Functional properties of interest, in different RMP bins (bin size: 25) based on control motif.

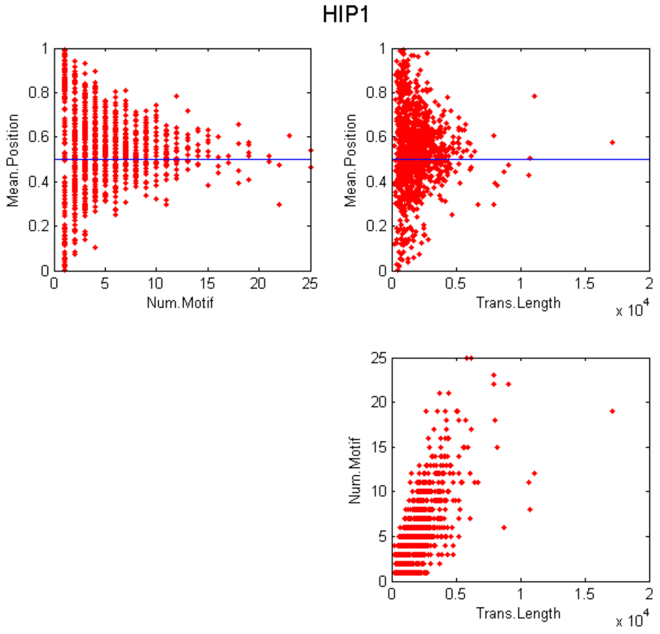


Figure 4.13: Scatterplots showing relationship between mean motif position, transcript length in bp, and per-transcript number of HIP1 motifs.

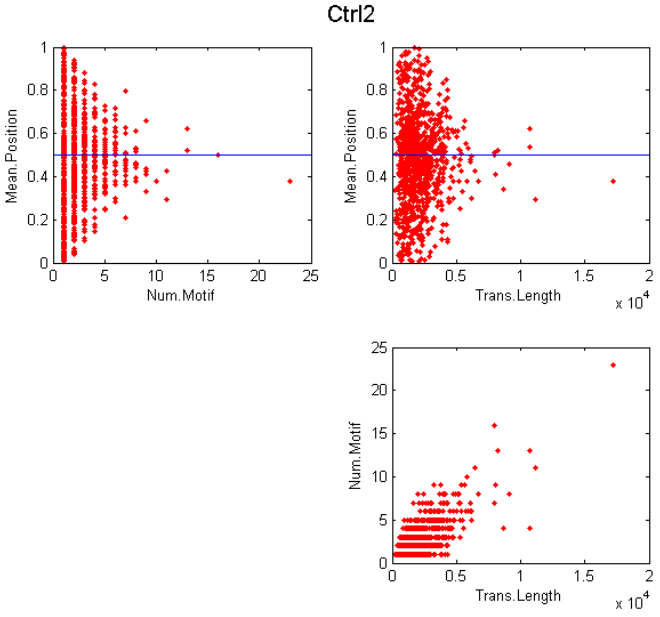


Figure 4.14: Scatterplots showing relationship between mean motif position, transcript length in bp, and per-transcript number of control motifs.

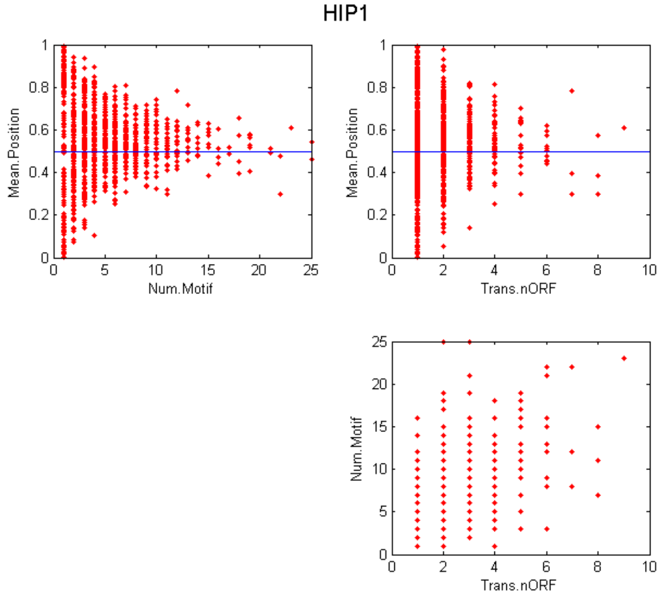


Figure 4.15: Scatterplots showing relationship between mean motif position, transcript length in nORF, and per-transcript number of HIP1 motifs.

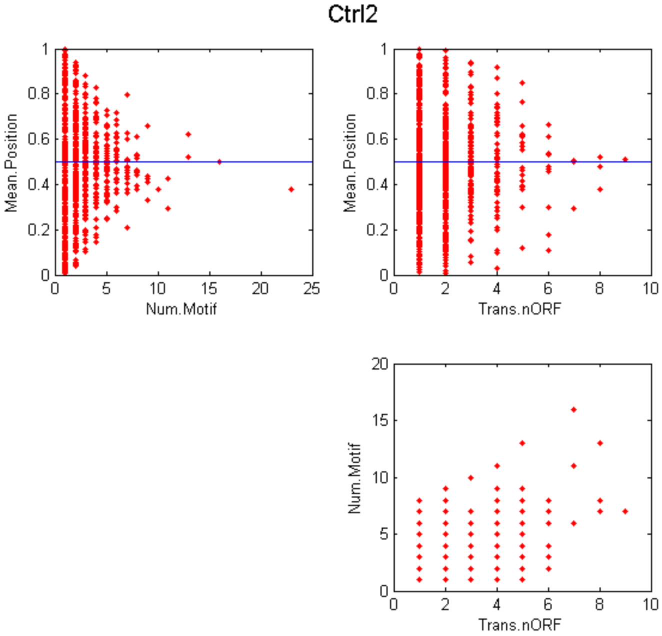


Figure 4.16: Scatterplots showing relationship between mean motif position, transcript length in nORF, and per-transcript number of control motifs.

transcript.

Visually, no observable correlation between mean position and ACE, or expression level, stands out in each of the subplots for both control and HIP1 motifs. Comparing plots generated from HIP1 motif with ones based on control motifs, no distinguishable difference was detected, except that there are more transcripts in the subsets with higher motif counts for HIP1 than for control. Similarly no trend was observed for other functional properties (not shown here), suggesting a lack of relationship between 3' bias and those functional properties tested.

In each of the 20 stratified subsets, I further compared the functional properties within the 20% of transcripts with the greatest mean position, and the 50% of the transcripts with the lowest mean position, using a two-sample Kolmogorov-Smirnov (KS) test. Table 4.4 shows the sample sizes, in terms of numbers of transcripts, of the top 20% and bottom 50% in each subset. The resulting *p*-values from the KS tests are listed in Table 4.5. As the KS *p*-values indicate, none of the KS tests show significant differences between the top 20% and bottom 50% of the sorted transcripts according to RMP, in all stratified subsets. Thus, this analysis provides no evidence to suggest that the observed 3' bias in HIP1 position is related to any of the functional properties tested. However, because of the stratification of the data, the sample sizes are extremely small (Table 4.4). It is possible that a relationship between mean HIP1 position and the various properties does exist, but that we do not have the statistical power to detect it.

	nHIP1=1	nHIP1=2	nHIP1=3	nHIP1=4	nHIP1>4
nORF=1	13 | 31	11 | 25	16 | 38	14 | 33	52 | 130
nORF=2	7 | 17	11 | 25	9 | 22	9 | 22	29 | 70
nORF=3	3 | 5	5 | 12	4 | 8	3 | 7	12 | 28
nORF>3	3 | 6	3 | 6	4 | 9	4 | 8	8 | 20

Table 4.4: Sizes of the stratified subsets (top 20% | bottom 50%).

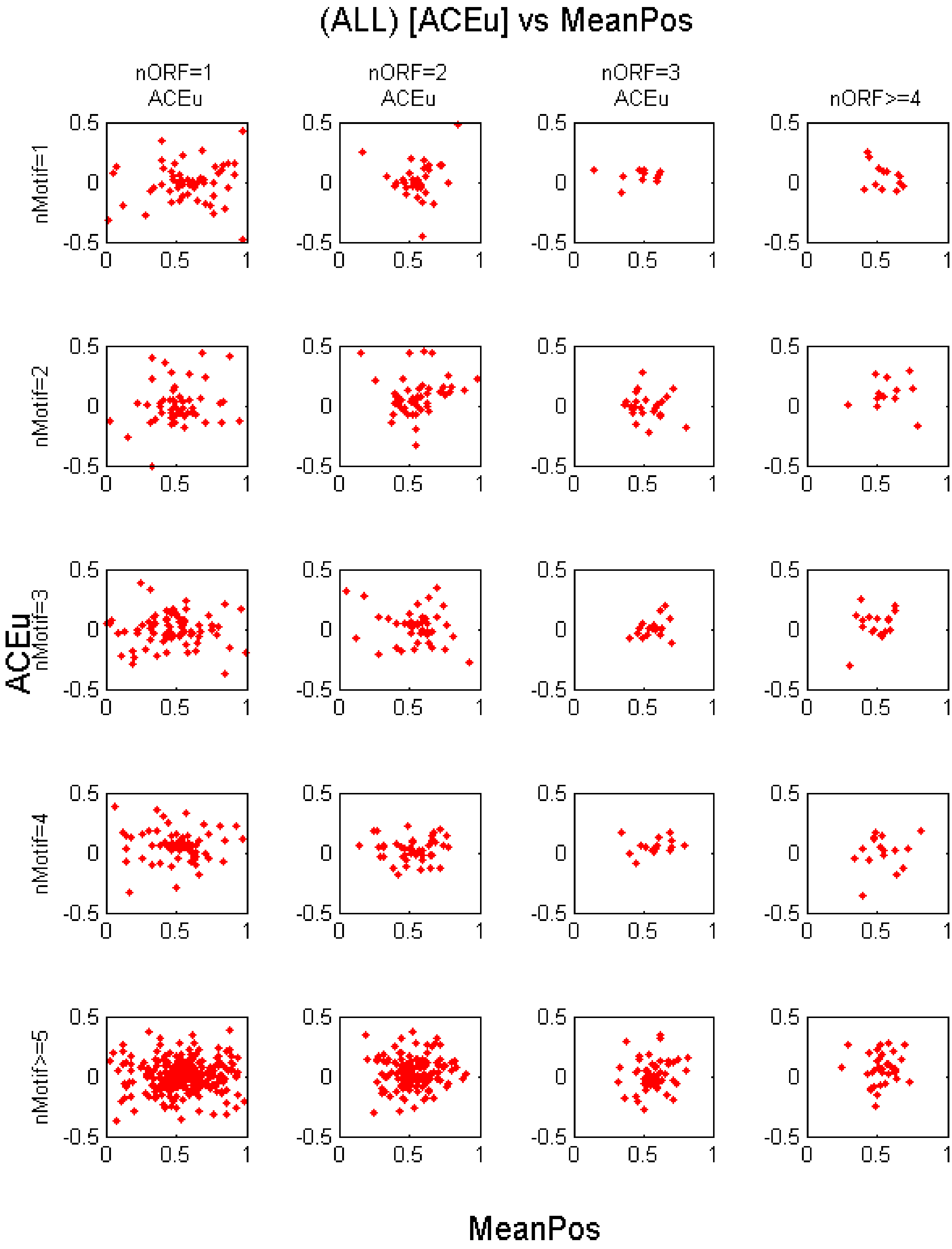


Figure 4.17: Relationship between ACEu and Mean HIP1 Motif Position.

