An Algorithmic Approach for Identifying Rearrangements in Multiple Bacterial Genomes

Noureen, Mehwish

Doctor of Philosophy

Department of Genetics

School of Life Science

The Graduate University for Advanced Studies,

SOKENDAI

2021

Declaration

I hereby declare that the work presented in the following thesis under the sincere guidance of my supervisor **(Prof. Masanori Arita)** is my own effort, except where otherwise acknowledged. This thesis is my own composition and no part of the thesis has been previously presented for any other degree.

Noureen Mehwish

Dedicated to

my Mother (Shah Sultan) and my Father (Ali Ahmad)

for their invaluable love, prayers, support and guidance

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Masanori Arita for his constant guidance, encouragement and support. I would like to thank him for his advices, comments and suggestions that paved the ways of this research for me.

I would also like to thank all my laboratory members for their useful comments and suggestions that helped me in refining my research. I am really grateful to the lab secretaries Ohunki-san and Murakata-san for always helping me out and making things easier for me.

I would like to thank Drs. Maria Constanza, Charles Rabkin and Andres Gutierrez Escobar at the Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA for making it possible for me to learn and experience the new avenues of research. I would like to deeply thank Dr. Maria Constanza for being such a great host and helping me out in everything during my visits to NCI.

I would take this opportunity to thank all my thesis committee members Prof. Kurokawa Ken, Prof. Niki Hironori, Prof. Miyagishima Shin-ya and Assoc. Prof. Ikeo Kazuho for their valuable comments and suggestions.

I would also like to thank Tanizawa Yasuhiro at National Institute of Genetics for his help and guidance. I am really grateful to Drs. Ikuo Uchiyama at National Institute for Basic Biology (JP) and Ichizo Kobayashi at Kyorin University (JP) for the great discussions and useful comments.

I would like to acknowledge the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan for providing me with the opportunity to pursue my doctoral studies in Japan. I would also like to acknowledge SOKENDAI for giving me the opportunity and making it possible for me to visit the National Cancer Institute under the Student Dispatch program. This opportunity enabled me to broaden the perspective of my research and enrich my experience.

I am very grateful to my friend Maria Altaf Satti for her sincere suggestions and support. I would like to deeply thank all my friends for making my stay in Japan a memorable one.

I am immensely thankful to my beloved parents for their unconditional love and care. I want to acknowledge them for their support and guidance that made be able to pursue my higher education. I can’t thank them enough for all the things that they have done for me. I am also very grateful to my sisters and brothers for their love, support and sincere advices.

Abstract

Each species has a unique genome structure that changes slowly with time. Genome of each species is subjected to both the local and global mutations during the evolution. Local mutations affect the genomes at a smaller scale and occur more frequently. On the other hand, genome rearrangements affect the large segments of the genome and occur less frequently. Increasing number of prokaryotic genomes and their comparison have revealed the presence of large number of genomic rearrangements. The dynamic nature of bacterial genome is the result of rearrangements, horizontal gene transfer and activity of the mobile genetic elements. Genome rearrangements not only change the orientation but also the order of the genes on the chromosomes. As genome rearrangements are rarer compared to the point mutations, they can reveal the important events that occurred during the course of evolution. Chapter 1 discusses in detail the different types of genomic variations. It also describes the possible causes that leads to the occurrence of the various genomic variations. In addition to this, it also focuses on the consequences of these variations on the genomes. It also discusses how the genome rearrangements help the organisms to cope up with the environmental challenges thus facilitating their adaptation to the rigorous niches.

Several approaches have been proposed to identify the genome rearrangements, however most of these approaches use only the pairwise comparison and consider the similar set of genes. I have developed an algorithmic approach to identify the genome rearrangements in multiple bacterial genomes considering highly conserved genes in a given set of genomes. Orthologous gene clusters were used to identify the gene order in each genome which was used as an input to identify the genome rearrangements. The obvious benefit of my approach is scalability: whole genome comparison is difficult for many genomes using previous approaches comparing two genomes. My method can handle hundreds of strains at the level of gene orders. Chapter 2 provides the brief overview of the previous approaches that were introduced to identify the genome rearrangements and discusses in detail the algorithm that I have developed to identify the genome rearrangements while comparing multiple bacterial genomes.

I have used *Helicobacter pylori* strains to demonstrate the use of my algorithm, as this bacterium has a very diverse genomic structure. Using my algorithm, the geographically region-specific rearrangements and those shared across continents were identified for 72 *H. pylori* strains in the public repository. Region specific breakpoints were overrepresented in

Asia and Australia whereas all breakpoints were detected in Europe. Total 41 inversions were identified, 23 were shared whereas 18 were strain specific. Strain from Europe and East Asia shared as many as 11 inversions. Some inversions occurred more frequently and were found in strains from all geographical locations except for Africa. Two specific inversions were associated with disease states such as cancer. Three genomic loci were frequently involved in rearrangements (*rearrangement hotspots*) in the analyzed strains. The pattern of inversions was most diverse in Japan probably because of the larger number of genome sampling. The North American region also had the diverse inversion pattern even though the number of samples was much smaller compared to Japan. This diversity occurred maybe because of human migration. Many inversions in *H. pylori* strains were shared across geographic regions, and only few were found to be geographically region-specific.

To identify the cause of rearrangements, the association of repeats, insertion sequences (IS) and genomic islands (GIs) were investigated. Among all the strains the largest number of direct and inverted repeats were found in the East Asian strain UM037. This strain contains six inversions, and three of them were associated with inverted repeats. The correlation between the number of repeats and inversions was weak, suggesting that not only the occurrence of repeats but their relative position is also important for the homologous recombination. Shared inversions tend to possess more inverted repeats compared to the strain specific inversions. Beside this, world-wide inversion breakpoints had more IS elements compared to others. No association was found between the well-known IS609 repeat and any type of the inversion. GIs were mostly associated with region-specific and strain-specific breakpoints. Most of the shared inversions breakpoints possessed the similar genomic elements with a few exceptions. This suggests that these elements are well conserved irrespective of the different geographical region. As the number of analyzed strains was small, a large-scale analysis can help us to understand the disease mechanism and reveal the migration pattern.

As some of the genome rearrangements were associated with the disease state such as cancer, I performed the analysis on a larger dataset, 123 *Helicobacter pylori* genomes to find the association of the genomic features more specifically the genome rearrangements with the disease outcome. Difference in the distribution of various genomic elements was investigated among the strains from different groups defined on the basis of the disease outcome. Comparative analysis of the strains revealed the presence of certain group-specific genes. Most of the identified inversions were shared and few were associated with the disease state. Three and six inversions were associated with the gastritis and chronic gastritis disease outcome,

respectively. Strains were more related based on their geographical locations rather than the disease outcome. Weak association between the genomic features and disease state might be because of the fact that the disease outcome depends on several other factors such as environment, diet and host. Besides this, several genomes with no disease state information also makes it difficult to draw some conclusions.

Publications

1. **Noureen M**, Tada I, Kawashima T, Arita M. Rearrangement analysis of multiple bacterial genomes. *BMC bioinformatics*. 2019; 20(23):1.

2. **Noureen M**, Kawashima T, Arita M. Genetic Markers of Genome Rearrangements in *Helicobacter pylori*. *Microorganisms*. 2021; 9(3):621.

3. Gutiérrez-Escobar AJ, Velapatiño B, Borda V, Rabkin CS, Tarazona-Santos E, Cabrera L, Cok J, Hooper CC, Jahuira-Arias H, Herrera P, **Noureen M**, Wang D, Romero-Gallo J, Tran B, Peek RM Jr, Berg DE, Gilman RH and Camargo MC. Identification of New Helicobacter pylori Subpopulations in Native Americans and Mestizos from Peru. *Frontiers in Microbiology*. 2020; 11:3118.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The basic unit of life “the cell”, contains the “blueprint” for the development of an organism [1, 2]. Everything on earth that has life is made of one or more cells. The cells can be classified into two types: prokaryotic cells and eukaryotic cells. The organisms in Bacteria and Archaea domains consist of the prokaryotic cells whereas all other organisms (animals, plants, fungi, protists) are made of eukaryotic cells. Every cell despite being different have something common that is they hold the organism’s entire genetic information [1, 3]. The complete genetic information of an organism is stored in the molecule termed as “deoxyribonucleic acid (DNA)”, that forms the genome of an organism. The DNA having two strands is present as a double helical structure in the cell. Each strand consists of the sequence of four nucleotides (bases), Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). The bases on one strand (leading strand) forms the hydrogen bonds with their complementary bases on the other strand (lagging strand), such that Adenine pairs with Thymine, and Cytosine pairs with Guanine [1, 2].

The sequence of bases on one strand can completely determine the other strand due to their complementary nature. As the genetic information is encoded by the sequence of bases on the strand therefore, it can be obtained by writing this sequence. In an organism, the DNA can be present as a single molecule (chromosome) or can be divided into two or more chromosomes. The chromosome can be in one of its two forms: linear or circular [1, 2]. Linear chromosome where DNA sequence has a beginning and end is present in most of the higher organisms whereas the circular chromosomes where DNA sequence is without the start and end is found in simpler organisms like bacteria [4]. The coding segments of chromosomes called the “genes” are transcribed and translated into proteins, that are the drivers of most of the functions performed by the cell [1, 2]. Each gene has a direction in which it should be transcribed that can be either forward or backward. A chromosome can then be viewed as a collection of the oriented genes in a particular order [4].

1.2 Genomic Variations

Each species has a specific genome structure that changes slowly with time. This change might be the result of the different factors that can be internal or external. Genomes are commonly considered dynamic and are believed to have undergone repeated alteration since the origin of life [5]. The dynamic nature of the genomes is the result of the collective effect of the small changes in the DNA sequence caused by mutation along with the changes that occur on a larger scale (rearrangements) because of recombination [6]. The term "mutation" is derived from a Latin word meaning "to change" [7]. It was first introduced by Hugo de Vries in 1905 to report the changes that he had observed in the phenotype of the evening-primrose plant (*Oenothera lamarckiana*) [5]. The DNA sequence can get altered at the local or global level during the evolution [8].

1.2.1 Point Mutations

Local mutations also known as point mutations, are the most common and studied mutations that occur in a DNA molecule. Only a small segment of the DNA is affected by these mutations [4]. These include substitution, insertion and deletion of a single nucleotide [8].

- **Substitution**

  A single nucleotide (base) is changed or substituted with another nucleotide. The substitution can either be classified as transition or transversion depending on the types of the bases involved (Figure 1.1a). The transition occurs when a purine base (A, G) is substituted by another purine or a pyrimidine (C, T) base by another pyrimidine. In contrast, transversion occurs when a purine base is substituted by pyrimidine or a pyrimidine base by a purine [5].
- **Insertion**

  The addition of a nucleotide into a DNA sequence (Figure 1.1b).
- **Deletion**

  The deletion of a nucleotide from a DNA sequence (Figure 1.1c).

Genes are altered by the mutations [7] that can be deleterious or beneficial resulting in the change in the genotype or phenotype of the organism [5]. Mutations can be caused either by the cellular processes or by the environmental mutagens [7]. Genes being the basic hereditary unit [2], carry these variations into the progeny leading to the differences among the organisms [1].

