DEVELOPMENT OF ALGORITHMS FOR METAGENOMICS AND APPLICATIONS TO THE STUDY OF EVOLUTIONARY PROCESSES THAT MAINTAIN MICROBIAL BIODIVERSITY

A Dissertation
Presented to
The Academic Faculty

by

Chengwei Luo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
December 2012

DEVELOPMENT OF ALGORITHMS FOR METAGENOMICS AND APPLICATIONS TO THE STUDY OF EVOLUTIONARY PROCESSES THAT MAINTAIN MICROBIAL BIODIVERSITY

Approved by:

Dr. Konstantinos T. Konstantinidis,
Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biomedical Engineering and Computer Science and Engineering
Georgia Institute of Technology

Dr. James M. Tiedje
Center for Microbial Ecology
Michigan State University

Dr. King I. Jordan
School of Biology
Georgia Institute of Technology

Date Approved: December 11, 2012

To my grandparents, and my wife, Jaclyn

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the generous help and advice of several individuals, and therefore I would like to thank them for their support. First of all, I would like to thank my thesis advisor, Dr. Kostas Konstantinidis for his guidance and support in exploring unknown realms, which were essential for my growth as a scientist. I also want to thank him for providing me with opportunities to get involved in various academic activities including conferences, proposal writing, *etc*. These experiences will be invaluable for my future career. Last, I have to thank him for being a great advisor and friend; his support and consideration had helped me through many difficult personal times.

I would also like to thank Dr. Mark Borodovsky, for introducing me to the Bioinformatics Ph.D. program and the Konstantinidis Lab. I am truly grateful to his tremendous help during my study in Georgia Tech.

During the three years of graduate study, my family has supported me unconditionally. I want to thank my wife, Jaclyn, for being always considerate, patient, and supportive along my pursuit of dreams. Without her encouragement, I would never have started this wonderful journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAI	Average Amino acid Identity
ANI	Average Nucleotide Identity
BH	Best Hit
B-H	Benjamini-Hochberg
dLGT	direct network of Lateral Gene Transfer
DNA	Deoxyribonucleic acid
EQDA	Embedded Quartet Decomposition Analysis
HGT	Horizontal Gene Transfer
HMP	Human Microbiome Project
JGI	Joint Genome Institute
KDE	Kernel Density Estimator
LCA	Lowest Common Ancestor
MV	Membrane Vesicle
MSA	Multiple-Sequence Alignment
NGS	Next-Generation Sequencing
NR	Non-Redundant
OTU	Operational Taxonomic Unit
PCoA	Principle Component Analysis
PE	Paired-End
RBM	Reciprocal Best Match
rRNA	ribosomal Ribonucleic Acid
SNP	Single Nucleotide Polymorphism

SSE	Summed Squared Error
TIGR	The Institute for Genome Research
WGS	Whole Genome Shotgun
WCS	Whole Community Shotgun

SUMMARY

Understanding microbial evolution lies at the heart of microbiology and environmental sciences. Numerous studies have been dedicated to elucidating the underlying mechanisms that create microbial genetic diversity and adaptation. However, due to technical limitations such as the high level of uncultured cells in almost every natural habitat, most of current knowledge is primarily based on axenic cultures grown under laboratory conditions, which typically do not simulate well the natural environment. How well the knowledge from isolates translates to *in-situ* processes and natural microbial communities remains essentially speculative.

The recent development of culture-independent genomic techniques (aka metagenomics) provides possibilities to bypass some of these limitations and provide new insights into microbial evolution *in-situ*. To date, most of metagenomic studies have been focused on a few reduced-diversity model communities, e.g., acid mine drainage. Highly complex communities such as those of soil and sediment habitats remain comparatively less understood. Furthermore, a great power of metagenomics, which has not been fully capitalized yet, is the ability to follow the evolution of natural microbial communities over time and environmental perturbations, i.e., times-series metagenomics. Although the recent developments in DNA sequencing technologies have enabled (inexpensive) time-series studies, the bioinformatics approaches to analyze the resulting data have clearly fallen behind. Taken together, to scale up metagenomics for complex community studies, three major challenges remain: 1) the difficulty to process and analyze massive short read sequencing data, often at the terabyte level; 2) the difficulty to

effectively assemble genomes from complex metagenomes; and 3) the lack of methods for tracking genotypes and mutational events such as horizontal gene transfer (HGT) through time. Therefore, developing efficient bioinformatics approaches to address these challenges represents an important and timely issue.

This thesis aimed to develop novel bioinformatics pipelines and algorithms for high performance computing, and, subsequently, apply these tools to natural microbial communities to generate quantitative insights into the relative importance of the molecular mechanisms creating or maintaining microbial diversity. The tools are not specific to a particular habitat or group of organisms and thus, can be broadly used to advance our understanding of microbial evolution in different settings.

In particular, the comparative whole-genome analysis of 24 *Escherichia* isolates form various habitats, including human and non-human associated habitats such as freshwater ecosystems and beaches, showed that organisms with more similar ecologies tend to exchange more genes, which has important implications for the prokaryotic species concept. To more directly test these findings from isolates and quantify the patterns of genetic exchange among co-occurring populations, three years of time-series metagenomics data from planktonic samples from Lake Lanier (Atlanta, GA) were analyzed. For this, it was first important to develop bioinformatics algorithms to robustly assemble population genomes from complex community metagenomes, identify the phylogenetic affiliation of assembled genome and contig sequences, and detect horizontal gene transfer among these sequences. Using these novel algorithms, *in situ* bacterial lineage evolution was quantitatively assessed, especially with respect to whether or not ecologically distinct lineages evolve according to the recently proposed fragmented

speciation model (Retchless and Lawrence, Science 2008). Evidence in support of this model was rarely observed. Instead, it appeared that rampant HGT disseminated ecologically important genes within the population, maintaining intra-population diversity.

By expanding the previous approaches to include methods to assess differential gene abundance and selection pressure between samples, it was possible to quantify how soil microbial communities respond to a decade of warming by 2 ^{0}C, which simulated the predicted effects of climate change. It was found that the heated communities showed significant shifts in composition and predicted metabolism, reflecting the release of additional soil carbon compared to the unheated (control) communities, and these shifts were community-wide as opposed to being attributable to a few taxa. These findings indicated that the microbial communities of temperate grassland soils play important roles in mediating the feedback responses to climate change.

Collectively, the findings presented here advance our understanding of the modes and tempo of microbial community adaptation to environmental perturbations and have important implications for better modeling the microbial diversity on the planet. The bioinformatics algorithms and approaches developed as part of this thesis are expected to facilitate future genomic and metagenomic studies across the fields of microbiology, ecology, evolution and engineering.

CHAPTER 1

An introduction to the mechanisms of microbial genome evolution and the limitations in quantifying their relative importance under natural settings

INTRODUCTION

Microbes represent convenient systems to understand evolution due to their short generation time and advantages related to (easy) manipulation and handling under laboratory settings (1). Numerous microbial studies have been conducted to better understand the molecular mechanisms underlying evolution and their impact on adaptation (2-4). It is now clear that the evolution of microbial genomes is realized by different types of mutational events such as spontaneous single nucleotide substitutions (5, 6), gene duplication, loss, gain, and innovation (7-9), synteny change (10-12), and horizontal gene transfer (HGT; import of genetic material from non-parents) (13-16). These mechanisms frequently act in parallel and while the final outcome and the speed of evolution is controlled by several essential parameters, such as mutation rate, effective population size, HGT rate, genome size, *etc* (17-22). The process of evolution is also deeply intertwined with the ecology/environment (e.g., selection pressure) (11, 23-25) and population/community structure (26, 27). However, most of the studies to date were limited to pure cultures or mixtures of a few species under laboratory conditions that rarely simulate well natural settings (4, 16, 17, 21, 28, 29). For instance, microbes rarely live in isolation in nature, from the deep ocean hydrothermal vents (30) to the human gut (31, 32), and usually a natural microbial community harbors hundreds, if not thousands, of species (33). It is likely that the lessons learned from laboratory incubations do not apply to natural conditions and within highly complex microbial assemblages (34). For instance, Weinberg *et al* found that in marine communities, large non-coding RNAs are abundant and carry out complex biochemical functions, which is rarely observed under laboratory conditions (35). Therefore, it is essential to directly assess the natural

processes in order to obtain a more realistic and accurate understanding of microbial evolution.

The advent of whole genome shotgun (WGS) metagenomics, which do not require cultivation in the laboratory (i.e., culture-independent), provides the means to sidestep several of the previous limitations (29), and has already revolutionized our knowledge of microbial community diversity, function and dynamics (36, 37). Several metagenomics studies have attempted to assess microbial evolution under natural settings using metagenomics and demonstrated that metagenomics is a powerful tool for such purposes (38-40). For example, Allen *et al* compared an environmental population with its corresponding isolate and identified genes that under positive selection *in situ*. Denef *et al* directly assessed the *in-situ* mutation rate for the first time ever for a natural population (41).

However, the communities studied to date were from extreme environments and their complexity was several orders of magnitudes lower than those in the major natural habitats (42) such as oceans, freshwater lakes, and soil. To study the latter communities, a large volume of sequence data is necessary, in excess of 10-20 Gbp per sample for adequate coverage. The recent advancement in sequencing technologies can now deliver tera-bytes of DNA sequences with a relatively low cost (a few thousand dollars). However, the volume and type of data, e.g., short, error prone sequencing reads, has created several new technical challenges. For instance, a typical Illumina HiSeq-2000 single lane yields ~50 Gbp, which makes analysis computationally expensive or even prohibitive for personal computers, even small computer clusters (43). It is therefore critical to design methods that can effectively handle and analyze metagenomic data.

Addressing these challenges would very likely provide unique opportunities to study *in situ* community processes and deepen our understanding of microbial evolution and ecology.

This chapter represents a literature review of the current understanding of microbial evolution and the remaining challenges and knowledge gaps, along with an introduction into the relevant state-of-the-art bioinformatics techniques. The chapter concludes with the outline of this dissertation and the specific scientific questions as well as the computational challenges that I specifically sought to address with my research.

BACKGROUND

Next generation sequencing (NGS) and metagenomics

With the release of the first bacterial genome (*Haemophilus influenzae*), sequenced by the traditional Sanger sequencer in 1995 (44), microbiology entered its genomic era. At that time, sequencing an average bacterial genome usually cost at least a couple hundred thousand U.S. dollars and a significant amount of time. The advent of next-generation sequencing (NGS) technologies around 2005 substantially changed the outlook of microbial genomics. The intense competition in the NGS market, driven primarily by human genetics and cancer research (e.g., the completion of the human genome), resulted in rapid drop in cost per base and fast increase in throughput. For example, the growth of NGS throughput has outpaced the Moore's law; and it is expected that in less than five years, a human genome would be completed for $1,000 in the commercial market, which could translate to $1 bacteria genomes. By providing low cost and ease of use, NGS has clearly revolutionized several aspects of microbiology.

In general, NGS platforms fall into two categories, template amplification-based and single-molecule sequencing (45). The template amplification based methods include some early stage technologies such as Roche 454 pyrosequencing. They usually require cloning and immobilization of a prepared DNA library, and have been widely used in profiling universally conserved housekeeping genes (e.g., those genes used in multi-loci genotyping) such as the 16S rRNA gene as well as lower-throughput metagenomes. The other category does not require library preparation; instead, it employs an initial DNA fragmentation step. In practice, the fragmentation step resembles a random sampling

across the length of the (fragmented) target DNA sequence, and this feature is critical for several types of analyses, including those presented in this thesis. The fragmented DNA molecules are then sequenced, often in pair-end manner, and the resulting reads are usually short (e.g., 75-150 bp for Illumina platforms) but massive in numbers, e.g., a run of Illumina HiSeq-2000 typically generates 200-500 Gbp of sequence data.

Previous studies have shown that the great majority (>99%) of the microorganisms in the natural environment are uncultured and thus, cannot be efficiently studied using conventional isolate-based approaches (46). Due to this limitation, many critical questions in microbiology have not been addressed. For instance, a robust understanding of the rates of genetic exchange among distinct bacterial species under natural conditions and the influence of the ecological settings on the rates remain elusive. It also not clear how the enormous bacterial species diversity is maintained under natural settings in light of high rates of horizontal gene transfer (HGT). For instance, introduction of ecologically advantageous genes into a recipient population via HGT is thought to result in two possible outcomes: either the individuals with these genes outcompete the remaining individuals of the population, decreasing intra-population diversity [population sweeps (47)]; or these genes are sweeping through the population via rampant HGT between the individuals of the population, maintaining intra-population diversity (sexual speciation). Due to the complexity of the ecological niche of a population and the possibility that several HGTs could occur simultaneously, the positive and negative advantages of different HGT events can also cancel each other out, preventing populations sweeps [balancing selection; (48, 49)]. However, these theories primarily originated from experiments with isolates in the laboratory; thus, it is

imperative to obtain the necessary experimental data from natural population to test the theories. Based on the capabilities provided by NGS technologies, metagenomics provides now the means to begin to collect the appropriate data to better understand population emergence and evolution (50). By directly sequencing the microbial genomes from the environment, bypassing isolation in the laboratory, metagenomics can provide new insights into (natural) population diversity, functions and dynamics..

Metagenomics started with the efforts to profile microbial communities by sequencing 16S rRNA gene amplicons, initially proposed by Pace and colleagues (51). However, more strictly speaking, metagenomics should be defined as the WGS-based sequencing of whole microbial communities. Although it is still debatable whether or not 16S rRNA gene amplicon-based approaches should be considered as metagenomics studies, in this thesis, metagenomics will refer only to community WGS data. The first such project was carried out in 2002 by Breitbart *et al* on a marine viral community (52), and soon followed by several milestone studies including the Global Ocean Sampling, the study on acid mine drainage communities, and the Human Microbiome Project (53-55). Several important discoveries have been made by metagenomics studies to date. Surveys of the oceanic communities showed that natural populations are not clonal but encompass higher intra-population diversity that previously anticipated (56, 57). Hehemann *et al* found that lateral transfer of polysaccharides-digesting enzymes from marine bacteria into gut microbiota of Japanese populations played an important role in improving nutrient absorption from seaweeds. Weinberg *et al* found that a few exceptionally large complex non-coding RNAs are abundant in marine bacterial communities, encoding yet-

to-be-determined functions that are highly likely to be responsible for survival in the marine environment (35).

For all metagenomics studies, bioinformatics represents an indispensible component of the analysis part. The most challenging bioinformatics task is probably to assemble individual reads into longer contigs or even draft genomes from complex community metagenomes. For longer-read sequencing (e.g., Roche 454 FLX, PacBio) overlap-based assemblers are generally used, which is also facilitated by the smaller throughput (compared to short-read sequencing) and thus lower requirement for computational resources. For example, Newbler (58) is generally used for Roche 454 reads, and assembling a typical metagenomic sample of moderate species complexity usually requires about 5Gbp of RAM and a few hours of running time. Other assembly software in this genre include Celera, CABOG, *etc* (59). However, for short reads such as Illumina GA II and AB SOLiD, *De Bruijn* graph-based algorithms are more suitable because overlap-based algorithms are not efficient with short reads and the large amount of data produced by the latter sequencers, typically a couple orders of magnitude more data compared to Roche 454 (60). ALLPATHS (61) and Velvet (62), and their derivatives [e.g., metaVelvet, Velvet-SC, meta-IDBA, ALLPATHS-LG (63-66)] represent the current state-of-the-art implementations for short read assembly. Aside from generalized assembling protocols that aim at resolving the assembly for the whole community, specialized approaches that focus on specific genes have also been developed recently. For example, EMIRGE (67) was designed to reconstruct 16S rRNA gene sequences from short read metagenomes. It is important to note, however, that some challenges cannot probably be resolved bioinformatically without further advancements

in sequencing technologies. For example, the complications introduced by genomic variation regions and repeats could not be explicitly resolved unless longer reads (longer than the repeat region for example) become available.

With the progress in bioinformatics approaches, direct recovery of genomes from metagenomes has become a possibility, which provides new means for more in-depth investigations. For example, Iverson *et al* demonstrated the successful recovery of the genome of a previously unknown, abundant euryarchaean directly from two marine datasets, sequenced by the SOLiD platform, and clarified the origin of the proteorhodopsin (68). Denef *et al* recovered the dominant chemolithoautotrophic *Leptospirillum* group residing in an acid mine drainage using a time series metagenomic dataset spanning a period of ten years, estimated *in situ* mutation rate, and reconstructed the ancestral genome (41). Wrighton *et al* reconstructed 49 partial or near-completed genomes from a temporal collection of samples from an acetate-simulated underground aquifer community and found unique physiological characteristics for three abundant yet uncultivated anaerobic bacterial populations (e.g., a hybrid RuBisCo for fermentation). Based on the previous examples, it is strongly anticipated that both time- and spatial-gradient collections of datasets from various habitats would become available in the immediate future. Such datasets can be used to tackle more complicated questions such as what is the role of HGT in lineage evolution and what is the relative importance of the different HGT mechanisms (e.g., viral- vs. conjugative pilus mediated) under natural settings and over periods of time that matter for human activities (e.g., days or months). However, the bioinformatic approaches that are necessary for these types of metagenomic data and analyses have not been developed yet, and therefore are in urgent need.

In situ bacterial lineage evolution

To better understand how bacterial lineages emerge, are maintained and evolve represents one of the most pressing questions in microbiology. A quantitative understanding of these questions will also lead to a better definition of bacterial species, a highly controversial yet, of great practical importance, issue for microbiology (69). Historically, isolate-based studies were employed to reveal the underlying molecular mechanisms of population evolution. After decades of research, it is now clear that the mechanisms include point mutations, intra-genome homologous recombination, and horizontal gene transfer and the strength of selection or neutral drift determine the outcome for the population. What is lacking is a complete understanding of how individual populations (or species) maintain their homogeneity and distinctiveness via the interplay of these mechanisms (70). For example, neutral drift and geographical or ecological barriers can diversify a population and drive the descendants into different lineages (candidates of novel species). On the other hand, homologous recombination or selective sweeps events purge population diversity and keep the individuals together as a distinct population (forces of population cohesion). HGT can potentially both homogenize and diversify a population (69). Researchers assessed the impacts of homologous recombination and horizontal gene transfer on population differentiation. Shapiro *et al* showed that inter-lineage gene exchanges are correlated with the relatedness between gene functions and ecological niches; and thus genes, instead of genomes, sweep through populations (11, 49). In natural settings, however, all these mechanisms are probably occurring simultaneously and their relative importance for bacterial genome evolution and adaptation is not well understood.

Among all the isolate-based studies, the most thoroughly investigated system is probably the long-term evolution study of *E. coli* lineages carried out in the Lenski Lab at Michigan State Univeristy, for more than twenty years now (translating to abuot fifty thousand generations) (71). By repeated daily transfers of 12 replicate *E. coli* populations into new flasks, these researchers have posed a strong selection pressure on the *E. coli* populations while, by studying frozen isolates from different time points of the experiment, they were able to reconstruct the mutations that occurred in the genome and compare these changes to the ancestor genomes. For example, Barrack *et al* studied the relationship between the rates of genomic evolution (e.g., accumulation of point mutations) and adaptation (e.g., reproducing rate) by complete genome sequencing of isolates from different time points during a 40,000-long generation period. They found that the genomic evolution rate was constant for the first 20,000 generation and most of the mutations were beneficial. However, after a frameshift mutation in the *mutT* gene at around generation 26,500, one lineage accumulated increased mutations and resulted in elevated whole genome mutation rate and fitness level compared to the other parallel lineages. This study demonstrated how mutation rate could change abruptly and affects the process of lineage adaptation (17). Other key issues investigated include the effect of co-existing lineages on the process of adaptation (72), and a functional innovation conferred by a mutation caused by tandem repeat (73, 74).

Although valuable insights were obtained from these and similar studies, the *E. coli* system described above cannot account for two important mechanisms that are highly relevant in nature. The first mechanism is HGT. By using pure cultures, it is impossible to evaluate how HGT would affect the path of evolution. The second

mechanism is lack of interactions with co-occurring (distinct) species. The existence of other species may alter the evolutionary path of a lineage significantly (e.g., by altering the strength of selection). Therefore, the lessons learned from pure cultures may not translate well to natural populations and studying the latter populations may provide new, and more practical, insights.

Toward this direction, a few studies appeared recently that targeted several housekeeping genes that are easy to amplify and compare. For example, by comparing the sequences of four housekeeping genes (*hsp60, mdh, adk,* and *pgi*), Hunt *et al* discussed the spatial and temporal resource partitioning among closely related, sympatric strains of *Vibrionaceae* retrieved from coastal bacterioplanktonic communities (75). They found that ecological specialization and differentiation within the same population might be a driving force or trigger speciation. Cordero *et al* further discussed the relationship between ecologically specialized *Vibrionaceae* populations and sensitivity to antibiotics (76). These authors found that groups with similar habitat associations tend to act as cohesive units with respect to resistance as well as production of antibiotics. These studies show that overlapping ecology (e.g., sympatric species) plays an important role in the evolution of bacterial lineages.

With the aid of high-throughput sequencing, it is now possible to begin genome-centric as opposed to gene-centric in these previous studies investigations of *in situ* microbial evolution using. A series of comparative metagenomic studies of ecologically distinct cyanobacterial *Prochlorococcus* populations uncovered several population-specific adaptations mediated by shifts in the relative abundance of genes within the population. For example, due to different nitrogen availability with depth in the oceans,

distinct, depth-stratified *Prochlorococcus* ecotypes (subgroups of a species that are adapted to a particular set of environmental conditions) have arisen and are easily discernible at the sequence level (77). Along the same line, Coleman and Chisholm reported that the relative abundance of phosphate-related gene differ between *Prochlorococcus* populations in the Pacific and Atlantic Oceans, consistent with stronger phosphorus limitation in the Atlantic Ocean (78).

Most, if not all studies to date, have focused on single lineages and intra-species diversity patterns, while in nature, many species often co-occur in the same inhabit. How to investigate natural populations within complex communities remains technically challenging however.

Horizontal gene transfer and its role in bacterial evolution

Horizontal gene transfer (HGT) represents a unique aspect of microbial evolution compared to higher eukaryotes and, as mentioned aboce, it can both diversify and homogenize a population, depending on the specific details. Various HGT mechanisms have been elucidated in the laboratory over the past few decades. The known mechanisms include transformation (naked DNA uptake from the environment), conjugation (intercellular DNA transfer mediated by conjugative pilli), transduction (phage-mediated DNA integration), gene transfer agents (phage-like DNA-vehicles), extracellular membrane vesicles (MVs; DNA-containing vesicles that could fuse into the cellular membrane of another species), and the more recently discovered inter-species nanotubes (DNA transfers between neighboring cells via tubular protrusions) (79). The factors that

determine HGT rates *in-situ* include the selection pressure on the gene transferred, the phylogenetic (sequence) and the ecological (physical) relatedness between the donor and the recipient, and the function of the transferred gene (79). What is missing is a quantitative understanding of the relative importance of these mechanisms and their cumulative impact on natural populations.

Substantial efforts have been made to evaluate the impacts of the previous factors. The rate of gene acquisition was estimated by Babic *et al* by visualizing conjugation events of a fluorescence gene in *E. coli* (80). These authors later carried out a similar experiment in *B. subtilis* to measure the dissemination of mobile elements (integrative and conjugative transposons?) within a bacterial colony (81). It was found that stochastic properties in the evolutionary history of a lineage could greatly impact the outcome of HGT and the likelihood of gene fixation. Further, it has been estimated that 18% of the total genes in *E. coli* were horizontally acquired after its split with *Salmonella* (estimated to have occurred about 100,000 years ago) and overall about 75% of the whole genome had an alien origin (82). Nonetheless, the *E. coli* genome size has probably remained relatively stable over this period. Therefore, equilibrium between gene acquisition and gene loss has been proposed. The high rate of HGT was further corroborated by the uneven distribution of age among the majority of genes present in the *E. coli* genome (82).

The impact of ecology and phylogeny on the rates of HGT are complicated, and a variety of methods have been developed to assess their relative importance. For instance, Popa *et al* developed a directed network of lateral gene transfer (dLGT) to integrate genomic similarity, phylogeny, and gene transfers into a graph representation (83). Smillie *et al* examined the network of recent HGT event among over two thousand

complete genomes of human-associated bacteria (84), and show that it is driven principally by ecology rather than phylogeny or geography. These authors also showed that, within the human microbiome, bacteria sharing a more similar ecological niche, evident, for instance, by simialr oxygen level tolerance or pathogenicity level, are significantly more likely to engage in HGT.

All these investigations have pointed to the importance of ecology on the frequency of HGT. Yet, few studies so far have considered a natural community in which species co-occur and share the same geography. For instance, the estimated HGT rate using complete genomes is likely an underestimate, since the genomes used were often recovered from geographically and/or temporary separated samples. To test these findings from isolates, it is important to investigate natural communities, and track natural populations over time. Simmons *et al* investigated the HGT events among a few closely related (94-99.5% average nucleotide sequence identity) subpopulations of *Leptospirillum* Group II within an acid mine drainage biofilm community, and found that horizontally exchanged plasmid/phage-like regions frequently contained functionally important genes (38). However, this study could not represent the average case in nature for two reasons. First, the acid mine drainage community is a reduced-diversity community, composed of (only) a few species. Second, the community is a closed, nearly perturbation free system, while in nature, the majority of the communities are constantly facing perturbations such as seasonal and weather changes (85) and migrations of exogenous species. Thus, investigating HGT events within complex natural communities is important.

OUTLINE OF DISSERTATION

The previous section underlined the importance of studying microbial evolution *in situ* as well as the need for developing bioinformatics/computational approaches that would enable such studies. Although substantial advancements in metagenomics have been made, several bottlenecks remain, especially with respect to the bioinformatics analysis of metagenomic data. During my Ph.D. research, I have undertaken several novel paths of research, and as a consequence, I was often faced with situations that no suitable solution was available. I developed a set of novel approaches, both conceptual and computational implementations, to achieve the objectives of my research. These novel approaches are expected to broadly benefit the scientific community. Using these novel approaches, several important discoveries have been made in the context of microbial evolution in complex communities.

Specifically, my Ph.D. research started with the project described in chapter 2, in which efforts were made to investigate how ecology played a role in speciation of the model bacterial species, the *Escherichia coli*. Nine *Escherichia* isolates recovered from diverse natural environments such as freshwater lake beaches, were sequenced (environmental genomes) and compared to available *Escherichia* isolates from the gut of warm-blooded animals (enteric genomes). The nine strains spanned four different clades, filling up the phylogenetic space between *E. coli* and its closest known relative, *E. fergusonii*. The apparent ecological difference between the environmental and the enteric strains offered an opportunity to evaluate the relative importance of ecology and phylogeny in shaping gene content and frequency of HGT. By comparing the clade-specific gene content and the patterns of inter-clade gene exchange, I found that ecology

played a more important role than phylogeny in driving the speciation of the *Escherichia* genus. The ecological barriers were not obvious previously due apparently to sampling and isolation biases (e.g., focus on clinical isolates). These findings have also major implications for the current bacterial species definition, which does not take into account ecological relatedness but is almost exclusively based on genetic relatedness.

In theory, time series metagenomics should address some of the critical questions in bacterial evolution under natural settings such as the rate and pattern of horizontal gene transfer. However, I soon realized that several major technical challenges remained before metagenomics can be employed for these purposes. Firstly, long-read sequencing technologies such as Roche 454 were not able to generate sufficient sequencing depth to fully recover the diversity and complexity of a medium or high complexity natural community such as the planktonic communities of freshwater Lake Lanier (Atlanta, GA). One appealing solution was to use the high throughput short-read sequencing technologies such as Illumina, but it was not clear whether these technologies could assemble long contigs and draft genomes from community metagenomic data.

Chapter 3 and 4 thoroughly assessed this issue. In chapter 3, Roche 454 and Illumina GA II technologies were directly compared based on the same community DNA sample. It was found that assemble sequences from the Illumina metagenome contained fewer errors and recovered more diversity compared to the Roche 454 ones, mostly due to the higher coverage obtained with Illumina. Roche 454 and Illumina showed strong agreement on the relative abundance of the sequences shared by the two assemblies. Using the principles learned from this work, I assessed how robustly a genome can be assembled from short-read community metagenomes, and investigated the relationship

between sequencing coverage and the quality of the recovered genome sequences in chapter 4. The analysis showed that, with ~15X coverage or better, a high quality genome sequence was obtained in terms of single base calling error rate, frameshift error rate, *etc*. Combining the conclusions from these two studies, it was clear that Illumina short-read sequencing technology was appropriate for addressing the biological questions of my thesis project.

Another challenge became obvious during my work; namely, how to accurately identify the taxonomic origins of assembled contig or raw metagenomic sequences. Because a metagenome is a mixture of numerous co-occurring species, the assembly represents a mixture of contigs from different genomes. It is difficult, yet important, to bin these contigs into population genomes, to facilitate downstream analysis. The available methods for these purposes were not adequate, especially for taxonomically classifying a large fraction of metagenomic sequences that represent novel taxa. Thus, I developed a fast and accurate classifier, MeTaxa, which showed improved performance compared to other algorithms. MeTaxa's advantage is rooted to the use of every gene in a query sequence as a classifier, weighting each gene differently based on its (predetermined) classification power. MeTaxa outperformed other available algorithms in accuracy and especially in determining the level of novelty (e.g., novel species, genus or phylum) of sequences representing novel taxa. MeTaxa's framework and performance benchmarking are introduced in chapter 5.

I applied these bioinformatics approaches on time series metagenomes from two real environmental systems. The first represented comparative metagenomic analysis to evaluate the responses of soil microbial communities to a decade of warming by 2°C

against the control communities (adjacent soils that underwent no warming), described in chapter 6. During this work, several novel computational pipelines were developed to handle and analyzed the massive data available (in excess of 50 Gbp in total). Community-wide adaptations to warming were observed and several essential pathways related to Carbon and Nitrogen cycles and the metabolism of greenhouse gases were enriched in the heated communities, suggesting potential exacerbating feedback from the soil microbial community to the greenhouse gas concentration. These discoveries collectively improved our understanding of how complex soil communities respond to environmental perturbations.

The second system represented a 2.5 years long collection of time series metagenomes from a freshwater lake planktonic community. The goal of this project was to quantify HGT *in situ* furing the period spanned by our samples. No suitable approach was available to carry out this task, and hence I developed two sets of computational solutions to address the bioinformatics challenges. The first challenge was to bin metagenomic contigs that represented the same population and, based on the binned contigs, to predict and quantify HGT events between the populations represented by the contigs. For the former, I developed an approach that utilizes the pair-end read links, contig coverage co-variance in the time series data, and tetranucleotide statistics to bin contigs into population genomes. For the latter, I developed an algorithm, metaHGT, to accurately predict HGTs based on pair-end read mapping information. Both approaches can be broadly applied to other systems such as the human microbiome and laboratory microcosms. Based on these novel approaches, I identified surprisingly high frequency of HGT between distantly related organisms, and some of them presumably underlay

important community adaptation to short-term environmental perturbations and population dynamics. These methods, approaches and discoveries are described in chapter 7.

REFERENCES

1. Madigan MT (2012) *Brock biology of microorganisms* (Benjamin Cummings, San Francisco) 13th Ed pp xxviii, 1043, 1077 p.
2. Cohen SN, Chang AC, & Hsu L (1972) Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. *Proceedings of the National Academy of Sciences of the United States of America* 69(8):2110-2114.
3. Chen I & Dubnau D (2004) DNA uptake during bacterial transformation. *Nat Rev Microbiol* 2(3):241-249.
4. Lederberg J & Tatum EL (1946) Gene recombination in Escherichia coli. *Nature* 158(4016):558.
5. Drake JW, Charlesworth B, Charlesworth D, & Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148(4):1667-1686.
6. Touchon M*, et al.* (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
7. Meyer JR*, et al.* (2012) Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science* 335(6067):428-432.
8. Moran NA, McLaughlin HJ, & Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323(5912):379-382.
9. Ochman H & Davalos LM (2006) The nature and dynamics of bacterial genomes. *Science* 311(5768):1730-1733.
10. Moreno-Hagelsieb G, Trevino V, Perez-Rueda E, Smith TF, & Collado-Vides J (2001) Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. *Trends in genetics : TIG* 17(4):175-177.
11. Caro-Quintero A*, et al.* (2011) Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea. *The ISME journal* 5(1):131-140.
12. Hasan NA*, et al.* (2010) Comparative genomics of clinical and environmental Vibrio mimicus. *Proceedings of the National Academy of Sciences of the United States of America* 107(49):21134-21139.
13. Chen I, Christie PJ, & Dubnau D (2005) The ins and outs of DNA transfer in bacteria. *Science* 310(5753):1456-1460.
14. Garcia-Vallve S, Romeu A, & Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome research* 10(11):1719-1725.
15. Gogarten JP, Doolittle WF, & Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution* 19(12):2226-2238.
16. Sorek R*, et al.* (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318(5855):1449-1452.
17. Barrick JE*, et al.* (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461(7268):1243-1247.
18. Perfeito L, Fernandes L, Mota C, & Gordo I (2007) Adaptive mutations in bacteria: high rate and small effects. *Science* 317(5839):813-815.
19. Kuo CH, Moran NA, & Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome research* 19(8):1450-1454.

20. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature reviews. Genetics* 10(3):195-205.
21. Nilsson AI*, et al.* (2005) Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America* 102(34):12112-12116.
22. Zhang Q*, et al.* (2011) Acceleration of emergence of bacterial antibiotic resistance in connected microenvironments. *Science* 333(6050):1764-1767.
23. Cooper VS & Lenski RE (2000) The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature* 407(6805):736-739.
24. Suerbaum S & Josenhans C (2007) Helicobacter pylori evolution and phenotypic diversification in a changing host. *Nature reviews. Microbiology* 5(6):441-452.
25. Kashtan N, Noor E, & Alon U (2007) Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America* 104(34):13711-13716.
26. Lieberman E, Hauert C, & Nowak MA (2005) Evolutionary dynamics on graphs. *Nature* 433(7023):312-316.
27. Kussell E & Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. *Science* 309(5743):2075-2078.
28. Elena SF & Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature reviews. Genetics* 4(6):457-469.
29. Mizunoe Y, Wai SN, Takade A, & Yoshida S (1999) Restoration of culturability of starvation-stressed and low-temperature-stressed Escherichia coli O157 cells by using H2O2-degrading compounds. *Archives of microbiology* 172(1):63-67.
30. Takai K & Nakamura K (2011) Archaeal diversity and community development in deep-sea hydrothermal vents. *Current opinion in microbiology* 14(3):282-291.
31. Ley RE, Lozupone CA, Hamady M, Knight R, & Gordon JI (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature reviews. Microbiology* 6(10):776-788.
32. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, & Gordon JI (2005) Host-bacterial mutualism in the human intestine. *Science* 307(5717):1915-1920.
33. Torsvik V, Goksoyr J, & Daae FL (1990) High diversity in DNA of soil bacteria. *Applied and environmental microbiology* 56(3):782-787.
34. Allen EE & Banfield JF (2005) Community genomics in microbial ecology and evolution. *Nature reviews. Microbiology* 3(6):489-498.
35. Weinberg Z, Perreault J, Meyer MM, & Breaker RR (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* 462(7273):656-659.
36. Dinsdale EA*, et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629-632.
37. Venter JC*, et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.
38. Simmons SL*, et al.* (2008) Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS biology* 6(7):e177.

39. Dick GJ, *et al.* (2009) Community-wide analysis of microbial genome sequence signatures. *Genome biology* 10(8):R85.
40. Morowitz MJ, *et al.* (2011) Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proceedings of the National Academy of Sciences of the United States of America* 108(3):1128-1133.
41. Denef VJ & Banfield JF (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336(6080):462-466.
42. Denef VJ, Mueller RS, & Banfield JF (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *The ISME journal* 4(5):599-610.
43. Lin Y, *et al.* (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 27(15):2031-2037.
44. Fleischmann RD, *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496-512.
45. Loman NJ, *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* 10(9):599-606.
46. Hughes JB, Hellmann JJ, Ricketts TH, & Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and environmental microbiology* 67(10):4399-4406.
47. Koeppel A, *et al.* (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences of the United States of America* 105(7):2504-2509.
48. Leffler EM, *et al.* (2012) Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS biology* 10(9):e1001388.
49. Shapiro BJ, *et al.* (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077):48-51.
50. Anonymous (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet,* The National Academies Collection: Reports funded by National Institutes of Health, Washington (DC)).
51. Hugenholtz P, Goebel BM, & Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology* 180(18):4765-4774.
52. Breitbart M, *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* 99(22):14250-14255.
53. Hugenholtz P & Tyson GW (2008) Microbiology: metagenomics. *Nature* 455(7212):481-483.
54. Tyson GW, *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37-43.
55. Rusch DB, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* 5(3):e77.

56. Venter JC*, et al.* (2001) The sequence of the human genome. *Science* 291(5507):1304-1351.
57. Konstantinidis KT, Braff J, Karl DM, & DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Applied and environmental microbiology* 75(16):5345-5355.
58. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
59. Miller JR, Koren S, & Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95(6):315-327.
60. Lin Y*, et al.* (Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 27(15):2031-2037.
61. Butler J*, et al.* (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* 18(5):810-820.
62. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18(5):821-829.
63. Gnerre S*, et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 108(4):1513-1518.
64. Namiki T, Hachiya T, Tanaka H, & Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*.
65. Peng Y, Leung HC, Yiu SM, & Chin FY (2011) Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27(13):i94-101.
66. Chitsaz H*, et al.* (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature biotechnology* 29(10):915-921.
67. Miller CS, Baker BJ, Thomas BC, Singer SW, & Banfield JF (2011) EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome biology* 12(5):R44.
68. Iverson V*, et al.* (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335(6068):587-590.
69. Papke RT & Gogarten JP (2012) How Bacterial Lineages Emerge. *Science* 336(6077):45-46.
70. Caro-Quintero A & Konstantinidis KT (2012) Bacterial species may exist, metagenomics reveal. *Environmental microbiology* 14(2):347-355.
71. Lenski RE (2011) Evolution in action: a 50,000-generation salute to Charles Darwin. *Microbe* 6(1):4.
72. Le Gac M, Plucain J, Hindre T, Lenski RE, & Schneider D (2012) Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 109(24):9487-9492.
73. Blount ZD, Barrick JE, Davidson CJ, & Lenski RE (2012) Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489(7417):513-518.
74. Blount ZD, Borland CZ, & Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli.

Proceedings of the National Academy of Sciences of the United States of America 105(23):7899-7906.
75. Hunt DE*, et al.* (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085.
76. Cordero OX*, et al.* (2012) Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* 337(6099):1228-1231.
77. Martiny AC, Kathuria S, & Berube PM (2009) Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. *Proceedings of the National Academy of Sciences of the United States of America* 106(26):10787-10792.
78. Coleman ML & Chisholm SW (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences of the United States of America* 107(43):18634-18639.
79. Popa O & Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Current opinion in microbiology* 14(5):615-623.
80. Babic A, Lindner AB, Vulic M, Stewart EJ, & Radman M (2008) Direct visualization of horizontal gene transfer. *Science* 319(5869):1533-1536.
81. Babic A, Berkmen MB, Lee CA, & Grossman AD (2011) Efficient Gene Transfer in Bacterial Cell Chains. *Mbio* 2(2).
82. Lawrence JG & Ochman H (1998) Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16):9413-9417.
83. Popa O, Hazkani-Covo E, Landan G, Martin W, & Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome research* 21(4):599-609.
84. Smillie CS*, et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.
85. Giovannoni SJ & Vergin KL (2012) Seasonality in ocean microbial communities. *Science* 335(6069):671-676.

CHAPTER 2

Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species

Parts of this chapter have been published in the article: C. Luo, S. T. Walk, D. M. Gordon, M. Feldgarden, J. M. Tiedje, and K. T. Konstantinidis. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl. Acad. Sci. U. S. A.* 2011, 108(17): 7200-5.

INTRODUCTION

The current bacterial species definition (*1*), while pragmatic and universally applicable within the bacterial world (*2*), remains controversial because it is difficult to implement due to technological limitations in identifying diagnostic traits, resulting frequently in species that are not adequately predictive of phenotype (*3, 4*). Further, and perhaps more importantly, it remains unclear whether the processes driving diversification and adaptation of bacteria produce sufficiently discrete groups of individuals (species) as opposed to a genetic continuum (*4, 5*) ["fuzzy" species, (*6*)]. An improved understanding of bacterial species and its definition has important broader impacts, such as for reliable diagnosis of infectious disease agents, intellectual property rights, international and national regulations for transport and possession of pathogens, bioterrorism agent oversight and reporting, and quarantine. Since the scientific, medical, regulatory, legal communities as well as the public expect species to reasonably reflect the phenotype and ecology of an organism, efforts towards a more refined definition of a bacterial species are needed.

The case of *E. coli* captures many of the problematic aspects of the bacterial species issue and has additional important ramifications for diagnostic microbiology and for assessing fecal pollution of natural ecosystems. Microbiological dogma is that *E. coli* strains live within the gastrointestinal tract of humans and other warm-blooded animals, are transmitted to susceptible host via the fecal-oral route, and do not survive for extended periods outside its host. *E. coli* is phylogenetically distinct (monophyletic) as are the other know *Escherichia* species, *E. fergusonii* and *E. albertii* (*7*). Despite their phylogenetic cohesiveness, however, *E. coli* strains are ecologically and phenotypically

heterogeneous (*3, 7*) and, in fact, a few strains have been assigned to a different genus (e.g. *Shigella flexneri*), based primarily on their distinct clinical presentation and importance as human pathogens (*8*). Whether pathogens, like *Shigella*, or other delineable groups of strains deserve their own taxonomic classification is currently based on subjective observations rather than empirical ecologic or phylogenetic data. This is attributable, at least in part, to the lack of data concerning truly innocuous (nonpathogenic) strains that are more relevant for comparisons to the life-threatening, pathogenic strains (e.g. negative controls). Furthermore, recent environmental surveys have repeatedly recovered substantial *E. coli* populations from soils and freshwater habitats (*9, 10*), indicating that "naturalized" (innocuous) strains (*11*) may be widespread in nature. These findings also imply that the current view of *E. coli* biodiversity and ecology might have been biased by the isolation procedures and/or the traditional focus on clinical samples. To what extend the latter populations represent truly autochthonous members of the natural communities sampled and how they differ genetically from host-associated *E. coli* remain elusive, however. Addressing these questions will have additional global consequences for the current practice of assessing fecal contamination based on *E. coli* cell counts (*10*).

MATERIALS AND METHODS

Information for each of the 25 *Echerichia* genomes used in this study is provided in Table 2.1. Twelve of the genomes (nine *Escherichia* spp. and two *E. albertii*) were sequenced as part of this study, using either the Illumina GA-II genome analyzer or the Roche 454 Sequencer available at the Genomic Facility at Michigan State University (Table 2.2). For sequencing, a pair-ended sequencing strategy (76-bp-long reads, 300-bp library insert size) was used that yielded ~300X coverage for each genome (one genome per Illumina lane). The accession numbers of the genomes sequenced in this study are provided in Table 2.2.

The 76-bp-long paired-ended reads first were clustered into two groups based on their quality score and length using the K-means algorithm, and the low-quality group was discarded. Sequences were trimmed further on both the 5'- and 3'-ends, baded on a threshold of Q=20, and were assembled using the Velvet algorithm (*12*). The K-mer parameter was varied to maximize the N50 of the resulting assembly for each genome (high stringency). Detailed statistics of each genome assembly are provided in Table 2.2. Comparisons of the assembly of genome TW10509 and the assembly performed at the Broad Institute based on independent, high-coverage 454 data revealed that our contigs had very low sequencing error (<0.01%) and contained no misassemblies or contaminating sequences (Figure 2.1). Our *in silico* evaluation also suggested that our assemblies recovered at least 98% of the core and 95% of the total genes in the genome (Figure 2.1). The few genes missing from our assemblies did not affect our conclusions because our analyses were based primarily on core genes recovered intact in all genome

sequences. Genes on the assembled contigs were identified by the GeneMark pipeline (*13*) and annotated as previously described (*14*).

Table 2.1. The genomes used in this study.

Strain	Lineage	Ecotype[a]	Pathotype[b]	Genome source[c]	Origin	Strain source	Sample type
MG1555	*E. coli*	GIT	commensal	NCBI	CA, USA	Human	feces
HS	*E. coli*	GIT	commensal	NCBI	MA, USA	Human	feces
SE11	*E. coli*	GIT	commensal	NCBI	Japan	Human	feces
IAI1	*E. coli*	GIT	commensal	NCBI	France	Human	feces
ED1a	*E. coli*	GIT	commensal	NCBI	France	Human	feces
Sakai	*E. coli*	GIT	EHEC	NCBI	Sakai, Japan	Human	feces
EDL933	*E. coli*	GIT	EHEC	NCBI	MI, USA	Food	ground beef
UTI89	*E. coli*	GIT/UT	UPEC	NCBI	unknown	Human	unknown
536	*E. coli*	GIT/UT	UPEC	NCBI	unknown	Human	unknown
CFT073	*E. coli*	GIT/UT	UPEC	NCBI	MA, USA	Human	blood
O1	*E. coli*	GIT/Other	APEC	NCBI	USA	Chicken	lung
ATCC	*E. fergusonii*	multiple	multiphle	NCBI	USA	Human	feces
TW08933	*E. albertii*	GIT	serotype 7	This study	Bangladesh	Human	feces
TW15818	*E. albertii*	GIT/Other	diarrheic	This study	Australia	Poultry	feces
B156	*E. albertii*	GIT/Other	avirulent	Broad Inst.	Australia	Magpie	feces
TW10509	*Escherichia* Clade I	GIT	ETEC	This study	India	Human	feces
TW15838 sediment	*Escherichia* Clade I	GIT	avirulent	This study	Australia	Environment	freshwater
TW09231 beach	*Escherichia* Clade III	ENV	avirulent	This study	MI, USA	Environment	freshwater
TW09276 beach	*Escherichia* Clade III	ENV	avirulent	This study	MI, USA	Environment	freshwater
H605	*Escherichia* Clade IV	ENV	avirulent	Broad Inst.	Australia	Human	feces
TW14182 beach	*Escherichia* Clade IV	ENV	avirulent	This study	MI, USA	Environment	freshwater
TW11588	*Escherichia* Clade IV	ENV	avirulent	This study	Puerto Rico	Environment	soil
E1118	*Escherichia* Clade V	ENV	avirulent	Broad Inst.	Australia	Environment	freshwater
TW09308 beach	*Escherichia* Clade V	ENV	avirulent	This study	MI, USA	Environment	freshwater
CT18	*S. typhi*	GIT	typhoid	NCBI	Vietnam	Human	unknown

[a]Ecotype designation is based on the frequency of isolation from various hosts (gastrointestinal tract, GIT, or urinary tract, UT) and the environment (ENV).

[b]Pathotype refers to the interaction between a particular strain and its host. Commensal strains do not cause disease and are commonly found in the GI tract of healthy humans, enterohemorrhagic *E. coli* (EHEC) strains cause bloody diarrhea in humans, urinary pathogenic *E. coli* (UPEC) cause urinary tract infections in humans and animals, avian pathogenic *E. coli* (APEC) cause a range of diseases in birds, enterotoxigenic *E. coli* (ETEC) cause watery diarrhea in humans, and avirulent strains have not been associated with a particular disease or a commensal phenotype.

[c]Publically available genomes were downloaded from the National Center for Biotechnology Information (NCBI) or the Broad Institute (Broad Inst.).

Table 2.2. The genomes sequenced as part of the study.

Item	TW08933	TW15818	TW10509	TW15838	TW09231	TW09276	TW14182	TW11588	TW09308
Project ID	56117	56131	56135	56127	56125	56123	56133	59765	59763
NCBI acc. No.	AEJU00000000	AEJY00000000	AEKA00000000	AEJX00000000	AEJW00000000	AEJV00000000	AEJZ00000000	AEMF00000000	AEME00000000
Collection time	3/13/03	9/27/01	1987	10/5/03	8/20/02	7/24/02	6/25/02	2/26/2003	7/9/02
Geographic location	Bangladesh	Australia	India	Australia	MI, USA	MI, USA	MI, USA	Puerto Rico	MI, USA
Strain source	human	poultry	human	environment	environment	environment	environment	environment	environment
Sample type	feces	feces	feces	freshwater sediment	freshwater beach (water)	freshwater beach (sand)	freshwater beach (sand)	Soil (0.5cm top soil, pH 4.5)	freshwater beach (sand)
Lineage	*E. albertii*	*E. albertii*	*Escherichia* C-I	*Escherichia* C-I	*Escherichia* C-III	*Escherichia* C-III	*Escherichia* C-IV	*Escherichia* C-IV	*Escherichia* C-V
DNA sequence source	Michigan State University	Michigan State University	Michigan State University	Michigan State University	Michigan State University	Michigan State University	Michigan State University	Michigan State University	Michigan State University
DNA preparation	Qiagen Gentra Puregene								
Sequencing method	Illumina GA II (2x76bp)							Roche 454 Titanium (avg. 240bp)	
Assembly software	*de novo*, Velvet 0.7.51							*de novo,* Newbler 1.1.01.20	
Coverage	~300X by 2 (coupled reads)							~20X	
Assembly error rate	~6bp per 100Kbp (0.006%)*							<0.5%	
Size (Mbp)	4.51	4.73	5.19	5.27	4.74	4.47	4.68	4.46	4.81
N50 (Kbp)	27.9	24.7	27.6	22.2	19.8	28.7	19.7	88.4	104.1
Contig number	370	392	453	517	495	328	533	241	194

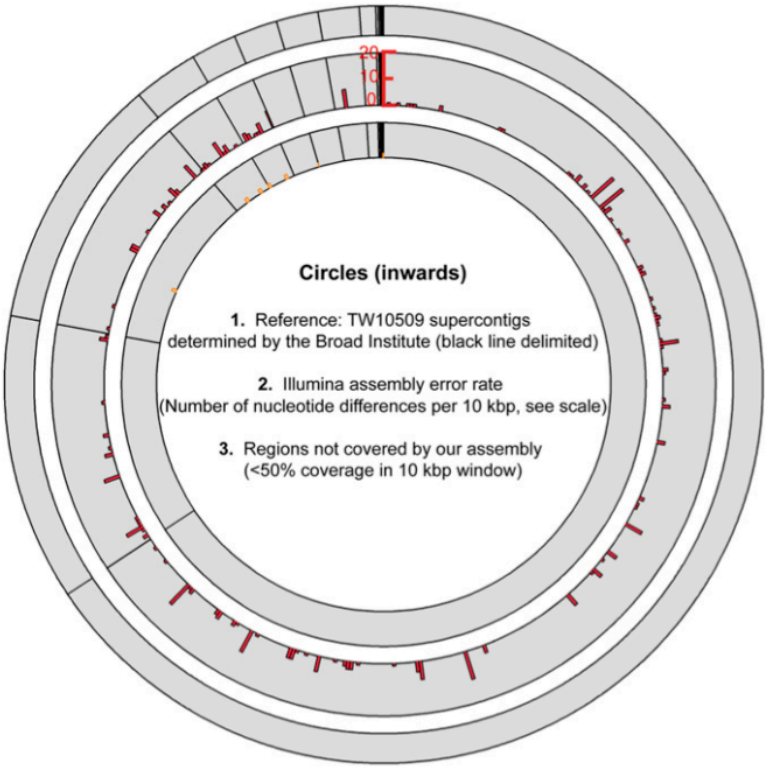


Figure 2.1. Assessing Illumina assembly quality against a reference genome. The genome of strain tw10509 was sequenced independently and assembled at the Broad Institute based on high-coverage 454 data (final genome is in nine supercontigs, each at least 2Kbp long). Comparisons of this genome assembly with the Illumina-based assembly determined by this study for the same strain showed that our assembly was of high quality and recovered almost all (>98-99%) of the core genes in the genome. In particular, the two assemblies typically disagreed in <20 nucleotide bases over a 10-Kbp window (second circle), and only 57 genes were missing from our assembly compared with the reference assembly. About half of the missing genes were integrases/transposons (24 genes); 20 genes were of hypothetical function; and eight genes had multiple copies in the genomes. Of the genes in our assembly, 229 (4.4% of total) were recovered truncated (i.e., <85% of the total gene sequence was recovered); only 25 of the truncated genes appear to be core genes (i.e., present in >18 genomes). The absence of misassemblies or genes

specific to our assembly indicates that the aliquot sequenced was free of contaminating DNA. Furthermore, our analysis showed that our TW10509 assembly was evenly covered by Illumina reads; i.e., 96% of the contigs had sequencing depth (coverage) between 280X and 320X. Similar results were obtained for the other sequenced genomes. Finally, the quality of our assemblies was evaluated further as follows: four Illumina sequencing datasets were generated *in silico* from the *Escherichia fergusonii ATCC 35469,* Escherichia coli O157:H7, EDL933, MG1655, and UTI89 complete chromosomal sequences using a custom Perl script (available from the authors upon request) and the same sequencing error, coverage, read length, and library insert size as in the real Illumina data. These *in silico* reads were assembled using the protocol and parameters described for real data. The resulting assembly recovered the complete sequence of 95.88% of the genes; 0.55% of genes were missed, and the remaining genes were recovered incomplete (truncated); sequencing error in the consensus sequence was limited to 0.009% bases, on average. These finding indicate that our draft genome sequences covered at least 95% of the genome of the isolates.

After all mobile elements (transposase, integrases, and so forth) and truncated gene sequences were removed, and all-versus-all BLAST search was carried out using all protein-coding genes annotated in all genomes. Alignments with coverage lower than 85% of the length of the query protein sequence were discarded. The analysis identified 1,910 genes that constituted reciprocal best matches in all pair-wise genome comparisons (core orthologs). These genes subsequently wer aligned using ClustalW2 (*15*), and the resulting alignments were concatenated to provide the whole-genome alignment and the Neighbor-Net algorithm (*16*) of the SplitsTree package and is shown in Figure 2.2. It should be noted that the set of 1,910 genes represents a subset of the total core genes

shared among the genomes analyzed (estimated to be around 2,200-2,500 genes, given that about 20-25 core genes were missed in each genome assembly and that we analyzed 12 draft genomes; Figure 2.2); it does not include truncated genes or genes not recovered in our assemblies. Nonetheless, the missing genes are highly unlikely to have a significant impact on the derived whole-genome phylogeny (because of the large number of genes included in the underlying alignment) or on the results of the horizontal gene transfer (HGT) analysis (see below), because they represented a small number of the total core genes in the genome and were distributed randomly around the genome (Figure 2.3).

To identify genes that recently were exchanged horizontally among the *Escherichia* clades, we used the approach outlined in Figure 2.4. In brief, the protein sequences of core orthologs (1,910 genes) were aligned using ClustalW2 (*15*). The corresponding nucleotide sequences of the aligned protein sequences subsequently were aligned, codon by codon, using the pal2nal script, with "remove mismatched codons" enabled and the protein alignement as the guide (*17*). Synonymous substitutions per site (Ks) were calculated based on the method described by Goldman and Yang (*18*) using KaKs_Calculator (*19*). To capture only recent HGT events, a Ks-based filter was applied to qualify orthologous genes that (i) has Ks values ! 0.02 (recentHGT events); ii) were not short (i.e., <300 bp) or truncated; and (iii) had a sequence that was not typically highly conserved within the *Escherichia* genus (i.e., the genes did not rank in the lower 15% of Ks values in all pair-wise genome comparisons). The cutoff Ks = 0.02 was used because it represented the average Ks among orthologs of genomes of the same lineage; hence, it was optimal for evaluating interlineage HGT events (we did not assess intralineage HGT). In addition, genes in the low Ks ranks that represented informational

genes, such as the ribosomal genes and DNA/RNA polymerases, were removed manually from further analysis because it could not be established whether the identity patterns observed were caused by genetic exchange or high sequence conservation. Fewer than 100 genes were removed. Embedded quartet decomposition analysis (EQDA) (*20*) was used subsequently to inferinterclade HGT events as follows. Embedded quartet analysis was applied to two clades at a time, using two genomes per clade (i.e., four genomes in total). The resulting phylogeny was bootstrapped and compared with the whole-genome tree topology. Only quartets incongruent with the genome topology and at least 95/100 bootstrap support were selected to represent HGT events. Noncore genes shared by at least two clades were assessed in the same way as core genes (Figure 2.4).

Although it is possible that our approach did not filter out a few informational genes that show high sequence conservation, this possibility should have no effect on our conclusion about the relative importance of HGT between commensal and environmental genomes, because HGT was assessed based on the same core genes for all genomes and genome quartets that showed comparable intergenome evolutionary relatedness. We also evaluated the extent to which EQDA analysis might be affected by the sequences used in the analysis; for instance, whether orthologous sets with high sequence similarity showed more false positives than more divergent orthologs because of the weak phylogenetic signal resulting from highly identical sequences. Our results which are summarized in Figure 2.6, suggested that our EQDA is impervious to such artifacts and that our approach did not underestimate the number of recently exchanged genes.

RESULTS AND DISCUSSION

Environmentally adapted *E. coli* lineages

We recently described five *Escherichia* clades (C-I to C-V) that were recovered primarily from environmental sources and are indistinguishable from typical *E. coli* based on traditional phenotypic tests included in either the API20E Identification System (biMerieux, Inc.) or the BBL Crystal Identification System (Becton, Dickinson and Company) (ref). To provide genomic insights into the phylogenetic diversity and metabolic potential of these clades, we sequenced the genome of nine representatives from clades C-I, -III, -IV, and -V (Table 2.1 and 2.2, and Figure 2.1). Whole-genome phylogenetic analysis confirmed our earlier observations based on multilocus sequence typing that the clades span the phylogenetic tree between *E. coli* and *E. albertii*, forming a genetic continuum within the *Escherichia* genus. In particular, C-I appears to be a sister clade of typical *E. coli*, being only slightly more divergent than the B2 phylogenetic lineage that includes uropathogenic *E. coli* (UPEC). The remaining clades are more divergent from typical *E. coli* (Figure 2.2). In agreement with previous phenotypic testing, the genomes of the strains of the four clades encode all genes shared by the available *E. coli* genomes (i.e., the *E. coli* core gene set) (Figure. 2.3A and Figure 2.4). Thus, the clades appear to be phenotypically and genetically (e.g., in gene content) indistinguishable from typical *E. coli*. Based on this information and the current genomic standards for species demarcation (*21*), these clades would be justifiably classified as *E. coli*.

The orders-of-magnitude higher abundances of these clades in environmental samples relative to those in human feces and the clinic (*10*) indicate that they represent

truly environmentally adapted organisms (meaning that they are not associated primarily with mammal hosts). Consistent with this interpretation, a recent study found that strains of clades C-III, -IV, and -V form biofilms more readily, outcompete typical *E. coli* strains at low temperatures (which characterize the environment compared with the gastronintestinal tract of warm-blooded hosts), and are nonpathogenic in a mouse model of septicemia (*22*). Furthermore, screening of 2,701 strains from humans, animals, and the environment identified an additional 57 environmental clade strains, and these strains were found more often in environmental and bird samples than in human samples (*10*). These studies consistently support the hypothesis that the environmental clades substantially expand the known ecological niche of *E. coli*.

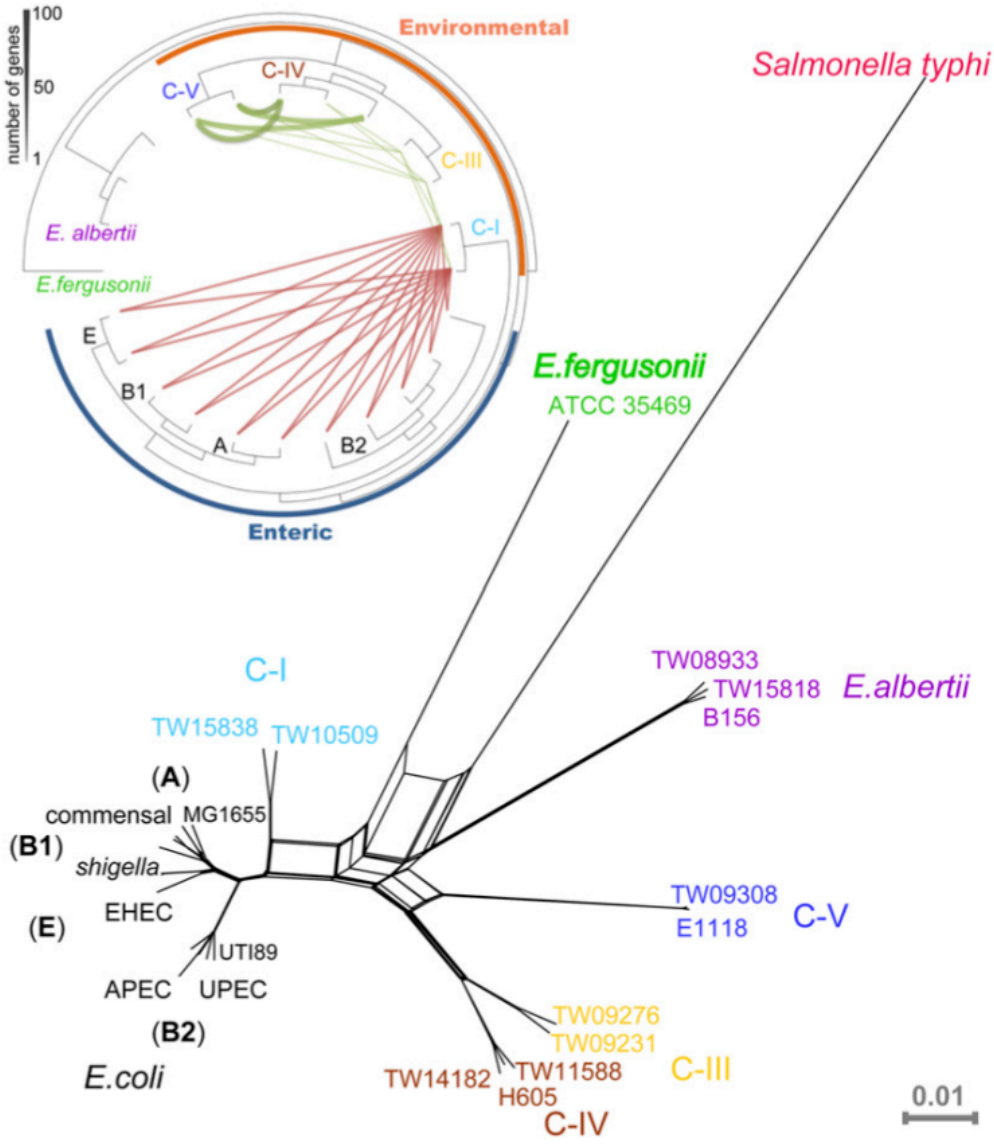


Figure 2.2. Whole-genome phylogeny of the *Escherichia* genomes used in this study. The phylogenetic network shown was constructed with the SplitsTree software (*16*), using as input the concatenated alignment of 1,910 single-copy core genes. (*Inset*) The graph represents the amount of recent horizontal transfer of core genes between the genomes of the clades. The thickness of the line is proportional to the number of genes transferred (scale at upper left in figure).