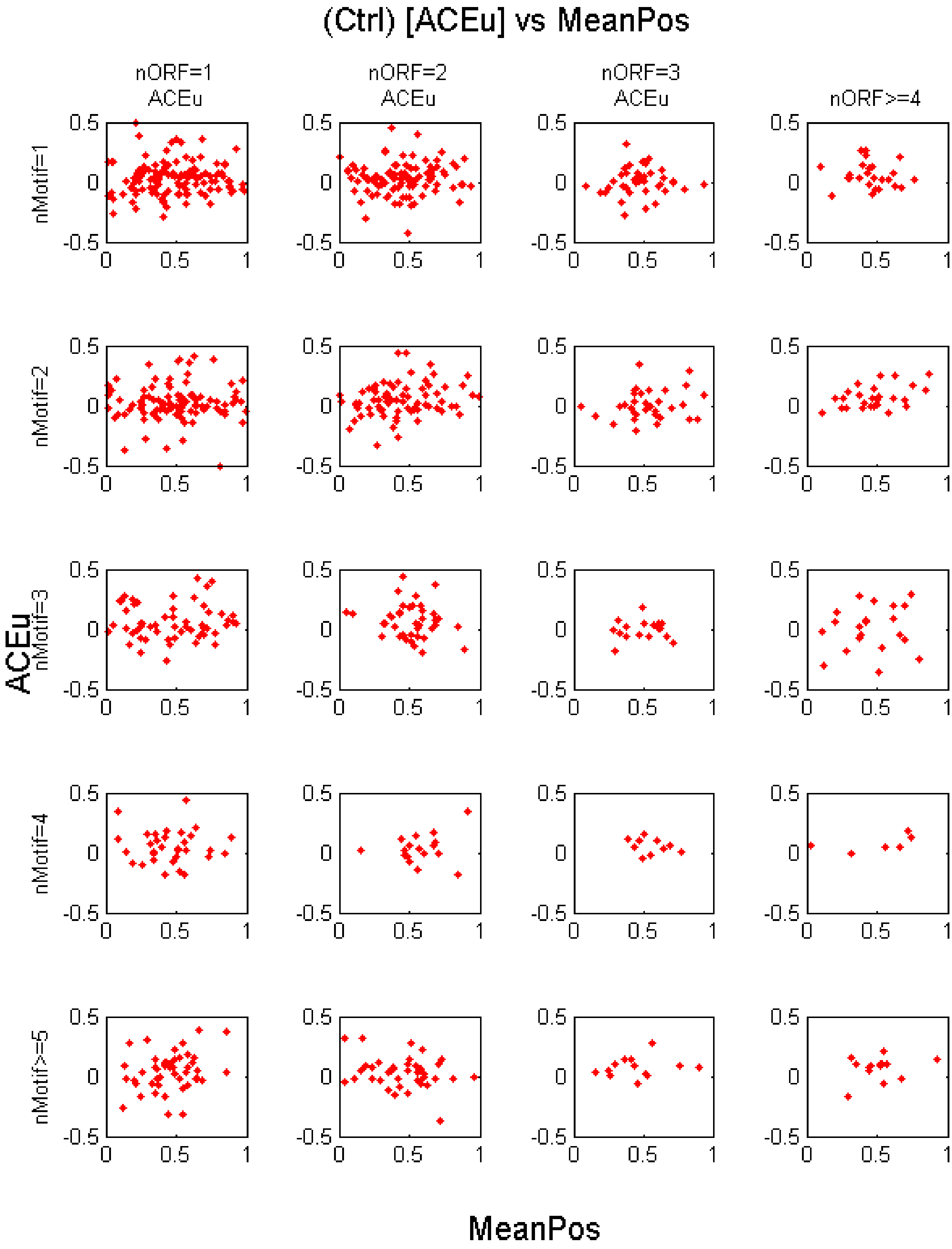


Figure 4.18: Relationship between ACEu and Mean Ctrl Motif Position.

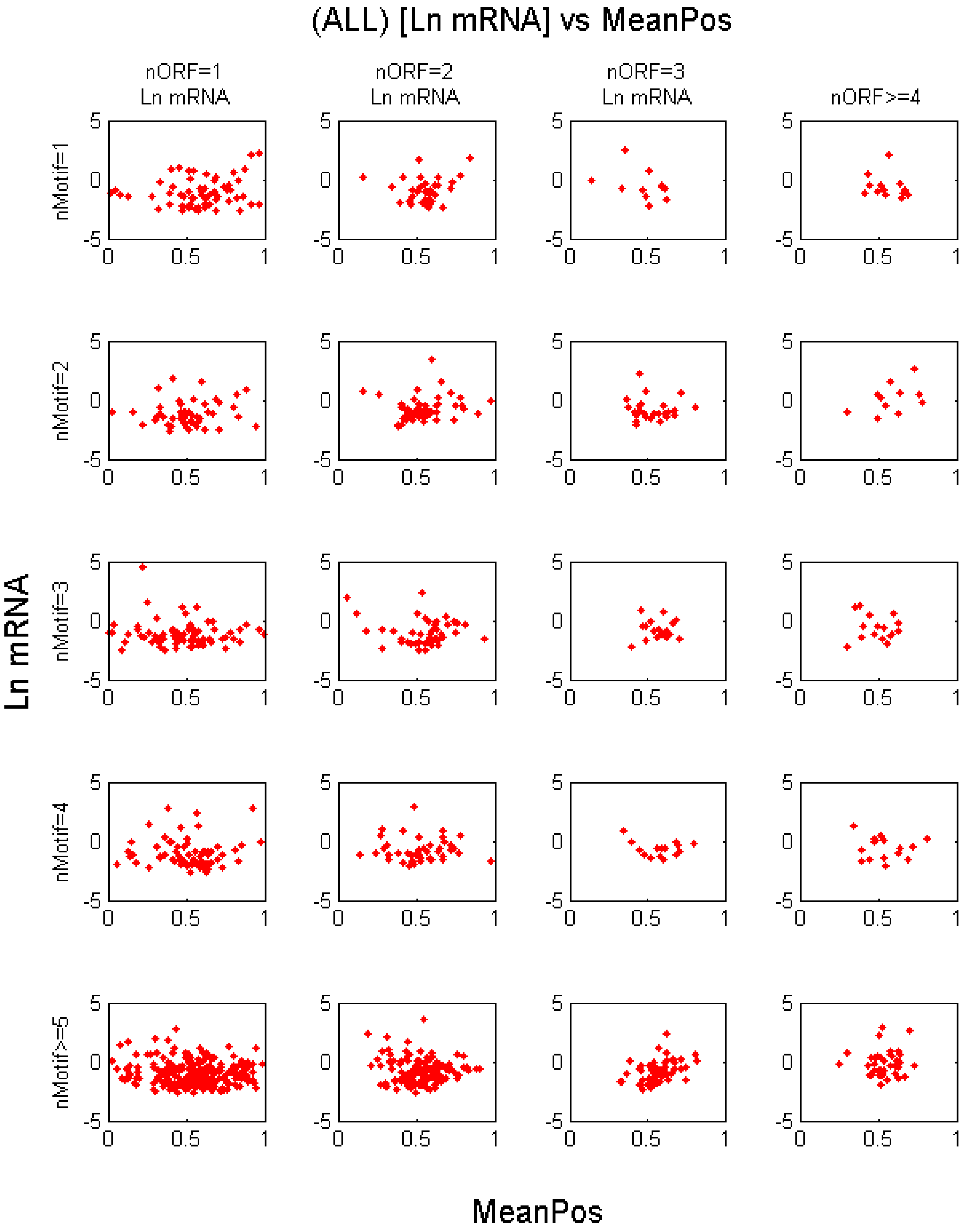


Figure 4.19: Relationship between transcript expression level and Mean HIP1 Motif Position. The unit for expression is log per-cell mRNA count.

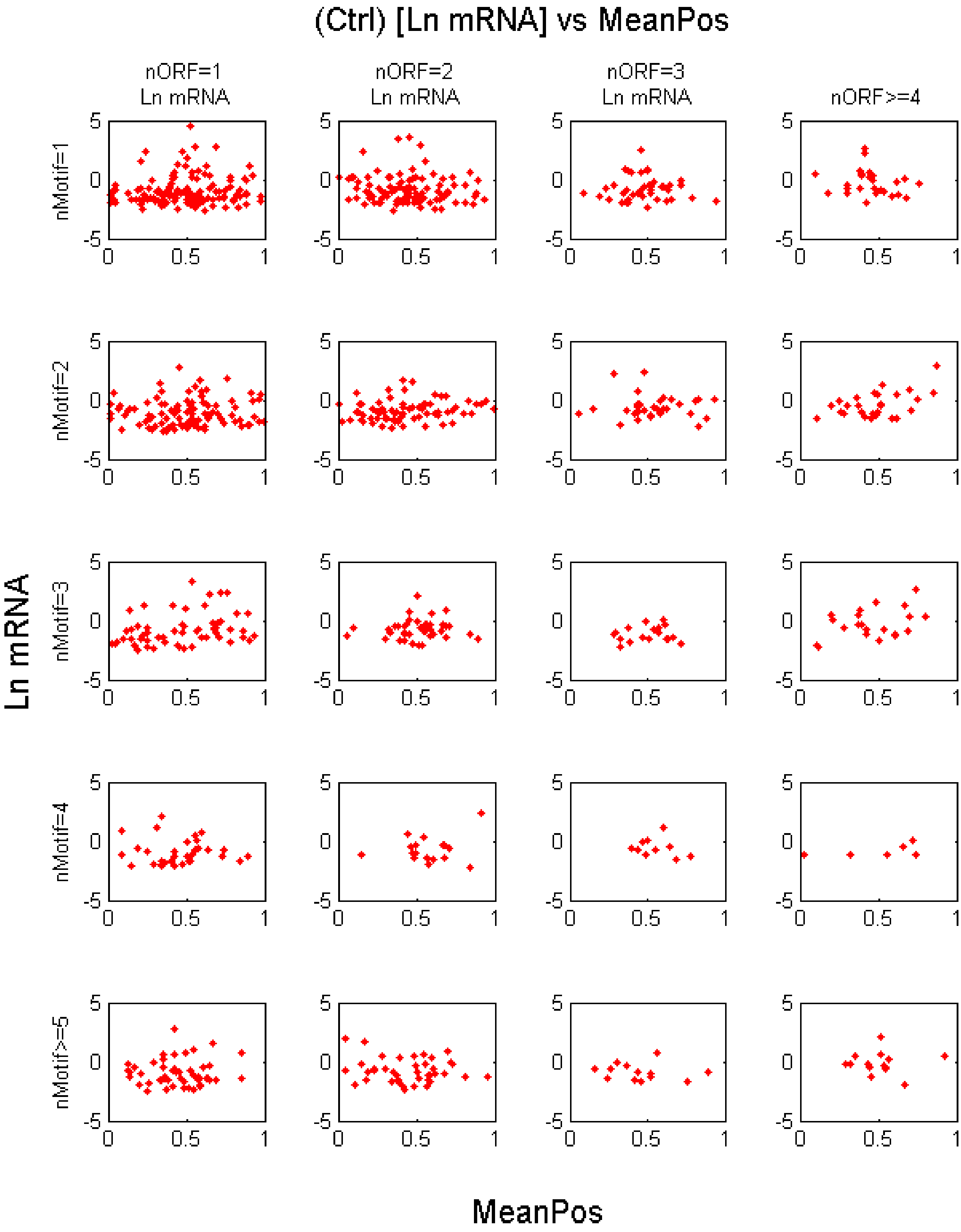


Figure 4.20: Relationship between transcript expression level and Mean Ctrl Motif Position. The unit for expression is log per-cell mRNA count.

4.3.2 Chapter Summary

In this chapter, I explored the intra-genome variation of the HIP1 motif distribution. Instances of regions with high HIP1 occurrence, as well as HIP1-free regions, can be observed in the cyanobacterial genomes. This is particularly interesting, as non-uniform spatial variation of sequence repeats can be indicative of various functional properties. For example, motif periodicity can be an indicator of function in DNA supercoiling, and chromosome organization (Mrazek, 2010).

I further looked into the distribution of HIP1 motifs within genetic units. Interestingly, a 3' bias was detected in the mRNA transcripts in *Synechococcus sp.* PCC 7942. This 3' bias was not observed with control motifs. When HIP1 distribution was assessed in the same transcripts, but with the UTRs removed, the bias was substantially reduced. This suggests that a substantial number of HIP1 motifs participating in this 3' bias are in transcripts but outside the protein coding region. This is consistent with a transcriptional role for HIP1 and further supports the hypothesis that HIP1 is unlikely to have a functional role on the amino acid level.

When the same analysis was applied to predicted operons, the observed 3' bias was also greatly reduced. This may be caused by the difficulty of predicting the coordinates of UTRs. This is unfortunate, because it makes it difficult to determine whether the 3' bias is specific to the *Synechococcus sp.* PCC 7942 genome or a phenomenon that occurs more broadly in cyanobacteria. If the latter, more accurate prediction of operons with UTRs and/or additional transcriptome data in other species would greatly improve our ability to investigate this trend.

In order to determine how individual transcripts contribute to the 3' bias, I designed methods to separate the transcripts dataset into batches with increasing 3' bias. The results show that transcripts represent a mixture of spatial distributions, but with an excess of transcripts in which HIP1 motifs are preferentially located at the 3' end. Having separated transcripts according to the mean HIP1 position, I analyzed the relationship between HIP1 spatial distribution within a transcript, and properties including GC content, codon usage bias, and transcript expression level. This analysis was confounded by a correlation between the length and the number of motifs in a transcript. I attempted to remove this confounding factor by stratifying the data by transcript length and by motif content. However, partitioning the data greatly reduced the statistical power and the results revealed no strong links between HIP1 position and any of the properties tested.