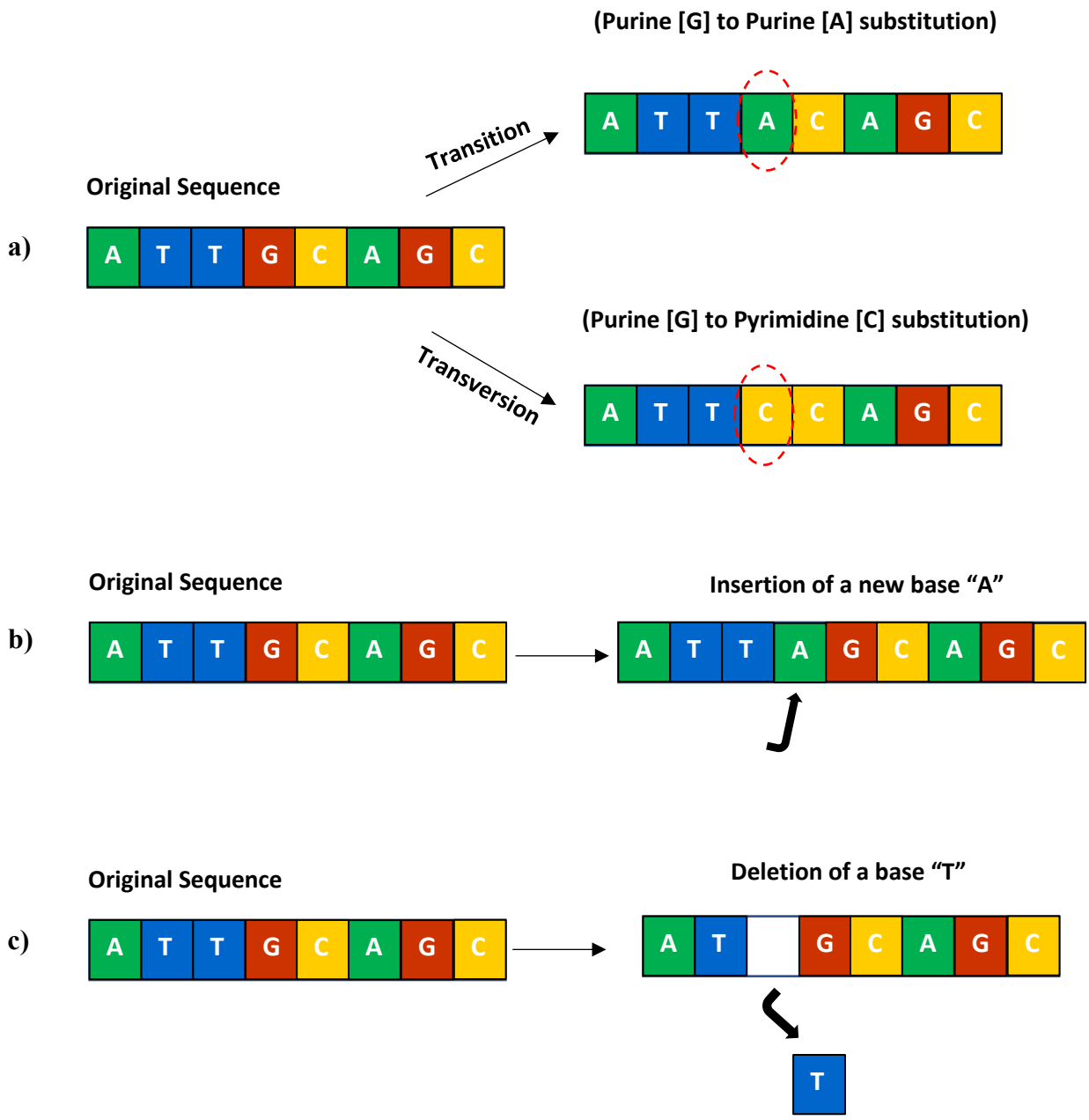


Figure 1.1: Schematic diagram of the different types of point mutations. **a**) Substitution [Transition and Transversion] **b**) Insertion of a purine (A) in the original sequence **c**) Deletion of a pyrimidine (T) from the original sequence

1.2.2 Genome Rearrangements

Global mutations (genome rearrangements) affect the DNA molecule at a larger scale. The breaks in the chromosome causes the occurrence of the genome rearrangements. As a result of these rearrangement events, the location of the genes is altered [4]. The most common types of genome rearrangement include inversion (also known as reversal), transposition, duplication, and translocation [9].

- **Inversion**
  A block of the genome is cut, inverted and is reinserted at the same location [8] (Figure 1.2a). The location remains the same but the direction in which the segment is inserted gets changed [1]. No gain or loss of genomic information occurs as a result of an inversion [10].
- **Translocation**
  End portion of the two chromosomes are exchanged (Figure 1.2b). In translocations either, the genetic information is maintained or changed by the gain or loss [11].
- **Transposition**
  A segment of a genome is inserted in another location on the chromosome (Figure 1.2c). If the segment is inverted along with the change in its location then it is called an inverse transposition.
- **Duplication**
  A segment of a genome is duplicated and is inserted in the genome (Figure 1.2d). If the duplicated region is adjacent then it is called the tandem duplication [12], else it is called insertional duplication if it is inserted in a different genomic region [10].
- **Fission**
  A chromosome is divided into two chromosomes (Figure 1.2e).
- **Fusion**
  Two chromosomes are fused to form one (Figure 1.2f).

Different molecular mechanisms like DNA repair, recombination and replication are the major cause of the genome rearrangements. Genome rearrangements could be of varying lengths, depending on the number of nucleotides involved which could range from a few thousand to a few million. The rearranged region could include few or more genes and operon depending on the genomic location that has been affected by the rearrangement [13].

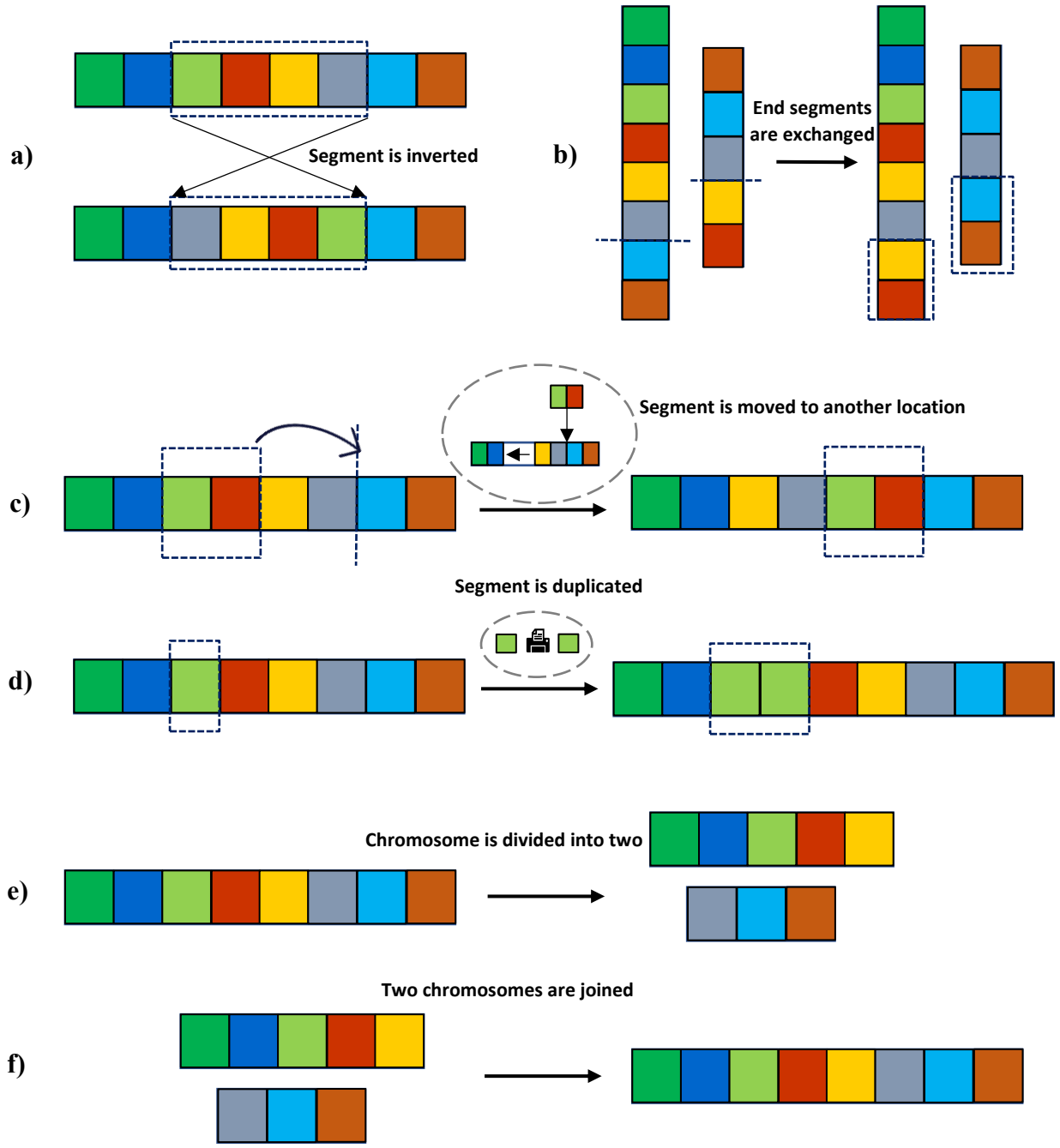


Figure 1.2: Schematic diagram of the genome rearrangements. **a**) Inversion **b**) Translocation **c**) Transposition **d**) Duplication **e**) Fission **f**) Fusion

1.3 Possible Drivers of Rearrangements in Bacteria

Bacteria are one the important form of life on earth. Approximately, 5×10^{30} bacteria have been reported to be present on earth that live in different environments like oceans, soil or land [14]. They can also exist as pathogens and symbionts, residing either inside or on the surface of the other organisms. The environment and evolution of the living organisms are greatly influenced by bacteria [15]. The genome of the bacteria is dynamic [16], which has been shaped by genome rearrangements, mobile genetic elements and horizontal gene transfer during evolution. Most of these changes are random, however few are planned [15]. In prokaryotes, the process that leads to the generation of different genome rearrangements seems to be the same as in eukaryotes [13, 17]. However, compared to the eukaryotes, genome rearrangements are not investigated extensively in prokaryotes [18-20]. There can be different molecular drivers (insertion sequences, repeat sequences, genomic islands, transposons and bacteriophages) of the genome rearrangements [15].

1.3.1 Insertion Sequences

Insertion sequence (IS) elements are the small mobile DNA segment, ranging in size from 0.7 kb to 2.5 kb [15]. They are one of the simplest mobile genetic elements and are widely distributed among the different species [21]. These elements consist of one or more open reading frames which encode the transposase enzyme involved in their movement. Most of these elements are bounded by the inverted repeats (10bp - 40bp) at their termini [22] (Figure 1.3).

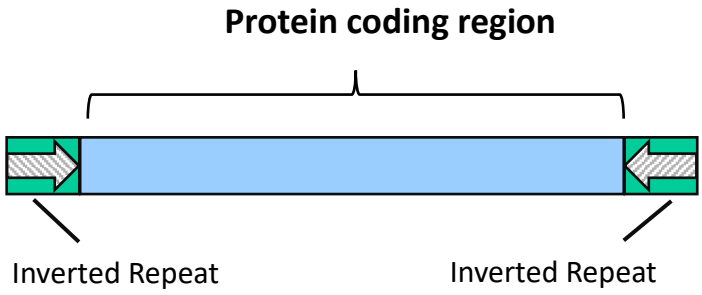


Figure 1.3: Schematic representation of an insertion sequence element

Insertion sequences are broadly distributed and in prokaryotic genomes they can occur in very large numbers [23]. They were first discovered in the late 1960s, leading to the identification

of more than 4000 different insertion sequences to date. They can be classified into different families depending on their transposases [24]. In prokaryotes the insertion sequences are quite diverse in terms of their structure, specificity and mechanism of transposable activity [25-27]. IS elements present in the strains of a species are not conserved but can be present in other species which reflects the transfer of DNA between different species [28]. Host genome gets changed whenever an IS element is inserted but it may regain its previous form or encounter a variation when an IS element is excised.

Insertion of an IS element can lead to the differential expression of some of the host genes. Gene inactivation can occur as a result of the disruption of the gene or its regulatory region. The insertion of the IS element may be beneficial or deleterious depending on the inactivated gene and its effects on the cell. Similarly, IS element can elevate the expression of some of the genes which can have the varying consequences [15]. The IS elements have been reported to affect the antimicrobial resistance and virulence of bacterial species. Some of the IS elements may be helpful to overcome the environmental challenges and help the bacteria to adapt to the new niche [21]. Besides this, IS elements can also cause genome rearrangements including inversions, duplications, fusion and deletions [22]. For example, the IS elements were found to be associated with the inversion in the bacterial species *Neisseria gonorrhoeae* [29]. Some of the IS elements like IS407A [30], ISBma2 [30], IS200[31] and IS905 [32] have been found to be involved in large deletion, large duplication and inversion respectively [15].

1.3.2 Repeat Sequences

DNA repeat can be mathematically defined as the two similar substrings present in a same genome [33]. Large number of repeat sequences having the recombination hotspots are present in the bacterial genomes [33, 34]. Genome plasticity results from amplification, recombination and deletion caused by the large number of repeat sequences in the DNA [34, 35]. Extensive sequence variation in various prokaryotes might be the consequence of the use of repeat sequences and homologous recombination. In addition to the intrachromosomal recombination, horizontal gene transfer (HGT) may be the cause of occurrence of a repeat in a genome. This happens when the new DNA fragment holds the information similar to that of the host genome and uses the site-specific recombination to integrate itself into the host genome [33]. High occurrence of repeats was found in the closely related bacterial genomes with the less conserved genomic structure by Rocha *et al*. Large number of repeats in a genome leads to the loss of gene order as a result of frequently occurring genome rearrangements [36]. One of the

characteristics of significant evolutionary mechanism are the repeat sequences that help the bacteria to cope up with the various environmental challenges thus leading to its adaptation [37].

The key factors that determine the degree and occurrence of recombination include the similarity between the two sequences, distance between them, their lengths and the mechanisms of recombination [33]. In most prokaryotes, repeats of length > 25 base pairs (bp) are considered statistically significant [38] and are thought to be involved in homologous recombination [39]. The average repeat size is 53 and 100 base pairs (bp) for *Methanococcus jannaschii* and *Helicobacter pylori*, respectively [37]. There are direct repeats (DRs) (Figure 1.4a) and inverted repeats (IRs) (Figure 1.4b), and the former is considered more common [40]. Genome rearrangements can occur as a result of recombination between the direct or inverted repeat sequences [41, 42]. Tandem duplication can occur as a result of recombination between the direct repeat sequences which have the same orientation. Besides this, direct repeat sequences can also recombine to give rise to the translocations [42].