Functions important in the gut

Comparisons between the environmental genomes and their commensal or pathogenic (enteric) counterparts provided insights into the functional differentiation of *E. coli* strains. Consistent with the core gene results described above, we found almost no genes specific to enterics when queried against all genomes of environmental clades (Figure 2.4). However, when the C-I clade was included in the enteric group (strains of C-I have been isolated from humans, and this clade does not appear to be overrepresented in environmental samples) and the stringency of the comparisons was relaxed to allow one or two genomes in each group not to encode the gene in question, we identified 84 and 120 genes as being specific to or highly enriched in the environmental and enteric groups, respectively (Figure 2.3B and Table A1). The environment-specific gene set included several genes of unknown function as well as the complete pathway for diol utilization (energy substrate) and the gene for lysozyme production (hydrolysis of bacterial cell walls). These functions apparently are important for resource acquisition and survival in the environment. In contrast, the enteric-specific functions included genes involved in the transport and use of several nutrients that are thought to be abundant in the gut, such as *N*-acetylglucosamine, gluconate, and 5-C and 6-C sugars such as fucose (*23*). The latter genes were significantly enriched in the recently determined human microbiome (*24*), further corroborating their importance for colonization of the gut. Therefore, these genes characterize enteric *E. coli* strains relative to their environmental counterparts and may represent robust biomarkers for the development of molecular assays to count commensal *E. coli* cells in environmental samples more accurately than done by current methods. The enteric gene set also includes several prophage genes,

consistent with recent finding from metagenomic studies indicating that the human virome is highly specialized to its host and differs from viromes of environmental ecosystems (*25*).

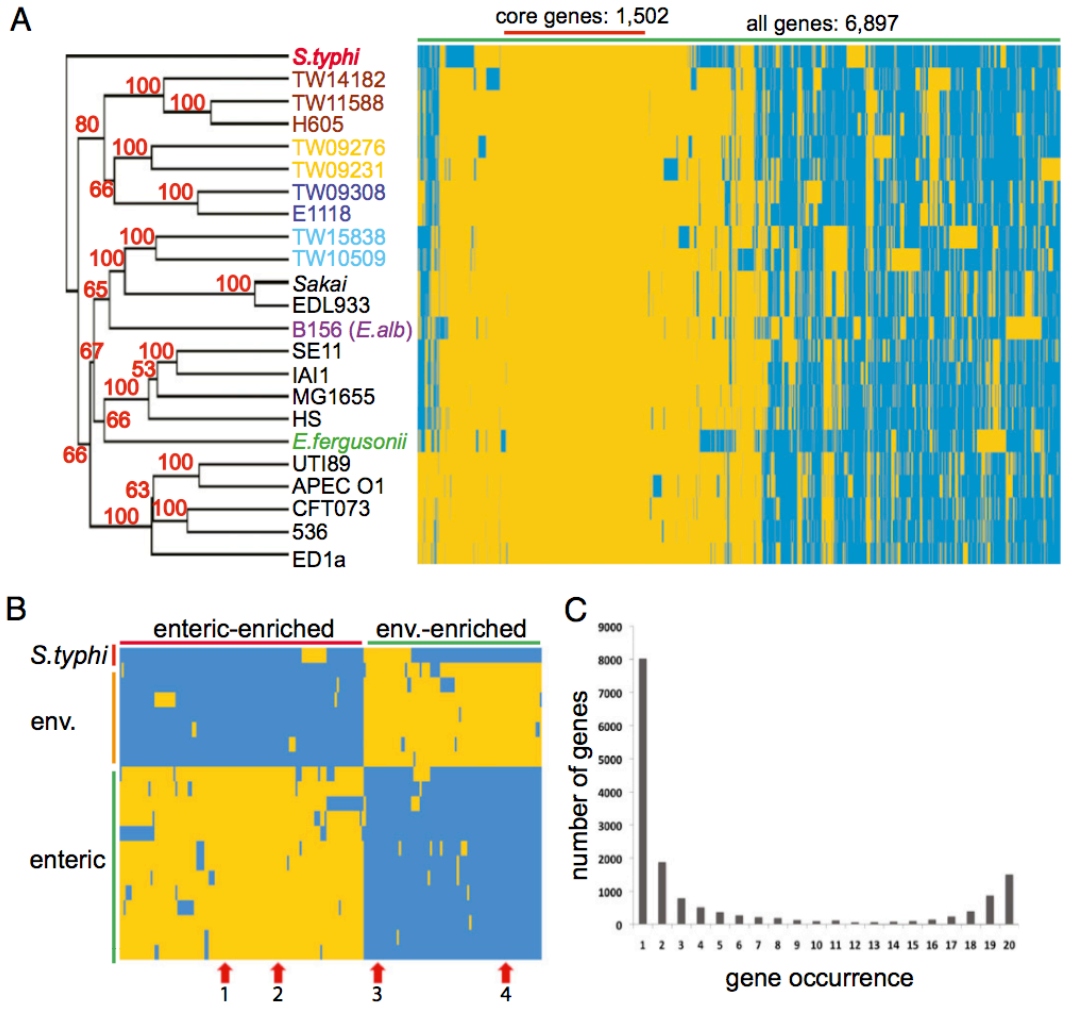


Figure 2.3. Gene-content signatures of *Escherichia* clades. Heatmap of gene presence (yellow) and absence (blue) in 20 selected genomes, using all nonredundant genes that were found in at least two of the genomes as reference. (**A**) Genomes were clustered based on the presence/absence of genes; values in red represent bootstrap support from Jackkinifing resampling with 100 replicates. (**B**) genes and pathways distinguishing enteric and environmental genomes were expanded (underlying data are provided in Table A2). 1, acetylglucosamine

transporter; 2, fructose transporter; 3, diol utilization operon; 4, lysozyme production. (**C**) Occurrence of the genes composing the *Escherichia* pangenome in the 20 genomes ranges from one (a genome-specific gene) to 20 (a core gene).

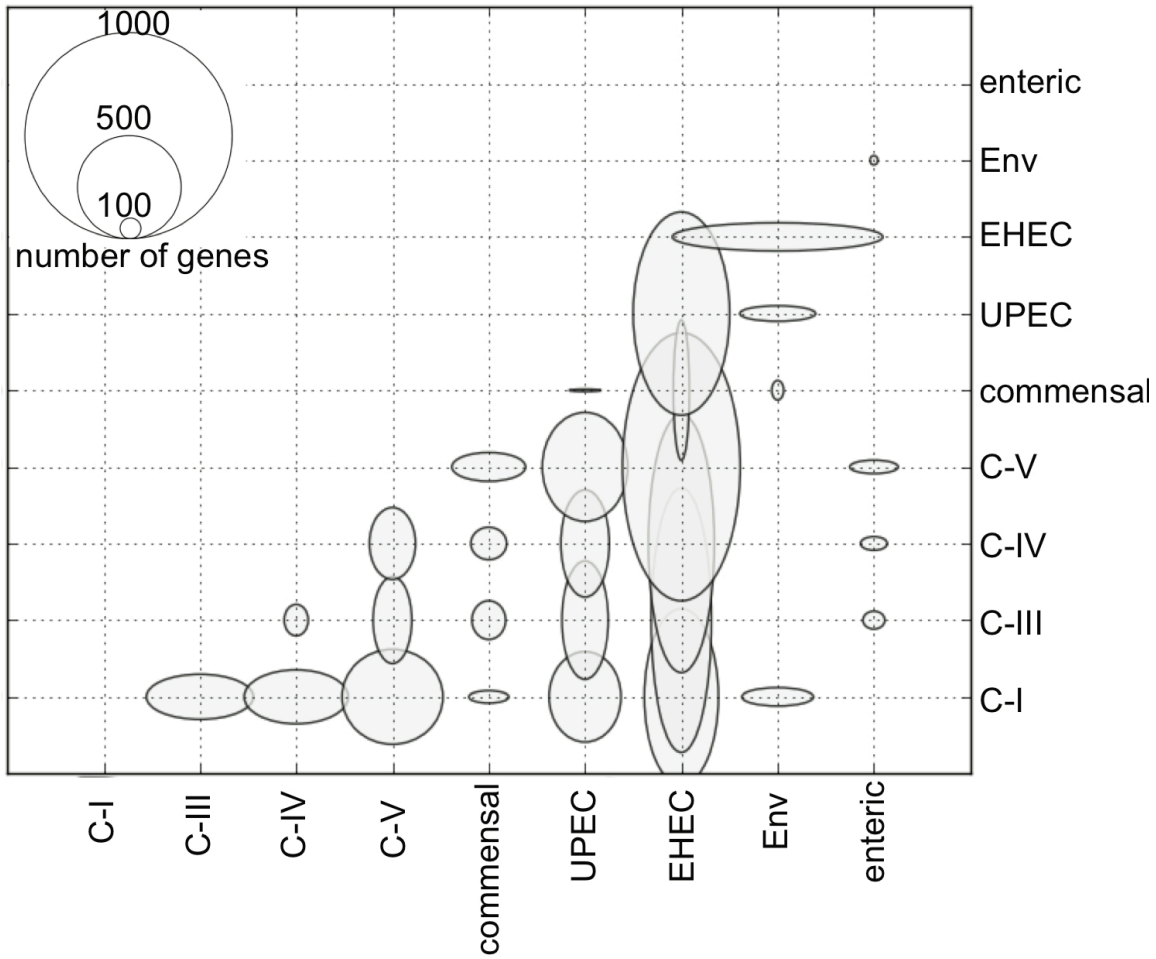


Figure 2.4. Summary of differences in gene content between *Escherichia* clades or groups of selected genomes. The centroid of the eclipse corresponds to the pair of clades in the comparison, and the height and width are proportional to the size (see key in figure) of the gene set that differentiates the clades (i.e., $a = X \setminus \overline{Y}$ and $b = \overline{X} \setminus Y$, where X and Y are the sets of shared genes within the corresponding clades on the x and y axis, respectively).

Ecologic barriers to gene flow within *Escherichia*

The availability of several genome sequences that span the *Escherichia* tree provided the opportunity to evaluate the importance of interclade genetic flow for *E. coli* evolution with greater phylogenetic coverage than previously achieved (*7, 26*). To this end, we devised a strategy to assess recent genetic exchange events based on embedded quartet decomposition analysis (EQDA) (refer to materials and methods for details; Figure 2.5 and 2.6). We focused on recent events only because historic genetic exchange of core genes (mediated by homologous recombination) frequently was impossible to detect robustly because of multiple (old) recombination events on the same segment of the genome and the process of amelioration of the newly introduced DNA sequence into the recipient cell (*27*).

We observed detectable genetic exchange of core genes within the environmental clades, within enterics, and between C-I and enterics but not between enterics and the remaining environmental clades or *E. albertii* (Figure 2.2 *Inset* and Figure 2.8). The core genes exchanged were distributed randomly in the genome and did not show any strong biases in terms of function when compared with the rest of the genome (Figure 2.9 and Table A2). These findings are consistent with a generalized mechanism for the transfer of genetic material (e.g., transformation or conjugation) and incorporation into the recipient genome via homologous recombination. They also confirm the closer affiliation of C-I with typical *E. coli* relative to the other clades and reveal reduced genetic flow between environmental and enteric genomes, presumably because of ecological barriers.

Nonetheless, the number of core genes exchanged within the evolutionary time that corresponded to 0.002 synonymous substitutions per site (the divergence time

typically separating the genomes of the same clade) accounted for only a small portion of the total core genes in the genome (0.06-2.33%). We also observed that noncore (auxiliary) genes were exchanged among the clades less frequently than core genes (Figure 2.2 and 2.8 and Table A2 and A3). Given also that more than 50% of the total unique genes of the *E. coli* pangenome are genome or clade specific (Figure 2.3C), our observations suggest that asexual divergence coupled with clade-specific gene acquisition or deletion dominates interclade recombination in driving *Escherichia* evolution.

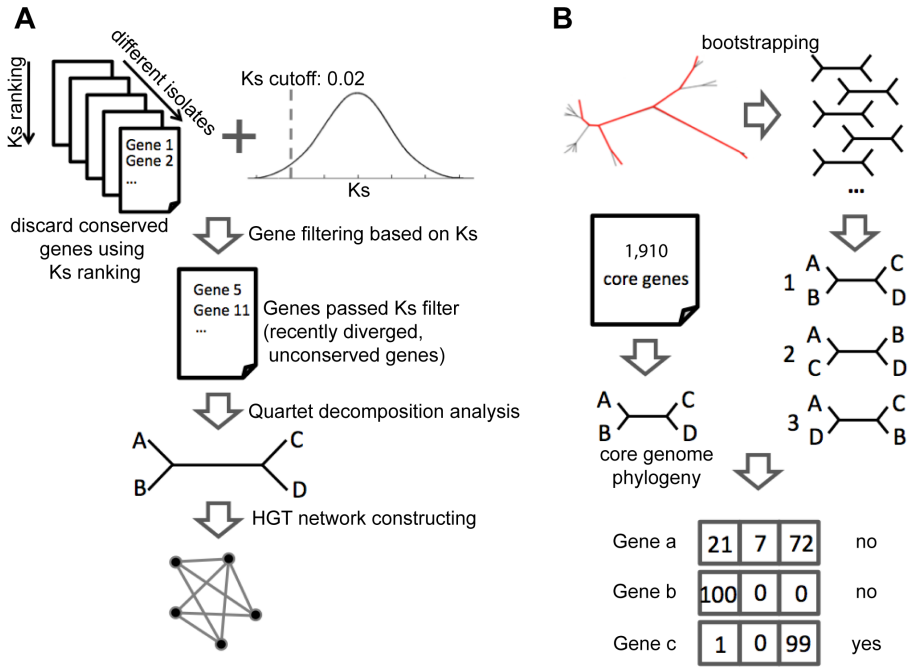


Figure 2.5. Flowchart of the horizontal gene transfer (HGT) network analysis. A shows the approach used to remove genes (<100) whose sequence typically was highly conserved within the

Escherichia genus from further analysis. **B** represents the embedded quartet decomposition analysis (EQDA), performed essentially as described in *Materials and Methods*.

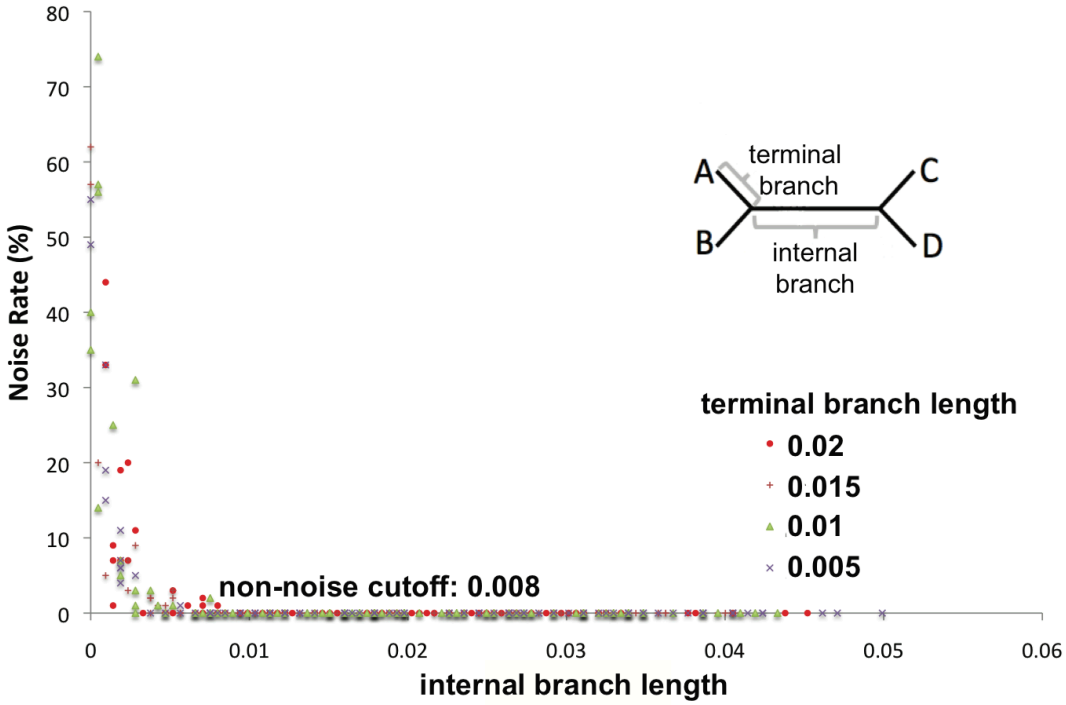


Figure 2.7. Robustness of the EQDA to phylogenetic noise. An *in silico* evaluation was performed based on the *recA* DNA sequences were mutated *in silico*. Subsequently, the mutated sequences were aligned by ClustalW2 and analyzed by EQDA, using the procedure used with real sequences. Thus, any bootstrapped tree based on a set of mutated sequences that was inconsistent with the [(A,B),(C,D)] topology was considered noise. The analysis revealed that when internal branch length exceeded 0.008 (*x* axis), no bootstrapped tree showed incongruence regardless of terminal branch length (*y* axis). In our study, we analyzed only orthologs that showed at least 0.01 internal branch length (e.g., the distance between the genomes compared was at least 0.02 in terms of Ks); thus, our EQDA was robust against phylogenetic noise.

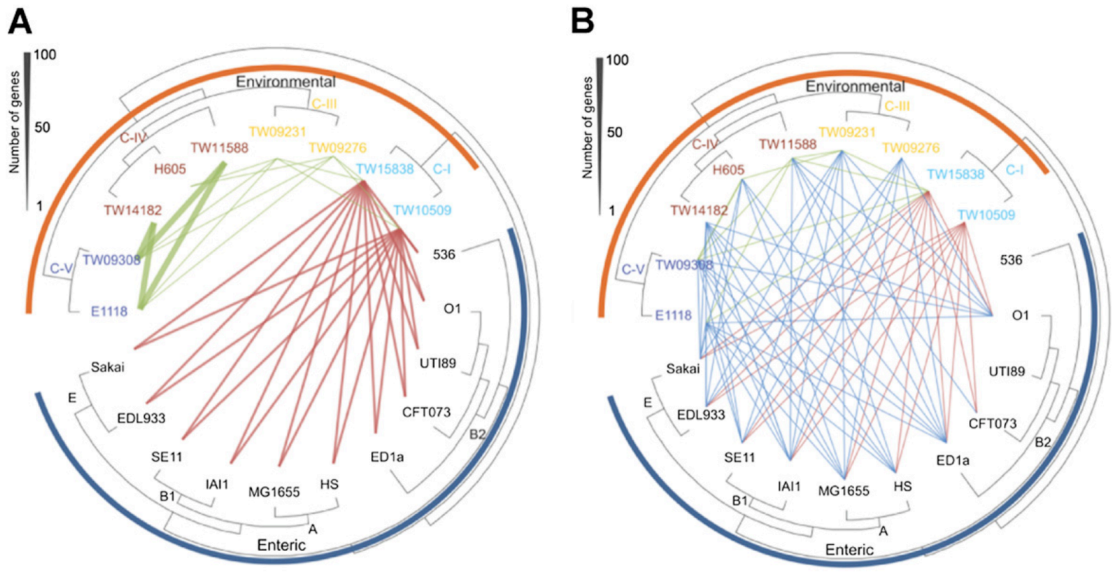


Figure 2.8. Recent genetic exchange of genes between clades. The graph represents the amount of recent horizontal transfer of core (**A**) and noncore (**B**) genes between the genomes of the clades (nodes on the tree). The thickness of the line is proportional to the number of genes transferred (see scaled in figure). The color of the line indicates whether the HGTs are within environmental (green) or between environmental and enteric (red) clades (note that no core gene transfer was observed between C-III, C-IV, or C-V and enteric clades). Panel **A** represents the full version of the figure shown in Figure 2.2 *Inset.*

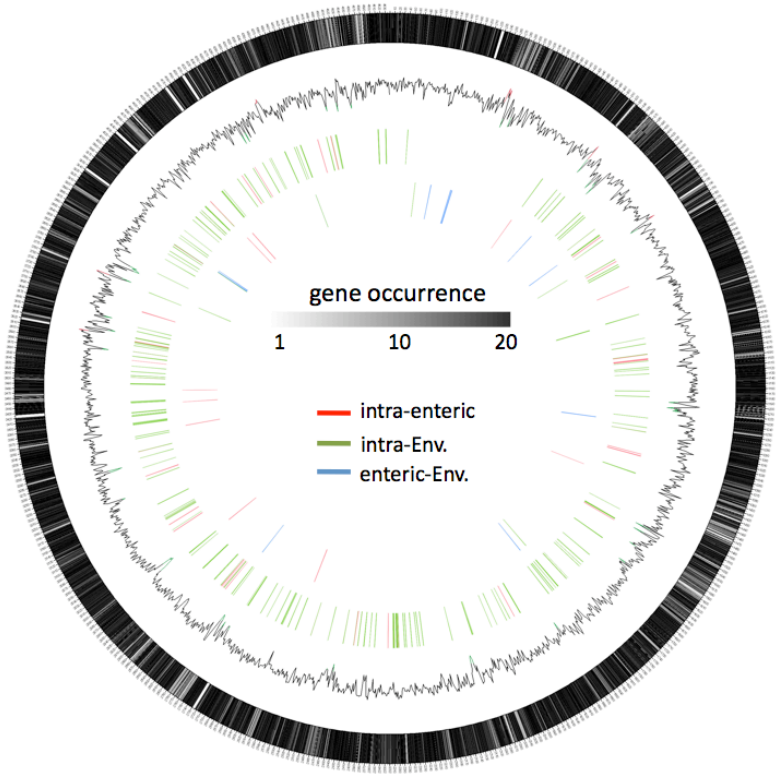


Figure 2.6. Chromosonal position of the genes identified as horizontally transferred between clades on the MG1655 (*E. coli* K-12) genome. From outermost of innermost, circles represent 1: *E. coli* MG1655 chromosome; 2: G+C content; 3: core genes involved in recent HGT events; 4: noncore genes involved in recent HGT events. Gene occurrence is indicated by gray scale in the outermost circle; occurrence ranges from one (a genome-specific gene) to 20 (a core gene). The possible partners in the HGT event are denoted by color (see key in figure). Intra-enteric, both clades involved in HGT are enteric; intra-Env., both clades involved in HGT are environmental; enteric-Env., HGT event is between enteric and environmental clades.

Test of the fragmented speciation model

It has been proposed recently that organism of the *Escherichia* genus evolve according to a fragmented speciation model (*28*) and that the model may be applicable to additional bacterial groups (*29*). If the model were true, one would expect that genomic islands that differentiate two ecologically distinct populations to be flanked by regions of increased nucleotide divergence, because such population-specific islands are free from the homogenizing effects of recombination. In other words, because interpopulation homologous recombination is halted around the genomic island (the sequence is not conserved in the population that does not carry the island), the genetic variation of the flanking DNA would be increased between the two populations compared to within either of the individual populations (Figure 2.10 gives a graphical representation for the expected signature of the model).

Our results strongly indicate that the environmentally adapted genomes are ecologically differentiated as compared with their enteric counterparts and thus are more appropriate for testing the model directly than are the divergent *Salmonella* genomes used previously (*28*). Although several candidate (ecologically relevant) genomic islands were identified (such as the islands encoding the fucose and gluconate utilization operons), and these islands were flanked by DNA sequences that were conserved and syntenic in the environmental strains, no island showed the predicted signature of the fragmented speciation model. Instead, the level of nucleotide divergence in the flanking regions of the islands covaried between the environmental and enteric genome (Figure 2.10). Similar patterns were observed when the analysis was restricted to commensal vs. pathogenic *E. coli* for the genomic islands that encode the know pathogenicity factors of

the latter genomes (Figure 2.11). Thus the predicted signature of the model was not observed even in comparisons of genomes that show both higher genetic relatedness and genetic flow than observed between environmental and enteric strains. In a few of the genomic islands examined, the flanking genes did show increased nucleotide divergence between ecologically distinct genomes. However, this pattern typically was associated with genes that were interrupted by the insertion of mobile elements; because of relaxed functional constraints, the truncation of gene(s), rather than the action of recombination, presumably caused an increased accumulation of mutations. Such truncated genes or their remnants may underlie some of the incongruent phylogenetic signal observed previously (*29*).

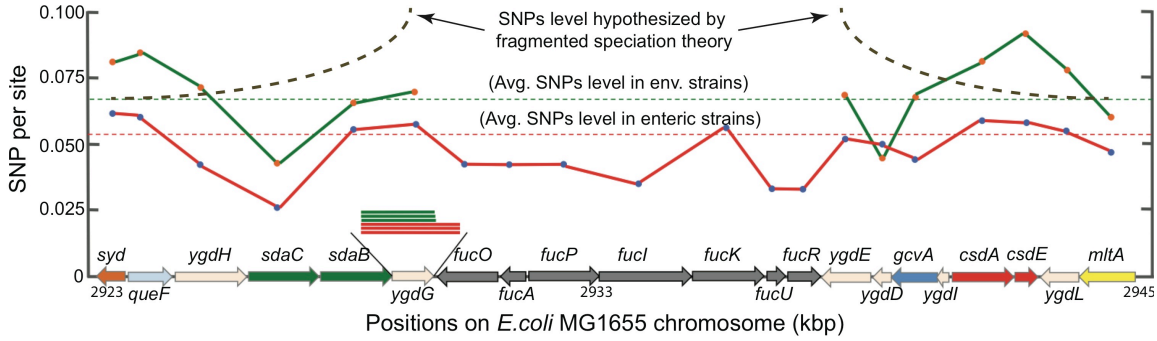


Figure 2.10. Lack of evidence in support of the fragmented speciation model. A representative example of the nucleotide divergence patterns, measured as the number of single nucleotide substitutions (or SNPs, y-axis), observed in flanking regions of a genomic island (x-axis) that differentiates environmental from enteric genomes. The island shown encodes the genes for utilization of fucose, a sugar commonly found in the glycan structures of the cell wall of animals.

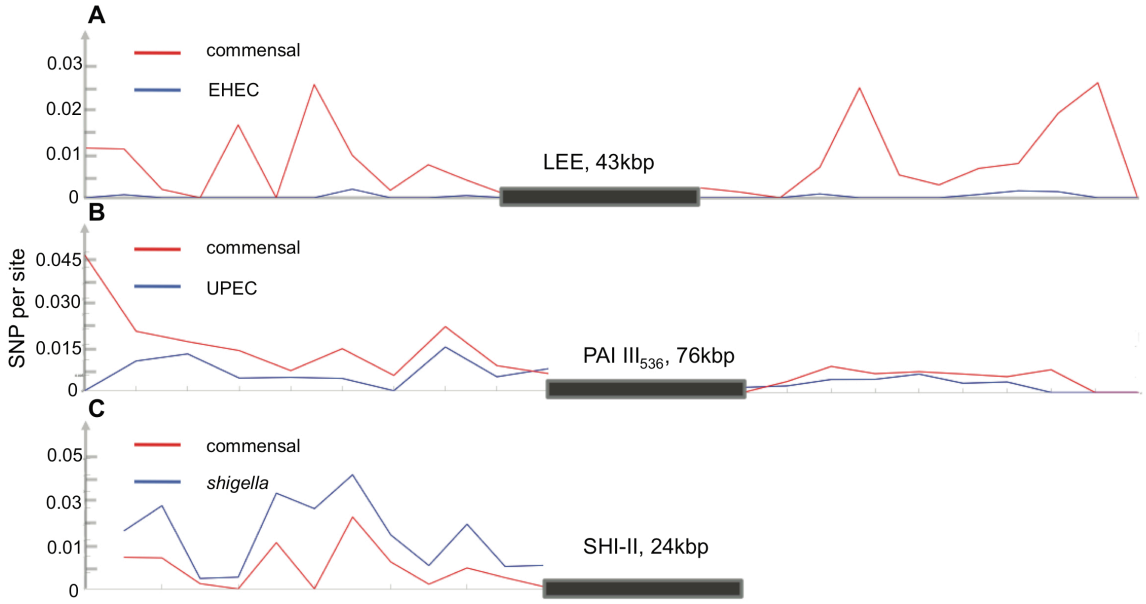


Figure 2.11. SNP levels in flanking regions of known pathogenicity islands. Three previously described pathogenicity islands are shown. (**A**) LEE in *E. coli* O157:H7 Sakai (GenBank accession No. NC_002965, coordinates 4580769..4623562). (**B**) PAI III_{536} in *E. coli* 536 (GenBank accession No. NC_008253, coordinates 294319..269466). (**C**) PAI SHI-II in *Shigella flexneri* (GenBank accession No. NC_004337, coordinates 3806404..3831722). Commensal genomes represent the average of four genomes (MG1655, IAI1, SE11, and HS).

Conclusions and perspectives

Our results collectively suggest that asexual divergence coupled with clade-specific gene acquisition or deletion has a much stronger influence on the evolution of the *Escherichia* genus than homologous recombination (sexual reproduction). These results differ quantitatively from those reported previously (*7, 26*). The difference is caused, at least in part, by the different genomes and methods used in the analysis. For instance, the previous studies evaluated the intraclade level, whereas our analysis was focused on more divergent genomes, an approach that is advantageous for unequivocally detecting recent gene exchange and recombination events (*14, 30*). Although our results do not rule out the existence of high levels of recombination within a clade, they do reveal that genetic exchange between incipient ecologically distinct clades of *E. coli* may not be as pronounced or prolonged as would be expected by the fragmented speciation model (*28*), and this reduced level of exchange probably accounts for the lack of evidence in support of the model.

Data described here concerning the environmental *Escherichia* clades show that justifiable species, which are ecologically distinct, sexually isolated, and phylogenetically tractable, may be identifiable even in cases of apparent phenotypic identity or a genetic continuum (such as revealed within the *Escherichia* genus in Figure 2.2). These findings, which also are consistent with recent metagenomic studies of natural populations, suggest that a more ecologic definition for species is more appropriate than the current definition that is heavily based on genetic distinctiveness alone. Comparative genomic analyses linked a substantial fraction of the clade-specific gene acquisitions (and deletions) to the unique ecology of the clade (e.g., Figure 2.3B). These findings further corroborate the

notion that it is time to start replacing traditional approaches of defining diagnostic phenotypes for new species with omics-based procedures.

What the preferred ecological niche or host (if any) of clades II-V is and whether the clades actually can persist in the external environment in the absence of fecal inputs (i.e., represent truly free-living bacteria) remain elusive, and additional data need to be collected before more robust conclusions can emerge. For instance, strains of clades II-V have been recovered occasionally from birds and ruminant mammals (*10*), but the extent to which these results are influenced by the processes of strain migration and extinction (as opposed to persistence within the host) is unclear. What our genomic data as well as data from physiological studies and environmental surveys performed previously (*10, 22*) suggest is that clades II-V are better at surviving in the external environment than is commensal *E. coli* and are poor competitors in the human gastrointestinal tract relative to successful clonal complexes such as those represented by CFT073 and MG1655 strains. Therefore, clades II-V are highly unlikely to represent a risk to public health.

Of practical significance, the cryptic clades represent microorganisms that show worldwide distribution (Table 2.2) and have been readily identified as typical *E. coli* by expert microbiologists in the laboratory and by managers of water quality who use this organism to assess fecal pollution of surface waters. However, these organisms probably should not be considered *E. coli* and are highly unlikely to represent an environmental hazard, according to our analyses. These findings underscore the need to reevaluate coliform testing and the microbiologic dogma that the niche of enteric microbes, such as *E. coli*, is the mammalian intestinal tract.

REFERENCES

1. E. Stackebrandt *et al.*, Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International journal of systematic and evolutionary microbiology* **52**, 1043 (May, 2002).
2. R. Rossello-Mora, R. Amann, The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39 (Jan, 2001).
3. K. T. Konstantinidis, A. Ramette, J. M. Tiedje, The bacterial species definition in the genomic era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **361**, 1929 (Nov 29, 2006).
4. D. Gevers *et al.*, Opinion: Re-evaluating prokaryotic species. *Nature reviews. Microbiology* **3**, 733 (Sep, 2005).
5. C. Fraser, W. P. Hanage, B. G. Spratt, Recombination and the nature of bacterial speciation. *Science* **315**, 476 (Jan 26, 2007).
6. W. P. Hanage, C. Fraser, B. G. Spratt, Fuzzy species among recombinogenic bacteria. *BMC biology* **3**, 6 (2005).
7. M. Touchon *et al.*, Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5**, e1000344 (Jan, 2009).
8. R. Lan, P. R. Reeves, Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol* **8**, 396 (Sep, 2000).
9. S. Ishii, W. B. Ksoll, R. E. Hicks, M. J. Sadowsky, Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. *Appl Environ Microbiol* **72**, 612 (Jan, 2006).
10. S. T. Walk *et al.*, Cryptic lineages of the genus Escherichia. *Applied and environmental microbiology* **75**, 6534 (Oct, 2009).
11. A. P. H. Association, *Standard Methods for the Examination of Water and Wastewater*. (American Public Health Association, Washington, D.C., ed. 18, 1992).
12. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821 (May, 2008).
13. J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607 (Jun 15, 2001).
14. K. T. Konstantinidis, J. Braff, D. M. Karl, E. F. DeLong, Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Applied and environmental microbiology* **75**, 5345 (Aug, 2009).
15. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947 (Nov 1, 2007).
16. D. Bryant, V. Moulton, Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**, 255 (Feb, 2004).
17. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609 (Jul 1, 2006).

18. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725 (Sep, 1994).
19. Z. Zhang *et al.*, KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259 (Nov, 2006).
20. O. Zhaxybayeva, J. P. Gogarten, R. L. Charlebois, W. F. Doolittle, R. T. Papke, Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**, 1099 (Sep, 2006).
21. J. Goris *et al.*, DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology* **57**, 81 (Jan, 2007).
22. D. J. Ingle *et al.*, Biofilm formation by and thermal niche and virulence characteristics of Escherichia spp. *Appl Environ Microbiol* **77**, 2695 (Apr, 2011).
23. D. E. Chang *et al.*, Carbon nutrition of Escherichia coli in the mouse intestine. *Proc Natl Acad Sci U S A* **101**, 7427 (May 11, 2004).
24. J. Qin *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59 (Mar 4, 2010).
25. A. Reyes *et al.*, Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334 (Jul 15, 2010).
26. T. Wirth *et al.*, Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol Microbiol* **60**, 1136 (Jun, 2006).
27. J. G. Lawrence, H. Ochman, Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution* **44**, 383 (Apr, 1997).
28. A. C. Retchless, J. G. Lawrence, Temporal fragmentation of speciation in bacteria. *Science* **317**, 1093 (Aug 24, 2007).
29. A. C. Retchless, J. G. Lawrence, Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proc Natl Acad Sci U S A* **107**, 11453 (Jun 22).
30. J. M. Eppley, G. W. Tyson, W. M. Getz, J. F. Banfield, Genetic exchange across a species boundary in the archaeal genus ferroplasma. *Genetics* **177**, 407 (Sep, 2007).

ACKNOWLEDGEMENTS

The authors thank the personnel of the Genomics Facility at Michigan State University for their help with sequencing the Escherichia genomes. This project was supported in part by the National Science Foundation (Award: DEB 0516252) and in part with Federal funds from the National Institute of Allergy and Infectious Diseases National Institutes of Health, Department of Health and Human Services (Contract No.: HHSN2722009000018C).

CHAPTER 3

Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample

Parts of this chapter have been published in the article: C. Luo, D. Tsementzi, N. Kyripides, T. Read, and K. T. Konstantinidis. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE,* 2012, 7(2): e30087.

INTRODUCTION

From the human gastrointestinal tract to the abyss of oceans, whole-genome shotgun metagenomics is revolutionizing our understanding of the structure, diversity, and function of microbial communities (1-4). The next generation sequencing (NGS) technologies, such as the Roche 454, Illumina/Solexa, and, to a lesser extent, ABI SOLiD, have been cornerstones in this revolution (5-7). The high coverage of the indigenous communities provided by NGS has made it possible, for instance, to quantitatively assess the impact of diet on human gut microbiota (8) and the diversity of metabolic pathways within marine planktonic communities (9). NGS platforms produce millions of short sequence reads, which vary in length from tens of base pairs (bp) to ~800 bp. Even though the read length increases as the technologies advance, it is still far shorter than the desirable length (e.g., the average bacterial gene length is ~950 bp) or the length of the traditional Sanger sequencing. Therefore, a desirable, first step in the analysis of metagenomic data frequently is to assemble sequences into longer contigs and, ultimately, into complete genome sequences. Analyzing raw (not assembled) reads, as opposed to assembled contigs, is typically restricted to cases of too high community complexity for the sequence coverage obtained or in specialized studies that aim to determine in-situ abundance and/or population genetic structure and recombination (4, 10).

It is critical to assess the quality of the derived assemblies and several studies have recently attempted to evaluate the sequencing errors and artifacts specific to each NGS platform. For instance, it has been established that Roche 454 has a high error rate in homopolymer regions (i.e., three or more consecutive identical DNA bases) caused by accumulated light intensity variance (5, 11) and up to 15% of the resulting sequences are

often products of artificial (*in vitro*) amplification (12). Illumina does not appear to share these limitations but it has its own systematic base calling biases (13). Most importantly, different tiles of the sequencing plate tend to systematically produce reads of different quality (14), the 3' end of the resulting sequences tends to have higher sequencing error rate compared to the 5' end (15), and increased single-base errors have been observed in association with GGC motifs (16). Algorithms that detect and correct these errors are being developed and incorporated into existing data processing pipelines.

It should be noted, however, that most of the previous error estimates and sequencing biases have been determined based on simple DNA samples (e.g., a single viral genome) and thus, their relevance for complex community DNA samples remains to be evaluated. More importantly, it is currently unclear how the above limitations affect the quality of the gene and genome sequences assembled from complex DNA samples, and whether the technologies differ with respect to the diversity recovered from a sample due to their different chemistries and protocols. To provide new insights into these issues, we evaluated the two most frequently used platforms for microbial community metagenomic analysis, the Roche 454 FLX Titanium and the Illumina GA II, by comparing and contrasting assemblies obtained from the same community DNA sample.

MATERIALS AND METHODS

Sampling, DNA extraction, and sequencing

Samples were collected from Lake Lanier, Atlanta, GA, below the Browns Bridge in August 2009 and community DNA was extracted as described previously (17). The DNA sample was divided into two aliquots of equal volume. One aliquot was sequenced with the Roche 454 FLX Titanium sequencer (average read length, 450 bp) and the other one with llumina GA II (100 ! 100 bp pair-ended reads) available at the Emory University's Genomics Facility.

Raw (not assembled) read comparisons

We compared the reads from the Lanier.Illumina dataset against the Lanier.454 dataset to identify the fraction of reads shared between the two datasets. Shared reads were defined as those that mapped on reads of the other dataset using Bowtie with default settings (18). For comparing gene calling accuracy on unassembled reads, we employed FragGeneScan (19) to predict genes on Lanier.454 and Lanier.Illumina reads, using 454 1% error rate model and Illumina 0.5% error model, respectively. We extracted the predicted gene sequences from the reads and the corresponding amino acid sequences were searched against the genes of the reference assembly of the same dataset using BLAT (20). The matching gene of the assembly from the protein search using BLAT was compared to the gene matched by the raw read using Bowtie and the cases of agreements (matched genes), disagreements (mismatched genes) and "not match found" (BLAT search did not match a gene while Bowtie mapping did) were counted and reported in Figure 3.1B.

To estimate the errors associated with GGC motifs in Illumina reads described previously (21), we selected as reference sequences the Roche 454 reads that were covered by at least 10 Illumina reads per base, on average, in Bowtie mapping (~86.6 Mbp of reads in total). An in-house package written in Python and Perl identified disagreements between Illumina and the reference Roche 454 reads associated with GGC motifs using the rules described previously (21) and counted the number of errors (scripts available upon request).

Metagenome assembly and contig error calculation

Lanier.454 and Lanier.Illumina reads were trimmed at both 5' and 3' ends using a Phred quality score cutoff of 20. Sequences shorter than 200 bp and 50 bp (after trimming) were discarded, respectively. Newbler (version 2.0) was used to assemble Lanier.454 with parameters set at 100 bp for overlap length and 95% for nucleotide identity. For Lanier.Illumina, the SOAPdenovo (22) and Velvet (23) *de novo* assemblers were used to pre-assemble short reads into contigs using different K-mers. We performed six independent assemblies, using K=21, 25, 29 for the three SOAPdenovo runs and K=23, 27, 31 for the three Velvet runs. The resulting contigs were merged into one dataset, and Newbler was used to assemble this dataset into longer contigs, using the same parameters as in the assembly of Lanier.454 data. Our previous evaluation showed that our hybrid protocol outperforms other approaches for assembling metagenomic and genomic data (24). Individual reads were mapped against the assemble contigs using Bowtie with default settings to calculate average contig coverage. Protein-coding genes encoded in the assembled contigs were identified by the MetaGene pipeline (25). Contigs were defined as shared between the assemblies of the Lanier.454 and Lanier.Illumina data when they

shared at least 95% nucleotide sequence identity and overlapped by at least 80% of their length (for the shorter contig). The same cut-off was used to map raw reads on contigs. The 95% identity cut-off was used to accommodate the maximum sequencing error observed in raw reads of an isolate genome (about 5%); other cut-offs are not as appropriate as the one used above and were not evaluated.

Homopolymer error rate

We assessed homopolymer error rate in metagenomic data using two different strategies. First, by examining disagreements in gene sequences annotated on contigs larger than 500 bp and shared between the Lanier.454 and Lanier.Illumian assemblies. For this, Blastn (26) was employed to search all gene sequences annotated in the Lanier.454 assembly against those in the Lanier.Illumina assembly. Reciprocal best matches (RBMs), when overlapping by at least 500 bp and showing higher than 95% nucleotide identity, were identified and re-aligned using ClustalW2 (27). Homopolymer disagreements between the sequences in the alignment were identified and counted using a custom Perl script (the same approach was applied to the isolate genome data as well). Second, by directly assessing homopolymer error rate against reference genomes from GenBank that represented close relatives (average amino acid identity >70%) of the microorganisms sampled in the lake metagenome. To select appropriate genomes, we first identified the putative phylogenetic affiliation of each assembled contig (genus level) of the Lanier.454 and Lanier.Illumina datasets and ranked genera in terms of their abundance. Abundance was determined based on the number and coverage of the contigs, as described elsewhere (17). Six genomes that represented abundant genera in the lake metagenome were identified this way. The genomes were: Candidatus *Pelagibacter*

ubique HTCC1062 (*!-Proteobacteria*), *Opitutus terrae* PB901 (*Verrucomicrobia*), *Polaromonas sp.* JS666 (*"-Proteobacteria*), *Polynucleobacter necessarius* STIR1 (*"-Proteobacteria*), *Synechoccocus sp.* RCC307 (*Cyanobacteria*), and *Synechoccocus sp.* PCC6803 (*Cyanobacteria*). The protein-coding sequences of these genomes were compared against their homologs from the two assemblies to determine homopolymer errors, as described above for direct comparisons between the two assemblies. In order to account for possible biases introduced by uneven genus abundance and provide statistically robust estimates, we employed a Jackknifing resampling method. We sampled 50% of the total homopolymers at random and estimated homolopolymer rate in this subset. The results reported represented averages from 100 iterations. A similar strategy based on reference genome sequences was used to identify and count non-homopolymer-related, single-base errors.

Analysis of isolate genome data

Assemblies of isolate genome sequences (closed or high-draft) were downloaded from the NCBI RefSeq database (called "reference assemblies" for convenience); raw Illumina and Roche 454 sequencing reads were available through the Joint Genome Institute (JGI, www.jgi.doe.gov). To compare the quality of Illumina vs. Roche 454 contigs assembled from isolate genome data the following approach was followed: Illumina data for each genome was randomly sampled to form several technical replicate datasets, each of which provided about 100X coverage of the reference assembly, on average. Velvet was used to assemble each of these Illumina datasets with K-mer set at 31. Newbler was used to assemble Roche 454 replicate datasets (about 20X coverage on average), using 50 bp minimal alignment length and 95% alignment identity. The amount

of Illumina and Roche 454 input sequence data was chosen so that the ratio of the two was similar to the ratio in the metagenomic analysis (i.e., 2.5 Gb Illumina reads versus 500 Mb Roche 454 reads, or 5:1). Between 10 and 15 replicate datasets for each genome and each sequencing platform where analyzed; the exact number depended on the amount of total data available for each genome. Gene sequences from assembled contigs were extracted and ClustalW2 (27) was used to align the sequences against their orthologs from the reference assembly. The alignments were used to count frameshift errors, separately for each Illumina or Roche 454 dataset. We also measured the percent of the reference genome recovered in each assembly and the degree of chimerism of contigs as follows: A 500 bp window was used to slide through all assembled contig sequences longer than 500 bp with a step of 100 bp. This resulted in a set of 500 bp long sequence fragments, which were subsequently mapped onto the reference assembly using Blastn. The percent of the reference genome recovered by these fragments as a fraction of the total length of the reference assembly was calculated using a custom Perl script. Similarly, the reference assembly sequence was cut into 500 bp long fragments and mapped onto assembled contigs longer than 500 bp; the unmapped regions of latter contigs were identified as chimeric sequences and their total length (as a fraction of the total length of the contigs) represented the degree of chimerism for each dataset. Finally, we calculated the average single-base call error rate and gap opening error rate of individual reads of each dataset as follows: raw reads were trimmed using the same standards as described above and subsequently mapped onto the corresponding reference assembly from RefSeq. Base call errors and gap opening errors were identified as discrepancies between the read sequence and the reference assembly sequence using a custom Perl script.

Assessing the effect of assembly parameters

We used the isolate genome data to evaluate the effect of the parameters of the assembly on the quality of the contigs as follows: a series of assemblies were obtained for genomes of low (*Arcobacter nitrofigilis*, 28%), medium (*Fibrobacter succinogenes*, 48%), and high (*Cellulomonas flavigena*, 74%) G+C% content. For each genome, we varied the amount of sequences input to the assembly and the primary parameters of assembly (K-mer for SOAPdenovo and Velvet, and minimal alignment length for Newbler). Assemblies were obtained for each possible combination and the base call error and gap opening error of the resulting assemblies were determined as described for individual reads above.

RESULTS

Genetic diversity recovered in raw (not assembled) reads and assembled contigs

We obtained a total of 513 Mbp (~450 bp long reads) and 3,640 Mbp (100 bp pair-ended reads) Roche 454 and Illumina sequence data, respectively, from the same community DNA sample. The sample represented mostly the prokaryotic fraction of a planktonic microbial community from a temperate freshwater lake (Lake Lanier, Atlanta, GA); the complexity of the community sampled (in terms of species richness and evenness) was estimated to be comparable to that of surface oceanic communities, but lower than that of soil communities (17). For convenience, we called the two sequence data sets Lanier.Illumina and Lanier.454, respectively. We applied widely used protocols to assemble both sets of reads (see Material and Methods for details), which substantially collapsed the datasets into 57 Mbp and 46 Mbp of total unique sequences, respectively; 57.7% and 49.5% of the total reads in each dataset were singletons (i.e., remained unassembled), respectively. For this analysis, we considered only contigs longer than 500 bp because shorter contigs were usually characterized by low coverage and thus, were error-prone (Figure 3.2A, inset; and in (24)). We found that about 90% of the Roche 454 unique contig sequences overlapped with Illumina contig sequences (Figure 3.1C). It is also possible that the remaining ~10% of the contigs might had been associated with the community DNA of the original sample not splitting perfectly in half in the two aliquots sequenced and the fact that the diversity in the sample was not saturated by sequencing (e.g., estimates based on rarefaction curves indicated that we sampled about 80-85% of the total diversity in the Illumina data). Consistent with the results based on assembled contigs, we obtained ~90% of overlapping sequences (~80% when the overlapping

sequences were expressed as a fraction of the total Illumina dataset) between the two datasets when we performed a similar analysis using all raw (not assembled) reads (Figure 3.1A). These results revealed that, in general, the two platforms sampled the same fraction of the total diversity in the sample. We also estimated the abundance of each contig shared between the two assemblies by counting the number of reads composing the contig, which can be taken as a proxy of the abundance of the corresponding DNA sequence in the sample (28). We found a strong linear correlation (r^2 > 0.99) between the Roche 454 and Illumina data with this respect (Figure 3.1D). Therefore, the two platforms provided comparable *in-situ* abundances for the same genes or genomes.

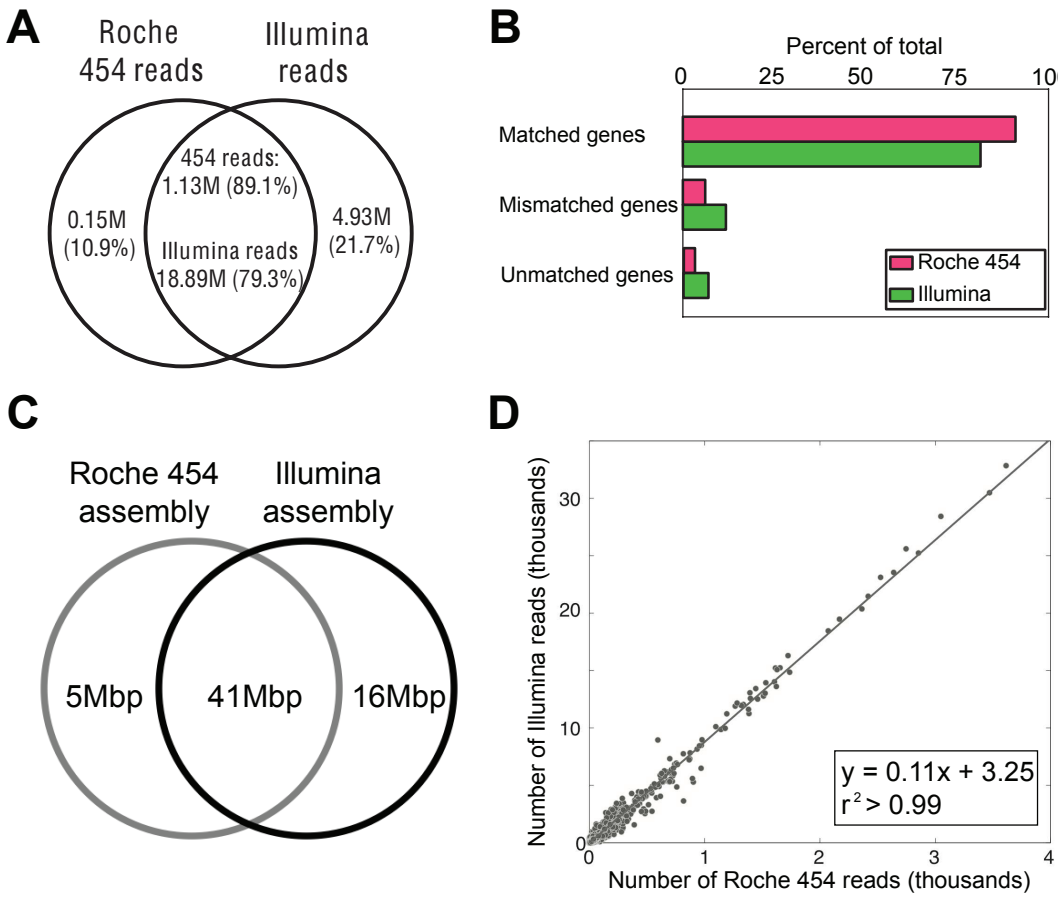


Figure 3.1. Genetic diversity and gene abundance in Roche 454 vs. Illumina data. (**A**) Venn diagram showing the extent of overlapping and platform-specific raw reads between the Lanier.454 and Lanier.Illumina datasets (without assembly). (**B**) Protein sequences annotated on raw (not assembled) reads matched genes in the reference assembly more frequently for the Roche 454 than the Illumina data. Conversely, protein sequences annotated on Illumina reads more frequently matched to the wrong protein sequence in the refernce assembly (mismatched genes) or did not match any reference gene (unmatched genes). (**C**) Assemblies were obtained from 502 Mbp of Roche 454 and 2,460 Mbp of Illumina data using established protocols. Venn diagram showing the extent of overlapping and platform-specific sequences of assembled contigs longer than 500bp. (**D**) Number of Roche 454 (*x* axis) and Illumina (*y* axis) reads mapping on the same contig shared between the two assemblies.

Illumina-specific unique contig sequences (16 Mbp) were more than three times as many as the Roche 454-specific ones (5 Mbp), and these additional contigs were attributed to the larger Illumina dataset rather than sequencing artifacts or errors. For instance, analysis of the assemblies of isolate genomes that were sequenced using both platforms (see also below), revealed that the extent of chimeric contigs, i.e., contigs that contained contaminating or in-vitro generated sequences, in the Illumina (or the Roche 454) assemblies was small, i.e., less than 0.2% of the total length of the assembled contigs, on average. Although low coverage contigs (i.e., 1 to 5X) are likely to contain a higher fraction of chimeric sequences than 0.2% according to our previous study (24) such contigs were rare in the results reported above, which included only contigs longer than 500 bp that showed, on average, 10X coverage or higher (only about 3% of the contigs showed less than 5X coveage; Figure 3.2A, inset). Illumina contigs were generally longer than Roche 454 contigs, i.e., the assembly N50 (the contig length that 50% of the entire assembly is contained in contigs no shorter than this length), was 1.6 Kb versus 1.2 Kb, respectively. Even when only a fraction of the total Illumina dataset was used in the analysis, which was comparable to the size of the Roche 454 dataset (i.e., 500 Mbp), the derived Illumina assemblies were comparable to the Roche 454 ones (e.g., N50 values were 990 bp for Illumina and 1193 bp for Roche 454, respectively; Figure 3.2B).

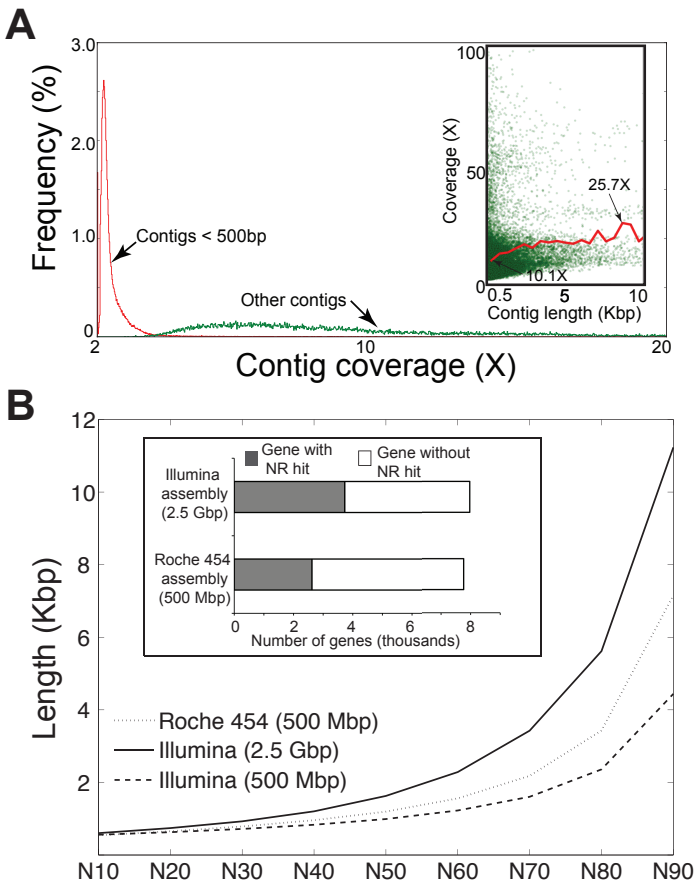


Figure 3.2. Average length and sequence accuracy comparisons of the Roche 454 and Illumina assembled contigs. (**A**) Length and coverage distribution of the contigs assembled from the Lanier.Illumina dataset. Note that contigs shorter than 500bp (red) were numerically more abundant than longer contigs (green) but were characterized by substantially lower coverage (inset). (**B**) Graph shows the comparison of the contig length of three assemblies plotted against the N statistic of the assembly [for instance, N40 (*x* axis) is equal to about 1 kbp (*y* axis), which means that (100-40=60) % of the entire assembly is contained in contigs no shorter than 1 Kbp]. Due to frameshifts caused primarily by homopolymer-associated errors in the derived consensus sequence of the contigs, genes from Roche 454 assembly had fewer complete matches in the NR database relatively to their Illumina counterparts (*inset*; results are based on a total of 72,709 gene sequences annotated on contigs that were shared between the two assemblies and were longer than 500bp).

Sequencing errors in assembled contigs

We evaluated the types and frequency of errors in assembled contigs from metagenomic data using both a comparative and a reference genome approach. In the former approach, we examined protein-coding sequences recovered in contigs longer than 500 bp that were shared between the Lanier.454 and Lanier.Illumina assemblies. We identified 0.4 million homopolymers (three identical consecutive nucleotide bases or more), 14 thousand of these (i.e., 3.3% of the total) disagreed on length between the two assemblies and resulted in alternative amino acid sequences for about 7% of the total 72,709 gene sequences evaluated. Among the latter genes, Roche 454 data appeared to have the wrong (artificial) sequence more often than Illumina data. For instance, searching all genes shared between the two assemblies against NCBI's Non Redundant (NR) protein database (Blastx) returned more complete matches with the Illumina than the Roche 454 data, regardless of the identity and e-value threshold used (14% more on average; Figure 3.2B, inset). These results were attributable to a higher number of (artificial) frameshifts, caused by homopolymer-associated or single base call errors, present in the Roche 454 versus the Illumina assembled sequences.

In the reference genome approach, genes annotated in Lanier.454 and Lanier.Illumina contigs were compared against their orthologs in publicly available genomes, and homopolymer errors were identified assuming the latter sequences contained no errors. We found that homopolymer errors affected 2.13% to 2.78% and 0.32% to 1.02% of the total genes evaluated for the Roche 454 and Illumina data, respectively (range was estimated from 100 replicates using Jackknife resampling), despite the fact that sequencing error in the raw reads of the two platforms was

comparable (~0.5% in our hands). These percentages were comparable to that reported above based on the comparative method (i.e., 3.3% of homopolymers disagreed between the two datasets, which includes both Roche 454- and Illumina-specific homopolymer errors). A closer investigation revealed that Roche 454 homopolymer sequence errors were biased toward A's and T's over C's and G's, and the errors were more frequent in homopolymers of higher length (Figure. 3.3). This pattern that was not as pronounced in the Illumina data, revealing that Illumina errors were (more) randomly distributed relative to Roche 454 errors (see for instance Figure 4, which is based on isolate genome data).

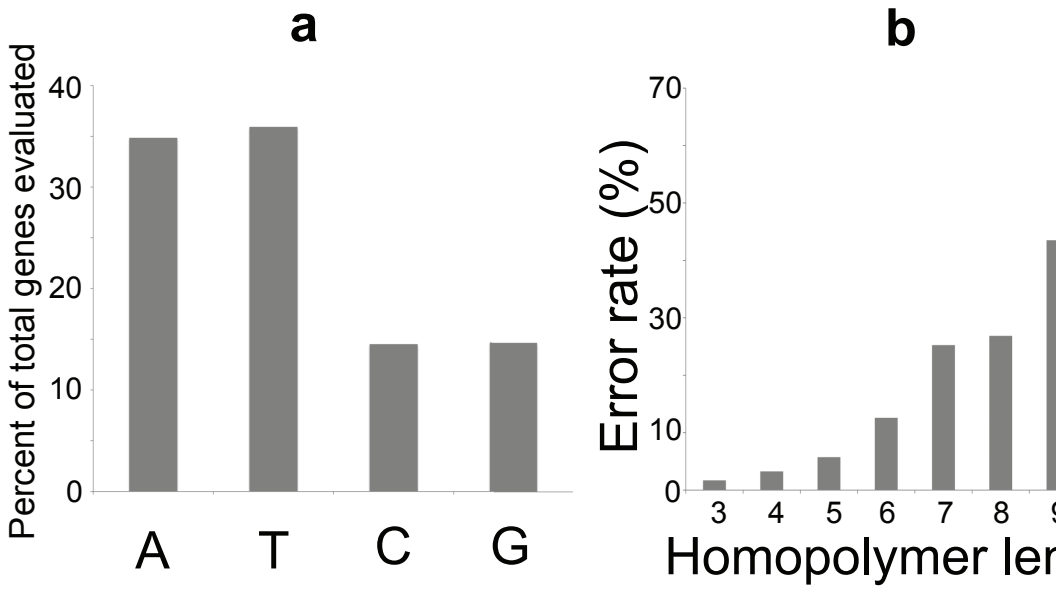


Figure 3.3. Characteristics of homopolymer-related sequence errors in Roche 454 metagenome assembly. (**A**) A's and T's contribute significantly more homopolymer errors than C's and G's. The average G+C% content of the metagenome was 47.4%; thus, our results are not simply attributable to higher abundance of A's and T's in the metagenome. (**B**) Error rate (as a

percentage of the total genes evaluated, *y* axis) increases as homopolymer length increases (*x* axis).

Single-base sequencing errors increased by about 2%, on average, for both platforms when non-homopolymer-associated errors were also taken into account. The frequency of single-base errors decreased with higher coverage of the corresponding contigs, i.e., the frequency dropped by about ten fold in contigs with 20X coverage relative to contigs with 2X coverage, and reached a plateau at about 20X coverage, i.e., we did not observed a significant difference in error frequency in contigs with higher than 20X coverage (see also our previous study, which defines standards on length and coverage for identifying error-prone Illumina contigs (24)). Given that the single-base error of individual reads was comparable between Lanier.454 and Lanier.Illumina (~0.5 %), our results reveal that the lower single-base error rate of Illumina contigs (i.e., ~3% vs. ~4.5% for Roche 454) is primarily due to the higher coverage obtained. Consistent with these interpretations, we found that the single-base error of Illumina contigs increased by about 0.07% when we reduced the average coverage of the Illumina contigs to the average coverage of the Roche 454 contigs (~8X). Obtaining, however, similar coverage to the Illumina data with the Roche 454 is economically unfavorable currently (see also Discussion below).