Properties		nHIP1=1	nHIP1=2	nHIP1=3	nHIP1=4	nHIP1>4
Ln mRNA	nORF=1	0.4609	0.2594	0.7615	0.9800	0.2159
Ln mRNA	nORF=2	0.2048	0.0287	0.0536	0.8092	0.0302
Ln mRNA	nORF=3	0.8254	0.9887	0.9857	0.7029	0.0357
Ln mRNA	nORF>3	0.5344	0.5344	0.5083	0.7399	0.4949
GCB	nORF=1	0.0243	0.6190	0.0555	0.6069	0.9437
GCB	nORF=2	0.1359	0.0305	0.0382	0.0538	0.3845
GCB	nORF=3	0.6816	0.8244	0.4623	0.7464	0.4933
GCB	nORF>3	0.8266	0.0235	0.9252	0.9026	0.4591
ACEu	nORF=1	0.2348	0.9401	0.6659	0.3309	0.9437
ACEu	nORF=2	0.0515	0.0053	0.3339	0.8794	0.3242
ACEu	nORF=3	0.1967	0.6176	0.0224	0.3728	0.5179
ACEu	nORF>3	0.1945	0.1441	0.8134	0.9801	0.2044
Trans Len	nORF=1	0.9744	0.7608	0.6464	0.5714	0.2385
Trans Len	nORF=2	0.4416	0.2942	0.6315	0.3970	0.6498
Trans Len	nORF=3	0.6567	0.5074	0.9857	0.4477	0.3325
Trans Len	nORF>3	0.9350	0.1984	0.5921	0.9857	0.2466
GC	nORF=1	0.2504	0.6728	0.9729	0.5135	0.0419
GC	nORF=2	0.8007	0.7608	0.1302	0.9621	0.3239
GC	nORF=3	0.2235	0.0950	0.3788	0.3399	0.6648
GC	nORF>3	0.1984	0.1984	0.1223	0.7399	0.7062
GC(avg ORF)	nORF=1	0.1795	0.9761	0.9874	0.9313	0.0792
GC(avg ORF)	nORF=2	0.8589	0.2278	0.2640	0.6315	0.0388
GC(avg ORF)	nORF=3	0.0854	0.5074	0.9857	0.7029	0.6648
GC(avg ORF)	nORF>3	0.1984	0.1984	0.8414	0.9857	0.4949
nCirc0(non)	nORF=1	0.9783	1.0000	1.0000	1.0000	0.9856
nCirc0(non)	nORF=2	1.0000	0.4974	0.9977	0.3970	1.0000
nCirc0(non)	nORF=3	1.0000	0.9242	0.7399	0.9245	0.2894
nCirc0(non)	nORF>3	0.1984	0.5344	1.0000	1.0000	0.5987
nCirc1(dusk)	nORF=1	0.8078	0.9761	1.0000	1.0000	1.0000
nCirc1(dusk)	nORF=2	0.4719	0.8105	0.6086	0.4978	1.0000
nCirc1(dusk)	nORF=3	1.0000	0.9242	0.7399	0.4477	1.0000
nCirc1(dusk)	nORF>3	0.9350	0.9350	0.5921	1.0000	0.1417
nCirc2(dawn)	nORF=1	0.1674	1.0000	1.0000	1.0000	1.0000
nCirc2(dawn)	nORF=2	0.9290	0.9936	1.0000	0.0498	1.0000
nCirc2(dawn)	nORF=3	1.0000	0.9242	0.3788	0.0089	0.7257
nCirc2(dawn)	nORF>3	0.1984	0.5344	0.9543	0.9857	0.0762

Table 4.5: The KS test p-values comparing the top 20% and bottom 50% (according to mean HIP1 position) transcripts in each category.

Chapter 5

Conclusion and Future Directions

Highly Iterated Palindrome-1 (HIP1) was first discovered in the mid 1990s (Robinson et al., 1995). Over the intervening two decades, HIP1 has been observed to be highly abundant in many cyanobacterial genomes. However, the forces maintaining HIP1 prevalence are still not understood.

This thesis focuses on the investigation of the origins and evolution of the HIP1 motif in cyanobacterial genomes. In Chapter 3, I characterized the taxonomic distribution, abundance, and enrichment of HIP1. The enrichment is based on the ratio of the observed and the expected number of HIP1 motifs in each genome. I estimated the expected number of motifs using a second order Markov model to reflect underlying tri-nucleotide frequencies, taking differences in sequence composition between intergenic and coding regions into account. Variation in tri-nucleotide frequencies across the three reading frames was also considered. In all genomes studied, except for the marine pico-cyanobacteria and *Gloeobacter violaceus* PCC7421 (*gvi*), the number of observed HIP1 motifs exceeds the expected number of motifs by a factor of at least 13. In some genomes, the O/E ratio exceeds 300.

Given these observed high levels of HIP1 enrichment, how HIP1 is maintained in cyanobacterial genomes is a key question. The hypothesis that HIP1 abundance is maintained by a neutral process that constantly replenishes the genome with new HIP1 instances, predicts that HIP1 sites will not be conserved. In contrast, my evidence shows that HIP1 sites are more conserved than would be expected given the degree of divergence in the genome pairs considered. In addition, analysis of codon conservation in HIP1 rules out the possibility that HIP1 is conserved due to selection acting on codon usage. These results support the hypothesis that selection is acting to maintain HIP1. Further, this raises intriguing questions about the targets of selection. By establishing the first evidence for selection acting on HIP1, my results open up the possibility for investigating HIP1's functional aspects.

My analyses led to the discovery of a novel variant of the HIP1 motif (5'-GGGATCCC-3') in two thermophilic strains isolated from a hot spring in Yellowstone National Park, *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*). In contrast, the canonical form (5'-GCGATCGC-3') was not enriched in those genomes. This putative HIP1 variant exhibits levels of abundance, enrichment, and conservation that are comparable to those of canonical HIP1 motifs in HIP1-rich genomes. This suggests that in the Yellowstone strains 5'-GGGATCCC-3' may play an equivalent role to the canonical HIP1 in other cyanobacterial genomes.

To explore the functional aspects of HIP1 motifs, in Chapter 4, I studied the spatial distribution of HIP1, and discovered a 3' bias in transcripts in *Synechococcus elongatus* PCC7942 (*sel*). A similar bias can also be detected in predicted operons in that genome, to a weaker extent. However, a 3' bias in predicted operons was not universally observed; it can only be detected in a selected set of the genomes (data not shown). This could be explained by poor operon prediction. In the transcript data, I observed that the 3' UTR contributes substantially to the strength of the 3' bias. Estimating the length of UTRs is more difficult than estimating the genes in an operon. In particular, in my study, the UTRs were predicted based on a single genome, and the lengths of UTRs in that genome vary greatly (Table 4.2). On the other hand, it is also possible that the 3'-bias is a specific feature in *Synechococcus elongatus* PCC7942 (*sel*). Last, but not least, the 3'-bias could also be an artifact due to error in the transcriptome dataset.

Given the 3' bias of the HIP1 distribution in *Synechococcus elongatus* PCC7942 (*sel*), I further attempted to establish a functional connection between transcripts wtih a strong 3' bias and various properties related to regulation, including transcript abundance, circadian gene expression, codon usage bias, and GC content. Unfortunately, no significant connection was detected. This could be due to the sparsity of accurate functional annotations in cyanobacterial genomes, a lack of statistical power, or the difficulty of distinguishing between transcripts that have a 3' bias for functional reasons and transcripts that have a 3' bias by chance. It could also suggest that HIP1 is not functionally related to any of the properties I tested.

The taxonomic distribution of HIP1 abundance provides insights into both the evolution and the function of HIP1. HIP1 is highly enriched in all genomes studied, except for the marine pico-cyanobacteria and *Gloeobacter violaceus* PCC7421 (*gvi*). Interestingly, *Gloeobacter violaceus* PCC7421 (*gvi*) is the deepest branching cyanobacterial species with a completely sequenced genome. Based on the phylogenetic distribution of HIP1 abundance, two models of HIP1 origin can be hypothesized, based on the principle of maximum parsimony:

1. The origination of HIP1 abundance occurred before the last common cyanobacterial ancestor. HIP1 prevalence was independently lost in *Gloeobacter violaceus* PCC7421 (*gvi*), and marine pico-cyanobacteria.

2. The origination of HIP1 abundance occurred in the common ancestor of all non-*Gloeobacter* cyanobacteria, after the divergence of *Gloeobacter violaceus* PCC7421 (*gvi*). HIP1 prevalence was lost once in the ancestor of all marine pico-cyanobacteria.

The discovery of the HIP1 variant in Yellowstone strains is particularly intriguing, considering the placement of *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) in the cyanobacterial phylogeny. Most cyanobacterial phylogenies in the literature (reviewed in Section 1.4.1) place these strains either as sister taxa to *Gloeobacter violaceus* PCC7421 (*gvi*), or as the sister group to all non-*Gloeobacter* cyanobacterial genomes. In either case, the placement of *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) has interesting implications for the ancestral form of the HIP1 motif, suggesting that the ancestral form of HIP1 could be either the canonical or the Yellowstone variant. If the former is true, the HIP1 form in *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*) possibly evolved to its current form when adapting to the thermophilic environment. Shifting from the canonical form to the 5'-GGGATCCC-3' variant, would have resulted in an alteration of free energy at both the

DNA and RNA level, with possible consequences for fitness.

The taxonomic distribution of HIP1 abundance may also contain clues to its function. Since HIP1 hyper-abundance is unique to cyanobacteria, it is tempting to seek functional hypotheses related to unique physiological features of cyanobacteria; i.e., oxygenic photosynthesis and circadian gene expression. In the genomes analyzed in this thesis, the species lacking HIP1 enrichment only from two clades, the basal species *Gloeobacter violaceus* PCC7421 (*gvi*) and the marine pico-cyanobacteria. Both groups have functional characteristics that set them apart from the "canonical" cyanobacterium.

Gloeobacter violaceus lack a thylakoid membrane, the location where the light-dependent reactions of photosynthesis occur in other cyanobacteria. Instead, these reactions occur in the plasma membrane. Several components of the canonical photosynthetic machinery are also lacking in *Gloeobacter violaceus* PCC7421 (*gvi*), suggesting a different photosynthesis scenario in this genome. In addition *Gloeobacter violaceus* PCC7421 (*gvi*) lacks a number of other features that are common to many other cyanobacterial genomes. For example, KaiABC, the three genes which encode for core components of the cyanobacterial circadian clock, are all missing in *Gloeobacter violaceus* PCC7421 (*gvi*), although LdpA, LabA, and RpaA are present.

Marine pico-cyanobacteria were also found to lack HIP1 abundance. Interestingly, they also lack many cyanobacterial pathways, and are believed to have undergone genome reduction in the process of adaptation to the nutrient-rich marine environment. It is likely that HIP1 prevalence was lost in the last common ancestor of all marine pico-cyanobacteria, along with the mechanism or pathway responsible for HIP1 abundance. The marine pico-cyanobacteria appear to have lost virtually all repetitive sequences (Treangen et al., 2009), so the loss of HIP1 in these genomes may be part of a larger trend.

In general, the photosynthetic machinery in marine pico-cyanobacteria (reviewed in Scanlan et al. (2009)) is similar to that of freshwater species, although most Prochlorococcus strains lack several extrinsic photosystem II proteins. The light harvesting systems in the pico-cyanobacteria differ substantially, however, from light harvesting strategies in non-pico species. There are also major differences between the light harvesting proteins in the two the two pico-cyanobacterial species, reflecting the differences in ecological adaptation.

Although Prochlorococcus species lack KaiA, as well as several proteins from the circadian input and output pathways, KaiBC and various other proteins associated with the circadian circuitry are still present. Axmann et al. (2014) have proposed that while pico-cyanobacteria lack a full oscillatory clock, their genomes do encode an “hourglass-like” timing mechanism.

In addition to the marine pico-cyanobacteria, my data set contains the genomes of two symbiotic strains that are in the process of genome reduction. Interestingly, low HIP1 abundance (37 HIP1 copies), but moderate HIP1 enrichment (28.46) was detected in cyanobacterium UCYN-A (*uca*) (Figure 3.3), an AT-rich reduced genome closely related to *Cyanothece*. Cyanobacterium UCYN-A (*uca*) is a symbiont to a prymnesiophyte, alga that is itself capable of photosynthesis. cyanobacterium UCYN-A (*uca*) lacks photosystem II and some key enzymes in the Calvin cycle. Some key genes for circadian regulation are also absent from the cyanobacterium UCYN-A (*uca*) genome, including KaiA, KaiB, Pex, LdpA, and LabA (Axmann et al., 2014).

Among the genomes analyzed, *Nostoc azollae* 0708 (*naz*)is another symbiont that is currently undergoing genome reduction. The *Nostoc azollae* 0708 (*naz*) genome is scattered with transposable elements and pseudogenes (Ran et al., 2010). Since *Nostoc azollae* 0708 (*naz*) still retains a large number of discernible pseudogenes, it may be possible to observe the decay process “in action”. Interestingly, despite its symbiotic life style, *Nostoc azollae* 0708 (*naz*) is capable of photosynthesis. As far as I know, its circadian clock machinery has not been studied.

Both *Nostoc azollae* 0708 (*naz*) and cyanobacterium UCYN-A (*uca*) have relatively low HIP1 frequencies (0.03 and 0.21, respectively), but HIP1 motifs are substantially enriched in both species. At this point, it is not possible to determine whether these genome are in the process of losing HIP1 enrichment, in tandem with genome reduction. Analysis of additional cyanobacterial symbionts with reduced genomes may reveal a link between the loss of genes or pathways and the loss of HIP1 abundance.