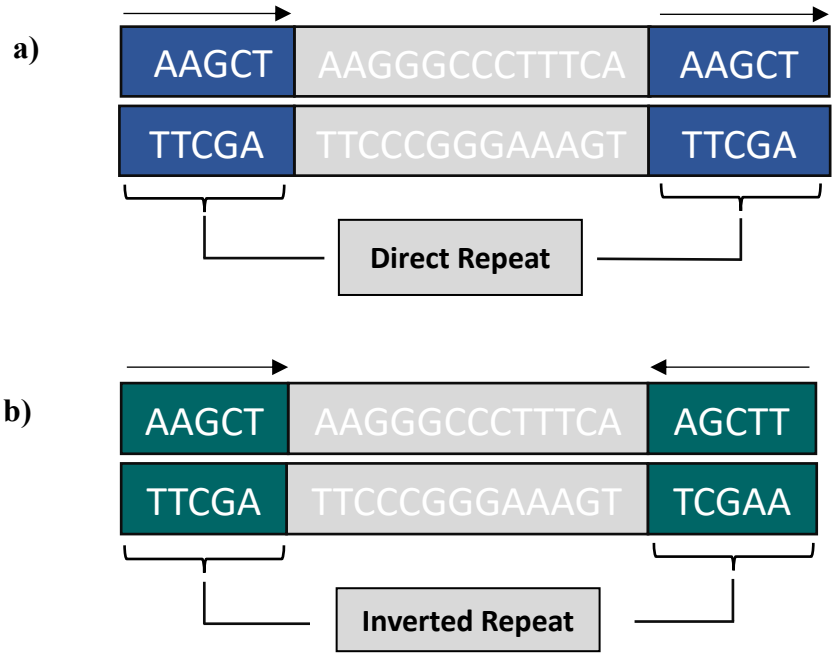


Figure 1.4: Hypothetical representation of repeats; **a)** Direct Repeat **b)** Inverted Repeat

On the other hand, recombination between the inverted repeat sequences (sequences that have the opposite orientation) can give rise to the inversions (Figure 1.5 [43]), thus causing the intermediate sequence to be inserted in an inverted orientation at the same genomic location [42]. Compared to the duplications which are usually unstable, inversions are mostly irreversible thus causing a permanent change in the bacterial genomes [44].

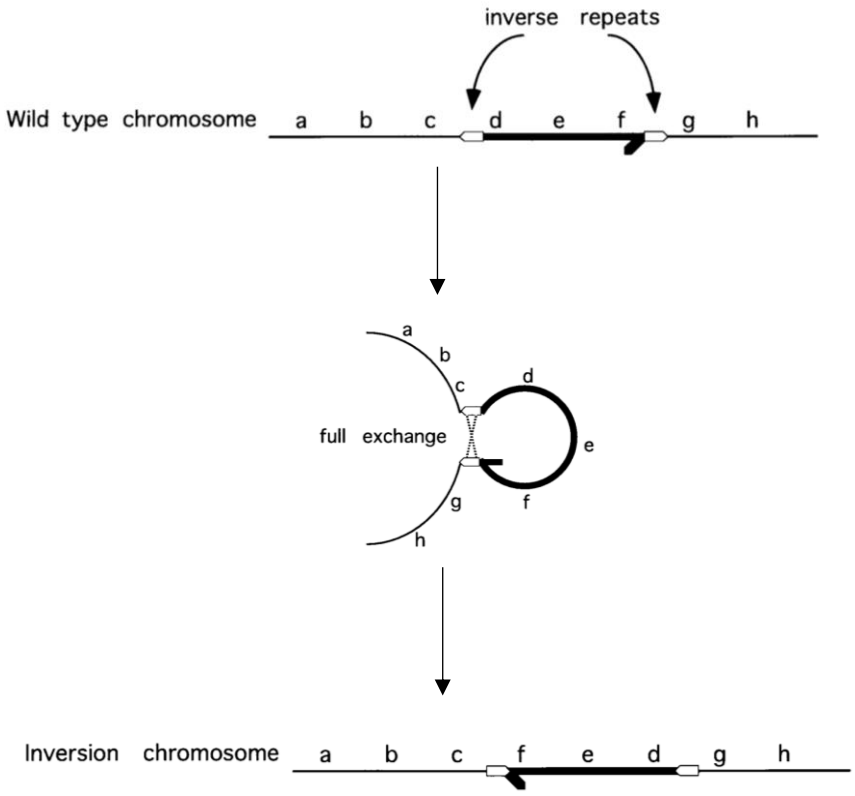


Figure 1.5: Formation of an inversion between the inverted repeat sequences [43]

Repeat sequences have some significant functional effects on the genome rearrangements [10]. Large number of repeat sequences including the IS elements and tandem repeats are found in the genome of *Neisseria* species. Repeats were found to be involved in three main inversions in the *Neisseria meningitidis* which has a greater number of repeats compared to the *Neisseria gonorrhoeae* [45]. Large repeat sequences play a significant role in the mechanism of antigenic variation in bacteria [37]. In bacteria, repeat-deficient genomes seem stable, and more repeats lead to more rearrangements [40, 46].

1.3.3 Genomic Islands

The segments of the genome that are acquired by the horizontal gene transfer (HGT) are known as genomic islands (GIs) [47] and they are present in several bacterial strains but missing in most of their closely related variants [48]. These regions range in size from 10 to 200 kb [47], similar regions smaller than 10 kb are called genomic islets [15]. Genomic features like GC content and skewness, the occurrence of small repeat sequences and codon usage of these regions (GIs) differ from the rest of the genome [47, 15]. Acquisition of the GI by the organism can affect its phenotype, activity and way of living. Genomic islands often include mobile genetic elements such as ISs, restriction modification or phage related genes, transposases and integrases. Some of the accessory genes are also encoded by GIs, which helps the bacteria in its adaptation to the new environment thus increasing the chances of its survival. The mobile genetic elements like insertion sequences present in the genomic islands can cause various genome rearrangements as a result of recombination [15]. For example, Piel *et al*., have shown that the mosaic structure of a '*ped* island' (genomic island that is present in one of the bacterial symbionts of beetles) might be because of the genome rearrangements [49]. Rearrangement hotspots have also been identified by Yan *et al.*, in the genomic islands of *Prochlorococcus*, a marine photosynthetic microorganism [50]. Figure 1.6 shows the schematic representation of how a genomic island looks like.

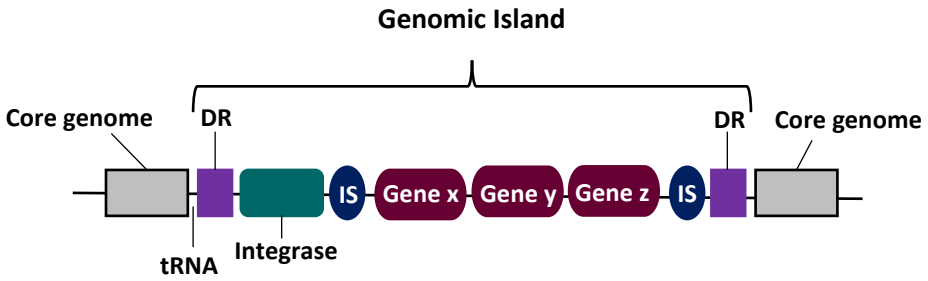


Figure 1.6: Schematic diagram showing the components of a genomic island

GIs can be classified into different types: a pathogenicity island, a metabolic island, a resistance island, a defense island, a symbiosis island, an ecological island, a fitness island or a saprophytic island, depending on the function of the genes encoded by them. Function of the similar GIs may vary depending on the bacterial species or the environmental conditions [15]. In short, GIs play an important role in the evolution of the bacteria and its adaptation to the challenging niche environment [51].

1.4 Effects of Genome Rearrangements

Genome rearrangements results in the change of ordering and orientation of the genes on the chromosomes [52]. As the genome rearrangements occur because of the breaks in the chromosome, it may or may not be deleterious depending on the genomic location of the breaks. If the breaks occur in the non-coding region it is likely to be non-deleterious however, if the breaks occur in the coding region then it might result in the loss of gene function causing the difficulty in the survival of an organism [4]. In prokaryotes, the effect of the genome rearrangements on the phenotype has been investigated [10]. Studies have reported that phenotypic outcomes are affected as a result of the genome rearrangements in variety of organisms [53-55]. The functional impact of the genome rearrangements is difficult to decipher because of their large size and complex nature [10]. In addition to this, the mechanisms involved in the formation of the genome rearrangements and their phenotypic impact is largely unclear [56]. However, some studies on prokaryotic genomes have shown the functional consequences of genome rearrangements [57]. Liang *et al.,* while investigating the strains of species *Yersinia pestis*, have shown that the large genome rearrangements are playing an important role in the evolution, pathogenicity and divergence of the organism. During evolution genome rearrangements can accumulate and can help in drug resistance as in the strains of *Yersinia pestis* [19]. However, in prokaryotic genomes some studies have reported the genome rearrangements to be deleterious [58]. As a consequence of genome rearrangements genes can be gained or lost, this can provide insight into the genome evolution at different time scale [59]. The functional impact of different genome rearrangements could vary greatly depending on the genomic region that is affected [10]. Among the genome rearrangements, inversions are the most frequently occurring [1]. They can be considered as one of the factors involved in the evolution of the genomes as reported by various studies [42]. The occurrence of inversion can vary greatly among the closely related species for example *Salmonella typhimurium* and *Escherichia coli* [10]. In general, the genome rearrangements result in the change of the ordering of the elements (e.g. genes) on the chromosome. Though, genome rearrangements are rare but with the passage of time they can accumulate and result in the completely different order of genes in the progeny compared to ancestral gene order. Therefore, the more conserved gene order can be observed for the closely related species compared to the distant ones [4]. Understanding the differences among the genomes in terms of gene orders can pave the way to comprehend the evolution of genomes [10].

1.5 Identifying Genome Rearrangements

The identification of genome rearrangements with the advanced techniques has improved our understanding of genomic structure and its organization in various organisms [60, 61]. The cytogenetic methods, like 'chromosome banding' and 'karyotyping' are among the classical approaches used to detect the genomic variations. The limitations of these methods include the low throughput and low resolution [62]. Genomic variants can also be detected by genotyping microarrays. One of the limitations of this approach is that only small copy-number variants can be detected [63]. The advent of 'optical mapping' technique has improved both the resolution and the scalability for detecting the genomic variations but it requires a reference genome [64]. Other techniques like 'DNA barcoding' and 'emulsion droplet PCR' have been used to detect the rearrangements and copy number variations, respectively [65, 66]. The understanding of genomic variations has been revolutionized by the availability of the whole genome sequences of large number of species and their strains [67]. Initially, methods for comparing the short sequences were designed to identify the similarities and difference by aligning the pair of sequences. These methods include the Needleman-Wunsch algorithm and Smith-Waterman algorithm for global and local alignment of the pair of sequences, respectively [68]. Similarly, for aligning more than two sequences various algorithms using the dynamic programming approach were introduced [69-71]. As the complexity of these algorithms increases with the increase in the length of the sequences, this makes these methods very time consuming [72]. Besides this, these alignment methods can only identify the local mutations but cannot detect the large-scale genome rearrangements [73]. The availability of the whole genome sequences of organisms have revealed that besides local mutations, their genomic structure varies greatly because of the global mutations. As a result of these large genomic variations, the structure of the chromosome differs in terms of gene orders among them. Through comparative genome analysis, differences in the gene order can not only be observed between species but also among the strains of same species [57]. Among the genome rearrangements, the inversions are the most common. As genome rearrangement events are rarer compared to the point mutations therefore they can provide insight into the important events that took place during the evolution of the organisms at different points of time [52]. Increase in the number of sequenced genomes along with the development of new algorithms for identification of the rearrangements facilitates the reconstruction of phylogenies [74]. Gene orders along with sequence data can facilitate more robust phylogenetic reconstruction.

1.6 Purpose and Organization of Dissertation

Several approaches for identifying the genome rearrangements have been proposed so far. These approaches have certain limitations that make the multiple genome comparison a bit difficult. This thesis focuses on providing an algorithmic approach to make the multiple genome rearrangement problem easier to solve. Using the gene order data, I provided an approach that has larger scalability in terms of number of genomes to be analyzed. This dissertation also focuses on identifying the possible drivers of genome rearrangement events and finding the association of these events with a particular variable.

Chapter 2, includes the overview of the theoretical approaches that were introduced to identify the genome rearrangements. It gives the brief introduction of most of the concepts that were introduced in the field of genome rearrangement identification. It discusses about the pair wise comparison approaches as well as the multiple genome rearrangement problem. Later, this chapter describes in detail the algorithm that I have developed to identify the genome rearrangements. The algorithm uses the gene order data to identify the genome rearrangements. The use of orthologous gene cluster information to determine the gene orders is described in detail in this chapter. The outputs of the algorithm are also discussed in the chapter.

Chapter 3 has two sections. The first section describes the demonstration of my algorithm using *Helicobacter pylori* genomes, as this bacterium has a diverse genomic structure. It also reports the classification of the inversions based on the geographical location of the genomes. It also reports the inversions associated with a particular disease state. Later, the section two of this chapter sheds light on identifying the underlying causes of the identified genome rearrangements. It describes in detail the association of various genomic elements with the identified inversions.