We also found that the systematic single-base errors associated with GGC-motifs in Illumina data reported recently (16) represented only a minor fraction of the non-homopolymer-associated errors (frequency estimated to 0.015% of total bases analyzed, which is consistent with the frequency reported in the original study). Hence, the great

majority of non-homopolymer-associated errors remain challenging to model and thus, correct. Finally, gene calling on individual reads (as opposed to assembled contigs) was found to be less error prone in Roche 454 reads relative to the Illumina reads, due mainly to the longer read length. For instance, protein sequences called on Roche 454 reads had more Blastp matches in reference genes from the Lanier.454 assembly compared to protein sequences from Illumina reads against the Illumina reference assembly by about 10%, on average (Figure 3.1B). Thus, Roche 454 is advantageous with respect to gene calling when working with unassembled reads.

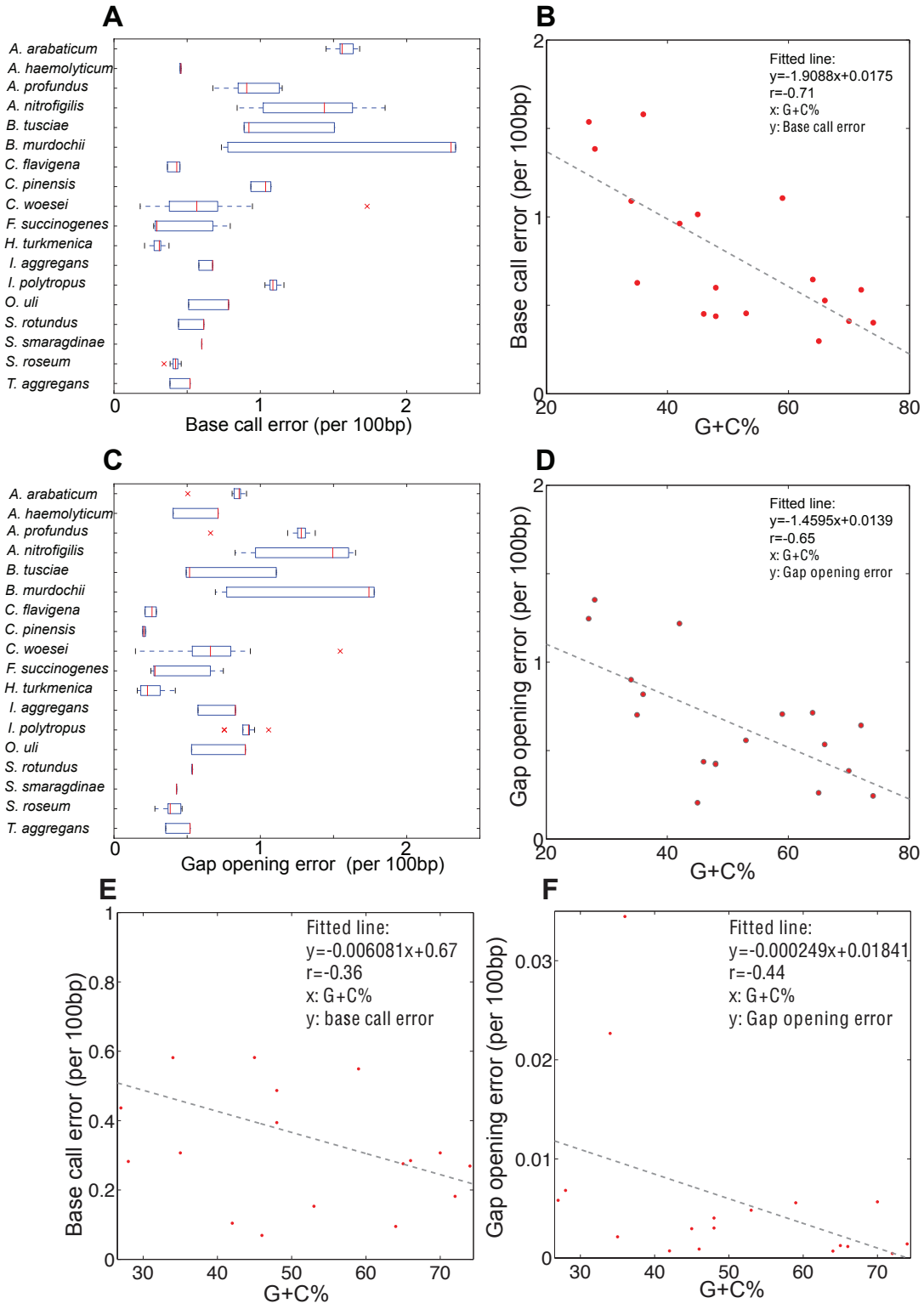

A
A. arabaticum
A. haemolyticum
A. profundus
A. nitrofigilis
B. tusciae
B. murdochii
C. flavigena
C. pinensis
C. woesei
F. succinogenes
H. turkmenica
I. aggregans
I. polytropus
O. uli
S. rotundus
S. smaragdinae
S. roseum
T. aggregans
0
1
2
Base call error (per 100bp)
B
Fitted line:
y=-1.9088x+0.0175
r=-0.71
x: G+C%
y: Base call error
Base call error (per 100bp)
0
1
2
20
40
60
80
G+C%
C
A. arabaticum
A. haemolyticum
A. profundus
A. nitrofigilis
B. tusciae
B. murdochii
C. flavigena
C. pinensis
C. woesei
F. succinogenes
H. turkmenica
I. aggregans
I. polytropus
O. uli
S. rotundus
S. smaragdinae
S. roseum
T. aggregans
0
1
2
Gap opening error (per 100bp)
D
Fitted line:
y=-1.4595x+0.0139
r=-0.65
x: G+C%
y: Gap opening error
Gap opening error (per 100bp)
0
1
2
20
40
60
80
G+C%
E
Fitted line:
y=-0.006081x+0.67
r=-0.36
x: G+C%
y: base call error
Base call error (per 100bp)
0
0.2
0.4
0.6
0.8
1
30
40
50
60
70
G+C%
F
Fitted line:
y=-0.000249x+0.01841
r=-0.44
x: G+C%
y: Gap opening error
Gap opening error (per 100bp)
0
0.01
0.02
0.03
30
40
50
60
70
G+C%

Figure 3.4. Roche 454 and Illumina GA II read sequence quality based on isolate genome data. Roche 454 sequencing quality is evaluated in panels A through D, which show: **(A)** base call error rate of individual reads (x-axis) for each genome evaluated (y-axis); **(B)** base call error rate (y-axis) plotted against the G+C% of the genome; **(C)** gap opening error rate of individual reads (x-axis) for each genome evaluated (y-axis); **(D)** gap opening error rate (y-axis) plotted against the G+C% of the genome. Illumina GA II sequencing quality is evaluated in panels E and F, which show: **(E)** base call error rate of individual reads plotted against the G+C% of the genome; and **(F)** gap opening error rate of individual reads plotted against the G+C% of the genome. Panels A and C represent the variation observed in reads from different (replicate) datasets of the same genome; red bars represent the median, the upper and lower box boundaries represent the upper and lower quartiles, and the upper and lower whiskers represent the largest and smallest observations. All 2D plots (panels B, D, E, and F) represent the arithmetic average of the medians of each dataset for the same genome; Illumina medians were identical among replicate datasets; that's why only one value is shown in panel E. The results show that Illumina sequence quality is affected less than that of Roche 454 by the G+C% content of the sequenced DNA (note the lower r-squared value and the slope in E). Thus, the results reported for Illumina based on the metagenome of Lake Lanier (47 G+C%) should be also applicable to metagenomes with different G+C% contents.

Analysis on isolate genome data

To validate our findings from metagenomics, we performed similar comparative analyses based on eighteen isolate genomes that were sequenced by both Illumina and

Roche 454 and showed a range of genome sizes and G+C% content (Table 2.1). Consistent with the metagenomic observations, we found that Roche 454 assemblies from genome data contained a significantly higher portion of frameshift errors compared to Illumina assemblies from the same genome. Specifically, in genomes of about 50% G+C% content (similar to the G+C% of the Lake Lanier metagenome, 47%), Roche 454 assemblies showed about 5% more frameshift errors than Illumina ones. This corroborated our estimated error rate in metagenomic data (i.e., Lanier.454 assembly was estimated to have 7% more frameshift sequences than Lanier.Illumina assembly, Figure 3.2). Noticeably, due to the inherent biases of the Roche 454 sequencing approach to produce more frameshifts in A and T rich DNA (Figure 3.3), low G+C% genomes may have 20% or more genes with frameshift errors compared to Illumina, which is not affected as much by the G+C% of the sequenced DNA (Figure 3.4). These findings call for special attention in cases where the sequenced DNA (e.g., community or isolate genome) is of low G+C%. Further, the single-base sequence error and gap opening error of individual reads were typically higher for the Roche 454 reads compared to the Illumina ones by 0.5% (i.e., 99% vs. 99.5%) and a factor of about 10, respectively, and despite the fact that reads were trimmed based on the same quality standard prior to the analysis (Figure 3.4). Similar gap opening errors were observed for the metagenomic reads of the two platforms while single-base accuracy was comparable among the two platforms, 99.34% vs. 99.46% for the Roche 454 and Illumina metagenomic reads, respectively, as also noted above. The slightly higher single-base accuracy of Roche 454 metagenomic reads relative to the average of the isolate genome reads is presumably due to the use of the latest, optimized Roche 454 protocol in the former case and slight

differences in the performance of the sequencers used. Finally, in all genomes analyzed, Illumina assemblies consistently recovered a larger percentage of the reference genome than Roche 454 assemblies (two tailed Whitney-Mann U test *p-value*=0.014; Figure 3.5), which was also consistent with our observations on the assembly N50 values of the metagenomes (Figure 3.2).

It should be mentioned that the RefSeq reference genome sequences (complete or high draft) used in our reference genome approach to detect errors in assembled contigs or genes were not independent of the Illumina and Roche 454 data used in our analysis but typically represented the consensus sequence assembled using all Illumina and Roche 454 data available for each genome (hybrid assembly). To eliminate the possibility that our results were biased by the selection of reference genomes, we used the reference assembly of *Fibrobacter succinogenes* subsp. *succinogenes* S85, which was sequenced independently by the Institute for Genomic Research (TIGR GenBank accession: CP002158.1; JGI GenBank accession: CP001792.1). We aligned the assembled contigs from 9 Illumina and 8 Roche 454 assemblies from JGI data for the same genome against the TIGR reference assembly and calculated base call error rate and gap open error rate as described above for JGI genomes. Although the use of the TIGR reference assembly resulted in slightly higher number of sequence errors for both Illumina and Roche 454 data, Illumina consistently showed a smaller number of sequencing errors and the relative error rate between the two platforms was similar, independent of the reference genome used (Figure 3.6). The higher sequence error rate observed for the TIGR reference genome might be due to the different strain of *F. succinogenes* sequenced or differences in the sequencing platform or the assembly protocol used between JGI and TIGR. Finally,

our evaluations showed that the choices of parameters and amount of input sequence of the assembly did not have any dramatic effect on the quality of the resulting contigs for both Illumina and Roche 454 assemblies (Figure 3.7); thus, the assembly step did not affect substantially downstream analyses and our conclusions.

Table 3.1. Isolate genomes used in the analysis.

Species	RefSeq	Genome size (Mbp)	GC (%)	% coding	Protein coding genes	Size of 454 data (Mbp)	Size of Illumina data (Mbp)
Acetohalobium arabaticum DSM 5501	NC_014378	2.47	36	85	2,282	603	2, 982
Arcanobacterium haemolyticum DSM 20595	NC_014248	1.99	53	86	1,731	252	2,871
Archaeoglobus profundus DSM 5631	NC_013741	1.56	42	91	1,819	600	4,479
Arcobacter nitrofigilis DSM 7299	NC_014166	3.19	28	92	3,126	504	6,087
Bacillus tusciae DSM 2912	NC_014098	3.38	59	84	3,150	124	2,285
Brachyspira murdochii DSM 12563	NC_014150	3.24	27	85	2,809	331	5,115
Cellulomona flavigena DSM 20109	NC_014151	4.12	74	90	3,678	563	3,394
Chitinophaga pinensis DSM 2588	NC_013132	9.13	45	88	7,192	161	3,769
Conexibacter woesei DSM 14684	NC_013739	6.36	72	93	5,914	303	2,578
Fibrobacter succinogenes substr. succinogenes S85	NC_013410	3.84	48	90	3,085	769	3,275
Haloterrigena turkmenica DSM 5511	NC_013743	3.89	65	84	3,739	205	2,581
Ignisphaera aggregans DSM 17230	NC_014471	1.88	35	86	1,930	258	2,739
Ilyobacter polytropus DSM 2926	NC_014632	2.95	34	85	1,889	210	5,854
	NC_014633 (plasmid)	0.96	34	83	992		
Olsenella uli DSM 7084	NC_014364	2.05	64	86	1,739	248	3,542
Segniliparus rotundus DSM 44985	NC_014168	3.16	66	90	3,006	245	3,170
Spirochaeta smaragdinae DSM 11293	NC_014363	4.63	48	92	4,219	509	3,306
Streptosporangium roseum DSM 43021	NC_013595	10.34	70	85	8,945	373	2,506
Thermosphaera aggregans DSM 11486	NC_014160	1.32	46	90	1,387	243	3,181

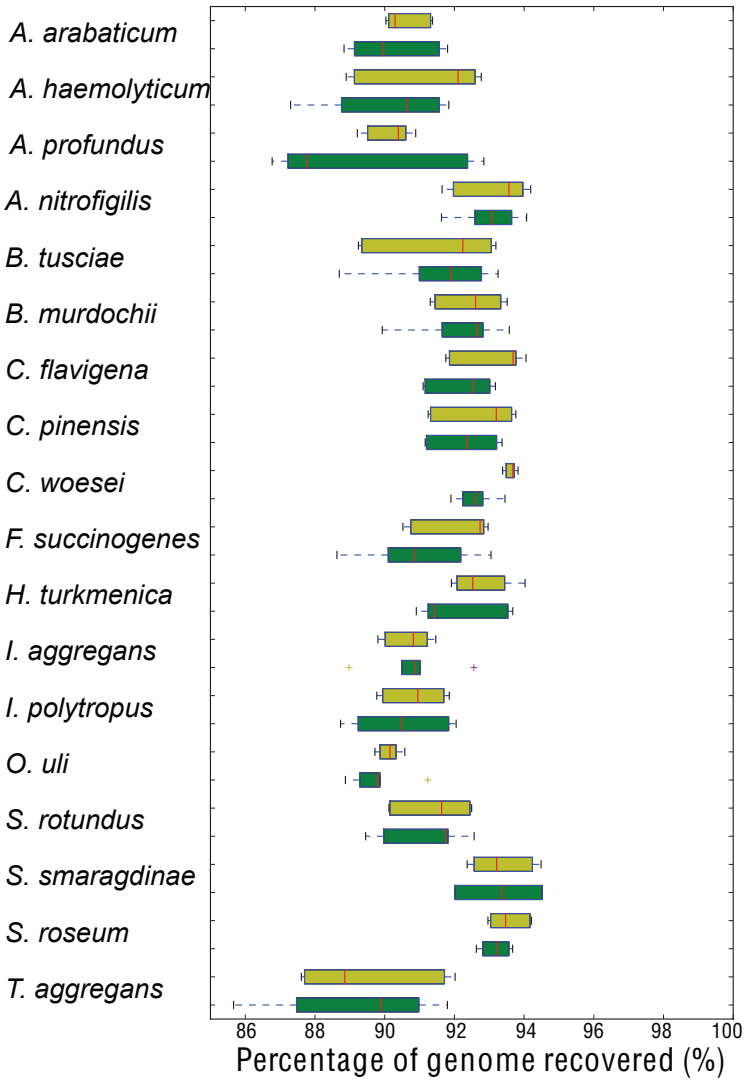


Figure 3.5. Percentage of reference genome recovered by Illumina (yellow) and Roche 454 (green) assemblies. Graph shows the variation observed in assemblies from different (replicate) datasets of the same genome; red bars represent the median, the upper and lower box boundaries represent the upper and lower quartiles, and the upper and lower whiskers represent the largest and smallest observations. Note that Illumina assemblies recover a significantly larger fraction of the reference genome than Roche 454 assemblies on average (two tailed Whitney-Mann U test p-value = 0.014), which is consistent with the results from the metagenomes (Figure 3.2). The results for the isolate genomes were based on Illumina input reads that were about 5 times more

compared to Roche 454 input reads to provide a ratio that was similar to the ratio in the metagenomic comparisons (5:1).

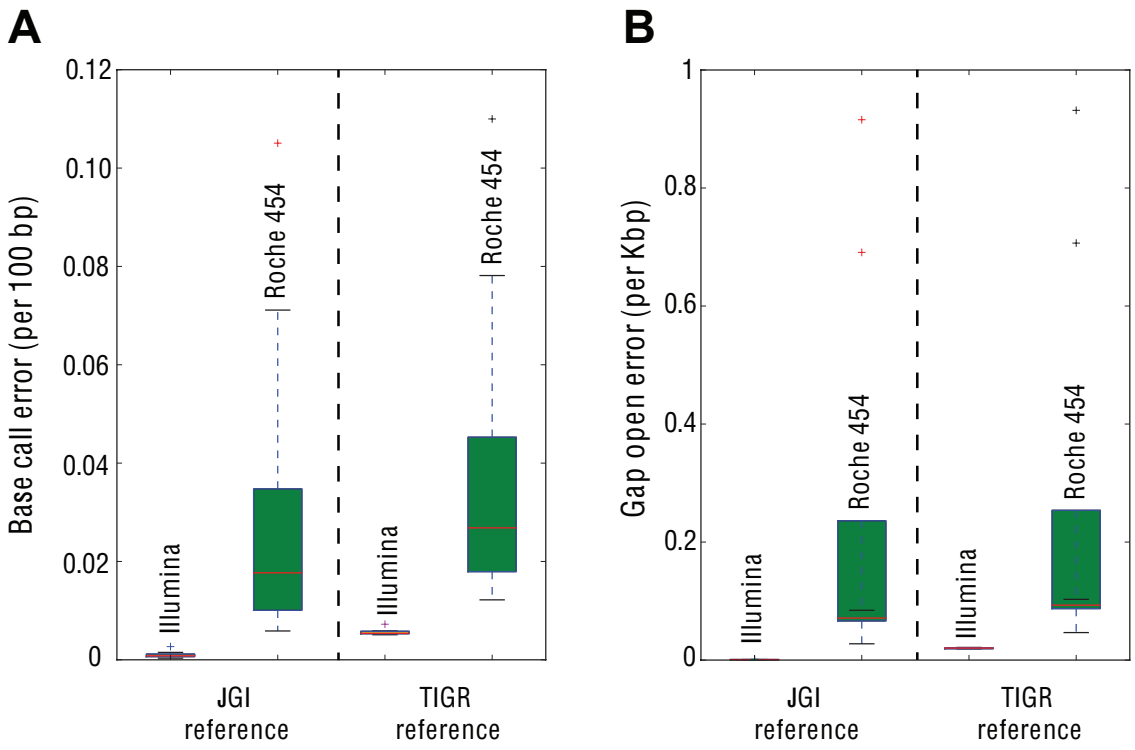
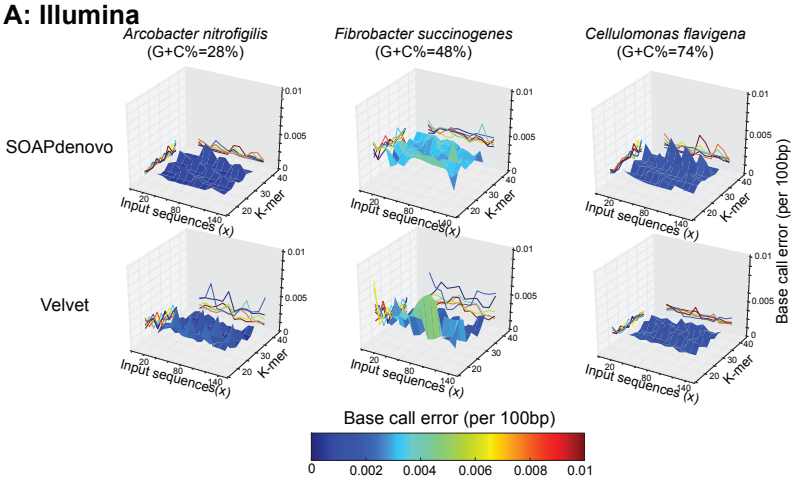
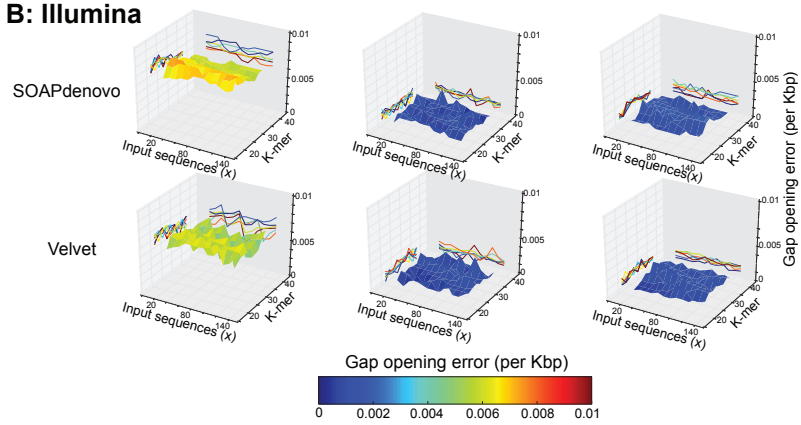
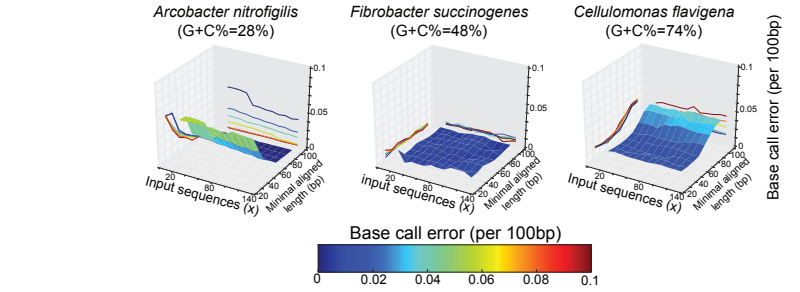


Figure 3.6. Comparisons of Illumina and Roche 454 assemblies against an independently sequenced reference genome. Nine Illumina and eight Roche 454 assemblies from independent replicate datasets of the *Fibrobacter succinogenes subsp. succinogenes* S85 genome sequenced at JGI were compared against the reference assembly from the JGI and TIGR genome projects of *Fibrobacter succinogenes subsp. succinogenes* S85. Graphs show the calculated base call error rate **(A)** and gap open error rate **(B)** for each comparison (figure key). Red bars represent the median, the upper and lower box boundaries represent the upper and lower quartiles, and the upper and lower whiskers represent the largest and smallest observations.


A: Illumina
Arcobacter nitrofigilis
(G+C%=28%)
Fibrobacter succinogenes
(G+C%=48%)
Cellulomonas flavigena
(G+C%=74%)
SOAPdenovo
Velvet
Input sequences (x)
K-mer
Base call error (per 100bp)
0
0.002
0.004
0.006
0.008
0.01


B: Illumina
SOAPdenovo
Velvet
Input sequences (x)
K-mer
Gap opening error (per Kbp)
0
0.002
0.004
0.006
0.008
0.01


C: Roche 454
Arcobacter nitrofigilis
(G+C%=28%)
Fibrobacter succinogenes
(G+C%=48%)
Cellulomonas flavigena
(G+C%=74%)
Input sequences (x)
Minimal aligned length (bp)
Base call error (per 100bp)
0
0.02
0.04
0.06
0.08
0.1

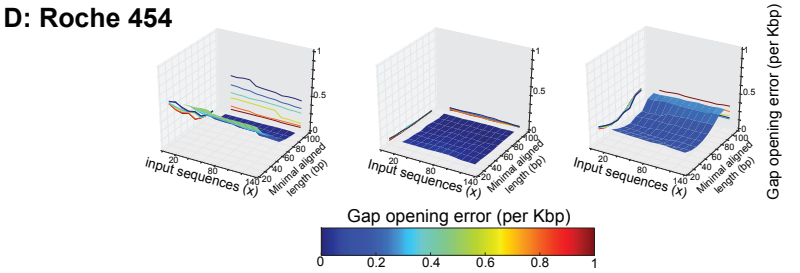

D: Roche 454
Input sequences (x)
Minimal aligned length (bp)
Gap opening error (per Kbp)
0
0.2
0.4
0.6
0.8
1

Figure 3.7. Dependence of the quality of assembled contigs on the parameters of the Illumina assembly. Assembly parameters (primary and secondary x-axes) were evaluated for low (*Arcobacter nitrofigilis*, 28%; left), medium (*Fibrobacter succinogenes*, 48%; middle), and high (*Cellulomonas flavigena*, 74%; right) G+C% genomes. For each genome, a 2D-grid assembly was performed, varying the size of input sequences (20X, 30X, 40X, …, 130X) and the K-mer (21, 23, 25, …, 37) of each of the assemblers used (SOAPdenovo and Velvet). The quality of the resulted contigs was examined in terms of base call error **(A)** and gap opening error **(B)**, which revealed that the combination of the parameters of the assembly did not have a dramatic effect on the quality of the contigs (see projected contours on x-z and y-z space). Similarly for the Roche 454 data, a 2D-grid assembly was performed, varying the size of input sequences (20X, 30X, 40X, …, 130X) and the minimal aligned length to merge contigs or reads (30bp, 40bp, …, 100bp) for Newbler. The quality of the resulted contigs was examined in terms of base call error **(C)** and gap opening error **(D)**, which revealed that the combination of the parameters of the assembly did not have a dramatic effect on the quality of the contigs except in the extreme values of the minimal aligned length (see projected contours on x-z and y-z space), which were avoided in our direct comparisons of Illumina versus Roche 454 assemblies.

DISCUSSION

Here we assessed the advantages and limitations of the Roche 454 and Illumina platforms for metagenomic studies based on the sequencing of the same community DNA sample. The two platforms agreed on over 90% of the assembled contigs and 80% of the unassembled reads as well as on the estimated gene and genome abundance in the sample (Figure 3.1). These findings suggest that NGS technologies represent reliable means for assessing quantitatively genetic diversity within natural communities. Moreover, Illumina yielded longer and more reliable contigs (e.g., fewer truncated genes due to frameshifts) despite the substantially shorter read length relatively to Roche 454 and the comparable average sequencing error in the raw reads of the two platforms (~0.5% in our hands; Figure 3.2B). Given also the cost savings (e.g., we obtained the Illumina data for about one fourth of the cost of the Roche 454 data), Illumina, and short-read sequencing in general, may represent appropriate methods for metagenomic studies. We also assessed quantitatively the errors in the consensus sequences of the derived assemblies. Roche 454 recovered 14% fewer complete genes than Illumina (Figure 3.2B, inset) and this was primarily attributable to a higher sequencing error rate associated with A- and T-rich homopolymers (Figure 3.3), which is in agreement with previous results (5, 11). These errors were not observed in the Illumina data, presumably due to the high sequence coverage that greatly facilitated the resolution of homopolymer ambiguities and less pronounced sequencing biases of Illumina (Figure 3.4). Nonetheless, about 1% of the total genes recovered in the Illumina assembly contained homopolymer-associated sequencing errors and this number increased to about 3% when non-homopolymer-associated errors were also taken into account (for contigs showing 10X coverage, on average). These results reveal the type and frequency of sequencing errors to expect when

performing NGS-enabled metagenomic studies. Although Illumina provided, in general, at least equivalent assemblies with Roche 454, there may be cases where Illumina might be inferior to Roche 454. For example, Roche 454 sequencing may be advantageous for resolving sequences with repetitive structures or palindromes or for metagenomic analysis based on unassembled reads, given the substantially longer read length.

Although our metagenomic analysis is based on a single community sample, we believe it is robust and informative. Our previous study (17) as well as those of others (29, 30) reported high reproducibility of Illumina-based and 454-based DNA sequencing of the same community sample, respectively. More importantly, most of our findings from metagenomic data were reproducible in data from isolate genomes, which were sequenced by both sequencing platforms and showed a range of G+C% content (Figures 3.4-6 and Table 3.1). Simulations with the isolate genome data also revealed that our conclusions were not affected substantially by the assembly protocols and amount of input data used (Figure 3.7). Some of our results (e.g., assembly N50 comparisons, Figure 3.2) should be independent of the NGS platform considered and be broadly applicable to short-read sequencing as well. Lastly, our preliminary evaluation indicates that the latest Illumina sequencer (Hi-Seq 2000) performs similar to Illumina GA-II in terms of read length and quality; hence, our results should be applicable to this sequencer as well.

NGS platforms continue to improve, while new major advancements in sequencing chemistries are on the horizon (31), creating a lot of excitement among microbial ecologists and engineers. The results presented here revealed the errors and limitations as well as the strengths in current metagenomics practice, and should

constitute useful guidelines for experiment design and analysis. Our work also provides a methodology for evaluating and comparing metagenomic data from NGS platforms.

REFERENCES

1. Nelson KE, *et al.* (A catalog of reference genomes from the human microbiome. *Science* 328(5981):994-999.
2. DeLong EF, *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311(5760):496-503.
3. Qin J, *et al.* (A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.
4. Konstantinidis KT, Braff J, Karl DM, & DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* 75(16):5345-5355.
5. Margulies M, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
6. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5(4):433-438.
7. Shendure J, *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309(5741):1728-1732.
8. De Filippo C, *et al.* (Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A* 107(33):14691-14696.
9. McCarren J, *et al.* (Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci U S A* 107(38):16420-16427.
10. Mackelprang R, *et al.* (Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*.
11. Quince C, *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6(9):639-641.
12. Gomez-Alvarez V, Teal TK, & Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3(11):1314-1317.
13. Erlich Y, Mitra PP, delaBastide M, McCombie WR, & Hannon GJ (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* 5(8):679-682.
14. Dolan PC & Denver DR (2008) TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics* 9:250.
15. Schroder J, Bailey J, Conway T, & Zobel J (Reference-free validation of short read data. *PLoS One* 5(9).
16. Nakamura K, *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic acids research* 39(13):e90.
17. Oh S, *et al.* (2011) Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of lake lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77(17):6000-6011.
18. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
19. Rho M, Tang H, & Ye Y (FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38(20):e191.

20. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome research* 12(4):656-664.
21. Nakamura K*, et al.* (Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39(13):e90.
22. Sun S*, et al.* (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research* 39(Database issue):D546-551.
23. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821-829.
24. Luo C, Tsementzi D, Kyrpides N, & Konstantinidis KT (2011) Individual genome assembly from complex community short-read metagenomic datasets. *The ISME Journal*:In press.
25. Noguchi H, Park J, & Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34(19):5623-5630.
26. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25(17):3389-3402.
27. Thompson JD, Gibson TJ, & Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2:Unit 2 3.
28. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* 2(10):1052-1065.
29. Rodriguez-Brito B*, et al.* (2010) Viral and microbial community dynamics in four aquatic environments. *ISME J* 4(6):739-751.
30. Stewart FJ, Ottesen EA, & DeLong EF (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* 4(7):896-907.
31. Eid J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133-138.

ACKNOWLEDGEMENTS

We would like to thank Chad Haase and Ryan Weil for their assistance. This research was supported, in part, by the U.S. Department of Energy (award DE-SC0004601). The Emory Genome Center acknowledges the Georgia Research Alliance and the Atlanta Clinical & Translational Sciences Institute for funding for major equipment purchases. DT acknowledges the support of the Onassis Scholarship Foundation.

CHAPTER 4

Individual genome assembly from complex community short-read metagenomic datasets

Parts of this chapter have been published in the article: C. Luo, D. Tsementzi, N. C. Kyrpides, and K. T. Konstantinidis. Individual genome assembly from complex community short-read metagenomic datasets. 2012, (6): 898-901.

INTRODUCTION

Next generation sequencing (NGS) technologies such as the Roche 454 and the Illumina/Solexa (1, 2) are revolutionizing the study of natural microbial communities (3-5). A major objective of metagenomic studies is to recover the genome sequence, complete or draft, of a genotype or species from a sample. Short-read (e.g., 50-100 bp) NGS technologies are becoming increasingly popular due to their high throughput, but it remains unclear whether these technologies can be used to robustly recover individual genomes from complex communities. Several recent studies have attempted to evaluate the sequencing errors and artifacts specific to each NGS platform (6-8); however, most of these studies have not assessed assembly quality and/or have employed simple DNA samples (e.g., single viral genomes) and thus, the relevance of their results for complex community samples remains to be evaluated. Moreover, the presence of closely related species in the sample may also complicate the assembly of a single genotype.

MATERIALS AND METHODS

Sampling, DNA extraction, and sequencing

Samples were collected from Lake Lanier, Atlanta, GA, below the Browns Bridge in August 2009. A horizontal sampler (Wildco Instruments) was used to collect samples of planktonic microbial communities at 5m depth. A total of 10 L of water was pre filtered through ~1.6 μm GF/A filters (Whatman) and cells were collected on 0.22 μm Sterivex filters (Millipore) using a peristaltic pump. DNA was extracted from the Sterivex as previously described (3), with minor modifications. Briefly, lysis buffer was added to the filters (50 mM Tris-HCl, 40 mM EDTA, and 0.75 M sucrose, 1mg/ml lysozyme, 150 mg/ml RNAse), followed by 30 min incubation at 37 ^{o}C. Lysates were further incubated at 55 ^{o}C for 2h after addition of 1% SDS and 10 mg/ml proteinase K. DNA extraction was performed with phenol and chloroform, and DNA was precipitated with ethanol at -20 ^{o}C overnight. DNA pellets were washed with 70% ethanol, and diluted in TE buffer. DNA extracted from four sterivex filters was combined and the resulting sample was sequenced with the llumina GA II (100 bp pair-ended reads) at the Emory University's Genomics Facility. For convenience, we called this dataset Lanier.Illumina. The insert size of this Illumina library was 150 bp (i.e., the two sister reads overlapped by ~50 bp). We also sequenced and evaluated larger insert size Illumina libraries from the Lake Lanier planktonic community, up to ~300 bp; we obtained similar results to the ones reported for the Lanier.Illumina dataset (data not shown but available from the authors upon request). The soil metagenome included in this study originated from a bulk soil sample from Norman (Oklahoma, USA), which was processed using the PowerSoil® DNA isolation kit from MO-BIO (Carlsbad, USA) and sequenced using a similar strategy and to the same coverage (~2.5 Gb of data) as the Lake Lanier sample. Its detailed analysis

will be presented elsewhere (data available from the authors upon request). The bioinformatic analyses described below for the Lanier.Illumina dataset were also applied to the soil dataset, when appropriate. The data of the cow rumen and human metagenomes (9, 10) were downloaded from NCBI (accession number: SRR094166) and EBI (accession number: ERR011333), respectively.

Assembly and gene calling

Lanier.Illumina reads were trimmed at both 5' and 3' ends using a Phred quality score cut-off of 20. Sequences shorter than 50 bp (after trimming) were discarded. The SOAPdenovo (11) and Velvet (12) *de novo* assemblers were used to pre-assemble short reads into contigs using different kmers (usually ranging from 21 to 31). We also evaluated other popular assemblers, such as ALLPATHS2 (13) and ABySS (14). The combination of Velvet and SOAPdenovo was chosen because of its overall higher accuracy, computational efficiency, and complementarities of the two assemblers. We performed six independent assemblies, using K=21, 25, 29 for the three SOAPdenovo runs and K=23, 27, 31 for the three Velvet runs. The resulting contigs were merged into one dataset, and Newbler (version 2.0) was used to assemble this dataset into longer contigs, with parameters set at 100 bp for overlap length and 95% for nucleotide identity. This hybrid protocol provided significantly longer contigs, of comparable or higher accuracy with the contigs of Velvet or SOAPdenovo for metagenome (Figure 4.1) and isolate genome data (data not shown). Different combinations of kmers, e.g., K=25, 31, 35 for Velvet, and K=29, 33, 37 for SOAPdenovo, did not provide substantially different assemblies with the hybrid protocol (data not shown).

Protein-coding genes encoded in the assembled contigs were identified by the MetaGene pipeline (15). The six *Escherichia* sp. genomes (Table 4.1) used in the *in silico* simulations were sequenced previously using the Illumina GA II platform at ~600X coverage (76 ! 76 bp pair-ended reads). The reads for each genome were first assembled into contigs by Velvet, using one fifth of the total data (i.e., ~100X coverage). The remaining reads (~500X coverage) were subsequently mapped onto the resulting scaffold to further improve base calling, using a consensus strategy (>80% agreement among overlapping reads). The assembled genomes were annotated using GeneMark (16). The genome sequences of the six *Escherichia* sp. strains have been described in more detail elsewhere (17).

Single and multiple genotype assembly simulations

To select appropriate reference genomes with abundant relatives in the Lake Lanier metagenome, Lanier.Illumina reads were searched against all fully sequenced genomes from NCBI (as of March, 2011) using blastn, and were assigned to the genus level based on best matches (minimum cut-off for a match: at least 50 bp alignment length and 85% nucleotide identity). Representative genomes of the genera that recruited the highest number of reads (Figure 4.4) were used as reference genomes. Reads of the reference genome(s) (Illumina reads for *Escherichia* sp. strains, and *in silico* generated reads for *Synechococcus* RCC307 and *Burkholderia ambifaria* MC40-6) were merged with (spiked in) the Lanier.Illumina dataset at different abundances (1X to 35X coverage for single genotype test, 5X to 35X for multiple genotype test) to form a series of *in silico* generated metagenomes of different abundance of the reference (target) genotype(s). Reads for *Synechococcus* RCC307 (GenBank accession: NC_009482) and *B. ambifaria* MC40-6 (GenBank accessions:

NC_010551, NC_010552 and NC_010557, the plasmid sequence was excluded) were generated *in silico* using a custom Perl script that simulated real Illumina data (i.e., the script reproduced the error rates on both 3'- and 5'-end, read quality, insertion length, *etc*). The same assembly protocol as described above for community metagenome was employed to assemble each *in silico* generated (mixed) metagenome. In order to eliminate possible biases introduced by different read length (100 bp vs. 76 bp for the Lake.Illumina and genome data, respectively), Lanier.Illumina reads were trimmed to 76 bp prior to the assembly. Subsequently, the assembled contigs were searched against the reference genome sequence to identify parts of the genome that were recovered in the assembly of the mixed metagenome. A sequence (e.g., a gene or a contig) was defined as "recovered" if it shared at least 80% of its length and 95% (for single genotype tests) or 90% nucleotide identity (for multiple genotype tests) with the corresponding sequence from the reference genome. The 80% length cut-off was employed to ensure that the same gene or contig, as opposed to just a conserved domain or a fragment of it, was recovered. The 95% identity cut-off (single genotype tests) was used to accommodate the maximum sequencing error observed in raw reads of an isolate genome (about 5%) and the sequence heterogeneity within populations frequently observed in nature (18); the 90% identity cut-off (multiple genotype simulations) was used to accommodate the sequence diversity among the genomes used (average genome-aggregate average nucleotide identity was about 95%) and the sequencing error. Other cut-offs are not as appropriate as the ones used above and were not evaluated. This way, non-target sequences, i.e., sequences that did not match or only matched the reference genome at a lower identity and thus, did not belong to the reference but were found in the same contig with target sequences (chimeric contigs), were also identified. Nucleotide mismatches and

frameshifts between recovered (target) gene sequences and their orthologous gene sequences of the reference genome were detected based on ClustalW2 alignments and counted using custom Perl scripts.

RESULTS AND DISCUSSION

Assembling genomes from metagenomes

To provide quantitative insights into the issues above, we generated a series of *in silico* metagenomes by spiking reference genome reads into a background metagenome (Lanier.Illumina) and compared the derived assembly against the assembly of the reference genome from the genome reads alone (Figure 4.1). For this analysis, we used the *Escherichia* sp. strain TW10509, whose genome sequence we described previously (19) and which has no close relatives in the Lake Lanier sample (Figure 4.2), as reference. The Lanier.Illumina dataset was described in detail elsewhere (20), originated from a freshwater planktonic community sample from Lake Lanier, Atlanta, GA, and represents a total of 3,640 Mbp sequence data (100 bp pair-ended reads; average G+C content ~50%) obtained using the Illumina GA-II sequencer. The community complexity in the sequenced sample (in terms of species richness and evenness) was comparable to that of previously characterized open ocean communities (20).

Table 4.1. The *Escherichia* sp. strains used to construct the *in silico* generated metagenomes. The genome sequences of these strains were described previously (17).

Strain	Lineage	Genome size (Mbp)	GC%	Strain source
TW10509	Clade I	5.19	50.29	Human feces
TW08933	*E.albertii*	4.51	49.74	Human feces
TW09276	Clade III	4.47	50.69	Freshwater
TW09231	Clade III	4.73	50.69	Freshwater
TW14182	Clade IV	4.68	50.43	Freshwater
TW15818	*E. albertii*	4.73	49.57	Poultry feces

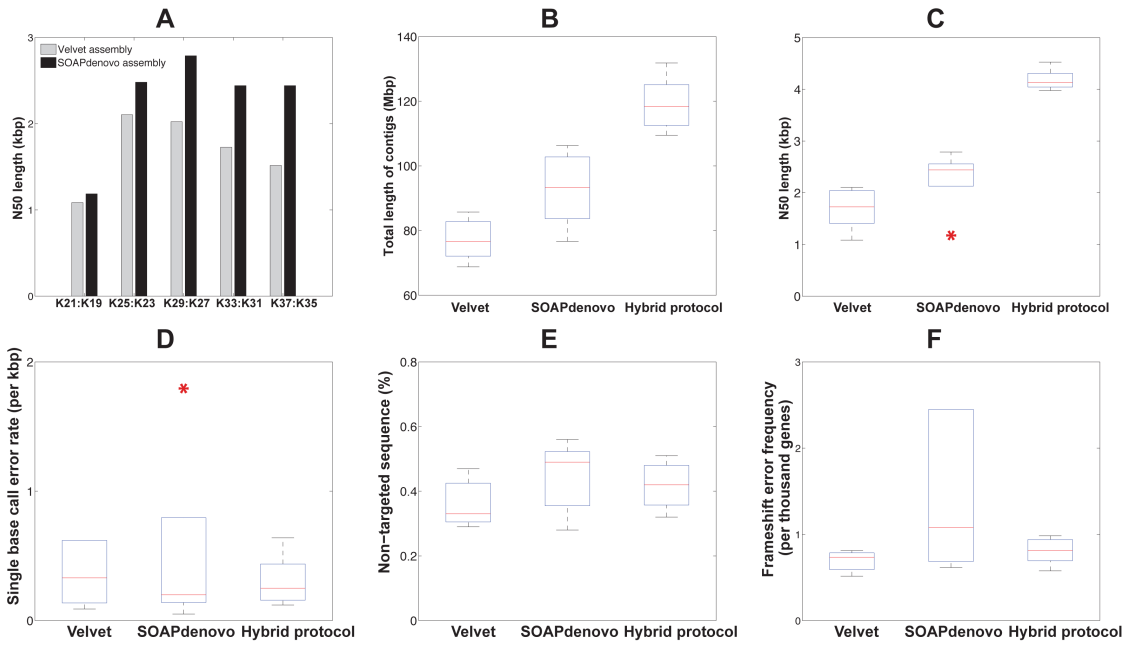


Figure 4.1. Comparisons of assemblies obtained using Velvet, SOAPdenovo, and the hybrid protocol developed in this study. An *in silico* synthetic metagenome was created from 100 randomly selected, fully sequenced bacterial genomes available in NCBI (plasmids excluded), using a custom Perl script (available from the authors upon request). The metagenome was composed of 76 bp pair-ended reads, which were characterized by 0.5% sequencing error and an average 300 bp long insert (insert sizes were normally distributed around the average with a standard deviation of 50 bp), to simulate a real metagenome. Velvet and SOAPdenovo were individually employed to assemble this metagenome with K=19, 23, 27, 31, 35 and K=21, 25, 29, 33, 37, respectively (kmer sets were as shown on the x-axis of Panel A). A hybrid protocol was also used to assemble the resulting contigs from three Velvet and three SOAPdenovo individual runs with different kmer as described in the Materials and Methods section. N50 of Velvet and SOAPdenovo assemblies varied, depending on the kmer sizes (**Panel A**; N50 values were grouped by kmer size); the corresponding assemblies of the hybrid protocol showed much smaller variation (data not shown). The hybrid protocol yielded substantially more assembled sequences compared to the Velvet and SOAPdenovo assemblies (**B**), and longer contigs (**C**), with

comparable single base call error (**D**), fraction of non-targeted (chimeric) sequence (**E**), and frameshift error frequency (**F**). Red bars represent the median, the upper and lower box boundaries represent the upper and lower quartiles, and the upper and lower whiskers represent the largest and smallest observations. Outliers are represented by red asterisks.

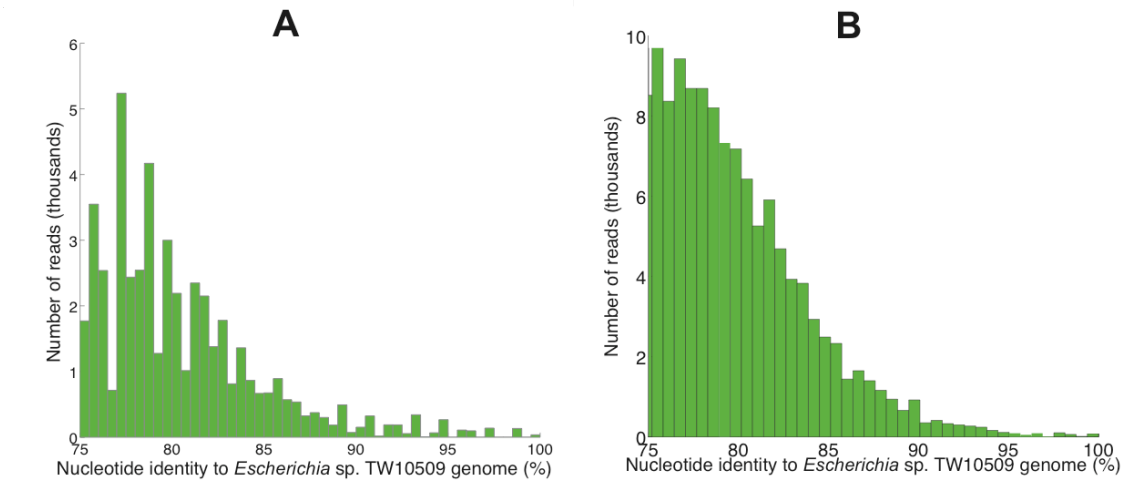


Fig. 4.2. The reference genome (TW10509) used in assembly simulations has no close relatives in the Lanier.Illumina (A) or the soil metagenome (B). Blastn was used to search all reads of the two metagenomes against the reference genome. Only matches with nucleotide identity higher than 75% are shown (~48,000 reads or ~0.2% of all total reads for the Lanier.Illumina dataset). The coverage plot was constructed as described previously (18) and clearly shows that there are no close relatives to TW10509 in the Lanier.Illumina or soil metagenomes (denoted by the shortage of reads with higher than 90% identity to the reference sequence). Even the few reads showing 80-90% identity to the reference genome encode highly conserved genes such as rRNA operon genes and thus, presumably originate from non-*E. coli* organisms.

We varied the reference genome abundance, measured by the average coverage of the genome in the final *in-silico* generated metagenome, more than thirty fold (i.e., 1X to 35X). As expected, the fraction of the reference genome recovered increased exponentially in the low coverage range and reached a plateau at about 20X coverage (Figure 4.3D). We also observed that greater than 20X coverage did not improve the recovery of the target genome substantially; thus, obtaining greater coverage is not recommended (unless a different library insert size is used for closing purposes). Surprisingly, more than 10% of the total assembled contigs that belonged to the reference genome (i.e., contained target sequences) were contaminated by non-target sequences at low coverage (1X), and this portion decreased to ~0.2% when coverage exceeded 15X (Figure 4.3C). Similar results were obtained when the reference genome represented an organism with close relatives in Lanier.Illumina (Figs. 4.4-9), albeit the sequences of the relatives generally had a positive effect on the quality of the derived assemblies (Fig. 4.6-7). We also quantitatively assessed the errors in the consensus sequences of the derived assemblies. About 1% of the total genes recovered in the Illumina assembly at 15X coverage contained homopolymer-associated sequencing errors (i.e., three or more consecutive identical DNA bases), resulting in truncated protein sequences or frameshifts. This number increased to about 3% when non-homopolymer-associated errors were also taken into account. Preliminary analyses revealed that most of the findings reported above were also applicable to a more complex soil metagenome, originating from a temperate (bulk) soil sample (Figure. 4.3B-C), although the average length of the assembled contigs of the reference genome was consistently shorter in the

soil spiked-in dataset (Fig. 4.3A) due to the higher complexity of the soil community (Fig. 4.10).

Table 4.2. Summary statistics of individual genome assembly from the Lake Lanier metagenome. The values shown are based on the single genotype spike in experiments described in the main text and should be applicable to other metagenomes of similar complexity to that of the Lake Lanier one.

Quality	Target genotype abundance in the sample			
	5X	10X	15X	20X
Genome recovery (%)	~50	~90	>95	>95
Gene recovery (%)	~80	~90	>95	>95
Base call error (%)	0.5	0.2	0.1	0.01
Frameshift error (per gene, %	3~5	1~2	0.5~1	<0.5
Chimerism (in genes, %)	~5	1~2	1	<1

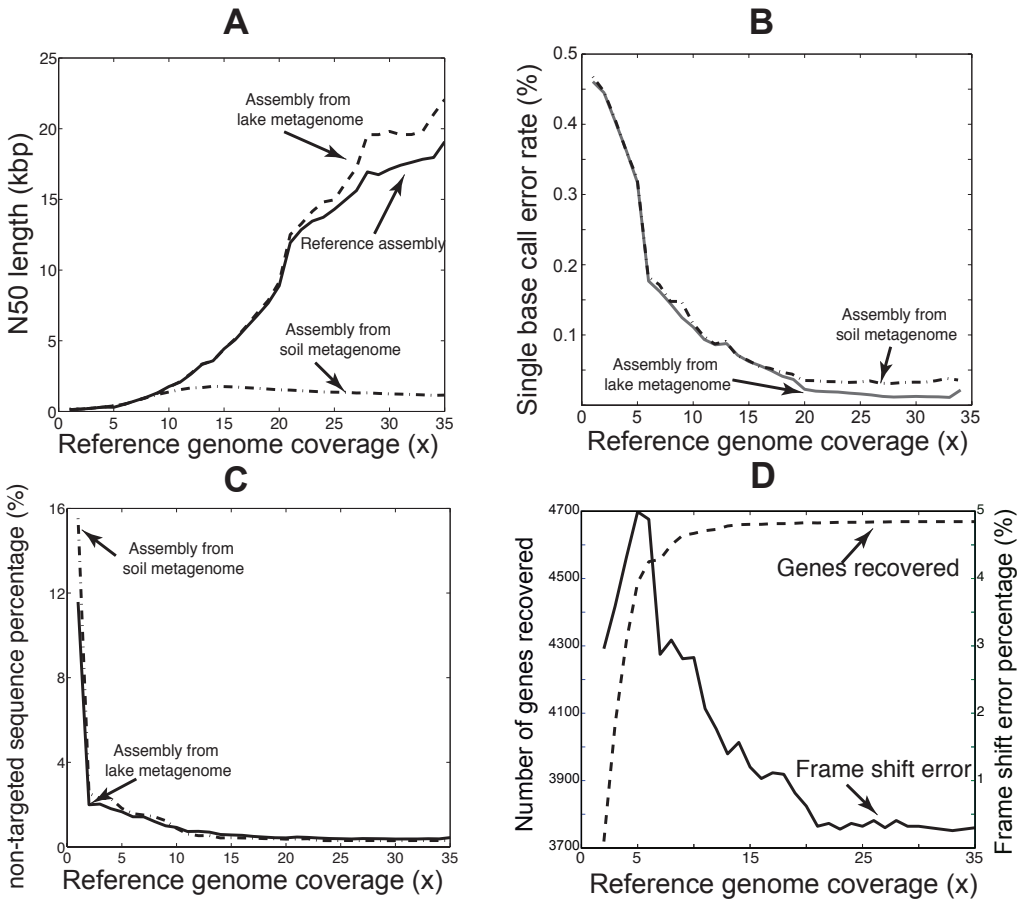


Figure 4.3. Sequence errors and artifacts in assembled contigs of a target genotype from a complex metagenome. The assembly of a reference genome (*Escherichia* sp. TW10509) based solely on its own reads (reference assembly) was compared to the assembly of the genome from the *in silico* metagenome, which was composed of Lanier.Illumina spiked in with reads of the reference genome. (**A**) Comparison of N50, i.e., the contig length that 50% of the entire assembly is contained in contigs no shorter than this length, between the latter and the reference assemblies over different reference genome coverage (abundance). (**B**) Single base call error rate decreased dramatically as reference genome abundance in the metagenome increased and reached a plateau at about 20X coverage. (**C**) At low coverage, contigs from the metagenome assembly had a substantial portion of non-targeted (chimeric) sequences. (**D**) Frequency of frameshift errors as a function of the reference genome abundance. Results from similar analyses using a higher-complexity (Figure 4.11) soil metagenome of similar size to the Lanier.Illumina metagenome are also shown for comparison.

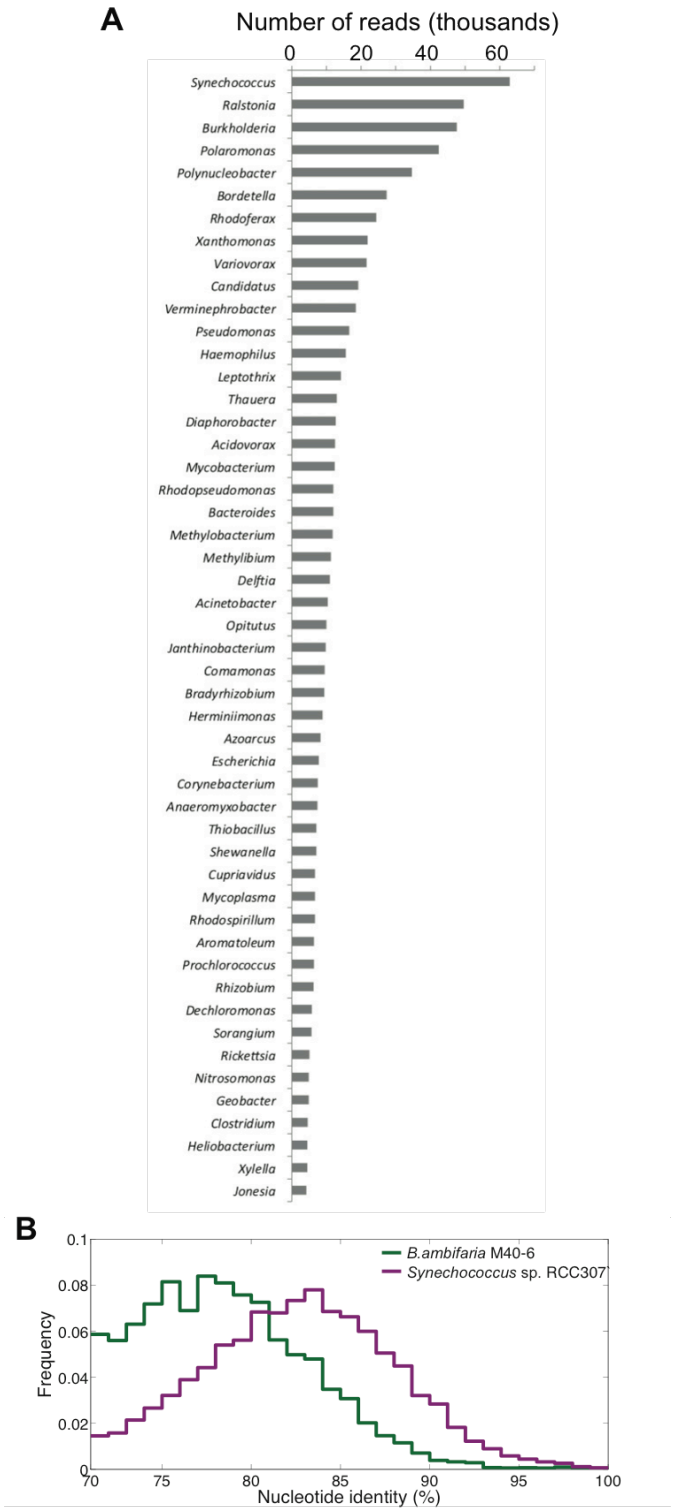


Figure 4.4. Bacterial genera present in Lake Lanier metagenome (A) and relatedness to *Synechococcus* and *Burkholderia* reference genomes (B). (Panel A) The Illumina reads were assigned to a genus based on blastn best match searches against all available fully sequenced

genomes (as of March, 2011). **(Panel B)** Assembled contigs of the two most abundant genera, *Burkholderia* and *Synechococcus*, were selected and the corresponding reference genomes (*B. ambifaria* MC40-6 and *Synechococcus* sp. RCC307, respectively) were fragmented into 100 bp pieces using sliding windows with a 10 bp step width. These 100 bp fragments were mapped on the contigs. The distributions of the nucleotide identities between the reference genome fragments and the contigs reveal that the reference genomes used in this study have relatives of varied genetic relatedness in the metagenome. *Escherichia* sp. TW10509 has no relatives (Figure 4.2); *B. ambifaria* MC40-6 has moderately related relatives (~75% ANI); and *Synechococcus* sp. RCC307 has close relatives (~85% ANI).

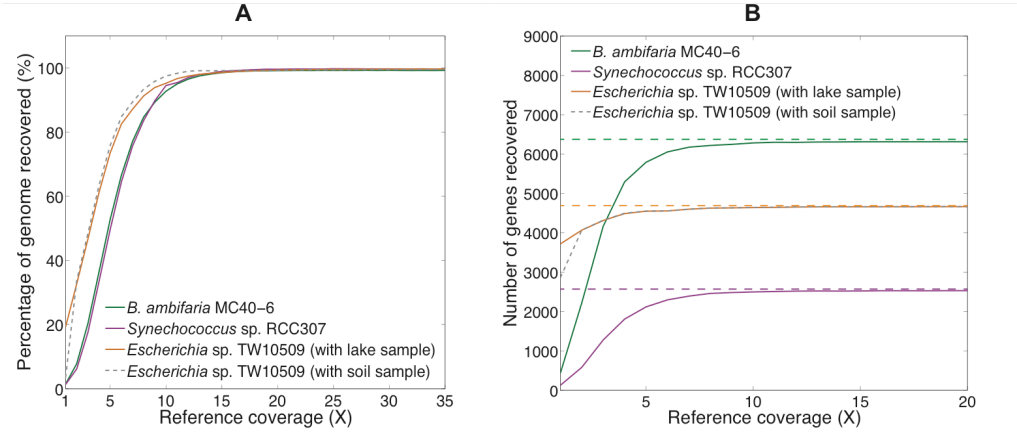


Figure 4.5. Recovery of the reference genome sequence (A) and number of genes (B) as a function of the abundance of the genome in the metagenome. The fraction of the reference genome sequence and number of gene sequences assembled from the *in silico* generated (spiked-in) metagenomes increased exponentially at low genome coverage and reached a plateau at about 15X coverage, for all three reference genomes used in the analysis. Horizontal dashed lines represent the total number of genes in each reference genome (panel B). The small difference in the slope between *Escherichia* sp. TW10509 and *B. ambifaria* MC40-6 or *Synechococcus* sp.

RCC307 is probably due to real Illumina data (*Escherichia* sp. TW10509) vs. *in-silico* generated (simulated) Illumina data (*B. ambifaria* MC40-6 or *Synechococcus* sp. RCC307).

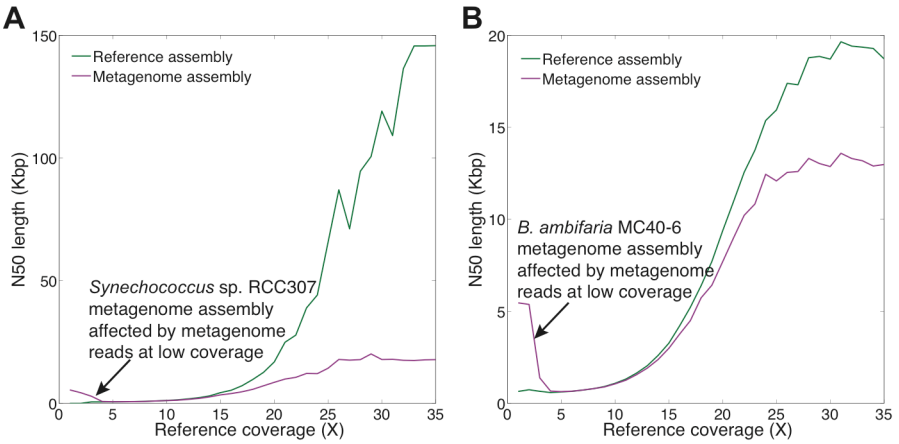


Figure 4.6. Assembly N50 of *Synechococcus* sp. RCC307 (A) and *Burkholderia ambifaria* MC40-6 (B) genomes as a function of coverage in metagenomic vs. genomic data. Note that assembly N50 is always larger in genome data alone (reference assembly) than genome data spiked in the Lanier.Illumina metagenome (metagenome assembly) except in the low coverage range, where N50 of the metagenome is larger due to the reads of the close relatives present in the metagenome. As coverage increases, however, the reads of the reference genome spiked-in the metagenome outweigh the reads of the natural population in the assembled consensus sequence and thus, N50 drops temporarily (5-10X coverage range) before it starts increasing steadily due to the higher abundance of spiked-in reference reads. This pattern at low coverage was not observed for the *Escherichia* sp. TW10509 reference genome (Figure 4.3A) due to lack of relatives in the metagenome.

Analysis of frameshift error

We also observed a non-monotonic relationship between frameshift frequency and the level of coverage of the target genome (Figure 4.4D and Figure 4.7). A possible explanation for this observation is that at low coverage, easy-to-assemble genes were recovered first, with low frameshift error; subsequently, difficult-to-assemble genes (e.g., due to repetitive sequence or conflicting reads covering the genes) were recovered. The latter genes tended to contribute more frameshifts than the average, resulting in a temporal increase in frameshift error as coverage increased. At higher coverage, all genes were covered by many reads, so the consensus sequence had fewer frameshifts overall. Such a non-monotonic curve was not observed for *Synechococcus* RCC307 and *Burkholderia ambifaria* MC40-6, presumably due to the additional coverage provided by the reads of relatives in the metagenome (equivalent to ~5X coverage). Therefore, if avoiding frameshifts is important, it is critical to obtain greater than 5X coverage of the target genome(s), which corresponded to the highest frequency of frameshifts in our study.