Several functional hypotheses are intriguing targets for future study. Given the weak spatial bias among reading frames, and between coding and intergenic regions, it is possible that HIP1 is involved in maintaining the structure or regulating the topological status of the chromosome. It is also possible that such a function can have a regional effect, such as adjusting the local chromosome relaxation status. It has been shown that the circadian change of chromosome topology, between relaxation and condensation, contributes to the circadian pattern of gene expression in *Synechococcus elongatus* PCC7942 (*sel*) (Vijayan et al., 2009).

Alternatively, HIP1 may function at a local level, consistent with the observed 3' bias in transcripts in *Synechococcus elongatus* PCC7942 (*sel*). Previous studies have reported examples of bacterial repetitive sequences with functional roles in transcription regulation through transcription termination of 3' to 5' degradation. The transcription termination factor Rho is reported to be associated with BIME for transcription attenuation (Espeli et al., 2001). There is evidence that REPs contribute to the protection of the 3' ends of mRNA molecules, from exonucleolytic degradation by exonuclease III (Khemici and Carpousis, 2004). Both of these functions are consistent with the observed HIP1 3' bias. However, HIP1 differs substantially from the REP sequence in motif size and structure. Because of HIP1's very short sequence length, clusters of HIP1 motifs would be required for such functions. Another hypothesis is that HIP1 motifs at the 3' end of transcripts may contribute to the function of stable secondary structures. A pair of neighboring HIP1 motifs could form a local hairpin structure at the mRNA level, similar to the REPIN sequence, where a pair of REP instances separated by a specific distance (e.g., $\approx$71 and 110 bp in *Pseudomonas fluorescens*) form a local hairpin structure (Bertels and Rainey, 2011). Investigating the impact of HIP1 on the secondary structure and local free energy at the 3' end of transcripts is a worthy direction for future study.

HIP1 may also be contributing to genomic plasticity. The highly conserved sequence pattern of HIP1 could cause intra-genome recombination, and thus promote genome rearrangements. This hypothesis could be investigated by looking at the HIP1 spatial distribution relative to breakpoints in pair-wise whole genome alignments of related genomes. It is interesting to see that the presence of HIP1 has been reported in cyanophage genomes (Delaye et al., 2011b). HIP1 can potentially provide a new angle to study how phage DNA interacts with cyanobacterial genomes.

Bibliography

Abed RM, Dobretsov S, Sudesh K (2009) Applications of cyanobacteria in biotechnology. J Appl Microbiol 106: 1–12.

Akiyama H, Kanai S, Hirano M, Miyasaka H (1998) A novel plasmid recombination mechanism of the marine cyanobacterium *Synechococcus sp.* PCC 7002. DNA Res 5: 327–34.

Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12: 402.

Aranda-Olmedo I, Tobes R, Manzanera M, Ramos JL, Marques S (2002) Species-specific Repetitive Extragenic Palindromic (REP) sequences in *Pseudomonas putida*. Nucleic Acids Res 30: 1826–33.

Axmann IM, Hertel S, Wiegard A, Dorrich AK, Wilde A (2014) Diversity of KaiC-based timing systems in marine Cyanobacteria. Mar Genomics 14: 3–16.

Azad RK, Lawrence JG (2012) Detecting laterally transferred genes. Methods Mol Biol 855: 281–308.

Bachellier S, Saurin W, Perrin D, Hofnung M, Gilson E (1994) Structural and functional diversity among Bacterial Interspersed Mosaic Elements (BIMEs). Mol Microbiol 12: 61–70.

Bandyopadhyay A, Elvitigala T, Welsh E, Stockel J, Liberton M, et al. (2011) Novel metabolic attributes of the genus cyanothece, comprising a group of unicellular nitrogen-fixing Cyanothece. MBio 2.

Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics 28: i283–291.

Bansal MS, Banay G, Gogarten JP, Shamir R (2011) Detecting highways of horizontal gene transfer. J Comput Biol 18: 1087–114.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709–12.

Baum D (2013) The origin of primary plastids: a pas de deux or a menage a trois? Plant Cell 25: 4–6.

Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A 102: 14332–7.

Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol 63: 240–50.

Bertels F, Rainey PB (2011) Within-genome evolution of REPINs: a new family of miniature mobile DNA in bacteria. PLoS Genet 7: e1002132.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, et al. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J 1: 703–13.

Bordewich M, Semple C, Talbot J (2004) Counting consistent phylogenetic trees is #P-complete. Advances in Applied Mathematics 33: 416 – 430, URL `http://www.sciencedirect.com/science/article/pii/S0196885804000107`.

Bourgon R, Delorenzi M, Sargeant T, Hodder AN, Crabb BS, et al. (2004) The serine repeat antigen (SERA) gene family phylogeny in Plasmodium: the impact of GC content and reconciliation of gene and species trees. Mol Biol Evol 21: 2161–2171.

Brinkman FS, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, et al. (2002) Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Genome Res 12: 1159–67.

Brocchieri L (2014) The GC Content of Bacterial Genomes. Journal of Phylogenetics and Evolutionary Biology 2: 1–3.

Chatterjee R, Chaudhuri K, Chaudhuri P (2008) On detection and assessment of statistical significance of Genomic Islands. BMC Genomics 9: 150.

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7: 429–47.

Collings CK, Fernandez AG, Pitschka CG, Hawkins TB, Anderson JN (2010) Oligonucleotide sequence motifs as nucleosome positioning signals. PLoS One 5: e10933.

Cormen TH, Leiserson CE, Rivest RL, Stein C, et al. (2001) Introduction to algorithms, vol. 2. MIT press Cambridge.

Courvalin P (1994) Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. Antimicrob Agents Chemother 38: 1447–51.

Criscuolo A, Gribaldo S (2011) Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol 28: 3019–32.

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14: 1394–403.

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147.

David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469: 93–96.

Delaye L, Gonzalez-Domenech CM, Garcillan-Barcia MP, Pereto J, de la Cruz F, et al. (2011a) Blueprint for a minimal photoautotrophic cell: conserved and variable genes in Synechococcus elongatus PCC 7942. BMC Genomics 12: 25.

Delaye L, Gonzalez-Domenech CM, Garcillan-Barcia MP, Pereto J, de la Cruz F, et al. (2011b) Blueprint for a minimal photoautotrophic cell: conserved and variable genes in *Synechococcus elongatus* PCC 7942. BMC Genomics 12: 25.

Delihas N (2011) Impact of small repeat sequences on bacterial genome evolution. Genome Biol Evol 3: 959–73.

Dessimoz C, Margadant D, Gonnet G (2008) DLIGHT Lateral Gene Transfer Detection Using Pairwise Evolutionary Distances in a Statistical Framework. In: Vingron M, Wong L, editors, Research in Computational Molecular Biology, vol. 4955 of Lecture Notes in Computer Science, pp. 315–330, Springer Berlin Heidelberg.

Donati B, Baudet C, Sinaimeri B, Crescenzi P, Sagot MF (2015) EUCALYPT: efficient tree reconciliation enumerator. Algorithms Mol Biol 10: 3.

Doyon JP, Ranwez V, Daubin V, Berry V (2011) Models, algorithms and programs for phylogeny reconciliation. Brief Bioinform 12: 392–400.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 33: e6.

Dvornyk V, Vinogradova O, Nevo E (2003) Origin and evolution of circadian clock genes in prokaryotes. Proc Natl Acad Sci U S A 100: 2495–500.

Ebersberger I, Galgoczy P, Taudien S, Taenzer S, Platzer M, et al. (2007) Mapping human genetic ancestry. Mol Biol Evol 24: 2266–76.

El Karoui M, Biaudet V, Schbath S, Gruss A (1999) Characteristics of Chi distribution on different bacterial genomes. Res Microbiol 150: 579–87.

Elhai J, Kato M, Cousins S, Lindblad P, Costa JL (2008) Very small mobile repeated elements in cyanobacterial genomes. Genome Res 18: 1484–99.

Espeli O, Moulin L, Boccard F (2001) Transcription attenuation associated with bacterial repetitive extragenic BIME elements. J Mol Biol 314: 375–86.

Eyre-Walker A (1996) Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol 13: 864–872.

Fattash I, Rooke R, Wong A, Hui C, Luu T, et al. (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56: 475–86.

Fewer DP, Halinen K, Sipari H, Bernardova K, Manttari M, et al. (2011) Non-autonomous transposable elements associated with inactivation of microcystin gene clusters in strains of the genus Anabaena isolated from the Baltic Sea. Environ Microbiol Rep 3: 189–94.

Fitch WM (2000) Homology a personal view on some of the problems. Trends Genet 16: 227–231.

Fitzpatrick DA (2012) Horizontal gene transfer in fungi. FEMS Microbiol Lett 329: 1–8.

Gilson E, Perrin D, Clement JM, Szmelcman S, Dassa E, et al. (1986) Palindromic units from E. coli as binding sites for a chromoid-associated protein. FEBS Lett 206: 323–8.

Gilson E, Saurin W, Perrin D, Bachellier S, Hofnung M (1991) Palindromic units are part of a new bacterial interspersed mosaic element (BIME). Nucleic Acids Res 19: 1375–1383.

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19: 2226–38.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, et al. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9: 585–98.

Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8: 172.

Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59: 2510–26.

Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23: 1089–97.

Hallett M, Lagergren J, Tofigh A (2004) Simultaneous Identification of Duplications and Lateral Transfers. In: Proceedings of the Eighth Annual International Conference on Resaerch in Computational Molecular Biology, RECOMB '04, pp. 347–356, New York, NY, USA: ACM.

Hendrickson H, Lawrence JG (2006) Selection for chromosome architecture in bacteria. J Mol Evol 62: 615–29.

Higgins CF, Ames GF, Barnes WM, Clement JM, Hofnung M (1982) A novel intercistronic regulatory element of prokaryotic operons. Nature 298: 760–2.

Higgins CF, McLaren RS, Newbury SF (1988) Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. Gene 72: 3–14.

Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. Genome Biol Evol 3: 23–35.

Ishiura M, Kutsuna S, Aoki S, Iwasaki H, Andersson CR, et al. (1998) Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. Science 281: 1519–23.

Jiang N, Feschotte C, Zhang X, Wessler SR (2004) Using rice to understand the origin and amplification of Miniature Inverted repeat Transposable Elements (MITEs). Current opinion in plant biology 7: 115–119.

Kaneko T, Nakajima N, Okamoto S, Suzuki I, Tanabe Y, et al. (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium Microcystis aeruginosa NIES-843. DNA Res 14: 247–56.

Karlin S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol 1: 598–610.

Karlin S, Brendel V (1992) Chance and statistical significance in protein and DNA sequence analysis. Science 257: 39–49.

Karlin S, Brocchieri L, Mrazek J, Campbell AM, Spormann AM (1999) A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. Proc Natl Acad Sci U S A 96: 9190–5.

Karlin S, Weinstock GM, Brendel V (1995) Bacterial classifications derived from recA protein sequence comparisons. J Bacteriol 177: 6881–93.

Katayama M, Tsinoremas NF, Kondo T, Golden SS (1999) cpmA, a gene involved in an output pathway of the cyanobacterial circadian system. J Bacteriol 181: 3516–24.

Katayama T, Okamoto S, Narikawa R, Fujisawa T, Kawashima S, et al. (2002) Comprehensive analysis of tandem repeat sequences in cyanobacteria genome. Genome Informatics 13: 400–401.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–9.

Keeling PJ (2010) The endosymbiotic origin, diversification and fate of plastids. Philos Trans R Soc Lond B Biol Sci 365: 729–48.

Kersey P, Bower L, Morris L, Horne A, Petryszak R, et al. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res 33: D297–302.

Khemici V, Carpousis AJ (2004) The RNA degradosome and poly(A) polymerase of Escherichia coli are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. Mol Microbiol 51: 777–90.

Kneip C, Voss C, Lockhart PJ, Maier UG (2008) The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. BMC Evol Biol 8: 30.

Kump LR (2008) The rise of atmospheric oxygen. Nature 451: 277–8.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: reconstructing the microbial phylogenetic network. Genome Res 15: 954–9.

Larsson J, Nylander JA, Bergman B (2011) Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol 11: 187.

Latysheva N, Junker VL, Palmer WJ, Codd GA, Barker D (2012) The evolution of nitrogen fixation in cyanobacteria. Bioinformatics 28: 603–6.

Lawrence JG, Hartl DL (1992) Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics 131: 753–60.

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383–97.

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95: 9413–7.

Li L, Stoeckert J C J, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–89.

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2: 150–174.

Libeskind-Hadas R, Wu YC, Bansal MS, Kellis M (2014) Pareto-optimal phylogenetic tree reconciliation. Bioinformatics 30: 87–95.

Lin S, Haas S, Zemojtel T, Xiao P, Vingron M, et al. (2011) Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. Gene 473: 139–49.