Chapter 4 focuses on understanding the genomic diversity of *Helicobacter pylori* genomes. It describes in detail the association of the genomic features more specifically the genome rearrangements with the disease outcome. It reports the difference in distribution of various genomic elements among the strains of the different groups defined on the basis of disease states.

Chapter 5 includes the general discussion and the conclusion of this dissertation.

Chapter 2

Algorithms

2.1 Overview

The availability of complete genome sequences of the large number of the organisms has broaden the scope of comparative genomic studies. In addition to the identification of point mutations by using the traditional alignment approaches, genome rearrangements can be identified by comparing the genomes at the level of gene orders [75]. In the late 1930s, Dobzhansky and Sturtevant became the pioneers of the genome rearrangement analysis in molecular biology [76]. They published an article describing the rearrangement scenario with the 17 reversals for the two *Drosophila* species: *Drosophila pseudoobscura* and *Drosophila miranda* [75]. In the late 1980s, Palmer *et al*., made a significant discovery by identifying the unique evolutionary patterns in the plant organelles. They compared the mitochondrial genomes of two closely related species (that have most of the genes with 99% identity): *Brassica oleracea* (cabbage) and *Brassica campestris* (turnip). They found that these genomes despite being almost identical at the sequence level differ greatly in terms of gene orders (Figure 2.1) [77]. This study along with several other studies in various organisms like virus, bacteria, plants and mammals have shown the genome rearrangements to be a common mode of evolution [78, 79].

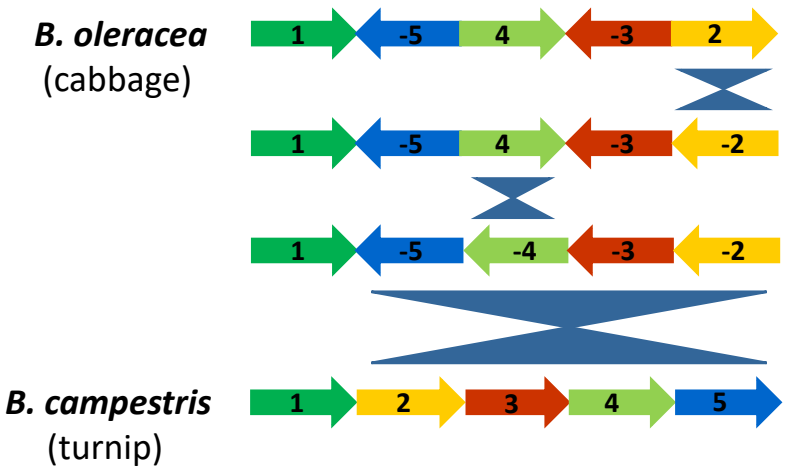


Figure 2.1: Cabbage to turnip transformation of conserved gene blocks as described in [77]

Genome rearrangements are one of the significant contributors in causing the diversity across genomes [75]. Comparison of the *Escherichia coli* and *Salmonella typhimurium* has shown their variable gene order to be the major difference between the two genomes [80]. Another study on herpes viruses has reported the series of genome rearrangements, especially the transpositions leading to the diverged genomes of these viruses [81]. Herpes viruses have a very rapidly evolving genome leading to the low similarity between genes thus making the traditional sequence comparison methods of no use for such diverged genomes. The number of genes in herpes viruses range from 70 to around 200 genes, forming the seven conserved blocks that are shared and rearranged in the different herpes viruses' genomes. For example, Cytomegalovirus (CMV) and Epstein-Barr virus (EBV) differ from each other with 5 reversals in terms of gene order which is way smaller compared to the point mutations identified between them. Therefore, analyzing the gene orders at the genomic level may complement the classical sequence level comparison [76].

Identification of the smallest number of reversals that transform one genome into another is like solving a combinatorial puzzle. It is possible to find the most parsimonious scenario of rearrangements for the genomes with the small number of blocks as Palmer *et al.*, were able to do for cabbage and turnip but for the genomes with larger number of blocks it is not possible to pen down all the possible scenarios [77]. Finding the most parsimonious evolutionary path is the interest of most of the scientists. The identified scenario may not represent the real evolutionary pattern but the number of identified rearrangement events can provide us with lower bound on the evolutionary events that took place [82]. Therefore, several approaches to solve this combinatorial puzzle has been proposed by a number of scientists.

2.2 Sorting by Reversals

A computational approach for comparing the genomes at the gene order level was pioneered by David Sankoff [83-85]. One of the combinatorial problems of sorting by reversals can be used to model the genome rearrangements [76]. In this approach, genes are represented as numbers from 1, …n and order of genes in the two genomes is represented in the form of permutations $\pi = (\pi_1 \pi_2 \dots \pi_n)$ and $\sigma = (\sigma_1 \sigma_2 \dots \sigma_n)$. A reversal affects the order of genes by reversing the segment at the same position and is represented as $\rho(i, j)$ [77]. When applied to a permutation it can transform it as follows [82]:

Permutation: $\pi = (\pi_1 \dots \pi_{i-1}\pi_i\pi_{i+1} \dots \pi_{j-1}\pi_j\pi_{j+1} \dots \pi_n)$ [82]

Reversals applied: $\pi \cdot \rho(i,j) = (\pi_1 \dots \pi_{i-1}\pi_j\pi_{j-1} \dots \pi_{i+1}\pi_i\pi_{j+1} \dots \pi_n)$ [82]

Below, the order of genes in a hypothetical genome are represented in a form of a permutation π and the transformed genome represented by π' indicates the effect of the applied reversals ρ (2,7) on π.

π = 1 3 7 10 5 8 4 2 6 9

ρ (2,7) = 1 3 7 10 5 8 4 2 6 9 (i = 2, j = 7)

π' = 1 4 8 5 10 7 3 2 6 9

i: 2 , j: 7

Indicate the index of number in permutation

2.3 Reversal Distance Problem

The problem of transforming one permutation into another with the shortest series of reversals is termed as *"reversal distance problem"*.

For example: π, σ are the two permutations and we want to transform π to σ. Thus, applying the series of reversals: $\rho_1, \rho_2, \rho_3, \dots, \rho_t$ such that $\pi \cdot \rho_1, \rho_2, \rho_3, \dots, \rho_t = \sigma$ and t should be minimum. t indicates the reversal distance between the two permutations π and σ and is denoted as $d(\pi, \sigma)$ [82].

In 1982, Watterson *et al*., for the first time defined the reversal distance problem for the circular permutations where the last gene is followed by the first gene [86].

2.4 Sorting by Reversal Problem

In this approach one of the genome's order represented by a permutation σ is arbitrarily set to the identity permutation 123…*n*. Later the other genome's order represented by a permutation π, is transformed to the identity permutation by applying the minimum number of reversals. Simple reversal sort approach is an example of a greedy algorithm that tries to bring every element to its position starting from 1 to *n*, thus sorting the permutation in *n*-1 steps. But, there

is no guarantee that the applied number of steps are the smallest to sort the given permutation or not [82].

For example: Given a permutation $\pi = 5\ 3\ 4\ 1\ 2$, applying the simple reversal sort will sort it in five steps as shown below:

$$5\ 3\ 4\ 1\ 2 \longrightarrow 1\ 4\ 3\ 5\ 2 \longrightarrow 1\ 2\ 5\ 3\ 4 \longrightarrow 1\ 2\ 3\ 5\ 4 \longrightarrow 1\ 2\ 3\ 4\ 5$$

The permutation π is sorted in *n*-1 steps, in this case 5-1=4 as show above. However, it can be solved in 3 steps:

$$5\ 3\ 4\ 1\ 2 \longrightarrow 5\ 3\ 4\ 2\ 1 \longrightarrow 5\ 4\ 3\ 2\ 1 \longrightarrow 1\ 2\ 3\ 4\ 5$$

Therefore, we can say that simple reversal sort might not give the correct solution as it takes *n*-1 steps to sort the permutation.

2.5 Concept of Breakpoints

Watterson *et al.*, in 1982 [86] and Nadeau and Taylor, in 1984 [87] introduced the concept of *breakpoint* for the first time in the field of computational studies of genome rearrangements. Some correlations between the number of breakpoints and the reversal distance was also observed by them [82].

In a permutation, a position where the two numbers occurring together are non-consecutive was defined as a *breakpoint* [88]. In a mathematical notation if the pair of numbers in a permutation π occur in such a way that π_i is followed by π_{i+1} then it is termed as *adjacency* but if π_i is not followed by π_{i+1} then it is called a *breakpoint* [76]. There are no breakpoints in an identity permutation. The number of breakpoints that can occur in any permutation π with *n* elements cannot be more than n+1 [89]. The example below shows the breakpoints in a permutation π.

$$\pi = 1\ 2|8\ 9|3\ 4\ 5|7\ 6|10$$

The red vertical lines in the above permutation indicate the occurrence of a breakpoint, so this permutation π has four breakpoints (2 8, 9 3, 5 7, 6 10) and five adjacencies (1 2, 8 9, 3 4, 4 5, 7 6).

Anchoring elements

Two elements 0 and n+1 are added in a permutation such that $\pi_0 = 0$ and $\pi_{n+1} = n + 1$, it is done to handle the boundaries [89]. Thus, 0 and n+1 can be called as anchoring elements [88]. While sorting the permutation both the anchoring elements 0 and n+1 should not be moved [82].

2.6 Greedy Algorithm

2.6.1 Breakpoint Reversal Sort Algorithm

It has been observed that at most two breakpoints are eliminated by every reversal that is applied on a permutation implying the reversal distance to be $d(\pi) \geq \frac{b(\pi)}{2}$, where $b(\pi)$ represents the total number of breakpoints in a permutation π. The breakpoint reversal sort algorithm, in order to get the identity permutation, tries to remove the number of breakpoints as many as possible in every step [82]. The simple workflow of the breakpoint reversal sort algorithm is shown below (Figure 2.2).

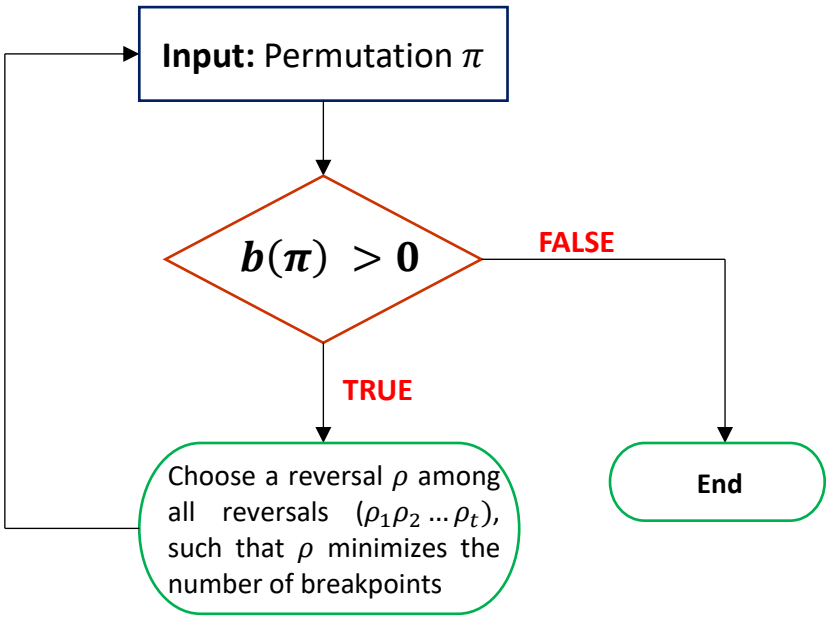


Figure 2.2: Flowchart of the breakpoint reversal sort algorithm adapted from [82].

One of the problems with this algorithm is that it might not terminate as it is not confirmed that when a breakpoint is removed it might not introduce the other breakpoints. Thus, the algorithm will get stuck in a never-ending cycle [82].

2.7 Concept of Strip

In a permutation a segment of consecutive numbers bounded by breakpoints is known as strip [88]. The strips can be further classified as increasing or decreasing strip. The strip consisting of a single element are always considered as decreasing strip but the anchoring elements 0 and n+1 are always defined as increasing strip [82]. Thus, there will be only one increasing strip in an identity permutation starting from 0 to n+1 [89].

The single increasing strip in an identity permutation will look like the one below:

$$\pi = 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11$$

The concept of increasing and decreasing strips is explained in the following example.

$$\pi = 0\ 1\ 2 \mid 8\ 9 \mid 5\ 4\ 3 \mid 7\ 6 \mid 10\ 11$$

In the permutation π written above the breakpoints are indicated by red vertical lines. There are five strips in this permutation. Three of them are increasing strips whereas two are decreasing strips. Below in the permutation π increasing strips are indicated by forwarded arrows (green) while the decreasing strips are shown by the backward arrows (blue).

$$\pi = 0\ 1\ 2 \mid 8\ 9 \mid 5\ 4\ 3 \mid 7\ 6 \mid 10\ 11$$

The increasing strips are 0 1 2, 8 9, 10 11 marked with green arrows and the decreasing strip are 5 4 3 and 7 6 marked with the blue arrows.

