PAN-GENOME ANALYSIS, VISUALIZATION AND EXPLORATION

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Wei Ding
aus Yueyang, Hunan, China

Tübingen, 2017

Wei Ding: *Pan-genome analysis, visualization and exploration*
Dissertation © 2017

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Eberhard Karls Universität Tübingen

TAG DER MÜNDLICHEN QUALIFIKATION:
29.01.2018

DEKAN:
Prof. Dr. Wolfgang Rosenstiel

BERICHTERSTATTER:
1. Dr. Richard Neher
2. Prof. Dr. Daniel Huson

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, 2017

Wei Ding

ABSTRACT

The dynamics of prokaryotic genomes are driven by the intricate interplay of different evolutionary forces such as gene duplication, gene loss and horizontal transfer. Even closely related strains can exhibit remarkable genetic diversity and substantial gene presence/absence variation. The *pan-genome*, namely the complete inventory of genes in a collection of strains, can be several times larger than the genome of any single strain.

Although several tools for pan-genome analysis have been published, there is still much room for algorithmic improvement, as well as needs for applications that better interactively visualize and explore pan-genomes. Therefore, we have developed panX, an automated computational pipeline for efficient identification of orthologous gene clusters in the pan-genome. PanX identifies homologous relationships among genes using DIAMOND and MCL and then harnesses phylogeny-based post-processing to separate orthologs from paralogs. Furthermore, we take advantage of a divide-and-conquer strategy to achieve an approximately linear runtime on large datasets.

The analysis result can be visualized by the accompanying software, an easy-to-use and powerful web-based visualization application for interactive exploration of the pan-genome. The visualization dashboard encompasses a variety of connected components that allow rapid searching, filtering and sorting of genes and flexible investigation of evolutionary relationships among strains and their genes. PanX seamlessly interlinks gene clusters with their alignments and gene phylogenies, maps mutations on the branches of gene tree and highlights gene gain and loss events on the core-genome phylogeny that can also be colored by metadata associated with strains. By using 120 simulated pan-genome datasets for benchmarking and comparing clustering results on real dataset between different tools, panX exhibits overall good performance across a large range of diversities.

PanX is available at pangenome.de, with a wide range of microbial pan-genomes established. Besides, user-provided pan-genomes can be visualized either via a web server or by running panX locally as a web-based application.

ZUSAMMENFASSUNG

Die Dynamik von prokaryotischen Genomen wird von einem komplizierten Zusammenspiel verschiedener evolutionärer Kräfte wie Genduplikation, Genverlust und horizontales Gentransfers bestimmt. Selbst nahe verwandte Bakterienstämme können eine beachtliche genetische Vielfalt und eine erhebliche Variation von der Anwesenheit/Abwesenheit bestimmter Gene aufweisen. Das Pan-Genom, nämlich der komplette Bestand von Genen in einer Sammlung von Bakterienstämmen, kann um ein Vielfaches größer sein als das Genom eines einzelnen Stammes.

Obwohl bisher bereits mehrere Tools für Pan-Genomanalyse veröffentlicht wurden, gibt es noch Spielräume für algorithmische Verbesserungen sowie auch das Bedürfnis nach neuen Anwendungen, die Pan-Genome besser interaktiv zu visualisieren und zu erkunden. Wir haben das Programm panX entwickelt, eine automatisierte Berechnungspipeline für die effiziente Identifikation von orthologen Genclustern in einem Pan-Genom. PanX identifiziert homologe Beziehungen zwischen Genen durch DIAMOND und MCL und benutzt anschließend auf Phylogenie basierendes Post-Processing zur Trennung von Orthologen und Paralogen. Wir nutzen dazu ein Divide-and-Conquer-Prinzip um eine annähernd lineare Laufzeit bei großen Datenmengen zu erreichen.

Die Analyseergebnisse können mit Hilfe einer dazugehörigen Software visualisiert werden, einer benutzerfreundlichen und effizienten webbasierten Visualisierungsanwendung für die interaktive Erkundung von Pan-Genomen. Das Visualisierungs-Dashboard beinhaltet eine Vielzahl miteinander verknüpfter Komponenten, die ein schnelles Suchen, Filtern und Sortieren von Genen sowie die Untersuchung evolutionärer Beziehungen zwischen Stämmen und ihren Genen erlaubt. PanX verbindet nahtlos Gencluster mit ihrem Alignment und Stammbäumen für Gene. Es zeigt Mutationen auf Zweigen eines Genbaums, und hebt Gene-Gewinne und -Verluste auf dem Kern-Genom-Baum hervor, der zudem mit Metadaten von Stämmen markiert werden kann. Wir haben 120 Pan-Genom-Datensätze zum Benchmarking simuliert und die Clustering-Ergebnisse realer Datensätze verschiedener Tools verglichen. PanX erzielt insgesamt gute Ergebnisse trotz eines breiten Spektrums unterschiedlicher Ausgangsdaten.

PanX ist verfügbar unter pangenome.de mit einem umfangreichen Angebot an bakteriellen Pan-Genomen. Zudem können benutzereigene Pan-Genome entweder durch einen Webserver oder lokal als eine webbasierte Anwendung visualisiert werden.

ACKNOWLEDGMENTS

In the summer of 2012, I decided to study for a Master's degree in Tübingen, a small picturesque and tranquil town far away from the hustle and bustle where one of my favorite authors Hermann Hesse lived. As the Chinese proverb says, "Time flies like an arrow, days and months as a weaver's shuttle." Now at the epilogue of my Ph.D., I feel extremely grateful for the wonderful people I have met here.

Foremost, I would like to express my sincere appreciation to the members of my thesis advisory and exam committee, who linked the interlocking plots that lead to my Ph.D. adventure: Prof. Dr. Daniel Huson and Prof. Dr. Kay Nieselt for recommending me as a student research assistant (HiWi) to Weigel Lab, Prof. Dr. Detlef Weigel for allowing me to conduct the Hiwi and master projects and later introducing me to Richard. That's how my Ph.D. story started.

Above all, I express my deepest gratitude to my supervisor Dr. Richard Neher for his excellent guidance and invaluable encouragement. He first aroused my interest in the visualization of complex bacterial pan-genome and always promptly provided me with insightful feedback, even during nights and weekends. Without him there would be no panX! His insatiable appetite for new ideas and his clear and concise way of communicating complex information are two of his hallmarks that are deeply engraved on my mind. I feel incredibly fortunate to meet such an enthusiastic and supportive advisor and look forward to continuing the journey with him.

I thank my lab mates and colleagues in Max Planck Institute for all the help. Among them are Xi, Dagmar, Rebecca und Moi. Furthermore, I would like to say a big thank you to Franz for numerous thought-provoking discussions - I really enjoyed working with you - , and to Talia, - I highly appreciate your improvement suggestions for my software and dissertation.

Moreover, I owe inexpressible gratitude to my parents for all the love and support. Last but not least, special thanks go to Horst, Hardy and Corlia, for the time we shared when reading poetry and philosophy together, contemplating splendid alpine view and Atlantic vistas, encountering incredibly delicate plants that were beyond my wildest imagination.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Genetic information can be readily exchanged via horizontal (or lateral) gene transfer between bacteria. This process acts as a driving force for prokaryotic genome plasticity, results in complex microbial evolutionary history and continuously contributes to bacterial adaptation and speciation [8, 17]. The horizontally transferred genetic information can be acquired by three main mechanisms: transformation (by taking up naked DNA from the environment), transduction (via bacteriophages) and conjugation (through plasmids and conjugative transposons) [77]. Furthermore, gene acquisition can be mixed with substantial gene loss and duplication. The resulting genome dynamics and diversity pose enormous challenges in the reconstruction of evolutionary relationships among bacterial isolates [82]. Moreover, even closely related strains can demonstrate a complex mosaic of presence/absence patterns and a significant fraction of strain-specific genes. The entire gene repertoire in a set of strains, often denoted as *pan-genome* [97], can be several times larger than the genome of a single strain.

With advances in sequencing technologies, the rapidly increasing amount of sequence data has spurred numerous genomic studies, spanning a broad range of bacterial species. A comprehensive investigation on genome diversity and patterns of gene gain and loss in a large collection of prokaryotic strains would necessitate the construction of the pan-genome. Characterizing the dynamics of the pan-genome helps to gain deeper insights into bacterial genomic evolution and facilitates the detection of genes associated with host adaptation, niche specificity, virulence and pathogenesis [64, 75].

Identifying the pan-genome from a collection of bacterial genomes typically relies on classifying genes into orthologous clusters based upon sequence similarity searches. Several tools for building such a pan-genome are already available, which harness different heuristic methods for clustering genes [28, 61, 78, 112]. Nevertheless, there remains much room for algorithmic improvement and many challenges still need to be tackled pertaining to better interpreting, interactively visualizing and exploring pan-genomes.

In order to address these concerns, we have developed panX, an automated and comprehensive pan-genome analysis pipeline and a web-based application for interactive visualization and exploration of bacterial pan-genomes. The analysis pipeline extracts genes from a group of annotated genomes (e.g. NCBI reference sequences) and groups them into orthologous clusters. Based on these clusters, panX identifies the core genome that contains genes present in all strains, reconstructs a core genome SNP phylogeny at the strain level, generates multiple alignments of sequences in gene clusters, builds individual gene phylogenies and displays the gene presence/absence pattern and gene gain and loss event on the core genome phylogeny.

In short, the panX analysis pipeline produces not only a set of entire gene clusters but also the corresponding *phylome* [91], namely, the collection of all gene phylogenies, thereby offering a comprehensive resource for delineating evolutionary relationships among strains and their genes.

By leveraging the output produced from the pipeline, the interactive web-based visualization application provides powerful exploration techniques for the above-mentioned results and allows for flexible filter, sort, and search features. Besides, panX visualization dashboard seamlessly connects gene clusters and the associated *phylome*, integrates metadata on core genome phylogeny, which serves as a user-friendly and powerful platform for investigating pan-genomes.

The beauty and utility of this application are showcased on `http://pangenome.de` with a wide range of microbial pan-genomes. Additionally, the application can also be hosted on web servers from users or used locally as a browser-based application with their own pan-genomes.

OUTLINE

This work is divided into six chapters. Chapter 2 provides an essential background on bacterial pan-genomes, which first centers on the two main compositions of pan-genome, namely core and accessory genome, and their roles in prokaryotic evolution. Then, after quickly reviewing the notion of orthology and introducing the key procedure required for building a pan-genome, i.e. identifying orthologous gene clusters, two main orthology inference schemes are succinctly described, including graph-based and tree-based methods. Next, examples of existing pan-genome tools are given, together with the limitations of these approaches and the motivations for computational enhancements and an innovative visualization solution.

Chapter 3 addresses the details of the panX analysis pipeline I have implemented. The essence of its computational approach can be summarized as follows: using a collection of annotated genomes as input, panX first infers homologous genes clusters and then harnesses phylogenetic information to post-process these gene clusters for ortholog identification. The first section describes how it identifies homologous relationships using the graph-based approach (DIAMOND and MCL) and how the panX divide-and-conquer strategy facilitates the all-against-all sequence alignment for large datasets with runtime scaling approximately linear. The second section goes through a three-step post-processing procedure, which takes advantage of branch length and gene duplication information to separate orthologs from paralogs based on an adaptive cutoff estimated from core genome diversity. The third section mainly covers phylogenetic analysis of gene clusters and the reconstitution of core genome SNP tree. The fourth section represents the analysis of branch association and presence/absence association between gene clusters and their metadata values.

Chapter 4 details how I have benchmarked panX's clustering accuracy using simulated pan-genomes and compared the clustering results produced by different tools on real data. The first part describes the process used to conduct the pan-genome simulation. Next, the assessment of clustering quality of five tools on these simulated data is presented. Lastly, using real data from highly diverse and less diverse species, agreement and disagreement of orthologous clusters between these methods are demonstrated.

Chapter 5 highlights the hallmark of the panX interactive visualization application I have developed, by describing its main components: (1) pan-genome statistical charts; (2) a searchable gene cluster table and an alignment viewer; (3) a comparative panel of core genome SNP tree and gene tree; (4) a strain-specific metadata table. Chapter 6 pertains to application cases based on the panX analysis pipeline for various collections of bacterial genomes, as well as a large-scale dataset in one of our collaborative projects. The final chapter aims to briefly discuss our future research directions.

2 BACKGROUND

The rapidly expanding volume of microbial genomic data has facilitated in-depth investigations into the complex genetic variation in bacteria. Extensive genetic diversity has been documented even for closely related isolates [49]. When progressively including a new genome sequence, the size of the gene pool among strains of a same species continues to grow [97]. These observations have inspired Tettelin et al. to introduce the concept of "pan-genome", which refers to the total repertoire of all genes in a given set of genomes. Afterwards, numerous studies have addressed different aspects of microbial pan-genomes across a broad range of species from human pathogens to the extremely diverse marine cyanobacterium Prochlorococcus [54, 84, 103].

These pan-genome studies have provided insights into bacterial genomic diversity and their evolutionary dynamics, and have advanced our understanding of the genes associated with clinically important pathogenic strains and their epidemiological markers [64, 72]. Recently, in light of pan-genomics, gene presence/absence patterns related to infections have been identified and new candidate virulence factors and novel genomic islands discovered [70, 101].

Based on the amount of genes shared in different strains, a pan-genome can be considered as the union of the core and accessory genome, as well as unique genes. The core genome typically comprises genes shared by all strains; the accessory genome consists of genes present in two or more but not all strains and unique genes are strain-specific.

2.1 CORE AND ACCESSORY GENOME

CORE GENOME

A large proportion of the core genome consists of genes that play crucial roles in maintaining basic cellular functions, such as housekeeping and regulatory genes [97].

The evolutionary mechanisms in the core genome are of vital importance to bacterial adaptation. It has been exemplified in different species that point mutations of conserved core genes can impart resistance to antibiotics [86, 96, 108], which significantly enhances bacteria survival under antibiotic pressure. In addition to mutational changes, horizontal gene transfer can contribute to the dynamics in core genome as well. Whereas plenty of core genes are highly conserved and less frequently horizontally transferred, several studies have also demonstrated that signals of horizontal transfer in the core genome are present in different bacterial species [105]. For example, Everitt et al. have discovered evidence of core genome transfer in Staphylococcus aureus, which is driven by the proximity of mobile elements [22].

Core genome as species backbone

Core genes can be regarded as relatively "stable" in comparison with accessory genes that are more frequently horizontally transferred. Considering that the composition of the core genome primarily encompasses vertically inherited genes, the SNPs in this core set of genetic material have been widely used to reconstruct phylogenetic relationships among bacterial isolates. Furthermore, core genome has been proposed as a fundamental genomic unit for defining bacterial species by Lan and Reeves [62]. Many studies have provided supportive evidence for the use of the core genome as a proxy for defining prokaryotic species. For example, Lefébure et al. have found a resistance to recombination between the core genome of two closely related sister species Campylobacter coli and C. jejuni, each of which carry unique core genes that are not identified in other species [65]. This test of core genome hypothesis suggests that core genome could be viewed as genomic backbone for the delineation of bacterial species. Notwithstanding horizontal transfer events involved in the core genes, the core genome still remains as a useful surrogate for understanding the evolutionary relationships among strains and capturing the principal features of the species phylogeny.

Strict and soft core genome

Although the core genome is often defined as a set of genes present in all strains (strict core), the stringency may need to be adjusted in different scenarios, depending on the species diversity, input data quality and research question. Based on this consideration, the delineation of core genome can be relaxed from strict core to soft core, which correspondingly comprises genes shared by the majority of strains (e.g., present in > 80% strains). As shown in [51], using a soft core cutoff 95% on the 186 Escherichia coli genomes yields a more reasonable size of core genome, while the strict core threshold 100% leads to only 55.7% of genes identified using the soft core cutoff. In order to avoid the omission of core genes, a reliable delimitation of core genome may necessitate a comparison among different cutoffs.

Moreover, when addressing issues concerning pathogen surveillance and epidemiological investigation, even finer resolution of core genome needs to be achieved. As illustrated in a recent study on Salmonella enterica serovar Typhimurium [30], the core genome of Salmonella has been revealed to be much smaller than that of S. Typhimurium, with the latter exhibiting a greater discriminatory power for epidemiological research on outbreaks. Thus, the choice of genomes to be included in an analysis can significantly influence the typing resolution. Dependent on the research topic, the appropriateness of the composition of core genome should always be examined with caution.

ACCESSORY GENOME

In contrast to the core genome, genes in the accessory genome, also described as the variable genome, undergo gene gain and loss more frequently [54]. As an important driving force in the evolution of accessory genome, genetic material acquired via horizontal transfer events from the surrounding environment or other organisms helps equip bacteria with new functional potentials and allows them rapidly adapt to adverse conditions [13, 27, 59]. Accessory genes are of great relevance to bacterial expansion to new ecological niches, facilitate the bacterial colonization of different hosts, contribute towards microbial pathogenicity and resistance [17, 77, 89]. Besides, the accessory genome encompasses mobile genetic elements such as transposons, plasmid, phage, pathogenic islands, integrative and conjugative elements (ICEs) [10, 32, 50].

There has been much interest in identifying genetic materials that make bacteria host-specific. For example, host specificity of Pseudomonas syringae has been revealed only weakly related to its core genome [87]. The highly virulent strain *PA14* of Pseudomonas aeruginosa with broad host range has been shown to possess pathogenicity islands which are not found in the less virulent strain *PAO1* [40].

The dynamics of pervasive horizontal transfer shape the evolutionary histories of accessory genes, which have been gained at different times and sometimes accompanied by loss events. Such complex phylogenetic relationships are difficult to represent in a single tree and yield significant incongruence between species phylogeny and gene phylogenies. This necessitates concepts such as phylogenetic forests [57] by Koonin et al. and phylogenetic networks [46, 47] by Huson and Bryant, to better delineate the intricate details of evolutionary pathways.

2.2 OPEN OR CLOSED PAN-GENOME

With new strains being added in pan-genome analysis, the size of the core genome may shrink quickly or slowly, while the volume of pan-genome can expand with a significant proportion of strain-specific genes or remain unchanged. Correspondingly, bacterial pan-genomes have been considered as either open or closed [63, 98]. An open pan-genome results when each additional strain contains novel, previously unobserved genes, even after a great quantity of strains have been analyzed. This observation suggests a seemingly infinite pool of genetic material. The number of accessory genes in this scenario can be massively larger than that of core genes. A closed pan-genome suggests that additional strains do not introduce any new genes and the total gene pool might be relatively stable.

As shown in the study of eight pathogenic strains of Streptococcus agalactiae [group B Streptococcus] [98], Tettelin et al. have observed that the size of corresponding pan-genome grows with 33 novel genes, on average, when a new genome sequence is included. They further found intriguing results in the comparison of similar analyses on five isolates of Streptococcus pyogenes [group A Streptococcus] and eight isolates of Bacillus anthracis. While the pan-genome of the S. pyogenes strains expanded with an average of 27 strain-specific genes from every new genome, the pan-genome of B. anthracis strains ceased to enlarge after the inclusion of only a fourth genome. The authors then interpreted the results in an intriguing way and came up with the conclusion that the current prokaryotic species definition is inconsistent and species showing an open pan-genome are actually species (though counterintuitive): because, as a clone of Bacillus cereus instead of a true species, B. anthracis has a very distinctive phenotype associated with anthrax toxin-coding plasmid.

Besides, whether a predicted pan-genome is actually open or closed would depend on the sampling. Only by including a large number of genome sequences, one could potentially delineate the entire pan-genome.

2.3 inferring orthologous gene clusters

Orthologs and paralogs

To construct the pan-genome for a collection of bacterial isolates, an essential step is to identify the orthologous relationships among genes in all strains. In 1970, Walter Fitch first introduced the concepts of orthology and paralogy to depict two types of homologous relationship driven by different evolutionary events, speciation and duplication, respectively [25]. Homologs pertain to characters (typically referring to genes) descending from a common ancestor [26]. Homologs may share significant similarity, whereas similar characters are not necessarily homologous [56].

Orthologs are homologs arising after a speciation event, namely when ancestral genes are separated into distinct species derived from the last common ancestor. Although orthologs generally perform similar biological functions [80], evidence on diverged functions between orthologs has also been found [18]. By contrast, paralogs are homologs resulting from a duplication event within a species and tend to evolve distinct functions [69]. The mix of speciation and duplication, together with gene gain, loss and rearrangements, leads to highly complex evolutionary relationships [55].

The inference of these two evolutionary events and putative orthologous relationships typically relies on computational approaches [58]. A variety of algorithms for inferring orthology has been developed, which can be generally classified into graph-based and tree-based methods [58, 60].