Reference genomes with relatives in the metagenome

In addition to evaluating genome assemblies from a complex metagenome using *Escherichia* sp. TW10509 genome sequence as a reference (Figure 4.3), which had no close relatives in the Lake Lanier sample, we analyzed reference genomes that had relatives in the Lake Lanier metagenome. We used *Burkholderia ambifaria* MC40-6 (heterotrophic *b-Proterobacterium*, 7.64 Mb genome size) and *Synechococcus* sp. RCC307 (photosynthetic *Cyanobacterium*, 2.2 Mb genome size) whose relatives in the metagenome showed ~75% and ~85% ANI (average nucleotide identity) to the reference

genome, respectively (Figure 4.4). In both of the cases, more than 95% of the reference genome was recovered at about 20X coverage, similar to the results reported for *Escherichia* sp. TW10509 (e.g., Fig. S4), although the quality of the assemblies, assessed, for instance, by the frequency of frameshift errors, was noticeably better in the reference genomes with relatives in the sample (Fig. S5-S7). The latter results were presumably attributed to the additional sequences originating from the relatives present in the community metagenome, which were included in the calculation of the assembled consensus sequences. In other words, the consensus sequence had higher coverage in the case of reference genomes with relatives compared to reference genomes with no relatives for the same amount of reference genome reads (coverage) spiked in the *in silico* metagenome, which improved the quality of the consensus sequence.

We were also able to recover sequence discrete populations with both *Synechococcus* and *Burkholderia* reference genomes in the multiple genotype analyses, similar to the results we obtained with *Escherichia* genomes (data not shown). These findings reveal that the results reported in Figure 2 are applicable to three different levels, i.e., no relatives in the sample; relatives that are moderately related (~75% ANI, *Burkholderia* case) and closely related relatives (~85% ANI, *Synechococcus* case). Simulations, which included *Salmonella* (80% ANI to *E. coli*) and *Yersinia* (70% ANI to *E. coli*) in addition to *E. coli* genomes spiked in a metagenome, showed that only when relatives show the whole gradient of nucleotide identities from 70% to 95% ANI to the reference genome, were we unable to retrieve sequence discrete populations. Finally, we did not assess the multiple genotype analyses with a higher number of reference (spiked in) genomes because our goal was to evaluate whether or not NGS can reveal intra-

population genetic structure. Performing the analysis with, for instance, 50 as opposed to just 5 genotypes (our study), should not differentiate our conclusions substantially as long as the additional genomes show similar genetic relatedness to the genomes used (i.e., 90-100% ANI). In all natural samples analyzed today (with the probable exception of soil samples, which have not been fully investigated yet), there are thousands of genomes comprising a population but these genomes are typically derived clonally from only one or a few distinct genotypes. Thus, our *in silico* experiments were designed to represent natural populations as closely as possible.

By spiking reads of a reference genome into a real metagenome dataset, we estimated that a typical bacterial genome (such as an *E. coli* genome) must have at least 15X sequence coverage (20X recommended) in order to recover more than 95% of its genome using Illumina short read sequencing (Figure 4.3). This is slightly higher than the coverage required to obtain the same assembly quality based on genome data alone, i.e., 10X coverage [Figure 4.3; and in (21)]. The difference is presumably due to the unavoidable interference of background reads in the metagenome case. In other words, reads with high sequencing errors or non-target (i.e., not derived from the reference genome) reads diverted the assembler to incorrect paths, which resulted in contig extension failure; such events were apparently less frequent in the assembly of the genome data compared to the metagenome data and in the Lake Lanier compared to the soil metagenome data (Figure 4.3A).

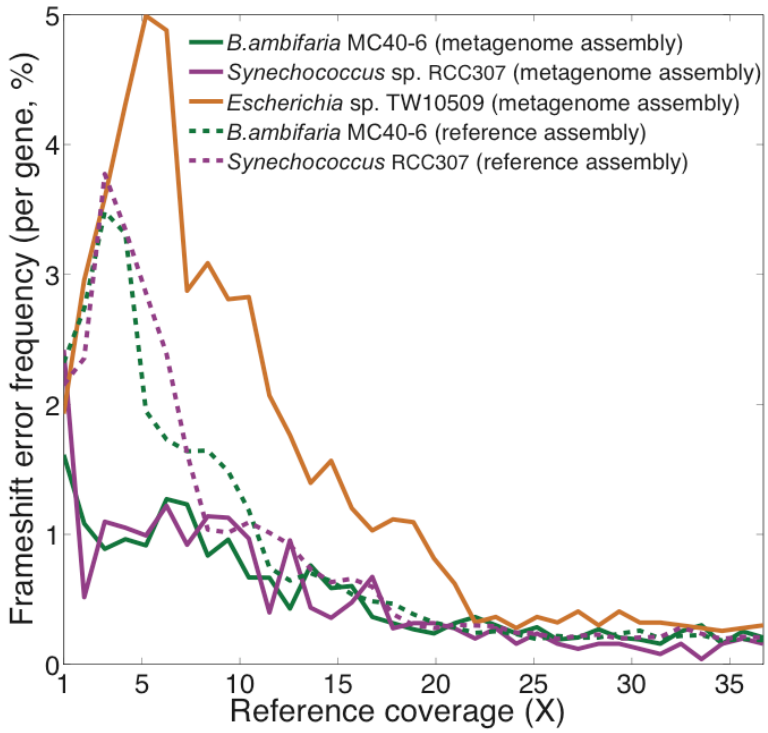


Figure 4.7. Frameshift error frequency with increasing genome coverage. Note that frameshift error frequency shows a non-monotonic behavior in Lanier.Illumina metagenome assemblies in the case of *Escherichia* sp. TW10509 but not with *Burkholderia* or *Synechococcus*. This is attributable to sequences of relatives in the metagenome, which improve the derived consensus (see also supplementary discussion). When these sequences of relatives were removed from the analysis prior to the assembly step, the monotonic behavior was observed for *Burkholderia* and *Synechococcus* as well (dashed lines).

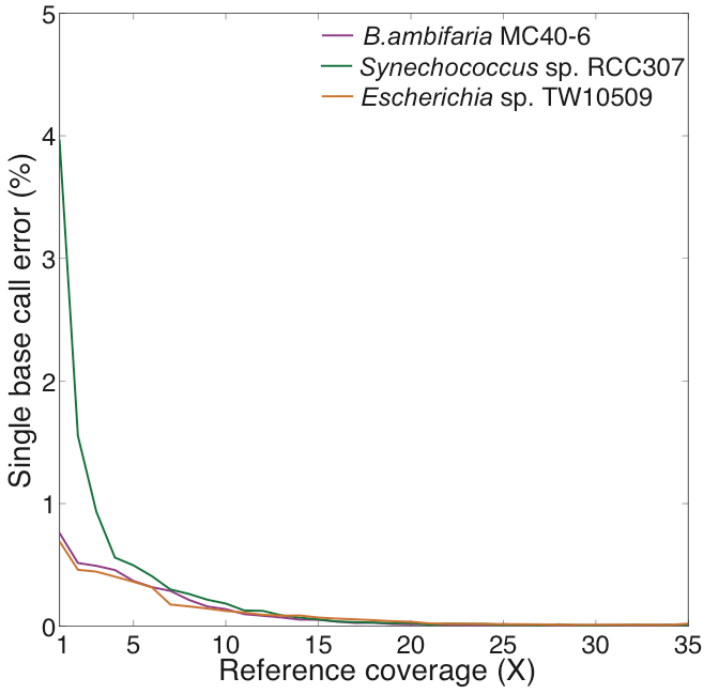


Figure 4.8. Frequency of single base error in assembled consensus sequence as a function of coverage. Note that the higher error rate in the case of *Synechococcus* sp. RCC307 is primarily due to sequences of close relatives in the metagenome. These sequences contained substantial single nucleotide polymorphisms (SNPs) relative to the reference genome sequence (Figure 4.4) that thus, contributed SNPs (errors) to the assembled consensus in the low coverage range (where they outweighed the reference sequences spike-in). In contrast, *Escherichia* sp. TW10509 has no relatives in the metagenome and the relatives of *B. ambifaria* MC40-6 were not of high enough identity to be included in the consensus sequence based on the cut-offs used in the assembly step (Figure range compared to *Synechococcus* sp. RCC307. In all cases, the rate of single base error decreased with higher coverage and converged to the same value.

Analysis of chimeric sequences

At low coverage, assembled contigs containing target sequences also contained a substantial fraction of non-target (chimeric) sequences (12% at 1X in this study, Figure 4.3C and Figure 4.9) and, as a result, *in silico* generated, artificial genes (6% of total predicted genes, about 40% of which occurred in contigs shorter than 500bp in the *Escherichia* sp. TW10509 case). The majority of the artificial genes were hypothetical, meaning that they did not have a significant match in nr database (1009/1548, or 65% in the *Escherichia* sp. TW10509 case); over 50% of the remaining genes with a match in nr database also matched hypothetical or conserved hypothetical genes (Figure 4.9). Therefore, special caution should be exercised when drawing conclusions based on low-coverage contigs. Many natural communities are characterized by low-abundance species, i.e., species than make up <0.1% of the community. Such species would be covered at the 1-2X level or less based on the amount of sequencing usually obtained in current studies. Thus, our findings explain, at least in part, why metagenomes typically have fewer genes with matches to the public databases than complete genomes of isolates (i.e., they contain more chimeric, artificial sequences). Such chimeric sequences likely contribute to the high sequence diversity observed in several assembled metagenomes, particularly those from soil, as well. On the other hand, when single genome abundance was higher than about 10X, which corresponded to a N50 longer than 2 Kb, the percentage of non-target sequences dropped below 1%. These results provided a practical threshold for identifying reliable contigs based on their length and coverage.

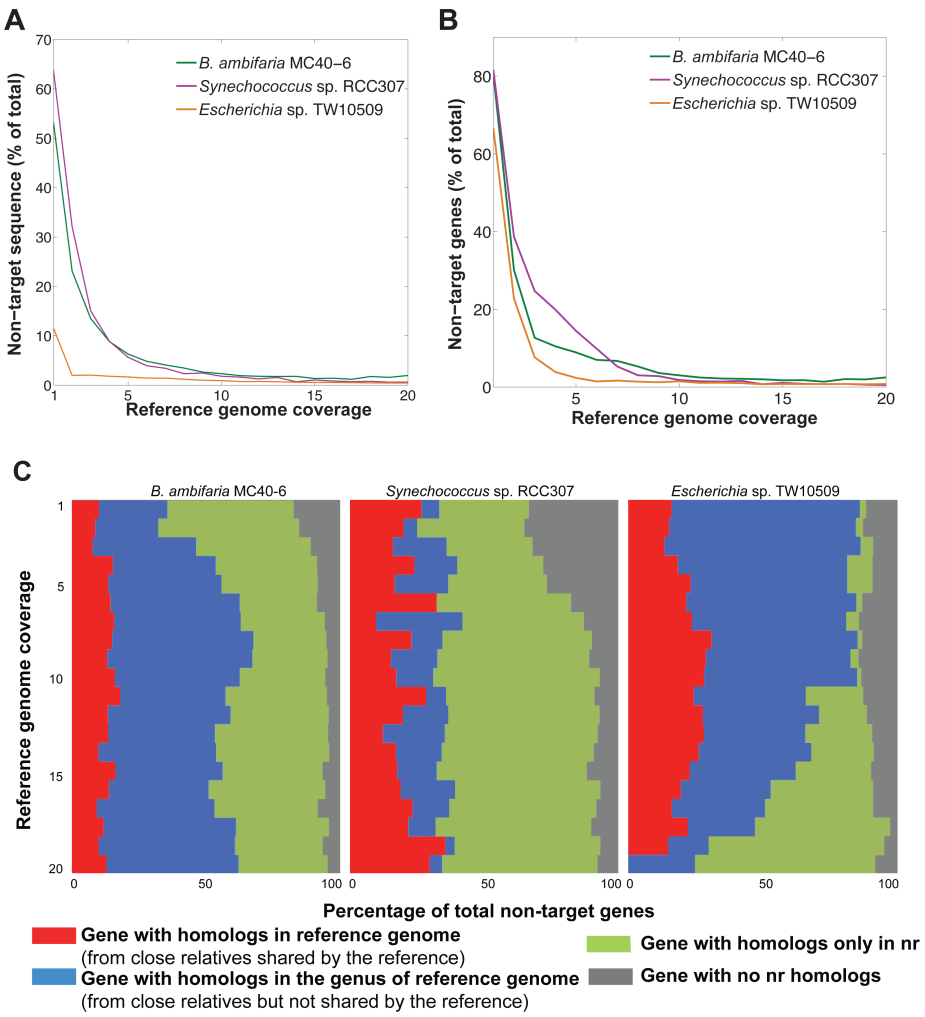


Figure 4.9. Analysis of the non-target (chimeric) sequences in assembled contigs. The fraction of non-target sequence (**A**) and non-target genes (**B**) in assembled contigs of the reference genome from the spiked-in Lanier.Illumina metagenome (y-axes) is shown as a function of the abundance of the reference genome (x-axes) for each reference genome (figure key). Target and non-target sequences were defined as described in the Materials and Methods. The distribution of the non-target genes (as a fraction on the total non-target genes, x-axis) in four functional categories (figure key) for assemblies of different coverage of the reference genome in the mixed metagenome (y-axis) is also shown (**C**). This analysis revealed that the close relatives present in the Lanier.Illumina metagenome contribute a larger fraction of non-target sequences in the *B. ambifaria* MC40-6 and *Synechococcus* sp. RCC307 compared to the *Escherichia* sp.

TW10509 assembly (no relatives in the metagenome). In contrast, most non-target sequences in the *Escherichia* sp. TW10509 case, especially in the high coverage range, were genes with no homologs in *Escherichia* sp. TW10509 or the non-redundant database (nr).

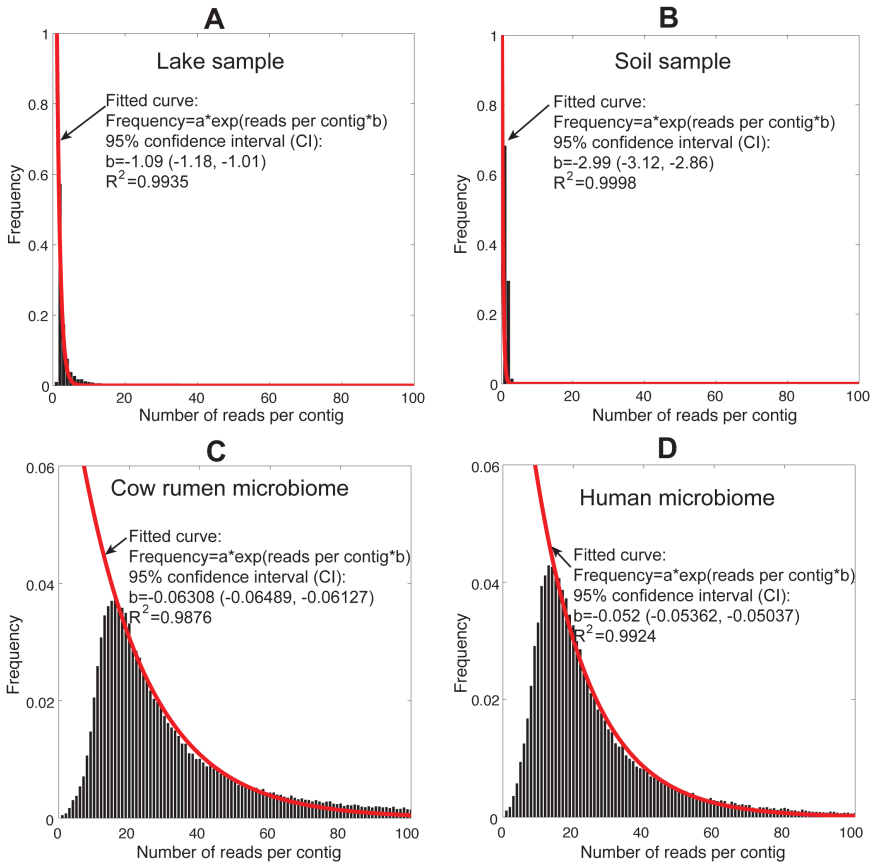


Figure 4.10. Comparison of the complexity of the Lake Lanier (A) and the soil (B) metagenomes used in this study to selected metagenomes reported previously. The graphs show the number of contigs (y-axes) plotted against the number of reads composing the contig (x-axes) resulting from the assembly of each corresponding metagenome. Exponential regression was fitted to the data using the cftool module in MatLab and the fitted curves are shown. The cow rumen and human metagenomes were reported previously (9, 10); a randomly drawn subset of the latter metagenomes, which was similar in size to the Lake Lanier and soil metagenomes, was

used in the analysis. Note that our approach takes into account both the relative abundance (evenness; represented by the number of reads per contig) as well as the number of unique populations (richness; represented by the number of contigs) of the communities. Therefore, the larger the absolute value of *b* (i.e., the steeper the curve) the higher the complexity of the metagenome. Also note the difference in scale of the y-axes between panels A and B relative to C and D.

Investigating intra-population genetic structure

Natural populations are frequently composed of several closely related genotypes as opposed to a single genotype. It remains challenging to use metagenomics for the robust assessment of intra-population genetic structure, e.g., to detect heterogeneous populations. To this end, we expanded the single genotype analysis to include five additional *Escherichia* sp. genomes, which showed pairwise genetic relatedness ranging from 90% to 95% average nucleotide identity [ANI, (22); Table 4.1 & Figure 4.11]. Regardless of the composition of the target population in the *in silico* generated metagenome, the six genomes were recovered as a discrete sequence cluster when all metagenomic reads were mapped on the reference *Escherichia* sp. strain TW10509 genome (Figure 4.12). The sequence-discrete clusters were obvious for other reference populations as long as no close relatives with higher than ~85% ANI to the population were present in the metagenome. Furthermore, the shape of the coverage plot reliably reflected the target population genetic structure: when the population was homogeneous (i.e., all genomes were spiked at similar abundances) the shape of the coverage plot approximated a normal distribution around the average ANI of the six genomes (~92%);

when the population structure was heterogeneous (e.g. one genome more abundant than the others), the shape of the coverage plot deviated from the normal-like distribution and quantitatively reflected the variations in individual genome abundance. However, we were unable to recover robust assemblies of individual genotypes, even in trials where the target genotype consisted more than 50% of the population (Figure 4.13) or when a high nucleotide identity cut-off in the assembly was used due to the fact that assemblers apply consensus strategy when encountering polymorphisms.

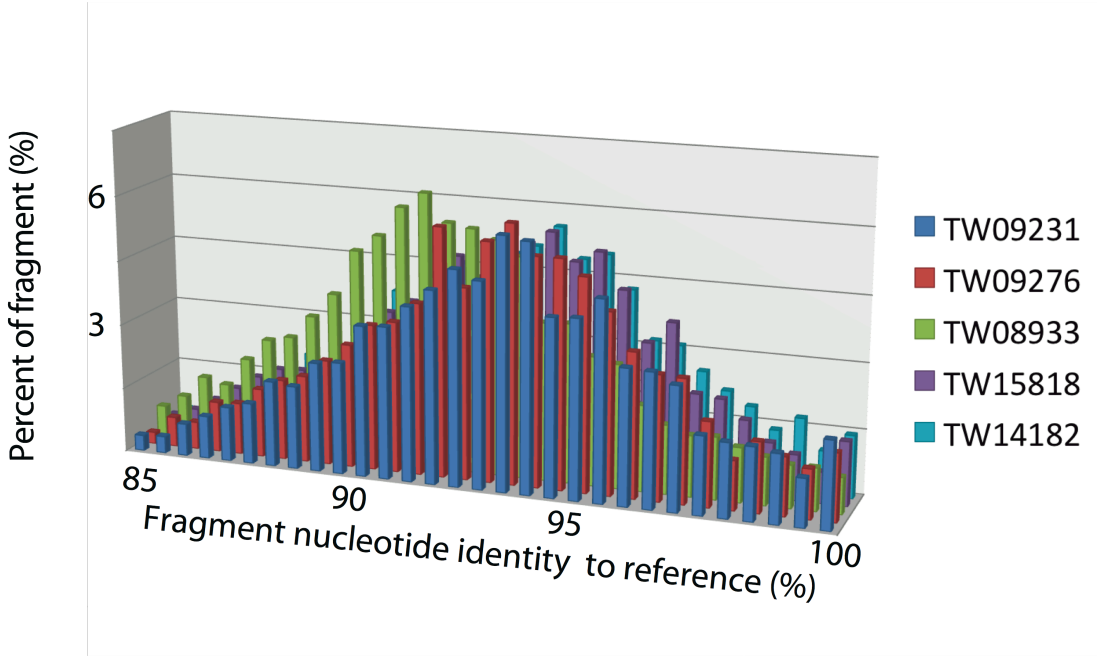


Figure 4.11. Genetic relatedness among the six *Escherichia* sp. genomes used in the study. The sequence of each genome was cut into non-overlapping consecutive 500 bp long fragments, and these fragments were searched against the TW10509 reference genome draft using blastn. The graph shows the number of fragments (y-axis) plotted against their nucleotide identity to the reference (x-axis) for each genome (graph key).

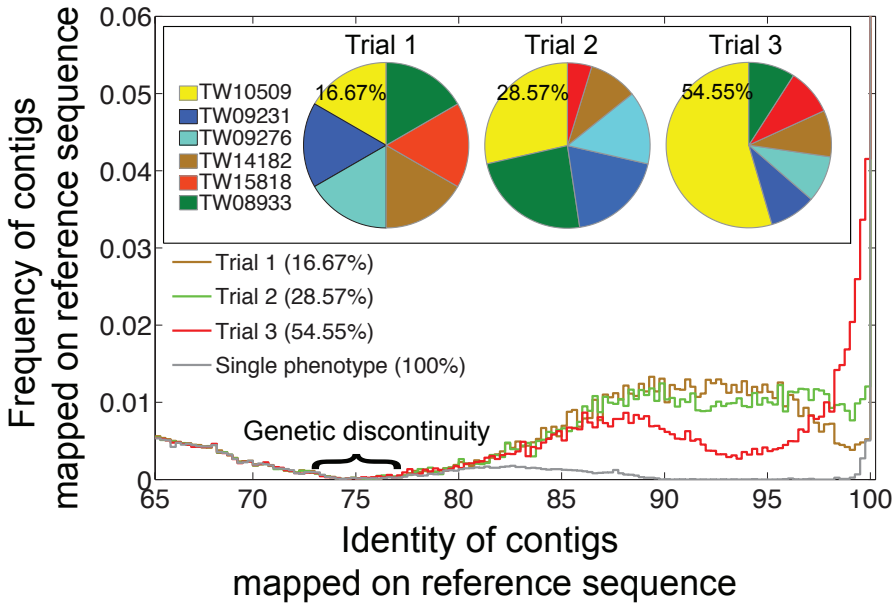


Figure 4.12. Assessment of intra-population genetic structure based on sequence coverage plots. The total number of reads of the reference population spiked into the Lanier.Illumina metagenome was fixed at 35X coverage, but the proportions of the different genotypes making up the population varied as represented by the pies (*inset*). The graph represents a coverage plot, constructed as previously described (18), and shows the nucleotide identity (x-axis) of all contigs from the *in silico* generated metagenome (target and non-target) that map on the TW10509 genome sequence (y-axis), which was used as reference. Note that a genetic discontinuity in the 75-80% nucleotide identity range was always observed, regardless of the genotype composition of the population. Also note that when the portion of TW10509 reads in the metagenome increased (from 16.67% in trial 1 to 54.55% in trial 3), the coverage plot reflected the shifts in the higher portion of reads in the 98-100% range (contributed by the TW10509 reads).

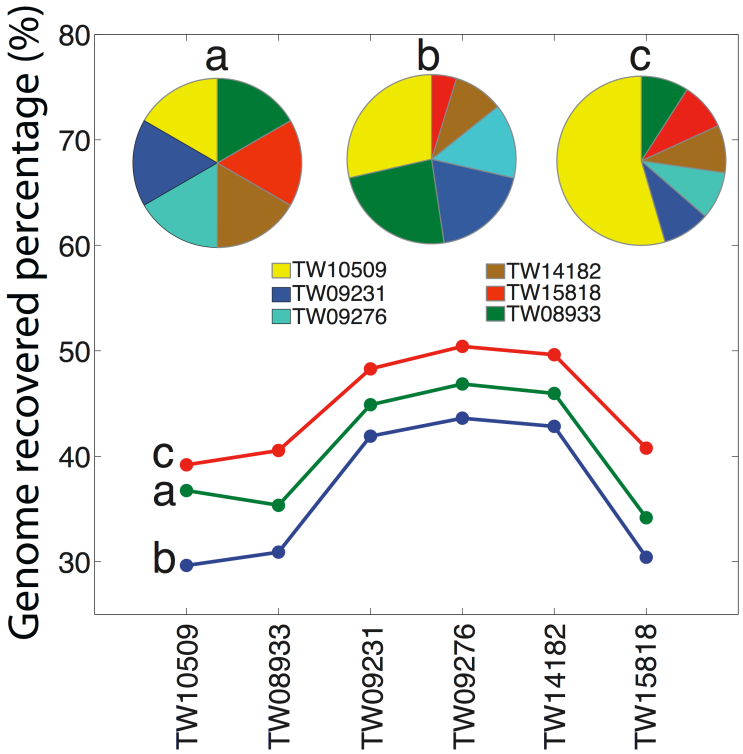


Figure 4.13. Recovery of the genome of a single genotype from a heterogeneous population spiked into a complex metagenome. The total number of reads of the reference population spiked into the Lanier.Illumina metagenome was fixed at 35X coverage, but the proportions of the different genotypes making up the population varied as represented by the pies (*inset*). The graph shows the fraction of the genome (y-axis) of a single target genotype (x-axis) that was assembled from the *in silico* generated metagenome. Note that we were unable to recover more than 50% of the genome of the target genotype regardless of the relative abundance of the genotype in metagenome or the genotype used in the analysis and that there was a marginal increase in the fraction of the genotype recovered with increase genotype abundance. This was mostly attributed to the fact that assemblers apply a consensus strategy when encountering polymorphisms and hence, the assembled contigs are weighted average sequences.

REFERENCES

1. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.
2. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5(4):433-438.
3. DeLong EF*, et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311(5760):496-503.
4. Qin J*, et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.
5. Konstantinidis KT, Braff J, Karl DM, & DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* 75(16):5345-5355.
6. Gomez-Alvarez V, Teal TK, & Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3(11):1314-1317.
7. Aird D*, et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12(2):R18.
8. Quince C*, et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 6(9):639-641.
9. Hess M*, et al.* (Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331(6016):463-467.
10. Qin J*, et al.* (A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.
11. Li R*, et al.* (De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265-272.
12. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821-829.
13. Maccallum I*, et al.* (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* 10(10):R103.
14. Simpson JT*, et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6):1117-1123.
15. Noguchi H, Park J, & Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34(19):5623-5630.
16. Borodovsky M, Mills R, Besemer J, & Lomsadze A (2003) Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics* Chapter 4:Unit4 5.
17. Luo C*, et al.* (Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A* 108(17):7200-7205.
18. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* 2(10):1052-1065.
19. Luo C*, et al.* (2011) Genome sequencing of environmental *E. coli* expands understanding of the ecology and speciation of the model bacterial species. *PNAS*:In press.

20. Oh S*, et al.* (2011) The metagenome of Lake Lanier provides new insights into the evolution, function and complexity of temperate freshwater microbial communities. *Appl Environ Microbiol*:Under revision.
21. Branscomb E & Predki P (2002) On the high value of low standards. *J Bacteriol* 184(23):6406-6409; discussion 6409.
22. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102(7):2567-2572.

ACKNOWLEDGEMENTS

We thank Rachel Poretsky for useful discussions related to the manuscript. This work was supported by the US Department of Energy under Award No. DE-SC0004601 (to KTK) and contract No. DE-AC02-0SCH11231 (to NCK).

CHAPTER 5

MeTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences

INTRODUCTION

Culture-independent whole genome shotgun (WGS) DNA sequencing has revolutionized the study of the diversity and ecology of microbial communities during the last decade (1, 2). However, the tools to analyze metagenomic data are clearly lagging behind the developments in sequencing technologies, with the probable exception of tools for sequence annotation and assembly (1, 3-5). Perhaps most importantly, the taxonomic identity of the majority of sequences assembled from a metagenomc dataset frequently remains elusive, severely impeding communication among scientists and scientific discovery across the fields of ecology, systematics, evolution, engineering and medicine. This is due, at least in part, to the fact that the great majority of microbial species in nature, >99% of the total in some habitats (6), resist cultivation in the laboratory and thus, are not represented by sequenced reference representatives that can aid taxonomic identification. Single-cell techniques can potentially overcome these limitations by providing the genome sequence of uncultured organisms (7). However these techniques are not amenable to all organisms or habitats and the 16S rRNA gene, which serves as the best marker for taxonomic identification due to the availability of a large database of 16S rRNA gene sequences from uncultured organisms (8, 9), is often missed or not assembled during single-cell (and WGS metagenomic) approaches. The 16S rRNA gene also provides limited resolution at the species level, which represents a major limitation for epidemiological and microdiversity studies (10). To overcome these limitations, whole-genome-based approaches and tools, comparable to those already available for the 16S rRNA gene, are highly needed. It is also important for these tools to scale with the increasingly large volume of sequence data produced by the new sequencers and to be

able to detect and categorize novel taxa, *e.g.*, determine if the taxa represent novel species or genera.

The previous methods to taxonomically identify metagenomic sequences fall into two categories: composition-based, such as PhylopythiaS and NBC (11, 12); and homology-based, such as CARMA and MEGAN4 (5, 13). Although composition-based methods do not depend on the availability of a reference database and are typically faster to compute, their accuracy is usually significantly lower than homology-based methods, especially for regions of the genome that are characterized by abnormal statistics compared to the genome average, due, for instance, to horizontal gene transfer (HGT) (14). On the other hand, homology-based approaches such as those employing BLAST (15) and HMMER3 (1) searches of assembled or unassembled sequences against known reference database(s), have become a nearly indispensible component of metagenomic studies (4). Even naïve implementations of simple classification algorithms such as best hit (BH) or lowest common ancestor (LCA) usually provide comparable accuracies with some sophisticated composition-based approaches (1). The main limitation of the homology-based approaches is the lack of a comprehensive database of reference genome sequences. Accordingly, query sequences representing novel taxa provide only low-identity matches or no matches to the reference sequences and, in a typical metagenomic study, the majority of sequences cannot be robustly classified. Low-identity matches represent a challenge to the identification of the degree of novelty of the query sequence, particularly for naïve classifiers, which are based on pre-set, and frequently arbitrary, thresholds. In such cases, a dynamic approach that takes into account the level of identity of the match and the classification power of the corresponding gene or sequence (e.g., the

16S rRNA gene provides robust resolution at the genus level and higher but poor resolution at the species level) are advantageous. However, most, if not all, of the dynamic approaches developed for these purposes rely on some unrealistic assumptions such as that genes within the same protein family are characterized by the same mutation rate, and lack a robust framework for determining the degree of novelty of a query sequence (4, 5, 13).

Here we present a novel framework, MeTaxa, which overcomes several of the previous limitations and can accurately classify metagenomic and genomic sequences with low computational requirements. MeTaxa considers all genes present in an unknown sequence as classifiers and quantifies the classifying power of each gene using predetermined weights. The weights are for i) how well the gene in question resolves the classification at a given taxonomic level based on its degree of sequence conservation, and ii) how frequently the gene phylogeny deviates from the species phylogeny due (primarily) to horizontal gene transfer. Based on these weights and the top homology matches of the genes in the query sequence against a pre-clustered reference gene database, a maximum likelihood analysis is performed to choose the most probable taxonomic assignment and to decide the lowest taxonomic rank for the query sequence. We show that MeTaxa significantly outperforms state-of-the-art tools for the same purposes in both sensitivity and specificity of the taxonomic assignments and can easily incorporate additional reference gene sequences as these become available through future isolate genome and single cell sequencing projects to provide for a more comprehensive coverage.

MATERIALS AND METHODS

Gene clustering

The predicted protein-coding genes of 1,480 completed and 1,687 drafted microbial genomes were downloaded from NCBI's FTP server (ftp.ncbi.nih.gov) in July 2012. An all-versus-all search of all genes was carried out using USearch (version 5.0) (16). Orthologs were defined as the reciprocal best match (RBM) genes between any two genomes, with percentage amino acid identity higher than 40%, no less than 70% coverage of the length of the shorter gene by the alignment, and e-value smaller than $1! 10^{-12}$. Neo4j (www.neo4j.org) was subsequently used to construct a graph in which the nodes were genes and the edges were RBM relationships. Genes were grouped in gene clusters based on the graph by an agglomerative hierarchical approach. Non-RBM (paralog) genes were searched against the resulting gene clusters using the same USearch search as described above; genes with matches above the previous cut-off were merged into the corresponding best-match gene cluster. In total, 850,629 gene clusters (singletons included) comprising 4,665,401 genes were obtained.

Genome-aggregate average amino acid identity (AAI)

To measure the overall genetic relatedness between any two genomes, we used the AAI, a robust and universal measure (17). AAI was calculated as the arithmetic average of the amino acid identity of all RBM conserved genes between two genomes. By comparing the AAI values among genome pairs grouped at different taxonomic ranks (e.g., phylum, class, *etc*), it became evident that phyla, genera and species are clearly distinguishable from

each other (Figure 1). Therefore, MeTaxa considers only these three taxonomic ranks when classifying query sequences.

Gene cluster-parameterization

We quantified the classifying power of each gene cluster at each of the three taxonomic ranks. The classifying power was defined by: i) how well the gene separates intra-group members from inter-group ones based on the degree of sequence conservation (measured by *D*). For instance, the 16S rRNA gene is highly conserved and thus can resolve well the phylum and genus levels but poorly the species level; several rapidly evolving protein-coding genes resolve well the species and genus levels but poorly the phylum level (*e.g.*, permissible mutations are saturated at the phylum level). And ii) how consistent the gene phylogeny is with the species phylogeny, the latter approximated by the AAI distance tree (measured by *M*).

To quantify *D*, the identities (or distances) among all gene sequences of a gene cluster were calculated in a pair-wise mode and categorized into "intra-group" (the two corresponding genomes that encode the genes were assigned to the same taxon) and "inter-group" (the two genomes were assigned to different taxa). The distributions of the distances of the two categories were then estimated by a kernel density estimator (KDE) with a Gaussian kernel function using bandwidths selected by Scott's rule (18). Therefore, the classifying power of a gene cluster *c* as a function of the amino acid identity of the gene obtained from the gene sequence comparison (*e.g.*, a query sequence against the database) *h,* denoted as $D_t^c(h)$, for a given taxonomic rank *t*, was calculated as:

$$D_t^c(h) = \frac{\int_h^1 f_t^c(s)ds}{\int_h^1 f_t^c(s)ds + \int_h^1 g_t^c(s)ds},$$

Where $f_t^c(s)$ and $g_t^c(s)$ denote the distributions of the intra- and inter-group distances for gene cluster *c* at taxonomic level *t*, respectively (Figure 5.1).

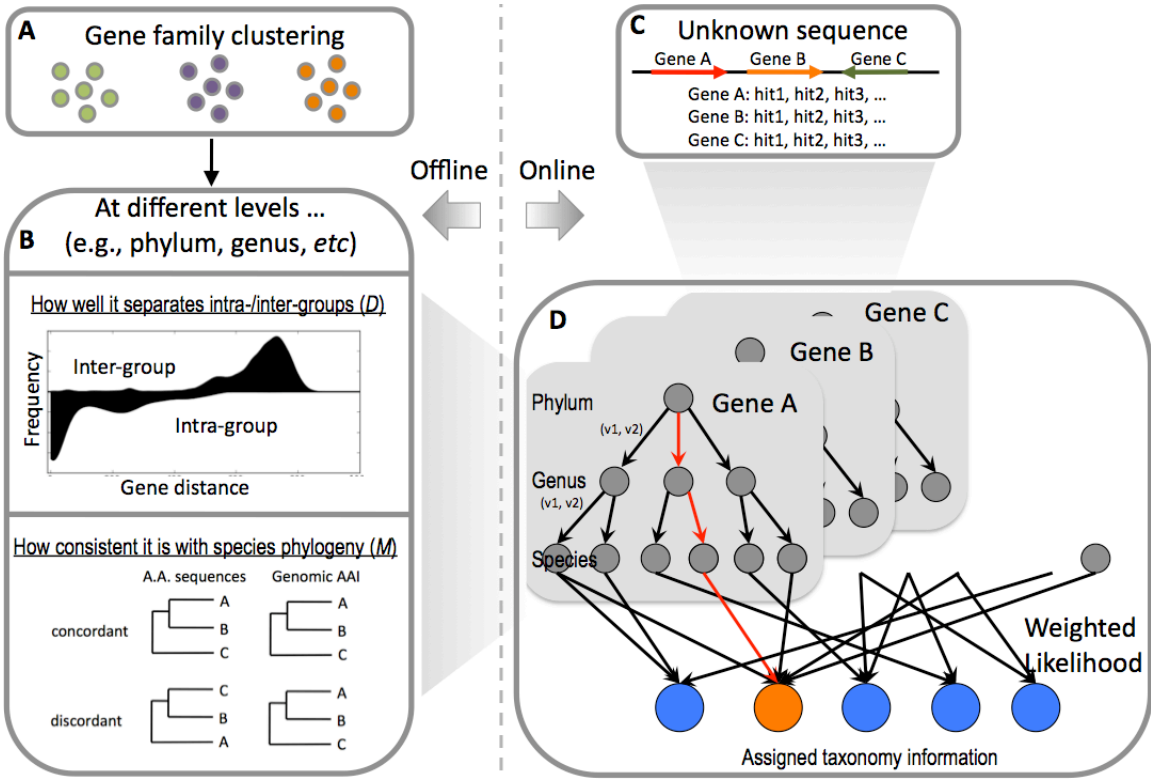


Figure 5.1. The workflow of the MeTaxa algorithm. The input to MeTaxa are the top *N* matches of the genes encoded in an unknown sequence (or gene fragments for shorter sequences) against a reference database of gene families. The identity of the match(es) and the gene family(ies) that are represented among the top *N* matches are modeled (online part) based on predetermined weights for the classifying power of each gene family considered (offline part) to determine the lowest taxonomic rank that the query sequence should be assigned to. More specifically, we grouped all genes from

available complete genomes into clusters (**box A**), and calculate the weights *D* (how well the gene resolves the taxonomic rank) and *M* (how consistent the gene phylogeny is to the species phylogeny) for each cluster and taxonomic rank considered (i.e., phylum, genus, and species). To quantify *D*, the identities (or distances) among all gene sequences of a gene cluster were calculated in a pair-wise mode and categorized into "intra-group" (the two corresponding genomes that encode the genes were assigned to the same taxon) and "inter-group" (the two genomes were assigned to different taxa). The larger the difference between the inter-group vs. the intra-group identities the larger the classifying power of the gene with respect to *D* (an example of the distribution of identities is represented by the histogram shown in **box B)**. To quantify *M*, we extracted all possible triplets from the phylogenetic tree of all sequences of a gene cluster, and compared them with the species tree, the latter approximated by the AAI tree (distance tree). Therefore, the triplets were either "concordant" or "discordant" with species tree (lower panel in **box B**). During the sequence assignment ("online" part), external users provide a list of matches of the genes in a query sequences and the corresponding identities from a similarity search (e.g., Blastn, BLAT) against a reference gene database such as GenBank (**box C**). MeTaxa takes this input and maps the matches onto the reference gene clusters generated from the offline part, based on the accession numbers of the (matching) genes from GenBank. The corresponding *D* and *M* weights are extracted for each rank that the taxon encoding the matching gene sequence is assigned to. If different genes of a query sequence or matches of a single gene suggest different classifications (i.e., matching taxon differs), each classification receives a likelihood score by merging the identity of the match and the corresponding *D* and *M* weights (see Online Methods for the exact equation used). If the total likelihood score of a classification (from the sum of the likelihoods of each match that supports the exact same classification) is below a minimum threshold, the classification is discarded. MeTaxa reports the classification that receives the largest likelihood score above the threshold at each taxonomy rank, together with its likelihood score (marked in red in **box D**).

To quantify *M*, we first constructed a phylogenetic tree of all gene sequences of a cluster using FastTree(19) with default settings, and then extracted all possible triplets from the tree. Each triplet was compared against the corresponding species tree constructed based on AAI values. Therefore, the triplets were either "concordant" (tree topology consistent with species tree), or "discordant" (topology inconsistent with species tree). Hence, the degree of gene cluster *c* being consistent with the species phylogeny at taxonomic level *t*, denoted as M_t^c, was calculated as:

$$M_t^c = \frac{N_t^c(concordant)}{N_t^c(concordant) + N_t^c(discordant)},$$

where $N_t^c(concordant)$ denotes the number of concordant triplets and $N_t^c(discordant)$ denotes the number of discordant triplets for gene cluster *c* at taxonomic level *t*.

Monte-Carlo method for estimating weights of large size gene clusters

For gene clusters with more than 5,000 members (40 such clusters were obtained, in total), it was computationally prohibitive to exhaust all possible triplets among the members. We employed a simple Monte-Carlo method to estimate the *M* values for these gene families. The method was applied as follows, separately for each of the three taxonomic ranks considered:

1. Initialization;

   Set *M*=*s*, *N*=number of genes in the cluster; set number for concordant triplet, *c*=0; and number of discordant triplet, *d*=0.

2. Random sampling;

10,000 gene triplets are sampled from the gene cluster uniformly at random without replacing; the corresponding species triplets are constructed from the species tree.

3. Calculate new M;

   The 10,000 gene triplets are compared against the corresponding species triplets. If they are concordant, then $c=c+1$; otherwise $d=d+1$. The new M, M', is calculated as $M'=c/(c+d)$.

4. Termination.

   If $|M'-M|<0.01$, then exit the algorithm and return M';

   Otherwise, $M=M'$, return to step 2.

To avoid being trapped at local maxima, we repeat the process with initializing M to be different values (in 0 to 1 with 0.1 increment, *e.g.*, 0, 0.1, 0.2, …, 1.0). If there were multiple estimated M's, we repeat the process until it converges to a single value.

Classification step and likelihood score calculation

For an unknown sequence U (*e.g.*, an assemble contig from a metagenome), we denote the genes encoded on it as $G=\{g_1, g_2, ..., g_n\}$. In the online part of the algorithm, these genes are searched against a reference database (*e.g.*, the gene clusters described above) and the returned m matching genes for g_i are denoted as $H_i=\{h_1, h_2, ..., h_m\}$, the corresponding percentage amino acid identities as $I_i=\{i_1, i_2, ..., i_m\}$, and the bit-scores as $S_i=\{s_1, s_2, ..., s_m\}$. For each match in H_i, we denote the corresponding taxonomic classification as $T_i=\{p_1, p_2, ..., p_m\}$, which represents the taxonomic affiliation of the genome that encodes the matching

gene; and at taxonomic rank *t*, the labels (taxa) are $T_i^t = \{p_1^t, p_1^t, ..., p_m^t\}$. We also denote the gene cluster each query sequence is assigned (matched) to as $C_i=\{c_1, c_2, ..., c_m\}$. The top *N* matches of the l^{th} gene against the reference database are used to weight different taxa (superscripted as *k*) at a specific rank (subscripted as *t*), p_t^k , and the weight is:

$$W_l(p_t^k) = \sum_{\{j | p_t^j = p_t^k\}}^{N} \left[D_t^{c_j}(i_j) w_t^D + M_t^{c_j} w_t^M \right] s_j ,$$

where w_t^D and $w_t^M = 1 - w_t^D$ are the weights for *D* and *M* at taxonomic level *t,* respectively. To select the optimal w_t^M and w_t^D, a grid search was carried out to maximize the algorithm's performance (see grid search below). Therefore, the relative weight of p_t^k for the l^{th} gene, $L_l(p_t^k)$ is normalized to the sum of weights over different *k* (*i.e.*, different taxonomic classifications):

$$L_l(p_t^k) = \frac{W_l(p_t^k)}{\sum_k W_l(p_t^k)}.$$

And, the likelihood score of a specific taxon *k* at rank *t*, p_t^k, over the whole query sequence is:

$$L(p_t^k) = \frac{\sum_{l=1}^{m} L_l(p_t^k)}{\sum_k \sum_{l=1}^{m} L_l(p_t^k)}.$$

If the top-scored taxon at this rank passes the likelihood score cutoff (see also score cut-off estimation below), MeTaxa predicts the query sequence to belong to this specific taxon and moves to the lower rank (if any) and calculates the likelihood score for this rank in a similar fashion. If likelihood is below threshold, MeTaxa marks the current and lower ranks (if any) as unknown (novel) taxon.

Test datasets and measuring accuracy

1,351 drafted microbial genomes were downloaded from NCBI's ftp site in February 2012. A custom Perl script was employed to randomly sample pieces of sequences from the drafted genomes at designated lengths (*e.g.*, 100bp, 800bp, *etc*; see Table S1). These sequences formed the synthetic test datasets. At a given taxonomic rank, each sequence was denoted as "known" if the same taxon (*i.e.*, species, genus or phylum) was also represented by at least one of the completed genomes, or "unknown" if the same taxon was not represented among the completed genomes. Assessing MeTaxa's performance on these test synthetic metagenomes was performed as described in the main text.

For evaluating MeTaxa on real metagenomes, a human stool sample microbiome (accession number: SRX023971, including both assembled scaffolds and the original paired-end Illumina reads) was downloaded from the Human Microbiome Project (HMP) Consortium webpage (www.hmpdacc.org). The gene sequences annotated on the scaffold sequences were searched against all completed genomes in NCBI using BLAT (16), and the taxonomy assignment was carried out by MeTaxa using default settings. Trimmed paired-end reads were mapped onto the scaffold to calculate the coverage (in-situ abundance) of the corresponding population using BLAT with default setting and a minimum cut-off of aligned length: 50bp, nucleotide identity: 70%, and e-value: 1e-10 for a match. The reads encoding fragments of the 16S rRNA gene were identified by a BLAT search against the reference 16S rRNA gene sequence from *E. coli* with the following cutoff: 70% nucleotide identity, 1e-10 e-value, and 50bp aligned length, both sister read matching above the cut-off. These sequences were extracted from the

metagenome using a custom PERL script and were searched against the GreenGenes database (20). Reads were assigned to taxa based on their GreenGenes match and the abundance of each taxon was approximated by the number of assigned reads, normalized by the rRNA copy number of the taxon reported in the literature. These data provided the community composition of the metagenome based on the 16S rRNA gene shown in Figure 4.

A sequence from the synthetic test datasets was labeled either "known" or "unknown" at a given taxonomic level. The taxonomic assignment of a "known" sequence by MeTaxa or another algorithm was denoted as "true prediction" (TP; predicted taxon matches the actual taxon), "wrong prediction" (WP; predicted taxon does not match the actual taxon), or "false negative" (FN; predicted as "unknown"); while the assignment of an "unknown" sequence was denoted either "false positive" (FP; predicted to match a specific taxon), or "true negative" (TN; predicted as "unknown"). Accordingly, the sensitivity of the algorithm was defined as:

$$Sn = \frac{1}{N+M}\left(\sum_{i=1}^{N}\frac{n_{TP}^{i}}{n^{i}} + \sum_{i=1}^{M}\frac{n_{TN}^{i}}{m^{i}}\right),$$

where N is the total number of taxa for the "known" sequences, and M is the total number of taxa for the "unknown" sequences. n^i is the number of "known" sequences for each taxon, and m^i is the number of "unknown" sequences for each taxon. n_{TP}^i is the number of TP in the i^{th} taxon and n_{TN}^i is the umber of TN in the i^{th} taxon. Similarly, the specificity of the algorithm was defined as:

$$Sp = \frac{1}{N}\sum_{i=1}^{N}\frac{n_{TP}^{i}}{n^{i}}.$$

Weight optimization based on a grid search

D and *M* weights were generated independently and thus, could not be integrated directly. To find the optimal combination of *D* and *M*, we defined (w^D, w^M) as the relative power of these two parameters, and the combined weight was $W = w^D D + w^M M$. The sum of w^D w^M should equal 1; therefore, we only need to optimize the algorithm performance over one of them, *e.g.*, w^D. A grid search was employed for this purpose and the 1000bp test dataset was used. For each possible (w^D, w^M) pair (w^D was set to be 0.05, 0.1, 0.15, …, 0.95), we sampled 10% of the 1000bp test dataset at random ten times (replicates) and make MeTaxa assignments. The assignments were evaluated by their accuracy as described above, and the corresponding weight pair with the highest accuracy was selected.

Impact of the number of matches and score cutoffs on classification accuracy

In MeTaxa, the top *N* number of matches of genes are used in predicting the taxonomic identity of the query sequence. When *N*=1, MeTaxa is equivalent to a weighted lowest common ancestor (LCA) algorithm; and when *N*= , MeTaxa considers all taxa in the reference database. Further, the larger the *N* value the larger the CPU and memory requirements. The choice of *N* has a complicated impact on prediction accuracy, and for practical reasons, we have tested *N*=1, 2, …, 10, and found that for most cases, *N*=5 offers optimal performance (Figure 5.2. Similarly, we evaluated the impact of likelihood score cutoffs on accuracy. We found that 0.5 usually provides the best performance (Figure 5.3).

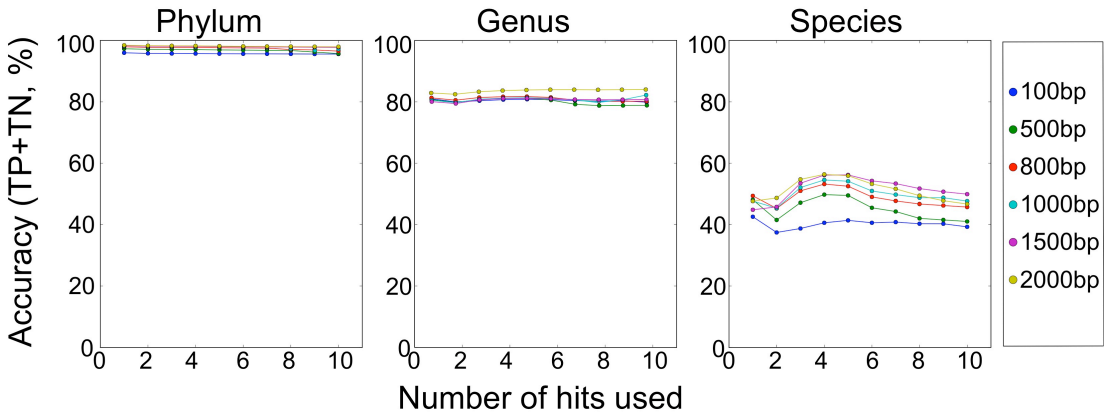


Figure 5.2. The impact of the number of matches used in the analysis on the classification accuracy of MeTaxa. The impact of the number of top matches analyzed, *N*, on MeTaxa accuracy was evaluated, for query sequences of varied length (figure key) and each taxonomic rank (labels on top). $N = 5$ typically performed the best, both in terms of accuracy (y-axes) as well as computational demand (data not shown).

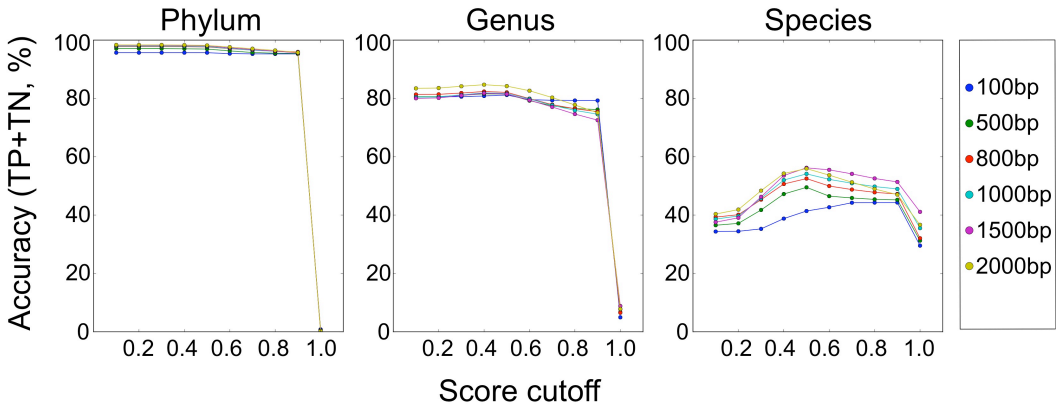


Figure 5.3. The impact of the likelihood score cutoff on the classification accuracy of MeTaxa. The impact of different likelihood cut-offs in the maximum likelihood analysis was evaluated in a mode similar to that shown in Figure 5.2. Scores of 0.5 typically perform the best (y-axes) and were efficient in terms of computational demand (data not shown).

RESULTS AND DISCUSSION

Standardizing novel taxa based on Average Amino-acid Identity

For correct and high-throughput taxonomic classification of an unknown sequence, it is essential to have a robust and standardized reference taxonomy system. The current taxonomic system, especially the ranks higher than the species rank, is primarily based on the grouping patterns of the 16S rRNA gene phylogeny but no standards exists on the degree of genetic relatedness of the organisms grouped at different ranks. Accordingly, adjacent ranks are highly overlapping with this respect. For instance, organisms representing different species of the same genus are often (>30% of the cases examined) as divergent from each other as many genera of the same family are (21). These inconsistencies can complicate taxonomic identification of unknown sequences. Indeed, several commonly used approaches including PhyloPythiaS and MEGAN have significantly lower specificity above the species level (12).

To examine in depth the inconsistencies in the current classification system, we analyzed 410 closed bacterial genomes (410 X 410 = 168,100 genome pairs, in total) using the genome-aggregate average amino acid identity (AAI) to measure the genetic relatedness among the genomes (17). Our results confirmed previous findings that high overlap exists among adjacent ranks (e.g., phylum vs. domain) but also revealed that the species, genus, and phylum ranks are rarely overlapping, *i.e.*, the inter-taxon divergence is typically higher than the intra-taxon diversity for these three ranks (Figure 5.4). In particular, organisms grouped at the "species" level typically show >95% AAI among themselves and are clearly distinguishable from those grouped at the genus (showing 60-80% AAI) and the phylum levels (showing <40% AAI). MeTaxa essentially employs

these AAI standards and examines the degree to which an individual gene reflects the genomic AAI, as described in the Online Methods, to determine the taxonomic rank of a sequence representing a novel organism, *i.e.*, a species, genus or phylum. The latter also represent the three most important ranks of prokaryotic taxonomy.

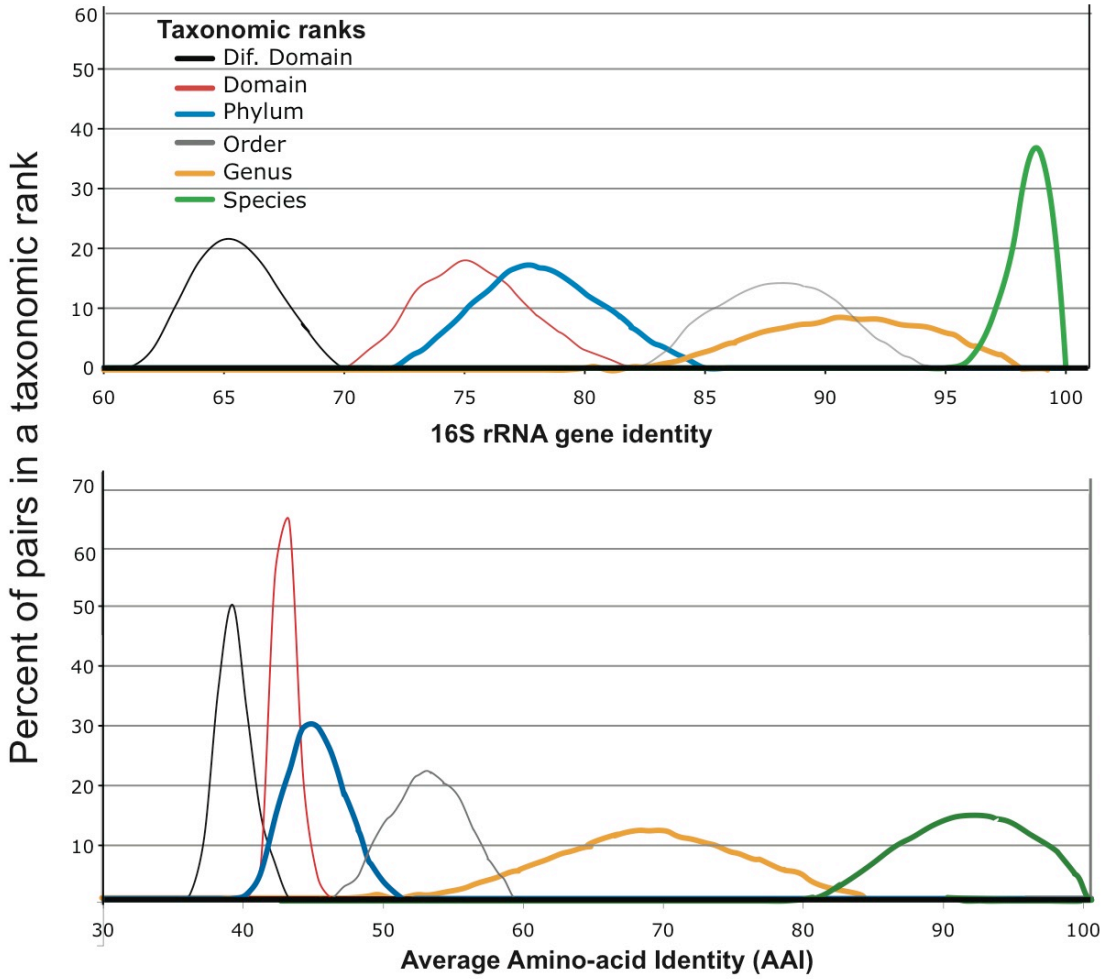


Figure 5.4. Relationships between taxonomic designations and genome-aggregate average amino-acid identity (AAI). The taxonomic designations of 410 fully sequenced genomes were compared to identify the lowest taxonomic rank shared by each pair of genomes (410 X 410 = 168,100 pairs, in total), essentially as described previously (21). For each taxonomic rank (figure

key), the corresponding line shown represents the distribution of the 16S rRNA gene identity (top) and AAI (bottom) values among all genomes grouped at the rank. Note that species, genera and phyla are clearly distinguishable from each other based on AAI and correspond to 95-100%, 60-80% and <45% AAI, respectively.

Computing the weights of the classifying power of each gene.

To determine the weights of each gene, we built clusters for all genes present in all completed bacterial and archaeal genomes as of August 2012 (n=1,480). We determined the classifying power of each gene cluster by comparing how well the identity between two genes of the cluster reflected the taxonomic rank of the genomes encoding the genes, separately for each of the three taxonomic ranks considered. The idea is analogous to the use of AAI above to examine overlap between the taxonomic ranks, applied to individual genes. A second weight was calculated for each gene cluster based on how frequently the ortholog gene phylogeny deviates from the species phylogeny, the latter approximated by the AAI-based tree, due (primarily) to horizontal gene transfer. The weights were stored in a structured database, and the preceding analysis is referred to as the "offline" part of MeTaxa (external users do not repeat this part). For the "online" part, an external user submits a file that contains the results of a search, by BLAST, HMMER3 or other algorithm, of the sequence in question against the reference database of gene clusters. In fact, the search is not necessary to be against our reference database as long as the input file contains the accession number of the best matching gene(s) in GenBank database and the amino acid identity of the match. MeTaxa then employs a maximum likelihood analysis of the pre-calculated weights for the gene cluster that

provided the best match of the query sequence and the identity of the best match to determine the taxonomic identity of the query sequence and provide a statistical probability for the assignment. Therefore, the most computationally intensive part is calculated offline, perhaps only once or twice a year (to update weights as more genomes become available), and MeTaxa requires significantly lower computational resources during the online part of the analysis compared to similar previous methods.

MeTaxa performance

We evaluated the performance of MeTaxa against that of other existing tools based on the following approach. For classifying sequences that represent organisms present in the database (100% AAI match) or close relatives of these organisms (*e.g.*, >95% AAI for organisms of the same species), the algorithm should (correctly) identify the sequence to the lowest level possible. The latter was typically the species level, unless the reference organism has not been assigned to a known species yet. For sequences representing, for instance, an unknown (novel) genus of a known phylum, the algorithm should ideally identify the correct phylum, predict the genus as the lowest taxonomic rank and denote it as a novel genus; similarly for novel phyla and species. Based on this framework, we sampled, at random, 1,687 draft genomes to produce six test, *in-silico* generated, metagenomic datasets that were composed of sequences of different length, ranging from 100bp (simulating Illumina reads), 500bp (simulating Roche 454 Titanium FLX reads), 800bp (simulating Roche 454 FLX+ reads), 1,000bp (representing the average bacterial gene length), 1,500bp, to 2,000bp. We employed 1,480 completed genomes from GenBank to serve as the reference database and build the gene clusters and

associated weights (Tables 5.1 and 5.2). Thus, the sequences in the test metagenomes were labeled "known" or "unknown" depending on whether or not a completed genome of the same species as the draft genome was available and the algorithms were evaluated on the number of correct assignments (predictions) made.

Table 5.1. The number of known and unknown taxa in draft genomes (synthetic metagenomic data) compared to completed genomes (reference database).

Rank	Number of known taxa; percentage (%)	Number of unknown taxa, percentage (%)
Phylum	20; 87.0%	3; 13%
Genus	143; 50.7%	139; 49.3%
Species	136; 18.5%	600; 71.5%

Table 5.2. The number of known and unknown sequences at different ranks in the synthetic metagenomic data used for performance evaluation.

Length (bp)	Phylum		Genus		Species	
	Known	Unknown	Known	Unknown	Known	Unknown
100	954676	7168	814224	121620	527684	408160
500	493493	4234	429146	69981	290555	208572
800	494576	3909	430989	67496	287334	211151
1000	494186	3534	423663	73757	268881	228539
1500	296648	1536	252148	46036	148360	149824
2000	198286	1242	172622	26906	110584	88944

For the homology-based algorithms, we ran a BLAT (16) search of the six test metagenomes against the reference database, and the search results were used as input for the algorithms. For composition-based algorithms, we classified test metagenomes using the default settings of each algorithm. Since composition-based methods tend to classify more sequences compared to homology-based methods (*e.g.*, they do not depend on the availability of a comprehensiveness reference database), we compared all methods based on sequences classified by all approaches, *i.e.*, sequences that had at least one significant match in the BLAT search.