Kececioglu and Sankoff proved through a theorem[1] that the endless cycle in the breakpoint reversal sort algorithm cannot happen using the concept of strips [82].

2.8 Approximation Algorithm

Kececioglu and Sankoff presented another algorithm termed as improved breakpoint reversal sort for sorting the permutation. They proved[2] it to be an approximation algorithm with a performance ratio of 4 [82]. Figure 2.3 shows the workflow of this algorithm.

1. For details, readers can look into Chapter 5, page 134 of reference 82
2. See Chapter 5, page 135 of reference 82 for the proof

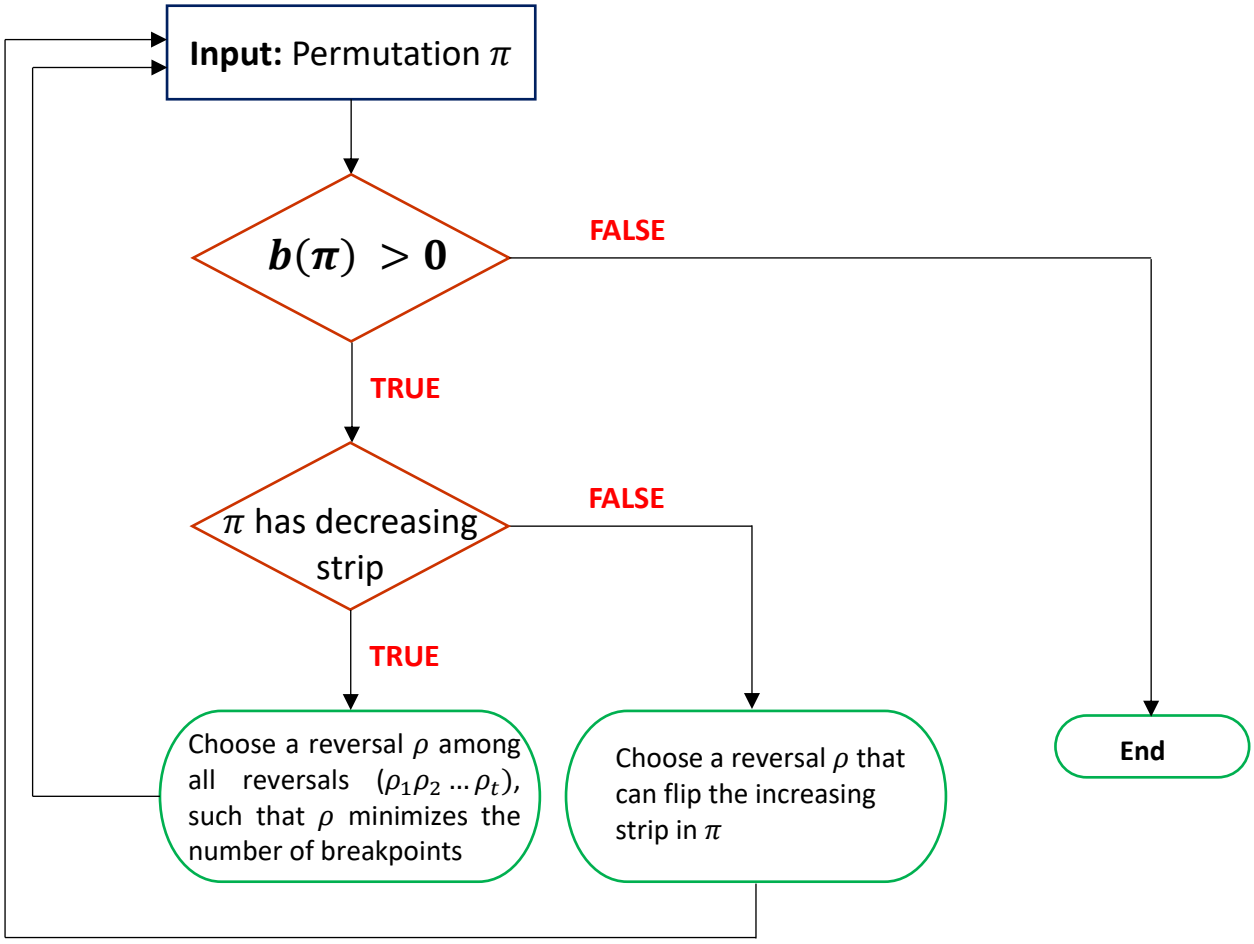


Figure 2.3: Flowchart of the improved breakpoint reversal sort algorithm adapted from [82].

In this theoretical approach they implied that the number of breakpoints in permutation π decreases until there exists a decreasing strip in π. Besides this, if there is no decreasing strip, a reversal might not reduce the number of breakpoints. Thus, by applying a reversal on the increasing strip will produce a decreasing strip which in turn will reduce the number of breakpoints in the next step as the reversal is applied on it. In the worst case this algorithm takes $2b(\pi)$ steps [82].

2.9 Types of Permutations

A permutation can be classified into two types, unsigned permutation or signed permutation. In a signed permutation each element has either a positive or negative sign [90]. As genes in a genome have a position as well as an orientation, they can be represented by the signed

permutation. In these permutations the reversal not only changes the position but also the direction of the numbers [88]. Beside this, in an unsigned permutation the number are without the positive and negative signs. All of the permutations previously mentioned in this chapter except the scenario shown in Figure 2.1 are the examples of unsigned permutation. Figure 2.1 shows the example of a signed permutation. It also shows how reversals effect the position and orientation of numbers in a signed permutation.

2.10 Exact Algorithm

An exact algorithm using the branch and bound approach was proposed by Kececioglu and Sankoff for sorting the unsigned permutation by reversals [89, 91]. In this branch and bound approach, among all the reversals they eliminated those reversals by which an optimal solution cannot be obtained [89]. In order to obtain the lower bound, they have introduced the linear programming technique. Besides this, for the identification of upper bound, among the reversals of the equal length consider those that remove the largest number of breakpoints [75]. By using this approach, Kececioglu and Sankoff were able to lower the number of steps considered but it remained exponential in the worst-case scenario. Exact solutions were found only up to n=30 by them [75]. They also gave the following two conjectures which they believed were true [89, 91].

Conjecture 1: An optimal set of reversals exists for every permutation that only cuts the strips at their first and last element.

Conjecture 2: An optimal set of reversals exists for every permutation that never increases the number of breakpoints.

These two conjectures were proved to be true by Hannenhalli and Pevzner [92]. For sorting the unsigned permutations by reversals, a polynomial time algorithm [92] was proposed by them for permutations that does not have a strip with one element. Thus, single element strips appear to be a major difficulty in obtaining an efficient algorithm for sorting the unsigned permutation [75].

In 1995 Kececioglu and Sankoff thought that problem of optimal sorting by reversals comes under the category of “NP-hard” computational problems [75]. Later, Alberto Caprara [93] proved that finding the optimal series of reversals for an arbitrary permutation is definitely NP-hard [75].

2.11 Breakpoint Graph and Cycle Decomposition

Bafna and Pevzner further studied the problem of sorting by reversals and introduced the concept of *breakpoint graph* of a permutation [77, 94]. They also identified the significant association between the reversal distance and the *maximum cycle decomposition* [94].

2.11.1 Breakpoint Graph

The *breakpoint graph* of a permutation π denoted by $G(\pi)$ is an edge-colored graph having n+2 vertices (0, 1, 2, …., n, n+1). Here, 1…. n represent the elements in a permutation whereas 0 and n+1 the two additional elements inserted are the anchoring elements which were described previously in this chapter. The two adjacent vertices (i, j) are connected by a *black* edge if it is a breakpoint (i and j are not consecutive) in a permutation. The two consecutive vertices (i, j) that are not adjacent in a permutation are connected by a *grey* edge [76, 77].

In an edge-colored graph G, a cycle in which every two consecutive edges are of a different color is called an *alternating cycle*. A graph is called a *balanced graph*, if for all the vertices 'v' the number of black and grey edges incident to every vertex are equal [76]. The number of black or grey edges in a cycle determines the length of the cycle C, represented by $l(C)$. A cycle of $l(C)$=2 is called short while a cycle of $l(C)$>2 is called a long cycle. A permutation with no long cycles in its breakpoint graph is called a *simple* permutation [77].

2.11.2 Cycle Decomposition

The edge-colored graph $G(\pi)$, being the balanced graph possesses an alternating Eulerian cycle. Therefore, a cycle decomposition of $G(\pi)$ must exist which decomposes it in such a way that every cycle has a distinct set of edges. The breakpoint graph should be decomposed into *maximum* number of cycles that are alternative as well as have a distinct set of edges. While estimating the reversal distance, cycle decompositions play an important role [76]. When a reversal is applied to a permutation, it might affect the number of breakpoints and the number of cycles in a maximum decomposition [94]. Bafna and Pevzner [95] showed that maximum cycle decomposition gives $d(\pi) \geq b(\pi) - c(\pi)$, which is a better bound for reversal distance as the parameter $b(\pi) - c(\pi)$ changes by at most 1 for every reversal that is applied to a permutation π [77].

Figure 2.4 illustrates the concept of breakpoint graph and cycle decomposition in a permutation π.

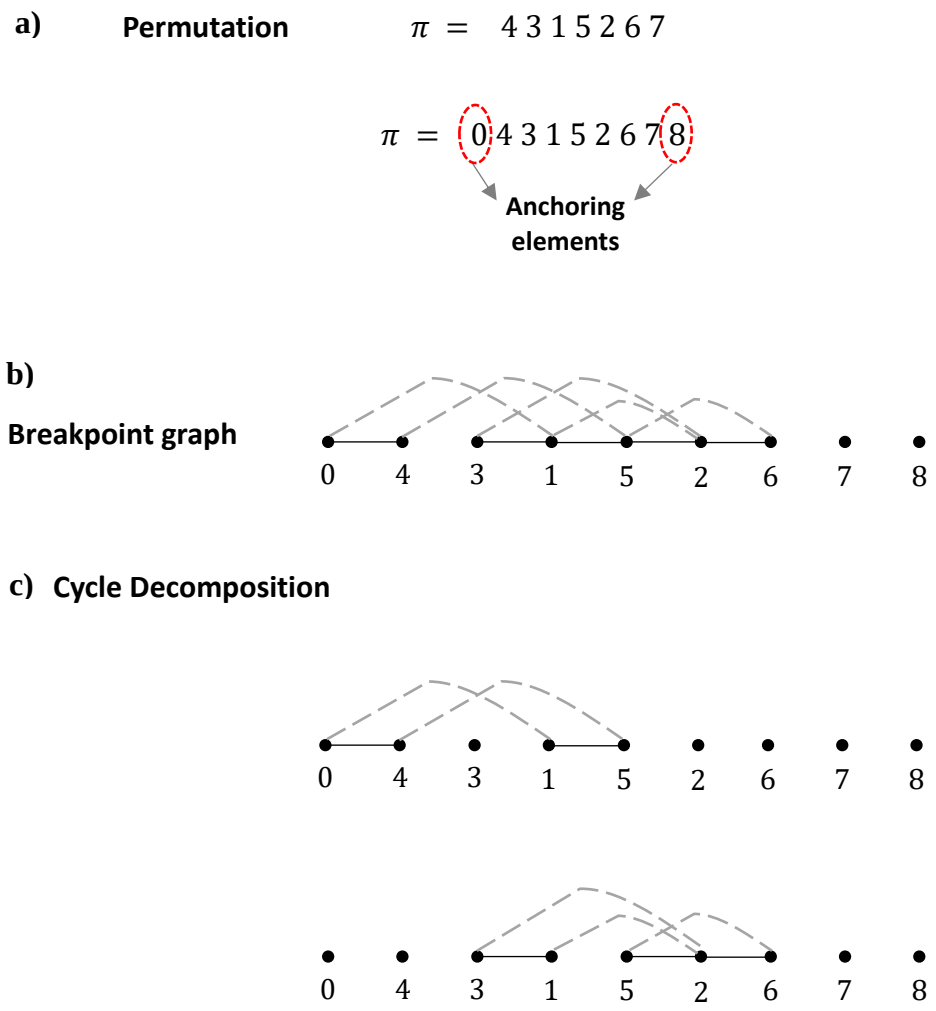


Figure 2.4: Breakpoint graph and cycle decomposition **a)** A permutation of numbers from 1, …,7 is represented by π. Anchoring elements 0 and 8 are added to the permutation. **b)** The breakpoint graph of π is shown, where black edges represent the breakpoints and grey edges represent the consecutive elements that are not adjacent. **c)** It shows the decomposition of the breakpoint graph of π into two cycles that are alternative as well as they have the distinct set of edges.

2.12 Breakpoint Graph for Signed Permutations

In 1996, Bafna and Pevzner observed that the signed permutations can be sorted by applying the concept of breakpoint graph. In order to do so, first a signed permutation must be transformed into an unsigned permutation. A signed permutation (π) consisting of n elements after transformation to an unsigned permutation (π') will have $2n$ elements. As in a signed permutation each element has either a + or – sign, so during transformation (to an unsigned permutation) considering the direction of elements the permutation is modeled in such a way that positive elements ($+x$) are replaced by $2x$-1 and $2x$, whereas the negative elements ($-x$) are replaced by $2x$ and $2x$-1 (Figure 2.5). The transformed unsigned permutation π' is called the image of the signed permutation [76,77].