Graph-based method

Graph-based approaches identify orthologous genes based on pairwise sequence similarity score in heuristic manners, which are computationally efficient and capable of analyzing considerably large datasets. Among them, the most commonly used tool orthoMCl [67] constructs a similarity score matrix as a graph, with nodes representing protein sequences and weighted edges denoting their relationships. This graph is then separated into sub-graphs to form orthologous clusters via the Markov clustering algorithm (MCL), which simulates random walks within the graph and iteratively calculates the transition probabilities among the nodes. [21]. Despite the obvious computational efficiency, graph-based approaches neglect the evolutionary details that can be obtained from a group of sequences and have proved more prone to erroneous clustering especially when differential gene loss events are involved [37, 88].

Tree-based method

Tree-based approaches, which are conceptually in accordance with the definition of orthology, chiefly utilize phylogenetic information to infer orthologs and paralogs via reconciliation between gene tree and species tree [73, 79]. They typically construct a gene phylogeny from multiple sequence alignment of homologous genes and then fit the gene tree to its species tree by maximum parsimonious approach [36] to detect speciation and duplication events. This is based on the assumption that the minimum number of evolutionary events is likely to represent the most probable paths of evolution. Several techniques and databases [19, 43, 102] have been developed. Yet, many of these methods depend on a well resolved species tree that is not always known [60] and the heavy computational burden from inferring large phylogenies hinders its application in large-scale analysis.

Besides, there are also hybrid approaches making use of both. Hybrid approaches aim at bypassing the limitations from graph-based and tree-based methods, by reducing the computational cost and harnessing evolutionary information from phylogenies [58]. For example, Ortholuge applies phylogenetic distance comparisons to evaluate orthologs generated by heuristic algorithms [31]. EnsemblCompara refines BLAST-based results by a tree-reconciliation technique to infer orthologs in vertebrate genomes [104]. HomoloGene refines pre-computed BLAST results by utilizing auxiliary information of gene neighborhood conservation and a guide tree to predict homologs for many eukaryotic species. [106].

Nonetheless, these methods are originally designed for handling distantly related organisms rather than for analyzing closely related bacterial strains. Therefore, further investigations are required to assess their performances in pan-genome analysis, which often primarily focuses on strains within a species.

2.4 A BRIEF OVERVIEW OF PAN-GENOME TOOLS

Building a pan genome mainly relies on heuristics for identifying orthologous gene clusters [103]. Several tools for pan-genome analysis are available [28, 78, 111, 112]. Among them, PanOCT is a graph-based ortholog searching program designed for pan-genome analysis of closely related bacterial genomes, which combines conserved gene neighborhood information and homology search results to partition recently diverged paralogs into clusters of orthologs [28]. PGAP pan-genome analysis pipeline relies on MultiParanoid [1] program and Gene Family [111] method to generate a pan genome from annotated genome sequences and includes other analysis modules such as genetic variation analysis of functional genes and function enrichment analysis of gene clusters.

PGAP and PanOCT utilize BLAST to perform all-against-all sequence alignment, which leads to a quadratic runtime in the dataset size [78]. Inefficient scaling with the number of genome sequences makes these tools impossible to handle large-scale datasets. On the contrary, Roary harnesses a pre-clustering technique (based on CD-HIT [29]) that makes the computation substantially faster than PanOCT and PGAP. The computational efficiency of Roary has been demonstrated on a large pan-genome analysis of one thousand strains [78].

Chiefly designed for closely related strains, both Roary and PanOCT use additional information from conserved gene neighborhood (CGN) to aid in clustering orthologs. However, reliance on synteny to conclude orthology can be problematic, because genome rearrangements can occur frequently and result in beneficial variations that promote invading bacteria to escape host immune responses [44]. Besides gene rearrangements, when taking also into account gene duplication and loss, gene orders are not always conserved [60]. Huynen and Bork has pointed out the limited potential of synteny for detecting orthologs [48], which is practicable mainly for closely related species. Additionally, there has also been a suggestion that the requirement for conserved gene order should be relaxed [92]. Therefore, the clustering accuracy of the approaches utilizing

CGN information might be affected when dealing with datasets involving considerably high genome variations.

3 PAN-GENOME ANALYSIS

Taking into account both the efficiency of graph-based clustering and evolutionary information from phylogeny, panX analysis pipeline performs two main procedures: (1) identifying homologous gene clusters based on all-against-all sequence alignment via DIAMOND [11] and Markov clustering by MCL [21]; (2) post-processing clusters to separate orthologs from paralogs using several phylogeny-based algorithms (see Figure 3.1).

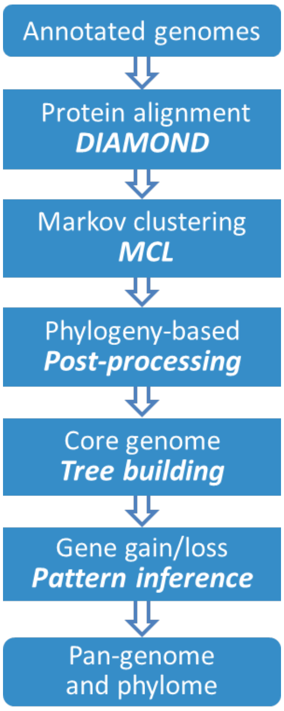


Figure 3.1: **PanX analysis pipeline**
Using annotated genomes as input (GenBank format), the pipeline extracts genes from all genomes and compares them using DIAMOND. The hits are processed by MCL for identifying homologous clusters. Then, panX harnesses phylogeny-based approaches to separate orthologs from paralogs. From these clusters, panX identifies the core genome, reconstructs the core genome phylogeny, builds an alignment and a gene phylogeny for each cluster, and infers the gene gain/loss events. The established pan-genome and its phylome (all gene phylogenies) can be interactively explored via the panX visualization tool.

3.1 IDENTIFYING HOMOLOGOUS GENE CLUSTERS

3.1.1 *Clustering homologs*

As a standard input, panX parses annotated genomes in GenBank format from the RefSeq database or the user's own genome data. All-against-all comparison of protein sequences is then computed by a rapid alignment program DIAMOND [11] (with default e-value threshold 0.001), which builds index on both query and reference sequences using a double-indexing algorithm. This yields a similarity table containing DIAMOND hits and corresponding bit score, which has been shown to have a good performance in the evaluation of different similarity measures for homology identification [34]. Subsequently, the table is processed by MCL to identify homologous relationships in the sequence similarity space. In addition, clusters of rRNA sequences from GenBank files can be produced using blastn and MCL. The sequences in these clusters are further aligned and corresponding rRNA trees built. When running on a computer cluster, panX uses 64 CPUs by default for DIAMOND protein alignment. Alternatively, user can also provide the output from other sequence similarity search tools such as blastx or blastn.

3.1.2 *Divide-and-conquer strategy for large datasets*

All-against-all pairwise alignment behaves quadratically with the number of input genomes. Nevertheless, considering the similarity among sequences in homologous clusters, many of them can be represented using a single representative sequence, which helps reduce the similarity search space, thereby significantly accelerating the homolog inference process. One approach is to harness a divide-and-conquer strategy to lower the quadratic cost by analyzing the large dataset first as subsets. As a widely used algorithm design paradigm, divide-and-conquer methods resolve a complex problem by partitioning it into simpler subproblems. The solution to the original problem can then be combined from the solutions to subproblems.

The entire collection of genomes is randomly divided into several subsets with the default size of 50 genomes. For each of the subsets, a "sub-pan-genome" is produced by the homology search pipeline based on DIAMOND and MCL. The representative sequences for all homologous clusters in each "sub-pan-genome" are then collected as a pseudo-genome. All generated pseudo-genomes are further processed in the final run using a similar homology search strategy. In the end, genes represented by the sequences in the pseudo-genomes are integrated together to construct the final clusters of the original sequences.

Based on testing results, the default subset size 50 has proved to be a good balance, since a subset size that is too small does not gain considerable computational advantage from DIAMOND's double-indexing approach and a subset size that is too large leads to significant performance slowdown in all-against-all sequence comparison. Moreover, the notably high number of hits to output in a collection of several thousand genomes correspondingly requires extremely large amounts of storage and memory. Therefore, a moderate subset size used in the divide-and-conquer method can significantly reduce the computational burden.

When running DIAMOND on the subset, panX applies a strict threshold of 90% sequence identity and 90% aligned length for both query and subject sequences, which aims at first clustering highly similar sequences and avoids generating less effective representatives for clusters with high diversity. At the final stage, panX calls DIAMOND with the default cutoff of e-value 0.001 and without further restriction to cluster representative sequences in pseudo-genomes from all subsets, ensuring that orthologous genes in diverse species are processed properly.

The simple but efficient application of the divide-and-conquer strategy lowers the complexity of all-against-all comparison from quadratic to almost linear (see Figure 3.2), thus making the inference of homologous sequences in large-scale pan-genome analysis computationally feasible.

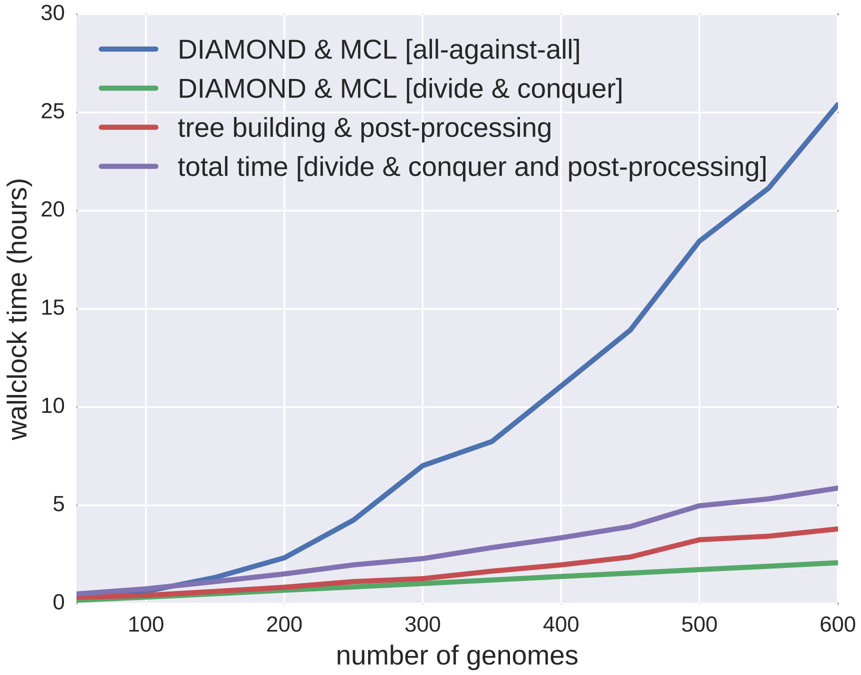


Figure 3.2: **Scalable clustering by divide-and-conquer strategy** PanX applies DIAMOND and MCL for the identification of homologous gene clusters in a large set of annotated genomes. These clusters are further separated into orthologous groups by phylogeny-based post-processing. The graph demonstrates the runtime required for identifying orthologous clusters in pan-genome datasets of different size with 64 CPUs. The initial all-against-all protein alignment with DIAMOND scales quadratically with the number of genomes (blue line, all-against-all comparison). The divide and conquer strategy significantly lowers the computational cost and scales approximately linear (green line, divide-and-conquer) by first clustering subsets of genomes and then combining their clustering results into the final clusters (details described in text).

For pan-genomes of 500 genomes, building gene trees and post-processing on them take a similar amount of time as the clustering procedure. This well-scalable computational pipeline makes pan-genome analysis on large-scale datasets feasible.

3.2 post-processing of gene clusters

The degree to which the homologous clusters may contain putative orthologs can vary significantly based on the cutoffs applied in the homolog search procedure. In datasets with low diversity or including many duplicated sequences, it has been observed that large clusters are formed due to either remotely homologous relationships or a substantial amount of gene duplications. By contrast, for datasets with much larger diversity, sequences in orthologous clusters show notably more variation. Accordingly, an appropriate cutoff needs to be carefully chosen for different scenarios.

In order to avoid repetitive tests for determining the suitable parameter settings, sequences can be first aggressively clustered and then split into orthologous groups via post-processing techniques. While over-clustering can be directly resolved by inspecting the phylogeny of erroneously merged gene cluster, under-clustering is much more difficult to detect and cope with. Therefore, panX performs the all-against all comparison with a less stringent e-value cutoff of 0.001 to filter the alignment hits. The output can then be further refined via an adaptive criterion based on the phylogenetic properties. This flexibility makes the analysis pipeline suitable not only for closely related strains from a species but also for the genome collection of diverse species.

The branch lengths in a phylogeny reflect evolutionary distances among a group of sequences. Breaking up prominent long branches can partition distantly related sequences into sub-groups of more closely related sequences. Together with the number of gene duplications shown in the gene phylogeny, the paralogs can be accurately ascertained, which helps separate homologous sequences into putative orthologs.

For each homologous gene cluster, panX first constructs an alignment of amino acid sequences by MAFFT [53]. Then, a codon-based alignment is inferred for the corresponding nucleotide sequences, with a three-nucleotide gap being added for each gap in the amino acid alignment. Based on the nucleotide alignment, panX builds a phylogenetic tree using FastTree [81]. The derived phylogenies are further examined by the following post-processing techniques, which split the original cluster into sub-clusters when evidence of paralogy has been found.

3.2.1 *Splitting distantly related homologs*

Remotely related homologous sequences can be readily identified in a gene phylogeny, which are linked via long branches. Especially for datasets with low diversity, splitting long branches can separate those distantly related sequences from an orthologous group. Nevertheless, the branch length cutoff often varies for different species. In order to automatically determine which branches are long enough to be split across a wide range of species, we apply an adaptive cutoff based on the estimated core genome diversity.

This cutoff b_c is computed from the average diversity d_c of single-copy genes in the core genome by the formula

$$b_c = \frac{0.1 + 2d_c}{1 + 2d_c} \ . \tag{3.1}$$

The cutoff value approximates to 0.1 for highly similar genomes and increases with the d_c until it reaches the saturation of 1. We have found the default factor 2 works well for a broad range of datasets, though this factor can be adjusted by corresponding parameter `-fcd` (–factor_core_diversity).

PanX harnesses core gene diversity as a surrogate estimate for the distance to the most recent common ancestor (MRCA) in a set of strains. Branch lengths longer than the estimated diversity can then be used to indicate duplication events prior to the MRCA, which aids in deciding whether a branch should be cut or not. Yet, when analyzing species with high genetic diversity (e.g. $d_c > 0.25$), genes in the same orthologous cluster can have significant genetic variation such that this long branch splitting step should be skipped to avoid under-clustering issues.

After separating sequences associated with long branches into individual clusters, panX builds alignments for the gene sequences of these new clusters and reconstructs the corresponding gene phylogenies. The new trees can be further split using the above-described process as long as it contains branches longer than the defined diversity cutoff.

3.2.2 *Splitting closely related paralogs*

While checking long branches can identify remotely related homologs (including ancient paralogs), relatively recent paralogs need to be examined more carefully. PanX applies a linear discriminator for examining evidence of paralogy, by combining information on both branch length and the number of gene duplications calculated along a branch.

The latter is computed as a paralogy score by traversing the tree twice. The score counts the number of strains linked to both sides of a branch in a gene tree. Namely, the paralogy score is the size of the intersection of two strain sets $B \bigcap (T - B)$, with B containing strains represented in the leaves on a branch and T including all strains shown on a gene tree.

The searching process for the best branch to split goes through the entire gene tree and marks a branch as a temporary best candidate, if it has a larger paralogy score or longer branch length with equal paralogy score. The decision for splitting a cluster into two sub-clusters is made, when the largest paralogy score ϕ_{max} and the corresponding branch length ℓ satisfy the following two conditions:

$$\phi_{max} > 0$$
$$\frac{\ell}{b_c} + \frac{\phi_{max}}{1.5 \times \#\text{strains}} > 1.0$$

b_c is the same cutoff on the estimated core genome diversity as defined in (3.1). The first condition $\phi_{max} > 0$ avoids splitting when no paralogy is present. Although more sophisticated criteria could be applied, this linear discriminator has demonstrated good performance across a large number of datasets.

PanX conducts the paralogy splitting procedure iteratively until the above-mentioned conditions cannot be fulfilled. Additionally, flexibility is provided to users such that branch cutoff for splitting paralogy b_c can be modified via the corresponding parameter, which influences the extent to which clusters are split.

3.2.3 *Identifying and resolving fragmented clusters*

Extensive tests on a large quantity of datasets have revealed an issue with the fragmentation of a small number of clusters, which mainly results from either a failure in homology search or a problem in Markov clustering. This has been reflected as suspicious peaks in the gene length distribution. These peaks are composed of numerous singleton clusters, all of which have identical gene length.

In order to automatically detect these peaks, panX first computes the average length of genes for all clusters. Next, panX scans the distribution of the average gene length to ascertain the positions of peaks, which identifies signals in the gene length distribution compared to a smoothed background distribution.

For each detected peak, genes of the clusters involved in the peak are collected into one temporary cluster. Their sequences are further aligned and the corresponding phylogeny is inferred. These merged clusters then undergo the above-explained branch splitting procedures. This approach has effectively resolved the fragmented clusters and combined them into correct orthologous group, though other strategies could be considered to check the surrounding clusters adjacent to the peak.

For low diversity simulated datasets, approximately 40% of homologous clusters went through post-processing, whereas for high diversity datasets, only a small number of clusters required splitting.

The clustering result is updated in each of the post-processing steps, during which the initial cluster records are cleaned and newly formed clusters are added. For each final cluster in the pan-genome, corresponding alignment and gene phylogeny are built as described below.

3.3 PHYLOGENETIC ANALYSIS OF GENE CLUSTERS

PanX generates alignments, trees and other summary statistics on the gene clusters for interactive visualization and exploration of the pan-genome, which by default takes the results of the aforementioned pipeline analysis as input. Alternatively, the clustering output file from `Roary` could be used. Besides, the output from other pan-genome tools can be accepted when providing the clustering results in the same format as panX. For each non-singleton gene cluster, panX aligns amino acid sequences in the cluster by MAFFT [53]. Next, a codon-alignment is constructed from the amino acid alignment for the corresponding nucleotide sequences, by adding a three-base-pair gap for each gap in the protein alignment.

Inferring phylogenetic trees and ancestral sequences

PanX builds a core-genome SNP matrix using all variable sites in the nucleotide alignments, which is constructed from all single copy core genes. A core genome SNP tree is then reconstructed from the SNP matrix by FastTree [81] and further refined by RaxML [93] based on a similar strategy developed in nextflu [76].

When taking the recombination into account [22], the core genome SNP phylogeny is not necessarily in accord with the phylogeny of each individual core gene. Moreover, the branch length does not reveal the real phylogenetic distance and reflect original sequence similarity [3], since the phylogeny is built using only variable positions. Nonetheless, branch length proportions are accurately represented. Notwithstanding these considerations, the core-genome SNP tree can still act as a useful surrogate for assessing the genetic relatedness among different strains.

Based on the phylogenetic trees built in the post-processing procedures, ancestral sequences of internal nodes on these trees are further computed by a joint maximum likelihood approach [24] as developed in treetime [85]. During the ancestral reconstruction, inferred mutations are mapped onto the branches of the gene tree.

Subsequently, panX applies a similar ancestral reconstruction technique to compute whether a gene is present or absent on the internal nodes of the core genome SNP tree. This ancestral inference strategy infers individual gene gain and loss events associated with branches, with the gain and loss rates being optimized by maximizing the likelihood for the observed gene presence/absence patterns [23, 110]. The estimated rates of gene loss consistently exceed the estimated rates of gene gain. The ratio of gene loss to gene gain varies with a median ratio of 22 (interquartile range, 9 to 35) in different bacteria.

Gene clusters and their corresponding phylogenies, as well as mutations and metadata, are stored in JSON files for the panX web-based interactive visualization.

3.4 ASSOCIATION ANALYSIS OF GENE CLUSTERS

Another useful feature of panX is the exploration of the gene distribution on the core genome SNP phylogeny and its association with specific metadata. When numerical metadata such as minimum inhibitory concentration (MIC) are provided, the stratification pattern of gene variants linked to specific phenotypes could be identified on a gene tree. Moreover, frequently gained and lost genes can also exhibit strong associations with different phenotypic traits. Based on these observations, two types of associations have been taken into account, of which the corresponding scores are computed for every gene cluster by panX:

(1) gene variants are linked to specific phenotypic characteristics;

(2) the presence/absence of a gene is associated with a phenotype.

3.4.1 *Branch association*

In order to estimate the degree to which specific gene variants are linked to a phenotype, panX takes advantage of a normalized difference of phenotypes of strains on each side of a branch in a gene tree, as shown in:

$$b_a = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{3.2}$$

where $\mu_{1/2}$ and $\sigma_{1/2}^2$ refer to the mean and variance of the phenotypes on each side of a branch. PanX computes this association score for each branch of a gene tree, from which the highest score is chosen as the final *branch association score*.