Liu Y, Tsinoremas NF, Johnson CH, Lebedeva NV, Golden SS, et al. (1995) Circadian orchestration of gene expression in cyanobacteria. Genes Dev 9: 1469–78.

Loza-Correa M, Gomez-Valero L, Buchrieser C (2010) Circadian clock proteins in prokaryotes: hidden rhythms? Front Microbiol 1: 130.

Ma B, Li M, Zhang L (2000) From Gene Trees to Species Trees. SIAM Journal on Computing 30: 729–752.

MacLeod D, Charlebois RL, Doolittle F, Bapteste E (2005) Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evol Biol 5: 27.

Mali P, Esvelt KM, Church GM (2013) Cas9 as a versatile tool for engineering biology. Nat Methods 10: 957–963.

Manzanera M, Aranda-Olmedo I, Ramos JL, Marques S (2001) Molecular characterization of *Pseudomonas putida* KT2440 rpoH gene regulation. Microbiology 147: 1323–30.

McFadden GI, van Dooren GG (2004) Evolution: red algal genome affirms a common origin of all plastids. Curr Biol 14: R514–6.

McLaren RS, Newbury SF, Dance GS, Causton HC, Higgins CF (1991) mRNA degradation by processive 3'-5' exoribonucleases in vitro and the implications for prokaryotic mRNA decay in vivo. J Mol Biol 221: 81–95.

Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222: 851–6.

Merkl R (2003) A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. J Mol Evol 57: 453–66.

Min H, Golden SS (2000) A new circadian class 2 gene, opcA, whose product is important for reductant production at night in Synechococcus elongatus PCC 7942. J Bacteriol 182: 6214–6221.

Moriyama EN, Powell JR (1998) Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res 26: 3188–3193.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130.

Mrazek J (2010) Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression. J Bacteriol 192: 3763–72.

Mrazek J, Gaynon LH, Karlin S (2002) Frequent oligonucleotide motifs in genomes of three *Streptococci*. Nucleic Acids Research 30: 4216–21.

Mulkidjanian AY, Koonin EV, Makarova KS, Mekhedov SL, Sorokin A, et al. (2006) The cyanobacterial genome core and the origin of photosynthesis. Proc Natl Acad Sci U S A 103: 13126–31.

Nakajima M, Imai K, Ito H, Nishiwaki T, Murayama Y, et al. (2005) Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. Science 308: 414–415.

Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, et al. (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids (supplement). DNA Res 10: 181–201.

Nakhleh L (2011) Evolutionary Phylogenetic Networks: Models and Issues. In: Problem Solving Handbook in Computational Biology and Bioinformatics, pp. 125–158, Springer US.

Nakhleh L, Ruths D (2009) Gene trees, species trees, and species networks. In: Guerra R, Goldstein D, editors, Meta-analysis and Combining Information in Genetics and Genomics, pp. 275–293, Boca Raton, FL, USA.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–26.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323–9.

Newbury SF, Smith NH, Robinson EC, Hiles ID, Higgins CF (1987) Stabilization of translationally active mRNA by prokaryotic REP sequences. Cell 48: 297–310.

Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, et al. (2004) Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 186: 6575–6585.

Nunvar J, Licha I, Schneider B (2013) Evolution of REP diversity: a comparative study. BMC Genomics 14: 385.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

Olson JM (2006) Photosynthesis in the Archean era. Photosynth Res 88: 109–17.

Ota T, Nei M (1994) Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol Biol Evol 11: 613–619.

Page RD, Charleston MA (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 7: 231–40.

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401: 877–884.

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53: 673–84.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5: 568–83.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–8.

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12: 32–42.

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet 2: e173.

Py B, Higgins CF, Krisch HM, Carpousis AJ (1996) A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. Nature 381: 169–172.

Ran L, Larsson J, Vigil-Stenman T, Nylander JA, Ininbergs K, et al. (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One 5: e11486.

Rasmussen MD, Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res 22: 755–65.

Ravenhall M, Skunca N, Lassalle F, Dessimoz C (2015) Inferring horizontal gene transfer. PLoS Computational Biology 0: 0–10.

Retchless AC, Lawrence JG (2010) Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A 107: 11453–8.

Retchless AC, Lawrence JG (2011) Quantification of codon selection for comparative bacterial genomics. BMC Genomics 12: 374.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Journal of General Microbiology 111: 1–61.

Robinson NJ, Robinson PJ, Gupta A, Bleasby AJ, Whitton BA, et al. (1995) Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. Nucleic Acids Res 23: 729–35.

Robinson PJ, Cranenburgh RM, Head IM, Robinson NJ (1997) HIP1 propagates in cyanobacterial DNA via nucleotide substitutions but promotes excision at similar frequencies in *Escherichia coli* and *Synechococcus* PCC 7942. Mol Microbiol 24: 181–9.

Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22: 1337–44.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, et al. (2009) Ecological genomics of marine picocyanobacteria. Microbiol Mol Biol Rev 73: 249–99.

Schirrmeister BE, Antonelli A, Bagheri HC (2011) The origin of multicellularity in cyanobacteria. BMC Evol Biol 11: 45.

Schmitz O, Katayama M, Williams SB, Kondo T, Golden SS (2000) CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. Science 289: 765–8.

Searls DB (2003) Pharmacophylogenomics: genes, evolution and drug targets. Nat Rev Drug Discov 2: 613–623.

Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK (2009) The continuing puzzle of the great oxidation event. Curr Biol 19: R567–74.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, et al. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A 110: 1053–8.

Smith GR (2012) How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev 76: 217–28.

Smith HO, Gwinn ML, Salzberg SL (1999) DNA uptake signal sequences in naturally transformable bacteria. Res Microbiol 150: 603–16.

Smith RM, Williams SB (2006) Circadian rhythms in gene transcription imparted by chromosome compaction in the cyanobacterium *Synechococcus elongatus*. Proc Natl Acad Sci U S A 103: 8564–9.

Stern MJ, Ames GF, Smith NH, Robinson EC, Higgins CF (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell 37: 1015–26.

Steunou AS, Bhaya D, Bateson MM, Melendrez MC, Ward DM, et al. (2006) In situ analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. Proc Natl Acad Sci U S A 103: 2398–403.

Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, et al. (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28: i409–i415.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609–12.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9: 322.

Tobes R, Pareja E (2006) Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. BMC Genomics 7: 62.

Tofigh A, Hallett M, Lagergren J (2011) Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans Comput Biol Bioinform 8: 517–35.

Treangen TJ, Abraham AL, Touchon M, Rocha EP (2009) Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev 33: 539–71.

Uno R, Nakayama Y, Arakawa K, Tomita M (2000) The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene 259: 207–15.

Versalovic J, Koeuth T, Lupski JR (1991) Distribution of repetitive DNA-Sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Research 19: 6823–6831.

Vijayan V, Jain IH, O'Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. Genome Biol 12: R47.

Vijayan V, Zuzow R, O'Shea EK (2009) Oscillations in supercoiling drive circadian gene expression in cyanobacteria. Proc Natl Acad Sci U S A 106: 22564–8.

Vioque A (2007) Transformation of cyanobacteria. Adv Exp Med Biol 616: 12–22.

Wang H, Fewer DP, Sivonen K (2011) Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. PLoS One 6: e22384.

Whitton BA (2012) Ecology of cyanobacteria II : their diversity in space and time. New York: Springer.

Whitton BA, Potts M (2000) The ecology of cyanobacteria : their diversity in time and space. Boston: Kluwer Academic.

Woelfle MA, Xu Y, Qin X, Johnson CH (2007) Circadian rhythms of superhelical status of DNA in cyanobacteria. Proc Natl Acad Sci U S A 104: 18819–24.

Wolk C, Ernst A, Elhai J (2004) Heterocyst Metabolism and Development. In: Bryant D, editor, The Molecular Biology of Cyanobacteria, vol. 1 of Advances in Photosynthesis and Respiration, pp. 769–823, Springer Netherlands.

Wu YC, Rasmussen MD, Bansal MS, Kellis M (2013) TreeFix: statistically informed gene tree error correction using species trees. Syst Biol 62: 110–120.

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.

Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23: 212–226.

Yu Y, Barnett RM, Nakhleh L (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Syst Biol 62: 738–51.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, et al. (2008) Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. Science 322: 1110–2.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, et al. (2006) KaKs Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4: 259–63.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res 16: 1099–108.

Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17: 821–8.

Appendix A

Supplementary Figures and Tables

A.1 Chapter 2 Phylogenetic Reconciliation with Transfers

Number of roots	Cycle-free solutions	Event counts	Source & target	Stats	heatmap
≥ 1	0	NA	NA	Discard tree	
≥ 1	1	NA	NA	OK	OK
≥ 1	≥ 2	Same	Same	OK	OK
		Same	Different	OK	Not
		Different	Different	Discard tree	
≥ 2	1 per root	For all roots:		For all roots:	
		Same	Same	OK	OK
		Same	Not	OK	No
		Not	Not	Discard tree	
≥ 2	At least one root has ≥ 1	For all roots and solns:		For all roots and solns:	
		Same	Same	OK	OK
		Same	Not	OK	No
		Not	Not	Discard tree	

Table A1: Protocol for handling the degeneracy in the empirical analyses.

Short name	Long name
Synechocys	Synechocystis sp. PCC 6803
Crocosphae	Crocosphaera watsonni WH 8501
Nostoc	Nostoc sp. PCC 7120
Anabaena	Anabaena variabilis ATCC29413
Trichodesm	Trichodesmium erythraeum IMS101
1Prochloro	Prochlorococcus marinus CCMP1375
2Prochloro	Prochlorococcus marinus CCMP1986
3Prochloro	Prochlorococcus marinus MIT9313
Synechococ	Synechococcus sp. WH8102
Thermosyne	Thermosynechococcus elongatus BP-1
Gloeobacte	Gloeobacter violaceus PCC7421

Table A2: Full names of cyanobacterial species used in the case study.

Short name	Long name
Ylip	Yarrowia lipolytica
Dhan	Debaryomyces hansenii
Calb	Candida albicans
Sklu	Saccharomyces kluyveri
Kwal	Kluyveromyces waltii
Klac	Kluyveromyces lactis
Agos	Ashbya gossypii
Cgla	Candida glabrata
Scas	Saccharomyces castellii
Sbay	Saccharomyces bayanus
Skud	Saccharomyces kudriavzevii
Smik	Saccharomyces mikatae
Spar	Saccharomyces paradoxus
Scer	Saccharomyces cerevisiae

Table A3: Full names of yeast species used in the case study.

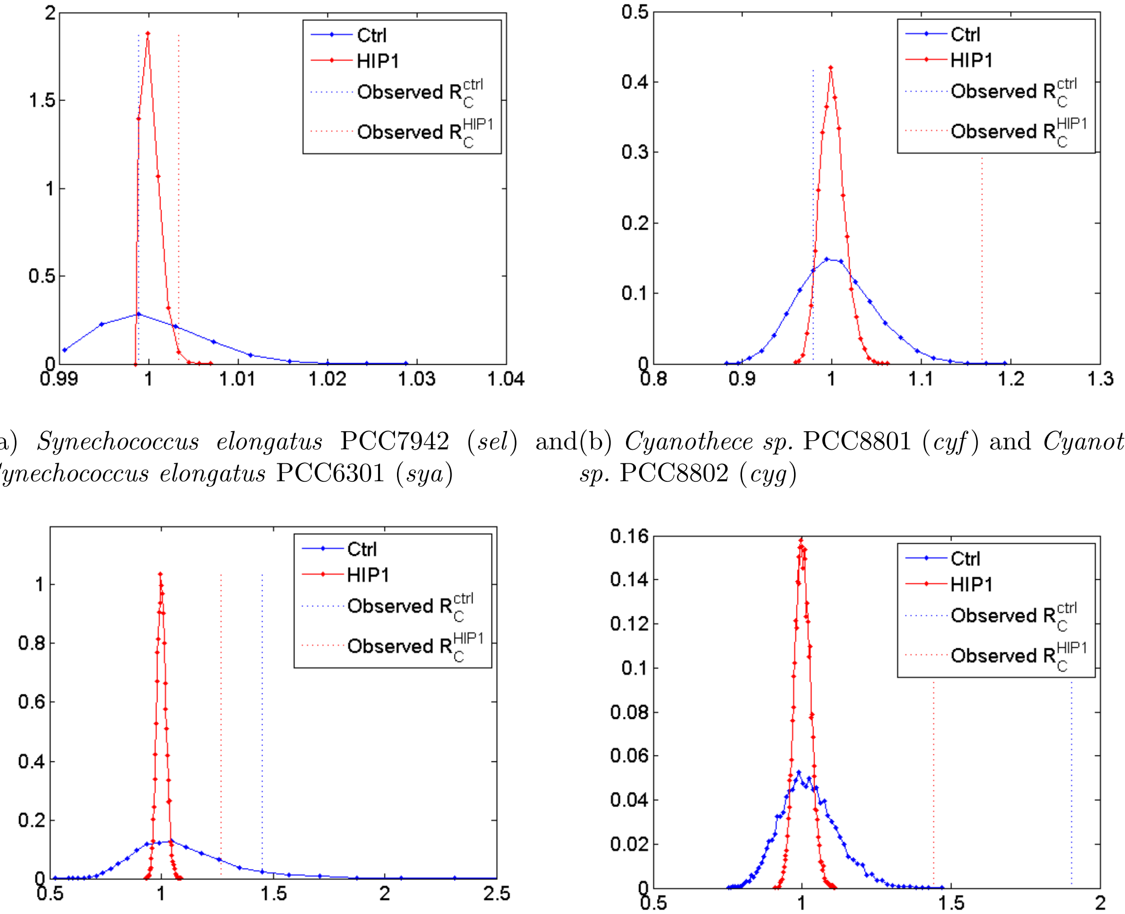
A.2 Chapter 3 Highly Iterated Palindrome-1 (HIP1) Motifs