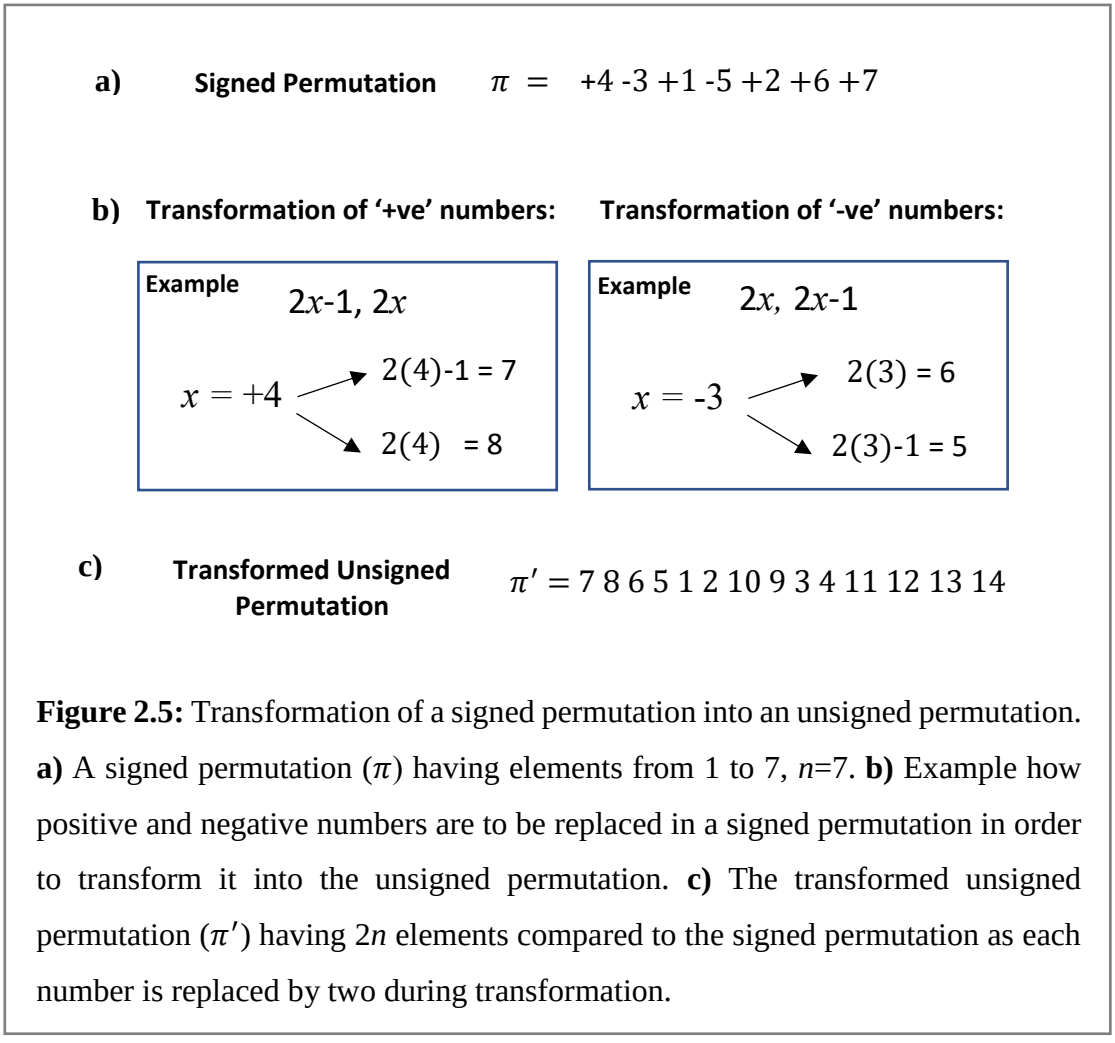


Figure 2.5: Transformation of a signed permutation into an unsigned permutation. **a)** A signed permutation (π) having elements from 1 to 7, n=7. **b)** Example how positive and negative numbers are to be replaced in a signed permutation in order to transform it into the unsigned permutation. **c)** The transformed unsigned permutation (π') having $2n$ elements compared to the signed permutation as each number is replaced by two during transformation.

When the breakpoint graph of a signed permutation is created there are both the black and grey edges for every element. A cycle of length two is created as a result of each of these pair of

black and grey edges. A maximum cycle decomposition of a breakpoint graph for a transformed permutation would definitely exist having all the cycles $c(\pi')$ of length two. The breakpoint graph of a signed permutation will have the disjoint cycles as each vertex has a degree of two (Figure 2.6) [76]. Bafna and Pevzner observed that the signed identity permutation of n elements can be mapped to an unsigned identity permutation having $2n$ elements. They implied that $d(\pi) \geq d(\pi')$ as they observed that the effect of a reversal applied on π can be imitated by the reversal on π'. The image π' of a signed permutation π can be sorted by applying those reversals $\rho(2i+1,2j)$ that cut only after the positions that are even. The effect of reversal $\rho(2i+1,2j)$ on π' can be imitated by a reversal $\rho(i+1,j)$ on π. Bafna and Pevzner also implied that if the cut by reversals are not allowed between π'_{2i-1} and π'_{2i} then $d(\pi) = d(\pi')$ [76, 77].

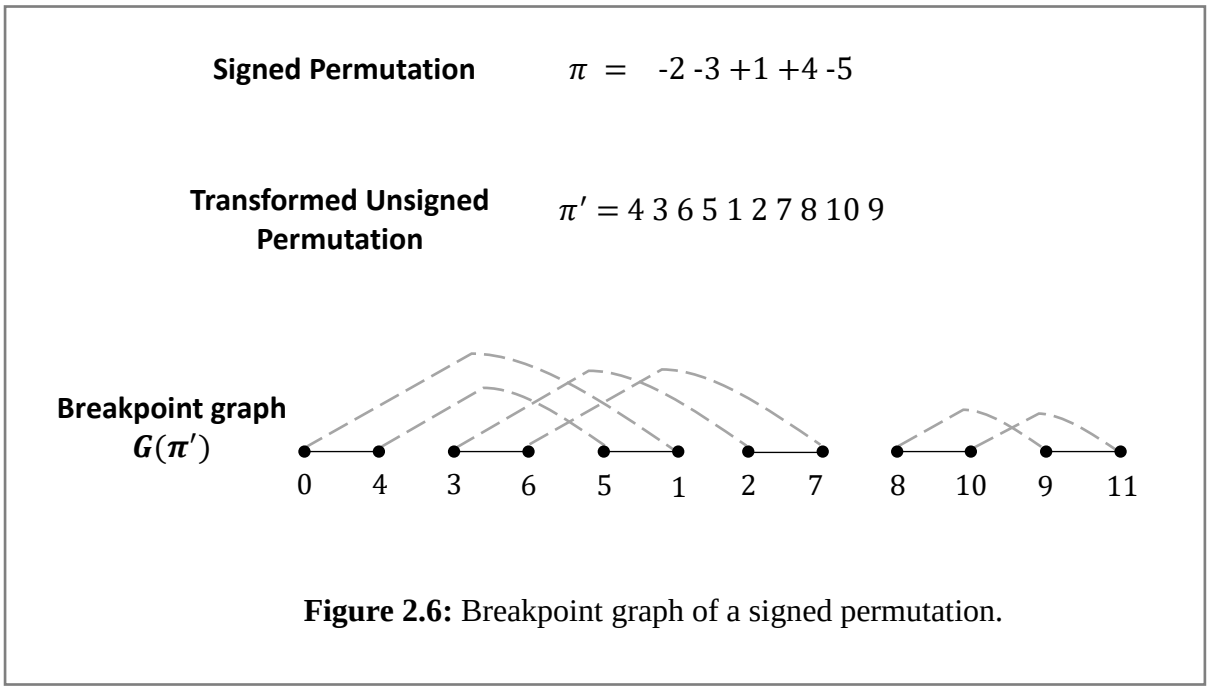


Figure 2.6: Breakpoint graph of a signed permutation.

They also proved [95] that the parameter $b(\pi) - c(\pi)$ is reduced by at most 1 for every reversal ρ that is applied on a permutation π [77]. A reversal was called *proper* by them if $\Delta c = 1$, where $\Delta c \equiv \Delta c(\pi, \rho) = c(\pi\rho) - c(\pi)$. Every permutation can be optimally sorted in n+1-$c(\pi)$ steps if a proper reversal can be found for a permutation. They also found that for some permutations a proper reversal might not exist so, they cannot be sorted in n+1-$c(\pi)$ steps [76]. They implied that there exist a third parameter beside number of breakpoints and maximum cycle decomposition which they called as "*hurdles*"[3], that makes it even more hard to sort a permutation. It was also observed that in a breakpoint graph the interleaving structure of long

3. For details, readers can look into Chapter 10, page 193-196 of reference 76

cycles present the important challenges in the analysis of genome rearrangements. To overcome these difficulties a new technique named *equivalent transformation* of permutations was developed by Hannenhalli and Pevzner [77].

2.12.1 Transformation of Permutations

They introduced the following technique to solve the problem of interleaving structure of long cycles. A permutation $\pi \equiv \pi(0)$ having a long cycle, can be transformed into another permutation $\pi(1)$ by breaking its long cycle into smaller ones. Repeat this process with the permutation $\pi(1)$ and continue this process which will result in a series of permutations $\pi \equiv \pi(0), \pi(1), \dots, \pi(k)$ and this process ends by giving a simple permutation having no cycles.

Let $C = \cdots, v_b, w_b, \dots, w_g, v_g, \dots$ be a cycle of the breakpoint graph $G(\pi)$ of a permutation π. The two edges, a black edge $b = (v_b, w_b)$ and a gray edge $g = (w_g, v_g)$ belong to the cycle C of a breakpoint graph $G(\pi)$ of a permutation π. A long cycle in a permutation can be transformed by a (g, b)-*split* of $\hat{G}(\pi)$ into a new $G(\pi)$ by following these steps [76, 77]:

i) Remove g and b edges from $G(\pi)$.
ii) Add the two new vertices w and v.
iii) Add (v_b, v) and (w, w_b), the two new black edges.
iv) Add (w_g, w) and (v, v_g), the two grey edges.

Figure 2.7 shows the (g, b)-*split* of $G(\pi)$ [76].

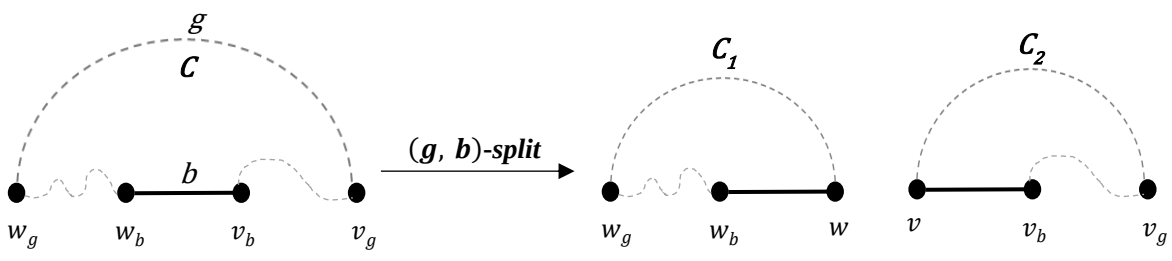


Figure 2.7: Graph splitting. Example showing the (g, b)-*split* of $G(\pi)$. The split has transformed a cycle C in $G(\pi)$ into two cycles C_1 and C_2 [76, 77]. The two new vertices w and v and the two new black as well as the two gray edges are added to transform one long cycle into two cycles.

Hannenhalli and Pevzner also introduced the concept of *generalized* [4] permutation and explained the procedure called *padding*[4] to identify it. They also explained how to search for the *safe*[5] reversals and *clear the hurdles*[6] in a breakpoint graph cycle decomposition [76, 77]. Thus, for sorting the signed permutations they provided the first polynomial time algorithm [77] with the time complexity of $O(n^4)$ for the permutation with n elements [75]. Later, many improved algorithms [96, 97] have been proposed for sorting the signed permutations by reversals.

2.13 Multiple Genome Rearrangement Problem

All the algorithms previously described in this chapter consider only two genomes and try to find the reversal distance between them. Sankoff *et al* were the first to investigate the molecular evolution using the rearrangement distance [85]. As previously mentioned in this chapter that for a given pair of genomes reversal distance can be computed in polynomial time however, its use in the multiple genome rearrangement problem is limited [98]. Later, multiple genome rearrangement problem was considered using breakpoint distance by Sankoff and Blanchette [99], whose objective is to find the most suitable tree which best represents the rearrangement scenario [8].

2.13.1 Breakpoint Distance

Breakpoint distance can be defined as the smallest number of breaks that when applied to one genome, transform it into another genome [8].

In 1999, Caprara [100] showed that the multiple genome rearrangement problem is NP-hard even in its simplest form, the Median Problem [98].