For example, when sorting the branch association score for benzylpenicillin MIC in the gene cluster table on the panX website http://pangenome.tuebingen.mpg.de/S_pneumoniae616, the cluster showing the highest score pertains to the penicillin binding protein *pbp2x*. This is confirmed when choosing benzylpenicillin MIC as the metadata coloring option for the core genome tree and the gene tree (see Figure 6.3). The resistant strains form a single clade in the gene tree for *pbp2x* and are separated from susceptible strains by a large number of amino acid substitutions, which are shown on the tooltip of the corresponding branch.

3.4.2 *Presence/absence association*

To measure the extent to which a gene presence and absence pattern is associated with a phenotype, panX calculates a *presence/absence association score* based on the average phenotype μ_p of strains having the gene and the average phenotype μ_a of strains not having the gene, the overall variance of the phenotype σ^2, and the number of gain and loss events n as represented by the following formula

$$p_a = \sqrt{n}\frac{\mu_a - \mu_p}{\sigma} \tag{3.3}$$

For example, high-score candidates in the gene cluster table indicate that susceptibilities to benzylpenicillin, trimethoprim, erythromycin and ceftiaxone are strongly associated with the presence of the gene *mefE* and the gene *mel*, which are expected for an efflux pump.

Although there are false positives mixed in computed association patterns, these effective association scores tremendously narrow down the gene search space and thus largely facilitate the inspection of potential genes of interest in further downstream analysis. We explicitly compared the candidate genes showing strong association with benzylpenicillin resistance in panX and the previously reported genes associated with beta-lactam resistance by [12]. In their study, many of the associations were considered false positives by the authors. The plausible biological candidates in the paper are pbp2x, pbp1a, pbp2a, mraY, mraF, ftsL, gpsB, recU, clpL, clpX, dhfR. Five of them (pbp2x, pbp1a, mraY, gpsB, recU) are among the strongly associated candidate genes found by the branch association score in panX. These scores largely confirm their reported results.

Furthermore, the web-based exploration platform provides users the flexibility to sort genes via their association scores, which helps rapidly identify genes exhibiting strong evidence of association with specific metadata. When using the corresponding metadata to color the species and gene tree, one can readily inspect the consistency of the association.

3.5 AVAILABILITY

The panX computational pipeline for the construction of a pan-genome and the phylogenetic analysis on all gene clusters is based on a collection of python scripts for individual modules and a master script controlling different analysis steps. Clustering results from other orthology inference programs can also be loaded into panX pipeline and explored in panX web-based visualization application, especially when user intends to visually compare the clustering results between different tools.

The visualization software is built upon the node.js server and takes advantage of D3.js [7], dc.js [107], BioJS [35], as well as other JavaScript libraries. The code for the analysis pipeline and the visualization application is available on github as repositories pan-genome-analysis and pan-genome-visualization under the GPL3 license. The browser-based application can either be hosted on a web server or run locally for interactivity exploring pan-genomes established by the panX analysis pipeline.

3.6 CONCLUSION

In order to build a computationally efficient pipeline for pan-genome analysis applicable to a wide range of bacterial species, we have developed panX software that combines graph-based approach and phylogeny-based post-processing for rapid identification of orthologous gene clusters in large-scale datasets. PanX harnesses DIAMOND and MCL to infer homologous relationships among genes and takes advantage of a divide-and-conquer strategy to achieve approximately linear scaling with the number of input genomes. In the divide-and-conquer algorithm, panX first partitions the entire dataset into subsets of 50 genomes and searches homologous relationships among genes in those subsets. Then, it further clusters the representative sequences derived from the subsets and integrates all clustering results into the final orthologous groups. As a substantial improvement for the naive all-against-all sequence comparison, this makes pan-genome analysis on a large collection of bacterial genomes feasible, with the runtime required for constructing a pan-genome of 1000 strains being less than a day on 64 CPUs.

Another hallmark of the panX computational pipeline is that its takes advantage of phylogenetic properties such as branch length and the number of gene duplications to split orthologous clusters from paralogous clusters. As a distinctive feature, the cutoff for branch splitting procedure in panX is adaptively estimated based on core genome diversity, which makes panX applicable for a wide range of species with different diversity levels.

Furthermore, panX constructs multiple sequence alignment for each gene cluster and builds the corresponding gene phylogeny. This produces a comprehensive phylome, which serves as an extremely useful resource for studying evolutionary relationships among strains and their genes. The comparison between the species tree built on the core genome SNPs and individual gene trees allows in-depth investigations of horizontal transfer events. Last but not least, panX computes scores for branch association and presence/absence association between gene clusters and numeric metadata, which facilitates the identification of candidate genes related to phenotypic characteristics such as antibiotic resistance phenotypes.

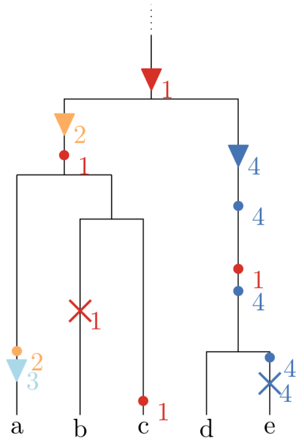
4 BENCHMARKING ORTHOLOG CLUSTERING ON PAN-GENOMIC DATA

Since the identification of putative orthologous relationships substantially depends on computational inferences that vary in their heuristics, it is not surprising to observe differences among the clustering results from different tools. Besides, lack of gold standard datasets for assessing the clustering accuracy has been a major hindrance for reliable benchmarking on the orthology inference in the pan-genome context. Therefore, we have applied a simulation approach based on prokaryotic reference genome data for generating pan-genome datasets.

4.1 SIMULATING PAN-GENOME

Here, we created 120 pan-genome datasets, with each containing 30 simulated genomes. Based on the genome sequence from Escherichia coli K-12 reference strain (NCBI accession number NC_000913), 2803 representative genes extracted from all its KEGG ortholog groups [52] were used to constitute an ancestral genome, with which the simulation process started. We then conducted the simulation by evolving the ancestral sequences along coalescent trees produced by the program ms [42] and also taking into account gene gain and loss and horizontal transfer events (an illustration of the simulation process is shown in Figure 4.1).

Using above-mentioned representative genes as ancestral sequences, we simulated pan-genomes by the following procedure. For each of the 2803 genes, a correlated tree was generated using the software ms [42] with different rates of horizontal transfer. If the gene transfer rate was zero, all 2803 genes evolved according to the same clonal genealogy of the population, namely, one common species tree. Notwithstanding the fact that the individual gene trees might differ if some genes were affected by gene transfer events, the gene trees were still significantly related to each other because they were linked to the common clonal genealogy. In order to inspect the impact of gene transfer on the accuracy of reconstruction, we decided to apply three different gene conversion rates for simulating the gene trees using ms, which corresponds to no, occasional and frequent gene conversion, respectively.



	gene 1	gene 2	gene 3	gene 4
refseq	RRRRR	RRRRR	RRRR	RRRRRRR
ind. a	-A---	--T--	----	∅
ind. b	∅	-----	∅	∅
ind. c	-AA--	-----	∅	∅
ind. d	T----	∅	∅	---AC--
ind. e	T----	∅	∅	∅

Figure 4.1: **Simulation of pan-genome datasets**
In the pan-genome comprising a total of 2803 gene clusters, we simulate the gene gain point, potentially subsequent gene losses, and mutations within these genes along a predefined genealogy. Without gene transfer, the events occur along the same clonal frame, simulated by the software *ms*. This clonal frame thus equals a single coalescent tree. Along each branch, a gene can be gained (down-triangle). Moreover, existing genes can undergo single point mutations (dot) or get lost by a gene loss event (cross).The left graph illustrates the simulation process for 4 different genes. Gene 1 has been placed at the root. The other genes are gained at random points along the tree. The index at each symbol indicates the affected gene. The table shows the resulting gene sequences. For example, gene 1 is lost in individual b and mutations have happened several times along different branches; gene 4, gained in the last common ancestor of individual d and e, has been again lost in individual e. Besides, for the scenarios involving gene transfer simulated based on options in *ms*, we do no longer use the same tree for all genes.

Whereas 2100 genes out of the 2803 genes were assigned to the most recent common ancestor (MRCA) at the root of the simulated gene tree, the rest of genes were gained at uniformly distributed points along the branches after the MRCA on the gene tree. While 300 random genes from those 2100 ancestral genes were set to be ever present, the remaining 2503 genes underwent a loss event at rate 2.1 along the branches of the corresponding gene tree as defined in [2, 45]. In addition, as a consequence of a loss event, the corresponding gene was then set to be always absent from all descendants from the branch of the loss event.

Given the gene trees and the presence/absence pattern for each of the representative K-12 sequences, mutations were allowed to happen along the branches of the corresponding gene tree. In order to simulate these mutations, we applied the program Seq-Gen [83] and the substitution model HKY [39], with the empirical *E. coli* base nucleotide frequencies set as the base frequencies and 1.1 as the transition-transversion bias.

For scenarios pertaining to frequent, occasional or no gene conversions, we simulated 5 different sets of trees. For each of these sets, pan-genome datasets were produced using 8 different substitution rate distributions: an exponential distribution with mean 0.06, uniform distributions between 0.05 and 0.1 and between 0.1 and 3, and constant substitution rates of $0.01, 0.05, 0.1, 0.2$ and 0.3. The substitution rate μ of each gene was based on the corresponding distribution. The mean number of substitutions per site between two strains was calculated by $1 - e^{-\mu T}$, with T being the distance between both strains in the gene tree.

4.2 BENCHMARKING AND COMPARING DIFFERENT TOOLS

Although real data can be used to compare the differences between clustering results among various methods, the accuracy assessment remains difficult to investigate in an absolute measure. In order to explicitly evaluate the performance of panX's ortholog clustering and compare it to other tools, it is of great importance to know which genes should be clustered in the same orthologous group. In addition, the benchmark dataset *orthoBench* based on manually curated protein families [100] has been designed for comparing proteins from different domains of life, thereby making it less appropriate for benchmarking methods on pan-genome analysis of closely related bacterial strains.

The orthology inference from pan-genome datasets by different approaches relies on their heuristics and underlying assumptions, where no ground truth is available to directly estimate the quality of the generated orthologous clusters. For the purpose of assessing accuracy of the clustering results among various programs panX, Roary [78], PanOCT [28], OrthoFinder

[20] and OrthoMCL [67], we established simulated pan-genome datasets, in which the ground truth is attainable for the evaluation on whether orthologous groups are correctly clustered. Moreover, we assessed agreement regarding to the orthologous clusters between these methods, using pan-genomes built from real datasets on highly diverse and less diverse bacterial species.

OrthoMCl and OrthoFinder were implemented for orthology inference on protein sequencing across different domains of life, not specially focused on gene clustering in bacterial pan-genome of similar strains. Nevertheless, these tools have also been used in pan-genome studies. Hence, we have included them in the benchmark tests. By contrast, Roary and PanOCT were developed for analyzing pan-genome of closely related bacterial strains. These tools are therefore expected to provide reasonable performance when tested on different parameter ranges.

Among all five methods, an outstanding characteristic of panX is that it harnesses phylogeny-based post-processing to improve the initial clustering results produced by MCL. This does not require users to examine how different cutoffs could affect the results, on account of the fact that the post-processing is designed to be highly adaptive by estimating the thresholds based on the core genome diversity. Hence, panX exhibits good performances across a large range of diversities for different species.

4.2.1 *Assessing clustering accuracy on simulated data*

40 simulated pan-genomes of size 30 were constructed by evolving 2803 genes from the E.coli K-12 strain along gene trees using the program ms. Gene sequences were mutated based on different substitution rate distributions across the genome, which could also be gained, lost, or horizontally transferred. For these simulated pan-genomes, we examined the clustering results by Roary, PanOCT, OrthoFinder, OrthoMCL and panX and measured the number of erroneous clusters for each tool.

To quantify the clustering errors compared with the ground truth, predicted clusters are classified into the following categories:

(i) if a cluster contains all and only genes from one true orthologous cluster, it is correctly assigned,

(ii) if a cluster contains only a subset of genes from one true cluster, it is under-clustered,

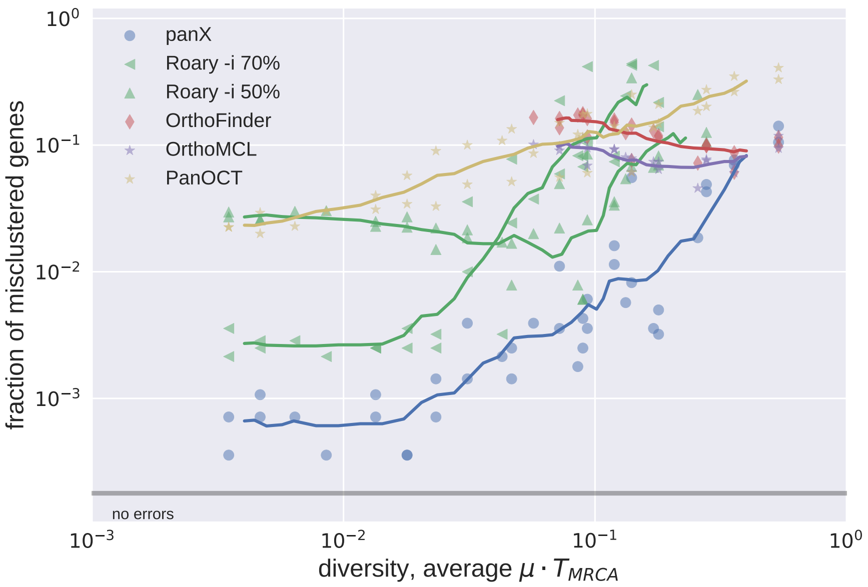


Figure 4.2: **Accuracy of clustering by different tools.**
Overall, the fraction of erroneously clustered genes increases with diversity of the pan-genome dataset. We have run Roary with sequence identity cutoffs `-i 70` and `-i 50`. At low diversity, panX and Roary (`-i 70`) demonstrate similar accuracy and generate only less than 0.01% erroneous clusters. At high diversities, the clustering accuracies of all tools are similar and miscluster 10% of genes. For clarity purposes, clustering results for tools designed for high diversity datasets (distantly related organisms) such as OrthoMCL and OrthoFinder are only shown for diversities above 0.02. Similarly, considering that Roary is not developed for very diverse datasets of genomes, results for Roary are not included at high diversity.

(iii) if a cluster contains not only all genes from one true cluster but also genes from other clusters, it is over-clustered,

(iv) a cluster may involve both over- and under-clustering errors. The performance of tools on the pan-genome at varying levels of diversity is illustrated in Figure 4.2. Furthermore, details on different type of errors (false merging/splitting) are demonstrated in Figure 4.7. Additionally, we conducted the same analysis on datasets with different gene conversion rates and found comparable results for frequent (Figure 4.7), occasional (Figure 4.8) and no (Figure 4.9) gene conversions.

OrthoMCL and OrthoFinder belong to orthology inference methods that compare cross-species protein sequences at large evolutionary distances. When applied to the simulated datasets of low diversity that mimicked pan-genome analysis of closely related strains within a species, these tools merged a large number of clusters that should be separated but combined with other clusters. In our further investigation, as shown in Figure 4.3, we found that this effect occurred mainly in rare accessory gene clusters, with the vast majority of them being singletons, which were mixed with other clusters. By contrast, they normally correctly assigned genes into core gene clusters and common accessory gene clusters. Their relatively high misclustering rates on very low diversity datasets were not presented in Figure 4.2 for the sake of clarity.

Regarding to the datasets with high diversities, OrthoMCL and OrthoFinder showed an accuracy similar to that of panX. Similar behaviors have been found in Roary and panX on a broad range of diversities from below 1% to 30%, with typically a factor of two fewer errors made by panX.

Nonetheless, it was not possible to find a parameter set for Roary that performed well across the entire range of diversities. Whereas a 70% identity threshold (`-i 70`) performed best for low diversity datasets, much lower thresholds were required to make reasonable inference (e.g. `-i 50`) from high diversity datasets.

In addition, PanOCT didn't work well on the simulated data, mainly because of a significant proportion of erroneously split clusters, see Figure 4.3.

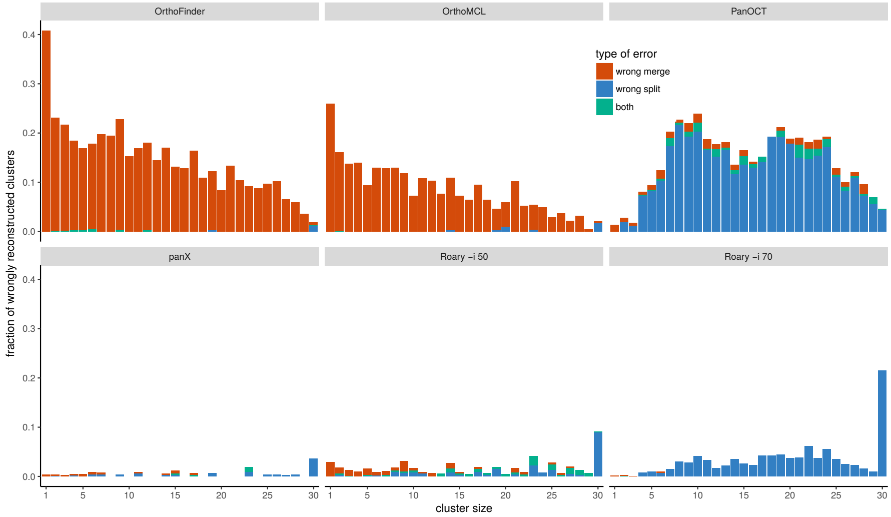


Figure 4.3: **Type of misclustering by tool and gene frequency.** The fraction of erroneously merged (red) and erroneously split (blue) clusters by gene frequency and clustering tool across all simulated datasets using exponentially distributed substitution rates with mean rate $\mu = 1/15$. OrthoMCL and OrthoFinder are tools primarily designed for comparing protein sequences at large diversities. They wrongly merged many rare accessory gene clusters, of which the vast majority pertained to singletons that were mixed with other genes. By contrast, they typicality correctly clustered genes that belonged to core genes and common accessory genes.

4.2.2 *Comparing clustering results on real pan-genomes*

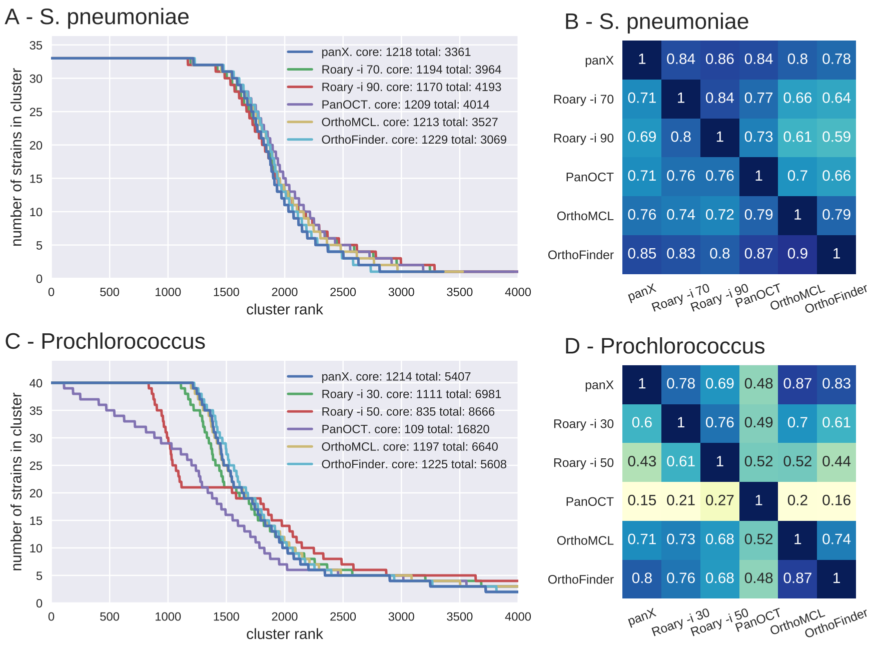
In comparison to the simulated pan-genomes, pan-genomes inferred from real data are generally more difficult to evaluate, since there is no straightforward way of assessing the accuracy of clustering results. Nevertheless, the difference in clustering behavior and the degree of agreement of the produced clusters can be investigated between different methods.

We selected all available *S. pneumoniae* strains (33 genomes) from RefSeq database as an example dataset of relatively low genetic diversity and 40 *Prochlorococcus* strains as a high diversity dataset to compare the clustering results generated by panX, Roary, PanOCT, OrthoMCL and OrthoFinder. These collections of genomes are chosen on account of the fact that only Roary and panX scale very well with large datasets.