The results revealed that MeTaxa consistently outperformed other tools (Table 5.3 for all results; Figure 5.5 shows the results for the 800bp dataset). For example, at the species level (800bp test dataset), it was, on average, 17.8%, 6.9%, 25.1%, and 9.2% more accurate than best hit (BH), lowest common ancestor (LCA), MEGAN4, and MG-RAST, respectively. Furthermore, as the length of query sequences increased, the advantage of MeTaxa was also more pronounced (Figure 5.6). NBC provided more correct classifications compared to the other composition-based methods, consistent with previous findings (11). MeTaxa outperformed NBC by 10.6%, 11.2%, 17.0% at the phylum, genus, and species levels, respectively (average of all test metagenomes).

We also calculated the sensitivity (Sn; portion of sequences from known taxa correctly assigned) and specificity (Sp; portion of sequences from unknown taxa correctly identified as unknown at the lowest rank possible) for all methods (Figure 5.7). MeTaxa showed both high sensitivity and specificity in all three taxonomic ranks evaluated, *e.g.*, at species level, MeTaxa is on average 5% more sensitive and 3% more specific than other methods. Moreover, the sensitivity and specificity of MeTaxa did not

seem to depend on the length of the input sequences, while most composition-based approaches showed strong length-dependent variance in both sensitivity and specificity (Figure 5.7).

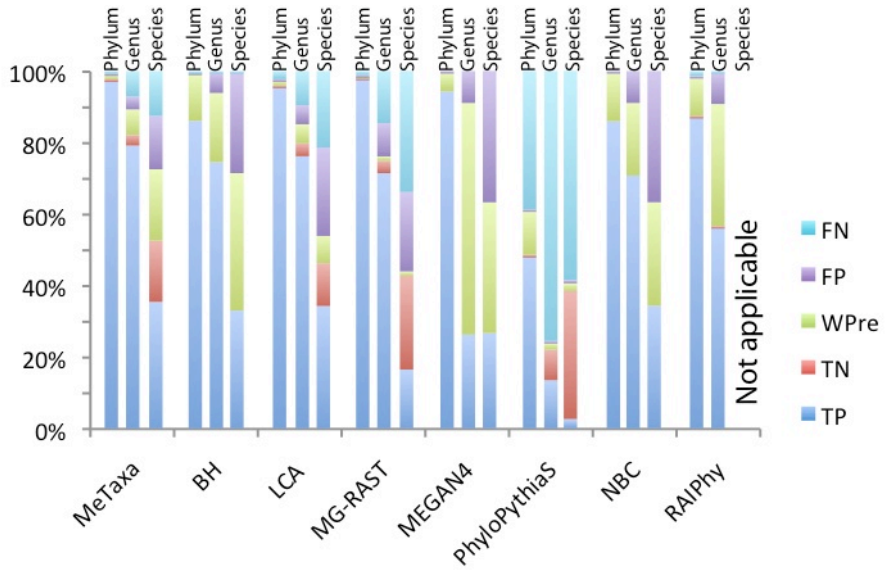


Fig. 5.5. MeTaxa performance and comparison with other methods. Each bar represents the relative distribution of the different types of predictions (figure key) made by each of the methods evaluated (*x* axis), at each taxonomic rank considered (labels on top). FN, false negative (sequence from a known taxon predicted as unknown); FP, false positive (sequence from an unknown taxon predicted as known); WPre, wrong prediction (the known taxon did not match the predicted taxon); TN, true negative (sequences from an unknown taxon were predicted as unknown); TP, true prediction (the known taxon matched the predicted taxon). RAIPhy is not applicable to the species level. The results are based on the 800 bp long test metagenome.

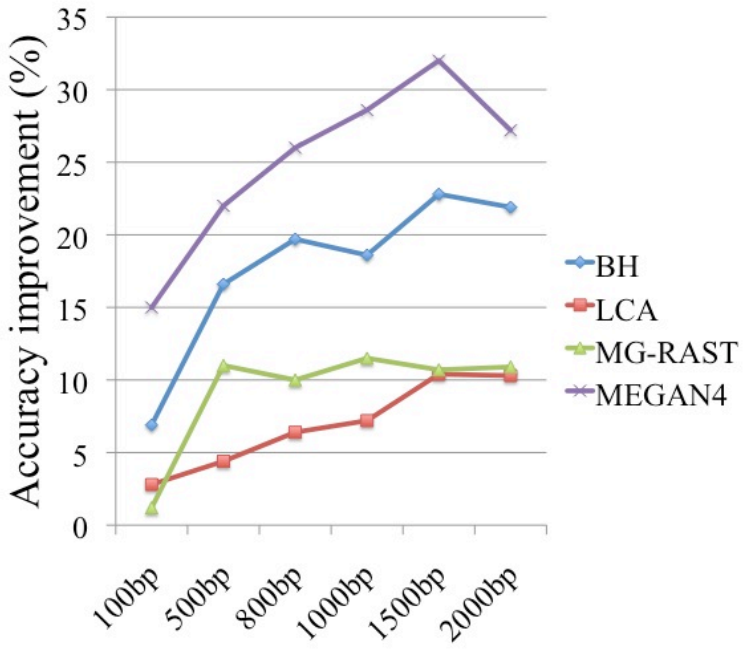


Figure 5.6. Accuracy of MeTaxa in comparison with other homology-based methods at the species level. The accuracy (TP+TN) of different methods (figure key; y-axis) as a function of the length of the query sequence (x axis) is shown. Note that MeTaxa correctly assigns at least 3%, and up to 32%, more sequences than any other method, depending on the length of the query sequences. Figure S4 is similar to Figure 3 but represents the sum of the true positives (TP) and true negatives (TN) results, zooming in at the species level.

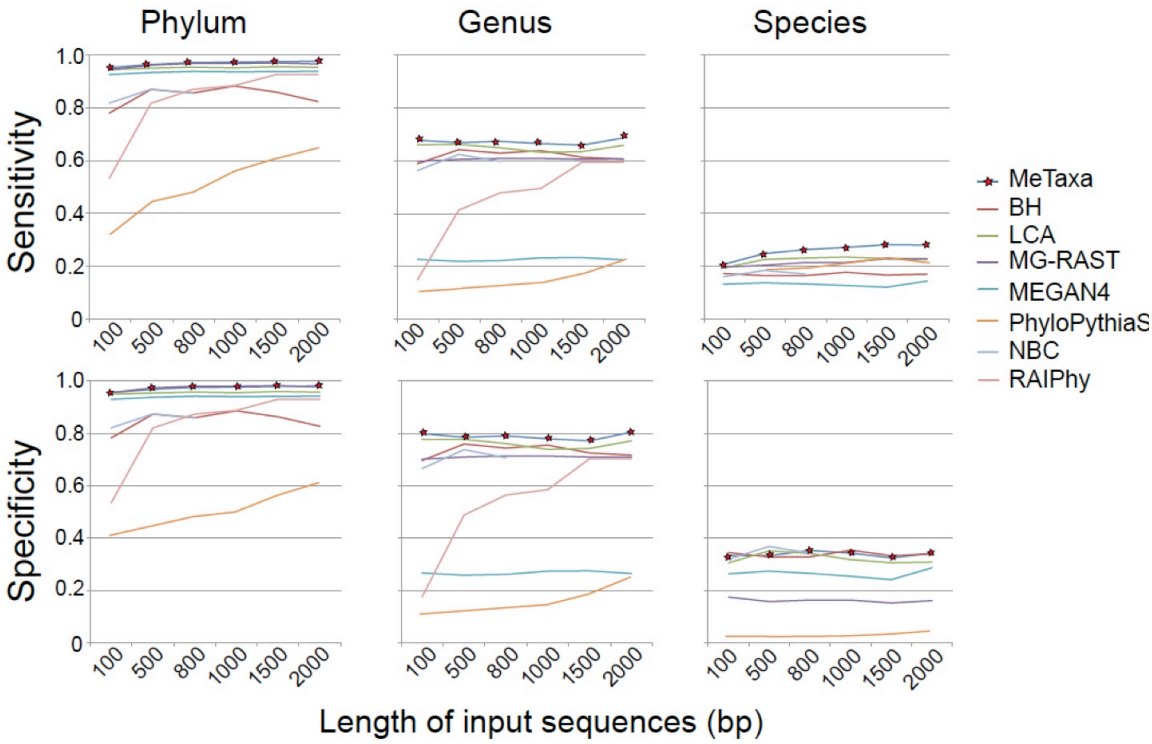


Figure 5.7. Sensitivity and specificity of MeTaxa and comparison with other methods. Each line represents the sensitivity (y-axes; upper panels) or specificity (y-axes; lower panels) of a method (figure key) on different lengths of input sequences (*x* axes). Sensitivity and specificity were defined as described in the text.

Table 5.3. Detailed performance on synthetic metagenomic datasets. This data underlie the results shown in figure 5.5.

	Phylum					Genus					Species					
	TP[6]	TN	WP	FP	FN	TP	TN	WP	FP	FN	TP	TN	WP	FP	FN	
100bp	95.6	0	3.1	0.8	0.5	80	1.2	10.4	3.8	4.7	33.4	8.0	30.8	21.6	6.2	MeTaxa
	78.4	0	20.8	0.7	0.2	69.7	0	26.1	4.0	0.2	34.5	0	41.2	24.0	0.2	BH[2]
	95.0	0	2.9	0.9	0.2	77.8	1.5	6.9	10.8	7.6	30.6	8.0	13.3	27.5	20.6	LCA[3]
	95.6	0.2	1.5	0.5	2.1	70.2	2.9	2.2	4.2	20.5	17.5	21.7	2.0	13.5	45.3	MGR[4]
	93.1	0	6.0	0.9	0	26.8	0	65.5	7.8	0	26.4	0	38.1	35.5	0	MEGAN4
	42.1	0.1	19.1	0.5	38.2	10.5	6.8	3.1	0.5	79.1	2.4	33.8	2.5	1.2	60.1	PPS[5]
	82.2	0	16.8	0.9	0	66.6	0	25.7	7.7	0	32.1	0	32.6	35.4	0	NBC
	53.5	0	43.4	1.0	2.1	17.6	0	73.9	7.7	0.7			N/A[1]			RAIPhy
500bp	96.9	0.1	1.9	0.4	0.7	78.6	2.9	7.5	3.7	7.3	33.4	16.1	23.2	15.1	12.2	MeTaxa
	87.4	0	11.8	0.4	0.4	75.9	0	17.7	5.9	0.3	32.9	0	38.7	28.0	0.3	BH
	95.5	0	2.0	0.5	2.0	77.8	3.3	6.1	5.7	7.1	35.2	9.9	9.9	25.8	19.1	LCA
	97.4	0.1	0.6	0.6	0	71.0	3.4	1.7	0.7	16.6	15.8	24.7	1.6	18.2	39.8	MGR
	93.8	0	5.6	0.4	1.5	25.9	0	65.1	9.0	0	27.5	0	36.8	35.7	0	MEGAN
	44.7	0.4	17.4	0.3	37.1	12.3	7.9	2.9	0.4	76.4	2.5	34.6	2.2	0.8	59.9	PPS
	87.5	0	11.9	0.6	0	73.8	0	17.2	9.0	0	36.9	0	27.4	35.7	0	NBC
	82.1	0	15.6	0.5	1.6	48.8	0	41.7	8.9	0.5			N/A			RAIPhy
800bo	97.6	0	1.4	0.3	0.6	79.1	3.0	7.2	3.6	7.1	35.4	17.2	20.0	14.9	12.6	MeTaxa
	86.0	0	13.2	0.3	0.4	74.4	0	19.4	5.7	4.2	32.9	0	38.5	28.2	0.3	BH
	95.9	0	1.4	0.4	2.3	76.1	3.6	5.3	5.3	9.6	34.2	12.0	7.6	24.8	21.4	LCA
	98.0	0.1	0.4	0.3	1.2	71.4	3.5	1.2	9.3	14.6	16.4	26.2	1.2	22.2	33.9	MGR
	94.3	0	5.3	0.4	0	26.2	0	64.8	9.0	0	26.6	0	36.6	36.8	0	MEGAN
	48.3	0.2	12.6	0.2	38.7	13.5	8.5	2.0	0.4	75.6	2.5	35.8	2.2	1.0	58.5	PPS
	85.9	0	13.7	0.4	0	70.7	0	20.3	9.0	0	34.3	0	29.0	36.8	0	NBC
	87.4	0	10.8	0.4		56.5	0.1	34.2	8.9	0.3			N/A			RAIPhy
1000bp	97.7	0	1.3	0.3	0.6	78.1	3.7	7.0	4.0	7.2	34.3	19.8	17.7	15.7	12.4	MeTaxa
	88.7	0	10.6	0.3	0.5	75.5	0	19.0	7.7	0.6	35.5	0	31.9	32.2	0.4	BH
	95.6	0	1.3	0.3	1.2	73.9	4.7	4.9	5.6	10.8	31.9	15.0	6.3	25.4	21.4	LCA
	98.0	0.1	0.4	0.3	1.2	71.4	3.5	1.2	9.3	14.6	16.4	26.2	1.2	22.2	33.9	MGR
	94.1	0	5.5	0.3	0	27.5	0	62.2	10.3	0	25.5	0	34.0	40.5	0	MEGAN
	50.0	0.2	11.1	0.1	38.5	14.6	9.8	1.8	0.5	73.2	2.8	39.4	2.3	1.1	54.4	PPS
	88.2	0	11.2	0.6	0	67.9	0	21.8	10.3	0	35.6	0	28.6	35.8	0	NBC
	88.9	0	9.5	0.3	1.3	58.6	0	30.8	10.3	0.3			N/A			RAIPhy
1500bp	98.0	0	1.2	0.2	0.6	77.2	4.3	6.2	4.6	7.7	32.5	23.7	14.0	17.4	12.3	MeTaxa
	86.4	0	10.6	0.3	0.5	72.3	0	19.0	7.7	0.6	33.4	0	29.8	36.3	0.4	BH
	95.6	0	1.3	2.9	2.8	74.3	4.3	4.7	6.3	10.4	30.7	15.1	5.5	29.3	19.4	LCA
	98.0	0.1	0.4	2.9	1.2	71.0	3.8	1.1	10.6	0	15.3	30.2	0.9	25.5	28.2	MGR
	94.1	0	5.5	0.3	0	27.6	0	61.7	10.7	0	24.2	0	31.4	44.4	0	MEGAN

Table 5.3 (continued)

	50.0	0.2	11.1	0.1	38.5	18.7	10.1	1.6	0.6	69.0	3.4	42.7	2.8	1.7	49.4	PPS
	88.3	0	11.3	0.4	0	67.3	0	22.0	10.7	0	31.7	0	23.9	44.4	0	NBC
	91.9	0	6.8	0.2	1.0	64.2	0	25.0	10.6	0.2			N/A			RAIPhy
2000bp	98.1	0	1.0	0.2	0.6	80.6	3.7	4.8	4.2	6.8	34.3	21.7	15.0	15.0	14.0	MeTaxa
	82.8	0	16.4	0.2	0.6	71.7	0	21.1	6.6	0.6	34.1	0	34.4	31.0	0.5	BH
	95.8	0	1.0	0.3	2.9	77.2	4.0	3.6	5.4	9.8	30.9	14.8	5.1	24.7	24.4	LCA
	97.8	0	0.5	0.3	1.3	71.0	4.3	1.5	10.0	13.1	16.2	28.9	1.7	25.1	28.1	MGR
	94.3	0	5.4	0.3	0	26.6	0	64.1	9.4	0	28.8	0	31.7	40.0	0	MEGAN
	61.2	0.2	7.1	0.1	31.3	25.1	8.8	1.4	0.6	64.1	4.6	37.7	4.2	1.9	51.7	PPS
	88.3	0	11.3	0.4	0	68.2	0	21.5	10.3	0	37.0	0	21.2	41.8	0	NBC
	93.1	0	5.6	0.3	1.0	70.4	0.1	20.0	9.3	0.2			N/A			RAIPhy

* The numbers represent percentages of the total;

[1] N/A, not available, RAIPhy does not provide species level prediction;

[2] BH, best hit; [3] LCA, lowest common ancestor; [4] MGR, MG-RAST (4); [5] PPS, PhyloPythiaS (12);

[6] TP, true prediction; TN, true negative; WP, wrong prediction; FP, false positive; FN, false negative.

Novel diversity revealed in the human microbiome

We also evaluated MeTaxa on real metagenomes, using the assembled scaffolds of the human gut microbiome project (GenBank accession number: SRX023971). In addition to MeTaxa analysis of the assembled scaffolds, we also mapped raw reads on the scaffolds to estimate relative *in-situ* abundance of the corresponding organisms, essentially as described previously (22). The resulting taxa abundance profiles were compared to those from the analysis of 16S rRNA gene fragments recovered in the metagenome, which represents the most popular approach to perform community composition analysis (23).

Community composition was highly similar between MeTaxa and 16S rRNA gene-based results, at both the phylum and genus levels. For example, the 16S rRNA gene analysis revealed that the relative abundance of the three most abundant genera, *i.e.*, *Bacteriodes, Prevotella,* and *Roseburia*, was 45.6%, 9.2%, and 5.6%, respectively; MeTaxa results for the same genera were 41.5%, 9.4%, and 6.1%, respectively (Figure 5.8). These findings demonstrate the high accuracy of MeTaxa in profiling real metagenomes. In addition, MeTaxa detected several organisms missed by the 16S rRNA gene survey and identified most sequences to the species level, which is typically inaccessible to 16S rRNA gene analysis. For example, at phylum level, MeTaxa identified that 0.16% of the overall reads were contributed by *Fusobacteria*, which have been shown to be associated with colon cancer and ulcerative colitis (24). This phylum was missed by the 16S rRNA gene analysis presumably due to the fact that rRNA genes represent only a small portion of the whole genome, and thus are likely to be missed during WGS sequencing by chance alone. In contrast, MeTaxa represents a genome-based approach and thus, it is less likely to miss low-abundance community members. MeTaxa analysis also showed that an average of 22.3% of the total sequences assigned to each of the top 20 most abundant genera were contributed by novel species, which are not currently represented by genome sequences, draft or complete. For instance, in *Prevotella* and *Paludibacter* genera, which represent keystone members of the three proposed human gut microbiome enterotypes (2), novel species represented 52.8% and 44.6% of the total sequences assigned to the genus level, respectively (Figure 5.8).

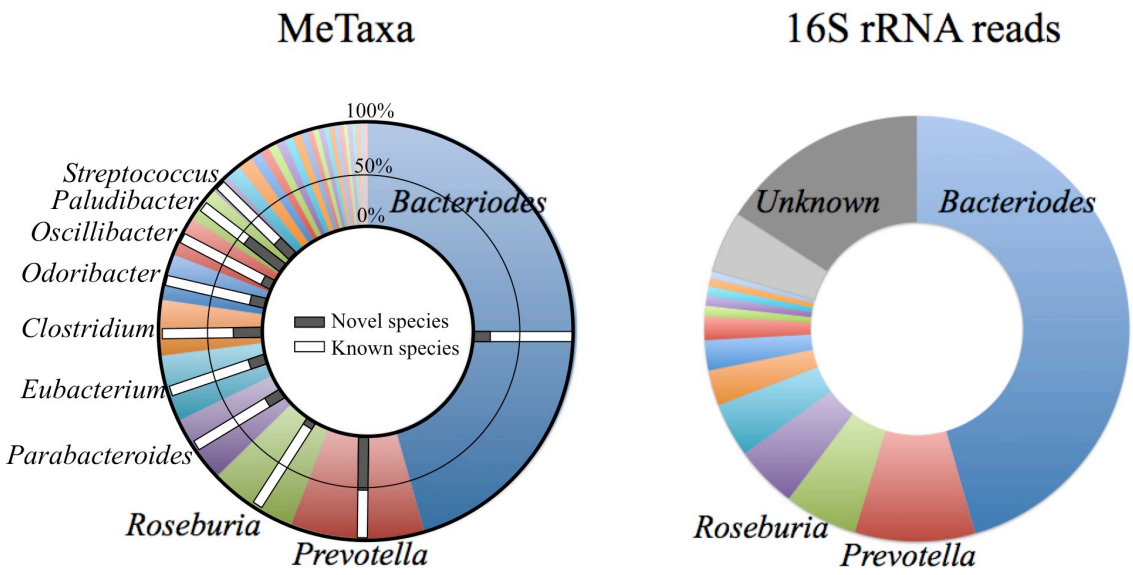


Figure 5.8. Genus-level community composition and abundance of novel taxa in the human microbiome based on MeTaxa and 16S rRNA genes. 16S rRNA gene results were based on all reads that were recovered in the WGS metagenome (accession number: SRX023971) and encoded fragments of the 16S rRNA gene. These reads were analyzed using established tools, as described in the Online Methods. MeTaxa analysis was performed as described in the text. Each color represents a different genus and the most abundant genera are labeled. Grey bars on the left panel represent the amount of sequences predicted by MeTaxa to represent novel species within each of the corresponding genera. The analysis revealed that MeTaxa shows comparable accuracy and higher sensitivity to detect low abundance taxa compared to the 16S rRNA gene-based analysis. For instance, note that a larger fraction of sequences remained unassigned (denoted as "Unknown" on the graph) for the 16S rRNA gene results, mostly due to the short length of 16S rRNA gene-encoding reads (lack of resolving power), which is typical of short-read-based shotgun or gene amplicon surveys.

DISCUSSION

We have shown that MeTaxa accurately classifies at least 5%, and up to 25%, more query sequences compared to other methods for sequences representing previously described taxa, independent of the length of the sequences (Figure 5.5, 5.7, 5.9-10). The advantage of MeTaxa is rooted to the construction of a likelihood framework that integrates the matches of individual genes with weights for the classifying power of each gene to achieve a better prediction. This approach is categorized into a broad genre of optimizations, often referred to as "the wisdom of the crowd". Indeed, a greater advantage of MeTaxa over other methods was observed when the query sequences were longer (Figure. 5.6), presumably due to more information (genes) available. Accordingly, MeTaxa can also facilitate taxonomic studies of whole genomes, complete or draft, and be complementary to 16S rRNA gene-based classifications since it provides higher resolution at the species level. MeTaxa has also a clear advantage over other methods in identifying the rank of sequences representing novel taxa due to the use of an AAI-based framework that emerges from the current classification system but it is more standardized (Figure 5.4). This is particularly useful to the study of communities that are not well represented by reference genome sequences (the majority of microbial communities) and can help identify abundant, and thus, presumably important, members of the community that should be targeted for single-cell or cultivation efforts. Indeed, MeTaxa analysis of a human microbiome sample revealed several abundant (novel) species that are not represented by genomes of isolates, despite the large number of isolates sequenced as part of the Human Microbiome Project. For instance, although a large number of *Prevotella* isolate genomes are available (50 of the total 1,569 used in this study, including draft

genomes), MeTaxa suggested that several key members of this genus are still awaiting genomic characterization.

Despite the significant improvements achieved by MeTaxa, assigning sequences at the species level remains problematic; mostly due to the lack of representative sequences for several species (*e.g.*, Figure 5.7). However, the recent developments in DNA sequencing technologies, especially single cell approaches, has greatly increased the number and phylogenetic diversity of available genomes so that the reference genome database will not represent such a major limitation in the near future. What, however, will still represent a limitation for automatic, high throughput taxonomic identification are the inconsistencies in the current classification system. While we employed a standardized AAI-based system to determine the degree of novelty of a sequencing representing a novel taxon, we relied on the existing system for sequences representing previously described taxa. The weights of gene clusters are expected to significantly improve if a standardized system, which will limit overlap between adjacent taxonomic ranks in terms of the genetic relatedness of the grouped organisms, will become available for previously described taxa. MeTaxa is also scalable to a higher volume of input data in that the computational demand for the online part of the algorithm represents a linear function of the number of input sequences. MeTaxa is not specific to the homology search algorithm used; thus, if new faster algorithms become available*, e.g.*, BLAT (25), they can be easily compatible with MeTaxa.

One advantage of composition-based methods is that they are able to classify sequences that show no significant homology to the reference database. However, it remains unclear how accurate these predictions are. Genes with no significant homology

to known genes are highly likely to represent taxon-specific functions and their evolutionary history is often inconsistent with that of the genome (*e.g.*, acquired via horizontal gene transfer) (26). In our datasets, NBC's accuracy on "unknown" sequences was on average 20% lower compared to those with matches in the reference database, which was largely attributable to a higher wrong prediction rate (Figure 5.10). Therefore, if classifying more sequences is more important than high accuracy in the classifications, a hybrid approach that combines composition-based and homology-based methods may be advantageous.

In addition to the applications mentioned above, MeTaxa can also be used to assist studies that aim to detect HGT between genomes or contigs assembled from a metagenome and the genomes represented in the reference database, *e.g.*, by scanning the query sequence in windows of specific length and compare the taxonomic affiliations of the resulting sequence fragments. It can also assist in validating the taxonomic identity of contigs binned into population genomes during metagenomic studies, especially for populations representing previously described taxa. Thus, MeTaxa can find several important applications in microbial identification and diversity studies and provide new insights into the tremendous complexity of microbial communities.

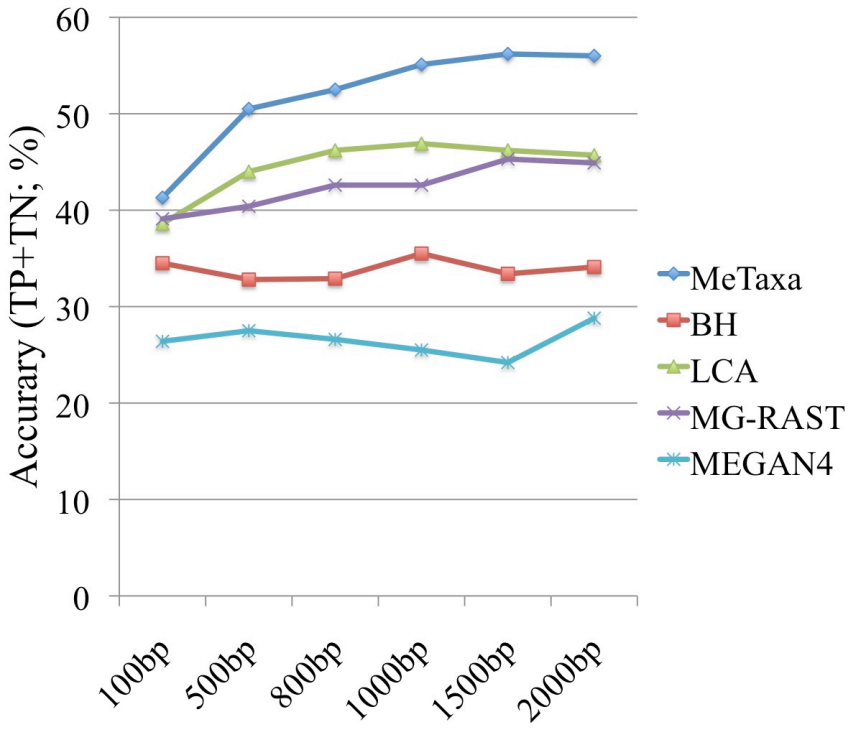


Figure 5.9. Accuracy of MeTaxa in comparison with other homology-based methods at the species level. The accuracy (TP+TN) of different methods (figure key; y-axis) as a function of the length of the query sequence (x axis) is shown. Note that MeTaxa correctly assigns at least 3%, and up to 32%, more sequences than any other method, depending on the length of the query sequences. This figure is similar to Figure 5.7 but represents the sum of the true positives (TP) and true negatives (TN) results, zooming in at the species level.

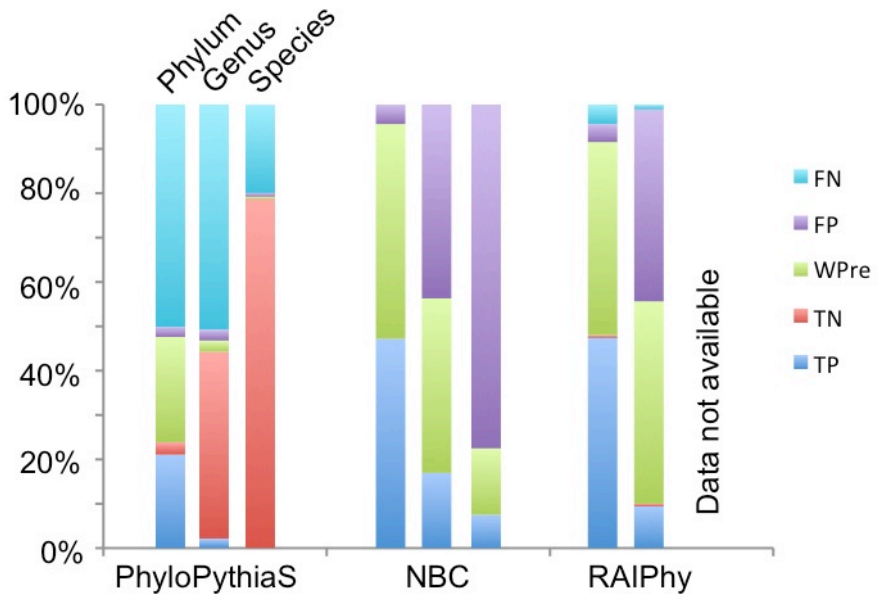


Figure 5.10. Performance of composition-based methods on sequences that did not have significant matches in the reference database. Significant lower accuracies were observed on these sequences compared to those with significant matches, e.g., NBC's accuracy was 16.9% versus 36.9% at the genus level, respectively; presumably due to the fact that sequences with no significant matches tend to represent taxon-specific genes, which are often the product of horizontal gene transfer (27). The types of predictions are color-coded in the same way as in Figure 5.5.

REFERENCES

1. Wrighton KC*, et al.* (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337(6102):1661-1665.
2. Arumugam M*, et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.
3. Iverson V*, et al.* (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335(6068):587-590.
4. Glass EM, Wilkening J, Wilke A, Antonopoulos D, & Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor protocols* 2010(1):pdb prot5368.
5. Sun S*, et al.* (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research* 39(Database issue):D546-551.
6. Amann RI, Ludwig W, & Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews* 59(1):143-169.
7. Stepanauskas R (2012) Single cell genomics: an individual look at microbes. *Curr Opin Microbiol*.
8. DeSantis TZ*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* 72(7):5069-5072.
9. Cole JR*, et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research* 37(Database issue):D141-145.
10. Konstantinidis KT & Tiedje JM (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 10(5):504-509.
11. Rosen GL, Reichenberger ER, & Rosenfeld AM (2011) NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27(1):127-129.
12. Patil KR*, et al.* (2011) Taxonomic metagenome sequence assignment with structured output models. *Nature methods* 8(3):191-192.
13. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, & Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome research* 21(9):1552-1560.
14. Krause L*, et al.* (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic acids research* 36(7):2230-2239.
15. Zhaxybayeva O & Doolittle WF (2011) Lateral gene transfer. *Current biology : CB* 21(7):R242-246.
16. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.
17. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102(7):2567-2572.

18. Bowman AW (1994) Multivariate Density-Estimation - Theory, Practice and Visualization - Scott,Dw. *J Classif* 11(2):261-262.
19. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* 5(3):e9490.
20. Larsen PE, Field D, & Gilbert JA (2012) Predicting bacterial community assemblages using an artificial neural network approach. *Nature methods* 9(6):621-625.
21. Konstantinidis KT & Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* 187(18):6258-6264.
22. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* 2(10):1052-1065.
23. Caporaso JG*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335-336.
24. Kostic AD*, et al.* (2012) Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome research* 22(2):292-298.
25. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome research* 12(4):656-664.
26. Linnaeus C (1753) *Systema naturae* (Haak, Lugduni Batavorum) 1st Ed.
27. Daubin V & Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res* 14(6):1036-1042.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. DOE Office of Science, Biological and Environmental Research Division (BER), Genomic Science Program, Award No. DE-SC0006662 and by U. S. National Science Foundation under Award No 1241046.

CHAPTER 6

Soil microbial community responses to a decade of warming as revealed by comparative metagenomics

INTRODUCTION

Extant biodiversity has been recognized as telling the evolutionary history of life while also providing an evolutionary scaffold for the future. Consequently, one of the great challenges in the natural sciences is to better understand how the inventory of biodiversity determines the evolutionary path(s) that will shape the future. Prokaryotes (bacteria and archaea) represent the largest component of the biodiversity on Earth, not only in terms of gene and sequence diversity but also in total biomass; yet, their natural communities remain the "black box" of biodiversity (1). Soil prokaryotic communities, in particular, are composed of thousands of distinct species (2-4), each of which typically makes up a rather small fraction (i.e., <0.1%) of the total community and encodes hundreds of species-specific genes of unknown function (5, 6). How such complex communities respond to natural as well as anthropogenic fluctuations in the environment, including major perturbations such as global climate change, is poorly understood. For instance, little is known about what genomic adaptations, interactions and feedback mechanisms occur among members of the community during perturbations that simulate the predicted effects of climate change such as increased ambient temperatures and carbon dioxide (CO_2) concentrations (7, 8). Advancing these issues would also lead to a more predictable understanding of the role of the soil ecosystem and its biota for models of climate change.

The recent advances in sequencing technologies provide an opportunity to comprehensively assess community-wide shifts in response to environmental perturbations. Several studies have recently attempted to quantify the impact of elevated CO_2 levels (9), input of exogenous organic matter of varied degrees of recalcitrance (10,

11), and different regimes of nitrogen fertilization (12) on soil microbial communities. Most of these studies analyzed small subunit ribosomal RNA (16S rRNA) gene sequences recovered from the indigenous communities and revealed important differences in community composition in response to the perturbations. Although the 16S rRNA gene successfully serves as the best phylogenetic marker to identify the taxa present in a sample, it represents just one of the genes in the genome while important levels of functional and ecological differentiation frequently underlie identical 16S rRNA gene sequences (13). Thus, in order to better understand and model the functional significance of the shifts in species composition observed, it is important to analyze the whole genome level. A recent study highlighted the power of whole genome approaches by linking methane (CH_4) emissions from a thawed permafrost soil to specific genes and species of the indigenous communities (14).

In this study, we report on the whole-genome shotgun metagenomic analysis of microbial communities of temperate grassland soils (well-aerated soil, in Oklahoma, USA) that experienced 2°C infrared heating for 10 years. Our analyses show that even such mild and relatively short-lived perturbations can induce significant changes in the composition and functional potential of the indigenous microbial community as well as the interactions among community members. In the Midwestern grassland soils studied here, these community changes appear to promote the respiration of the additional plant-derived soil carbon fixed as an effect of warming, which has important implications for better understanding and modeling the effects of global climate change.

MATERIALS AND METHODS

Experimental setup and sampling

This study was conducted at the Kessler Farm Field Laboratory (KFFL) located at the Great Plain Apiaries in McClain County, Oklahoma, USA (34°58'54"N, 97°31'14"W). This is an old field tallgrass prairie that had been abandoned from agriculture for more than 30 years. The herbivores were excluded at this site in 2002 to prevent grazing. The grassland is dominated by C4 grasses (*Andropogon gerardii*, *Sorghastrum nutans*, *Schizachyrium scoparium*, *Panicum virgatum*, and *Eragrostis* spp.), C3 forbs (*Ambrosia psilostachyia* and *Xanthocephalum texanum*), and C3 annual grass (*Bromus japonicas*) (15, 16). Based on Oklahoma Climatological Survey from 1948 to 1999, the mean annual temperature at this site was 16.3ºC with the lowest, 3.3ºC, in January and the highest, 28.1ºC, in July, while the mean annual precipitation was 967mm, which was highest in May and June (240 mm) and lowest in January and February (82 mm). The soil is silt loam (36% sand, 55% silt, and 10% clay in the top 15 cm) and part of Nash–Lucien complex, which typically has high fertility, neutral pH, high available water capacity, and a deep moderately penetrable root zone.

The experiment was established in November 1999 with a blocked split-plot design, in which warming is a primary factor. Two levels of warming (ambient and +2ºC) were set for six pair of 1 m ! 1 m subplots by utilizing a “real” or “dummy” infrared radiator (Kalglo Electronics, Bethlehem, Pennsylvania) as the heating device, suspended 1.5m above the ground in warming plots. In control plots, the dummy infrared radiator is also suspended (but not functional) to exclude the shading effect of the device itself. The 12 soil samples were taken from 0-15 cm layer in 6 warming and 6 control plots in

October 2010. Each sample was composited from two soil cores (2.5 cm diameter ! 15 cm deep) and was sieved by 2mm sieves prior to being transported to the laboratory and stored at -80^{o}C. The annual temperature measured on actual soil samples was on average 1.2 ^{0}C higher in heated vs. control plots at 15cm depth (Table 6.3), confirming that our heating strategy was effective.

DNA extraction and sequencing

Ten grams (10 g) of soil was used for DNA extraction for each sample. DNA was extracted by freeze-grinding mechanical lysis as described previously (17) and was purified using a low melting agarose gel followed by phenol extraction. DNA quality was assessed based on the ratios of 260/280 nm and 260/230 nm absorbance by NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE) and the yield of purified DNA was in the range of 1-2 " g/g, depending on the sample considered. Final DNA concentrations were quantified by PicoGreen (18) using a FLUOstar Optima (BMG Labtech, Jena, Germany). Library preparation, cluster generation, and sequencing followed the instructions of the manufacturer (Illumina, Inc., San Diego, CA). Briefly, 1-2 " g of genomic DNA from each of the 12 soil samples was used to construct 350 bp long insert sequencing libraries (hence, the derived metagenomes represent the microbes present in about 1g of soil sample). DNA was first fragmented using the Covaris system (Covaris, Woburn, Massachusetts), and was end-repaired and ligated with adapters using the TruSeq DNA Sample Prep Kit. The resulting libraries were subjected to cluster generation using TruSeq PE Cluster Kit v2 on cBot and massively parallel sequencing using TruSeqSBS 200 cycle kit on the Hi-Seq 2000 instrument.

Physicochemical measurements

The soil pH, soil moisture, total soil C and N, soil labile and recalcitrant C, soil microbial biomass, soil ammonia and nitrate content, soil temperature, soil nitrification and denitrification potentials were measured as previously described (19-23).

Sequence processing

Illumina read trimming was carried out using an in-house python script (available from the authors upon request). The script trimmed each pair-ended read based on the following sequential steps: i) trim the read from both 5'- and 3'-ends until a base with Phred score >20 is met; ii) use a sliding 3bp-long window to examine the quality of the remaining sequence from step (i). If the average Phred score of a window is lower than 20, create a cut at that position; iii) keep the remaining sequence if it is longer than 50bp and contains no more than 1 N (ambiguous base); otherwise discard the sequence. Only reads that both sister reads passed the trimming cut-offs were used for further analysis (Table 6.1). This method was applied on all samples, including the publicly available metagenomes used in this study for consistency purposes.

Table 6.1. Sequencing and assembly statistics for each soil metagenome. Sample replicates (denoted by "rep" in the Table) denote technical replicates, i.e., same library sequenced independently in two different lanes of the Illumina HiSeq2000 instrument.

Sample	Size (! 10^6 reads)		Assembly (Kbp)[a]			Taxonomy composition (%)				
	Raw	Trimmed	N50	Max	Total	Bacterial	Archaeal	Viral	Eukaryotic	Other
C1.rep1	47.1	38.3	0.79	11.47	517	90.8	1.7	0.2	7.1	0.2
C1.rep2	39.8	32.9	0.71	5.89	305	90.8	1.7	0.1	7.3	0.1
C2	190.1	134.3	0.76	39.94	3,320	90.9	1.7	0.2	7.0	0.2
C3.rep1	96.6	70.8	0.68	6.59	976	91.4	1.8	0.2	6.6	0.1
C3.rep2	81.1	62.8	0.73	7.13	1,482	89.7	1.9	0.2	8.1	0.2
C4	150.4	118.8	0.69	4.71	1,223	90.6	1.6	0.2	7.2	0.1
C5	134.9	102.7	0.73	7.67	3,656	91.2	1.8	0.2	6.9	0.1
C6	176.3	127.0	0.81	18.16	1,741	90.8	1.7	0.2	7.2	0.1
H1	107.0	83.9	0.72	27.19	4,616	91.0	1.7	0.2	7.0	0.1
H2	110.1	85.8	0.72	9.03	860	90.9	1.8	0.2	7.0	0.2
H3	103.9	81.5	0.67	3.6	936	91.0	1.7	0.2	6.9	0.1
H4.rep1	58.3	45.7	0.69	3.36	449	91.1	1.8	0.2	6.9	0.1
H4.rep2	57.9	45.6	0.72	3.63	459	91.0	1.7	0.1	6.9	0.2
H5	104.4	82.1	0.71	15.84	2,554	90.6	1.7	0.2	7.4	0.1
H6	113.4	89.9	0.67	3.43	752	90.8	1.7	0.2	7.2	0.1

[a] Only contigs longed than 500bp were used.

Table 6.2. Site information where samples were taken.

Item	Unit, if applicable	Content
Investigation type		Metagenomics
Location		Oklahoma, USA
Latitude, longitude		34°58#54$N, 97°31#14$N
Collection date		August, 2010
Environment (biome)		Grassland
Environmental package		Soil
Sample collection device		Soil corer
Sampling depth	cm	0-15
Sequencing method		Illumina HiSeq-2000
Current land use		Grassland
Current land vegetation		Grass
Previous land use		Pasture
Sample weight for DNA extraction	g	10
Mean annual temperature	°C	16
Total organic carbon, avg	%	1.24
Total organic C method		Shimadzu TC analyzer
Total nitrogen, avg	%	0.11
Total nitrogen method		Shimadzu TC analyzer

Table 6.3. Physicochemical measurements of soil samples.

Item	Year	H1	H2	H3	H4	H5	H6	C1	C2	C3	C4	C5	C6
Moisture (%)	08	17.5	16.0	13.2	13.0	14.5	13.2	15.0	16.3	15.5	14.5	14.8	15.2
Average annual soil temperature (°C)	07	16.1	16.3	15.7	15.6	17.0	16.7	14.7	15.0	15.0	14.8	15.6	15.0
pH	08	8.03	8.08	8.03	7.64	6.61	7.07	7.98	7.87	8.02	7.75	6.57	7.70
NO_3	08	3.17	4.97	3.23	8.23	5.29	1.45	5.77	4.06	2.12	14.8	2.86	2.61
NH_4	08	6.02	4.25	3.96	7.43	22.5	9.9	8.31	7.17	5.21	6.66	11.9	8.12
Total soil N (%)	08	0.13	0.13	0.09	0.14	0.17	0.11	0.13	0.16	0.09	0.22	0.14	0.21
Total soil C (%)	08	2.81	3.96	2.56	2.54	2.12	1.28	2.75	3.63	2.58	3.27	1.57	2.44
Total soil organic matter (%)	08	4.84	6.83	4.41	4.38	3.66	2.21	4.74	6.26	4.45	5.64	2.71	4.21
Microbial N (mg/kg)	03	90.9	67.3	90.6	125.6	90.7	74.4	87.4	28.0	67.0	67.1	71.5	72.5
Microbial C (mg/kg)	03	676.3	770.8	847.5	715.3	570.7	810.7	352.2	174.1	418.9	488.4	457.7	591.3
Soil labile C pool 1 (mg/kg)	08	3.75	2.80	1.66	2.49	3.59	2.23	2.26	4.66	2.31	3.18	2.43	1.79
Soil labile C pool 2 (mg/kg)	08	3.83	3.79	2.33	4.95	6.64	4.49	2.00	4.36	2.88	4.66	5.39	3.76
Recalcitrant C pool (mg/kg)	08	4.34	3.87	3.41	10.5	5.32	4.66	4.51	3.90	3.98	10.0	4.15	2.47
Soil labile N pool 1 (mg/kg)	02	32	32	24	28	28	24	28	28	16	25	24	28
Soil labile N pool 2 (mg/kg)	02	84	72	68	75	75	76	80	88	64	76	88	60
Soil Respiration	07	2.86	2.13	2.48	3.36	1.97	3.07	2.10	1.89	2.23	2.71	1.64	1.68
Nitrifiction rate	10	0.01	0.01	0.07	0.15	0.01	0.01	0.01	0.01	0.01	0.10	0.02	0.02
Denitrification rate	10	7.30	6.35	8.63	11.3	-	10.1	6.31	6.90	5.45	11.3	12.3	5.75

Table 6.4. Plant information for the soil samples analyzed in this study.

Item	Year	H1	H2	H3	H4	H5	H6	C1	C2	C3	C4	C5	C6
Total peak biomass (July or August, g/m^2)	07	281	268	262	334	288	347	192	173	181	453	200	196
5 year averaged total peak biomass (July or August, g/m^2)	03-07	225	217	240	373	327	315	217	227	240	317	307	223
Forbs (C3; g/m^2)	07	75	72	71	86	76	88	22	5	12	256	29	25
Grasses (C4; g/m^2)	07	206	195	191	248	212	259	170	168	169	198	171	171
5 year averaged forbs (C3; g/m^2)	03-07	49	43	50	39	50	56	37	24	26	172	41	42
5 year averaged grasses (C4; g/m^2)	03-07	176	174	190	334	277	259	180	203	214	145	266	181
C3 plant leave N (%)	08	0.83	1.27	1.19	1.67	1.61	1.04	1.07	1.32	1.00	2.26	1.29	1.64
C4 plant leave N (%)	08	0.67	0.48	0.65	0.76	0.88	0.76	0.75	0.61	0.80	0.77	0.85	0.82
C3 litter N (%)	06	0.51	0.40	0.48	0.99	0.45	0.41	0.59	0.75	0.46	0.75	0.60	0.73
C4 litter N (%)	06	0.57	0.46	0.63	0.54	0.52	0.43	0.58	0.56	0.83	0.72	0.54	0.48
Root biomass (g/m^2)	05	290	111	131	139	335	215	150	185	116	92	187	128
Root N (%)	05	0.58	0.51	0.60	0.77	0.49	0.68	0.65	0.49	0.67	0.94	0.71	0.60
Normalized difference vegetation index	07	0.49	0.39	0.47	0.45	0.50	0.55	0.39	0.42	0.47	0.53	0.51	0.45
Leaf area index	07	1.45	0.95	1.28	1.13	1.46	1.59	1.45	1.35	1.45	1.24	1.13	1.05

16S rRNA gene encoding read analysis

To identify reads encoding 16S rRNA gene fragments, we used the nucleotide sequence of the 16S rRNA gene of *E. coli* strain K-12 as reference (GenBank accession number: NC_000913). Reads were identified using a Blastn (24) search (settings: "-m 8 –v 1 –b 1 –X 150 –q -1 –F F –e 1e-12", remaining parameters at default settings) and the *E coli* sequence as a reference database and a cut-off for a match of at least 70% nucleotide sequence identity and 50bp alignment length. The matching metagenomic reads were extracted and searched against a reduced version of the GreenGenes database (25), in which all GreenGenes sequences were first pre-clustered at the 99% nucleotide sequence identity level (OTUs). Pair-ended reads with both ends matching the same OTU with higher than 99% nucleotide identity were assigned to that OTU. The relative abundance of different genera/phyla in each sample was quantified by the number of reads assigned to each taxon, normalized by the sample size (assuming each community/sample is characterized by the same rRNA copy number per genome, on average). The normalized counts for genera/phyla were subjected to PCoA analysis as implemented in MatLab and genera/phyla that were significantly differentially present were identified using the paired *t*-test from statlib in Python (Figure 6.1).

To perform FastUniFrac analysis (26), the 16S rRNA gene encoding reads from the previous step were aligned to *E. coli* 16S rRNA gene sequence using CLUSTALW2 (27). A 50 bp window was used to count the number of aligned reads at different positions across the gene sequence (Figure 6.7A). Following visual inspection, four regions with high read coverage (755-804bp, 882-931bp, 1061-1110bp, and 1187-1236b; in the

proximity of the V5 to V8 regions) were selected and the corresponding 50 bp-long, fully overlapping, aligned sequences were extracted and served as the input alignment for FastUniFrac. The four regions targeted are not fully overlapping with the V5 to V8 regions and typically show higher sequence conservation thus, underestimating OTU diversity. However, our FastUniFrac analysis was restricted to the phylum level, where the varied degree of sequence conservation among the regions does not have a significant effect.

Non 16S rRNA gene encoding read analysis and sequence discrete populations

Reads not encoding 16S rRNA gene fragments were searched against all the complete and draft bacterial genomes available in NCBI at the end of 2011 (www.ncbi.nlm.nih.gov) using BLAT (cut-off for a match: E-value <1e-10, alignment length >50bp, and nucleotide identity >80%). Only sister reads that had the same genome sequence as their best match were considered for further analysis. Reads were assigned to a genus based on the taxonomic classification of the genome that provided the best match. The number of reads recruited by each genome was normalized for the sample size by dividing by the total number of reads of the sample. The normalized read counts were used as a proxy of the genus abundance in the corresponding sample. The correlation of the abundances of any two genera between all samples was calculated using the Pearson correlation (statlib in Python), and a genus co-occurrence network was built using pairs of genera with correlation coefficient > 0.7 and *P*-value < 0.01 (Figure 6.12). To identify discrete populations, reads were mapped onto assembled contigs from

each metagenome (see below), using Blastn and a cut-off for a match of at least 70% nucleotide sequence identity and 50bp alignment length. We also used MG-RAST (28) on unassembled reads to estimate the fraction of bacteria, archaea, viruses, and eukaryotes in each metagenome. The results showed that the non-bacterial fraction was rather small (<10% of the total) and the viral and eukaryotic fractions were similar across the samples (Table 6.1). It is unlikely that genome size or domain-level differences in abundance have significantly affected our results and hence we have preferentially focused on the bacterial fraction in our results and discussions.

Metagenome assembly and gene annotatio

The assembly of metagenomes was carried out using a hybrid protocol that combines Velvet (29), SOAPdenovo (30), and Newbler 2.0 (Table 6.1), as described previously (31). MetaGeneMark (32) was employed to identify protein-coding genes in assembled contigs, and a Blastx search against nr database was used to functionally annotate the genes. Protein-coding genes on individual reads were identified by FragGeneScan (33) using Illumina 0.5% error model and default settings. The amino acid sequences of these genes were searched against the SEED database (34) by BLAT (35) using default settings. The best match for each read, when better than a minimum cut-off of e-value <1e-10, alignment length >20 amino acids (a.a.), and a.a. identity >30%, against the SEED genes was recorded and the number of best-matching reads was taken as a proxy of the abundance of the SEED genes and subsystems in each sample, after normalizing for the sample size.

Differentially present pathways

To identify pathways that were significantly differentially present between the control and the heated samples, we employed an approach combining re-sampling techniques, the DESeq package (36), and binomial testing (Figure 6.9). A jackknife method was used to generate all combinations of three control and three heated samples (from a total of six samples in each set). For each combination, a count table was generated. Each row of the table represented a SEED subsystem, each column represented a sample, and each element was the normalized number of reads from the sample assigned to the SEED subsystem from the previous BLAT search (counting reads that were assigned to all genes that constitute the subsystem). DESeq was then used to detect the difference between heated and control samples for each SEED subsystem. For the same SEED subsystem, the log2 fold changes from the DESeq analysis of all combinations of samples followed a distribution, the mean of which represented the best estimate of fold change, while the variance reflected the reliability of the estimate. A binomial test was carried out to test the significance of the log2 fold changes, and the *P*-value was adjusted for false discovery rate using the Benjamini-Hochberg method. SEED subsystems that recruited at least 100 reads in one sample with *P*-value <0.01 and a fold change > 5% are reported in Table B1.

To test whether the observed changes in a SEED subsystem relative abundance were community-wide or instead attributable to the emergence/disappearance of a few taxa, we first extracted the nucleotide sequences of the genes that constitute the pathway in question. A Blastx search was carried out to identify the homologs of each gene in nr database (cut-off: 80% alignment length and 70% a.a. identity). The potential nr

homologs with ambiguous tags such as “putative”, “hypothetical”, “possible” were removed. The remained sequences and the SEED genes were merged to form a reference database, against which the pair-ended reads were searched using Blastx (same cutoff). For each a gene, a multi-sequence alignment (MSA) was constructed of all reference homolog sequences by MUSCLE (37), and the metagenomic reads were aligned to the MSA based on their Blastx search results. A 20 amino acid window was applied to scan the final MSA read alignment. Regions with coverage (number of reads per amino acid) two standard deviations higher than average coverage of the whole gene sequence were discarded, because it is likely that such regions encode universally conserved motifs and thus recruit false positive reads (i.e., reads representing other genes). The regions with coverage of at least 300 reads were extracted, and the aligned a.a. sequences were used as a guide to produce the corresponding codon by codon nucleotide sequence alignment. FastTree (38) was employed to reconstruct the phylogenetic tree of all reads based on their nucleotide alignments. CD-HIT (39) was used to collapse the reads into OTUs based on the level of nucleotide sequence identity (three levels were used, 80%, 85% and 90% identity). The results based on different identity thresholds were highly similar (data not shown); thus, 80% was used for further analysis. Note that we did not use higher identity thresholds (e.g., 95%), even though such thresholds would have captured better the nucleotide identity range that corresponded to the genetic discontinuities observed between sequence-discrete populations, because only a few clades containing enough sequences for downstream analysis were observed with higher thresholds. The percent of the total reads in each OTU that represented reads of control samples was counted and reported in Figure 6.12-13.

Soil comparative metagenomic analysis

Publically available metagenomes used in this study included: Alaska permafrost samples (14), which were downloaded from the ftp site of the Joint Genome Institute (www.jgi.gov); ocean samples (GOS257, GOS258, GOS259, GOS262, GOS263, and GOS264) from the Global Ocean Sampling expedition (40), which were downloaded from the trace archive database (www.ncbi.nlm.nih.gov/Traces/trace.cgi); and planktonic samples from Lake Lanier (Atlanta, GA) from our previous study (41). The permafrost and freshwater lake raw reads (Illumina) were trimmed as described above, while the ocean samples (Sanger sequences) were trimmed with comparable standards and randomly sampled to produce 100bp long reads that were comparable to the Illumina reads.

To compare the complexity of samples from different environments, we applied an in-house developed parallelized algorithm. The algorithm subsamples and calculates the portion of non-unique reads at a given amount of reads (sequencing throughput). By varying the input amount of reads to this algorithm (to the amount that is computational tractable to search all reads against themselves) and performing multiple replicates (resampling), a saturation curve that resembles a rarefaction curve is produced, which reflects the complexity of the sample. An exponential regression is subsequently fit to the calculated "read uniqueness" values, with the goal to minimize the summed squared error (SSE). The degree of complexity of different samples is quantitatively assessed by comparing the slopes of the regression lines (the lower the slope, the more diverse the community sampled, Figure 6.1A). The exponential function used in the regression analysis was:

$$y = (1 - e^{-\alpha x})^{\beta},$$

where y is the percentage of non-unique reads, x is the amount of reads in gigabytes. The fitting was carried out in R.

To compare the datasets in terms of gene relative abundance, all trimmed reads in each dataset were then searched against the genes recovered on the assembly of the dataset (cut-off for a match: Blastn; e-value: <1e-12; nucleotide identity: >70%, and alignment length: >50bp), and the number of reads matching each gene represented the relative abundance of the gene in the corresponding sample (normalized for the sample size). A hierarchical clustering of the samples was carried out based on the abundances of all genes found in all samples (Figure 6.4).

The Oklahoma and Alaska samples were also compared at the read (read vs. read) and contig coverage (read vs. contig) levels, using the same Blastn parameters and cut-offs as described above. In the read comparison, we assessed what fraction of the reads of one dataset was shared by the other dataset. The taxonomic information of the read was obtained from its best match against the available genome sequences in nr database as described in the previous section as long as the match was of at least 95% nucleotide sequence identity. In the contig comparison, the percent of a contig covered by a metagenome was estimated based on the length of the contig covered by metagenomic reads divided by the total length of the contig.

RESULTS AND DISCUSSION

Community complexity, comparisons to other habitats and sequence-discrete populations.

Community DNA was extracted and sequenced from six replicate samples representing the heated soils (H1 to H6) and six samples representing the adjacent unheated soils (controls; C1 to C6). Sequencing was performed using the Illumina HiSeq-2000 platform and yielded about 10-15 Gb of short pair-ended (PE) sequence data per sample (100 X 100 bp; Tables 6.1 and 6.2). Prokaryotes represented the great majority of each community sampled based on the fact that an average of 92% of the total genes recovered had best matches (average amino acid identity ~35%) against bacterial and archaeal genomes in MG-RAST database (42). Due to the high complexity of the soil communities in terms of species richness, the assembly of the metagenomes using established algorithms such as Velvet (29) or our recently described hybrid protocol (31) yielded only short contigs (e.g., N50 = 500 to ~1,000bp) while the majority of the reads remained unassembled (Table 6.1). The community complexity was quantitatively evaluated based on the fraction of unique reads in randomly drawn subsets of the data and was compared to the complexity of metagenomes from permafrost soils and aquatic habitats. The soil community was estimated to be eight times more complex in terms of species richness than the planktonic communities of the open ocean or Lake Lanier (Atlanta GA), assuming an average genome size three times larger in the soil than the aquatic communities (43), and two and a half times more complex compared to the permafrost soil community reported recently (14) (assuming similar genome sizes; Figure 6.1A). Lake Lanier was previously estimated to contain about 500 operational taxonomic

units (OTUs; defined at the 97% 16S rRNA gene sequence identity level) and to show comparable diversity to the open ocean (41). Therefore, the number of OTUs in a gram of soil for our samples was extrapolated to be about 4,000, which is in

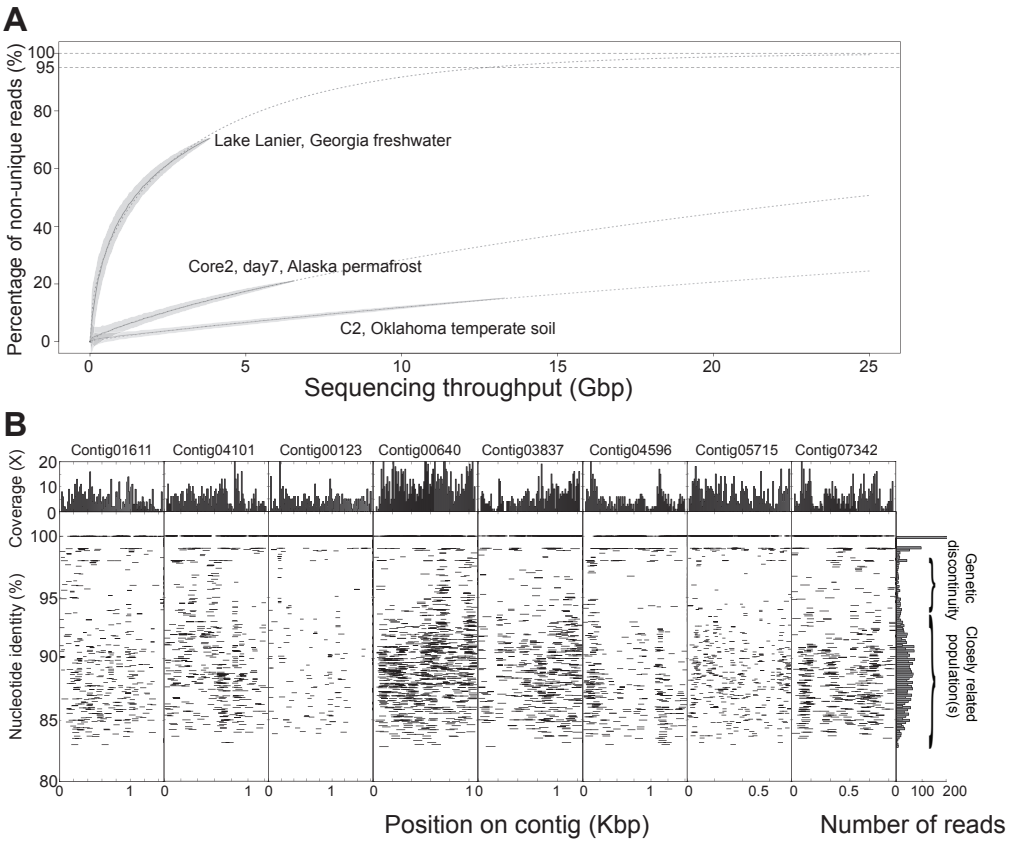


Figure 6.1. Soil community complexity and dominance of sequence-discrete populations. (**A**) The percentage of non-unique reads (defined as reads with at least one match at 95% nucleotide identity level; y-axis) as a function of the size of randomly drawn subsamples from whole-genome shotgun metagenomes of different habitats (x-axis) is shown. Solid lines represent averages, shadowed regions represent 1 standard deviation from the average based on 1,024 random subsamples, and dashed lines represent the fitted exponential curves. (**B**) Eight contig sequences, assembled from a control metagenome (C5), were used as references to recruit reads, essentially as described previously (44). The graph shows the identity of each read against the

reference sequence (y-axes) plotted against the position of the read on the reference sequence (x-axes). The histogram on the top represents the read coverage across the length of the contigs; the histogram on the right represents the number of reads recruited per unit of nucleotide identity. Note the genetic discontinuity typically observed in the 95-98% nucleotide identity range.

agreement with previous estimates (2), but it likely represents an underestimate of species richness due to the high conservation of 16S rRNA gene sequences (45). Although the average amount of sequence diversity (or community complexity) was comparable between heated and control samples, heated samples showed significantly less variability in their diversity estimates (Figure 6.2), indicating that warming drove the community to a more defined (non-random) direction. We also estimated that about 337 Gb of sequencing would be required to cover 95% of the sequence diversity within each sample used in the study (95% confidence interval: 331.6-341.7 Gb). Finally, the twelve samples typically shared >90% of the OTUs recovered based on 16S rRNA gene fragments encoded on metagenomic reads and the majority of the non-16S rRNA gene encoding reads (Figure 6.3). Given also that sequencing did not saturate the total diversity in the samples, these results revealed that the communities sampled by the twelve samples were highly overlapping in terms of species present.