	sya	sel	syq	ava	ana	naz	npu	syr	cyd	ama	syb	syp	mae	cya	cye	uca	cyb	cyf	cyg	syg	syh
AGGGCCCT	5	5	9	2	2	0	11	10	21	19	3	39	3	1	7	0	6	0	1	22	27
GAGGCCTC	14	13	22	3	2	3	1	5	8	22	6	4	11	0	7	1	0	1	1	19	29
GGAGCTCC	14	14	11	0	1	2	3	69	16	10	8	11	4	0	1	4	0	10	12	40	64
GGGATCCC[1]	36	36	57	0	1	8	1	12	14	12	33	37	92	4	0	1	0	1	1	4099	3401
AGGCGCCT	6	6	20	0	1	3	0	2	6	12	5	1	1	0	1	2	1	0	0	13	31
GAGCGCTC	37	36	24	22	22	46	40	25	29	57	15	9	56	38	52	1	0	0	0	86	88
GGACGTCC	19	19	11	0	0	2	0	12	14	37	16	4	1	1	0	8	0	0	0	7	14
GGCATGCC	42	42	23	0	0	3	1	3	48	62	6	27	0	0	2	4	0	1	1	43	65
AGCGCGCT	21	20	20	11	12	7	5	6	19	21	16	0	25	7	5	2	6	2	2	13	27
GACGCGTC	7	7	2	14	10	0	9	1	8	20	8	1	6	0	0	0	1	2	1	0	2
GCAGCTGC	175	171	44	82	76	67	138	32	36	158	9	3	60	0	0	24	10	2	4	98	45
GCGATCGC[2]	7277	7323	3659	5227	5253	1102	7128	2907	3395	2147	5083	3160	1821	2252	647	37	2390	2956	2977	57	67
AGCCGGCT	40	37	33	21	29	6	22	13	16	15	8	18	25	89	134	3	19	4	7	115	153
GACCGGTC	23	23	24	2	1	4	7	2	7	22	3	10	25	12	15	0	4	15	10	10	10
GCACGTGC	7	7	12	0	0	19	8	4	6	6	9	2	0	4	7	4	1	0	0	13	15
GCCATGGC	107	106	138	3	1	51	3	26	182	188	151	407	134	22	37	11	58	77	81	144	171
ACGGCCGT	25	25	10	10	9	3	5	6	30	18	4	23	27	0	8	2	23	62	58	2	19
CAGGCCTG	32	32	53	5	3	5	1	12	79	50	23	34	14	0	9	3	2	3	3	49	99
CGAGCTCG	30	31	10	1	0	2	3	12	5	11	9	7	0	0	1	2	0	8	5	7	21
CGGATCCG	54	54	34	2	2	0	0	5	10	58	44	19	27	23	2	3	0	1	0	109	121
ACGCGCGT	6	6	1	6	6	8	11	1	1	9	3	0	6	2	6	1	2	7	7	5	7
CAGCGCTG	288	289	117	73	75	5	145	15	105	161	31	37	33	15	22	3	0	2	2	173	228
CGACGTCG	18	18	8	0	0	1	0	6	5	30	9	1	8	0	1	2	1	1	0	11	16
CGCATGCG	18	18	9	0	0	1	0	0	16	15	2	14	0	0	3	3	0	0	0	20	26
ACCGCGGT	12	12	21	0	0	6	2	5	15	11	3	11	21	1	0	0	0	0	0	21	13
CACGCGTG	4	4	2	3	2	0	2	0	4	5	2	1	3	0	0	0	0	0	0	2	0
CCAGCTGG	53	53	20	25	11	17	30	22	87	123	3	10	13	0	0	6	4	0	0	79	42
CCGATCGG	82	82	8	2	4	1	43	7	519	86	5	9	57	87	13	0	6	5	6	28	22
ACCCGGGT	10	10	17	1	1	5	2	19	11	26	10	15	44	14	44	2	5	21	24	15	15
CACCGGTG	51	51	94	37	26	22	19	6	10	33	36	45	82	41	63	1	30	5	6	105	53
CCACGTGG	19	19	27	2	1	8	11	4	15	15	22	22	7	6	4	3	2	0	0	50	55
CCCATGGG	36	36	61	1	1	23	0	28	87	106	65	114	29	0	0	5	11	55	55	60	45
TGGGCCCA	14	14	22	7	9	4	37	19	68	34	21	80	14	21	43	7	22	0	0	68	68
GTGGCCAC	38	38	141	24	3	24	19	24	103	72	54	130	62	31	40	0	47	23	21	143	138
GGTGCACC	27	27	39	0	1	17	5	13	21	38	18	19	3	1	3	9	0	19	15	25	52
GGGTACCC	14	14	76	4	3	4	7	8	31	52	34	49	9	0	14	2	3	0	0	30	20
TGGCGCCA	37	37	93	1	2	4	2	29	21	61	17	30	17	0	0	3	1	1	0	57	85
GTGCGCAC	22	21	66	0	1	0	0	48	11	14	21	5	2	0	0	0	2	0	0	79	79
GGTCGACC	30	31	4	1	0	1	4	8	6	34	2	5	3	6	4	0	8	1	2	9	10
GGCTAGCC	63	64	41	36	34	1	51	14	73	110	20	66	0	16	3	2	2	2	3	39	62
TGCGCGCA	9	10	14	34	29	4	5	3	8	22	5	2	18	1	4	1	5	4	4	7	14
GTCGCGAC	51	51	18	1	0	4	12	39	6	44	12	8	23	0	2	1	0	1	1	28	14
GCTGCAGC	274	273	69	2	3	33	4	73	244	285	31	12	22	4	0	27	58	6	8	309	447
GCGTACGC	4	4	20	5	3	3	0	2	4	8	4	0	1	0	3	2	5	4	3	2	1
TGCCGGCA	65	62	87	40	40	13	42	11	39	28	11	69	92	98	172	1	30	1	2	254	262
GTCCGGAC	6	6	12	11	6	4	1	5	9	7	12	6	10	12	18	1	2	7	6	5	3
GCTCGAGC	69	68	34	0	0	9	3	10	31	18	5	1	3	3	22	2	1	4	4	28	29
GCCTAGGC	50	50	35	1	3	3	8	12	8	45	38	43	14	10	25	2	3	6	4	35	41
TCGGCCGA	36	36	3	6	6	4	8	4	29	22	4	11	80	0	21	0	23	8	8	6	9
CTGGCCAG	89	90	93	39	39	27	33	30	438	116	52	96	61	33	43	2	30	32	42	238	292
CGTGCACG	10	10	8	0	0	1	0	2	4	3	3	0	0	0	0	2	0	4	5	3	7
CGGTACCG	48	47	39	7	8	1	15	6	20	52	52	26	9	11	19	0	11	4	5	33	25
TCGCGCGA	75	74	22	24	30	3	36	4	9	14	11	0	12	22	22	0	2	31	32	6	8
CTGCGCAG	97	97	68	2	0	4	0	8	25	36	14	5	3	1	1	1	0	1	1	106	112
CGTCGACG	28	27	3	1	0	2	3	6	3	26	2	4	1	6	4	4	7	2	1	7	17
CGCTAGCG	77	77	32	27	34	3	63	4	18	57	10	10	3	9	4	0	0	5	5	35	38
TCCGCGGA	9	9	6	0	0	1	5	9	16	13	4	8	12	3	0	5	2	2	2	15	20
CTCGCGAG	48	48	13	0	0	2	8	11	9	32	15	2	34	1	10	0	3	4	8	27	32
CCTGCAGG	125	126	86	2	0	6	1	29	373	137	8	40	0	1	0	15	10	3	3	311	345
CCGTACGG	19	19	25	7	0	5	0	2	34	23	7	6	4	0	5	0	9	13	13	10	3
TCCCGGGA	6	6	10	0	0	0	3	16	5	8	26	35	49	10	15	1	0	6	4	24	24
CTCCGGAG	9	9	11	15	12	4	10	37	16	13	41	45	19	19	16	1	3	7	9	34	35
CCTCGAGG	42	42	48	0	0	4	2	1	9	9	7	10	10	1	32	2	1	0	0	29	68
CCCTAGGG	43	44	166	0	0	73	4	25	25	115	122	141	124	62	47	2	63	118	127	53	55
Total Control	2825	2813	2346	625	565	596	914	882	3145	2966	1250	1899	1528	749	1036	196	545	604	625	3471	4005

Table A4: Count of all 64 8-mer palindromes with 75% GC content in the 20 HIP1-rich genomes. [1]Yellowstone HIP1 variant. [2]Canoncial HIP1 motif.

	G_a	G_b	K_S	n_a	n_b	n_{ab}	S	C	$\hat{p}_u^1$	95% C.I.^{2}	p-value^3
HIP1	sya	sel	0.00	7356	7399	7348	0.992	0.993			1.06×10^{-13}
	cyf	cyg	0.02	2947	2972	2825	0.913	0.951			2.34×10^{-39}
	ana	ava	0.18	4814	4837	4211	0.774	0.871			0
	syh	syg	0.59	3181	3767	2409	0.531	0.640			0
Control	sya	sel	0.00	2825	2808	2796	0.986	0.996	0.998	0.996-1.000	
	cyf	cyg	0.02	602	623	535	0.775	0.859	0.881	0.855-0.907	
	ana	ava	0.18	532	576	236	0.271	0.410	0.444	0.402-0.486	
	syh	syg	0.59	3710	3190	1015	0.172	0.318	0.332	0.317-0.347	

Table A5: Genome-wide Positional conservation of HIP1 and control motifs. C scores are calculated as $C = n_{ab}/n_b$. [1] $\hat{p}_u$, the estimate of C calculated based on the upper bound of the likelihoood interval. [2,3] 95% confidence intervals and p-values based on the binomial distribution estimated using the $\hat{p}_u$.



(a) *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*)

(b) *Cyanothece sp.* PCC8801 (*cyf*) and *Cyanothece sp.* PCC8802 (*cyg*)

(c) *Nostoc sp.* PCC 7120 (*ana*) and *Anabaena variabilis* ATCC29413 (*ava*)

(d) *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*)

Figure A1: The distribution of permutated $\mathcal{R}_C$ for HIP1 and control motif in the 4 genome pairs. (a) *Synechococcus elongatus* PCC7942 (*sel*) and *Synechococcus elongatus* PCC6301 (*sya*), (b) *Cyanothece sp.* PCC8801 (*cyf*) and *Cyanothece sp.* PCC8802 (*cyg*), (c) *Nostoc sp.* PCC 7120 (*ana*) and *Anabaena variabilis* ATCC29413 (*ava*), (d) *Synechococcus sp.* JA-3-3Ab (*syh*) and *Synechococcus sp.* JA-2-3B'a(2-13) (*syg*).