In order to detect the congruence and incongruence of clustering behaviors between different tools, summary statistics on the size distribution of clusters, the number of the core genes and the total number of clusters are calculated for each dataset and each tool. As shown in panel A&C in Figure 4.4, we take advantage of an inverse cumulative distribution of clusters sizes for characterizing the clustering output. Namely, clusters are sorted in descending order by the number of strains that appear in the cluster such that the inconsistency among clustering results can be easily distinguished for the core genome (indicated by the first plateau of the curve) and the rest (i.e. from common accessory genes to rare accessory genes, as well as singleton genes). Furthermore, the fractions of identical gene clusters between tools are also computed by comparing the size and the content of clusters, see panel B&D in Figure 4.4.

For the low diversity dataset shown in panel A, different tools exhibit similar clustering results for *S. pneumoniae* strains regarding the cluster size distributions with a difference in the number of core genes being less than 10%, see Figure 4.4 A. Panel B illustrates that 78%-86% of clusters detected by panX are in agreement with the clusters identified by other tools. Among them, Roary, using identity threshold `-i 90`, shows the best agreement with panX.

For the scenario on the higher diversity dataset, panel C and D show more dissimilarity for the inferred pan-genomes of 40 *Prochlorococcus* strains, see Figure 4.4 C&D. When running with the default parameter setting (`-i 95`), Roary detected only 10 core genes and split the vast majority of gene clusters (31762 clusters in total). Although applying a much lower identity cutoff (`-i 30`) led to the warning message showing that it's not designed to support very diverse dataset, Roary generated 1111 core genes and 6981 genes in total, with 60% of clusters being identical to the clustering results from panX. Whereas OrthoFinder and OrthoMCL largely agreed with the clusters identified by panX and Roary with the parameter (`-i 30`), PanOCT showed a $2-3$-fold increase of the pan-genome size and produced $16,820$ gene clusters, with a large number of clusters containing merely $1-3$ genes and the core genome including only 109 gene clusters. While tools developed for closely related bacterial strains such as PanOCT showed significantly weaker agreement with others, tools designed for diverse datasets such as OrthoMCl and OrthoFinder demonstrated similar clustering behavior to that of panX (see panel D in Figure 4.4).



B - S. pneumoniae

	panX	Roary -i 70	Roary -i 90	PanOCT	OrthoMCL	OrthoFinder
panX	1	0.84	0.86	0.84	0.8	0.78
Roary -i 70	0.71	1	0.84	0.77	0.66	0.64
Roary -i 90	0.69	0.8	1	0.73	0.61	0.59
PanOCT	0.71	0.76	0.76	1	0.7	0.66
OrthoMCL	0.76	0.74	0.72	0.79	1	0.79
OrthoFinder	0.85	0.83	0.8	0.87	0.9	1

D - Prochlorococcus

	panX	Roary -i 30	Roary -i 50	PanOCT	OrthoMCL	OrthoFinder
panX	1	0.78	0.69	0.48	0.87	0.83
Roary -i 30	0.6	1	0.76	0.49	0.7	0.61
Roary -i 50	0.43	0.61	1	0.52	0.52	0.44
PanOCT	0.15	0.21	0.27	1	0.2	0.16
OrthoMCL	0.71	0.73	0.68	0.52	1	0.74
OrthoFinder	0.8	0.76	0.68	0.48	0.87	1

Figure 4.4: **Pan-genome statistics on real datasets:**
In panel A&C, the distribution of the number of strains was computed for the pan-genomes inferred from different methods. For *S. pneumoniae*, all 5 methods showed similar agreement on the number of core genes. Among them, panX, OrthoFinder, and OrthoMCL shared great similarity on the cluster size distribution and the total size of the pan-genome. Roary required an appropriate identity cutoff `-i 70` to agree with the other methods, whereas panOCT yielded many more singleton genes. For *Prochlorococcus*, panX, OrthoFinder and OrthoMCL showed high agreement on the cluster size distribution, the number of core genes and the total size of the pan-genome. While Roary needed a much lower identity cutoff `-i 30` to infer a comparable core genome, PanOCT detected only a small number of core genes and split many genes aggressively. Panels B&D exhibit the degree of congruence between each pair of clustering methods. The number in each row presents the fraction of clusters inferred by one tool, which perfectly accord with the clusters inferred by another tool. The same type of comparison on simulated data is demonstrated in Figure 4.10

4.2.3 *Benchmarking divide-and-conquer strategy*

In order to test the divide-and-conquer strategy, we generated a simulated pan-genome encompassing 500 strains based on exponentially distributed substitution rates with mean 0.06 per coalescent time scale. We analyzed this large pan-genome dataset by panX with divide-and-conquer, panX without divide-and-conquer and Roary. PanX correctly identified 2780 out of 2803 gene clusters (0.08% error rate), while panX yielded only 4 additional errors when switching off the divide-and-conquer strategy. By contrast, Roary generated 75 mis-clustered genes with parameter `-i 50` and 146 false gene clusters with parameter `-i 70`. This demonstrates that the divide-and-conquer strategy does not affect the clustering accuracy significantly (see Figure 4.5).

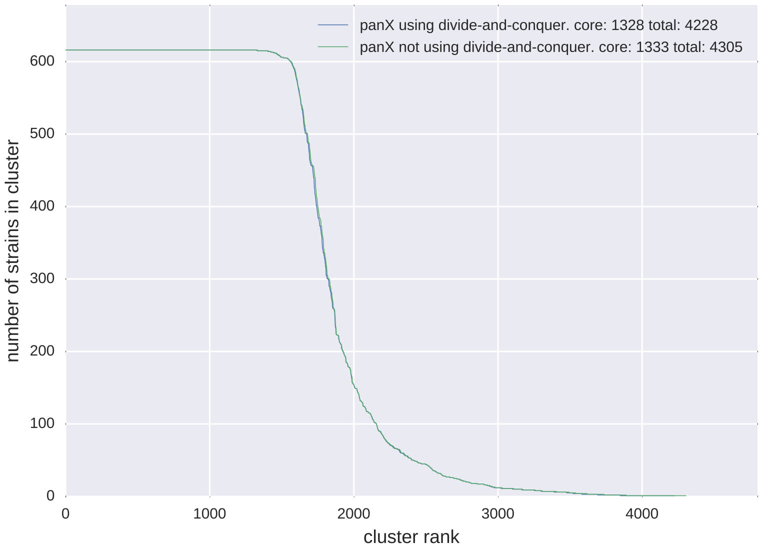


Figure 4.5: **Benchmarking divide-and-conquer strategy.** We analyzed the large dataset of 616 *S. pneumoniae* strains using panX with divide-and-conquer and panX without divide-and-conquer and found comparable clustering results.

4.2.4 *Testing clustering results on incomplete datasets*

Since draft genomes assembled from short reads are often incomplete and fragmented, it is essential to know how incomplete genomes may affect the performance of orthology inference. By design, panX infers orthologous gene clusters irrespective of neighboring genes, which avoids erroneous clustering when conserved gene neighborhood information is obscured by missing genes.

In order to explicitly investigate the effect of genome incompleteness on the accuracy of panX's ortholog clustering, we generated different incomplete datasets by deleting random genes from a random subset of genomes in the simulated datasets. In this procedure, 10%, 20%, 30% of genes from 10%, 20%, 30% of strains, respectively, were eliminated from the initial simulated datasets, which yielded three datasets with different degrees of incompleteness. Additionally, we generated another dataset where all strains missed 10% of genes. In all these incomplete datasets, the mis-clustering rate of panX is comparable to that of panX on complete data (see Figure 4.6), which demonstrates that genome incompleteness has negligible effect on panX's clustering accuracy when tested on the simulation data.

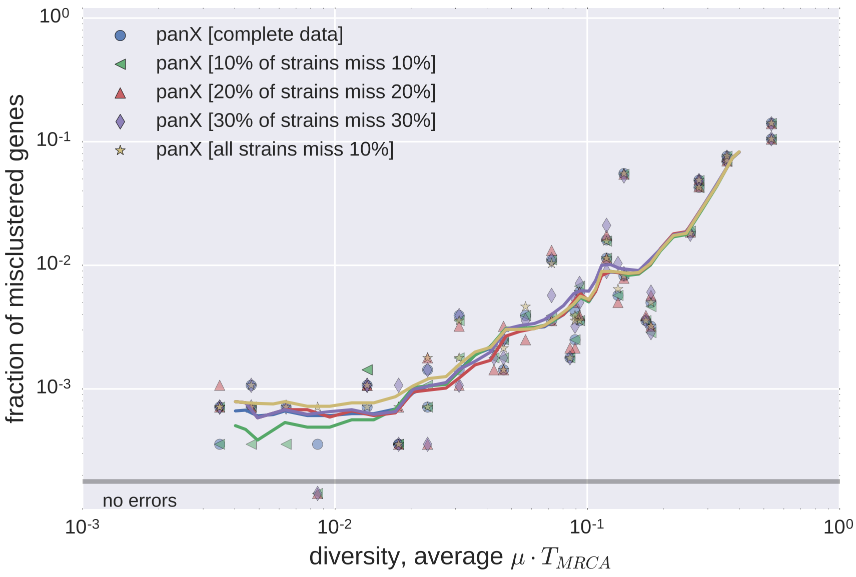


Figure 4.6: **Accuracy of clustering on different incomplete datasets.** We generated four incomplete simulation datasets with 10%, 20%, 30% of genes missing from 10%, 20%, 30% of strains, respectively, and with 10% of genes missing in all strains. The graph shows the sum of all clustering errors for different incomplete datasets as a function of the pangenome diversity, namely the sum of merge errors (clusters that contain extra genes), split errors (incomplete clusters) and merge/split errors (clusters that miss genes and contain additional genes), respectively.

4.3 CONCLUSION

To benchmark the clustering accuracies of 5 different methods on pan-genome analysis, we simulated 120 pan-genome datasets, with each containing 30 genomes built on the *E. coli* reference strain K-12. The simulation was well designed with consideration of different mutation rates, gene gain and loss, and horizontal transfer. Whereas other tools showed different performances on different diversities, panX worked well across a large range of diversities, due to its adaptive paralog splitting based on the core genome diversity during the phylogeny-aware post-processing.

On real datasets, we investigated the agreement of clustering results between different tools on a *S. pneumoniae* pan-genome and a *Prochlorococcus* pan-genome. Whereas the cluster size distributions built for the different tools were similar for the low diversity dataset of *S. pneumoniae* strains, more variation was demonstrated on the high diversity dataset of *Prochlorococcus* genomes and OrthoMCl, OrthoFinder and panX showed similar results to each other.

Moreover, we benchmarked the divide-and-conquer strategy and found comparable results between clustering analysis using and not using divide-and-conquer. In addition, tests for incomplete data showed that genome incompleteness did not reduce the panX's clustering performance on simulated datasets.

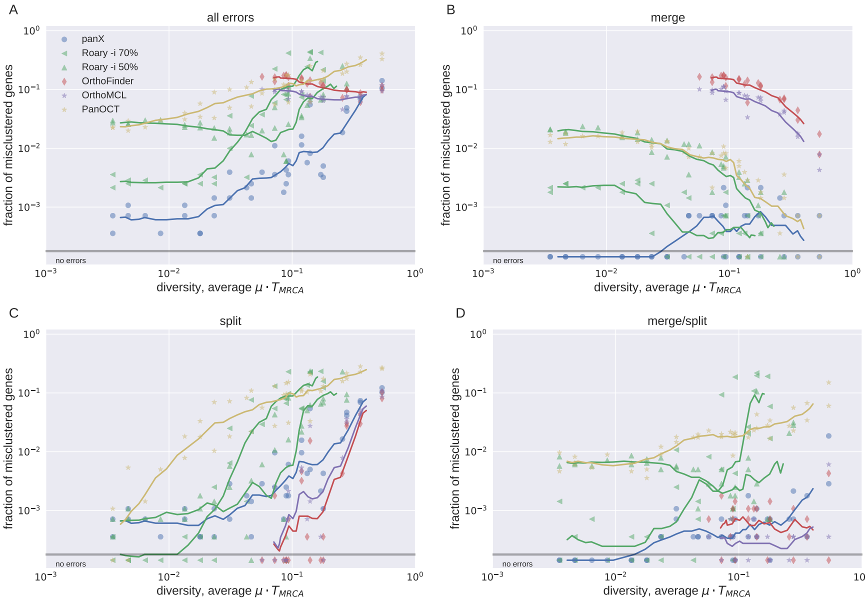


Figure 4.7: **Accuracy of clustering by different tools for frequent gene conversion rate.**

Panel A illustrates the sum of all three types of clustering errors for different tools as a function of the diversity of the pan-genome dataset.

Panels B-D show the fraction of clusters that contain extra genes (merge errors), incomplete clusters (split errors), and clusters that miss genes and contain additional genes (merge/split errors), respectively.

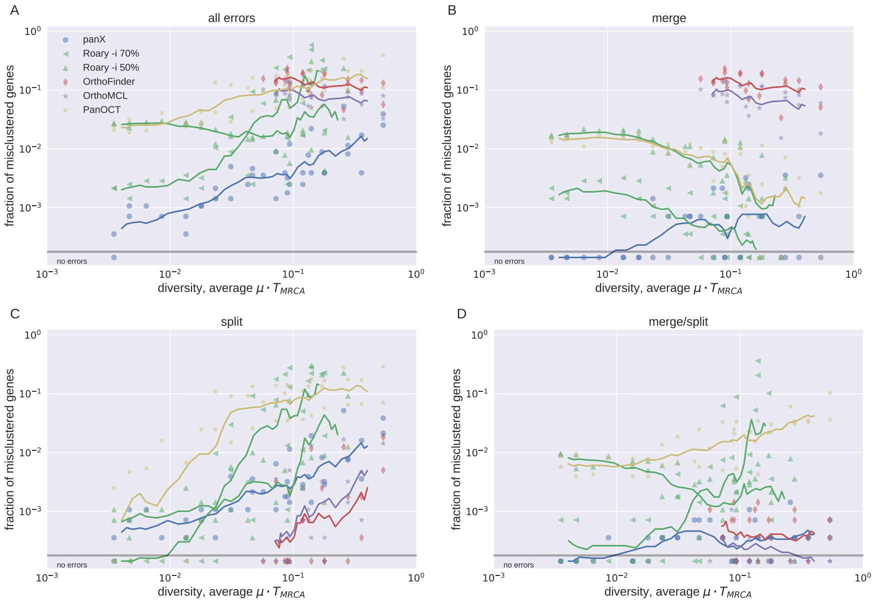


Figure 4.8: **Accuracy of clustering by different tools for occasional gene conversion rate.**
Panel A shows the sum of all clustering errors for different tools as a function of the pan-genome diversity.
Panels B-D show the fraction of clusters that contain extra genes (merge errors), incomplete clusters (split errors), and clusters that miss genes and contain additional genes (merge/split errors), respectively.

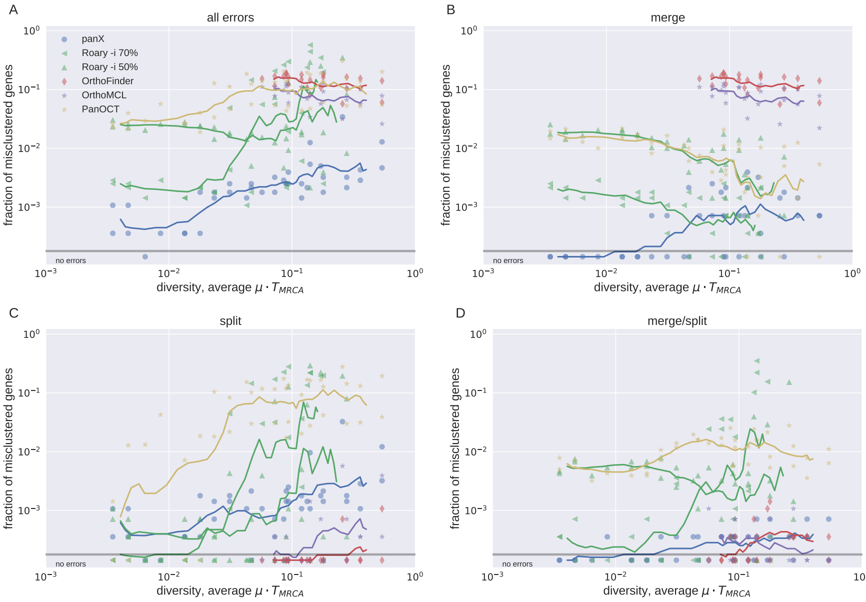
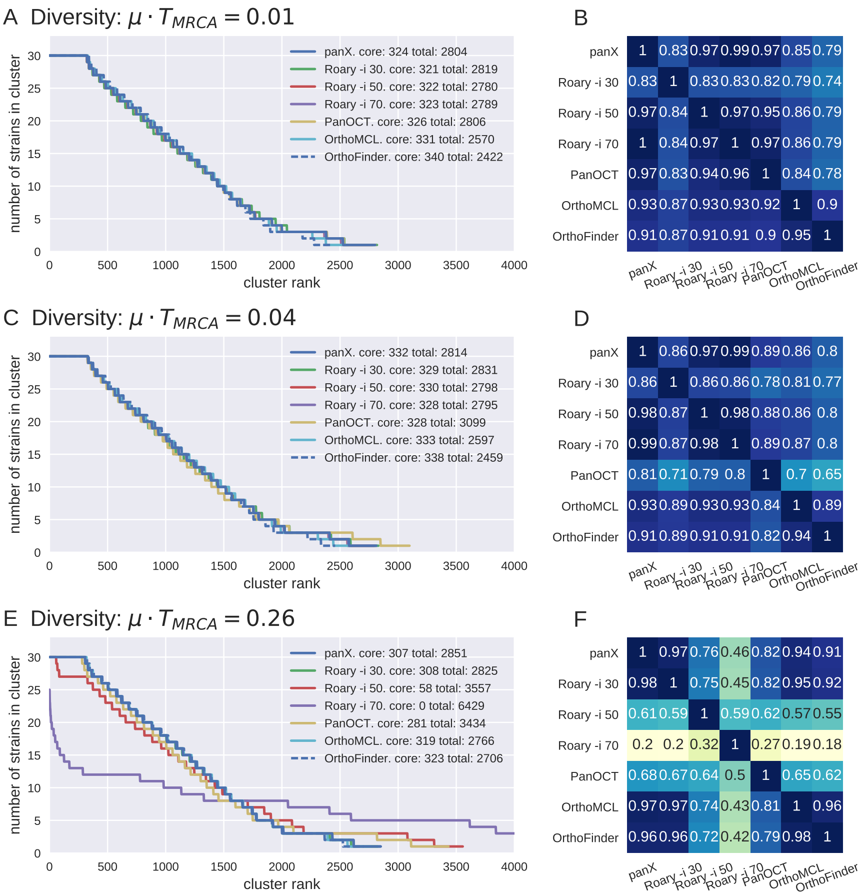


Figure 4.9: **Accuracy of clustering by different tools without gene conversion.**

Panel A shows the sum of all clustering errors for different tools as a function of the pan-genome diversity.

Panels B-D show the fraction of clusters that contain extra genes (merge errors), incomplete clusters (split errors), and clusters that miss genes and contain additional genes (merge/split errors), respectively.



B

	panX	Roary -i 30	Roary -i 50	Roary -i 70	PanOCT	OrthoMCL	OrthoFinder
panX	1	0.83	0.97	0.99	0.97	0.85	0.79
Roary -i 30	0.83	1	0.83	0.83	0.82	0.79	0.74
Roary -i 50	0.97	0.84	1	0.97	0.95	0.86	0.79
Roary -i 70	1	0.84	0.97	1	0.97	0.86	0.79
PanOCT	0.97	0.83	0.94	0.96	1	0.84	0.78
OrthoMCL	0.93	0.87	0.93	0.93	0.92	1	0.9
OrthoFinder	0.91	0.87	0.91	0.91	0.9	0.95	1

D

	panX	Roary -i 30	Roary -i 50	Roary -i 70	PanOCT	OrthoMCL	OrthoFinder
panX	1	0.86	0.97	0.99	0.89	0.86	0.8
Roary -i 30	0.86	1	0.86	0.86	0.78	0.81	0.77
Roary -i 50	0.98	0.87	1	0.98	0.88	0.86	0.8
Roary -i 70	0.99	0.87	0.98	1	0.89	0.87	0.8
PanOCT	0.81	0.71	0.79	0.8	1	0.7	0.65
OrthoMCL	0.93	0.89	0.93	0.93	0.84	1	0.89
OrthoFinder	0.91	0.89	0.91	0.91	0.82	0.94	1

F

	panX	Roary -i 30	Roary -i 50	Roary -i 70	PanOCT	OrthoMCL	OrthoFinder
panX	1	0.97	0.76	0.46	0.82	0.94	0.91
Roary -i 30	0.98	1	0.75	0.45	0.82	0.95	0.92
Roary -i 50	0.61	0.59	1	0.59	0.62	0.57	0.55
Roary -i 70	0.2	0.2	0.32	1	0.27	0.19	0.18
PanOCT	0.68	0.67	0.64	0.5	1	0.65	0.62
OrthoMCL	0.97	0.97	0.74	0.43	0.81	1	0.96
OrthoFinder	0.96	0.96	0.72	0.42	0.79	0.98	1

Figure 4.10: **Comparison of pan-genome inference from simulated datasets.**

Panel A,C&E show the inverse cumulative cluster size distribution for simulated datasets with different diversities. The gene clusters are identical, only the degree to which members of gene clusters are mutated is different between panels. While all tools inferred similar distributions for low diversity dataset, several tools estimated fewer core genes and more small gene clusters for high diversity dataset. The panels B,D&F show the fraction of clusters that are also identified by other tools.