2.13.2 Median Problem

The problem involving the three unichromosomal genomes is termed as *Median Problem*. This problem involves the identification of ancestral genome having the minimum total reversal distance [98].

Blanchette *et al*; [101] and Sankoff and Blanchette [102], used the breakpoint analysis for minimizing the breakpoints in order to solve the Median Problem. Here, breakpoint occurs at those locations where the pair of elements are consecutive in one permutation but are not consecutive in the other permutation. The limitation of this analysis is that minimum number

4. For details, readers can look into Chapter 10, page 197-200 of reference 76
5. How to search for safe reversals, look into Chapter 10, page 200-204 of reference 76
6. How to clear the hurdle, look into Chapter 10, page 204-209 of reference 76

of reversal events cannot be determined by the breakpoint distance. Therefore, the median obtained by this analysis may not represent the ancestral median [98].

Bourque and Pevzner presented a greedy heuristic to create phylogenetic tree to find a reversal median. Their algorithm tries to identify the *good* reversals from all the possible reversals that can be applied on the given set of three genomes. They consider a reversal as a *good* reversal, if it reduces the distance between a genome and its ancestor. However, as the ancestor genome is unknown they claimed that a good reversal is the one that brings one genome closer to the other two. Thus, applying the good reversals in an iterative manner will transform all the three genomes into the ancestral genome [98].

2.13.3 Perfect Triple

The median problem that can only be solved by the good reversals was termed as the *perfect triple* by them. However, there may not be a good reversal for some cases, in that case their algorithm searches for a *best reversal* that minimizes the total pairwise reversal distance. Only if the good reversal is not found then the search for best reversal is done.

For more than three genomes (m>3), the good reversal is the one that minimizes the distance of a genome with m-1 genomes. In this case, again the good reversals are applied iteratively until the two genomes become identical, after this one of the identical genomes is removed and the process is repeated until the number of genomes becomes three meaning the median problem is obtained. When the number of genomes is large their algorithm may not be able to find the good reversal, for that case they have developed the heuristic[7] to resolve this problem [98].

2.14 Limitations of Previous Methods

Most of the above approaches use pairwise comparison, which transforms one genome into another assuming one as a reference and performing permutations on the other [75]. In terms of evolution, however, both genomes might have been affected by rearrangements in parallel. Therefore, multiple genome comparison is needed to identify which rearrangements are more ancestral [103]. Moreover, it is also assumed that all genomes have the same set of genes while comparing the multiple genomes. This indicates another limitation of the previous methods that they consider only the fully conserved genes while identifying the genome rearrangements.

7. For details, see reference 98

2.15 Multiple Genome Comparison

Increasing number of prokaryotic genomes and their comparison have revealed the presence of large number of genomic differences [20, 104]. Among many genomic variations, rearrangements are the most difficult to identify [10]. Multiple genome comparison for identification of genome rearrangements is of great importance as it not only helps to identify the most commonly occurring rearrangements but also the ones that are rare or specific for certain genomes. This in turn makes it easy to understand the role of genome rearrangements during the course of evolution.

The comparison of large number of genomes becomes easier using the gene order data in contrast to the gene sequence data. Genome rearrangements change the ordering of genes. Two genomes might appear functionally identical on the basis of the gene content but their gene order can be quite different because of rearrangements (Figure 2.8). As degree of genome rearrangements increases with time, therefore gene orders can be used to identify the genome rearrangements. Identification of the time course of rearrangements can provide insights into evolution.

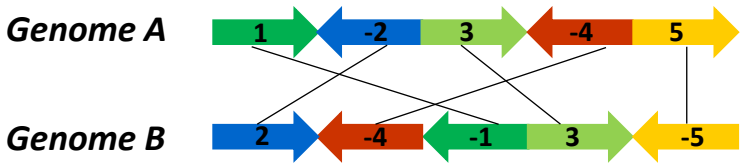


Figure 2.8: Gene order of two genomes. Two hypothetical genomes A and B having the same gene content but different gene ordering is shown. There exists an inversion of gene 5, inverse transposition of gene 1 and transposition of gene 3 in genome B, when compared with the genome A.

In the above figure 2.8 only the two genomes are compared but if we consider such a scenario for the multiple genomes, we can easily imagine the complexity of the problem.

I have developed an algorithm for identifying genome rearrangements while comparing multiple genomes of closely related strains. I have used the gene order data to identify the genome rearrangements in multiple genomes. Besides this, my approach considers not only the

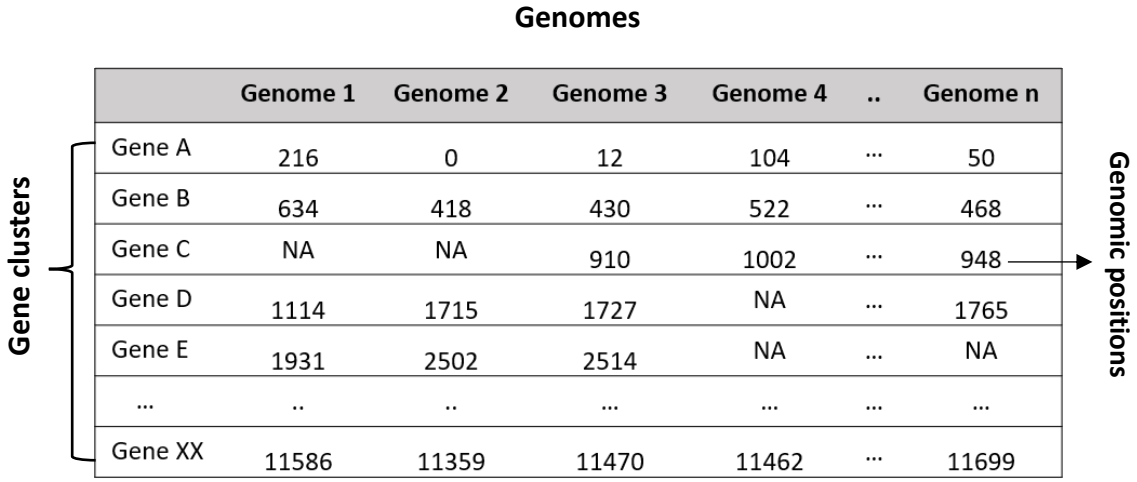
fully conserved genes but also the missing genes present in different genomes. Next section describes the details of my algorithm.

2.16 My Algorithm

The algorithm that I have developed follows a series of steps. Shortly, it compares the gene order of multiple genomes and identify the reversals that are shared by several genomes. Besides this, it also identifies the reversals that are specific for certain genomes. Following steps are carried out to identify the reversals while comparing the multiple genomes.

2.16.1 Orthologous Gene Clustering

Orthologs are one of the major types of homologs that are evolved by speciation from a common ancestor [105]. In the comparative genomics approach, identification of the orthologs or orthologous gene clusters can be used to elucidate the evolutionary patterns. To understand the variations in the genomic structure of the organisms it's important to compare the organization of the orthologous gene clusters [106]. In my approach, first the protein blast search is used and orthologous gene clusters are identified using the bidirectional best-hits criterion and the genomic position are recorded and represented as a gene table (Figure 2.9), as described by Tada *et. al* [107].



	Genome 1	Genome 2	Genome 3	Genome 4	..	Genome n
Gene A	216	0	12	104	···	50
Gene B	634	418	430	522	···	468
Gene C	NA	NA	910	1002	···	948
Gene D	1114	1715	1727	NA	...	1765
Gene E	1931	2502	2514	NA	...	NA
...	..	..	...	...	...	...
Gene XX	11586	11359	11470	11462	···	11699

Figure 2.9: Gene cluster table. Each row represents one cluster; each column represents a genome and each cell represents the genomic position of a gene in a particular genome. NA indicates the gene is absent in particular genome [171].

Gene table with the genomic position for each cluster is used in the next step to classify the gene clusters that are fully conserved, almost conserved and non-conserved among the several genomes that are being compared.

2.16.2 Selection of Gene Clusters

In the initial implementation of the algorithm, among all the clusters that are obtained after the orthologous gene clustering, only the *'almost conserved'* clusters were selected. Here, almost-conserved indicates clusters that are present in *n*-1 genomes (present in all the genomes except one) (Table 2.1).

Table 2.1: Example of almost conserved gene clusters in four genomes. The rows (clusters) encircled in red are the almost conserved gene clusters as they are present in all genomes except one [171].

	Genome 1	Genome 2	Genome 3	Genome 4
Gene A	216	0	NA	104
Gene B	634	418	430	522
Gene C	NA	NA	910	1002
Gene D	1114	1715	1727	NA
Gene E	1931	2502	2514	2434

The updated version of my algorithm can handle more than one missing gene. For example, it can run on the different set of genes that are conserved in ~85 to 100 percent of the genomes. This indicates that at these percentages there will be a large number of missing genes. The gene cluster are selected and filtered from the rest of the clusters by using my python script named removing_nonconserved.py.

2.16.3 Gene Order Identification

The identification of gene orders is important as it makes it easier to identify the genome rearrangements while comparing multiple genomes. In my approach, gene clusters that are selected in the previous step are used to identify the order of genes in multiple genomes. Gene clusters in one genome are numbered from 1 to *n* in the order of their genomic positions where *n* represents the total number of genes. Genes absent in this genome will obtain serial numbers

larger than *n*. Order of genes in this genome is used to assign the gene orders to the other genomes. In this way the genes in one cluster are represented by a same number in all of the genomes.

2.16.3.1 Rotation and Flipping

Ideally, for all the genomes gene order should start at gene '1' and end at the gene 'n', but some of the genomes might not have gene 1 at the start and gene n at the end. For these genomes, in order to bring gene 1 at the start and gene n at the end rotation and flipping of the gene orders is carried out. Figure 2.10 explains how the gene orders are rotated and flipped.

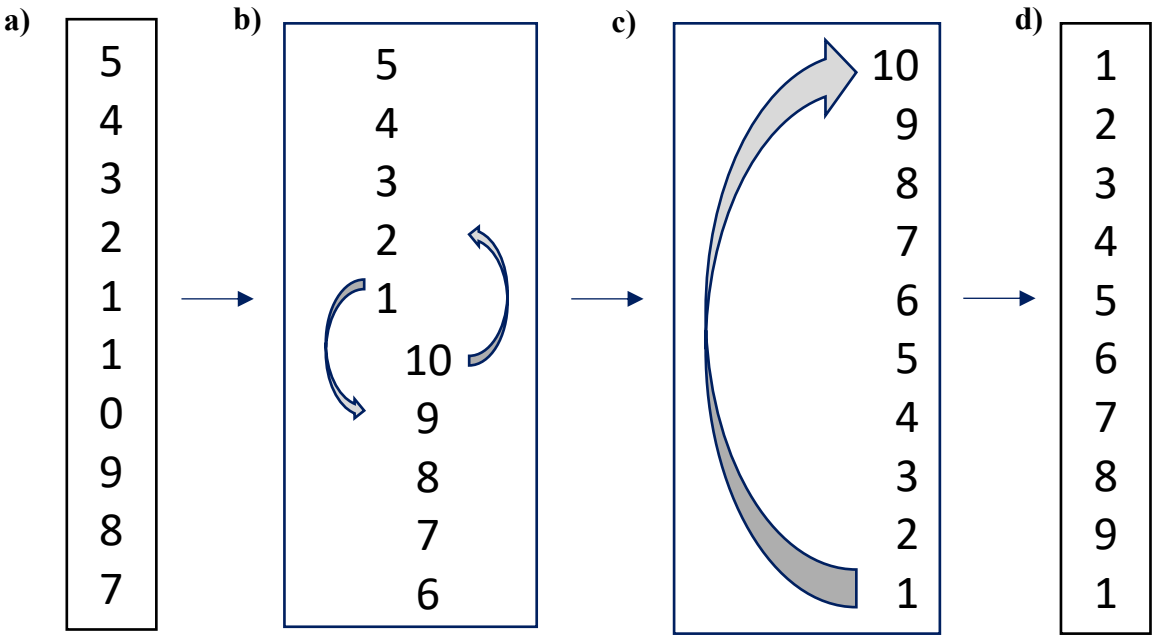


Figure 2.10: Gene order rotation and flipping example. **a**) Gene order neither starting from 1 nor ending at 10 (maximum number here). **b)** Gene order is flipped. **c)** Gene order is rotated. **d)** Gene order after flipping and rotation [171].

Gene order of some of the genomes might require both the rotation and flipping, however, some only require rotation or flipping in order to bring gene 1 at start and gene n at the end. After the gene orders are rotated and flipped, all the genome will have gene 1 at the start and gene n at the end of their gene orders.

Rotation and flipping makes it easier to align the gene orders which in turn helps to identify the genome rearrangements. In this step, first the gene orders are identified, then they are checked if the rotation or flipping is required or not using my python script named

gene_order.py. After this step the genomic position data of all the genomes transformed into the gene orders is obtained as shown by an example in Table 2.2.

Table 2.2: Example of gene order data. After the rotation and flipping all genomes have gene 1 at the start and gene n at the end of their gene orders [171].