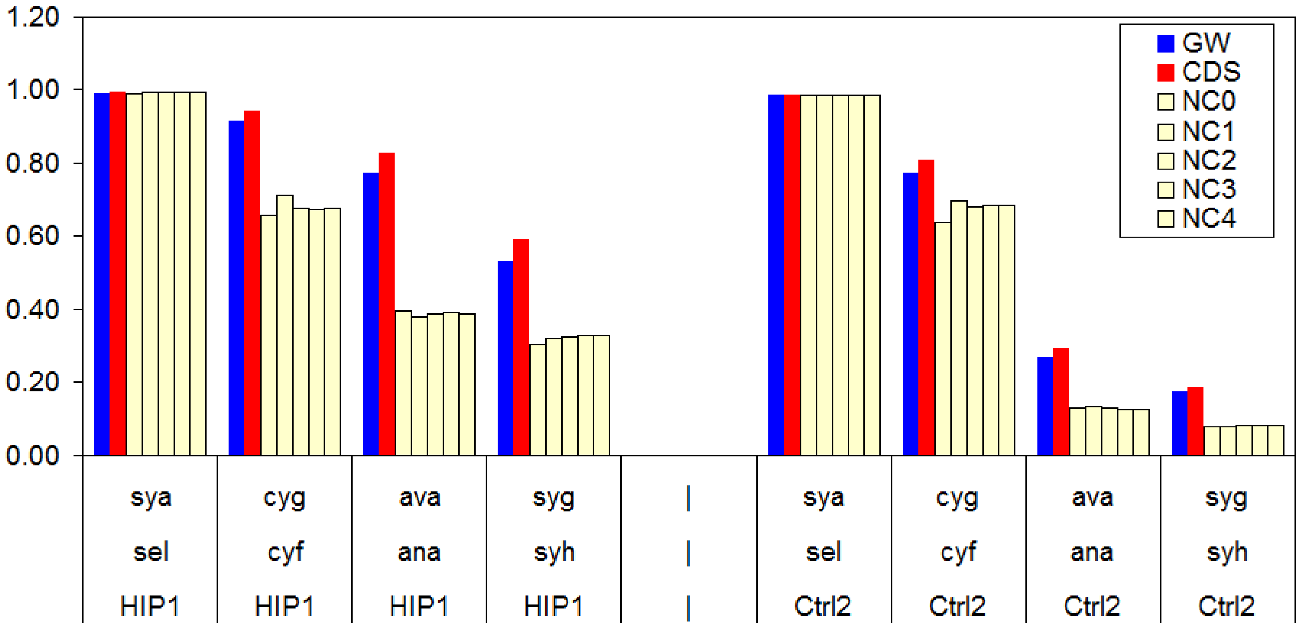


Figure A2: The motif conservation (S scores) calculated using different criteria for non-coding regions. GW: genome wide; CDS: coding regions; NC0: full length inter-ORF regions; NC1: inter-ORF regions less than 500 bp; NC1: inter-ORF regions less than 1000 bp; NC1: inter-ORF regions less than 2000 bp; NC1: inter-gene regions less than 3000 bp. For NC1-4, the 10 bp immediately next to annotated protein or RNA genes were excluded.

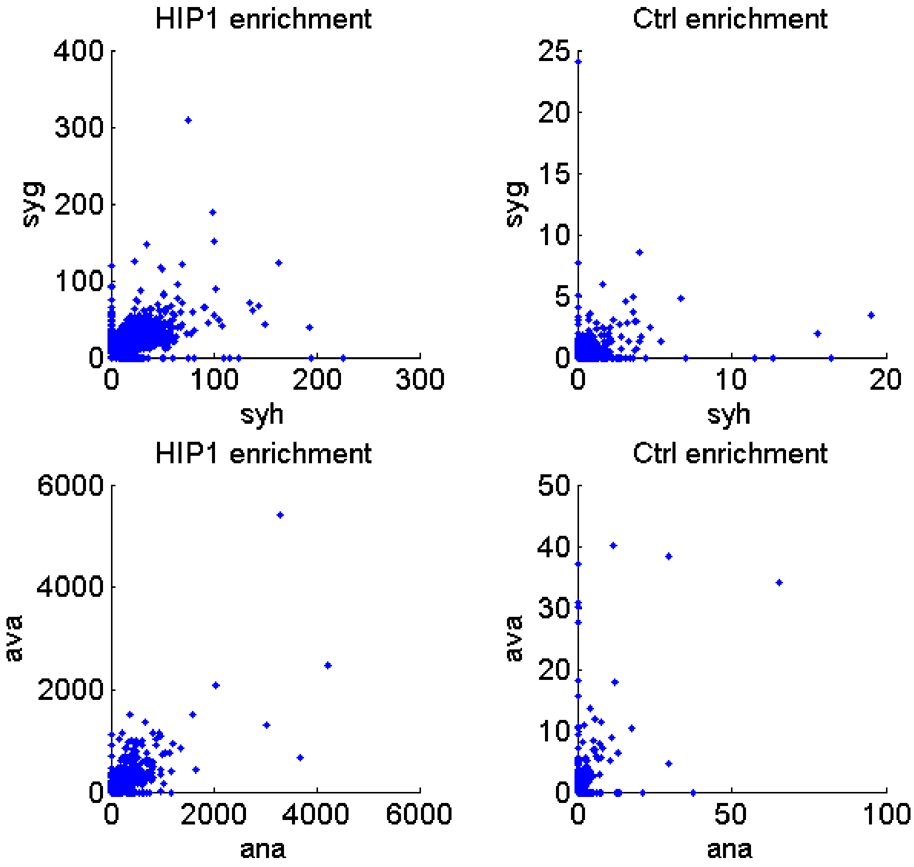


Figure A3: Full range scatter plots showing the HIP1 enrichment in orthologous genes. Each dot represents an ortholog pair.

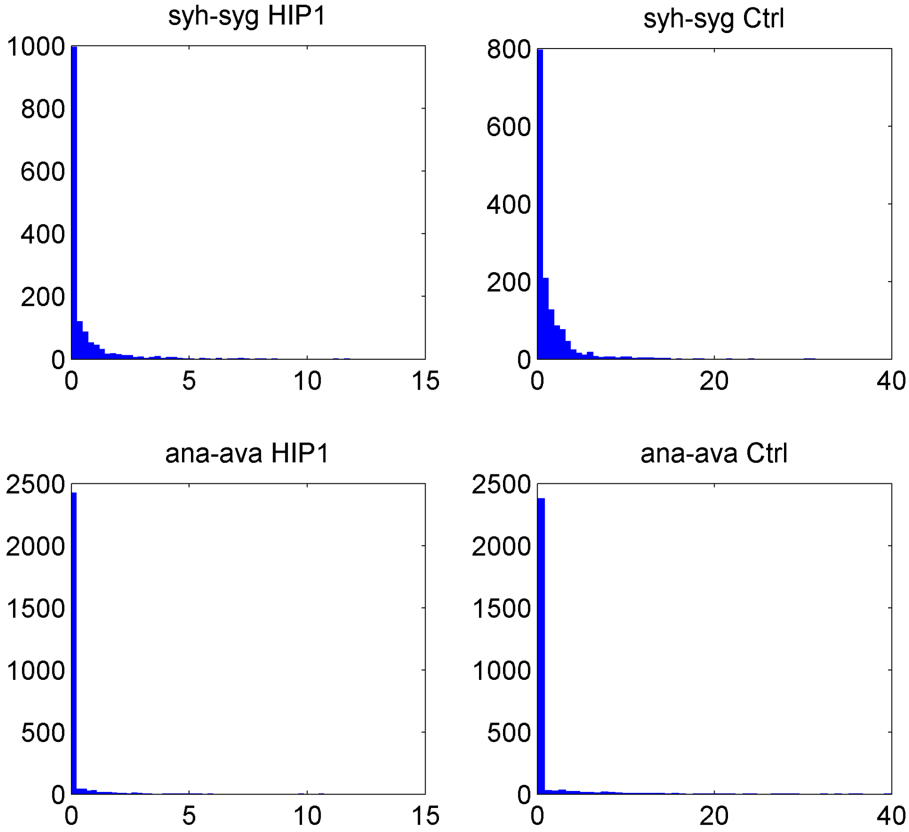


Figure A4: Histograms (full range view) of d_s[HIP1] and d_s[Ctrl] based on motif frequency for the genome pairs *Nostoc sp.* PCC 7120 (*ana*)-*Anabaena variabilis* ATCC29413 (*ava*) and *Synechococcus sp.* JA-3-3Ab (*syh*)-*Synechococcus sp.* JA-2-3B'a(2-13) (*syg*).

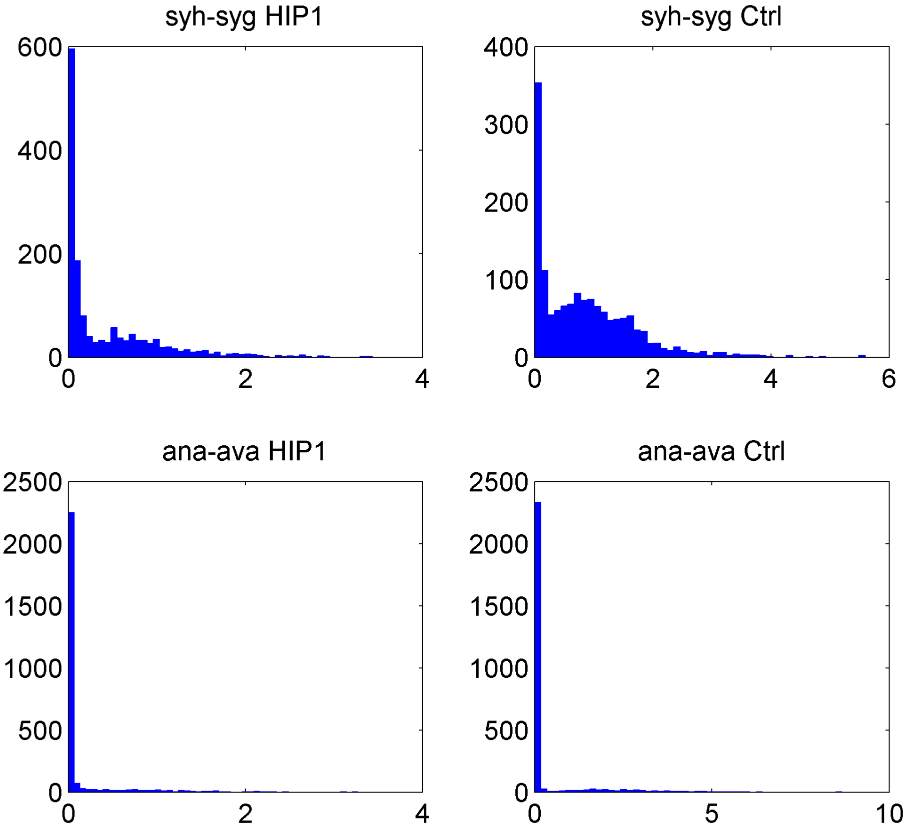


Figure A5: Histograms (ull range view) of d_a[HIP1] and d_a[Ctrl] based on motif frequency for the genome pairs *Nostoc sp.* PCC 7120 (*ana*)-*Anabaena variabilis* ATCC29413 (*ava*) and *Synechococcus sp.* JA-3-3Ab (*syh*)-*Synechococcus sp.* JA-2-3B'a(2-13) (*syg*).

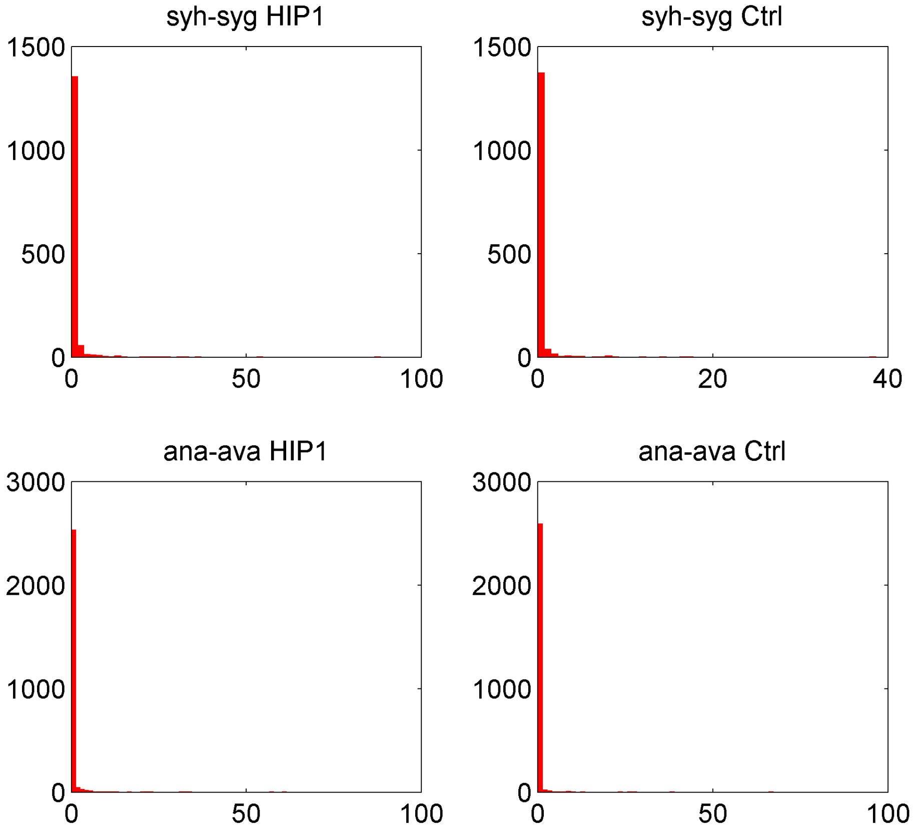


Figure A6: Histograms (ull range view) of d_s[HIP1] and d_s[Ctrl] based on motif enrichment for the genome pairs *Nostoc sp.* PCC 7120 (*ana*)-*Anabaena variabilis* ATCC29413 (*ava*) and *Synechococcus sp.* JA-3-3Ab (*syh*)-*Synechococcus sp.* JA-2-3B'a(2-13) (*syg*).