5 PAN-GENOME VISUALIZATION

Most of the software [28, 78, 111, 112] have focused on computational aspects of pan-genome analysis, yet interactive visual exploration of the results still remains as a challenge. Moreover, various programs are required to generate summary statistics, alignments and phylogenies for gene clusters in the pan-genome, which necessitates substantial experiences in these analyses. In order to provide an easy-to-use and powerful platform for exploring the pan-genomic data, panX displays results in an interactive web-based visualization application and seamlessly connects gene clusters with their alignment and gene phylogenies. Using panX, user can readily inspect the evolutionary relationships among strains and their genes, as well as presence/absence patterns associated with metadata displayed on the core genome phylogeny. The dynamic dashboard offers a comprehensive view of the pan-genome analysis results and allows rapid high-dimensional filtering and searching for genes in large-scale datasets.

5.1 OVERVIEW OF PANX VISUALIZATION

The interactive pan-genome dashboard consists of four main sections and six interconnected visual components, whose links are illustrated in Figure 5.1 and details demonstrated in Figure 5.2. They pertain to: (1) pan-genome summary statistics on core and accessory genome, inverse cumulative distribution of clusters sizes and gene length distribution; (2) a gene cluster table and an alignment viewer for cluster sequences; (3) a comparative panel of core genome phylogeny and gene phylogeny; (4) a table for strain-specific metadata.

5.2 INTERACTIVE PAN-GENOME STATISTICAL CHARTS

The top section is comprised of three interactive pan-genome statistical charts, which are organized adjacent to the species selector containing a wide range of species. These summary charts are designed as an overview on the size of core and pan-genome and gene length distribution. They can be utilized to

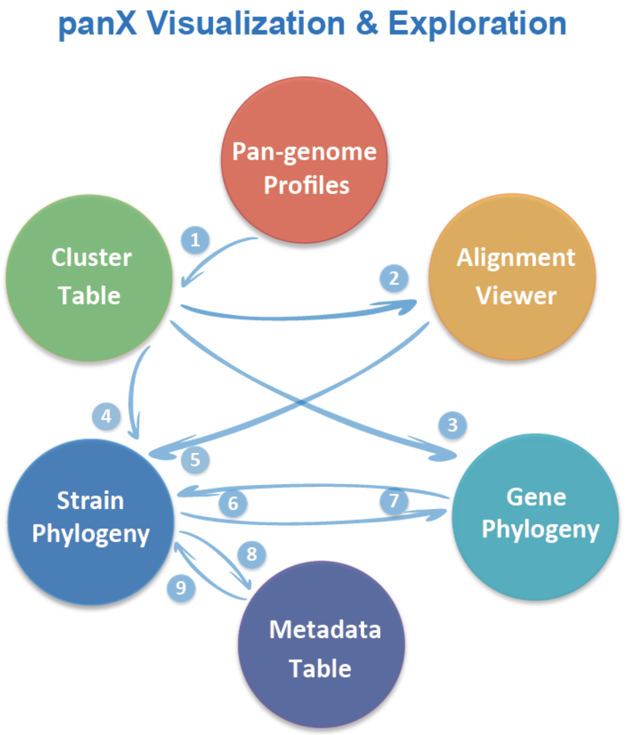


Figure 5.1: **PanX visualization components and their links:** (1) Charts with summary statistics allow rapid filtering and selection of gene subsets in the cluster table; Clicking a gene cluster in the cluster table loads the (2) corresponding alignment, (3) gene tree and (4) gene presence/absence and gain/loss pattern on strain tree; (5) Selecting sequences in the alignment highlights the associated strains on strain tree; (6)&(7) Strain tree interacts with gene tree in various ways; (8) Zooming into a clade on strain tree screens strains in metadata table; (9) Searching in metadata table displays strains pertinent to specific meta-information and highlights them on the strain tree.

rapidly select gene subsets based on a range of gene abundance and gene size.

The first pie chart depicts the relative proportion of core and accessory genome. Clicking core and accessory portion screens for the corresponding genes in other interlinked charts. Considering different core genome criteria used in the literature [51] , core genome can contain either strictly core genes present in all strains or soft core genes shared by the majority of strains (e.g. present in > 80% strains). Especially for datasets including incomplete genomes and potential outliers, applying an appropriate soft core cutoff helps determine a more realistic set of core genes. Based on this consideration, panX displays a slider bar of the core genome cutoff below the pie chart, which automatically updates the core genome composition based on user-triggered threshold change. This flexible feature allows easy visual comparison of core genome size among different core genome cutoffs.

The second line chart describes the number of strains shared in each gene cluster, which is sorted descendingly so that all core genes are arranged on the left of the curve. The flat region until the first drop presents the core genome. The gradual decline in the number of strains indicates common and rare accessory genes in the accessory genome. The remaining part shown as a tail covers strain-specific singleton genes. Using mouse to select a span of genes allows quick examination of a gene subset in core and accessory genome.

The third bar chart displays the distribution of average length of genes in each cluster, which can be used to select specific ranges of gene size and identify patterns of mis-clustering, especially when unclustered genes of identical length exhibit peaks in the distribution.

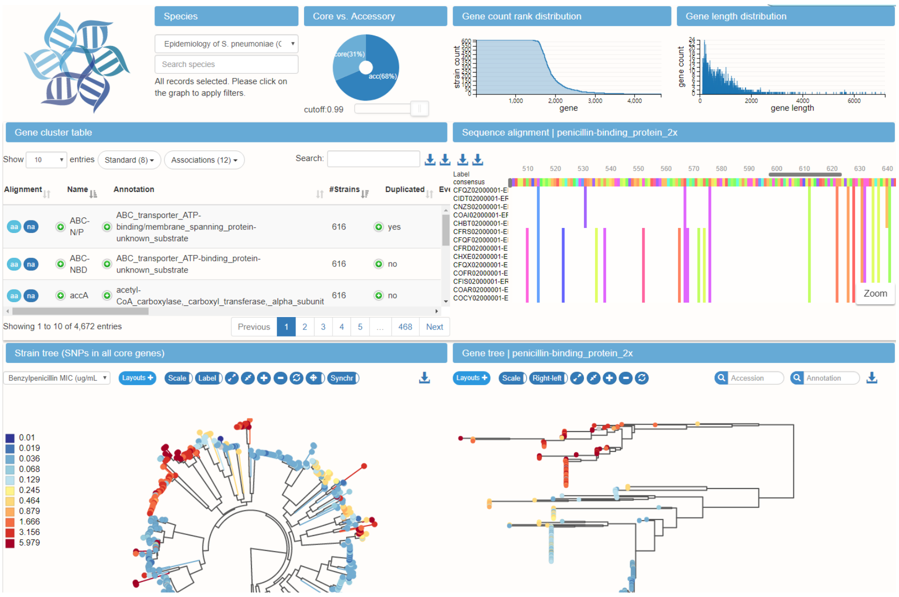
The filter of genes triggered by a single selection event in one chart simultaneously updates the genes in other charts, as well as the table of gene clusters below. In addition, all filtering actions can be switched back to the initial state such that user can flexibly explore combinations of different filtering settings.

5.3 RAPID GENE SEARCH IN GENE CLUSTER TABLE

The second section consists of a sort- and search-able table that contains all gene clusters and an alignment viewer that demonstrates the multiple sequence alignment of the selected gene cluster.

Gene cluster table

The gene cluster table allows rapid interrogation of large datasets. Various summary statistics presented in different columns can be readily investigated via flexible sort and search function. The statistics are organized into the following categories: the first part presents gene name and annotation that conveys fundamental information on the type and function of a gene; the second part covers different phylogenetic properties, including the number of gene gain and loss events, gene diversity and gene duplications; additionally, optional data columns such as links to expert database (e.g. The Pseudomonas Genome Database [9]) and other association statistics can be easily integrated based on user-customized setting. Particularly, when metadata such as drug concentration measurements are available, calculating scores on the degree to which branches of a gene tree are associated with the metadata can be eminently useful for the quick identification of drug resistance candidate genes.



Epidemiology of S. pneumoniae (C

Search species

All records selected. Please click on the graph to apply filters.

cutoff:0.99

Show 10 entries Standard (8) Associations (12) Search:

Alignment	Name	Annotation	#Strains	Duplicated	Ev
aa na	ABC-N/P	ABC_transporter_ATP-binding/membrane_spanning_protein-unknown_substrate	616	yes	
aa na	ABC-NBD	ABC_transporter_ATP-binding_protein-unknown_substrate	616	no	
aa na	accA	acetyl-CoA_carboxylase,_carboxyl_transferase,_alpha_subunit	616	no	

Showing 1 to 10 of 4,672 entries Previous 1 2 3 4 5 … 468 Next

Benzylpenicillin MIC (ug/mL

Figure 5.2: **A screenshot of the panX web-based visualization application:**
The top section offers a statistical characterization of the pan-genome and allows rapid filtering of gene clusters by gene abundance and gene length. The second section includes a searchable and sortable gene cluster table, which helps users to quickly select gene clusters by gene name, annotation, diversity, etc. Clicking an individual cluster in the cluster table loads the alignment of gene cluster into the alignment viewer beside the table and simultaneously updates the strain phylogeny and the gene phylogeny in the tree viewer at the bottom. Gene presence/absence patterns and gain/loss events of the selected gene cluster are mapped onto the core genome phylogeny. The core genome tree can also be colored by metadata such as year of isolation, sampling location, resistance phenotypes.

Especially for the columns including the expand buttons such as *Annotation*, panX exhibits the majority of annotations of all sequences in a gene cluster. The less frequent annotations can be inspected by the expand button to compare annotations from different tools. All annotations are searchable so that user can easily find out the cluster in which a gene belongs to. For example (`http://pangenome.tuebingen.mpg.de/Escherichia_coli`), searching *mcr-1* instantly highlights 11 out of 307 *E. coli* strains (from RefSeq database) that carry a mobile colistin resistance gene annotated as "phosphoethanolamine–lipid A transferase MCR-1". Searching *cmr* on `http://pangenome.tuebingen.mpg.de/Yersinia_pestis` shows that 32 strains are annotated as "multidrug transporter MdfA" without gene name, whereas 1 strain is annotated as "multidrug translocase" with gene name *cmr* on the column *Name*. In a similar way to *Annotation*, the column *Duplicated* denotes whether a gene cluster involves more than one gene per strain. Using the expand button lists strains carrying duplicated genes and the corresponding gene copy number.

Moreover, each row also includes triggers exhibiting the nucleotide or amino acid sequence alignment of the selected cluster via MSAViewer [109] from BioJs [35]. Clicking the trigger shows the majority of annotations on the top of the alignment and updates both phylogenetic trees by loading the gene presence and absence pattern on the core genome SNP phylogeny and the individual gene phylogeny described below.

Alignment viewer

With the aim of underlining sequence differences and illustrating gene haplotype groups as clearly as possible, panX reduces the redundancy in the alignment by keeping only consensus sequences and variable positions, see Figure 5.3. The "reduced" alignment can be flexibly magnified to show mutation details or minimized to demonstrate the haplotype structure. The corresponding original alignment can be fetched via the download button. The vertical and horizontal sliders display genes and positions in the alignment, respectively, in which user can click and highlight sequence of interest or specific position.

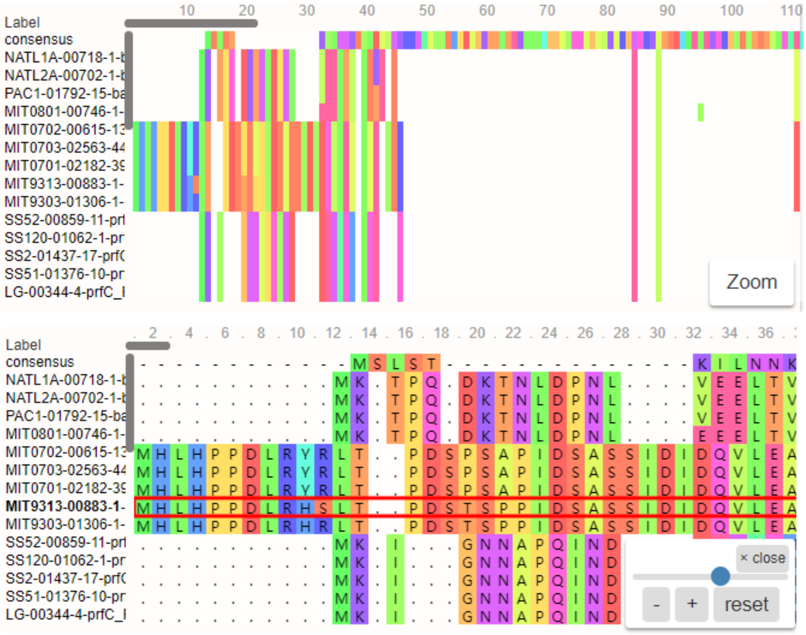


Figure 5.3: **Flexible zooming function in the alignment viewer** PanX displays a "reduced" alignment by default for emphasizing sequence variations and potential haplotype patterns. The first row in the alignment pertains to the consensus sequence. Mutations are highlighted in each gene. Users can flexibly zoom into base pair level or zoom out to investigate haplotype structure. Besides, sequence of interest or specific position can also be highlighted and the original "unreduced" alignment can be downloaded.

5.4 COMPARATIVE PHYLOGENETIC PANEL

The phylogenetic viewer contains a strain tree built on all core gene SNPs and a gene tree inferred from sequences of a gene cluster. The terminal nodes (strains) on the core genome SNP phylogeny can be colored by different meta-information, which facilitates the inspection on phylogenetic subgroups associated with metadata such as country, host, collection date and antibiotics resistance. The metadata can be either extracted from GenBank files or provided by user using a tab-separated value (TSV) file.

In order to enable easy visual comparison between core genome SNP phylogeny and gene phylogeny, panX interactively connects two trees in various ways. Touching the mouse to a leaf node of one tree simultaneously highlights the corresponding strain on the core genome tree and the gene carried by the selected strain on the gene tree, as well as gene duplications when multiple copies of a gene are present. Placing the mouse on the internal branch selects the sub-tree in both phylogenies and marks the corresponding nodes with different colors for each strain, which allows rapid visual inspection on the degree of congruence between the core genome tree and the gene tree and gene duplications in a gene cluster, see Figure 5.4. When user clicks an internal branch, the tree will be zoomed into a specific clade to reveal finer details of the sub-tree structure. Moreover, clicking on an inner branch on the strain tree updates the strain metadata table and enables easy interrogation of meta-information associated with selected strains.

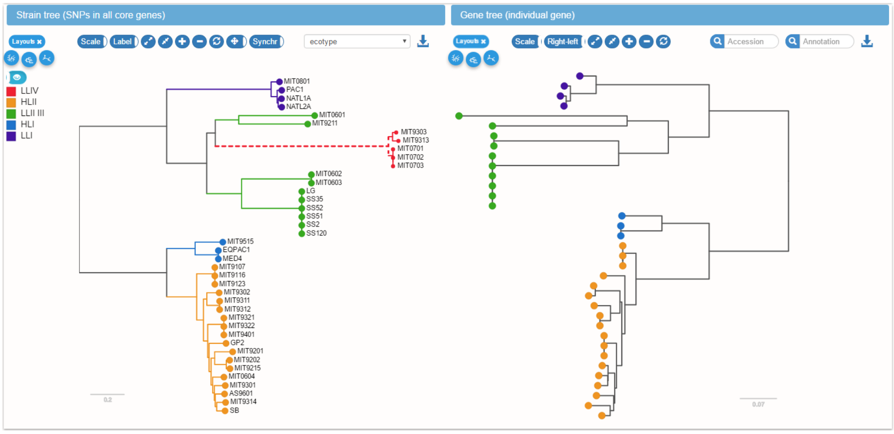


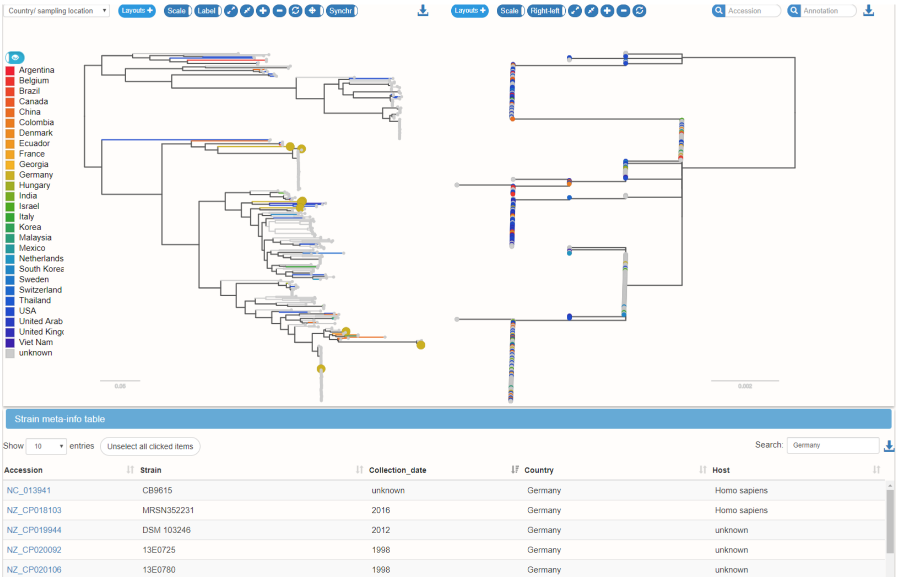
Figure 5.4: **Interconnected strain tree and gene trees.** The core genome tree shows the presence or absence of the selected gene in different strains. Placing the mouse on an internal node or branch in one of the trees highlights all strains in the corresponding clade in both trees. This provides users a rapid impression of phylogenetic discordance and likely gene gain and loss events. Gene gain and loss events are indicated by thickened lines or dashed lines, as shown on the clade *LLIV* in the strain tree. When there are gene duplications in the selected cluster, placing mouse over a leaf node in any of the trees highlights the corresponding strain on the core genome phylogeny and all gene copies on the gene phylogeny.

Genes present or absent in some strains are marked as blue and gray, respectively. The visualization of gene presence/absence patterns facilitates the investigation of genes associated with specific phenotypes. Whereas some gene trees are consistent with the species tree, others might differ significantly from the species tree. The inconsistencies between the species tree and the gene tree could indicate the dynamics of gene gain and loss. Based on the ancestral reconstruction algorithm [85], the most probable gene loss and gain events are precomputed and highlighted on the strain tree by dashed or thickened lines, respectively.

Moreover, mutations in the nucleotide or amino acid sequences of a gene cluster are mapped onto the gene tree, which can be easily examined by the tooltips linked to the branches.

5.5 METADATA INTEGRATION AND INTERROGATION

The core genome SNP tree and the metadata table are mutually connected (see Figure 5.5). Searching meta-information such as the age range of patients or specific niche filters the metadata table for displaying strains paired with the interrogated item and simultaneously highlights them on the strain tree. Correspondingly, selecting a clade by clicking on an internal branch on the strain tree updates the metadata table to show the strains only presented in that clade. For example, on `http://pangenome.tuebingen.mpg.de/P_marinus_meta`, searching *HL* in the metadata table displays all strains living in the ecotypes associated with high light (both HLI and HLII), whereas moving the mouse over an item listed in the metadata legend on the core tree highlights either only HLI strains or only HLII strains. Hence, it helps easily pinpoint the locations of the strains with similar metadata on the core genome phylogeny. Moreover, since the drop-down menu for metadata on the strain tree viewer demonstrates only one type of metadata at a time, the metadata table listing all metadata categories is designed to be complementary so that searching host information in the table can display the related strains and other associated metadata such as pathogenicity. This allows interrogating multiple metadata at the same time. Additionally, metadata can be downloaded as a TSV table.



Strain meta-info table

Show 10 entries Unselect all clicked items

Search: Germany

Accession	Strain	Collection_date	Country	Host
NC_013941	CB9615	unknown	Germany	Homo sapiens
NZ_CP018103	MRSN352231	2016	Germany	Homo sapiens
NZ_CP019944	DSM 103246	2012	Germany	unknown
NZ_CP020092	13E0725	1998	Germany	unknown
NZ_CP020106	13E0780	1998	Germany	unknown

Figure 5.5: **Interactions between strain tree and metadata table** Metadata associated with different strains can be used as coloring options on the core genome tree. The metadata legend on the core genome phylogeny presents one type of metadata at a time. Clicking an internal branch on the core tree triggers the zooming into a specific clade and simultaneously updates the strain metadata table, which allows easy interrogation of meta-information associated with selected strains. Metadata table is interlinked with the core genome tree such that selecting strains in the table highlights corresponding strains in the phylogeny. As a complementary functionality to the selection of single metadata type on the core tree, metadata table displays entire metadata categories and can be used to compare multiple metadata types among different strains. For example, searching country Germany on the metadata table highlights all isolates collected in Germany. One can further inspect the collection date or other meta-information associated with these strains.