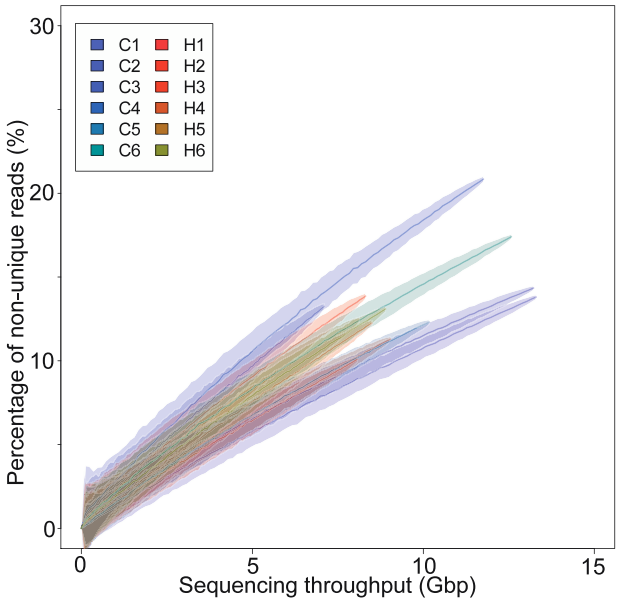


Figure 6.2. Community complexity in the samples used in this study. The percentage of non-unique reads (defined as reads with at least one match at 95% nucleotide identity level; y-axis) as a function of the size of randomly drawn subsamples of the metagenomes used in tis study (x-axis) is shown. Solid lines represent averages, shadowed regions represent 1 standard deviation from the average based on 1,024 random subsamples. Note that, although the average level of community complexity was comparable between heated and control samples, heated samples showed significantly less variability in their diversity estimate (red ribbons are contained within the blue ribbons), indicating that warming drove the community to a more defined (non-random) direction.

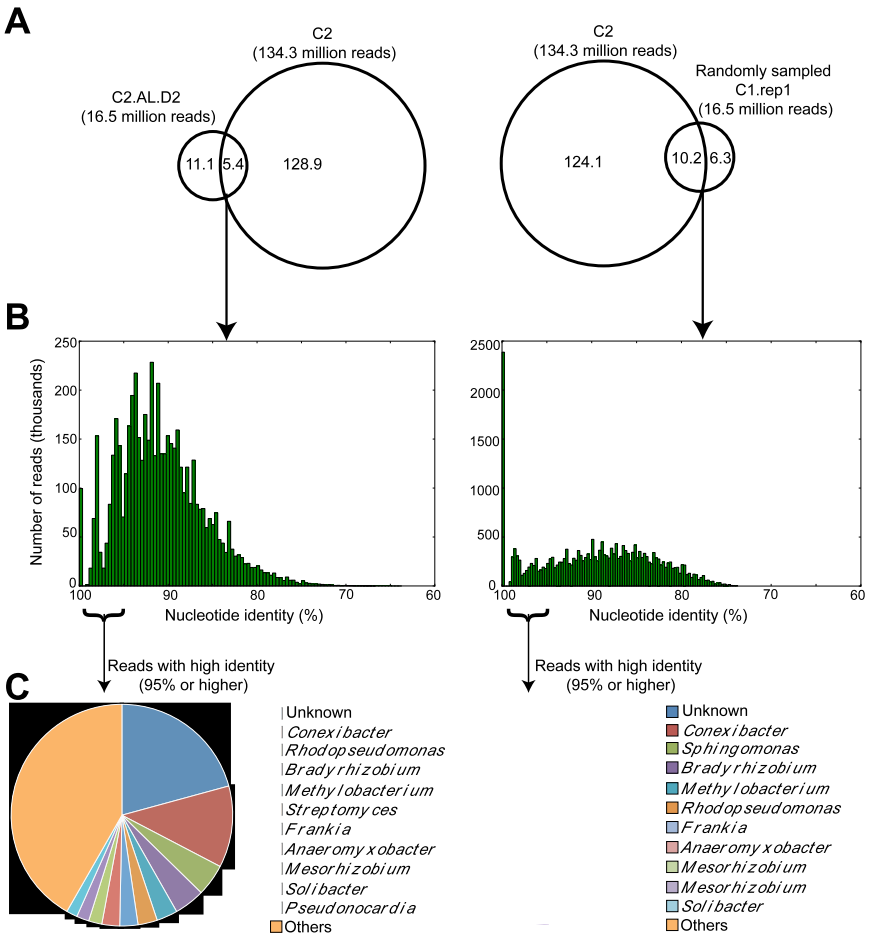


Figure 6.3. Overview of read-based comparisons between Oklahoma temperate soil and Alaska permafrost soil samples. The graph on the top (**A**) shows that about 1/3 of the reads from the Alaska sample (C2.AL.D2) have a match in Oklahoma sample C2 (left panel; cut-off, e-value: 1e-12 and alignment length: 50bp). The shared reads showed an average nucleotide sequence identity of ~90% (**B**). The reads that showed higher than 95% nucleotide sequence identity were extracted and searched against completed and drafted genomes from NCBI, and the relative abundances of the top 10 most abundant genera are shown in **C**. For comparison, a randomly drawn subsample from the Oklahoma C1 sample, which was similar in size to the Alaska sample C2.AL.D2, was also searched against the Oklahoma C2 (right panels). Note that the reads of the Oklahoma samples overlapped two times more frequently compared to the Alaska

sample (61.2% vs. 32.7% of reads overlapped, normalized to the smallest dataset; panel **A**) and the overlapping reads were of significantly higher nucleotide identity (panel **B**).

The temperate soil microbial communities sampled in this study were also compared to previously characterized communities from other habitats based on the abundance of genes recovered in the corresponding metagenomes. The analysis revealed that the former communities were distinctly different from those of the permafrost soil or aquatic environments. In fact, temperate and permafrost soil communities were only slightly more similar to each other compared to aquatic communities (Figure 6.4), indicating greater diversity among soil microbial communities compared to communities within other environments. Despite the high diversity among soil communities, a substantial part of the permafrost metagenomic reads (about 33% of the total) had high sequence identity (average ~90%) matches to reads of the temperate soil metagenomes (Figure 6.3), revealing the existence of a core set of (closely related) organisms that are present in soils of different type and geographic regions. Fragment recruitment plots against available genome sequences confirmed these interpretations and revealed that, although the majority of organisms in soils are still not represented by genome sequences in the public databases, several species with sequenced representatives appear to represent cosmopolitan soil inhabitants such as *Conexibacter, Rhodopseudomonas,* and *Bradyrhizobium*. These findings also corroborated previous 16S rRNA gene based findings (46). Interestingly, the permafrost community that underwent thawing (14) resembled more the temperate soil communities than the original (un-thawed) community

(Figure 6.5), indicating that temperature represents a major driver of community diversity in these soil ecosystems.

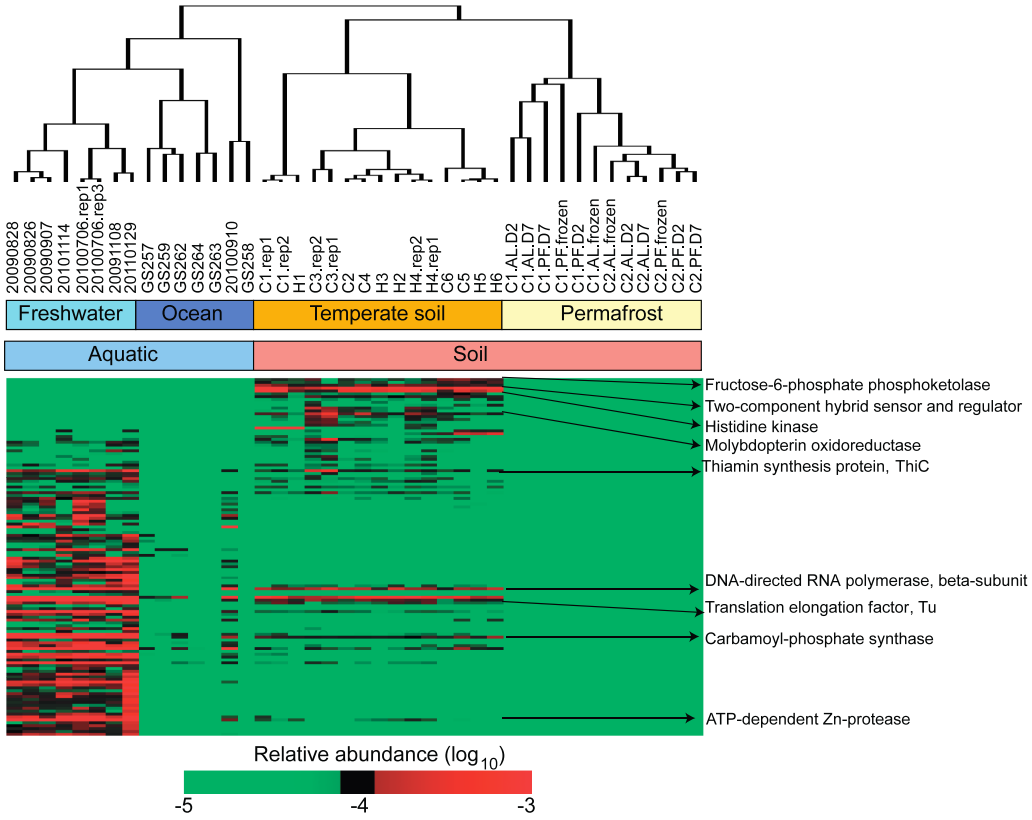


Figure 6.4. Clustering of metagenomics samples from different environments. Freshwater samples: Lake Lanier time series; Ocean samples: selected samples from the Global Ocean sampling Survey (GOS); Temperate soil samples: samples of this study; Permafrost samples: samples reported by Mackelprang and colleagues, using the naming of the samples provided by the authors, i.e., C1/C2, core 1/2; AL, active layer; PF, permafrost; D2/7, day 2/7). Clustering was based on the abundance of the top 500 most abundant genes (averaged across all samples). Gene abundance represented the number of reads matching the gene, normalized by the total number of reads in the sample.

A recent synthesis of the findings from previous metagenomic studies has revealed that microbial communities of many habitats such as the open ocean, freshwater ecosystems, the human gut, iron-reducing biofilms, and phosphorus removal bioreactors are predominantly composed of sequence-discrete populations and these populations may represent the important units of microbial diversity (47). Due to unavailability of appropriate datasets from soil communities, it has not yet been possible to test the applicability of these findings to soil ecosystems. Fragment recruitment plots using the contigs assembled from the temperate soil metagenomes as references revealed that sequence-discrete populations dominate the soil microbial communities, similar to other habitats (Figure 1B). Using the number of reads recruited by each contig (at the >95% nucleotide identity level) as a proxy for in-situ abundance and the phylogenetic affiliation of the housekeeping genes encoded on the contigs, we found that *Burkholderia sp*. *Conexibacter sp.* and *Rhizobacter sp.* were the most abundant species in the samples and no single species recruited more than 0.1% of the total reads.

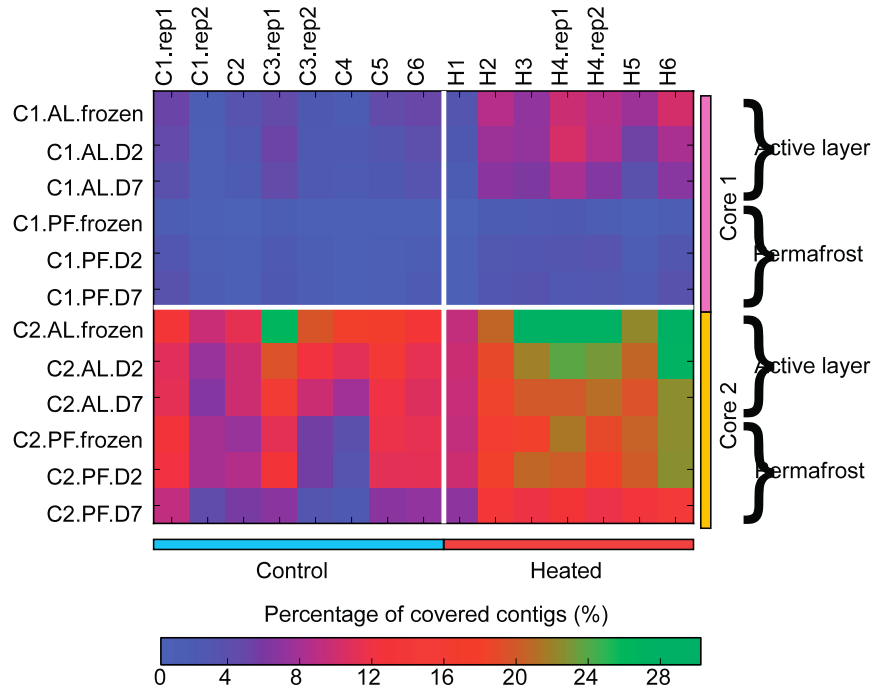


Figure 6.5. Coverage of assembled Oklahoma soil contigs by Alaska sample reads. Each row represents an Alaska permafrost metagenome and each column represents an Oklahoma soil metagenome; the percentage of Oklahoma soil contigs covered by Alaska reads is represented by the color intensity (see scale). For the naming of samples, see Figure S3. Note that, after quality read trimming, core 1 metagenomes were % of the size of the corresponding core 2 metagenomes, on average, and this probably accounts for the higher number of reads shared between core 2 and Oklahoma metagenomes and that the samples of the Alaska active (thawed) layer were more similar to Oklahoma soil samples compared to the deeper, permafrost layer.

Taxa distribution and co-occurrence patterns as an effect of warming.

At the sequencing depth of about 100 million PE reads per sample, no domain-level abundance differences were observed as an effect of warming (Table 6.1). Within the bacterial domain, however, several significant differences were observed. Based on

the 16S rRNA gene fragments recovered in the metagenomes, the most abundant phyla, i.e., *Proteobacteria, Acidobacteria, Planctomycetes,* and *Bacteroidetes*, were significantly differentially present between heated and control datasets [$P<0.05$ paired *t*-test, Benjamini-Hochberg (B-H) adjusted for false discovery in multi-testing; Figure 6.6A], albeit the difference was not dramatic, 2% on average. PCoA projection using phylum relative abundance confirmed that the samples from the two different treatments clustered separately, which was primarily attributable to the differences in the four aforementioned phyla (Figure 6.6B). These results were reproducible (Figure 6.7) when the analysis was performed using the FastUniFrac algorithm (26) and for different regions of the 16S rRNA gene sequence.

To determine the taxa whose abundance correlated (potential synergistic interactions) or anti-correlated (potential antagonistic interactions) as an effect of warming, a genus co-occurrence network of was constructed based on the abundances of all bacterial genera present in all 12 samples, the latter defined by the number of reads recruited by available representative genomes of each genus. The resulting network was composed of four major well-connected subgraphs, representing -*Proteobacteria*, #, /%*Proteobacteria*, *Actinobacteria*, and *Acidobacteria*/*Verrumicrobia* (Figure 6.6C). We observed mostly positive correlations within a subgraph whereas only negative correlations were typically observed between genera from different subgraphs. These findings indicate that genera of the same subgraph (corresponding

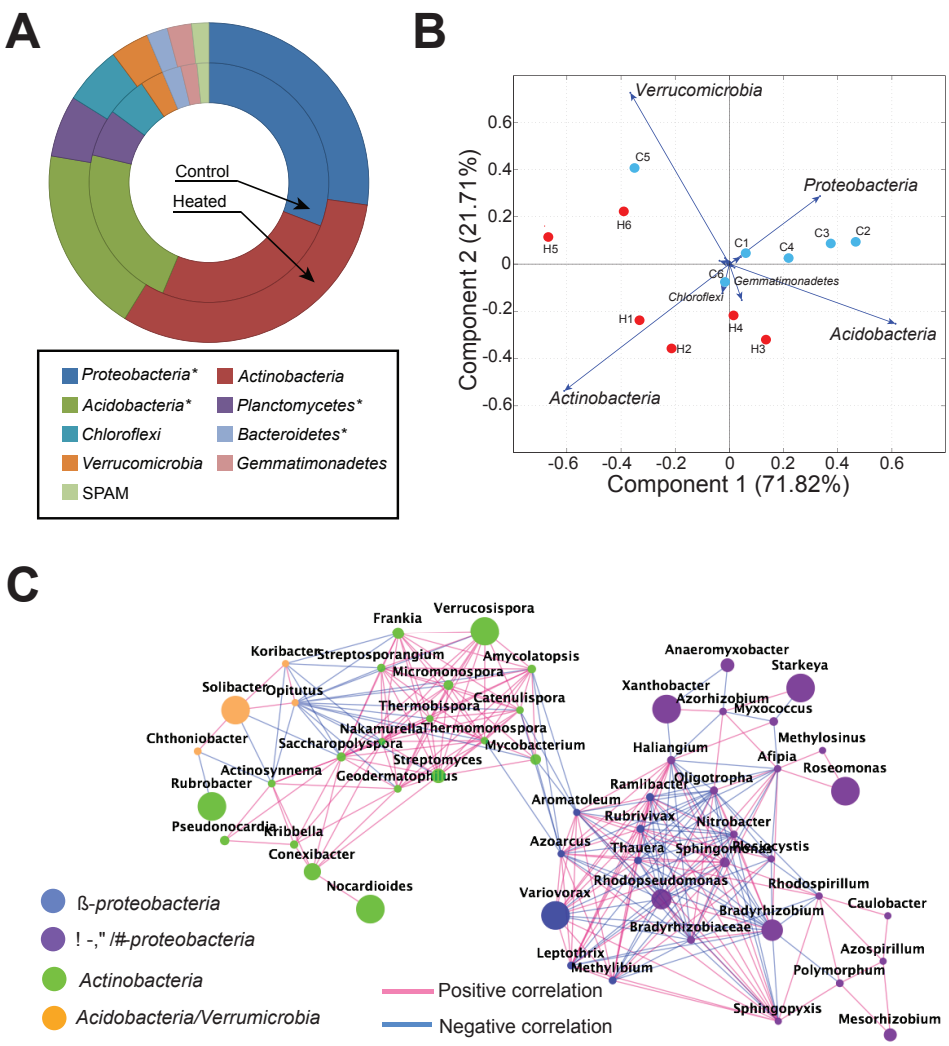


Figure 6.6. Taxa abundance shifts and co-occurrence network as an effect of warming. (**A**) Rings represent the average abundance (from six replicate samples) of phyla that made up at least 1% of the whole community; phyla that were significantly different in abundance between heated and control samples are marked by asterisks (P<0.05, two-tailed paired t-test). (**B**) PCoA analysis of the phylum abundance values separated heated from control samples. Arrows point to the direction that the corresponding phylum distinguishes the samples, and their length is proportional to the degree of the differences in abundance observed. (**C**) Co-occurrence network based on Pearson correlation analysis of the relative abundance of genera in the twelve soil samples (only genera with *Correlation coefficient* > 0.7 and P < 0.01 are shown). Each node

represents a genus, color-coded for the phylum the genus is assigned to; the size of the node is proportional to the genus average relative abundance across the twelve samples. Each line represents a significant correlation between the two genera it connects; red denotes positive correlations, blue denotes negative ones.

usually to the phylum or order levels) act synergistically among themselves, as a cohesive unit, and antagonistically to genera of different subgraphs upon environmental perturbations. Similar patterns were reported previously for soil communities, using different approaches. For example, based on BrdU-labeled 16S rRNA gene quantification by PhyloChip, Goldfarb and colleagues reported antagonistic interactions among bacterial phyla in response to carbon substrate addition (10); and Barberán and colleagues concluded that taxa within the same phylum tend to co-occur more often than expected based on the analysis of 151 soil 16S rRNA gene amplicon datasets (48). These observations across different soil and data types collectively reveal a hierarchical structure within soil microbial communities. It is important to note, however, that co-occurrence does not necessarily indicate direct interactions between the taxa, as co-occurrence may be due to hidden (indirect) factors. For instance, the more abundant phyla in heated samples include many copiotrophic (r-strategists) members (e.g., *Proteobacteria*) compared to phyla more abundant in the control (e.g., *Actinobacteria*, and *Verrucomicrobia*) and a higher concentration of available labile organic carbon was observed in heated samples (see also below). Functional experiments will be necessary to elucidate the mechanisms that underlie the correlation patterns revealed here.

Interestingly, a significant increase in the G+C% content of the 16S rRNA gene encoding sequences (~0.3% on average) or all reads (~1% on average) assigned to the

four most abundant phyla was observed in the heated metagenomes (P<0.05 paired t-test; B-H adjusted; Figure 6.8). The increased G+C% likely represents a genomic adaptation to exposure to higher temperatures as suggested in previous comparative genomics studies based on the higher thermo-stability of GC bonds relative to AT ones (49). Other factors, such as shifts in organic nitrogen availability in heated samples (see also below), might have also contributed to the higher G+C% content of heated datasets (50).

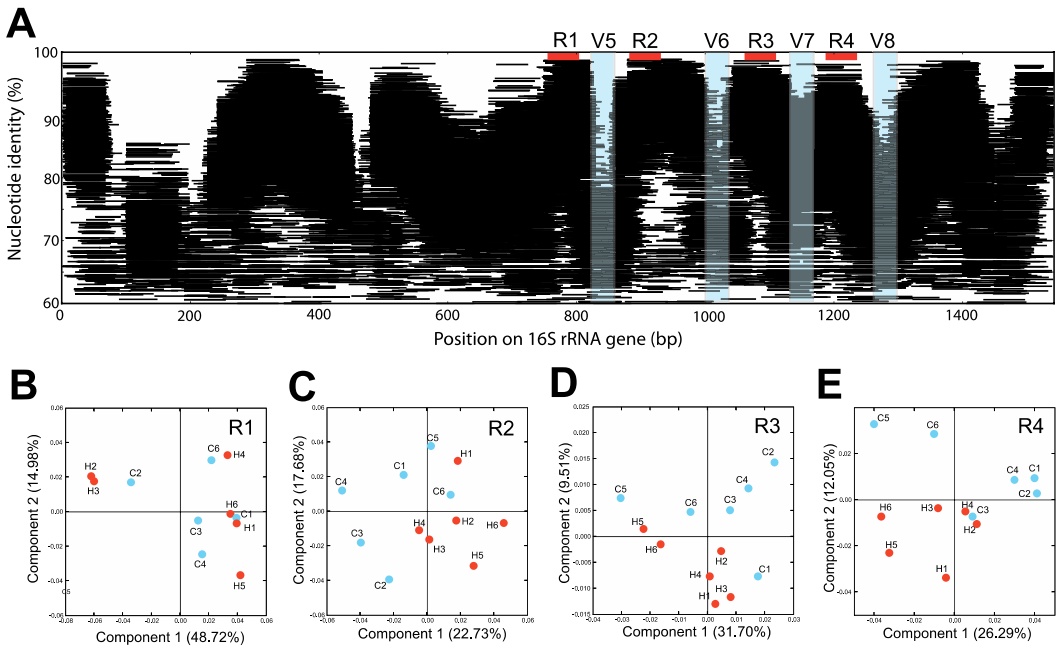


Figure 6.7. 16S read recruitments and FastUniFrac analysis. (**A**) 16S rRNA gene encoding reads were recruited to *E. coli* 16S rRNA gene reference sequence and are plotted based on their nucleotide identity (y-axis) and aligned position (x-axis) on the reference. (**B-E**) Four 50bp long regions with the highest number of recruited reads were selected for PCoA analysis in FastUniFrac (marked by the red bars in **A** and referred to as R1 to R4; these regions are adjacent to the V5 to V8 regions as shown in Panel A, respectively), and the results for each region are shown in the bottom panels.

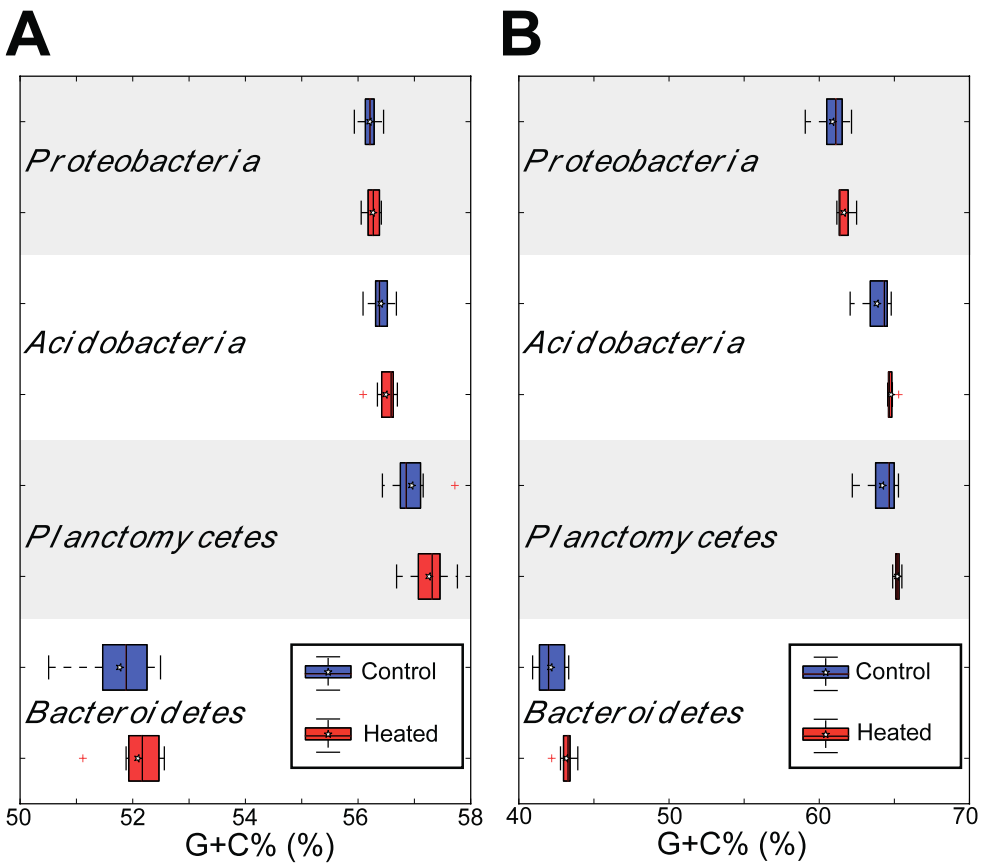


Figure 6.8. G+C% differences between control and warming samples over different phyla. Changes in abundance of the major phyla and their G+C% content were consistent between the 16S rRNA gene (**A**) and the whole genome levels **(B)**. Black bars represent the median and white stars represent the mean, the left and the right box boundaries represent the first and the third quartiles, respectively, and the left and the right whiskers mark the 1.5 inter-quartile range. Outliers are plotted using red crosses.

Relative abundance of metabolic pathways in heated vs. control metagenomes

The protein-encoding PE reads were assigned to pathways in the SEED database (34) based on homology searches of the protein sequence encoded on the reads, and the number of reads was taken as a proxy of the relative abundance of the pathway in the corresponding sample (between 30 to 40% of the total reads in each sample were assignable to the SEED database). A statistical approach was developed, which employed Jackknife resampling and the B-H method for adjusting *P*-values, to identify pathways differentially presented between the heated and control datasets (Figure 6.9). Consistent with the low-impact perturbation applied (2 ^{0}C warming, for 10 years), the differences in pathway abundances between heated and control samples were small, typically <5% change. Nonetheless, several significant changes were also noted and these were reproducible in biological and technical replicates (Figure 6.10). A large portion of pathways involved in carbon source utilization and degradation, and nitrogen cycle showed significant changes in relative abundance (Figure 6.11). In particular, several pathways that are involved in labile carbon source metabolism were enriched in heated samples, such as glycerate metabolism (+13%), cellulose degradation (+13%) and -glucuronide utilization (+22%). The opposite trend was observed for pathways related to (more) recalcitrant carbon sources, such as chitin utilization (-9%) and lignin degradation (-18%). These observations were consistent with field physicochemical measurements, which showed higher labile organic carbon content and higher primary production in heated soils (Figure 6.11; Table 6.3), driven mostly by aboveground plant communities (Table 6.4), and previous results based on GeoChip analysis of samples from the same site but different years (51). Furthermore, the enriched pathways in

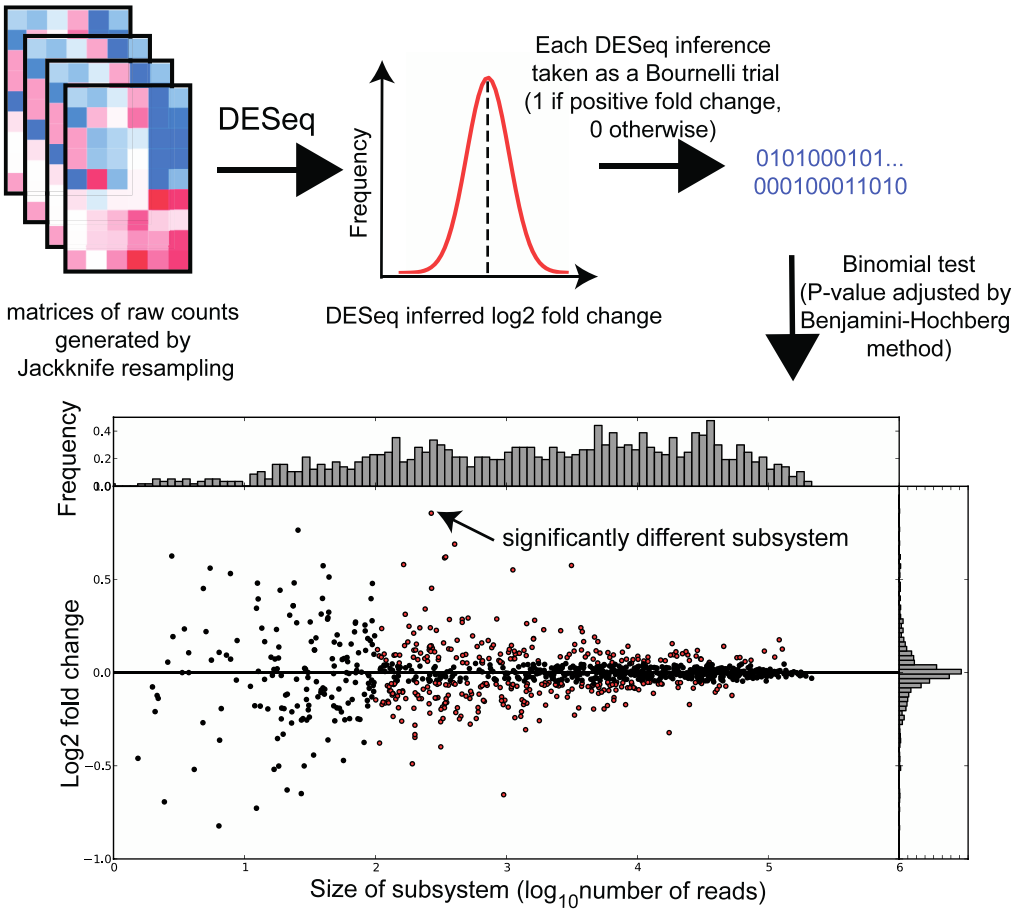


Figure 6.9. Flowchart of identifying significantly shifted pathways. The number of reads recruited to each subsystem was converted to a raw count table (with rows representing subsystem abundance and columns representing samples) and Jackknife resampling was then carried out to resample three heated samples and three controls each time without replacement (three samples were used instead of the total six samples to provide enough replicate datasets for statistical analysis). Each of the sub-sampled tables was fed into R package DESeq (36) to infer the log2 fold change. The inference followed a normal distribution with a mean of 0 if no change between heated and control samples, positive if increased in heated samples, negative otherwise. To test if the fold changes were significant, a binomial test was carried out (1 for DESeq inference with positive log2 fold change, and 0 otherwise), and the *P*-values were adjusted using the Benjamini-Hochberg method. The subsystems with more than 5% change, *P*-value <0.05, and

more than 100 recruited reads (averaged across all samples) were identified as significantly differentially present (red dots on the scatter plot).

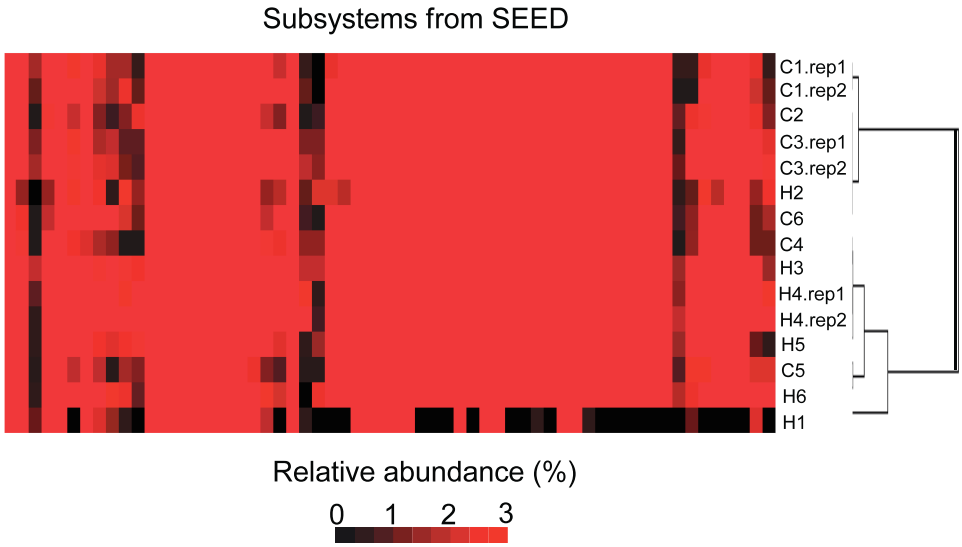


Figure 6.10. Clustering of samples and replicates based on SEED subsystem relative abundance. Each column represents a SEED subsystem (with at least ten thousand reads recruited per sample, on average), and each row represents a sample. The abundance of the subsystem, normalized for the sample size by dividing by the total number of reads in the sample, is represented by the color intensity (see scale). Hierarchical clustering was carried out to group samples using Euclidean distance. Note that the technical replicates (e.g., C1.rep1 and C1.rep2) and control vs. heated samples were clustered together, consistent with our expectations. The exceptions to this pattern were the clustering of control C5 sample with the heated samples and the heated sample H2 with the control samples, which may be due to soil heterogeneity and the pH, e.g., the pH value of the C5 sample was lower and more similar to the pH of the heated samples compared to the other control samples. Differentially present subsystems are annotated in Table 6.5.

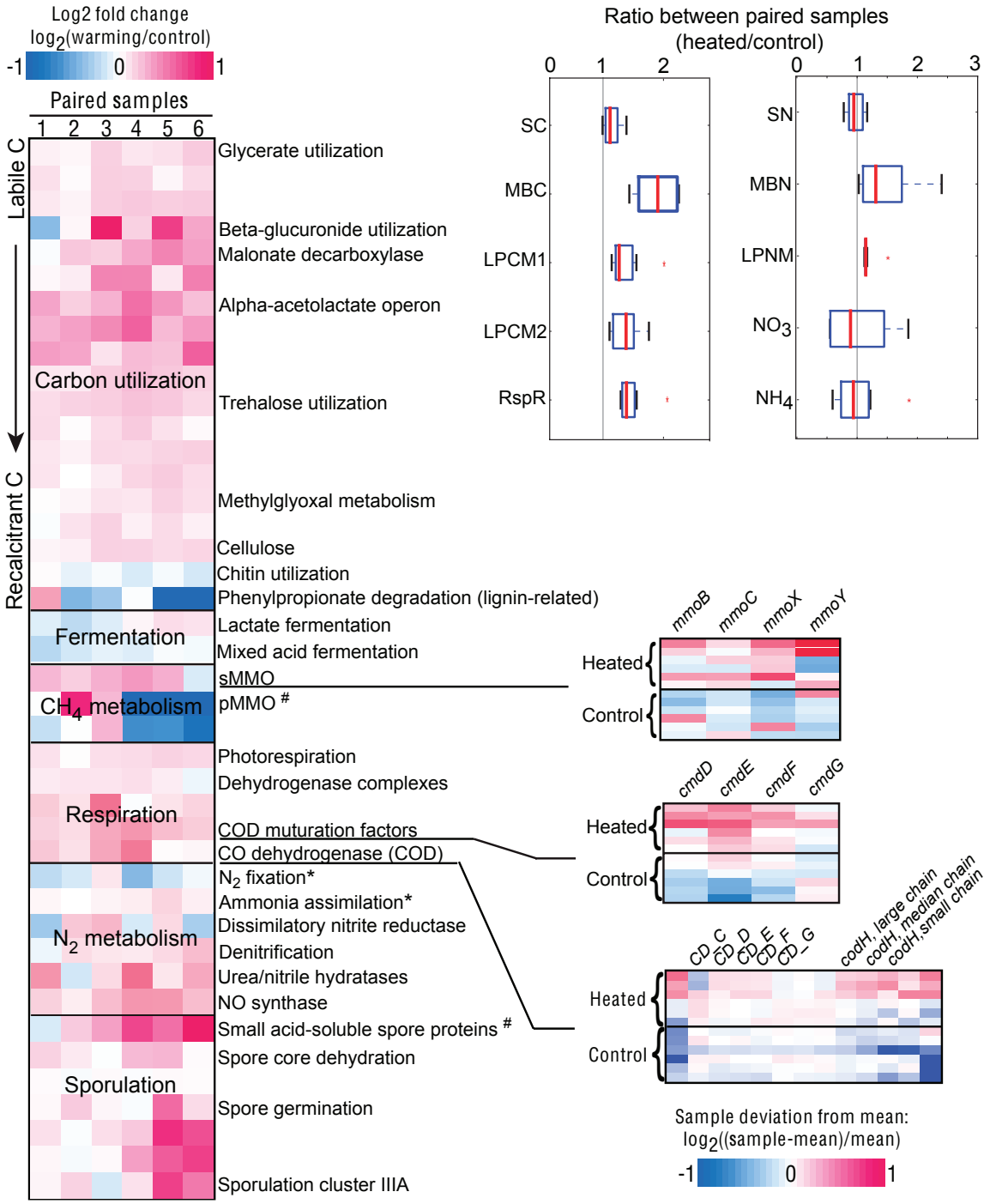


Figure 6.11. Changes in pathway relative abundance as an effect of warming. The heatmap on the left represents changes in abundance of different pathways (rows) for each pair of samples (columns), color-coded based on the magnitude of the change (see scale on the top left). For selected pathways related to the emission of greenhouse gases, the relative abundance of the

individual genes that constitute the pathway is shown on the right (small heatmaps; rows represent samples, and columns represent genes). In this case, changes in abundance represent deviation from the average abundance of the gene in all twelve samples, are color-coded based on the magnitude of the difference (see scale on the bottom right), and are generally consistent with the results for the whole pathway. The results of physicochemical measurements are represented by box-plots on the top-right. The vertical line at ratio 1 indicates no change between heated and control samples; the median of six paired replicate samples is marked by the red bar; the first and third quartiles are represented by the left and right boundaries of the box, respectively; the left and right whiskers represent the 1.5 inter-quartile range; outliers are marked by red asterisks. Abbreviations denote: SC, total soil carbon; MBC, microbial carbon; LPCM1/2, labile pool 1/2 of carbon, microbial; RspR, respiration rate; SN, total soil nitrogen, MBN, microbial nitrogen, LPNM, labile pool of nitrogen, microbial.

heated samples included carbon monoxide (CO) dehydrogenases, their maturation factors, and various respiratory pathways. These findings were in agreement with higher respiration and elevated carbon dioxide (CO_2) emissions measured in the heated vs. the control soils (51). In contrast, fermentation pathways, e.g., lactate and mixed acid fermentation, were typically less abundant in the heated metagenomes, apparently due to the prevalence of oxidative (aerobic) metabolism and the availability of additional, plant-derived labile organic soil carbon as an effect of warming.

In terms of nitrogen metabolism, significantly higher abundance of denitrification and dissimilatory nitrite reductase in heated samples were observed while nitrogen

fixation genes did not significantly differ in relative abundance (Figure 6.11). These observations indicated higher turnover and decreased content of organic nitrogen in heated soils, in agreement with higher labile carbon concentration and physicochemical measurements (Figure 6.11, and Table 6.3). It should be also mentioned that soil moisture, which is typically positively associated with the prevalence of anaerobic conditions and processes (such as denitrification), was lower in heated vs. control samples by about 4% (Table 6.3) but the difference was not statistically significant.

Taken together, our results revealed that warming induces higher primary production and microbial respiration rates in the temperate soils studied here; microbial respiration appears to release most, if not all, of the soil organic carbon fixed by (primarily) aboveground plant activity to the atmosphere. In agreement with these interpretations, we found that although aboveground plant biomass was 10 to 30% higher in heated vs. control sites, depending on the site considered (Table 6.4), the total soil carbon concentration was not significantly different between the sites (Figure 6.11). Finally, a higher abundance of sporulation-related genes and pathways, e.g., spore core dehydration (5% difference) and spore germination (12% difference) was observed in the communities that underwent warming (Figure 6.11 and Table B1), which was consistent with our expectations.

Community-wide vs. taxon-specific shifts.

We also evaluated whether the shifts observed between heated and control datasets were due to systematic community-wide adaptations or instead to the differential presence of a few taxa. To this end, all (control and heated) overlapping PE reads

encoding a gene that was found to be differentially abundant in heated vs. control datasets were clustered at the 80% sequence identity level, providing the operational taxonomic units (OTUs) present in the samples for each gene. The percentage of heated vs. control reads constituting each OTU was compared to determine the OTU(s) that contributed to the higher abundance of genes and pathways in the heated samples. Overall, most of the gene content shifts were attributable to many OTUs, typically more that 50% of the total OTUs observed for each gene analyzed, revealing that warming induced community-wide adaptations (Figure 6.12-13). These systematic responses and the shifts in G+C% and gene content mentioned above indicate that the differences between heated and control samples were likely attributable to long-term adaptations as opposed to short-term, pulse-like responses of a few taxa.

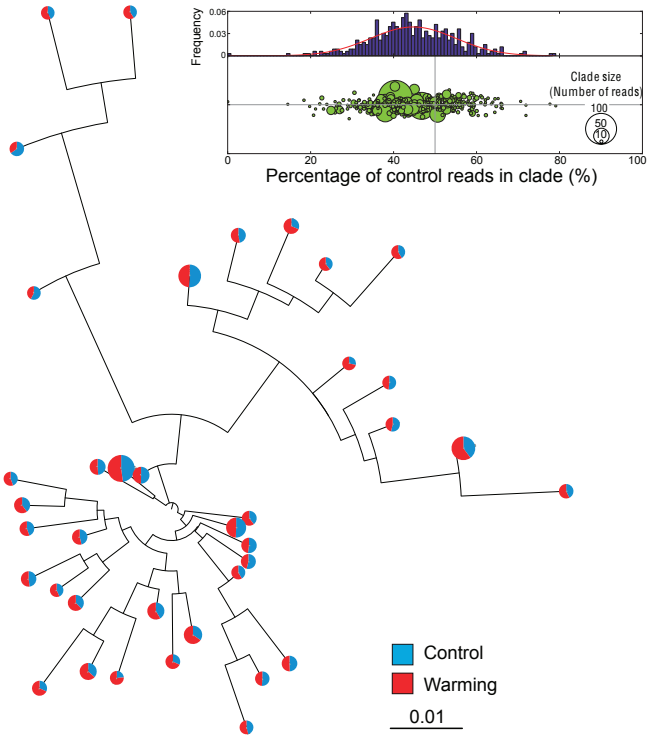


Figure 6.12. Changes in pathway abundance are community-wide and not attributable to a few taxa. Representative sequences from all OTUs (or clades) of a specific gene (in this case, a CO_2 dehydrogenase, *CD_D*) were analyzed to produce the distance-based phylogenetic tree shown. Pie charts at the tips of the tree represent the percentage of heated vs. control reads that made up each OTU and the size of the chart is proportional to the number of reads in the OTU; only OTUs with at least 50 reads are shown for simplicity purposes. Note that no OTU was heat- or control-specific and about ~60% of the pie charts had a higher number of heated vs. control reads, revealing that many distinct taxa are responsible for the higher abundance of CO_2 dehydrogenase in heated metagenomes. This is also evident in the graph shown on the top (***inset***). In the latter graph, each circle represents an OTU; the x-coordinate represents the percent of the total reads of the OTU that are control reads, the y-coordinate represents a random value for visualization purposes. Note that more OTUs have lower than 50% control reads relatively to OTUs with more than 50% control reads. The histogram on the top shows the distribution of the percentages of control reads in all OTUs, with a fitted Gaussion curve in solid line.

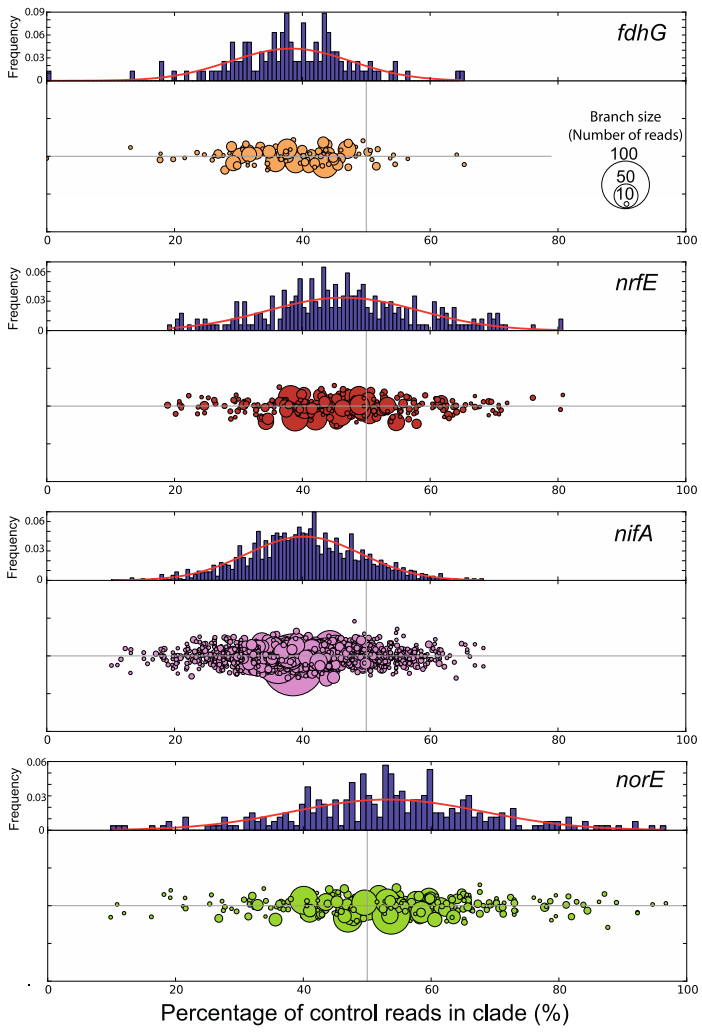


Figure 6.13. Systematic changes in relative abundance in different pathways To evaluate whether the shifts in pathway relative abundance observed in Fig. 3 were attributable to community-wide changes as opposed to changes in the abundance of a few taxa, we identified all reads encoding the genes of the pathways and built phylogenetic trees of all fully overlapping reads from both control and heated samples (combined in one tree). The sequences in the tree were subsequently grouped in OTUs, using a cut-off of 80% nucleotide sequence identity, i.e., intra-clade nucleotide sequence diversity <20%, inter-clade >20%. Each circle represents an OTU and its size is proportion to the number of reads the clade contains (see figure key). The x-axis value represents the portion of control read in clades for characteristic genes in methanogenesis (*fdhG*), ammonification (*nrfE*), nitrogen fixation (*nifA*), and denitrification (*norE*). The y-

coordinate is a noise added at random for visualization purposes. In other words, clades appearing on the left of the y=0 line are made up of more heated than control reads and thus account for the higher abundance of the corresponding subsystem in heated samples. The histogram on the top shows the distribution of the portions of control clades for all OTUs of a gene, with a fitted Gaussion curve in red. Note that many clades accounted for the differential presence of the pathways shown, suggesting that warming induced community-wide shifts in microbial communities.

Conclusions and perspectives for the future

The results reported here demonstrate that metagenomics and related molecular techniques represent powerful means to monitor the genomic adaptations and functional responses of complex soil microbial communities to long-term perturbations such as the predicted effects of climate change. Metagenomic data obtained from replicate samples were quantitative (e.g., Figure 6.11), highly reproducible at the subsystems or individual gene levels (e.g., Figure 6.10-11), and consistent with macroscopic, biochemical and physicochemical measurements of soil and aboveground plant biota. These data revealed that soil microbial communities adapt fast to perturbations, even low-impact ones, perhaps faster than previously anticipated. In the case of (mild) warming, adaptation was evident, for instance, by significant shifts in G+C% content and metabolic pathway abundance in the genomes of the indigenous microbes. These adaptations apparently took place in less than 10 years and we find it remarkable that features like G+C% content, which are thought to represent stable properties of the genome and community, can change in such a (relative) short period of time. Our findings indicated that microbial communities of temperate grassland soils play important roles in feedback responses to exposure to elevated temperatures, at least in the short term (e.g., a decade). In the soils studied here, this was evident by a significantly higher abundance of respiration and labile carbon metabolism genes in heated vs. control samples; control samples showed instead a higher abundance of recalcitrant carbon degradation genes (e.g., Figure 6.11). This feedback appears to represent a community-wide response, as opposed to being attributable to the activity of a few taxa (e.g., Figure 6.12), and is presumably driven by complex interactions among community members (e.g., Figure 6.6). Our study also

highlighted the complex interactions and feedbacks between belowground microbial communities and aboveground plant communities and the importance of the former (in addition to the latter) for the models of climate change. Nonetheless, disentangling the direct effect of warming on the belowground microbial communities from the indirect effect of warming due to the stimulation of aboveground plant communities remains challenging. Additional samples across time and soils of different types and latitudes need to be examined before more robust conclusions can emerge with respect to the importance of the belowground microbial communities for mitigating or exacerbating the effects of climate change.

REFRENCES

1. Handelsman J*, et al.* (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies Press, Washington, DC).
2. Torsvik V, Goksoyr J, & Daae FL (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* 56(3):782-787.
3. Whitman WB, Coleman DC, & Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95(12):6578-6583.
4. Curtis TP, Sloan WT, & Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A* 99(16):10494-10499.
5. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102(7):2567-2572.
6. Nelson KE, Paulsen IT, Heidelberg JF, & Fraser CM (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nat Biotechnol* 18(10):1049-1054.
7. Bond-Lamberty B & Thomson A (2010) Temperature-associated increases in the global soil respiration record. *Nature* 464(7288):579-582.
8. Heimann M & Reichstein M (2008) Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature* 451(7176):289-292.
9. Deng Y*, et al.* (2012) Elevated carbon dioxide alters the structure of soil microbial communities. *Appl Environ Microbiol* 78(8):2991-2995.
10. Goldfarb KC*, et al.* (2011) Differential growth responses of soil bacterial taxa to carbon substrates of varying chemical recalcitrance. *Frontiers in microbiology* 2:94.
11. Dunbar J*, et al.* (2012) Common bacterial responses in six ecosystems exposed to 10 years of elevated atmospheric carbon dioxide. *Environmental microbiology*.
12. Fierer N*, et al.* (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME journal* 6(5):1007-1017.
13. Konstantinidis KT & Tiedje JM (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current opinion in microbiology* 10(5):504-509.
14. Mackelprang R*, et al.* (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480(7377):368-371.
15. Cheng X, Luo Y, Xu X, Sherry R, & Zhang Q (2011) Soil organic matter dynamics in a North America tallgrass prairie after 9 yr of experimental warming. *Biogeosciences* 8(6):1487-1498.
16. Sherry RA*, et al.* (2008) Lagged effects of experimental warming and doubled precipitation on annual and seasonal aboveground biomass production in a tallgrass prairie. *Global Change Biol* 14(12):2923-2936.
17. Zhou J, Bruns MA, & Tiedje JM (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* 62(2):316-322.
18. Ahn SJ, Costa J, & Emanuel JR (1996) PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic acids research* 24(13):2623-2625.

19. Belay-Tedla A, Zhou XH, Su B, Wan SQ, & Luo YQ (2009) Labile, recalcitrant, and microbial carbon and nitrogen pools of a tallgrass prairie soil in the US Great Plains subjected to experimental warming and clipping. *Soil Biol Biochem* 41(1):110-116.
20. Luo Y, Wan S, Hui D, & Wallace LL (2001) Acclimatization of soil respiration to warming in a tall grass prairie. *Nature* 413(6856):622-625.
21. Luo Y, White L, & Hui D (2004) Comment on "Impacts of fine root turnover on forest NPP and soil C sequestration potential". *Science* 304(5678):1745; author reply 1745.
22. Sherry RA*, et al.* (2007) Divergence of reproductive phenology under climate warming. *Proc Natl Acad Sci U S A* 104(1):198-202.
23. Zhou X, Wan SQ, & Luo YQ (2007) Source components and interannual variability of soil CO2 efflux under experimental warming and clipping in a grassland ecosystem. *Global Change Biol* 13(4):761-775.
24. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
25. DeSantis TZ*, et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72(7):5069-5072.
26. Lozupone C, Hamady M, & Knight R (2006) UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics* 7:371.
27. Thompson JD, Gibson TJ, & Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 2:Unit 2 3.
28. Glass EM, Wilkening J, Wilke A, Antonopoulos D, & Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor protocols* 2010(1):pdb prot5368.
29. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18(5):821-829.
30. Li R*, et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* 20(2):265-272.
31. Luo C, Tsementzi D, Kyrpides NC, & Konstantinidis KT (2011) Individual genome assembly from complex community short-read metagenomic datasets. *The ISME journal*.
32. Zhu W, Lomsadze A, & Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic acids research* 38(12):e132.
33. Rho M, Tang H, & Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research* 38(20):e191.
34. Overbeek R*, et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research* 33(17):5691-5702.
35. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome research* 12(4):656-664.
36. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11(10):R106.

37. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792-1797.
38. Price MN, Dehal PS, & Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution* 26(7):1641-1650.
39. Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658-1659.
40. Williamson SJ*, et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS one* 3(1):e1456.
41. Oh S*, et al.* (2011) Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77(17):6000-6011.
42. Meyer F*, et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 9:386.
43. Konstantinidis KT, Braff J, Karl DM, & DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* 75(16):5345-5355.
44. Konstantinidis KT & DeLong EF (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* 2(10):1052-1065.
45. Cole J, Konstantinidis KT, Farris RJ, & Tiedje JM (2010) Microbial diversity and phylogeny: extending from rRNAs to genomes. *Environmental Molecular Biology*, eds Liu W-T & Jansson J (Horizon Scientific Press, Norwich, UK), pp 1-20.
46. Nacke H*, et al.* (2011) Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PloS one* 6(2):e17000.
47. Caro-Quintero A & Konstantinidis KT (2012) Bacterial species may exist, metagenomics reveal. *Environmental microbiology* 14(2):347-355.
48. Barberan A, Bates ST, Casamayor EO, & Fierer N (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal* 6(2):343-351.
49. Nakashima H, Fukuchi S, & Nishikawa K (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *Journal of biochemistry* 133(4):507-513.
50. Rocap G*, et al.* (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424(6952):1042-1047.
51. Zhou J*, et al.* (2012) Microbial mediation of carbon-cycle feedbacks to climate warming. *Nature Climate Change* 2:106–110.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Energy (award DE-SC0004601). We thank the personnel of the Los Alamos National Laboratory Genomics Facility for their assistance with DNA sequencing.

CHAPTER 7

Quantifying the role of horizontal gene transfer in maintaining microbial population biodiversity with time-series metagenomics

INTRODUCTION

Our understanding of how bacteria evolve is mainly based on laboratory studies with pure cultures of a few model species (1-3). Although the genetic mechanisms responsible for evolution and adaptation (e.g., horizontal gene transfer, intra-genomic recombination, point mutation etc.) have been well documented (4-6), their relative importance in shaping bacterial lineages *in-situ* remains elusive. A major reason accounting for the latter is that laboratory conditions do not simulate well natural conditions, including the biotic interactions among co-occurring populations (7-9). Studying microbial populations *in-situ* is thus critical for advancing our understanding of the mechanisms that create and maintain biodiversity and for preserving the biodiversity on the planet.

Horizontal gene transfer (HGT) represents an important mechanism for diversification and adaptation of bacteria, and the main reason that eukaryotic concepts do not frequently translate well in bacteria. Analysis of the genomes of isolates has shown that bacterial genomes are more dynamic and fluid than previously anticipated due to HGT. Accordingly, it is nowadays thought that HGT represents a mechanism that creates genomic diversity; for instance, by introducing new genes into a subset of a population and enable the resulting subpopulation to explore a new ecological niche (speciation) or outcompete its co-occurring relatives (purging population diversity) (10-12). Yet, more recent studies have indicated that HGT, when combined with homologous recombination and affecting all genes in the genome (11, 13, 14), can serve as a population homogenizing force and maintain a population (15). The relative importance of these two different faces of HGT for natural populations remain essentially unknown.

Despite the obvious need to detect and quantify HGT in natural populations and the recent development of culture-independent genomic techniques to study *in-situ* processes [aka metagenomics (7, 16, 17)], there has been no effective approach to accomplish these tasks. The latter is primarily attributed to the high complexity of microbial communities, frequently composed of hundreds, if not thousands, of distinct species, and the fragmented sequence information provided by metagenomics techniques (e.g., only short and unlinked fragments of a genome are typically recovered). Therefore, it is essential to develop new approaches to investigate HGT and genomic adaptation within natural microbial communities, in real time. Such approaches will greatly facilitate studies of pollutant biodegradation under natural or engineering settings and spreading of infectious diseases and antibiotic resistance.

Towards closing these gaps in knowledge and enabling technologies, we developed a robust bioinformatic pipeline, called metaHGT, to detect HGTs in time series metagenomic datasets. Subsequently, we applied this pipeline to metagenomes originating from planktonic samples collected at the same, well-oxygenated (5m depth) site of the mesotropic Lake Lanier (Atlanta, GA) and spanning a period of almost three years (August 2009 to January 2012). Our results revealed the frequency of HGT was at least three orders of magnitude higher compared to previous estimates, especially among distantly related populations of different phyla, and indicated that HGT frequently maintains population diversity by facilitating spreading of advantageous genes.

MATERIALS AND METHODS

Sampling, DNA extraction, and DNA sequencing

Freshwater planktonic samples were collected between August 2009 and January 2011 (Table 7.1). All samples were collected at the same location, below the Brown's Bridge in Lake Lanier (Atlanta, GA), at 5m depth (oxygenated water). Samples were filtered and DNA was extracted using the same protocol as described previously (18). High-throughput short-read sequencing were carried out on Illumina platforms; samples collected in 2009 were sequenced at Emory University's Genomic Facility on an Illumina GA II technology; all other samples were sequenced at Los Alamos National Laboratory's sequencing facility on an Illumina HiSeq-2000 technology (Table 7.2). The physicochemical measurements of each sample were taken at the time of sampling (Table 71).

Table 7.1. Physicochemical characteristics of samples.

	09/08/26	09/08/28	09/09/07	09/11/08	10/07/06	10/09/10	10/11/14	11/01/29
Temperature (°C)	28.5	28.5	N/A	19.2	30.5	28.8	17.9	7.8
pH	7.61	7.71	N/A	6.8	7.5	7.6	5.73	6.56
Dissolved solids (g/L)	0.032	0.033	N/A	0.030	0.027	0.03	0.03	0.03
Dissolved O_2 (mg/L)	7.9	7.8	N/A	7.3	8.9	7.5	5.5	4.2

Table 7.2. DNA sequencing and assembly information for each sample

Sample ID	Sequencing Platform	Number of reads, raw (! 10^6)	Number of reads, after trimming (! 10^6)	Assembly length (>500bp, Mbp)	Assembly N50 (Kbp)	Number of contigs (! 10^3)	Percentage of read recruited by assembly (%)
09/08/26	Illumina GA II 100x100bp	32.1	27.9	93.4	1.71	70.5	55.3
09/08/28	Illumina GA II 100x100bp	29.6	26.0	96.3	1.78	70.9	56.2
09/09/07	Illumina GA II 100x100bp	27.1	23.6	86.5	1.39	71.8	50.5
09/11/08	Illumina GA II 100x100bp	34.2	27.8	97.7	1.08	96.7	38.9
10/07/06.rep1	Illumina HiSeq-2000 100x100bp	67.6	58.0	147.0	1.29	129.2	53.3
10/07/06.rep3	Illumina HiSeq-2000 100x100bp	68.2	59.2	145.8	1.29	128.4	47.6
10/09/10	Illumina HiSeq-2000 100x100bp	42.6	36.3	89.9	1.71	67.1	55.2
10/11/14	Illumina HiSeq-2000 100x100bp	47.3	41.1	77.1	1.27	68.6	39.1
11/01/29	Illumina HiSeq-2000 100x100bp	52.3	41.2	149.8	1.01	155.4	35.7

Sequencing read processing and quality check

Read trimming was carried out using an in-house python script (available from the authors upon request). The script trimmed each pair-ended read based on the following sequential steps: i) trim the read from both 5'- and 3'-ends until a base with Phred score >20 is found; ii) use a 3bp-long sliding window to examine the quality of the

remaining sequence from step (i). If the average Phred score of a window is lower than 20, create a cut at that position; iii) keep the remaining sequence if it is longer than 50bp and contains no more than 1 N (ambiguous base); otherwise discard the sequence. Only reads that both paired-end reads passed the trimming cut-offs were used for further analysis. This method was applied in all samples used in the study, including the publicly available metagenomes, for consistency purposes.

Metagenome assembly and binning contigs into population genomes

The trimmed pair-end (PE) reads were first pre-assembled, separately for each sample, using a hybrid protocol combining Velvet, SOAPdenovo, and Newbler 2.0 (19, 20) as described previously (21, 22). Assembled contigs longer than 500 bp were subsequently binned into candidate genomes. For this, reads were mapped back to these contigs by BLAT (23) with a length cutoff of 50bp, a percent nucleotide identity cutoff of 97, and an e-value cutoff of 1e-10. The number of reads recruited per 100 bp of contig sequence was used as a proxy of the relative abundance (coverage) of each pre-contig. Contigs representing presumably the same population were binned together based on their (similar) pair-wise tetra-nucleotide frequency correlation, linkages by PE read, and relative coverage, similarly to what reported previously (24, 25), and as detailed below.