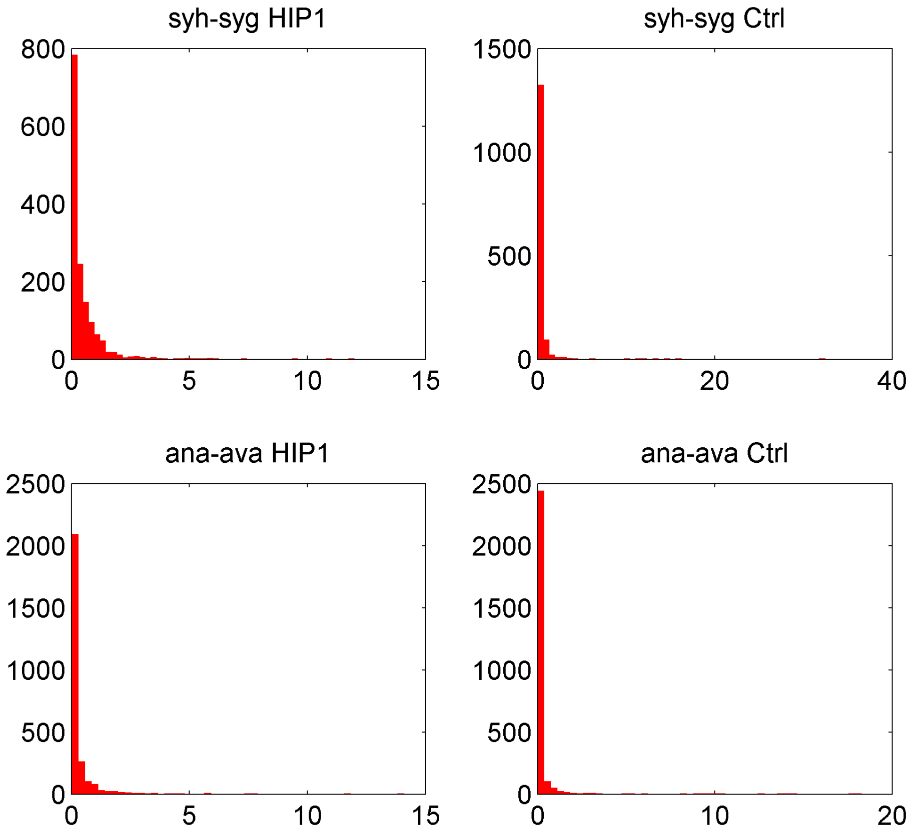


Figure A7: Histograms (ull range view) of d_a[HIP1] and d_a[Ctrl] based on motif enrichment for the genome pairs *Nostoc sp.* PCC 7120 (*ana*)-*Anabaena variabilis* ATCC29413 (*ava*) and *Synechococcus sp.* JA-3-3Ab (*syh*)-*Synechococcus sp.* JA-2-3B'a(2-13) (*syg*).

A.3 Chapter 5 Intra-Genome Variation of HIP1 Motifs

Rank	Location	Strand	Length	nORF[1]	nHIP1[2]	Mean HIP1 position
1	1824612-1825011	+	400	1	2	0.94
2	2485537-2486679	-	1143	1	3	0.93
3	2340127-2340927	+	801	1	2	0.92
4	2499675-2500957	-	1283	2	4	0.89
5	1716762-1717589	+	828	1	2	0.89
6	1173872-1174990	-	1119	1	3	0.89
7	2032100-2032718	-	619	1	2	0.88
8	1137758-1138034	-	277	1	2	0.88
9	294788-296809	+	2022	1	2	0.88
10	473353-474207	-	855	1	2	0.88
11	1990819-1991436	-	618	1	2	0.87
12	1911935-1913842	-	1908	1	2	0.85
13	2492192-2493019	+	828	1	2	0.85
14	585681-587529	+	1849	1	4	0.85
15	1497152-1498624	+	1473	2	2	0.84
16	621248-622240	+	993	1	2	0.84
17	2245016-2245713	+	698	1	3	0.84
18	2018328-2019540	+	1213	1	3	0.84
19	959818-960692	+	875	1	3	0.83
20	2445310-2447361	+	2052	2	4	0.83
21	1852455-1853695	+	1241	1	3	0.82
22	974053-975417	+	1365	2	4	0.82
23	1157386-1157996	-	611	1	2	0.82
24	2619075-2620994	-	1920	3	4	0.82
25	466721-469366	-	2646	2	4	0.82
26	2044382-2047420	-	3039	3	4	0.81
27	1377140-1379998	+	2859	4	3	0.81
28	1518990-1521363	+	2374	3	3	0.81
29	226210-227342	-	1133	1	7	0.81
30	1311801-1313283	+	1483	1	4	0.81
31	1484524-1485033	+	510	1	2	0.80
32	1583328-1586106	+	2779	2	6	0.80
33	590846-592822	-	1977	2	2	0.80
34	1780025-1781245	+	1221	3	3	0.80
35	671621-672657	-	1037	1	3	0.80
36	2112719-2114132	-	1414	1	2	0.80
37	1276672-1277247	-	576	1	2	0.80
38	966374-967577	-	1204	1	2	0.80
39	595366-596808	+	1443	2	4	0.80
40	1172694-1174011	+	1318	1	3	0.79
41	483622-484524	-	903	2	3	0.79
42	1687415-1689171	-	1757	1	3	0.79
43	1973439-1984570	-	11132	7	12	0.78
44	2085761-2090061	-	4301	2	6	0.78
45	1118636-1120443	+	1808	2	4	0.78
46	2570956-2573066	+	2111	1	3	0.78
47	1431081-1432779	-	1699	2	3	0.77
48	1671435-1672079	+	645	1	2	0.77
49	763606-764877	-	1272	1	4	0.77
50	1463651-1464861	-	1211	2	5	0.77

Table A6: The top 50 transcripts with highest mean HIP1 motif position. [1] Number of annotated ORFs within the transcript. [2] Number of HIP1 motifs within the transcript.

Trans.[1]	Location	Strand	Length	Locus tag	Gene annotation
1	1824648-1824935	+	288	Synpcc7942_1757	hypothetical protein
2	2485595-2486650	-	1056	Synpcc7942_2414	hypothetical protein
3	2340153-2340920	+	768	Synpcc7942_2273	hypothetical protein
4	2499798-2500244	-	447	Synpcc7942_2428	biopolymer transport ExbD like protein
4	2500278-2500928	-	651	Synpcc7942_2429	biopolymer transport ExbB like protein
5	1716788-1717504	+	717	Synpcc7942_1649	rubrerythrin
6	1173913-1174890	-	978	Synpcc7942_1149	dTDP-glucose 46-dehydratase
7	2032190-2032696	-	507	Synpcc7942_1960	hypothetical protein
8	1137860-1138015	-	156	Synpcc7942_1120	hypothetical protein
9	294872-296713	+	1842	Synpcc7942_0297	FtsH peptidase
10	473510-474172	-	663	Synpcc7942_0487	thylakoidal processing peptidase
11	1991020-1991412	-	393	Synpcc7942_1915	chorismate mutase
12	1911978-1913807	-	1830	Synpcc7942_1846	hypothetical protein
13	2492215-2492949	+	735	Synpcc7942_2420	serine O-acetyltransferase
14	585701-587473	+	1773	Synpcc7942_0598	peptidoglycan-binding LysM
15	1497217-1498290	+	1074	Synpcc7942_1443	fructose-1,6-bisphosphate aldolase
16	621277-622137	+	861	Synpcc7942_0628	spermidine synthase
17	2245037-2245525	+	489	Synpcc7942_2163	hypothetical protein
18	2018389-2019417	+	1029	Synpcc7942_1944	pyruvate dehydrogenase (lipoamide)
19	959837-960685	+	849	Synpcc7942_0951	nicotinate-nucleotide pyrophosphorylase
20	2445387-2446568	+	1182	Synpcc7942_2378	cell division protein FtsZ
20	2446568-2447359	+	792	Synpcc7942_2379	phosphomethylpyrimidine kinase
21	1852529-1853311	+	783	Synpcc7942_1784	RNA polymerase sigma factor SigF
22	974099-975061	+	963	Synpcc7942_0967	porphobilinogen deaminase
22	975143-975358	+	216	Synpcc7942_0968	hypothetical protein
23	1157413-1157931	-	519	Synpcc7942_1135	cation transporter
24	2619093-2619626	-	534	Synpcc7942_2536	heat shock protein DnaJ-like protein
24	2619632-2620231	-	600	Synpcc7942_2537	ATP-dependent Clp protease proteolytic subunit
24	2620274-2620960	-	687	Synpcc7942_2538	ATP-dependent Clp protease-like protein
25	467004-468773	-	1770	Synpcc7942_0482	peptidoglycan glycosyltransferase
25	468776-469219	-	444	Synpcc7942_0483	hypothetical protein
26	2044388-2045224	-	837	Synpcc7942_1974	condensin subunit ScpA
26	2045410-2045763	-	354	Synpcc7942_1975	hypothetical protein
26	2045793-2047397	-	1605	Synpcc7942_1976	NAD(P)H-quinone oxidoreductase subunit 4
27	1377217-1378335	+	1119	Synpcc7942_1343	NADH dehydrogenase subunit H
27	1378378-1378986	+	609	Synpcc7942_1344	NADH dehydrogenase subunit I
27	1378990-1379595	+	606	Synpcc7942_1345	NADH dehydrogenase subunit J
27	1379614-1379925	+	312	Synpcc7942_1346	NADH dehydrogenase subunit K
28	1518994-1519449	+	456	Synpcc7942_1465	BadM/Rrf2 family transcriptional regulator
28	1519664-1520668	+	1005	Synpcc7942_1466	cysteine synthase
29	226263-227309	-	1047	Synpcc7942_0230	hypothetical protein
30	1311841-1313181	+	1341	Synpcc7942_1289	putative modulator of DNA gyrase
31	1484549-1484986	+	438	Synpcc7942_1431	peptidylprolyl isomerase
32	1583407-1585200	+	1794	Synpcc7942_1525	GTP-binding protein TypA
32	1585287-1586078	+	792	Synpcc7942_1526	hypothetical protein
33	591459-592751	-	1293	Synpcc7942_0603	glucose-1-phosphate adenylyltransferase

34	1780048-1780278	+	231	Synpcc7942_1710	DNA-directed RNA polymerase subunit omega
34	1780275-1780847	+	573	Synpcc7942_1711	hypothetical protein
34	1780853-1781218	+	366	Synpcc7942_1712	hypothetical protein
35	672082-672462	-	381	Synpcc7942_0677	PadR family transcriptional regulator
36	2112977-2113990	-	1014	Synpcc7942_2044	hypothetical protein
37	1276940-1277221	-	282	Synpcc7942_1253	hypothetical protein
38	966530-967561	-	1032	Synpcc7942_0959	GTPase ObgE
39	595381-595707	+	327	Synpcc7942_0607	hypothetical protein
39	595756-596298	+	543	Synpcc7942_0608	hypothetical protein
40	1172695-1173912	+	1218	Synpcc7942_1148	metal dependent phosphohydrolase
41	483634-484323	-	690	Synpcc7942_0496	hypothetical protein
41	484328-484504	-	177	Synpcc7942_0497	hypothetical protein
42	1687584-1689152	-	1569	Synpcc7942_1621	Elongator protein 3/MiaB/NifB
43	1973458-1975839	-	2382	Synpcc7942_1901	putative glycosyltransferase
43	1975932-1976978	-	1047	Synpcc7942_1902	putative glycosyltransferase
43	1976983-1978113	-	1131	Synpcc7942_1903	hypothetical protein
43	1978129-1979454	-	1326	Synpcc7942_1904	hemolysin secretion protein-like protein
43	1979496-1982498	-	3003	Synpcc7942_1905	cyclic nucleotide-binding domain-containing protein
43	1982513-1983253	-	741	Synpcc7942_1906	hypothetical protein
43	1983417-1984493	-	1077	Synpcc7942_1907	magnesium-protoporphyrin IX monomethyl ester cyclase
44	2085893-2086762	-	870	Synpcc7942_2019	hypothetical protein
44	2086845-2089937	-	3093	Synpcc7942_2020	translation initiation factor IF-2
45	1118661-1119818	+	1158	Synpcc7942_1101	hypothetical protein
45	1119830-1120258	+	429	Synpcc7942_1102	hypothetical protein
46	2571012-2572949	+	1938	Synpcc7942_2491	DNA gyrase subunit B
47	1431285-1431506	-	222	Synpcc7942_1385	hypothetical protein
47	1431627-1432760	-	1134	Synpcc7942_1386	hypothetical protein
48	1671451-1672011	+	561	Synpcc7942_1606	Beta-Ig-H3/fasciclin
49	763640-764854	-	1215	Synpcc7942_0771	hypothetical protein
50	1464368-1464817	-	450	Synpcc7942_1412	hypothetical protein

Table A7: Annotated ORFs associated to the top 50 transcripts with highest mean HIP1 position. [1] Transcript rank as shown in Table A6.