5.6 CONCLUSION

In order to gain deeper insights into the evolution of prokaryote genomes, one often needs to visually explore high dimensional data space to identify intriguing genetic patterns. PanX is designed to allow such an interactive exploration for a large-scale collection of bacterial genome data.

PanX integrates a wealth of summary statistics from the pan-genome computational pipeline and enables users to flexibly select interesting gene sets or rapidly search for genes. Each gene cluster is seamlessly linked to its corresponding alignment and gene phylogeny, which allows an in-depth analysis with gene gain and loss events and mutations integrated into the core genome tree and the gene tree, respectively. The panX phylogenetic panel facilitates easy visual comparison between the core genome tree and the gene tree. The details of tree incongruence offer important clues about the dynamics of horizontal transfer, gene duplication and loss, from which we can further conduct downstream analyses to infer the evolutionary histories of genes of interests.

Furthermore, incorporating genome sequences and epidemiological metadata into a hospital surveillance system helps monitor the transmission of pathogenic bacteria and enhance outbreak detection. Flexible exploration features for evolutionary relationships among strains and their genes, as implemented in the panX visualization application, would facilitate the identification of mutations or presence/absence patterns of genes associated with phenotypic traits and clinical manifestation.

When rich metadata such as co-morbidities, pathogenicity, sampling date and location are provided, panX serves as a powerful platform for investigating how pathogens adapt and spread. There are already similar approaches showing great utility for tracking the spread and evolution of seasonal influenza virus or Ebola virus in a recent outbreak [33, 76]. Even though well-annotated metadata records associated with genome sequences is a rarity, we exhibited an application case based on the exceptional dataset of 616 *S. pneumoniae* strains generated by Croucher et al. [14].

Moreover, the construction of a pan-genome encompassing a wide range of drug-resistant strains and clinically relevant antibiotic resistant genes can provide an extremely useful resource for deeper understanding on the mechanisms of antibiotic resistance. Comparing new sequences from nosocomial isolates against such a pan-genome database might contribute to effective evaluation of potential risks or even help predict resistance emergence.

6 APPLICATIONS

In order to provide comprehensive summary statistics on pan-genomes of different bacteria, we selected complete genomes from NCBI Reference Sequence Database and conducted the panX analysis on 93 microbial species and 3 bacterial orders including Pseudomonadales, Enterobacteriales and Vibrionales. The results are publicly explorable and downloadable via http://pangenome.de. Detailed descriptions on the panX analysis pipeline and visualization are documented in repository *pan-genome-analysis* and *pan-genome-visualization* on https://github.com/neherlab/, respectively. Although the collection of genomes accompanied by rich and carefully crafted metadata is considerably rare, we showcase one application in which panX aids in finding genes associated with antibiotic resistance based on 616 genome sequences of Streptococcus pneumoniae isolates in an epidemiological study from Croucher et al. [14]. Moreover, we applied panX analysis pipeline on a large dataset of 1524 Pseudomonas isolates associated with Arabidopsis thaliana populations and found significant variability of gene content in the mainly pathogenic Pseudomonas syringae species complex.

6.1 PAN-GENOMES OF COMMON BACTERIAL GROUPS

We applied panX analysis on collections of complete genomes of bacterial species which has at least 10 genomes in RefSeq database. This yields 93 pan-genomes involving many human pathogens. A subset of statistics is presented in Table 6.1. The core genomes of most species listed in the table have considerably low diversity with nucleotide differences of merely a few percent or even less than 1.0×10^{-3}, as shown in *Bacillus anthracis* (1.0×10^{-4}), *Bordetella pertussis* (4.1×10^{-6}), *Mycobacterium tuberculosis* (2.0×10^{-4}), *Yersinia pestis* (1.0×10^{-4}). Even with similar number of strains, the proportion of core genes and singletons can be substantially different. In the collection of 68 *Chlamydia trachomatis* strains from multiple countries (average genome size about 1 Mb), 82.7% of genes are strictly core genes and 0.01% are singletons, while in the dataset of 70 *Pseudomonas aeruginosa* strains (average genome size about 6 Mb), core genes constitute only 25.5% of the pan genome and singletons up to 25.0%. *Chlamydia trachomatis*, as an obligate intracellular pathogen, is ecologically isolated and much less likely to expose to foreign genetic materials, which suggests a closed pan-genome [71] as also shown in the *Bacillus anthracis* pan-genome.

Moreover, the median core genome size is about 1800, whereas the median pan-genome size is around 5000 genes. Besides, not surprisingly, the most abundant photosynthetic organism Prochlorococcus has a relatively smaller core genome and a larger pan-genome, with the variable part of the pan-genome (accessory genome together with unique genes) being four-fold larger than the strict core genome, which implies important roles the accessory genome plays in the distinct physiological functions for such a widespread diverse bacterial group.

To facilitate the downstream analysis on these pan-genomes, several downloading options are made available: the alignments of core genes and entire genes can be downloaded by the down-arrow buttons on the gene cluster table. Alignments, gene tree for individual cluster and the core genome SNP tree can be downloaded via buttons on the alignment viewer and tree viewers, respectively. All buttons are associated with tooltips that explain the content. Besides the downloading options for gene clusters, their summary statistics generated by the panX analysis pipeline can be easily reloaded and processed in downstream analysis.

6.2 PAN-GENOME OF A DIVERSE BACTERIAL GROUP

While the majority of these datasets is collected among closely related genomes, we also added a diverse set of *Prochlorococcus* genomes. Prochlorococcus, a dominant oxygenic phototrophic cyanobacterium in tropical and subtropical oceans, can survive at a variety of depths under different high- and low-light intensities and exhibits distinct physiological and genetic characteristics [74].

We re-annotated 40 *Prochlorococcus* genome sequences [5] by Prokka [90] based on a customized database containing the following annotated *Prochlorococcus* strains $CCMP1375$, $MED4$, $MIT9313$, $NATL2A$, $MIT9312$, $AS9601$, $MIT9515$, $NATL1A$, $MIT9303$, $MIT9301$, $MIT9215$ and $MIT9211$. *Prochlorococcus* is much more diverse than the other species we have investigated,

Table 6.1: **Summary statistics of a subset of pan-genomes available at `pangenome.de`.** Diversity calculation: average number of pairwise differences per nucleotide in core gene alignments.

Species	genomes	core genes	all genes	singletons	diversity
Acinetobacter baumannii	71	1701	8334	1558	0.010
Bacillus anthracis	43	4156	5980	62	$1.0e-04$
Bacillus cereus	36	2979	13364	3486	0.048
Bordetella pertussis	291	2437	3743	158	$4.1e-06$
Burkholderia pseudomallei	59	4098	11580	1966	0.003
Campylobacter jejuni	113	935	3166	526	0.014
Chlamydia trachomatis	68	809	978	12	0.005
Clostridium botulinum	23	795	9083	2294	0.147
Corynebacterium pseudotuberculosis	59	1133	2316	65	0.005
Enterobacter cloacae	22	2971	10783	3211	0.087
Escherichia coli	307	778	23107	6339	0.015
Francisella tularensis	35	838	2339	302	0.007
Helicobacter pylori	85	694	2371	328	0.042
Klebsiella pneumoniae	109	2545	15978	4004	0.007
Listeria monocytogenes	95	1907	4947	485	0.031
Mycobacterium tuberculosis	51	2665	4350	93	$2.0e-04$
Neisseria meningitidis	78	1071	3375	426	0.015
Prochlorococcus marinus	40	1047	5407	1262	0.291
Pseudomonas aeruginosa	70	3264	12768	3195	0.006
Salmonella enterica	260	1327	15521	3996	0.009
Staphylococcus aureus	146	1229	5206	731	0.008
Streptococcus pneumoniae	33	1188	3361	540	0.010
Streptococcus pyogenes	50	970	2856	341	0.008
Vibrio cholerae	28	2412	5156	771	0.005
Xanthomonas citri	26	3385	5261	291	0.001
Yersinia pestis	33	2557	4587	172	$1.0e-04$
Pseudomonadales	119	966	42520	20577	0.194
Enterobacteriales	33	1998	16413	6988	0.112
Vibrionales	66	716	30461	15643	0.193

see Table 6.1, which presents a challenging case in pan-genome analysis.

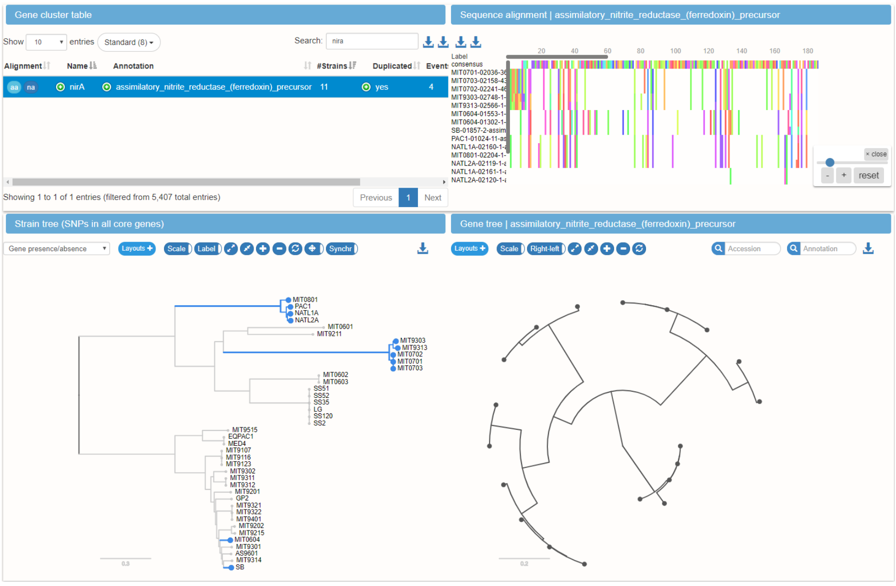
Event though it has been found that the difference among the 16S ribosomal RNA sequences of all 40 *Prochlorococcus* strains is no more than around 3%, *Prochlorococcus* can be separated into various ecotypes and shows significant differences in GC content and genome size [6]. These ecotypes are associated with *Prochlorococcus* populations that are adapted to high- and low-light intensities and can be readily explored on the species and gene trees via the panX visualization application.

Whereas gene losses seem to be strong selective forces that have streamlined the genomes of *Prochlorococcus* [94], gene gains and duplications have been frequently found in all *Prochlorococcus* lineages. Examples for gene acquisition and duplication are gene *nirA* (Figure 6.1) coding for "assimilatory nitrite reductase (ferredoxin) precursor" [4] and gene *psbA* (Figure 6.2) annotated as "Photosystem II reaction center D1" [68], which has been duplicated and involved in phage-mediated gene transfer. The ancestral insertions of these genes are placed on the species tree in panX and can be easily investigated in the strain phylogeny when searching for the corresponding gene name and selecting gene presence/absence as metadata, see `http://pangenome.tuebingen.mpg.de/P_marinus_meta`.

Moreover, genes can also be flexibly searched via their annotations. In addition, one can quickly find frequently gained and lost genes by sorting the table header *Events* (the number of gene gain and loss events), duplicated genes via the header *Duplicated* or diverse gene clusters by the header *Diversity*. When gene name is missing from the annotation or discordant gene annotations are present, an alternative solution provided by panX is to search locus_tag available from GenBank files. For example, searching locus_tag *MIT*9515_00842 shows the gene coding for ferredoxin-sulfite reductase in the strain *MIT*9515.

Furthermore, user can search all genes that are present at least in a particular clade. For example, searching the string of strain names from the highlight clade [1] displays about 150 genes present at least in all highlight-associated strains (both HLI and HLII). When combined with the range selection functionality on the gene count rank distribution to select gene clusters shared in approximately 21 strains, the gene cluster table

[1] *MIT*9515 *EQPAC*1 *MED*4 *MIT*9107 *MIT*9116 *MIT*9123 *MIT*9302 *MIT*9311 *MIT*9312 *MIT*9201 *GP*2 *MIT*9321 *MIT*9322 *MIT*9401 *MIT*9202 *MIT*9215 *MIT*0604 *MIT*9301 *AS*9601 *MIT*9314 SB



Gene cluster table

Show 10 entries Standard (8) Search: nira

Alignment	Name	Annotation	#Strains	Duplicated	Events
aa na	nirA	assimilatory_nitrite_reductase_(ferredoxin)_precursor	11	yes	4

Showing 1 to 1 of 1 entries (filtered from 5,407 total entries)

Sequence alignment | assimilatory_nitrite_reductase_(ferredoxin)_precursor

Strain tree (SNPs in all core genes)

Gene tree | assimilatory_nitrite_reductase_(ferredoxin)_precursor

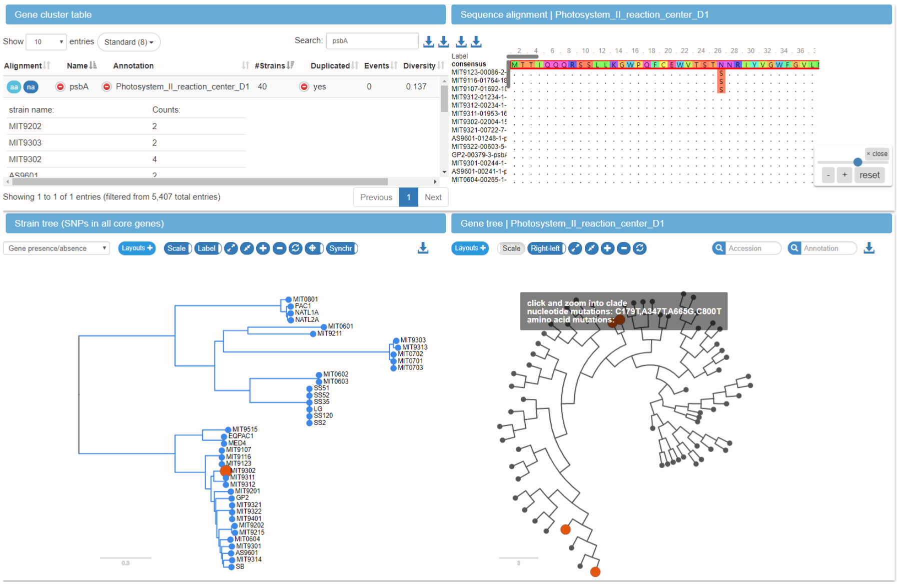
Figure 6.1: **Visualization on the gain of gene nirA in *Prochlorococcus***

As demonstrated on http://pangenome.tuebingen.mpg.de/P_marinus_meta, searching nirA in the gene cluster table quickly displays the cluster for nitrate assimilation gene *nirA*, with its gene gain events highlighted on the core genome phylogeny. Multiple gene gains and the ancestral insertions of these genes can be readily observed in different lineages, which are denoted by the blue lines. Many of these genes have also duplications, which can be easily explored via mouse-hovering over the leaves on species tree or gene tree for showing the corresponding gene copies.

exhibits immediately 100 genes present only in these strains (e.g. Phosphomethylpyrimidine kinase).

6.3 pan-genomes of bacterial orders

We also conducted pan-genome analysis on a wide range of genomes that are presented in three bacterial orders *Pseudomondales*, *Enterobacteriales* and *Vibrionales* from the RefSeq database. For the sake of avoiding over-representing well-sampled human pathogenic species, we selected no more than 10 genomes for each species according to the species assignment recorded in the RefSeq GenBank files for each of the above-mentioned orders. The panX analysis using default parameters on these collections yielded around 1000 core genes for *Pseudomondales*



Gene cluster table

Show 10 entries Standard (8) Search: psbA

Alignment	Name	Annotation	#Strains	Duplicated	Events	Diversity
aa na	psbA	Photosystem_II_reaction_center_D1	40	yes	0	0.137

strain name:	Counts:
MIT9202	2
MIT9303	2
MIT9302	4
AS9601	2

Showing 1 to 1 of 1 entries (filtered from 5,407 total entries)

Sequence alignment | Photosystem_II_reaction_center_D1

Strain tree (SNPs in all core genes)

Gene presence/absence

Gene tree | Photosystem_II_reaction_center_D1

click and zoom into clade
nucleotide mutations: C179T,A347T,A665G,C800T
amino acid mutations:

Figure 6.2: **Visualization on the duplication of gene psbA in *Prochlorococcus***
Searching psbA in the gene cluster table rapidly exhibits the cluster for photosynthesis gene *psbA*, with its gene duplications highlighted on the corresponding gene phylogeny. The details on which strains carry duplicated genes and the number of copies can be easily inspected by expanding the column *Duplicated* in the cluster table. Clicking a specific row in the nested sub-table for duplication (e.g. "$MIT9302$ 4") updates both phylogenies by highlighting the selected strain $MIT9302$ on the species tree and the corresponding four gene copies on the gene tree. Besides, mouse-hovering over an internal branch displays the mutations among genes. As shown in the tooltip on the gene tree, there are four synonymous mutations between the two investigated gene copies. The single base-pair substitutions and their positions are demonstrated and can be further inspected in multiple sequence alignment.

and *Vibrionales*, and about 2000 core genes for the smaller collection of *Enterobacteriales*. Whereas core genes of *Enterobacteriales* were approximately 11% diverged from each other, the core genome of *Pseudomondales* and *Vibrionales* were more diverse (about 20%) . While the core genome SNP tree of the *Vibrionales* clearly shows the groups of genomes that are in accord with the assigned species from RefSeq GenBank files, the core genome trees of the other two orders display considerable mixture of species.

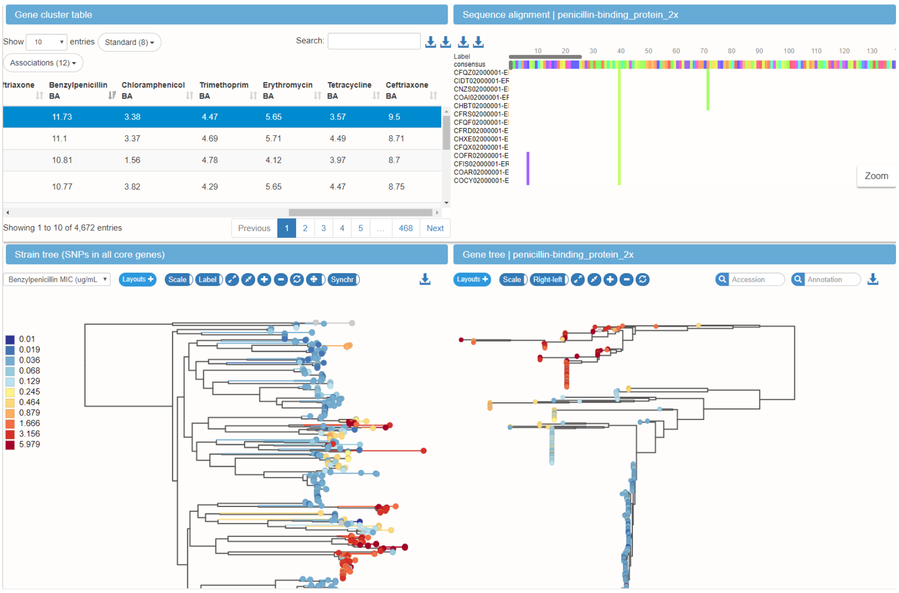
6.4 LARGE STREPTOCOCCUS PNEUMONIAE PAN-GENOME

Pan-genome of 616 *epidemiology-related Streptococcus pneumoniae isolates*

When combined with rich metadata, the utility of interactive pan-genome exploration is most pronounced. Notwithstanding the fact that the meta-information recorded for genome collections is scarce or patchy, there exists an ideal dataset produced by Croucher et al. [14] in their epidemiological study after the introduction of pneumococcal vaccine. This dataset encompasses 616 annotated genomes of Streptococcus pneumoniae isolates and a wealth of metadata such as levels of susceptibility/resistance to different antibiotics and patient age.

Even for this large collection of genomes, the panX interactive visualization application performs smoothly and rapidly. All orthologous clusters in the *S. pneumoniae* pan-genome can be quickly searched and sorted according to gene name, annotation, sequence diversity and the number of gene gain and loss events. The population structure of the collected strains is illustrated on the core genome SNP tree, which can be readily colored with various metadata including serotype, sequence cluster, antibiotic concentrations. All gene phylogenies are integrated in the visualization platform such that user can easily inspect the evolutionary relationships among genes of interests. Besides, metadata associated with specific strains can be interrogated in the metadata table. Moreover, panX pre-computed scores on branch association and presence/absence association between gene clusters and antibiotic concentration and made them available in the gene cluster table, which helps facilitate the investigation of drug resistance candidate genes.