If two contigs had tetra-nucleotide Spearman's correlation larger than 0.85, and/or if they had three or more PE reads linking them, the contigs were linked, and thus, binned together. The weight of the link was 1 if qualified by one of the two measures and 2 if qualified by both. It is expected that two contigs representing the same population would show the same relative abundance in multiple sampling points. Based on this assumption,

a linear regression was fit to the coverage values of every pair of contigs from the nine available metagenomic samples. Pairs with R^2 larger than 0.85 and a slope ranging from 0.9 to 1.1 were linked together (if they were already linked together by the previous step, then the link weight increased by 1). Subsequently, the individual contigs were projected onto a large graph, in which each contig was a node and each link was a weighted edge. An iterative partitioning algorithm was then carried out to bin these contigs in population genomes. We first used a quick *k-means* algorithm to find possible centroids of partitions by initializing the number of clusters, *k*, to be 10 (because our previous study (26) established that at least ten abundant populations were present in the 2009 samples), and we increased *k* until the *Calinski-Harabasz* (CH) Index (27) was maximized (converge condition was [*CH(k+1)-CH(k)*]/*CH(k)*<0.01). This procedure was similar to the approaches used by Arumugam and colleagues (28). The CH index was defined as:

$$CH(k) = \frac{B(k)/(k\ !\ 1)}{W(k)/(n\ !\ k)},$$

where *n* was the total number of contigs; *B(k)* was the inter-cluster sum of squares of distances between contigs (the distance was defined as *d=4-w*, where *w* was the weight of the links); and *W(k)* was the intra-cluster sum of squares of distances between contigs (clusters in this case represented the potential genome bins).

We retrieved 45 bins, with a total length of the binned contigs longer than 500 Kbp for each bin. PE reads were recruited to those bins by BLAT mapping using the same cutoffs as described above. For each bin, we used the recruited reads to re-assemble them into contigs by Velvet (19) with K-mer length optimized for the longest N50. This approach substantially improved the quality of the assembled sequence due to the reduction of sequence complexity during the assembly step compared to the original

assembly of the whole metagenomic sample. The resulting assembly represented the final genome sequence of each bin (population) used for further analysis.

Genome draft validation and estimate of completeness

To validate the final assembly of each contig bin, the mapping patterns of PE reads were visually inspected for consistency (e.g., distance between PE reads to be similar to the expected distance based on library insert size, i.e., 200-300 bp) and even coverage. Contigs with uneven coverage were not used in downstream analysis. We employed other metagenomic datasets (i.e., other than the one used for the assembly) from this study as well as datasets from independent studies to validate our population genomes partitions, as detailed below. A Roche 454 metagenome was previously generated from the same DNA sample as the 2009/08/26 Illumina dataset used here (18, 22). Roche 454 reads were mapped to the final contigs of bins from the latter Illumina dataset using BLAT (cutoff: 80% length aligned and 95% nucleotide identity) to calculate average per base coverage, and the coverage was visually inspected, as described above. 18 partitions were of high quality were selected based on this analysis (one example given in Figure 7.1). Therefore, the rest of the study was focused on these 18 genome drafts.

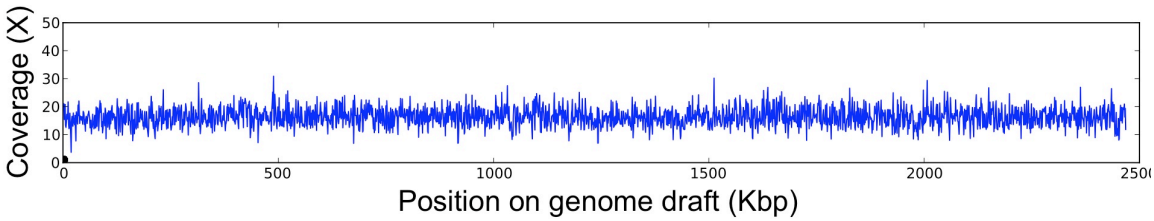


Figure 7.1. Validation of draft *LL3* genome by read mapping. Metagenomic reads from the same DNA sample (Roche 454 reads from Lake Lanier, GA, August 2009 sample) were mapped against the *LL3* genome assembled from Lake Lanier (Atlanta, GA). Read mapping revealed even coverage along the genome. For instance, Roche 454 reads covered 100% of the draft genome with average coverage 15X (standard deviation <3 X).

Gene calling, annotation, and phylogenetic analysis

Protein-coding genes in the draft genomes were predicted by MetaGeneMark (29), and the corresponding amino acid sequences were searched against the KEGG (30), eggNOG (31), COG (32), and SEED (33) databases (as of May 2012) using Blastp (34). Only matches with at least 40% amino acid identity and 70% of the length of the query sequence covered in the alignment were considered further. Predicted genes sequences were annotated with the function of their best match when the best three matches had the same (predicted) function.

To identify orthologous genes, we performed an all-vs-all Blastp search using protein-coding genes. Only matches with 70% or longer gene coverage in the alignment for both query and target gene sequences and 30% or higher amino acid identity were considered further. Orthologs were defined as the reciprocal best matches (RBMs) above

the previous cut-off. The pan-genomes of the 18 drafted population genomes were constructed using orthologous genes.

While the 18 genomes represented several divergent phyla, they shared at least five genes (*dnaK, polA, recA, lon,* and *mtf*). Sequence alignments of these five genes from the 18 genomes and 29 close relatives available in NCBI were performed by MUSCLE 3.8 (35), using default settings. The individual gene sequence alignments were subsequently concatenated, and a maximum likelihood phylogenetic tree was built based on the concatenated alignment using PhyML3.0 (36) with gamma set to be 4, the HYK85 model, and 1,000 bootstraps. The tree (Figure 7.4) was visualized by SplitTree 4.0 (37).

The metaHGT algorithm for HGT detection in time series metagenomes

We developed a novel algorithm, called metaHGT, to detect HGT events in time series metagenomic data. The basic idea of the algorithm is to detect abnormal patterns in pair-end read mapping onto assembled contigs. The idea is illustrated with the following simple example. Consider that we wish to test if a HGT event took place between species A and B between time points 1 and 2, and involved contig S_A and S_B from species A and B, respectively, so that $S_A = g_1^A g_2^A ... g_n^A$ and $S_B = g_1^B g_2^B ... g_m^B$, where n and m are respectively the number of genes in genome A and genome B, and g_i^γ is the i^{th} gene in genome $! \quad \{A,B\}$. The intergenic distances are denoted as $D_A = \{d_i^A \mid 1 < i < n \,!\, 1\}$ and $D_B = \{d_i^B \mid 1 < i < m \,!\, 1\}$ for S_A and S_B, respectively. We also defined a "PE linkage", $f_! (g_i, g_j)$, as the PE read with one end aligned to g_i and the other end aligned to g_j in time point τ. Assuming that the HGT occurred between time points 1 and 2 and created a novel genotype in time point 2, denoted $S_{HGT} = ... g_i^A g_j^B ...$, we then expect that, with

adequate sequencing depth, $f_2(g_i^A, g_j^B) > 0$ and $f_1(g_i^A, g_j^B) = 0$. We employed this concept to interrogate metagenomic data and patterns of PE read mapping to detect HGT events.

However, in real metagenomes, the above ideal scenario is often complicated by two factors: i) local similarity between two genomes (e.g., due to paralogs or multiple gene copies), which leads to misleading mapping of PE reads; and ii) difference between the insert size of the library sequenced for each sample, which leads to incomparable PE linkages. To account for (i), we defined a weight α_k for PE read linkage $f_!(g_i, g_j)$ as:

$$!_k = \frac{2I(g_i, g_j)}{\max\{I(g_i, g_s), I(g_j, g_t)\} + I(g_i, g_j)} \#1,$$

where $I_!(g_i, g_j)$ is the product of the percent nucleotide identities of the two sister reads of a PE pair mapped to gene g_i and gene g_j. Therefore, if a strong local similarity was introduced by other genes in the same genome (denoted by gene g_s and gene g_t), we have $\max\{I_!(g_i, g_s), I_!(g_j, g_t)\} \quad I_!(g_i, g_j)$, and $!_k = 0$; otherwise, $\max\{I_!(g_i, g_s), I_!(g_j, g_t)\} = 0$, and $!_k = 1$. Therefore, the impact of local similarity from adjacent gene sequences on HGT detection is normalized by introducing a weight $0 \le \alpha \le 1$, and converting the PE read linkage count *K* into the weighted sum $\sum_{k=1}^{K} !_k$. Note that the sister reads have to be mapped legally on gene g_i and g_j to be considered in calculation of weight i.e., being concordant with the α_k, upposed orientation (for Illumina sequencing libraries, one read represents the forward strand while its sister read represents the reverse strand). If no such legal mapping is found, then $I_!(g_i, g_j) = 0$).

To account for (ii) above, we introduced the alignment length cutoff *e* and the intergenic distance *d*. In order to extend a PE linkage between two genes *d*-bp apart with

minimal aligned length *e*, the corresponding insert size *L* of the PE read pair has to satisfy *L>d+2e*. Thus, the probability that one PE read pair, selected at random, could extend a PE linkage between the two genes would be the cumulative probability Pr(L ! $d+2e$). Assuming for time points 1 and 2, the corresponding insert sizes for the paired-end (PE) reads follow Gaussian distributions $N_1(\mu_1,!\ ^2_1)$ and $N_2(\mu_2,!\ ^2_2)$, respectively, the cumulative probability can be transformed into ! (Z), where $Z=\frac{d+2e\ !\ \mu}{}$. Assuming also that the reads were uniformly sampled, the probability that a PE read pair links two genes then is 1! (Z). Therefore, we introduce a scaling factor $!_{12}=\frac{1\quad \#\ (Z_1)}{1\quad \#\ (Z_2)}$ to account for any differences in library insert size between time point 1 and time point 2.

Next we considered $h^{!}_{x,y}$, the normalized and weighted PE linkage count in time point τ between two genes, g_x and g_y. In the case of $S_{HGT}=...g^A_i g^B_j...$, let x ! $\{i,i+1\}$and y ! $\{j\quad 1,j\}$, the null hypothesis is that no HGT occurred between the two sampling points. Under this null hypothesis, $(h^1_{i!\,1,i},h^1_{i,j})$ and $(h^2_{i!\,1,i},h^2_{i,j})$ were drawn from the same binomial distribution with p_0 denoting the probability that a randomly drawn PE read links g_x and g_y. Obviously, the maximum likelihood estimate of p_0 for the null hypothesis is $p_0=(\frac{h^1_{i,j}}{h^1_{i,j}+h^1_{i!\,1,i}}+\frac{h^2_{i,j}}{h^2_{i,j}+h^2_{i!\,1,i}})/2$.

The alternative hypothesis is that HGT occurred between the two sampling points. In the latter case, $(h^1_{i!\,1,i},h^1_{i,j})$ and $(h^2_{i!\,1,i},h^2_{i,j})$ were drawn from two different binomial distributions with parameters $p_1=\frac{h^1_{i,j}}{h^1_{i,j}+h^1_{i!\,1,i}}$ and $p_2=\frac{h^2_{i,j}}{h^2_{i,j}+h^2_{i!\,1,i}}$, respectively. We test the null hypothesis using a combined two-tailed binomial test such that:

$$P = B(h^1_{i!1,i} + h^1_{i,j}, h^1_{i,j}; p_0)\ \ B(h^2_{i!1,i} + h^2_{i,j}, h^2_{i,j}; p_0),$$

where $B(a,b;p)$ gives the two-tailed *P*-value of the binomial test with success probability p for b successes among a trials. The *P*-values were adjusted for multi-testing by Benjamini-Hochberg approach (38), and only cases where $|p_2\text{-}p_1|>0.1$ were identified as horizontally transferred between A and B species.

Population diversity calculation

An 803 Kbp contig (contig18) from *LL3* was selected for testing the impacts of HGT on bacterial population diversity. Genes that were exchanged between year 2009 and 2010 were detected by metaHGT. For each gene on the contig (both non-HGT and HGT genes), single nucleotide polymorphism sites (SNPs) were identified as follows: if a base was consistent (same nucleotide) within a year (both 2009 and 2010) but different between 2009 and 2010, it was identified as SNP site. For example, if a base in the year 2009 was T throughout all samples, then it was A among all samples in 2010, we would identify this as a T A transversion that occurred between 2009 and 2010. With this approach, the number of SNPs was counted for every gene on the contig (normalized to gene length). Genes were then grouped into two clusters: the ones that were upstream or downstream to an HGT gene; and the remaining genes. We varied the number of upstream/downstream genes considered as adjacent genes to HGT-genes from one to three, and compared the distributions of the number of SNPs per gene between the two resulting gene clusters using a two-sample *t*-test each time.

RESULTS AND DISCUSSION

Performance of metaHGT algorithm

We measured the accuracy of metaHGT using *in silico* generated HGT events and simulated Illumina 100bp paired-end reads. In particular, we assessed the impact of relative abundance, genome relatedness, and intra-population genetic diversity on the performance of the algorithm by measuring false positives (FP; defined as predicted HGT event that did not actually take place) and false negatives (FN; defined as true HGT event that was not predicted).

i) Experimental set up

We used the following genomes as the recipient of HGT events (main chromosomes only): *E. coli* MG1655 (NCBI accession number: NC_000913), *Salmonella enterica* serovar *typhi* CT18 (NCBI accession number: NC_003198), *Vibrio fischeri* ES114 (NCBI accession number: NC_006840), and *Pseudomonas aeruginosa* LESB58 (NCBI accession number: NC_011770); and we used *E. coli* MG1655 as the donor of the HGT event. The recipient genomes were selected to show showed a gradient of genetic relatedness to the donor, measured by the genome-aggregate average amino acid identity [AAI (39)], ranged from 41% to 87% (Figure 7.2). To assess the impact of intra-population genetic diversity (i.e., the existence of distinct sub-populations of the same species), we also used a real Illumina dataset of an *Escherichia* genome (TW09308) that showed ~95% AAI to *E. coli* MG1655 and was sequenced recently (40).

We used an in-house python script to cut the original genomes into Illumina paired-end reads with insert size following a normal distribution with a mean size of 350bp and standard deviation 50bp and 1% base-calling error rate (which was similar to

the Illumina data in this study). We then used an in-house python script to pick 100 genes at random to be (*in-silico*) horizontally transferred from the donor to the recipient genome. A control dataset was also produced where no HGT occurred, which represented time point 1. We generated *in-silico* Illumina reads in the same manner as mentioned above from the hybrid genomes (i.e., the ones encoding the *in-silico* horizontally transferred gene), and denoted these reads to represent time point 2. metaHGT was then applied to detect HGT genes on the data from time point 1 vs. 2, using the same settings as in the analysis of real metagenomes. The predicted horizontally transferred genes were then compared against the known (*in-silico* generated) HGT events, and the FP and FN errors were calculated. We repeated the whole process ten times per donor-recipient pair to obtain statistically robust estimates.

ii) Impact of relative abundance.

We varied the relative abundance of both donor and recipient genomes from 1X to 15X coverage. Therefore, for each donor-recipient pair, we created 15x15=225 datasets. We observed that, at lower abundance (1-3X coverage), the false negative error rates were high, which was caused by insufficient coverage of the assembled contigs by PE reads. However, false negatives dropped rapidly as the coverage increased, and converged at around 1% at 10X coverage. False positive errors increased as the coverage increased but stabilized at around 7-8% after 10X coverage (Figure 7.3). In our time-series metagenomes, all draft genomes had >10X coverage at the time points analyzed by metaHGT.

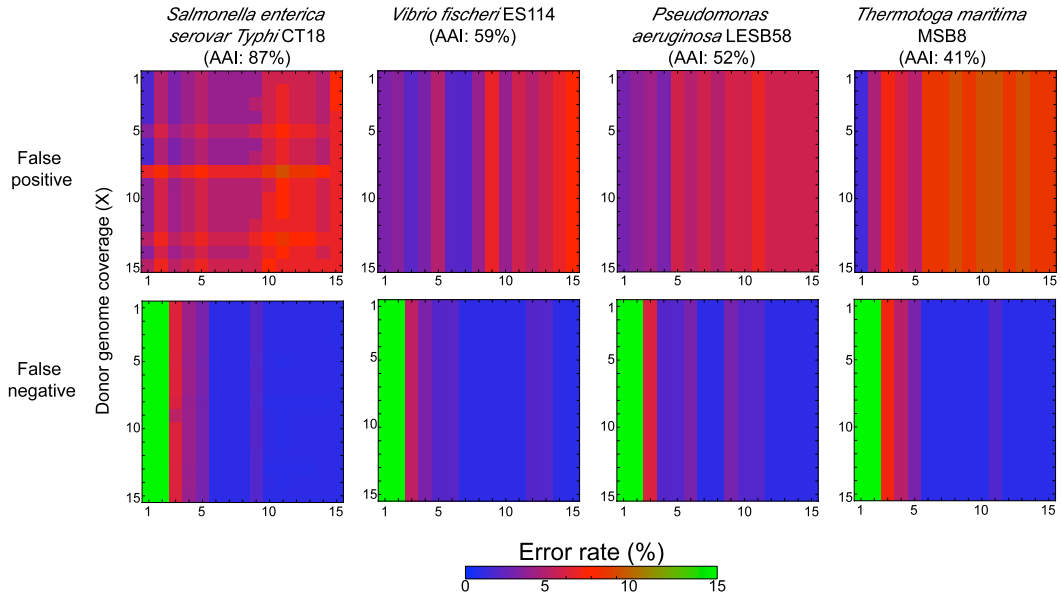


Figure 7.2. Performance of the metaHGT algorithm. The heatmaps show the frequency of false positives (FP; upper panels) and false negatives (FN; lower panels) as a function of the abundance of the pair of genomes participating in (the *in-silico* generated) HGT (x-axis: recipient genome coverage; y-axis: donor genome coverage). Results for different levels of genetic relatedness between the two genomes are shown (see AAI values on top). Note that FP and FN frequencies did not differ dramatically for genomes of different relatedness.

ii) Impact of genome relatedness.

We also explored the impact of the degree of genetic relatedness between the recipient and donor genomes on genome on metaHGT performance. We used recipient genomes that showed varied genome relatedness to the donor genome, measured by AAI. We found no significant correlation between FP/FN error rates and genome relatedness, in the 41% to 87% AAI range ($P > 0.95$; Wilcoxon rank-sum test), confirming that our analysis is unlikely to be biased by the degree of genome relatedness. However, when the

recipient genome belonged to the same or closely related species as the donor genome, the false positive error increased rapidly. For example, when using *E. coli-E. fergusonii* pair (AAI ~91%), false positive error rate increased to ~35%, while false negative error rate remained low (~1%). Therefore, we do not recommend metaHGT for HGT predictions between genomes with higher than 90% AAI. The difficulty in detecting HGT events among closely related genomes is not specific to metaHGT but represents a limitation of most, if not all, approaches.

iii) Impact of closely related subpopulation(s).

Co-occurring closely related genomes to a target genome are sometimes observed in metagenomic dataset, and such genomes may confound HGT detection due to high sequence conservation. Therefore, we investigated the accuracy of metaHGT when a closely related genome to the recipient genome was present in the sample. We simulated this scenario by spiking *Escherichia* sp. TW09308 reads together with the (target) *E. coli* MG1655 reads into the *in-silico* generated metagenomic datasets (spiking in relatives of the recipient genomes did not differentiate our conclusions significantly). The total amount reads for the *"Escherichia* population" was fixed at 10X; however, the relative abundance of the reads from each of the two *Escherichia* genomes used varied from 10% to 90% in each trial. No significant correlations between error rate and the abundance of TW09308 were observed (P>0.95; Wilcoxon rank-sum test), which suggested that our analysis was not biased the presence of closely related strains (Figure 7.3).

Though the estimated error rates were sufficiently low for both false positives and negatives, we examined the genes that caused errors to obtain further insights into what

factors underlying the errors. We found that more than 50% of the false positive genes were ribosomal protein genes, and when excluded from further analysis, the false positive error rate dropped to ~3%. Therefore, ribosomal proteins were excluded in our time-series metagenome analysis.

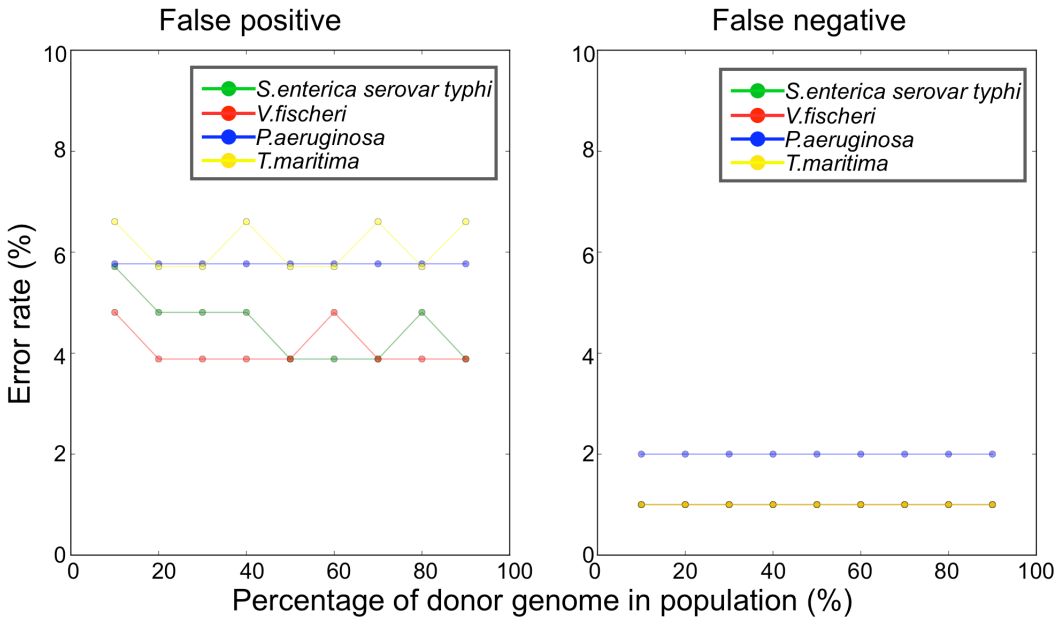


Figure 7.3. Co-occurring closely related genomes had no significant impact on the accuracy of metaHGT. *Escherichia* sp. TW09308 reads together reads from the (target) *E. coli* MG1655 were spiked into the in-silico metagenome. The total amount of reads from these two genomes was fixed at 10X but the relative abundance of reads from each genome varied. The results show that no significant variation was observed in the error rate of metaHGT algorithm (y-axes) in terms of both false positive (left) and false negative (right) when the relative abundance of the

donor/target genome (*E. coli* MG1655) varied (x-axes), for all donor-recipient pairs evaluated (figure legend).

HGT events occur frequently among distantly related populations

Our previous study characterized the planktonic microbial community of Lake Lanier based on both 16S ribosomal RNA gene (16S rRNA) amplicon and whole-genome shotgun sequencing (WGS) and revealed that the species complexity of this community is comparable to oceanic communities. It also showed that, with the exception of a few uncharacterized populations phylogenetically related to the *Burkholderiales* order and *Cyanobacteria* phylum, which together comprised about ~10% of the total community, all individual populations represented <1% of the total community (26). In the present study, we extended the previous efforts to sequence nine additional temporal samples (>40 Gb of Illumina 100-bp paired-end read data, in total) and devise a new bioinformatic pipeline to bin assembled contigs into population genomes for relatively abundant populations (making >0.1% of the community) (see *materials and methods* for details). 18 high quality (e.g., N50 > 10Kbp and estimated genome coverage > 85%) draft genomes were selected for further analysis (*LL1* to *LL18*; Figure 7.4 and Table 7.3). Phylogeny analysis based on six universally shared genes revealed that these genomes represented distinct lineages of several phyla, including the previously detected *Burkholderiales* and *Cyanobacteria* populations (populations *LL3, LL4*, and *LL8*). These 18 genomes also included three related low G+C% (28-32%) populations (*LL9, LL35,* and *LL38* in Figure 7.4) that represent a novel phylum-level lineage according to current classification standards (41).

metaHGT analysis revealed 256 HGT events among the 18 genomes during the period of 2.5 years spanned by our samples (Table S4), affecting up to 2.9% of the total genes in a genome (*LL3*). Intriguingly, most of the HGT events detected were among different phyla and occurred at higher frequencies (e.g., up to 56 genes per year for *LL3*) than estimated previously based on available complete genomes. For instance, Lawrence and Ochman estimated that HGT affected about 16 Kbp / Myr when comparing *E. coli* and *Salmonella sp.* Genomes. Extrapolations from our estimates for the *LL3* and *LL4* genome pair, which show similar genetic relatedness to that between *E. coli* and *Salmonella sp.* (i.e., 80% average amino acid identity), suggested up to $6.5! 10^7$ Kbp / Myr affected by HGT, i.e., about four million times higher HGT frequency. However, it is important to note that the HGT events detected here do not necessarily represent fixed mutational events but likely represent neutral and/or ephemeral mutations (see also below) whereas it is possible that most HGT events detected in (13) were fixed. It should be also mentioned that our estimates are likely underestimates of HGT frequency since our method considers only events that both the donor and the recipient are among the genomes analyzed and the genome sequences used here were not completed.

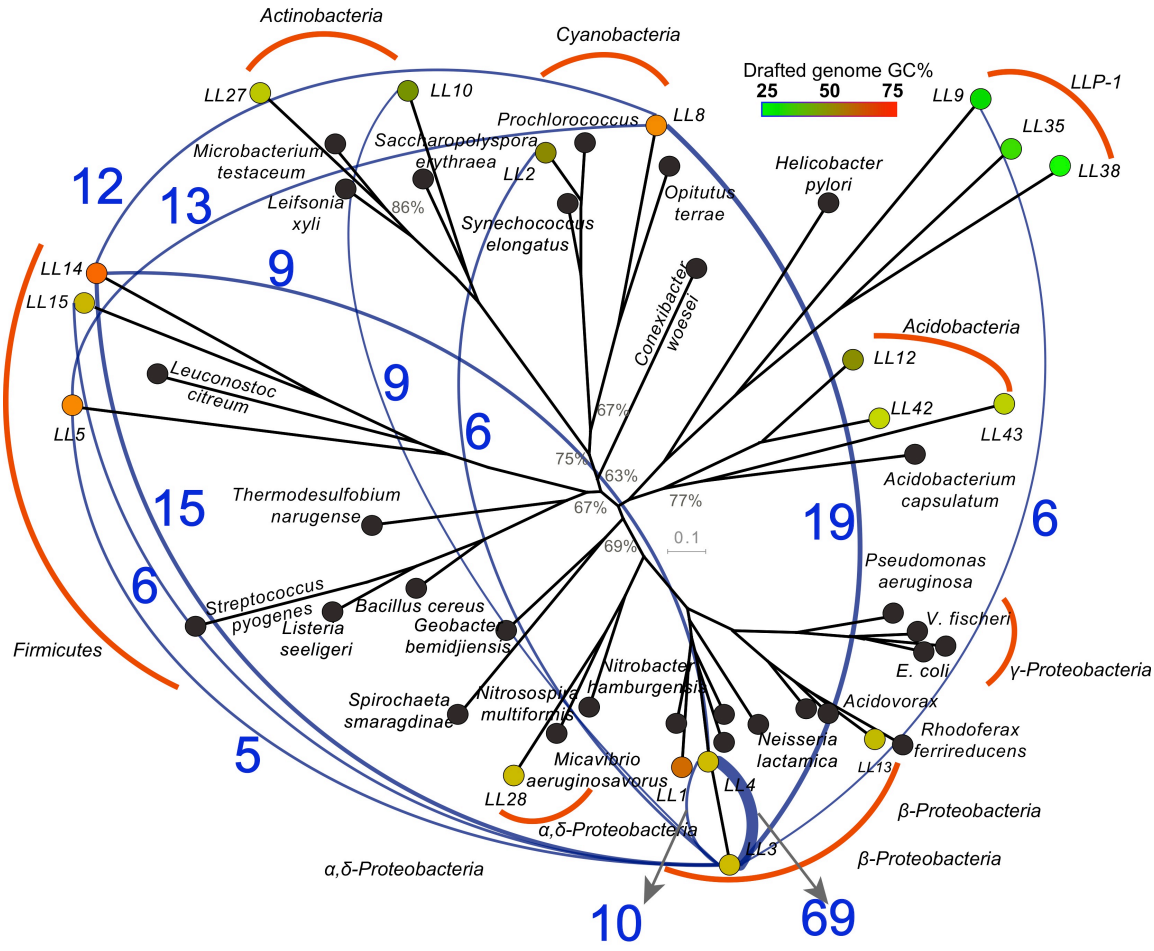


Figure 7.4. Phylogenetic relationships and network of horizontal gene transfer among the 18 population genomes recovered from Lake Lanier time series metagenomes. The phylogeny of the 18 genomes (colored circles; color also reflects genomic G+C% content, see figure key on top) and selected relatives from GenBank (black circles) is based on the concatenated multi-sequence alignment of five universally shared genes (*polA*, *dnaK*, *recA*, *lon*, and *mtf*). Branches with less than 90% bootstrap support from 1,000 replicates are shown in grey numbers. Blue lines connecting genomes denote HGT events detected between the genomes; the thickness of the line is proportional to the number of genes transferred (numbers are shown on the lanes in blue).

Table 7.3. Statistics of the 18 draft population genomes.

ID	Assembly size (Mbp)	Number of contigs	Assembly N50 (Kbp)	Max contig (Kbp)	Protein-coding genes	G+C%
LL1	1.869	316	10.9	59.7	2,356	63.48
LL2	2.053	99	38.3	125.9	2,315	50.38
LL3	2.469	33	206.0	803.3	2,391	56.76
LL4	2.644	72	94.9	157.6	2,573	54.96
LL5	1.378	499	3.7	26.4	1,934	60.64
LL8	1.318	100	21.4	64.1	1,412	58.86
LL9	2.186	308	13.5	214.5	2,625	30.06
LL10	0.979	95	16.5	71.1	1,173	44.61
LL12	2.402	253	15.2	44.7	2,598	41.92
LL13	1.533	219	11.2	33.3	1,714	55.26
LL14	2.678	551	7.0	29.2	2,719	62.19
LL15	1.341	384	5.85	65.8	1,969	52.19
LL27	0.882	76	17.0	91.7	970	46.27
LL28	1.138	128	12.7	39.5	1,223	56.37
LL35	2.638	327	15.9	70.4	2,792	31.59
LL38	2.264	559	4.0	18.5	1,996	28.27
LL42	3.375	600	11.8	81.7	4,373	45.31
LL43	2.747	389	13.2	67.2	2,924	37.39

Factors driving HGTs in natural settings

To provide insights into the molecular mechanisms that facilitated HGT, the (predicted) function of the transferred genes and their adjacent genes was examined. Transferred genes were significantly associated with mobile genes, primarily transposases, compared to the average gene in the genome ($P < 0.05$; G-test). About 25% of the transferred genes were found adjacent (within 1Kbp) to a transposon or integrase gene, while only 2/256 were found adjacent to a prophage gene (Figure 7.5A). The two genes adjacent to a prophage were also highly enriched in a viral metagenome originating from the same samples (unpublished observations), further supporting that they were indeed transferred via a bacteriophage. A similarly high proportion of mobile elements among HGT genes was observed previously based on the analysis of genomes of isolates from human samples (42).

Functional annotation of the transferred genes revealed that these were biased compared to the genome average toward four major categories, i.e., energy production and conversion, cell cycle control, amino acid and ion transport and metabolism (Figure 7.6). For example, antibiotic-related genes, including penicillin amidase (between *LL3* and *LL14*), multi-drug resistance protein B (between *LL2* and *LL4*), and penicillin-binding protein 1A (between *LL3* and *LL4*) and chemotaxis-related genes such as flagellar biosynthesis proteins (*flhB*; between *LL2* and *LL4*) were enriched among the horizontally transferred genes. We found that functions that require fewer genes were more likely to be transferred. For instance, permeases and transporters of the cell membrane, which are typically composed of only 2-3 genes such as ammonium transporter and sulfate permease, were over-represented among HGT events. In contrast, information-processing genes such as those involved in translation apparatus were significantly under-represented. The strong enrichment in energy and transport related functions indicate that at least some of the transferred genes might offer a selective advantage to the recipient population and are consistent with the complexity hypothesis raised by Jain and colleagues (43).

We next examined whether the degree of overlapping ecology and genetic relatedness among the partners significantly correlated with the frequency of HGTs. Previous studies have indicated that the these factors are important (11, 42) but a quantitative understanding of their relative importance is lacking. Though the number of highly related genome pairs (e.g., related at the genus or higher, corresponding to 50-80% AAI] was limited, a weak positive correlation between genetic relatedness and frequency of HGTs was observed (Figure 7.5B). We also evaluated how temporal co-occurrence

played a role in driving HGTs. A strong positive correlation was observed between the average relative abundance of the partners (across all sampling points) and the number of gene exchanged (Figure 7.5C). Given also that co-variance in abundance in the time-series data is likely to reflect (more) overlapping ecology, our results show that both ecological and genetic factors determine the HGT frequency.

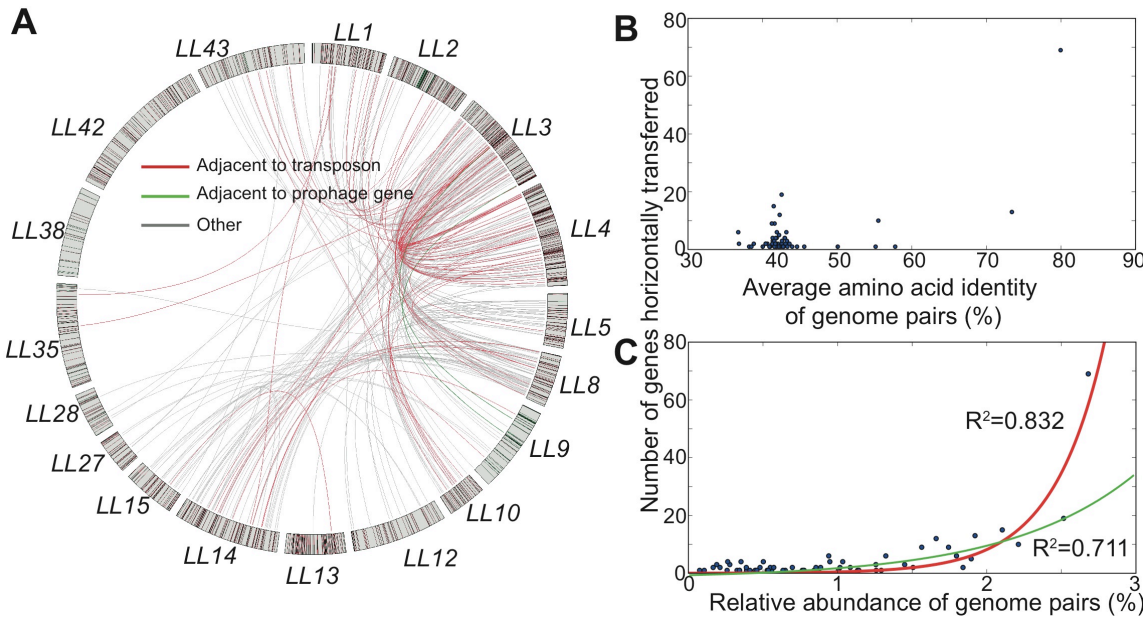


Figure 7.5. Factors driving HGT among the 18 population genomes. (A) Each draft genome is represented by a distinct piece of the ring. Lines represent HGT events between the genomes, colored-coded based on whether a transoposon/integrase (red lines; 25% of total cases) or a prophage gene (green lines, <1% of the cases) was found within 1Kbp from the transferred gene. All other cases are represented by gray lines. All transposons/integrases and prophage genes encoded on each genome are also marked with the same colors on the ring. **(B)** A weak positive trend between genome relatedness (measure by AAI; x-axis) and the number of HGTs (y-axis) was observed in the 50-80% AAI range. **(C)** A strong positive correlation between the relative abundance of the partners involved in HGT (see methods for details) and the number of genes exchanged was observed. Red line, all genomes pairs included; green line, the most closely related pair (*LL3-LL4*) was removed from the regression analysis.

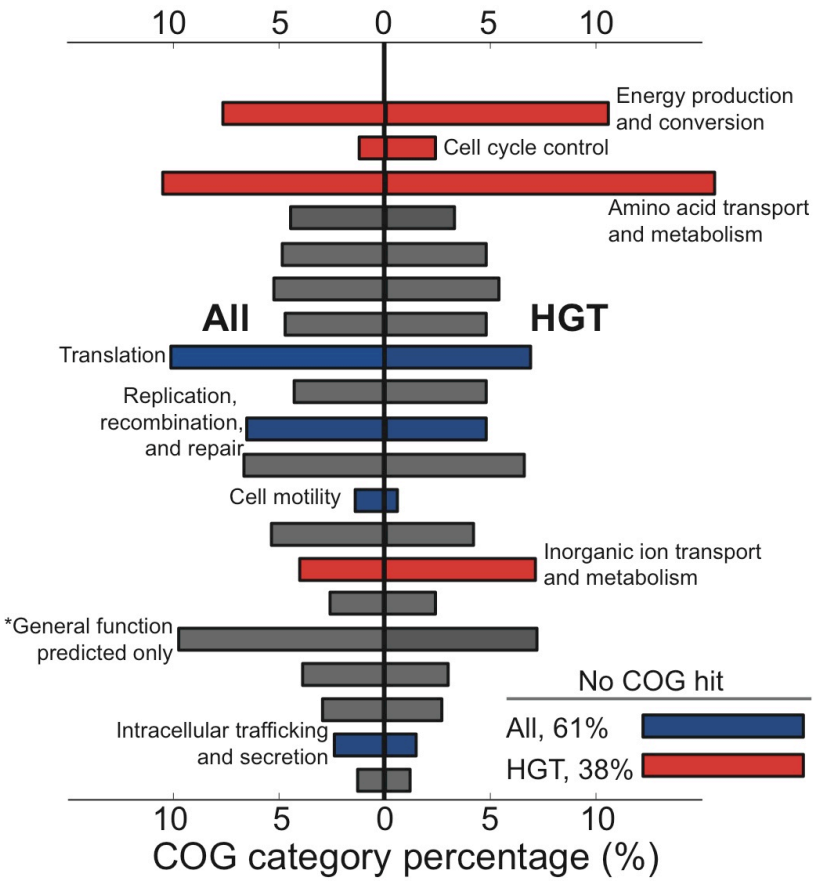


Figure 7.6. Functional biases of the horizontally transferred genes. Note that genes predicted to have undergone HGT (right) were enriched in functions related to energy production/conversion, and substrate transportation and metabolism (highlighted in red) compared to the genome average of the 18 population genomes (left).

The impacts of HGT on maintaining population diversity

It has been postulated that HGT could serve as both a diversifying and a homogenizing mechanism for microbial populations (15). If the transferred gene(s) provides a strong advantage to the recipient cell, this cell will outcompete co-occurring relatives, sweep through the population, and thus, cause the loss of intra-population diversity. Rampant HGT of the selected gene(s) will instead make the genes sweep though the population, maintaining intra-population genome diversity except at the loci of the transferred genes. Variants of the previous two theories have been also proposed recently. For instance, the fragmented speciation model suggests that the recipient organism(s) begin to accumulate mutations around the transferred gene or island compared to ecologically differentiated organisms that do not possess the gene because the inserted gene acts as a barrier to homologous recombination (12). Due to the complexity of the ecological niche of a population and the possibility that several HGTs could occur simultaneously, the positive and negative advantages of different HGT events can also cancel each other out, preventing populations sweeps and maintaining intra-population diversity [balancing selection; (10)].

To test these different models, we compared the 2009 to 2010 nucleotide diversity patterns of different regions of a large contig of *LL3* (contig18, 803Kbp). Our analyses revealed that more than 50% of the genes immediately adjacent to a transferred gene, which are present in all genomes of the recipient population unlike the gene(s) that was predicted to be transferred and was typically present in only part of the population based on coverage by raw metagenomic reads, showed almost no sequence divergence (Figure 7.7A). This contrasted with the remaining genes of the genome that showed significantly

higher divergence (*t*-test two tailed *P*-value: 1.98! 10^{-6}; Figure 7.7A). The level of sequence divergence in the latter genes was, however, comparable among the different samples (i.e., maintenance of intra-population diversity).

Moreover, the adjacent genes showing almost no divergence (<0.1 SNPs per Kbp) were highly enriched (19/25 genes) in secondary metabolism functions (e.g., D-lactate dehydrogenase) and substrate transporters (e.g., amino acid permease). For instance, the *sox* operon (sulfur oxidation function) involved in HGT between *LL3* and *LL5* showed no SNPs compared to the other genes further upstream or downstream (Figure 7.7B left). This finding implied that the *sox* operon, not the whole genome, swept through *LL3* population, presumably due to the selective advantage it offers. Further confirming these interpretations, genes adjacent to transferred genes with hypothetical or unknown functions, which are less likely to be functional and/or offer selective advantages (44), showed higher sequence divergence levels and hence no evidence of recombination-mediated sweeps through the population (e.g., Figure 7.7B right).

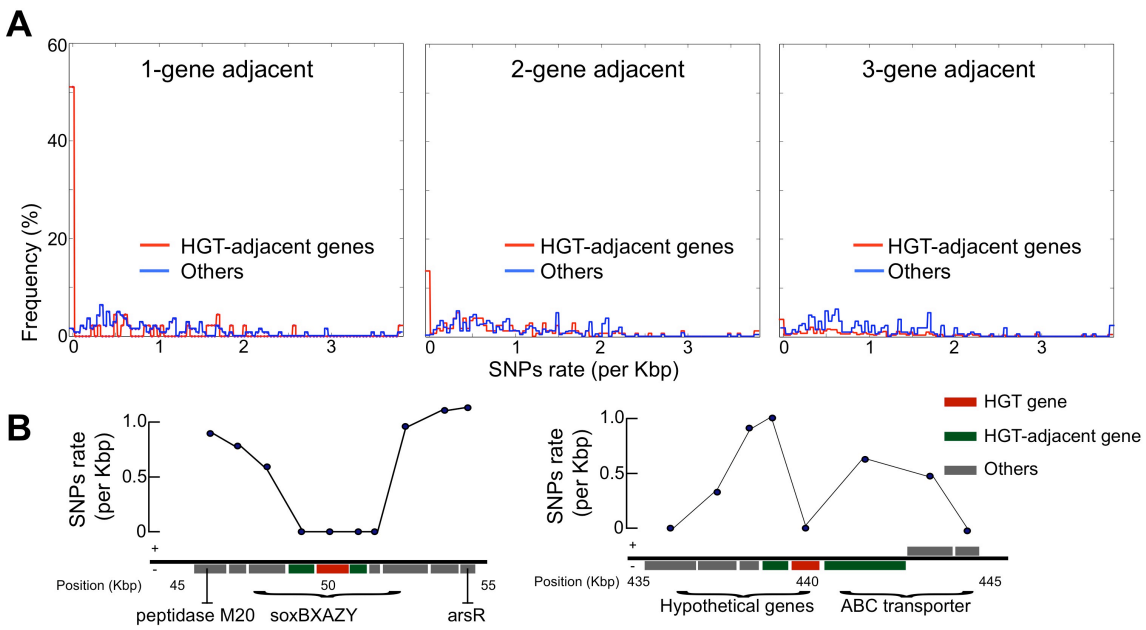


Figure 7.7. Genes adjacent to horizontally transferred genes showed significantly lower divergence level than other genes. An 803 Kbp long contig from *LL3* was used, which encoded 60 genes predicted to be exchanged between 2009 and 2010 with other genomes. **(A)** Distributions of SNP rate (proxy for gene divergence; y-axis) for HGT-adjacent genes (red line) and other genes (blue line) are shown. Data are shown separately for the first, second, and third upstream/downstream gene adjacent to the HGT event (top panels). Note the lack of sequence divergence (SNPs) in the adjacent genes, especially the first one. **(B)** Two representative examples of the SNP patterns observed around a horizontally transferred gene or operon. On the left, the sulfur oxidation *sox* operon exchanged between LL3 and LL5, which showed no SNPs in its adjacent genes, indicating recombination-mediated spreading within the population. On the right, the genes adjacent to a horizontally transferred hypothetical gene showed SNP levels similar to that of the whole genome. See also text for more details.

Horizontally transferred genes correlated with community dynamics

Carotenoid biosynthesis gene clusters were frequently observed among the horizontally transferred genes. Carotenoids represent a widely used class of molecules that carry out a variety of functions (45), including protection against sunlight radiation and oxidation (46, 47). This protection is presumably more important in the summer time (when sunlight is stronger) compared to wintertime. MetaHGT predicted that *LL3* donated carotenoid cleavage oxygenase and phytoene (precursor of carotenoid) synthase to *LL4* in summer 2009. Specifically, two independently predicted genes, phytoene synthase (precursor of carotenoids) and an osmolality sensor protein gene, flanking a 9 Kbp region in a large contig (206 Kbp) of *LL3* genome (Figure 7.8A), had two to three times higher coverage (Figure 7.8B) than the rest of the genome *LL3* ($P < 0.01$; G-test) and many PE reads links to *LL4* genome, indicating recent acquisition of these genes from *LL3*. Visual inspection of the paired-end read mapping along the flanking regions confirmed that these patterns were not due to misassembles or multiple gene copies, which would have been evident, for instance, by many outward reads (i.e., not mapping within the region of interest) mapping on other *LL3* contigs. Such patterns were not observed.

We further investigated the recipient(s) of this region and found that about 70% of *LL4* cells might acquired it in 2009, Consistent with the idea that the carotenoid biosynthesis genes represent an ecologically important island, we observed a high portion of outward reads mapping on other, lower abundance populations in 2010 that were not among the 18 populations primarily used in the analysis (Figure 7.8C). For instance, our analysis revealed that these low abundance populations contributed less than 5% of the

total reads mapping on the island in 2009, whereas their percentage of reads increased to ~30% in 2010. A search against NCBI-NR protein revealed that most of these reads were contributed by populations affiliated with *Thiorhodospira* and *Isosphaera* in 2009, and *Thioalkalivibrio* in 2010.

Interestingly, other than the phytoene synthase genes, this region also contained a *radA* homolog, which encodes repair enzymes for UV light-induced DNA damages. The Lake community is under higher solar radiation stress and more metabolically active, suggesting a higher level of oxidation stress in the summer- versus the winter-time (Figure 7.8C). Taken together, our findings suggested that the transferred island encodes ecologically important functions that can protect from sunlight radiation and oxidation and its horizontal transfer during the summertime might be associated with the observed community dynamics.

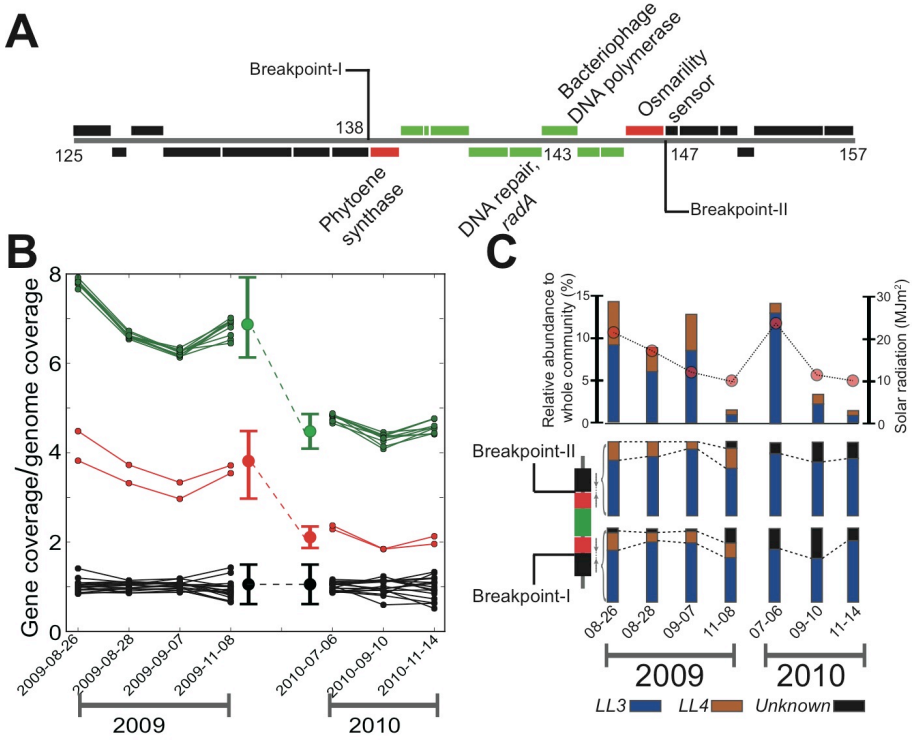


Figure 7.8. Carotenoid biosynthesis genes are frequently transferred during summer time in Lake Lanier microbial community. (**A**) The gene organization of the 9 Kbp region found in *LL3* genome and predicted to be horizontally transferred to *LL4* prior to August 2009. The flanking genes identified by metaHGT to be subjected to HGT are highlighted in red. (**B**) Each line represents the coverage of a gene, normalized to the abundance of the *LL3* genome, color-coded as in panel A. Vertical lines mark the range of coverage variation for 2009 and 2010 and dots mark the mean. Note that this region had 4~8 times higher coverage than the genomic background, further suggesting that it was transferred in other population(s). (**C**) The best match among all assembled genomes of an outward read, whose sister read mapped on the flanking genes (red in **A**), was determined and the graph represents the identity of the best matching genome (see figure key). Note that most 2010 reads mapped on many different genomes compared to 2009 ("Unknown" segments in black), indicating that this region has spread to other organisms.

Conclusions and future perspectives

This study represents, to the best of our knowledge, the first attempt to detect and quantify HGT within complex microbial communities, over time scales that are relevant for adaptation and human activities (e.g., a couple years), and to assess its impact on population diversity. Several novel bioinformatic approaches were developed to enable our study; most notably, how to directly recover draft genomes from time series metagenomes and how to detect HGTs. These approaches would be applicable to other natural or engineered systems, including the human microbiome.

Our study contrasts with most, if not all, previous studies that assessed historical HGT, occurring over a period of thousands of years. The frequency and number of genes exchanged among distantly related organisms such as members of different phyla were substantially higher compared to what was previously anticipated, and indicated that barriers to HGT flow might not be as important as previous analysis of cultured organisms indicated (11). In other words, cultivation biases have likely biased our view of the role of HGT, and other evolutionary processes, for community evolution and adaptation. Therefore, our findings are relevant to better understand and model the microbial diversity on the planet, including how HGT-mediated antibiotic resistance and highly virulent strains emerge from within a predominantly benign, naturally-occurring diversity.

At least some of the detected HGT events presumably underlay important community adaptation to short-term environmental perturbations and population dynamics. We showcased the transferred of carotenoid-related pigmentation and DNA repair genes that may provide protection against oxidative damage and intense solar

radiation during summertime. More generally speaking, functions with high probability to be ecologically important, such as chemotaxis genes, substrate permease/transporters, and antibiotics genes, were subjected to frequent HGT but also deletion. Thus, the obvious conflict between high frequency of HGTs and the conserved size of the genomes must be reconciled via low rate of fixation. Future investigations should focus to better understand the mechanisms of HGT among distantly related organisms, and how frequently the horizontally transferred genes are actually functional and confer an ecological advantage. The methods developed and the lessons learned here will greatly facilitate such studies.

REFERENCES

1. Barrick JE*, et al.* (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461(7268):1243-1247.
2. Blount ZD, Barrick JE, Davidson CJ, & Lenski RE (2012) Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489(7417):513-518.
3. Elena SF & Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature reviews. Genetics* 4(6):457-469.
4. Garneau JE*, et al.* (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468(7320):67-71.
5. Martincorena I, Seshasayee AS, & Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485(7396):95-98.
6. Saks ME, Sampson JR, & Abelson J (1998) Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science* 279(5357):1665-1670.
7. Venter JC*, et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.
8. Caporaso JG*, et al.* (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* 108 Suppl 1:4516-4522.
9. Yooseph S*, et al.* (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 468(7320):60-66.
10. Leffler EM*, et al.* (2012) Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS biology* 10(9):e1001388.
11. Popa O & Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Current opinion in microbiology* 14(5):615-623.
12. Retchless AC & Lawrence JG (2007) Temporal fragmentation of speciation in bacteria. *Science* 317(5841):1093-1096.
13. Lawrence JG & Ochman H (1998) Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16):9413-9417.
14. Kloesges T, Popa O, Martin W, & Dagan T (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular biology and evolution* 28(2):1057-1074.
15. Papke RT & Gogarten JP (2012) How Bacterial Lineages Emerge. *Science* 336(6077):45-46.
16. Dinsdale EA*, et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629-632.
17. Wrighton KC*, et al.* (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337(6102):1661-1665.
18. Oh S*, et al.* (Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77(17):6000-6011.

19. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18(5):821-829.
20. Gilbert JA*, et al.* (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Standards in genomic sciences* 3(2):183-193.
21. Luo C, Tsementzi D, Kyrpides NC, & Konstantinidis KT (2011) Individual genome assembly from complex community short-read metagenomic datasets. *The ISME journal*.
22. Luo C, Tsementzi D, Kyrpides N, Read T, & Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one* 7(2):e30087.
23. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome research* 12(4):656-664.
24. Iverson V*, et al.* (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335(6068):587-590.
25. Denef VJ & Banfield JF (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336(6080):462-466.
26. Oh S*, et al.* (2011) Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77(17):6000-6011.
27. Calinski T (1968) A Dendrite Method for Cluster Analysis. *Biometrics* 24(1):207-&.
28. Arumugam M*, et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.
29. Zhu W, Lomsadze A, & Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic acids research* 38(12):e132.
30. Ogata H*, et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* 27(1):29-34.
31. Jensen LJ*, et al.* (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic acids research* 36(Database issue):D250-254.
32. Tatusov RL, Galperin MY, Natale DA, & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research* 28(1):33-36.
33. Overbeek R*, et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research* 33(17):5691-5702.
34. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25(17):3389-3402.
35. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792-1797.
36. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* 59(3):307-321.

37. Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23(2):254-267.
38. Holm S (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat* 6(2):65-70.
39. Konstantinidis KT & Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2567-2572.
40. Luo C*, et al.* (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America* 108(17):7200-7205.
41. Konstantinidis KT & Tiedje JM (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 10(5):504-509.
42. Smillie CS*, et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.
43. Jain R, Rivera MC, & Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 96(7):3801-3806.
44. Daubin V & Ochman H (2004) Bacterial Genomes as new gene homes: The genealogy of ORFans in E-coli. *Genome research* 14(6):1036-1042.
45. Goodwin TW (1986) Metabolism, nutrition, and function of carotenoids. *Annual review of nutrition* 6:273-297.
46. Slouf V*, et al.* (2012) Photoprotection in a purple phototrophic bacterium mediated by oxygen-dependent alteration of carotenoid excited-state properties. *Proceedings of the National Academy of Sciences of the United States of America* 109(22):8570-8575.
47. Dieser M, Greenwood M, & Foreman CM (2010) Carotenoid Pigmentation in Antarctic Heterotrophic Bacteria as a Strategy to Withstand Environmental Stresses. *Arct Antarct Alp Res* 42(4):396-405.

ACKNOWLEDGEMENTS

This work was supported by U. S. National Science Foundation under Award No 1241046.

CONCLUTIONS AND PERSPECTIVES FOR THE FUTURE

As a model organism, *E. coli* represents the most thoroughly studied species; it is among the earliest genome sequenced and is now represented by a large collection of sequenced representative strains (n > 100). However, due to the fact that most of the *E. coli* studies have been conducted under laboratory conditions, it is still poorly understood how this model organism evolves under natural settings and how ecology plays a role in shaping its genome. Therefore, I started my research by comparing the genomes of nine environmental (i.e., adapted to live outside human or animal hosts) free-living *Escherichia* strains, closely related to and phenotypically indistinguishable from typical *E. coli,* against 15 selected gut-associated (enteric) *E. coli* strains to determine whether or not gene-signatures specific to the habitat of isolation (free-living vs. gut-associated) were detectable and assess the importance of ecology in driving gene content differences among these genomes and speciation. It was found that among the dispensable genes of the *Escherichia* pan-genome, ecologically advantageous genes such as genes for fucose utilization and acetylglucosamine transports were enriched in enteric genomes; a pattern that correlated with the nutrients available in the human gut and hence, ecology. Moreover, a higher frequency of genetic exchange was observed within enteric or environmental genomes compared to between genomes from the two groups, indicating that overlapping ecology also favors more frequent horizontal gene transfer.

These observations suggested that it is important to consider bacterial populations under natural settings to better understand and quantify the role of ecology in genome and lineage evolution. For instance, the results from the analysis of the *Escherichia* genomes

indicted that co-occurring relatives may be directly involved in more horizontal gene transfer than previously thought, shaping the evolutionary trajectory of a specific lineage or population. It is important to note that bacteria were thought to evolve primarily asexually and only rarely exchange DNA about a decade ago. With these considerations in mind, it became obvious to me early on that studying natural communities over time would provide new quantitative insights into the evolution process and the underling mechanisms. Metagenomics is a promising approach for these purposes; however, several technical challenges remained at that time.

One challenge was the inherent tradeoffs between read length and sequencing output among the available DNA sequencing technologies. My initial assessment indicated that a deep sequencing coverage was needed to robustly quantify the genetic events within a complex microbial community (1), which only short read sequencing (e.g., Illumina) could provide. To account for the limitations of the short read length and robustly assemble short reads into longer pieces, I developed a hybrid protocol that combined different assembly algorithms and cut-offs. This protocol typically offered ~30% improvement in terms of the number of sequences assembled into longer contigs compared to the current state-of-the-art methods such as Velvet (2), while still maintaining comparable assembly quality (1). This protocol has been tested on various metagenomes, with robust performance (3).

Recent developments in DNA sequencing technologies may further improve the assemblies. For example, the single molecule sequencing platform PACBIO RS by Pacific Biosciences (4, 5) offers higher sequencing depth (~150Mbp per SMRT cell, highly parallelizable) and longed reads (~1.5 Kbp) at low cost. However, its current high

error rate in base calling (~15% error rate) has limited its application in metagenomic studies. The highly parallelizable nanopore-based sequencing modules, such as the GridION and MinION provided by Oxford Nanopore Technologies (6), represent another promising solution. Longer reads, especially when longer than the typical bacterial/archaeal repeat regions, could help to resolve ambiguous assembly paths and thus, extend contig length. Longer reads will also substantially reduce the computational effort required during assembly (especially memory requirements) and thus, make metagenomics more accessible to laboratories with limited computational infrastructure. However, a careful assessment of sequencing biases and artifacts such as base call errors or G+C% biases, similar to those described in Chapter 4 for Roche 454 and Illumina platforms, is necessary before the new technologies can be used in metagenomics.

Aside from sequencing technologies, single cell genomics (SCG) offers an alternative technology, highly complementary to metagenomics. By isolating and sequencing cells directly from environmental samples, single cell technologies bypass the complication of mixing genomic fragments from different organism during assembly, which represents the greatest challenge for metagenomics (7-9). Because SCG produces reference genomes without cultivation, it can be also integrated with other culture-independent omics technologies (e.g., transcriptomics) for more robust results and interpretations (10). For instance, SCG could provide the reference genomes of low abundance (rare) members of a natural community, which are economically impractical to cover adequately and assemble with metagenomics, and therefore offer new insights into the role and activity of these organisms. Rare organisms frequently make up a large part of the community in natural aquatic or soil habitats (the “rare biosphere”) (11), but

their importance for community function remains essentially elusive (11, 12). There are still many questions and challenges to be addressed regarding SCG, such as biased cell sorting and lysis, DNA contamination (13). However, SCG represent a very promising technology, especially when integrated with high coverage metagenomics data.

Another major challenge for metagenomics is to determine the taxonomic affiliation of assembled or raw sequences since more than 50% of the reads remain unassigned to a species in a typical metagenome. To overcome this limitation, a new approach, MeTaxa, was developed. In addition to taxonomic classification of sequences with previously characterized (known) close relatives, MeTaxa can identify novel organisms and their degree of novelty (e.g., novel species, genus or phylum). Therefore, MeTaxa can help identifying novel organisms that are abundant in an environment and thus, presumably important and preferred targets for isolation efforts and/or single cell genomics. MeTaxa outperformed all previously published methods in terms of the number of sequences accurately assigned, based on both *in-silico* (test) and real metagenomes, and requires low computational resources for the analysis part. The remaining challenges in this area originate from the lack of a comprehensive reference genome database, which renders the taxonomic assignment of a larger number of environmental sequences practically impossible. Certainly, the abovementioned single cell genomic technologies as well as international sequencing efforts to sequence a large number of isolates [e.g., the Genomic Encyclopedia of Bacteria and Archaea project, or GEBA; (14)], will very likely outgrow the previous limitation in the long term. In the short term, a possible solution could be to create a framework that integrates prediction results from different methods and thus achieves higher accuracy and prediction coverage

as well as to devise a framework to classify new organisms without necessarily cultivating them first. An analog could be found in the field of prokaryotic gene prediction. It was a major challenge to predict genes based on a single approach/tool when the number of available prokaryotic genomes was less than 200 around year 2006, and thus, a consensus strategy that combined several different tools was proposed and implemented (15). This strategy showed substantial improvement over any single method at the time (16), and played an important role in important new discoveries on novel gene functions (17).

To quantify horizontal gene transfer (HGT) frequency within a natural microbial community (Chapter 7), novel methods for reconstructing genomes from metagenomes and detecting HGT events among these genomes were developed. My genome recovering pipeline is broadly applicable to time-series metagenomic data from various habitats and levels of community complexity, and, if combined with SCG techniques, it could greatly assist in expanding the collection of uncultured reference genomes. The HGT detecting algorithm, metaHGT, is based on changes in pair-ended read mapping and coverage across time-series metagenomic data and was the first tool to predict HGT events, with high confidence, in complex metagenomes. Thus, MetaHGT, combined with the lessons learned from the work described in chapters 3 and 4 on genome assembly from complex metagenomes, can greatly facilitate future metagenomic experiments and guide experimental design.

With the tools developed as part of this thesis, I tackled several important evolutionary and ecological questions about natural microbial communities, using both a soil and a freshwater microbial system. In the comparative soil metagenomics study

(Chapter 6), read-centric analyses were carried out to identify the responses of the microbial communities inhabiting a temperate grassland soil to increased temperatures that simulated the global climate change (i.e., 2 ^{0}C infrared warming for ten years). The analyses revealed that the heated communities showed significant shifts in composition and predicted metabolism compared to control (un-heated) communities and these shifts were community-wide as opposed to being attributable to a few taxa. Key metabolic pathways related to the turnover of carbon, e.g., cellulose degradation (~13%) and CO_2 production (~10%), and nitrogen, e.g., denitrification (~12%), were enriched under warming, and these community shifts were interlinked, in part, with higher primary productivity of the aboveground plant communities stimulated by warming. Collectively, these results indicated that the microbial communities of the temperate soils play important roles in mediating the feedback responses to climate change and advance understanding of the modes and tempo of community adaptation to environmental change.