For example, sorting Benzylpenicillin BA (branch association) in descending order immediately shows that the highest corre-



Gene cluster table

Show 10 entries Standard (8) Associations (12)

Search:

ftriaxone	Benzylpenicillin BA	Chloramphenicol BA	Trimethoprim BA	Erythromycin BA	Tetracycline BA	Ceftriaxone BA
	11.73	3.38	4.47	5.65	3.57	9.5
	11.1	3.37	4.69	5.71	4.49	8.71
	10.81	1.56	4.78	4.12	3.97	8.7
	10.77	3.82	4.29	5.65	4.47	8.75

Showing 1 to 10 of 4,672 entries

Sequence alignment | penicillin-binding_protein_2x

Label
consensus
CFQZ02000001-E
CIDT02000001-EF
CNZS02000001-E
COAI02000001-EF
CHBT02000001-E
CFRS02000001-E
CFQF02000001-E
CFRD02000001-E
CHXE02000001-E
CFQX02000001-E
COFR02000001-E
CFIS02000001-EF
COAR02000001-E
COCY02000001-E

Strain tree (SNPs in all core genes)

Benzylpenicillin MIC (ug/mL

Gene tree | penicillin-binding_protein_2x

Figure 6.3: **Using association score to identity genes associated with drug resistance**
For the dataset of 616 S. pneumonia strains accompanied by rich metadata, we computed branch association scores and presence/absence association scores for different types of drugs. Sorting the branch association score for benzylpenicillin minimum inhibitory concentration (MIC) shows that the cluster having the highest branch association with benzylpenicillin MIC refers to the penicillin binding protein *pbp2x*. This is confirmed using the coloring of the gene tree by the benzylpenicillin MIC. The resistant and susceptible isolates are grouped into two distinct clades separated by a large number of amino acid substitutions. Whereas resistant strains are scattered across the species tree, they form a single clade in the tree of *pbp2x*.

sponding branch association score is calculated for the gene cluster *pbp2x* annotated with penicillin-binding protein 2x (see Figure 6.3), which has been reported as an important resistance determinant for beta-lactam antibiotics [38]. Selecting Benzylpenicillin MIC (minimum inhibitory concentration) in the menu of metadata coloring readily displays the strains with different concentration levels that are scattered on the core genome tree and demonstrates the clear stratification pattern of protein sequences on the gene tree, which are partitioned into two distinct groups related to relatively high and low antibiotic concentration, respectively (see Figure 6.3). Mouse-hovering over the branch that separates strains associated with antibiotic susceptibility/resistance demonstrates plenty of amino acid mutations and their corresponding positions on the tooltip. In a similar manner, sorting Erythromycin PA (presence/absence association) exhibits that the strongest associated gene cluster is *mefE* annotated as Mef efflux pump protein, which has been shown to be related to erythromycin resistance [95].

Based on careful inspection, there are false positives involved in the list of clusters exhibiting relatively high association scores (discussed in Chapter 3). Nevertheless, the primary goal of this association approach is to narrow down thousands of gene clusters into a short list of candidates such that users can easily dig into the details of the alignments and trees linked to the candidate genes and conduct downstream analyses for a further validation. When results from more sophisticated methods are available, they can also be readily integrated into the gene cluster table and easily explored via the rapid sorting function for identifying genes associated with different phenotypes.

Besides, a significant portion of the gene clusters showing strong association is annotated as hypothetical proteins, which raises the question whether the unknown candidates play actually crucial biological roles that have not yet been properly recognized by the existing annotation programs. In another study on the genes associated with pathogenicity in Pseudomonas syringae isolates (unpublished data), at least some hypothetical candidate genes are aligned considerably well (>90% sequence identity, >90% alignment length) to sequences in NCBI nr (non-redundant) protein database, suggesting that more detailed annotation found there would be used to denote their potential biological functions.

6.5 large pseudomonas pan-genome

Pan-genome of 1524 *plant-associated Pseudomonas isolates*

We have also conducted the panX analysis on a large collection of newly sequenced *Pseudomonas* genomes produced by TL Karasov et al. (in submission). *Pseudomonas* is among the most abundant genera in plant populations including the model plant Arabidopsis thaliana. This genus is composed of species that are well-known crop pathogens (such as P. syringae and P. viridiflava belonging to the Pseudomonas syringae species complex), as well as species that can act as biocontrol agents. In microbiome studies of A. thaliana populations in Southwestern Germany conducted by Karasov, a single operational taxon unit

(OTU) was found to be both abundant across populations and persistent within populations, accounting for more than 40% of the endophytic Pseudomonas. Taxonomical assignment of the OTU predicted the OTU to belong to P. viridiflava, a pathogenic species, though taxonomical resolution of the microbiome analysis was low. It was unclear whether A. thaliana populations were colonized by single semi-clonal expansions of Pseudomonas strains, or the parallel expansions of different strains that belonged to the same OTU.

To determine the composition of strains that classify as the same OTU, Karasov collected 1524 Pseudomonas strains from Southwestern Germany during winter 2015 and spring 2016. We applied pan-genome analysis on this large collection of sequenced isolates via the panX pipeline. Functional assays determined that the P. syringae complex in which P. viridiflava was found (purple) was composed of primarily plant-pathogenic strains, whereas the isolates outside of this species complex (blue) did not significantly influence plant growth. Grouping isolate genomes that shared 99.99% sequence identity throughout the concatenated core genome determined abundant strains.

As shown in Figure 6.4, the abundance for each strain within one population in December 2015 and the same plant population in April 2016 is denoted by circle radius for a given strain. The phylogenies of isolates collected in each of 20 different plant individuals indicate that plants are colonized by several lineages, though the P. syringae lineages show putative in plant expansions. Although the presence/absence (dark blue/light blue respectively) pattern for the 25,000 most abundant orthology gene clusters across isolates shows that gene content varies within the P. syringae species complex, a single sequence homolog of the effector AvrE, known to be important for survival in the leaf, is broadly conserved among the most abundant and persistent lineages.

These results suggest that the OTU found to be most abundant across A. thaliana populations is composed of dozens of distinct pathogenic P. viridiflava strains, which might have colonized in parallel. Furthermore, single strains seem to dominate within a plant, whereas the identity of the dominant strain differs across plants. Intriguingly, we have observed evident differences of gene content between these P. viridiflava strains. These primary findings pose new questions that we will address in further analyses, with the main focus on the genetic variation of the pathogenic complex. In light of pan-genome information, we will attempt to identify patterns between the proliferation of specific strains and their gene content and detect orthology clusters that are associated with the success of these lineages.

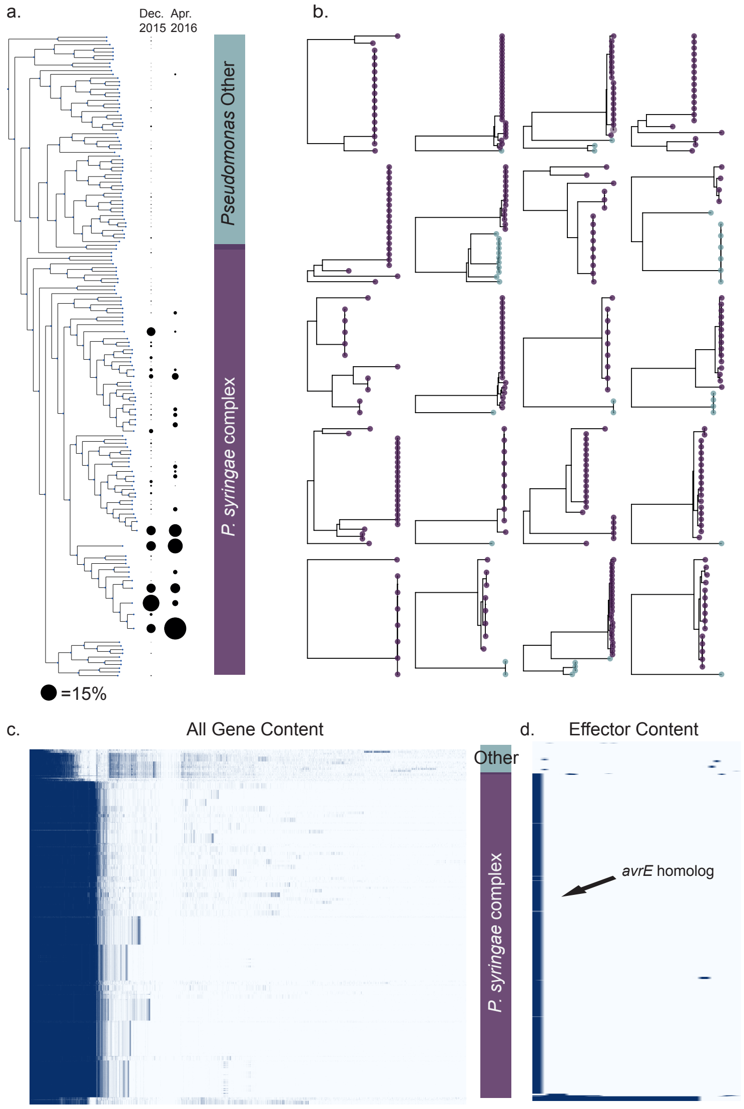


Figure 6.4: **Pan-genome analysis of** 1524 **Pseudomonas isolates collected from A. thaliana populations**
Isolates collected during winter 2015 and spring 2016 in Southwestern Germany were sequenced and compared using the panX pipeline. Abundant strains were identified by clustering together genomes that shared 99.99% sequence identity throughout the concatenated core genome (a) illustrates the abundance of each of the identified strains within one population in Dec. 2015 and April 2016. (b) presents the phylogenies of isolates collected per plant in each of 20 different plants. (c) shows that gene content varies within the P. syringae species complex. (d) reveals that, however, a single effector implicated in strain pathogenicity is conserved among the most abundant and persistent lineages.

7 FUTURE DIRECTIONS

In this doctoral work, we have implemented an efficient pan-genome analysis pipeline that works well across a large range of diversities based on an adaptive post-processing procedure and harnessed a divide-and-conquer algorithm to achieve approximately linear runtime, which allows a pan-genome of 1000 strains to be constructed in less than one day with 64 CPUs. This has been extensively tested using 120 simulated pan-genome datasets and real bacterial pan-genomes. Furthermore, we have developed a user-friendly and powerful visualization application for interactively exploring microbial pan-genomes and provided a wealth of pan-genomes from 93 bacterial species in RefSeq database.

As the number of prokaryotic genome sequences continues to increase rapidly, it can be anticipated that more microbial pan-genome studies will be instigated with this wealth of genomic data. Besides the above-mentioned wide range of applications of pan-genome analysis, we present some thoughts pertaining to our main interests in future studies and the conception of a pan-genome browser.

7.1 READS MAPPING ON PAN-GENOME REFERENCE

Building pan-genome reference

Selecting a single reference genome for reads mapping has an obvious drawback: sequences not present in the reference remain erroneously unmapped. Building a more complete pan-genome reference achieves better reads mapping without neglecting species-wide diversity and mobile genetic elements that can exhibit large presence/absence variation between single genomes, thereby being more appropriate for capturing genomic heterogeneity among different strains (see Figure 7.1). One approach to construct such a reference is to select representative sequences from all gene clusters of an established pan-genome (unpublished work).

The choice for gene representatives hinges on the sequence diversity of each cluster:

(1) for highly conserved gene (e.g. > 90% similarity among sequences in a gene cluster), one representative sequence could be sufficient and randomly selected;

(2) for genes exhibiting a high level of diversity (e.g. a large portion of genes in *Prochlorococcus*), gene cluster should be split into sub-clusters based on phylogenetic properties such as branch length and each representative of the sub-clusters is treated as a reference sequence.

This generates a pan-genome reference as a list of representative sequences for all gene clusters.

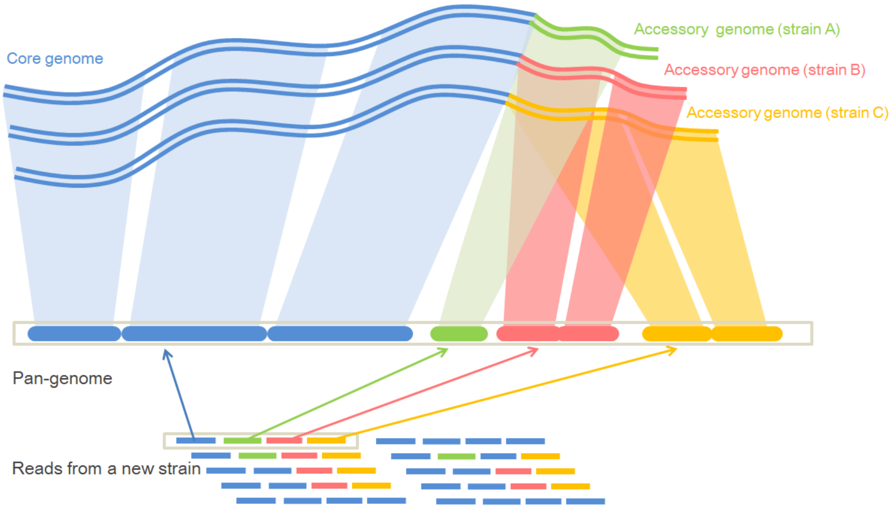

Core genome
Accessory genome (strain A)
Accessory genome (strain B)
Accessory genome (strain C)
Pan-genome
Reads from a new strain

Figure 7.1: **Schematic representation of reads mapping on pan-genome.**
The top part of the figure shows three strains and their core (blue region) and accessory genomes (regions with strain-specific colors). The symbolic clusters are indicated by the thick segments. The reads from a newly sequenced strain comprise genetic materials that are orthologous to accessory genes in different strains, which can be mapped on a pan-genome reference.
By contrast, using only a single genome as the reference leads to false unmapped reads especially for genes not present in the given reference genome.

Integrating mapping results into the interactive dashboard

After pan-genome construction, reads can be mapped on the pan-genome reference using software BWA [66]. The mapped reads and SNPs can be further integrated in the panX interactive dashboard. The dashboard combined with the mapping results has several adjusted features (unpublished work): the corresponding gene table contains all genes mapped by reads with a pre-defined threshold (e.g. at least 75% of genes are covered); the alignment viewer includes new sequences of the investigated isolates that are built on the mapped reference sequence with the called SNPs substituting the original nucleotides.

The gene sequences made for the newly sequenced isolate can also be organized as a GenBank file, by annotating them using the major annotation from their clusters. Additionally, unmapped reads can be assembled and the corresponding contigs examined for novel gene identification. The new genes that are not present in the initial reference records can be further integrated into the pan-genome reference to extend its genetic reservoir.

Another useful feature is the incorporation of the mapping results into the phylogenetic viewers and the illustration of which strain in the collection of reference genomes is the closest to the user-sequenced isolate. This can be either calculated rapidly on the fly by a browser-side JavaScript or computed by server-side reconstruction of the core genome tree. In order to find the most similar strain on the core genome phylogeny to the newly sequenced isolate, the concatenated SNP sequence identified in the reads mapped onto the core genes can be compared to all sequences in the core genome SNP matrix. The closest strain with the highest pairwise similarity can then readily be highlighted on core genome SNP phylogeny.

Alternatively, the phylogenetic reconstruction approach builds a new core genome SNP matrix and infers the corresponding core genome phylogeny, by inserting an additional sequence for the isolate into all core genes alignment and filling the missing positions with dashes when some core genes are absent in the isolate. However, this approach has its limitations, when taking into account the computational burden involved in building the phylogeny on thousands of strains.

Furthermore, gene phylogenies can also be updated using the same similarity search as mentioned above. When selecting a gene cluster in the table, the closest sequence match on the gene tree can be searched, on which the location of new sequences from the isolate can be easily marked. This helps to quickly ascertain the phylogenetic position of new sequences on the gene tree.

7.2 ADDING NEW GENOMES IN A CONSTRUCTED PAN-GENOME

New sequencing technologies have greatly facilitated genomic studies and continue to flourish and advance. We undoubtedly expect a further rapid expansion of genome sequence databases. This raises new scaling challenges that would hinder the feasibility of existing computational methods. Besides, one of the central interests will still be the comparison of new genome sequences against the vast amount of reference genomes in databases.

The construction of a pan-genome rests on pairwise comparisons among all genes in a collection of annotated genomes. We have demonstrated that the divide-and-conquer algorithm facilitates the sequence similarity search and makes it feasible for building a pan-genome of a large-scale dataset.

Yet, in an extreme scenario of incorporating a new genome sequence into an established pan-genome comprising 10000 isolates, it would be computationally intensive to generate a new pan-genome consisting of 10001 isolates from scratch. Our suggestion for a potential solution to this issue is to directly compare the gene sequences from the new genome against the sequences in the gene clusters of the pan-genome. In a similar way as shown in our divide-and-conquer strategy, representative sequences can be extracted to stand for each gene cluster, which are then collected in a single reduced pan-genome. In addition, the selection of representatives for this purpose has more nuances: instead of selecting only one representative per cluster as applied in the primary run of homolog clustering in the divide-and-conquer approach, more representatives for a cluster would need to be used especially when dealing with highly diverged orthologs.

The translation of an all-against-all alignment into a one-against-one (a new genome against a reduced representative pan-genome) alignment would avoid redundancy and allow fast comparison between a new genome and a large pan-genome. The same should be applicable for more new genomes as a query against a pan-genome reference database.

7.3 IDENTIFYING SYNTENY BLOCKS USING GENE CLUSTERS

One approach to detect conserved syntenic blocks in a set of genomes is to search shared blocks in whole genome alignment. However, existing multiple genome aligners [15, 16] fail to address the challenge on large collections of genomes. Besides scalability limitations, the alignment of genomic sequences of a highly diverse species is not trivial since many clusters are very diverged and involve significant insertion and deletion polymorphisms.

Nonetheless, instead of directly aligning entire genome sequences, comparing multiple genomes by leveraging knowledge derived from gene clustering could achieve better computational efficiency on large-scale datasets. Hence, searching synteny can be largely accelerated, when combining the information of gene clusters inferred from pan-genome analysis and gene orders available in annotated genomes. Genes with hundreds or thousands of kilobase pairs from different genomes that are grouped into a cluster can be compactly denoted using a simple cluster ID.

This helps simplify the problem of aligning whole genomes by transforming it into the task of aligning reduced genomes represented by a list of gene cluster IDs. Correspondingly, synteny region can therefore be detected in a faster and more memory-economic manner. Besides, considering tools for multiple genome alignment require genomic sequences as input, the cluster ID would need to be coded as a pseudo sequence based on the DNA alphabet (ACGT). Additionally, the pseudo sequences should also be significantly distinct from each other such that the genome aligner will not erroneously treat any two pseudo sequences (two different gene clusters) as the same one during the comparison.

7.4 ENVISIONING A PAN-GENOME BROWSER

Visualizing multiple genomes is essential for interpreting comparative genomic data. Two widely used approaches refer to linear and circular representations: for example, the Integrative Genomics Viewer (IGV) allows real-time exploration of diverse genomic data at different scales of genome resolution [99]; GenomeRing [41] takes advantage of a circular layout to interactively visualizing multiple genome alignment and has demonstrated its application in the identification of genomic islands.

Pan-genomics introduces new visualization challenges. Whereas multiple sequence alignment among genes from different strains is presented in panX, gene neighborhood information is currently not yet integrated. This might necessitate a pan-genome browser to allow visual inspection of shared regions such as conserved synteny blocks. Since a large proportion of genetic information can be condensed based on information from orthologous clusters, one could harness a cluster-centered scheme to organize pan-genomic data in a pan-genome browser, for the sake of better visually highlighting genomic variation and selectively hiding information redundancy.

The conceptual pan-genome browser can be delineated as follows. First, gene clusters act as display unit by default and all genes in one cluster that are located on different genomes are connected by links denoting orthologous relationships. Once user searches for a particular gene cluster, the positions of corresponding orthologous genes can be highlighted, auto-adjusted by automatically aligning the locations of the related genes, and flexibly panned to inspect the gene neighborhood. Besides, recent paralogous relationships (recent duplications on the same genome) should also be inspectable via "paralogous links" in a different color.

Furthermore, each genome in the browser is associated with a strain checkbox and metadata coupled with the corresponding strain. The genomes can be flexibly sorted using genome information or metadata. Using the strain checkbox to select a subset of strains can highlight those clusters that are only present in selected strains (e.g., present only in strains belonging to a pathogenic group or absent in strains showing reduced virulence).

Another important functionality is the flexible switch between different zoom levels, which allows easy inspection of details of synteny blocks, gene clusters and base pairs. Except the selection of multiple genomes, user can choose a genome as reference and the browser automatically sorts the order of genomes by a checked criterion, such as sequence similarity, same phenotypic traits and pathogenicity.

Moreover, a pan-genome browser should enable a moderate reduction of information, by either applying visualization transparency to emphasize variations and deemphasize redundancy or providing flexible options to display and hide information in the viewed region. Pop-up windows can also be utilized as a useful extension for the content that is difficult to be accommodated in the main window. Such a pan-genome browser could offer useful complementary functionalities to the panX interactive dashboard, by enabling users to investigate conserved synteny and genome-wide variation.

BIBLIOGRAPHY

[1] Andrey Alexeyenko, Ivica Tamas, Gang Liu, and Erik LL Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22 (14):e9–e15, 2006.

[2] Franz Baumdicker, Wolfgang R Hess, and Peter Pfaffelhuber. The diversity of a distributed genome in bacterial populations. *The Annals of Applied Probability*, 20(5):1567–1606, 2010.

[3] Frederic Bertels, Olin K. Silander, Mikhail Pachkov, Paul B. Rainey, and Erik van Nimwegen. Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Mol Biol Evol*, 31(5):1077–1088, May 2014. doi: 10.1093/molbev/msu088.

[4] Paul M Berube, Steven J Biller, Alyssa G Kent, Jessie W Berta-Thompson, Sara E Roggensack, Kathryn H Roache-Johnson, Marcia Ackerman, Lisa R Moore, Joshua D Meisel, Daniel Sher, Luke R Thompson, Lisa Campbell, Adam C Martiny, and Sallie W Chisholm. Physiology and evolution of nitrate acquisition in Prochlorococcus. *The ISME Journal*, 9(10):1195–1207, 2014.

[5] Steven J Biller, Paul M Berube, Jessie W Berta-Thompson, Libusha Kelly, Sara E Roggensack, Lana Awad, Kathryn H Roache-Johnson, Huiming Ding, Stephen J Giovannoni, Gabrielle Rocap, Lisa R Moore, and Sallie W Chisholm. Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Scientific Data*, 1:140034, 2014.

[6] Steven J. Biller, Paul M. Berube, Debbie Lindell, and Sallie W. Chisholm. Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13 (1):13–27, 2014.

[7] Mike Bostock. D3: Data-driven documents, 2016. URL http://d3js.org.

[8] Luis Boto. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1683):819–827, 2010.

[9] Fiona SL Brinkman. The pseudomonas genome database. *Edited by Juan-Luis Ramos, CSIC, Granada, Spain*, page 341, 2006.

[10] Harald Brüssow, Carlos Canchaya, and Wolf-Dietrich Hardt. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and molecular biology reviews*, 68(3):560–602, 2004.

[11] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

[12] Claire Chewapreecha, Pekka Marttinen, Nicholas J Croucher, Susannah J Salter, Simon R Harris, Alison E Mather, William P Hanage, David Goldblatt, Francois H Nosten, Claudia Turner, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS genetics*, 10(8):e1004547, 2014.

[13] Daniel Croll and Bruce A McDonald. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS pathogens*, 8(4):e1002608, 2012.

[14] Nicholas J. Croucher, Jonathan A. Finkelstein, Stephen I. Pelton, Julian Parkhill, Stephen D. Bentley, Marc Lipsitch, and William P. Hanage. Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. *Sci Data*, 2, October 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.58.

[15] Aaron CE Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403, 2004.

[16] Aaron E Darling, Bob Mau, and Nicole T Perna. progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one*, 5(6):e11147, 2010.

[17] Fernando de la Cruz and Julian Davies. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in microbiology*, 8(3):128–133, 2000.

[18] Rafael Díaz, Carmen Vargas-Lagunas, Miguel Angel Villalobos, Humberto Peralta, Yolanda Mora, Sergio Encarnación, Lourdes Girard, and Jaime Mora. argc orthologs from rhizobiales show diverse profiles of transcriptional efficiency and functionality in sinorhizobium meliloti. *Journal of bacteriology*, 193(2):460–472, 2011.

[19] Jean-François Dufayard, Laurent Duret, Simon Penel, Manolo Gouy, François Rechenmann, and Guy Perrière. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603, 2005.

[20] David M. Emms and Steven Kelly. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16:157, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0721-2.

[21] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.

[22] Richard G Everitt, Xavier Didelot, Elizabeth M Batty, Ruth R Miller, Kyle Knox, Bernadette C Young, Rory Bowden, Adam Auton, Antonina Votintseva, Hanna Larner-Svensson, et al. Mobile elements drive recombination hotspots in the core genome of staphylococcus aureus. *Nature communications*, 5, 2014.

[23] J Felsenstein. Phylogenies from Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46:159–173, 1992.

[24] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer, 2004. URL http://www.sinauer.com/inferring-phylogenies.html.

[25] Walter M Fitch. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113, 1970.

[26] Walter M Fitch. Homology: a personal view on some of the problems. *Trends in genetics*, 16(5):227–231, 2000.

[27] Gregory P. Fournier and J. Peter Gogarten. Evolution of Acetoclastic Methanogenesis in Methanosarcina via Horizontal Gene Transfer from Cellulolytic Clostridia. *J. Bacteriol.*, 190(3):1124–1127, February 2008. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.01382-07.

[28] Derrick E. Fouts, Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research*, 40(22):e172, December 2012. ISSN 1362-4962. doi: 10.1093/nar/gks757.

[29] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[30] Songzhe Fu, Sophie Octavia, Mark M Tanaka, Vitali Sintchenko, and Ruiting Lan. Defining the core genome of salmonella enterica serovar typhimurium for genomic surveillance and epidemiological typing. *Journal of clinical microbiology*, 53(8):2530–2538, 2015.

[31] Debra L Fulton, Yvonne Y Li, Matthew R Laird, Benjamin GS Horsman, Fiona M Roche, and Fiona SL Brinkman. Improving the specificity of high-throughput ortholog prediction. *BMC bioinformatics*, 7(1):270, 2006.

[32] Ohad Gal-Mor and B Brett Finlay. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cellular microbiology*, 8(11):1707–1719, 2006.

[33] Jennifer Gardy, Nicholas J. Loman, and Andrew Rambaut. Real-time digital pathogen surveillance - the time is now. *Genome Biology*, 16:155, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0726-x.

[34] Theodore R. Gibbons, Stephen M. Mount, Endymion D. Cooper, and Charles F. Delwiche. Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm. *BMC Bioinformatics*, 16:218, 2015.

[35] John Gómez, Leyla J García, Gustavo A Salazar, Jose Villaveces, Swanand Gore, Alexander García, Maria J Martín, Guillaume Launay, Rafael Alcántara, Noemi Del Toro Ayllón, et al. Biojs: an open source javascript framework for biological data visualization. *Bioinformatics*, page btt100, 2013.

[36] Morris Goodman, John Czelusniak, G William Moore, AE Romero-Herrera, and Genji Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 1979.

[37] Jean-François Gout, Laurent Duret, and Daniel Kahn. Differential retention of metabolic genes following whole-genome duplication. *Molecular biology and evolution*, 26 (5):1067–1072, 2009.

[38] Thorsten Grebe and Regine Hakenbeck. Penicillin-binding proteins 2b and 2x of streptococcus pneumoniae are primary resistance determinants for different classes of beta-lactam antibiotics. *Antimicrobial agents and chemotherapy*, 40(4):829–834, 1996.

[39] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22 (2):160–174, oct 1985. ISSN 0022-2844.

[40] Jianxin He, Regina L Baldini, Eric Déziel, Maude Saucier, Qunhao Zhang, Nicole T Liberati, Daniel Lee, Jonathan Urbach, Howard M Goodman, and Laurence G Rahme. The broad host range pathogen pseudomonas aeruginosa strain pa14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2530–2535, 2004.

[41] Alexander Herbig, Günter Jäger, Florian Battke, and Kay Nieselt. Genomering: alignment visualization based on supergenome coordinates. *Bioinformatics*, 28(12):i7–i15, 2012.

[42] R Hudson. Ms a Program for Generating Samples Under Neutral Models. *Bioinformatics*, (2002):337–338, 2002.

[43] Jaime Huerta-Cepas, Salvador Capella-Gutierrez, Leszek P Pryszcz, Ivan Denisov, Diego Kormes, Marina Marcet-Houben, and Toni Gabaldón. Phylomedb v3. 0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, 39(suppl_1): D556–D560, 2010.

[44] Diarmaid Hughes. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome biology*, 1(6):reviews0006–1, 2000.

[45] D H Huson and M Steel. Phylogenetic trees based on gene content. *Bioinformatics*, 20(13):2044–2049, 2004.

[46] Daniel H. Huson. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[47] Daniel H. Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267, 2006.

[48] Martijn A Huynen and Peer Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856, 1998.

[49] Nadeeza Ishmael, Julie C Dunning Hotopp, Panagiotis Ioannidis, Sarah Biber, Joyce Sakamoto, Stefanos Siozios, Vishvanath Nene, John Werren, Kostas Bourtzis, Seth R Bordenstein, et al. Extensive genomic diversity of closely related wolbachia strains. *Microbiology*, 155(7):2211–2222, 2009.

[50] Robert W Jackson, Boris Vinatzer, Dawn L Arnold, Steve Dorus, and Jesús Murillo. The influence of the accessory genome on bacterial pathogen evolution. *Mobile genetic elements*, 1(1):55–65, 2011.

[51] Rolf S Kaas, Carsten Friis, David W Ussery, and Frank M Aarestrup. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse escherichia coli genomes. *BMC genomics*, 13(1):577, 2012.

[52] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, jan 2016. ISSN 0305-1048.

[53] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[54] Gregory C Kettler, Adam C Martiny, Katherine Huang, Jeremy Zucker, Maureen L Coleman, Sebastien Rodrigue, Feng Chen, Alla Lapidus, Steven Ferriera, Justin Johnson, et al. Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS genetics*, 3(12): e231, 2007.

[55] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338, 2005.

[56] EV Koonin and MY Galperin. Sequence-evolution-function, computational approaches in comparative genomics (2003). *Norwell, Massachusetts: Kluwer Academic Publishers*.

[57] E.V. Koonin, Y.I. Wolf, and P. Puigbo. The Phylogenetic Forest and the Quest for the Elusive Tree of Life. *Cold Spring Harb Symp Quant Biol*, 74:205–213, 2009. ISSN 0091-7451. doi: 10.1101/sqb.2009.74.006.

[58] David M Kristensen, Yuri I Wolf, Arcady R Mushegian, and Eugene V Koonin. Computational methods for gene orthology inference. *Briefings in bioinformatics*, 12(5):379–391, 2011.

[59] Vanderlene L Kung, Egon A Ozer, and Alan R Hauser. The accessory genome of pseudomonas aeruginosa. *Microbiology and molecular biology reviews*, 74(4):621–641, 2010.

[60] Arnold Kuzniar, Roeland CHJ van Ham, Sándor Pongor, and Jack AM Leunissen. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551, 2008.

[61] Chad Laing, Cody Buchanan, Eduardo N Taboada, Yongxiang Zhang, Andrew Kropinski, Andre Villegas, James E Thomas, and Victor PJ Gannon. Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC bioinformatics*, 11(1):1, 2010.

[62] Ruiting Lan and Peter R Reeves. When does a clone deserve a name? a perspective on bacterial species based on population genetics. *Trends in microbiology*, 9(9):419–424, 2001.

[63] Pascal Lapierre and J Peter Gogarten. Estimating the size of the bacterial pan-genome. *Trends in genetics*, 25(3):107–110, 2009.

[64] Tristan Lefébure and Michael J Stanhope. Evolution of the core and pan-genome of streptococcus: positive selection, recombination, and genome composition. *Genome biology*, 8(5):R71, 2007.

[65] Tristan Lefébure, Paulina D Pavinski Bitar, Haruo Suzuki, and Michael J Stanhope. Evolutionary dynamics of complete campylobacter pan-genomes and the bacterial species concept. *Genome biology and evolution*, 2:646–655, 2010.

[66] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[67] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.

[68] Debbie Lindell, Matthew B Sullivan, Zackary I Johnson, Andrew C Tolonen, Forest Rohwer, and Sallie W Chisholm. Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30): 11013–8, 2004. ISSN 0027-8424.

[69] Michael Lynch and Vaishali Katju. The altered evolutionary trajectories of gene duplicates. *TRENDS in Genetics*, 20(11):544–549, 2004.

[70] Mylène M Maury, Yu-Huan Tsai, Caroline Charlier, Marie Touchon, Viviane Chenal-Francisque, Alexandre Leclercq, Alexis Criscuolo, Charlotte Gaultier, Sophie Roussel, Anne Brisabois, et al. Uncovering listeria monocytogenes hypervirulence by harnessing its biodiversity. *Nature genetics*, 48(3):308–313, 2016.

[71] Duccio Medini, Claudio Donati, Herve Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pangenome. *Current opinion in genetics & development*, 15(6): 589–594, 2005.

[72] Guillaume Méric, Koji Yahara, Leonardos Mageiros, Ben Pascoe, Martin CJ Maiden, Keith A Jolley, and Samuel K Sheppard. A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic campylobacter. *PloS one*, 9(3):e92798, 2014.

[73] Boris Mirkin, Ilya Muchnik, and Temple F Smith. A biologically consistent model for comparing molecular phylogenies. *Journal of computational biology*, 2(4):493–507, 1995.

[74] Lisa R Moore, Gabrielle Rocap, and Sallie W Chisholm. Physiology and molecular phylogeny of coexisting prochlorococcus ecotypes. *Nature*, 393(6684):464, 1998.

[75] Alessandro Muzzi, Vega Masignani, and Rino Rappuoli. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug discovery today*, 12(11):429–439, 2007.

[76] Richard A. Neher and Trevor Bedford. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, November 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv381.

[77] Howard Ochman, Jeffrey G Lawrence, and Eduardo A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299, 2000.

[78] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.

[79] Roderic DM Page and Michael A Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular phylogenetics and evolution*, 7(2):231–240, 1997.

[80] Mark E Peterson, Feng Chen, Jeffery G Saven, David S Roos, Patricia C Babbitt, and Andrej Sali. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, 18(6):1306–1315, 2009.

[81] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009490.

[82] Pere Puigbò, Alexander E. Lobkovsky, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology*, 12:66, 2014. ISSN 1741-7007. doi: 10.1186/s12915-014-0066-4.

[83] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997. ISSN 1367-4803.

[84] David A Rasko, MJ Rosovitz, Garry SA Myers, Emmanuel F Mongodin, W Florian Fricke, Pawel Gajer, Jonathan Crabtree, Mohammed Sebaihia, Nicholas R Thomson, Roy Chaudhuri, et al. The pangenome structure of escherichia coli: comparative genomic analysis of e. coli commensal and pathogenic isolates. *Journal of bacteriology*, 190(20):6881–6893, 2008.

[85] Pavel Sagulenko, Vadim Puller, and Richard Neher. TreeTime: maximum likelihood phylodynamic analysis. *bioRxiv*, page 153494, June 2017. doi: 10.1101/153494.

[86] Andreas Sandgren, Michael Strong, Preetika Muthukrishnan, Brian K Weiner, George M Church, and Megan B Murray. Tuberculosis drug resistance mutation database. *PLoS medicine*, 6(2):e1000002, 2009.

[87] Sara F Sarkar and David S Guttman. Evolution of the core genome of pseudomonas syringae, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*, 70(4):1999–2012, 2004.

[88] Devin R Scannell, Kevin P Byrne, Jonathan L Gordon, Simon Wong, and Kenneth H Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–345, 2006.

[89] Herbert Schmidt and Michael Hensel. Pathogenicity islands in bacterial pathogenesis. *Clinical microbiology reviews*, 17(1):14–56, 2004.

[90] Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[91] Thomas Sicheritz-Pontén and Siv GE Andersson. A phylogenomic approach to microbial evolution. *Nucleic acids research*, 29(2):545–552, 2001.

[92] Cedric Simillion, Klaas Vandepoele, and Yves Van de Peer. Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, 26(11):1225–1235, 2004.

[93] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014.

[94] Zhiyi Sun, Jeffrey L. Blanchard, A Bleasby, R Chenna, and PA McGettigan. Strong Genome-Wide Selection Early in the Evolution of Prochlorococcus Resulted in a Reduced Genome through the Loss of a Large Number of Small Effect Genes. *PLoS ONE*, 9(3):e88837, 2014.

[95] Amelia Tait-Kamradt, Joanna Clancy, Melissa Cronan, Fadia Dib-Hajj, Lillian Wondrack, Wei Yuan, and Joyce Sutcliffe. mefe is necessary for the erythromycin-resistant m phenotype in streptococcus pneumoniae. *Antimicrobial agents and chemotherapy*, 41(10):2251–2255, 1997.

[96] Mayumi Tanaka, Tong Wang, Yoshikuni Onodera, Yoko Uchida, and Kenichi Sato. Mechanism of quinolone resistance in staphylococcus aureus. *Journal of Infection and Chemotherapy*, 6(3):131–139, 2000.

[97] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.

[98] Herve Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Current opinion in microbiology*, 11(5):472–477, 2008.

[99] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[100] Kalliopi Trachana, Tomas A Larsson, Sean Powell, Wei-Hua Chen, Tobias Doerks, Jean Muller, and Peer Bork. Orthology prediction methods: A quality assessment using curated protein families. *Bioessays*, 33(10):769–780, October 2011. ISSN 0265-9247. doi: 10.1002/bies.201100062.

[101] Ikuo Uchiyama, Jacob Albritton, Masaki Fukuyo, Kenji K Kojima, Koji Yahara, and Ichizo Kobayashi. A novel approach to helicobacter pylori pan-genome analysis for identification of genomic islands. *PloS one*, 11(8):e0159419, 2016.

[102] Rene TJM Van der Heijden, Berend Snel, Vera Van Noort, and Martijn A Huynen. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC bioinformatics*, 8(1):83, 2007.

[103] George Vernikos, Duccio Medini, David R Riley, and Herve Tettelin. Ten years of pan-genome analyses. *Current opinion in microbiology*, 23:148–154, 2015.

[104] Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327–335, 2009.

[105] Michiel Vos and Xavier Didelot. A comparison of homologous recombination rates in bacteria and archaea. *The ISME journal*, 3(2):199, 2009.

[106] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36 (suppl_1):D13–D21, 2007.

[107] Gordon Woodhull, Nick Zhu, et al. dc.js - dimensional charting javascript library, 2016. URL https://dc-js.github.io/dc.js/.

[108] Gerard D Wright. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology*, 5(3):175–186, 2007.

[109] Guy Yachdav, Sebastian Wilzbach, Benedikt Rauscher, Robert Sheridan, Ian Sillitoe, James Procter, Suzanna E Lewis, Burkhard Rost, and Tatyana Goldberg. Msaviewer: interactive javascript visualization of multiple sequence alignments. *Bioinformatics*, 32(22):3501–3503, 2016.

[110] Seyed Alireza Zamani-Dahaj, Mohamed Okasha, Jakub Kosakowski, and Paul G. Higgs. Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Molecular Biology and Evolution*, 33(7):1843–1857, 2016.

[111] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. Pgap: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418, 2012.

[112] Yongbing Zhao, Xinmiao Jia, Junhui Yang, Yunchao Ling, Zhang Zhang, Jun Yu, Jiayan Wu, and Jingfa Xiao. Pangp: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, 30(9):1297–1299, 2014.

APPENDIX

PAPERS

Wei Ding, Franz Baumdicker, and Richard A Neher. panx: pangenome analysis and exploration. *bioRxiv*, page 072082, 2017.

TL. Karasov, J. Almario, C. Friedemann, W. Ding[1], D. Lundberg, M. Neumann, J. Regalado, S. Kersten, R. Neher, E. Kemen, and D. Weigel. Single abundant OTU composed of dozens of pathogenic lineages colonizing in parallel within A. thaliana populations. (in submission)

M. Exposito-Alonso, F. Vasseur, W. Ding[2], G. Wang, H. A. Burbano, and D. Weigel. Genomic basis and evolutionary potential for extreme drought adaptation in arabidopsis thaliana. *bioRxiv*, page 118067, 2017.

C. Lee, H. Svardal, A. Farlow, M. Exposito-Alonso, W. Ding[3], P. Novikova, C. Alonso-Blanco, D. Weigel, and M. Nordborg. On the post-glacial spread of human commensal Arabidopsis thaliana. *Nature Communications*, 8, 2017.

1001 Genomes Consortium[4] et al. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell, 166(2):481–491, 2016.

[1] I conducted pan-genome analysis on 1524 Pseudomonas isolates.

[2] I provided analysis scripts and haplotype analysis results on donor-recipient pairs of 1135 Arabidopsis thaliana accessions.

[3] I provided haplotype sharing statistics among different groups of Arabidopsis thaliana accessions.

[4] I contributed to the pipeline analysis of IBD sharing on 1135 Arabidopsis thaliana genomes.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. `classicthesis` is available for both LaTeX and LyX:

`http://code.google.com/p/classicthesis/`