The analyses of 3 years of metagenomic data originating from planktonic samples collected at the same site in Lake Lanier (Atlanta, GA) revealed a high frequency of gene transfer events among distantly related members of the community (e.g., inter-phylum gene transfers), much higher than previously anticipated (18, 19). Several of these transfer events were related to specific population dynamics and intra-population diversity patterns. For instance, it was found that genes in the carotenoid biosynthesis pathway, which protects cells from solar radiation and oxidative DNA damage, were frequently transferred horizontally and these HGT events were directly associated with changes in the relative abundance of several populations, especially during the summertime when the sunlight intensity was high. This study also provided experimental

data to test several of the prevailing theories on how diversity within bacterial populations is maintained. It was found that when an exchanged gene (or pathway) possibly introduced an ecological (selective) advantage the gene and its flanking region swept through the population via further genetic exchange, thus, maintaining intra-population diversity. In contrast, when the exchanged genes were likely to be of low advantage to the population (e.g., hypothetical proteins), the gene and its flanking regions usually presented higher diversity than the genome average, indicating relaxed purifying selection. These findings are consistent with a more sexual evolution of bacterial populations than previously anticipated and do not support the fragmented speciation model (20) or frequent population sweeps caused by adaptive evolution (21, 22). However, in order for more robust conclusions to emerge, more studies that cover additional microbial groups and habitats are necessary. The work presented here provided several important findings and the enabling bioinformatics tools toward this direction.

While my studies provided important new insights, they also brought into more sharp focus several important questions that remain to be addressed toward a better understanding of the microbial world. In the first case, soil microbial communities harbored extreme diversity (e.g., I estimated >4,000 species per gram of soil; >350 Gb of data require to capture 95% of this species diversity) that cannot be efficiently covered based on current sequencing practices and technologies. Accordingly, the soil metagenomes remained largely unassembled, e.g., the assembled contigs typically recruited only <1% of the overall reads, despite the relatively large sequencing effort applied (>10 Gb of data per sample). Capturing the total diversity within communities is at the heart of answering several important questions such as how extremely complex

communities assemble (e.g., stochastic vs. driven by environmental parameters) and how the low abundance, "rare" members contribute to the resilience/robustness of the community upon perturbations. The currently proposed bioinformatics solutions to the assembly problem focus on reducing the read space complexity by applying compressing techniques such as Bloom Filter (23) and digital normalization (24) in order to provide the assembler with a less complex, and easily to assemble dataset. In other words, these approaches scale the metagenome assembly by reducing resource usage and removing sequencing noise. Significant improvements have been achieved; however, not at the level necessary to robustly assemble draft genomes from highly complex microbial communities such as the soil ones. New computational efforts as well as new sequencing technologies are still needed to further push the frontiers of soil metagenomics. For instance, single-molecule sequencing with longer reads (e.g., 1-2Kbp) will greatly facilitate the assembly; single cell genomics can recover difficult-to-culture yet important reference genomes. The availability of a more comprehensive collection of soil reference genomes will be instrumental in addressing several remaining questions such as those reported above but also what genetic mechanisms maintain the extremely high community complexity in soils and what is the extent of interactions among community members upon environmental perturbations.

In the Lake Lanier time-series metagenomics study, HGT was found to frequently occur between distantly related yet co-habiting organisms. Although these findings advanced our understanding of the mechanisms and rates of HGTs *in-situ*, they also highlighted several interesting questions for future investigation. First the mechanisms underling HGT events between highly divergent organisms require further attention, as

most known agents of HGT are thought to operate among closely related organisms (19). Recent advancements in fluorescent *in situ* hybridization (FISH) and polymeric sequence probe technologies (25) might be helpful in this direction. On the other hand, the importance of the horizontally transferred genes for community evolution and adaptation to perturbations remains unclear for most HGT events. For example, are the transferred genes functional and how frequent are they fixed (selected) in the recipient population? How do the transferred genes, especially those that could directly alter the interactions among community members, impact the adaptation and evolution of the community? To answer these questions, high-level integration of multi-omics data and mathematical modeling will be necessary. HGT is an important component of understanding how microbial lineages evolve, and it is intertwined with other processes including functional innovation, diversification, and population dynamics. It remains challenging, however, to disentangle several of these processes; for instance, to distinguish genomic adaptation from the effect of interactions (e.g., competition) with other co-occurring populations and assess the impact of genotypes coming from outside the system (e.g., through aerial dispersion), as this was exemplified in the work presented in chapter 7. Novel experiment designs and long time series data, are expected to provide new insights into these remaining questions and advance our understanding of *in situ* microbial evolution within complex natural communities.

REFERENCES

1. Luo C, Tsementzi D, Kyrpides NC, & Konstantinidis KT (2012) Individual genome assembly from complex community short-read metagenomic datasets. *The ISME journal* 6(4):898-901.
2. Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18(5):821-829.
3. Luo C, Tsementzi D, Kyrpides N, Read T, & Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one* 7(2):e30087.
4. Eide K, Miller-Morgan T, Heidel J, Bildfell R, & Jin L (2011) Results of total DNA measurement in koi tissue by Koi Herpes Virus real-time PCR. *Journal of virological methods* 172(1-2):81-84.
5. Eid J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133-138.
6. Olasagasti F*, et al.* (2010) Replication of individual DNA molecules under electronic control using a protein nanopore. *Nature nanotechnology* 5(11):798-806.
7. Ghai R*, et al.* (2011) New abundant microbial groups in aquatic hypersaline environments. *Scientific reports* 1:135.
8. Woyke T*, et al.* (2009) Assembling the marine metagenome, one cell at a time. *PloS one* 4(4):e5299.
9. Woyke T*, et al.* (2010) One bacterial cell, one complete genome. *PloS one* 5(4):e10314.
10. Stepanauskas R (2012) Single cell genomics: an individual look at microbes. *Current opinion in microbiology*.
11. Sogin ML*, et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* 103(32):12115-12120.
12. Pedros-Alio C (2007) Ecology. Dipping into the rare biosphere. *Science* 315(5809):192-193.
13. Chitsaz H*, et al.* (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature biotechnology* 29(10):915-921.
14. Wu D*, et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462(7276):1056-1060.
15. Luo C, Hu GQ, & Zhu H (2009) Genome reannotation of Escherichia coli CFT073 with new insights into virulence. *BMC genomics* 10:552.
16. Petty NK (2010) Genome annotation: man versus machine. *Nature reviews. Microbiology* 8(11):762.
17. Aoki SK*, et al.* (2010) A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. *Nature* 468(7322):439-442.
18. Lawrence JG & Ochman H (1998) Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16):9413-9417.

19. Popa O & Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Current opinion in microbiology* 14(5):615-623.
20. Retchless AC & Lawrence JG (2007) Temporal fragmentation of speciation in bacteria. *Science* 317(5841):1093-1096.
21. Cohan FM & Koeppel AF (2008) The origins of ecological diversity in prokaryotes. *Current biology : CB* 18(21):R1024-1034.
22. Koeppel A*, et al.* (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences of the United States of America* 105(7):2504-2509.
23. Pell J*, et al.* (2012) Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America* 109(33):13272-13277.
24. Titus Brown CH, Adina; Zhang, Qingpeng; Pyrkosz, Alexis B.; Brom, Timothy H. (2012) A reference-free algorithm for computational normailization of shortgun sequencing data. *eprint arXiv* (1203.4802).
25. Teeling H*, et al.* (2012) Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* 336(6081):608-611.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A1. List of genes distinguishing environmental from enteric genomes. This data underlies the heatmap shown in Figure 2B.

Gene ID	Gene name	COG ID	Category	COG annotation	Enriched section
PAN000006					enteric
PAN000191	ldcC	COG1982	E	Arginine/lysine/ornithine decarboxylases	enteric
PAN000222					enteric
PAN000224	ECs0224	COG3455	S	Uncharacterized protein conserved in bacteria	enteric
PAN000226	ECs0226	COG3521	S	Uncharacterized protein conserved in bacteria	enteric
PAN000233	ECs0234	COG3157	S	Hemolysin-coregulated protein (uncharacterized)	enteric
PAN000346					enteric
PAN000347					enteric
PAN000348	ECs0324	COG2771	K	DNA-binding HTH domain-containing proteins	enteric
PAN000358					enteric
PAN000369					enteric
PAN000377	betA	COG2303	E	Choline dehydrogenase and related flavoproteins	enteric
PAN000378	betB	COG1012	C	NAD-dependent aldehyde dehydrogenases	enteric
PAN000379	ECs0359	COG1309	K	Transcriptional regulator	enteric
PAN000381	ECs0360	COG1292	M	Choline-glycine betaine transporter	enteric
PAN000392	ECs0379	COG1064	R	Zn-dependent alcohol dehydrogenases	enteric
PAN000410	yaiM	COG0627	R	Predicted esterase	enteric
PAN000440					enteric
PAN000534	ECs0537	COG2217	P	Cation transport ATPase	enteric
PAN000638					enteric
PAN000644	ybeF	COG0583	K	Transcriptional regulator	enteric
PAN000777	ybhM	COG0670	R	Integral membrane protein, interacts with FtsH	enteric
PAN000977	ycdG	COG2233	F	Xanthine/uracil permeases	enteric
PAN000978	ECs1253	COG1853	R	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	enteric
PAN000979	ECs1254	COG0778	C	Nitroreductase	enteric
PAN000980	ycdJ	COG0596	R	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	enteric
PAN000981	ycdK	COG0251	J	Putative translation initiation inhibitor, yjgF family	enteric
PAN000982	ycdL	COG1335	Q	Amidases related to nicotinamidase	enteric
PAN000983	ycdM	COG2141	C	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase	enteric
PAN000992					enteric
PAN000993	ycdQ	COG1215	M	Glycosyltransferases, probably involved in cell wall biogenesis	enteric
PAN000994	ycdR	COG0726	G	Predicted xylanase/chitin deacetylase	enteric
PAN001028	ECs1442	COG2999	O	Glutaredoxin 2	enteric
PAN001169	ycgE	COG0789	K	Predicted transcriptional regulators	enteric
PAN001170	ycgF_2	COG2200	T	FOG: EAL domain	enteric
PAN001184	ZypjA	COG3468	MU	Type V secretory pathway, adhesin AidA	enteric
PAN001190					enteric
PAN001331	ycjM	COG0366	G	Glycosidases	enteric
PAN001332	ycjN	COG1653	G	ABC-type sugar transport system, periplasmic component	enteric
PAN001333	ycjO	COG1175	G	ABC-type sugar transport systems, permease components	enteric
PAN001334	ECs1891	COG0395	G	ABC-type sugar transport system, permease component	enteric
PAN001335	ycjQ	COG1063	ER	Threonine dehydrogenase and related Zn-dependent dehydrogenases	enteric
PAN001336	ycjR	COG1082	G	Sugar phosphate isomerases/epimerases	enteric
PAN001337	ECs1894	COG0673	R	Predicted dehydrogenases and related proteins	enteric
PAN001338	ycjT	COG1554	G	Trehalose and maltose hydrolases (possible phosphorylases)	enteric
PAN001339	ycjU	COG0637	R	Predicted phosphatase/phosphohexomutase	enteric
PAN001342					enteric
PAN001343	ycjW	COG1609	K	Transcriptional regulators	enteric
PAN001352	ycjZ	COG0583	K	Transcriptional regulator	enteric
PAN001381	ECs2010	COG0558	I	Phosphatidylglycerophosphate synthase	enteric

Table A1 (continued)

PAN001382	ynbB	COG4589	R	Predicted CDP-diglyceride synthetase/phosphatidate cytidylyltransferase	enteric
PAN001383	Z2317_2	COG0500	QR	SAM-dependent methyltransferases	enteric
PAN001384	ECs2013_1	COG0671	I	Membrane-associated phospholipid phosphatase	enteric
PAN001391					enteric
PAN001420	ECs2058	COG0625	O	Glutathione S-transferase	enteric
PAN001459	ECs2104	COG2207	K	AraC-type DNA-binding domain-containing proteins	enteric
PAN001460					enteric
PAN001461	ydeP	COG0243	C	Anaerobic dehydrogenases, typically selenocysteine-containing	enteric
PAN001462	ydeS	COG3539	NU	P pilus assembly protein, pilin FimA	enteric
PAN001463	Z2203	COG3188	NU	P pilus assembly protein, porin PapC	enteric
PAN001677	ydjE	COG0477	GEPR	Permeases of the major facilitator superfamily	enteric
PAN001678	ECs2479	COG1349	KG	Transcriptional regulators of sugar metabolism	enteric
PAN001679	ydjG	COG0667	C	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	enteric
PAN001680	ydjH	COG0524	G	Sugar kinases, ribokinase family	enteric
PAN001681	ydjI	COG0191	G	Fructose/tagatose bisphosphate aldolase	enteric
PAN001682	ydjJ	COG1063	ER	Threonine dehydrogenase and related Zn-dependent dehydrogenases	enteric
PAN001683	ydjK	COG0477	GEPR	Permeases of the major facilitator superfamily	enteric
PAN001684	ECs2485	COG1063	ER	Threonine dehydrogenase and related Zn-dependent dehydrogenases	enteric
PAN001804					enteric
PAN001853					enteric
PAN001858	yedV	COG0642	T	Signal transduction histidine kinase	enteric
PAN001859	ECs2707	COG0745	TK	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	enteric
PAN001860	ECs2708	COG2351	R	Transthyretin-like protein	enteric
PAN001864	yodA	COG3443	R	Predicted periplasmic or secreted protein	enteric
PAN002081					enteric
PAN002341	emrY	COG0477	GEPR	Permeases of the major facilitator superfamily	enteric
PAN002342	ECs3247	COG1566	V	Multidrug resistance efflux pump	enteric
PAN002343	ECs3248	COG2197	TK	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	enteric
PAN002344	ECs3249_1	COG0834	ET	ABC-type amino acid transport/signal transduction systems, periplasmic component/domain	enteric
PAN002345	yfdE	COG1804	C	Predicted acyl-CoA transferases/carnitine dehydratase	enteric
PAN002346	ECs3252	COG0679	R	Predicted permeases	enteric
PAN002347	ECs3253	COG0028	EH	Thiamine pyrophosphate-requiring enzyme	enteric
PAN002348	yfdW	COG1804	C	Predicted acyl-CoA transferases/carnitine dehydratase	enteric
PAN002349					enteric
PAN002351					enteric
PAN002713	ygcY	COG4948	MR	L-alanine-DL-glutamate epimerase and related enzymes of enolase superfamily	enteric
PAN002726	fucP	COG0738	G	Fucose permease	enteric
PAN002727	fucI	COG2407	G	L-fucose isomerase and related proteins	enteric
PAN002728	fucK	COG1070	G	Sugar (pentulose and hexulose) kinases	enteric
PAN002729	ECs3664	COG4154	G	Fucose dissimilation pathway protein FucU	enteric
PAN002730	fucR	COG1349	KG	Transcriptional regulators of sugar metabolism	enteric
PAN003148					enteric
PAN003163	ECs4015	COG1820	G	N-acetylglucosamine-6-phosphate deacetylase	enteric
PAN003654	ECs4507	COG0859	M	ADP-heptose:LPS heptosyltransferase	enteric
PAN003875	ECs4696	COG0477	GEPR	Permeases of the major facilitator superfamily	enteric
PAN003895	ECs4721	COG1088	M	dTDP-D-glucose 4,6-dehydratase	enteric
PAN003896	rffH	COG1209	M	dTDP-glucose pyrophosphorylase	enteric
PAN004231	ECs5075	COG1235	R	Metal-dependent hydrolases of the beta-lactamase superfamily I	enteric
PAN004232	phnO	COG0454	KR	Histone acetyltransferase HPA2 and related acetyltransferases	enteric
PAN004234	ECs5078	COG3454	P	Metal-dependent hydrolase involved in phosphonate metabolism	enteric
PAN004235	phnL	COG4778	P	ABC-type phosphonate transport system, ATPase	enteric

Table A1 (continued)

PAN004236	ECs5080	COG4107	P	ABC-type phosphonate transport system, ATPase component	enteric
PAN004237	Ecs5081	COG3627	P	Uncharacterized enzyme of phosphonate metabolism	enteric
PAN004238	Ecs5082	COG3626	P	Uncharacterized enzyme of phosphonate metabolism	enteric
PAN004240	Ecs5084	COG3624	P	Uncharacterized enzyme of phosphonate metabolism	enteric
PAN004241	phnF	COG2188	K	Transcriptional regulators	enteric
PAN004242	Ecs5086	COG3639	P	ABC-type phosphate/phosphonate transport system, permease component	enteric
PAN004243	phnD	COG3221	P	ABC-type phosphate/phosphonate transport system, periplasmic component	enteric
PAN004244	phnC	COG3638	P	ABC-type phosphate/phosphonate transport system, ATPase component	enteric
PAN004533	Ecs5271	COG0582	L	Integrase	enteric
PAN004550	yjiE	COG0583	K	Transcriptional regulator	enteric
PAN004553	Ecs5288	COG3314	S	Uncharacterized protein conserved in bacteria	enteric
PAN004966					enteric
PAN004997					enteric
PAN005007	Ecs2293	COG0243	C	Anaerobic dehydrogenases, typically selenocysteine-containing	enteric
PAN005008	ynfF	COG0243	C	Anaerobic dehydrogenases, typically selenocysteine-containing	enteric
PAN005009	Ecs2295	COG0437	C	Fe-S-cluster-containing hydrogenase components 1	enteric
PAN005010	Ecs2296	COG3302	R	DMSO reductase anchor subunit	enteric
PAN005305					enteric
PAN005773	relE	COG2026	JD	Cytotoxic translational repressor of toxin-antitoxin stability system	Env.
PAN006230	BB0604	COG1620	C	L-lactate permease	Env.
PAN006350					Env.
PAN006400	YOL164w	COG2015	Q	Alkyl sulfatase and related hydrolases	Env.
PAN006404	Ecs2095_2	COG2199	T	FOG: GGDEF domain	Env.
PAN006558	STM2036_1	COG4936	TK	Predicted sensor domain	Env.
PAN006559	STM2037	COG0580	G	Glycerol uptake facilitator and related permeases (Major Intrinsic Protein Family)	Env.
PAN006560	STM2039	COG4816	E	Ethanolamine utilization protein	Env.
PAN006561	STM2040	COG4909	Q	Propanediol dehydratase, large subunit	Env.
PAN006562	mll6722	COG4909	Q	Propanediol dehydratase, large subunit	Env.
PAN006563	STM2042	COG4910	Q	Propanediol dehydratase, small subunit	Env.
PAN006564					Env.
PAN006565					Env.
PAN006566	STM2045	COG4577	QC	Carbon dioxide concentrating mechanism/carboxysome shell protein	Env.
PAN006567	STM2046	COG4577	QC	Carbon dioxide concentrating mechanism/carboxysome shell protein	Env.
PAN006568	STM2047	COG4869	Q	Propanediol utilization protein	Env.
PAN006570	STM2049	COG4576	QC	Carbon dioxide concentrating mechanism/carboxysome shell protein	Env.
PAN006571	STM2050_1	COG2096	S	Uncharacterized conserved protein	Env.
PAN006572	STM2051	COG1012	C	NAD-dependent aldehyde dehydrogenases	Env.
PAN006573	STM2052	COG1454	C	Alcohol dehydrogenase, class IV	Env.
PAN006574	STM2053	COG4656	C	Predicted NADH:ubiquinone oxidoreductase, subunit RnfC	Env.
PAN006575	STM2054	COG4577	QC	Carbon dioxide concentrating mechanism/carboxysome shell protein	Env.
PAN006576	STM2055	COG4810	E	Ethanolamine utilization protein	Env.
PAN006577	STM2056	COG4917	E	Ethanolamine utilization protein	Env.
PAN006734					Env.
PAN006868					Env.
PAN006880					Env.
PAN006881					Env.
PAN006882	rfaJ	COG1442	M	Lipopolysaccharide biosynthesis proteins, LPS:glycosyltransferases	Env.
PAN006883	rfaI	COG1442	M	Lipopolysaccharide biosynthesis proteins, LPS:glycosyltransferases	Env.
PAN006884	rfaB	COG0438	M	Glycosyltransferase	Env.
PAN006886	rfaQ	COG0859	M	ADP-heptose:LPS heptosyltransferase	Env.

Table A1 (continued)

PAN006887					Env.
PAN006927					Env.
PAN006928					Env.
PAN007850					Env.
PAN007949	PA5083	COG0251	J	Putative translation initiation inhibitor, yjgF family	Env.
PAN007950	PA5084	COG0665	E	Glycine/D-amino acid oxidases (deaminating)	Env.
PAN007951	PA5085	COG0583	K	Transcriptional regulator	Env.
PAN007952					Env.
PAN008511					Env.
PAN008512	PA1068	COG0326	O	Molecular chaperone, HSP90 family	Env.
PAN008565	STM4449	COG3077	L	DNA-damage-inducible protein J	Env.
PAN008634					Env.
PAN008640					Env.
PAN008651					Env.
PAN008652	STM0266	COG3515	S	Uncharacterized protein conserved in bacteria	Env.
PAN008653	STM0267	COG3520	S	Uncharacterized protein conserved in bacteria	Env.
PAN008654	STM0268	COG3519	S	Uncharacterized protein conserved in bacteria	Env.
PAN008655	STM0269	COG3518	S	Uncharacterized protein conserved in bacteria	Env.
PAN008656	STM0270	COG4455	R	Protein of avirulence locus involved in temperature-dependent protein secretion	Env.
PAN008657					Env.
PAN008660	STM0274	COG3517	S	Uncharacterized protein conserved in bacteria	Env.
PAN008663	STM0276	COG3157	S	Hemolysin-coregulated protein (uncharacterized)	Env.
PAN008716	RSc0632	COG1064	R	Zn-dependent alcohol dehydrogenases	Env.
PAN008729	Cj0599_2	COG2885	M	Outer membrane protein and related peptidoglycan-associated (lipo)proteins	Env.
PAN008730					Env.
PAN008731	VC1760	COG0553	KL	Superfamily II DNA/RNA helicases, SNF2 family	Env.
PAN008749					Env.
PAN008754					Env.
PAN008807	SMa2301	COG2199	T	FOG: GGDEF domain	Env.
PAN008808	YPPCP1.07	COG4571	M	Outer membrane protease	Env.
PAN008813					Env.
PAN008829	Z0414	COG4405	S	Uncharacterized protein conserved in bacteria	Env.
PAN008835					Env.
PAN008838					Env.
PAN008847	YPO0852	COG1874	G	Beta-galactosidase	Env.
PAN008848	YPO0849	COG1609	K	Transcriptional regulators	Env.
PAN008849	ECs0396	COG0477	GEPR	Permeases of the major facilitator superfamily	Env.
PAN008878	AGc3846	COG3757	M	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)	Env.
PAN008879	BMEII0782	COG3757	M	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)	Env.
PAN008902					Env.
PAN008922					Env.
PAN008934					Env.
PAN008966	STM1236	COG4461	S	Uncharacterized protein conserved in bacteria, putative lipoprotein	Env.
PAN008974					Env.
PAN009011					Env.
PAN009026					Env.
PAN009031	yjhH	COG0329	EM	Dihydrodipicolinate synthase/N-acetylneuraminate lyase	Env.
PAN009050	STM2396	COG2204	T	Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	Env.
PAN009053	STM2398	COG1840	P	ABC-type Fe3+ transport system, periplasmic component	Env.
PAN009054					Env.
PAN009055	STM2399	COG2271	G	Sugar phosphate permease	Env.

Table A2. List of core genes found to be horizontally transferred between clades. Genes in close proximity to repeat regions or mobile elements and plasmid genes are denoted by * and &, respectively.

Gene ID	Distance from mobile element (bp)	Annotation
PAN002200		Uncharacterized protein involved in formation of periplasmic nitrate reductase
PAN003853		F0F1-type ATP synthase, epsilon subunit (mitochondrial delta subunit)
PAN001208		Predicted membrane protein
PAN001547		Predicted periplasmic protein
PAN002200		Uncharacterized protein involved in formation of periplasmic nitrate reductase
PAN002260		Uncharacterized conserved protein
PAN002841		Uncharacterized protein conserved in bacteria
PAN002891*	33	Uncharacterized protein conserved in bacteria
PAN003250		Uncharacterized protein conserved in bacteria
PAN003274		Biotin carboxyl carrier protein
PAN003288		Factor for inversion stimulation Fis, transcriptional activator
PAN003317		DNA-directed RNA polymerase, alpha subunit/40 kD subunit
PAN003406		Shikimate kinase
PAN003878		Uncharacterized protein conserved in bacteria
PAN003889		Thiol-disulfide isomerase and thioredoxins
PAN004055*	30	Transcriptional regulator of met regulon
PAN004097&		Preprotein translocase subunit SecE
PAN004200		Predicted membrane protein
PAN001208		Predicted membrane protein
PAN001547		Predicted periplasmic protein
PAN002891*	33	Uncharacterized protein conserved in bacteria
PAN003250		Uncharacterized protein conserved in bacteria
PAN003274		Biotin carboxyl carrier protein
PAN003288		Factor for inversion stimulation Fis, transcriptional activator
PAN003317		DNA-directed RNA polymerase, alpha subunit/40 kD subunit
PAN003406		Shikimate kinase
PAN003878		Uncharacterized protein conserved in bacteria
PAN004055*	30	Transcriptional regulator of met regulon
PAN004097&		Preprotein translocase subunit SecE
PAN000050*	17	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases
PAN000490*	13	Predicted thioesterase
PAN000542		ABC-type uncharacterized transport system, ATPase component
PAN000550		Predicted ATPase
PAN000563		Malate/L-lactate dehydrogenases
PAN000599*	29	Phosphopantetheinyl transferase component of siderophore synthetase
PAN000607		ABC-type Fe3+-siderophore transport system, permease component
PAN000612		Isochorismate hydrolase
PAN000656		Rare lipoprotein B
PAN000667		Predicted metal-dependent hydrolase
PAN000669		2-methylthioadenine synthetase
PAN000690*	5	Flavodoxins
PAN000731		Biopolymer transport proteins
PAN000747*	19	ABC-type molybdenum transport system, ATPase component/photorepair protein PhrA
PAN000755		3-carboxymuconate cyclase
PAN000886		Permeases of the major facilitator superfamily
PAN000931		3-hydroxymyristoyl/3-hydroxydecanoyl-(acyl carrier protein) dehydratases
PAN000949		Ni,Fe-hydrogenase I small subunit
PAN000968*	9	Uncharacterized component of anaerobic dehydrogenases
PAN000988*	211	High-affinity Fe2+/Pb2+ permease
PAN001000*	34	-
PAN001010		Predicted phosphatase homologous to the C-terminal domain of histone macroH2A1
PAN001022		Cytochrome B561
PAN001075		Transcriptional regulator
PAN001089		ABC-type spermidine/putrescine transport system, permease component I
PAN001237		Uncharacterized protein involved in cation transport

Table A2 (continued)

PAN001246		Nitrate reductase delta subunit
PAN001362		FOG: GGDEF domain
PAN001415		Sortase and related acyltransferases
PAN001438		Anaerobic dehydrogenases, typically selenocysteine-containing
PAN001473		NAD-dependent aldehyde dehydrogenases
PAN001509		Predicted permease
PAN001531		Adenosine deaminase
PAN001542		Predicted EndoIII-related endonuclease
PAN001592		-
PAN001595		Predicted permease
PAN001626		Threonyl-tRNA synthetase
PAN001634		Predicted phosphatase/phosphohexomutase
PAN001640		Alpha-galactosidases/6-phospho-beta-glucosidases, family 4 of glycosyl hydrolases
PAN001646*	46	NAD synthase
PAN001659		Uncharacterized conserved protein
PAN001754		Dihydroxyacid dehydratase/phosphogluconate dehydratase
PAN001784		FOG: CheY-like receiver
PAN001792		-
PAN001819		-
PAN001842		DNA-binding HTH domain-containing proteins
PAN002008		DNA gyrase inhibitor
PAN002010		Exonuclease I
PAN002061		Uridine kinase
PAN002130*	6	Uncharacterized membrane-associated protein
PAN002142*	60	ABC-type glucose/galactose transport system, permease component
PAN002148		GTP cyclohydrolase I
PAN002161		Sugar kinases, ribokinase family
PAN002174		ABC-type uncharacterized transport system, permease component
PAN002179		16S rRNA uridine-516 pseudouridylate synthase and related pseudouridylate synthases
PAN002193		ABC-type transport system involved in cytochrome c biogenesis, permease component
PAN002198		Ferredoxin
PAN002199&		Anaerobic dehydrogenases, typically selenocysteine-containing
PAN002201		Ferredoxin
PAN002207		Alkylated DNA repair protein
PAN002208		Adenosine deaminase
PAN002210		Outer membrane protein (porin)
PAN002246		Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis
PAN002271		NADH:ubiquinone oxidoreductase subunit 1 (chain H)
PAN002285		Acetate kinase
PAN002298		ABC-type amino acid transport/signal transduction systems, periplasmic component/domain
PAN002328		Phosphohistidine phosphatase SixA
PAN002356		Response regulator of the LytR/AlgR family
PAN002399		-
PAN002436*	113	Uncharacterized lipoprotein
PAN002445		Uracil phosphoribosyltransferase
PAN002461		FOG: WD40-like repeat
PAN002464		Enzyme involved in the deoxyxylulose pathway of isoprenoid biosynthesis
PAN002476		DnaJ-domain-containing proteins 1
PAN002499*	16	Nitrogen regulatory protein PII
PAN002557		Putative Mg2+ and Co2+ transporter CorB
PAN002561		Small protein A (tmRNA-binding)
PAN002583		Protein involved in ribonucleotide reduction
PAN002588		ABC-type proline/glycine betaine transport system, permease component
PAN002615*	22	Uncharacterized NAD(FAD)-dependent dehydrogenases
PAN002629		Formate hydrogenlyase subunit 4
PAN002670		Predicted acid phosphatase
PAN002672		2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
PAN002711*	75	Signal transduction histidine kinase
PAN002779		Prolipoprotein diacylglyceryltransferase
PAN002787		Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II
PAN002877		Endonuclease I
PAN002880*	466	Putative transcriptional regulator
PAN002881*	50	Predicted endonuclease involved in recombination (possible Holliday junction resolvase in Mycoplasmas and B. subtilis)
PAN002886		Xanthosine triphosphate pyrophosphatase
PAN003014*	274	Zn finger protein HypA/HybF (possibly regulating hydrogenase expression)

Table A2 (continued)

PAN003016		Ni,Fe-hydrogenase maturation factor
PAN003097*	32	Predicted transcriptional regulators
PAN003135*	14	Predicted membrane protein
PAN003160		Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IIC
PAN003179		Predicted endonuclease containing a URI domain
PAN003215*	304	ABC-type transport system involved in resistance to organic solvents, auxiliary component
PAN003224		Uncharacterized protein conserved in bacteria
PAN003245		Glutathione S-transferase
PAN003249		Predicted ATPase
PAN003252		Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain
PAN003270		FOG: EAL domain
PAN003291		Transcriptional regulator
PAN003322		Preprotein translocase subunit SecY
PAN003388		NAD(P)H-nitrite reductase
PAN003402		Predicted phosphatases
PAN003417		Disulfide bond chaperones of the HSP33 family
PAN003430*	23	Thioredoxin-like proteins and domains
PAN003457		Aspartate-semialdehyde dehydrogenase
PAN003471		-
PAN003488		N6-adenine-specific methylase
PAN003497		-
PAN003501		ABC-type dipeptide transport system, periplasmic component
PAN003502		ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
PAN003507*	9	ABC-type multidrug transport system, permease component
PAN003519		Protein involved in catabolism of external DNA
PAN003539		Uncharacterized conserved protein
PAN003618*	228	Glutathione S-transferase
PAN003670		Guanylate kinase
PAN003811		Molecular chaperone (small heat shock protein)
PAN003835		Phosphopantetheinyl transferase
PAN003849*	45	ABC-type phosphate transport system, permease component
PAN003854		F0F1-type ATP synthase, beta subunit
PAN003926		Uncharacterized protein, possibly involved in aromatic compounds catabolism
PAN003951		Dienelactone hydrolase and related enzymes
PAN003952		Uridine phosphorylase
PAN003975		Flavodoxin
PAN003976*	7	Molybdopterin-guanine dinucleotide biosynthesis protein
PAN003990*	132	Signal transduction histidine kinase, nitrogen specific
PAN003991*	70	Glutamine synthetase
PAN004005		Predicted membrane protein
PAN004047		1,4-dihydroxy-2-naphthoate octaprenyltransferase
PAN004049		ATP-dependent protease HslVU (ClpYQ), peptidase subunit
PAN004117		-
PAN004142		16S rRNA uridine-516 pseudouridylate synthase and related pseudouridylate synthases
PAN004165		tRNA-dihydrouridine synthase
PAN004204		-
PAN004251*	136	Signal transduction histidine kinase
PAN004252		Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
PAN004280		Protein affecting phage T7 exclusion by the F plasmid
PAN004296		Succinate dehydrogenase/fumarate reductase, flavoprotein subunit
PAN004377		Predicted Zn-dependent proteases and their inactivated homologs
PAN004388		Aspartate carbamoyltransferase, catalytic chain
PAN004399*	53	DNA polymerase III, chi subunit
PAN004401		Predicted permeases
PAN004587*	20	16S RNA G1207 methylase RsmC
PAN004589		Acetyltransferases
PAN004610*	11	ATPase components of ABC transporters with duplicated ATPase domains
PAN000542		ABC-type uncharacterized transport system, ATPase component
PAN000563		Maltate/L-lactate dehydrogenases
PAN000599*	29	Phosphopantetheinyl transferase component of siderophore synthetase
PAN000656		Rare lipoprotein B
PAN000669		2-methylthioadenine synthetase
PAN000672*	52	Predicted sugar phosphatases of the HAD superfamily
PAN000731		Biopolymer transport proteins
PAN000737		Quinolinate synthase

Table A2 (continued)

PAN000813		Pyruvate-formate lyase-activating enzyme
PAN000968*	9	Uncharacterized component of anaerobic dehydrogenases
PAN000985*	8	Delta 1-pyrroline-5-carboxylate dehydrogenase
PAN001246		Nitrate reductase delta subunit
PAN001312		Enoyl-(acyl-carrier-protein) reductase (NADH)
PAN001322*	249	ABC-type oligopeptide transport system, periplasmic component
PAN001438		Anaerobic dehydrogenases, typically selenocysteine-containing
PAN001444*	483	Zn-dependent alcohol dehydrogenases
PAN001595		Predicted permease
PAN001773		SAM-dependent methyltransferases
PAN002193		ABC-type transport system involved in cytochrome c biogenesis, permease component
PAN002197		Polyferredoxin
PAN002198		Ferredoxin
PAN002199&		Anaerobic dehydrogenases, typically selenocysteine-containing
PAN002201		Ferredoxin
PAN002207		Alkylated DNA repair protein
PAN002208		Adenosine deaminase
PAN002210		Outer membrane protein (porin)
PAN002219		Uncharacterized protein conserved in bacteria
PAN002285		Acetate kinase
PAN002297		ABC-type arginine transport system, permease component
PAN002328		Phosphohistidine phosphatase SixA
PAN002399		-
PAN002557		Putative Mg2+ and Co2+ transporter CorB
PAN002611*	458	Transcriptional regulators of sugar metabolism
PAN002615*	22	Uncharacterized NAD(FAD)-dependent dehydrogenases
PAN002617*	87	Fe-S-cluster-containing hydrogenase components 2
PAN002624*	72	Ni,Fe-hydrogenase maturation factor
PAN002626		Ni,Fe-hydrogenase III small subunit
PAN002629		Formate hydrogenlyase subunit 4
PAN002836		Glycine cleavage system T protein (aminomethyltransferase)
PAN002854		Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase
PAN002880*	466	Putative transcriptional regulator
PAN002881*	50	Predicted endonuclease involved in recombination (possible Holliday junction resolvase in Mycoplasmas and B. subtilis)
PAN003105		Amino acid transporters
PAN003178		Uncharacterized protein conserved in bacteria
PAN003402		Predicted phosphatases
PAN003440		Rhodanese-related sulfurtransferase
PAN003471		-
PAN003501		ABC-type dipeptide transport system, periplasmic component
PAN003502		ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
PAN003503		ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
PAN003507*	9	ABC-type multidrug transport system, permease component
PAN003572		ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
PAN003853		F0F1-type ATP synthase, epsilon subunit (mitochondrial delta subunit)
PAN003951		Dienelactone hydrolase and related enzymes
PAN003952		Uridine phosphorylase
PAN003975		Flavodoxin
PAN003976*	7	Molybdopterin-guanine dinucleotide biosynthesis protein
PAN003991*	70	Glutamine synthetase
PAN004006		D-Tyr-tRNAtyr deacylase
PAN004108		Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2
PAN004147		-
PAN004357		Permeases of the drug/metabolite transporter (DMT) superfamily
PAN004386*	59	Putative translation initiation inhibitor, yjgF family
PAN004387		Aspartate carbamoyltransferase, regulatory subunit
PAN004587*	20	16S RNA G1207 methylase RsmC

Table A3. List of non-core genes found to be horizontally transferred between clades. Genes in close proximity to repeat regions or mobile elements and plasmid genes are denoted by * and &, respectively.

Lineages	Gene ID	Distance from mobile element (bp)	Annotation
C-I, C-IV	PAN004282		Co-chaperonin GroES (HSP10)
C-I, C-V	PAN001225		Phosphoribosylpyrophosphate synthetase
	PAN000487		Bacterial nucleoid DNA-binding protein
	PAN001369		Universal stress protein UspA and related nucleotide-binding proteins
	PAN002435*	55	Phosphoribosylaminoimidazolesuccinocarboxamide (SAICAR) synthase
	PAN002788		Transcriptional regulators
	PAN003233		Uncharacterized protein involved in an early stage of isoprenoid biosynthesis
C-I, GI	PAN003308		N-formylmethionyl-tRNA deformylase
	PAN003370		Uncharacterized protein conserved in bacteria
	PAN004015		Fe-S-cluster-containing hydrogenase components 1
	PAN004032		Predicted Co/Zn/Cd cation transporters
C-III, C-IV	PAN000080		Uncharacterized protein conserved in bacteria
	PAN001620		Bacterial nucleoid DNA-binding protein
C-III, C-V	PAN001620		Bacterial nucleoid DNA-binding protein
C-III, GI	PAN000202		Uncharacterized conserved protein
	PAN000207		Histidinol phosphatase and related phosphatases
	PAN000569		Phosphoribosylcarboxyaminoimidazole (NCAIR) mutase
	PAN000717		Succinate dehydrogenase, hydrophobic anchor subunit
	PAN000888		Pyruvate-formate lyase
	PAN001620		Bacterial nucleoid DNA-binding protein
	PAN001644		Phosphotransferase system cellobiose-specific component IIB
	PAN002593		Transcriptional regulators
C-IV, C-V	PAN003623		Transcriptional regulator
	PAN003857		F0F1-type ATP synthase, delta subunit (mitochondrial oligomycin sensitivity protein)
	PAN004282		Co-chaperonin GroES (HSP10)
	PAN000207		Histidinol phosphatase and related phosphatases
	PAN001644		Phosphotransferase system cellobiose-specific component IIB
	PAN002593		Transcriptional regulators
C-IV, GI	PAN003861		F0F1-type ATP synthase, subunit I
	PAN003865		Transcriptional regulators
	PAN000136*	274	Aspartate 1-decarboxylase
C-V, GI	PAN000666		Putative Mg2+ and Co2+ transporter CorC
	PAN001644		Phosphotransferase system cellobiose-specific component IIB

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

Table B1. Differentially abundant SEED subsystems between warming and control metagenomes.

SEED subsystem	Mean number of reads	Enriched group	Log2 fold change	Fold change	P-value (B-H adjusted)
Malonate_decarboxylase	844	H	0.288	22%	1.43E-72
Ribosome_LSU_eukaryotic_and_archaeal	459	H	0.286	22%	3.56E-35
Polysaccharide_deacetylases	367	H	0.285	22%	2.34E-53
Phage_packaging_machinery	1644	H	0.281	21%	2.76E-86
Cholesterol_catabolic_operon_in_Mycobacteria	460	H	0.270	21%	4.05E-31
Phage_tail_fiber_proteins	1429	H	0.263	20%	3.20E-119
NAD_consumption	180	H	0.249	19%	3.20E-119
Coenzyme_PQQ_synthesis	3704	H	0.241	18%	1.49E-52
Tryptophan_catabolism	218	H	0.240	18%	1.60E-22
Phosphorylcholine_incorporation_in_LPS	111	H	0.237	18%	4.67E-85
DNA_replication,_archaeal	1928	H	0.230	17%	7.08E-93
Terminal_AA3-600_quinol_oxidase	291	H	0.227	17%	7.17E-27
CBSS-316273.3.peg.227	2599	H	0.225	17%	1.49E-87
Pyrroloquinoline_Quinone_biosynthesis	3799	H	0.220	16%	5.59E-51
Threonine_anaerobic_catabolism_gene_cluster	196	H	0.212	16%	3.13E-59
Sporulation_Cluster_III_A	118	H	0.225	17%	2.69E-02
Spore_germination	307	H	0.157	12%	4.00E-14
SeqA_and_Co-occurring_Genes	552	H	0.197	15%	2.86E-80
Glycerate_metabolism	34650	H	0.177	13%	2.76E-05
CBSS-320388.3.peg.3759	4715	H	0.191	14%	2.13E-24
Denitrification	5164	H	0.163	12%	6.68E-11
Carbon_monoxide_induced_hydrogenase	422	H	0.189	14%	3.73E-97
Biflavanoid_biosynthesis	1613	H	0.183	14%	5.13E-105
Tannin_biosynthesis	1610	H	0.180	13%	3.70E-100
Cobalt-zinc-cadmium_resistance	121204	H	0.176	13%	3.71E-90
CRISP_Cmr_Cluster	420	H	0.174	13%	8.88E-36
Polymyxin_Synthetase_Gene_Cluster_in_Bacillus	726	H	0.172	13%	1.75E-29
beta-glucuronide_utilization	143	H	0.287	22%	1.35E-43
Cellulosome	1354	H	0.171	13%	6.47E-16
CBSS-224911.1.peg.435	3894	H	0.167	12%	5.50E-61
Phosphonoalanine_utilization	176	H	0.165	12%	4.05E-31
ESAT-6_proteins_secretion_system_in_Actinobacteria	648	H	0.160	12%	6.47E-16
Streptococcus_pyogenes_Virulome	331	H	0.159	12%	6.00E-29
Terminal_cytochrome_oxidases	18448	H	0.156	11%	9.20E-52
Bacterial_Caspases	205	H	0.153	11%	3.63E-54
Lipopolysaccharide-related_cluster_in_Alphaproteobacteria	3328	H	0.151	11%	1.34E-65
Bacillibactin_Siderophore	796	H	0.148	11%	1.60E-22
Glycine_reductase,_sarcosine_reductase_and_betaine_reductase	1093	H	0.146	11%	3.20E-119
Sporulation_draft	167	H	0.145	11%	3.20E-119
Photosystem_II	267	H	0.144	11%	2.87E-38

Table B1 (continued)

Multidrug_efflux_pump_in_Campylobacter_jejuni_(CmeABC_operon)	261	H	0.142	10%	1.27E-98
Lipid_A-Ara4N_pathway_(_Polymyxin_resistance_)	8485	H	0.141	10%	3.73E-97
Marinocine,_a_broad-spectrum_antibacterial_protein	598	H	0.140	10%	1.40E-34
Dot-Icm_type_IV_secretion_system	258	H	0.136	10%	1.75E-05
Tricarballylate_Utilization	1572	H	0.136	10%	1.23E-116
Translation_elongation_factors_eukaryotic_and_archaeal	772	H	0.134	10%	1.94E-17
Phosphoglycerate_transport_system	272	H	0.131	10%	1.40E-34
p-cymene_degradation	362	H	0.127	9%	8.37E-63
Pentose_phosphate_pathway	127971	H	0.126	9%	1.18E-64
Alpha-acetolactate_operon	378	H	0.125	9%	5.76E-07
Trans-envelope_signaling_system_VreARI_in_Pseudomonas	1053	H	0.125	9%	1.02E-45
P38_MAP_kinase_pathways	103	H	0.123	9%	2.72E-44
Predicted_mycobacterial_monooxygenase	257	H	0.122	9%	5.50E-34
Archease	202	H	0.122	9%	2.12E-33
Chlorophyll_Biosynthesis	3745	H	0.121	9%	7.88E-18
Cytochrome_B6-F_complex	196	H	0.120	9%	1.47E-39
Ferrous_iron_transporter_EfeUOB,_low-pH-induced	1363	H	0.119	9%	1.92E-13
Alkanesulfonate_assimilation	28741	H	0.119	9%	8.37E-63
Methionine_Salvage	4843	H	0.117	8%	1.27E-82
Benzoate_degradation	3290	H	0.116	8%	1.47E-39
cAMP_signaling_in_bacteria	93367	H	0.116	8%	1.07E-74
Amidase_clustered_with_urea_and_nitrile_hydratase_functions	1149	H	0.115	8%	2.74E-16
Nitric_oxide_synthase	3089	H	0.125	9%	1.26E-02
rRNA_modification_Archaea	264	H	0.114	8%	7.02E-06
Selenocysteine_metabolism	23549	H	0.113	8%	2.28E-58
Menaquinone_Biosynthesis_via_Futalosine_--_gjo	19720	H	0.112	8%	4.67E-85
Peripheral_Glucose_Catabolism_Pathways	283	H	0.111	8%	2.04E-28
Omega_peptidases_(EC_3.4.19.-)	7698	H	0.110	8%	3.63E-54
Hypothetical_Related_to_Dihydroorotate_dehydrogenase	1039	H	0.109	8%	1.27E-21
Alkanesulfonates_Utilization	5962	H	0.107	8%	1.40E-34
Proton-dependent_Peptide_Transporters	5187	H	0.105	8%	4.05E-31
Clavulanic_acid_biosynthesis	221	H	0.105	8%	3.35E-11
CBSS-288000.5.peg.1793	1242	H	0.104	7%	4.72E-17
Putative_sulfate_assimilation_cluster	1257	H	0.104	7%	3.13E-59
Calvin-Benson_cycle	22126	H	0.101	7%	6.35E-48
Teichoic_and_lipoteichoic_acids_biosynthesis	4190	H	0.100	7%	3.20E-119
Unknown_carbohydrate_utilization_(_cluster_Ydj_)	116	H	0.100	7%	8.79E-13
Sporulation-related_Hypotheticals	1169	H	0.100	7%	3.20E-119
A_Gammaproteobacteria_Cluster_Relating_to_Translation	900	H	0.099	7%	3.20E-119
Polyunsaturated_Fatty_Acids_synthesis	3024	H	0.098	7%	7.17E-27
Pterin_metabolism	489	H	0.098	7%	7.88E-18

Table B1 (continued)

Serotype_determining_Capsular_polysaccharide_biosynthesis_in_Staphylococcus	142	H	0.096	7%	6.47E-16
LOS_core_oligosaccharide_biosynthesis	24302	H	0.095	7%	2.76E-86
Formate_dehydrogenase	6569	H	0.095	7%	1.07E-74
Methylglyoxal_Metabolism	30581	H	0.122	9%	3.50E-15
Cobalamin_synthesis	17924	H	0.094	7%	2.12E-33
Mediator_of_hyperadherence_YidE_in_Enterobacteria_and_its_conserved_region	153	H	0.094	7%	2.27E-25
Vibrioferrin_synthesis	138	H	0.094	7%	2.16E-36
CO_Dehydrogenase	32957	H	0.117	8%	2.16E-02
Biofilm_Adhesin_Biosynthesis	463	H	0.093	7%	1.64E-57
Carbon_monoxide_dehydrogenase_maturation_factors	10192	H	0.135	10%	1.14E-07
Trehalose_Uptake_and_Utilization	11513	H	0.105	8%	4.80E-04
Dipeptidases_(EC_3.4.13.-)	1416	H	0.092	7%	1.28E-73
CBSS-222523.1.peg.1311	106	H	0.090	6%	5.07E-30
At3g21300	28131	H	0.089	6%	3.70E-100
Molybdopterin_cytosine_dinucleotide	20027	H	0.088	6%	1.22E-37
Zinc_resistance	10081	H	0.088	6%	7.30E-26
Dehydrogenase_complexes	96393	H	0.145	11%	1.34E-65
Soluble_methane_monooxygenase_(sMMO)	1090	H	0.086	6%	3.51E-47
CBSS-243277.1.peg.4359	248	H	0.086	6%	1.75E-29
CBSS-251221.1.peg.1863	2579	H	0.085	6%	1.27E-98
CBSS-52598.3.peg.2843	4008	H	0.084	6%	7.88E-18
Utilization_of_glutathione_as_a_sulphur_source	12063	H	0.084	6%	6.34E-24
Iojap	112063	H	0.083	6%	2.86E-80
USS-DB-4	110	H	0.083	6%	2.24E-04
Archaeal_Flagellum	195	H	0.083	6%	3.28E-40
PA0057_cluster	303	H	0.082	6%	2.07E-08
Indole-pyruvate_oxidoreductase_complex	621	H	0.082	6%	4.16E-13
Group_II_intron-associated_genes	6161	H	0.082	6%	1.14E-07
Alkaloid_biosynthesis_from_L-lysine	2909	H	0.081	6%	9.58E-21
Formate_hydrogenase	60735	H	0.081	6%	1.02E-63
Photorespiration_(oxidative_C2_cycle)	40992	H	0.119	9%	5.76E-07
Multidrug_Resistance_Efflux_Pumps	57620	H	0.080	6%	1.93E-114
Oxygen_and_light_sensor_PpaA-PpsR	414	H	0.080	6%	1.75E-108
mycolic_acid_synthesis	736	H	0.076	5%	4.26E-05
CBSS-316273.3.peg.922	238	H	0.075	5%	2.65E-06
Spore_Core_Dehydration	1158	H	0.075	5%	1.61E-70
CBSS-198094.1.peg.4426	1132	H	0.074	5%	7.88E-18
CBSS-269801.1.peg.1725	1075	H	0.074	5%	1.61E-70
N-heterocyclic_aromatic_compound_degradation	30410	H	0.074	5%	6.52E-39
Allantoin_Utilization	4601	H	0.072	5%	1.18E-64
YgjD_and_YeaZ	8263	H	0.072	5%	5.50E-61
CBSS-314267.3.peg.390	30742	H	0.072	5%	5.17E-37
Siderophore_Pyoverdine	10139	H	0.071	5%	6.68E-11
Phenylpropionate_Degradation	262	C	0.241	18%	9.50E-107
Hyperosmotic_potassium_uptake	953	C	0.655	58%	2.41E-112
SpoVS_protein_family	191	C	0.489	40%	1.02E-101
Spliceosome	315	C	0.398	32%	2.28E-58

Table B1 (continued)

Methanogenesis_from_methylated_compounds	107	C	0.378	30%	9.50E-107
Sucrose_utilization_Shewanella	200	C	0.348	27%	1.61E-70
Xyloglucan_Utilization	200	C	0.333	26%	1.08E-95
CBSS-49338.1.peg.459	17499	C	0.322	25%	2.83E-94
Citrate_Utilization_System_(CitAB,_CitH,_and_tctABC)	1402	C	0.306	24%	1.27E-82
Energy-conserving_hydrogenase_(ferredoxin)	200	C	0.282	22%	1.68E-91
Ketoisovalerate_oxidoreductase	529	C	0.276	21%	1.02E-101
RNA_polymerase_III	315	C	0.264	20%	5.30E-45
Iron_acquisition_in_Vibrio	158	C	0.261	20%	1.75E-108
P_uptake_(cyanobacteria)	1785	C	0.258	20%	7.88E-18
Tocopherol_Biosynthesis	438	C	0.249	19%	7.21E-41
RNA_polymerase_II_initiation_factors	588	C	0.246	19%	3.13E-59
L-ascorbate_utilization_(and_related_gene_clusters)	328	C	0.246	19%	3.20E-119
V-Type_ATP_synthase	1161	C	0.242	18%	2.28E-58
Aminoglycoside_adenylyltransferases	136	C	0.235	18%	8.08E-56
Spore_pigment_biosynthetic_cluster_in_Actinomycetes	1289	C	0.235	18%	1.49E-66
Cadmium_resistance	449	C	0.234	18%	2.86E-80
CBSS-316273.3.peg.448	3313	C	0.228	17%	1.40E-34
RNA_polymerase_I	260	C	0.218	16%	7.01E-25
Melibiose_Utilization	410	C	0.214	16%	7.08E-93
Protein_secretion_by_ABC-type_exporters	609	C	0.211	16%	5.51E-55
Proteasome_eukaryotic	1305	C	0.210	16%	1.40E-34
VC0266	317	C	0.197	15%	2.48E-103
Mannitol_Utilization	4636	C	0.195	15%	5.13E-105
Periplasmic-Binding-Protein-Dependent_Transport_System_for_945;-Glucosides	5574	C	0.195	14%	7.83E-89
CBSS-393131.3.peg.612	196	C	0.191	14%	1.43E-72
Fatty_Acid_Biosynthesis_FASI	3567	C	0.189	14%	5.57E-23
Ribosome_LSU_mitochondrial	336	C	0.189	14%	1.55E-41
CBSS-176279.3.peg.1262	150	C	0.189	14%	1.55E-41
Dissimilatory_nitrite_reductase	607	C	0.187	14%	4.02E-79
Flavodoxin	1033	C	0.185	14%	1.56E-71
HtrA_and_Sec_secretion	325	C	0.185	14%	1.27E-21
ECF_class_transporters	5913	C	0.177	13%	1.93E-114
Sucrose_utilization	1993	C	0.174	13%	3.20E-119
Iron_acquisition_in_Streptococcus	8142	C	0.172	13%	7.17E-77
WhiB_and_WhiB-type_regulatory_proteins_	946	C	0.172	13%	2.31E-26
Inositol_catabolism	25008	C	0.171	13%	1.93E-114
O-antigen_capsule_important_for_environmental_persistence	112	C	0.171	13%	3.28E-40
CBSS-344610.3.peg.2335	337	C	0.168	12%	1.93E-114
tRNA-dependent_amino_acid_transfers	256	C	0.167	12%	5.59E-51
Na(+)-translocating_NADH-quinone_oxidoreductase and rnf-like_electron_transport_complexes	1433	C	0.164	12%	2.16E-36

Table B1 (continued)

CBSS-288681.3.peg.1039	118	C	0.163	12%	1.55E-41
Sulfate_reduction-associated_complexes	342	C	0.163	12%	1.59E-67
Cresol_degradation	126	C	0.162	12%	2.22E-27
RNA_polymerase_II	350	C	0.160	12%	1.14E-16
CBSS-261594.1.peg.2640	114	C	0.158	12%	1.34E-65
Erythritol_utilization	1366	C	0.156	11%	2.28E-58
L-fucose_utilization	679	C	0.154	11%	1.60E-22
Choline_Transport	124	C	0.154	11%	6.67E-43
USS-DB-1	918	C	0.150	11%	1.35E-43
Plastoquinone_Biosynthesis	667	C	0.149	11%	2.72E-44
Glutathione-regulated_potassium-efflux_system_and_associated_functions	6615	C	0.149	11%	1.61E-70
Lysine_biosynthesis_AAA_pathway_2	974	C	0.148	11%	8.00E-15
Control_of_Swarming_in_Vibrio_and_Shewanella_species	100	C	0.146	11%	5.50E-61
CBSS-216592.1.peg.3534	1316	C	0.146	11%	1.61E-70
Anaerobic_Oxidative_Degradation_of_L-Ornithine	143	C	0.141	10%	2.34E-53
Fermentations:_Lactate	45686	C	0.141	9%	1.02E-63
Bacitracin_Stress_Response	346	C	0.140	10%	1.64E-57
D-allose_utilization	224	C	0.140	10%	1.34E-65
Pseudaminic_Acid_Biosynthesis	170	C	0.139	10%	2.22E-27
Ribosome_SSU_mitochondrial	215	C	0.139	10%	1.94E-17
Siderophore_Aerobactin	526	C	0.139	10%	1.35E-43
Fermentations:_Mixed_acid	24974	C	0.165	11%	8.89E-76
CBSS-83332.1.peg.3803	290	C	0.136	10%	1.00E-03
2-methylcitrate_to_2-methylaconitate_metabolism_cluster	285	C	0.136	10%	5.07E-30
Steroid_sulfates	300	C	0.134	10%	1.79E-09
Inteins	6636	C	0.134	10%	1.07E-74
Outer_membrane	2921	C	0.132	10%	1.86E-12
Phage_shock_protein_(psp)_operon	919	C	0.130	9%	3.28E-40
Resistance_to_Vancomycin	2425	C	0.128	9%	7.30E-26
963;-Fimbriae	491	C	0.126	9%	1.96E-81
O-Methyl_Phosphoramidate_Capsule_Modification_in_Campylobacter	147	C	0.125	9%	6.76E-28
Chitin_and_N-acetylglucosamine_utilization	33163	C	0.144	9%	6.52E-39
Copper_homeostasis	52831	C	0.122	9%	3.20E-119
CBSS-261594.1.peg.788	360	C	0.121	9%	1.49E-87
Cell_envelope-associated_LytR-CpsA-Psr_transcriptional_attenuators	1548	C	0.121	9%	1.35E-43
CBSS-196620.1.peg.2477	210	C	0.121	9%	8.00E-15
Salmochelin-mediated_Iron_Acquisition	134	C	0.121	9%	7.30E-26
CBSS-83333.1.peg.946	207	C	0.119	9%	2.83E-94
CBSS-235.1.peg.567	618	C	0.116	8%	1.75E-29
CBSS-159087.4.peg.2189	457	C	0.115	8%	1.86E-12
Conjugative_transposon,_Bacteroidales	542	C	0.114	8%	5.76E-07
Pseudomonas_quinolone_signal_PQS	135	C	0.113	8%	1.40E-34
p-Aminobenzoyl-Glutamate_Utilization	136	C	0.112	8%	4.21E-60
Pterin_carbinolamine_dehydratase	13266	C	0.111	8%	6.34E-24
Triacylglycerol_metabolism	5053	C	0.111	8%	7.88E-84
Transport_of_Iron	1790	C	0.110	8%	6.52E-39

Table B1 (continued)

Dimethylsulfoniopropionate_(DMSP)_mineralization	201	C	0.108	8%	5.30E-45
CBSS-280355.3.peg.2835	258	C	0.105	8%	4.99E-10
L-rhamnose_utilization	39375	C	0.105	8%	3.20E-119
pH_adaptation_potassium_efflux_system	161	C	0.105	8%	4.00E-14
Glutathione_analogs:_mycothiol	5338	C	0.105	8%	4.72E-17
Glutathionylspermidine_and_Trypanothione	160	C	0.104	7%	5.17E-37
Alpha-Amylase_locus_in_Streptococcoccus	2964	C	0.103	7%	3.71E-90
Streptococcus_pyogenes_recombinatorial_zone	151	C	0.102	7%	9.20E-52
Unspecified_monosaccharide_transport_cluster	1417	C	0.101	7%	9.50E-107
Ectoine_biosynthesis_and_regulation	533	C	0.100	7%	1.16E-56
D-galactonate_catabolism	2198	C	0.100	7%	7.88E-84
RNA_polymerase_III_initiation_factors	128	C	0.099	7%	4.99E-10
p-Hydroxybenzoate_degradation	4482	C	0.099	7%	1.65E-68
Inner_membrane_protein_YhjD_and_conserved_cluster_involved_in_LPS_biosynthesis	131	C	0.098	7%	1.47E-39
Siderophore_Desferrioxamine_E	161	C	0.098	7%	1.02E-45
ABC_transporter_alkylphosphonate_(TC_3.A.1.9.1)	6056	C	0.096	7%	9.50E-107
CBSS-188.1.peg.6170	6866	C	0.094	7%	7.21E-41
Na(+)_H(+)_antiporter	5567	C	0.093	7%	8.89E-76
MLST	10256	C	0.092	7%	1.14E-16
Na+_translocating_decarboxylases_and_related_biotin-dependent_enzymes	1504	C	0.092	7%	3.35E-11
Nitrate_and_nitrite_ammonification	40663	C	0.091	6%	1.56E-71
CBSS-376686.6.peg.291	232	C	0.089	6%	2.58E-20
Biogenesis_of_cbb3-type_cytochrome_c_oxidases	3175	C	0.088	6%	1.56E-71
Sugar_utilization_in_Thermotogales	9550	C	0.084	6%	3.56E-35
Choline_and_Betaine_Uptake_and_Betaine_Biosynthesis	26768	C	0.084	6%	1.16E-56
CBSS-342610.3.peg.1536	8422	C	0.082	6%	9.50E-107
Lipopolysaccharide_assembly	8779	C	0.081	6%	1.92E-13
CBSS-196164.1.peg.1690	9342	C	0.080	6%	7.17E-27
MazEF_toxin-antitoxing_(programmed_cell_death)	896	C	0.080	6%	8.79E-13
L-Arabinose_utilization	20485	C	0.080	6%	3.73E-97
LMPTP_YfkJ_cluster	5081	C	0.080	6%	4.05E-31
Proteorhodopsin	250	C	0.077	5%	1.86E-12
Isobutyryl-CoA_to_Propionyl-CoA_Module	635	C	0.077	5%	6.68E-11
CBSS-290633.1.peg.1906	4458	C	0.076	5%	1.02E-45
CBSS-228410.1.peg.134	7588	C	0.076	5%	9.50E-107
CBSS-342610.3.peg.1794	1096	C	0.076	5%	4.05E-31
ESAT-6_proteins_secretion_system_in_Firmicutes	127	C	0.076	5%	4.82E-19
Sodium_Hydrogen_Antiporter	4210	C	0.075	5%	1.94E-17
CBSS-320372.3.peg.6046	3754	C	0.074	5%	1.49E-52
Methanogenesis	1204	C	0.073	5%	7.21E-41
Glycerol_and_Glycerol-3-phosphate_Uptake_and_Utilization	47415	C	0.073	5%	1.43E-72
Methanopterin_biosynthesis2	1425	C	0.072	5%	1.75E-29
Mannose-sensitive_hemagglutinin pilus	2516	C	0.072	5%	7.01E-25

Table B1 (continued)

BlaR1_Family_Regulatory_Sensor-transducer_Disambiguation	42282	C	0.071	5%	1.02E-63

VITA

Chengwei Luo

Chengwei was born in Jiangxi, China. He received a B.S. in Mechanic Engineering from Beijing University of Aeronautics & Astronautics, Bejing, China in 2006, and an M.S. in Condensed Matter Physics from Peking University, Beijing, China in 2009. He then joined the Bioinformatics Ph.D. program in Georgia Institute of Technology in August 2009 to pursue his doctorate. He has a broad interest in applying mathematical models and quantitative approaches in solving complex problems. In his spare time, he enjoys traveling, hiking, soccer, saxophone, reading, and writing.