Genome 1	Genome 2	Genome 3	..	..	Genome n
1	1	1	…	…	1
5	5	5	…	…	5
3	4	3	…	…	3
4	3	4	…	…	4
2	2	2	…	…	2
…	…	…	…	…	…
…	…	…	…	…	…
n	n	n	…	…	n

2.16.4 Rearrangement Identification

Gene orders identified in the previous step are used to identify the genome rearrangements. The identification of genome rearrangements involves the several steps shown in Figure 2.11.

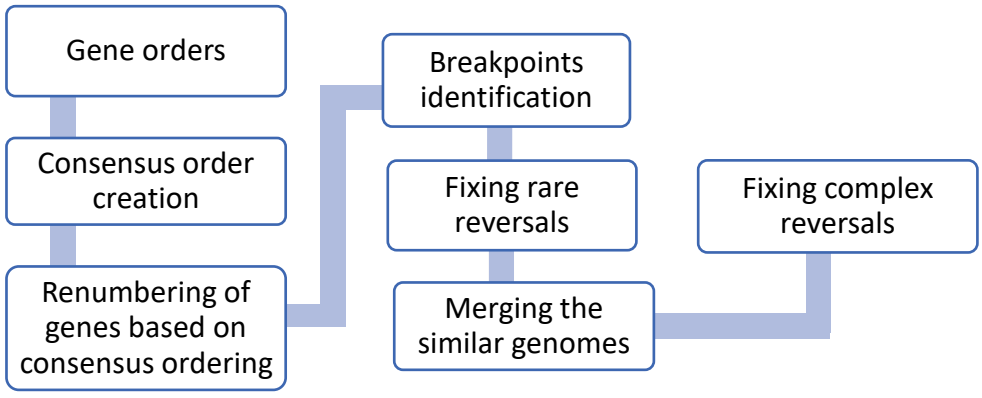


Figure 2.11: Workflow of the genome rearrangement identification process.

2.16.4.1 Creation of Consensus Ordering and Renumbering of Genes

In order to obtain the consensus ordering, for each gene: the most common upstream and downstream gene are identified and the consensus gene ordering for all almost-conserved genes is created by the majority rule (Figure 2.12a). Then, all genes are renumbered according to this consensus ordering (Figure 2.12b). Identification of the consensus gene order is important in finding the average ordering. Renumbering of all genes using the consensus ordering reveals the positional differences of orthologous genes, which correspond to the rearrangement events.

a)

Genome 1	Genome 2	Genome 3	..	..	Genome n	Consensus
1	1	1	...	...	1	1 → 1
5	5	5	...	...	5	5 → 2
3	4	3	...	...	3	3 → 3
4	3	4	...	...	4	4 → 4
...	...	...	...	...	...	.....
...	...	...	...	...	...	.....
n	n	n	...	...	n	n → n

b)

Genome 1	Genome 2	Genome 3	..	..	Genome n
1	1	1	...	...	1
2	2	2	...	...	2
3	4	3	...	...	3
4	3	4	...	...	4
...	...	...	...	...	...
...	...	...	...	...	...
n	n	n	...	...	n

Figure 2.12: Creation of consensus gene ordering. **a)** Consensus gene order is determined by majority rule for adjacent genes. **b)** Renumbering of genes in all the genomes on the basis of consensus gene ordering [171].

2.16.4.2 Identification of Breakpoints

In each genome, locations where gene numbers are gapped more than two are identified as breakpoints. Gain or loss of a single gene is not considered a breakpoint (Figure 2.13).

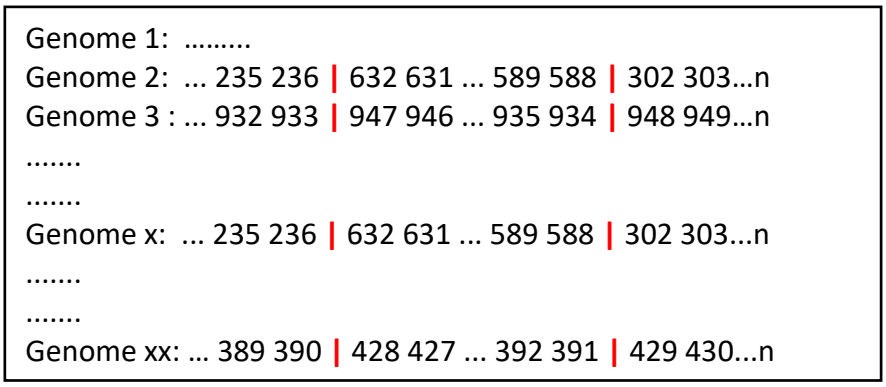


Figure 2.13: Breakpoints identification. Hypothetical gene order of multiple genomes where breakpoints are marked by the bold red vertical lines [171].

2.16.4.3 Detection of Rare Reversals

In this step the algorithm finds the reversals that are observed only in a single genome and fixes them. After fixing the rare reversal, the algorithm checks if multiple genomes share the same gene ordering, then they are merged and represented as one genome (Figure 2.14a).

2.16.4.4 Iteration of the Merger

The algorithm repeats the step 3 (detection of rare reversals) until all the rare reversals are fixed. After this the reversals that are shared are obtained. Some of the complex reversals are not resolved in this step (Figure 2.14b).

2.16.4.5 Complex Reversals

Initially, this step was performed manually but later I automated this step to resolve the complex reversals. The complex pattern of reversals is basically fixed by identifying and resolving the simplest of the reversals in the complicated gene order (Figure 2.15). After this step all the reversals are fixed. The comparison of the reversals identified in various genomes can show which reversals are shared among most of the genomes.

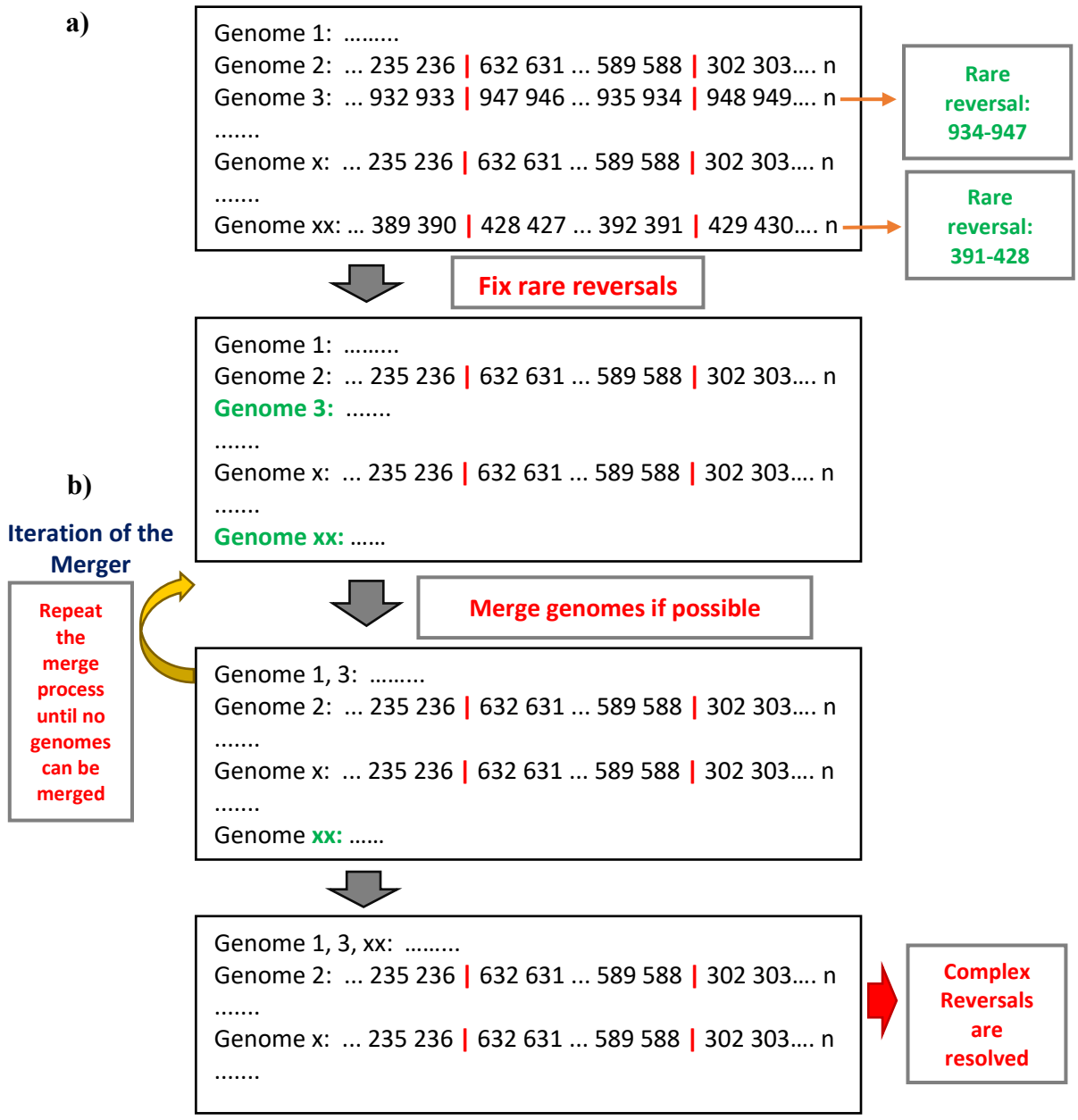


Figure 2.14: Step 3 and step 4 of the rearrangement identification process. **a)** The rare reversals shown in the hypothetical genome 3 and genome xx are identified and fixed. After the rare reversal is fixed genome 3's gene order becomes identical as that of genome's 1 so they are merged. **b)** Merging process continues to find if other genomes have the same gene order else the shared reversals are reported. Initially, complex reversals were resolved manually later it was automated. The hypothetical example of resolving complex reversals is shown in the Figure 2.15 [171].

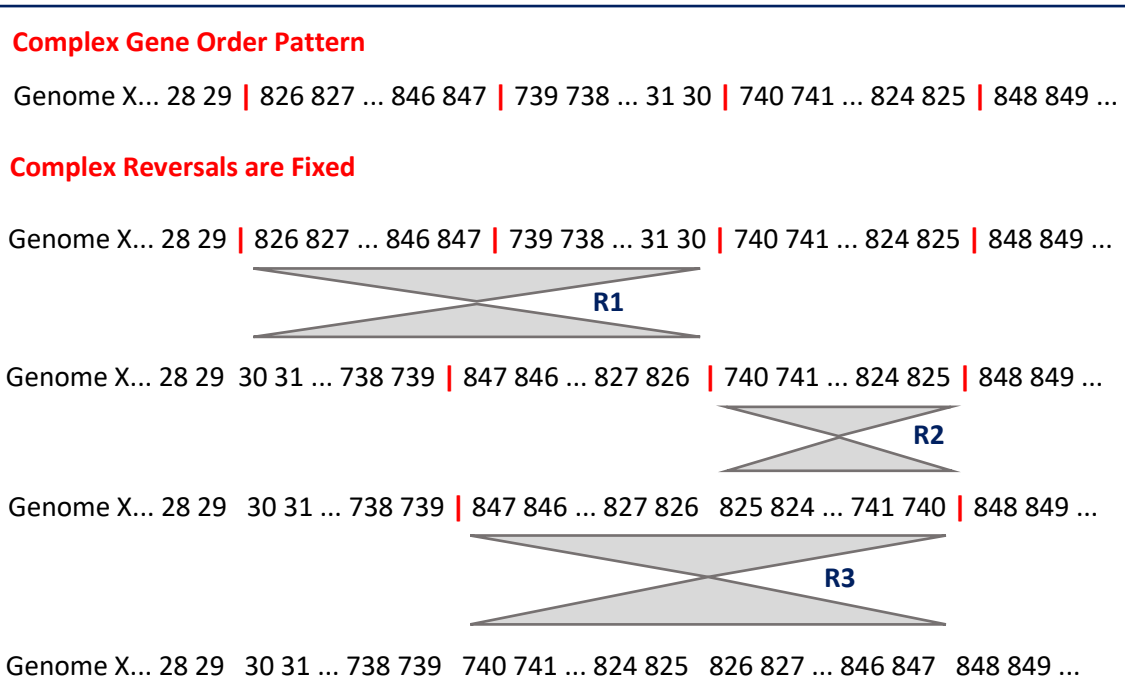


Figure 2.15: Complex pattern of reversals. Example showing the complex gene order pattern. This complex pattern is resolved by looking for the simplest possible reversal that can be applied. Here, R1 shows one of the simplest reversals that can be applied. Resolving of the simplest reversal can reduce the complexity of the gene order thus making it easier to fix the other reversals. Applying the two other reversals labeled as R2 and R3 the complex gene order pattern of genome X is fixed completely [171].

The output of the rearrangement identification process includes the following files:

i. The gene order of all the genomes renumbered based on the consensus gene ordering (Con_order.txt).

ii. The information about the location (gene number) after which the breakpoint occurs (outputbp.txt).

iii. The rare reversals that are fixed (Removed_Reversals.txt).

iv. Information of the iteration process and the genomes that are merged in every iteration (Merged_genomes.txt).

v. Reversals that are identified in all genomes (Shared_Reversals.txt).

vi. Information of the missing genes (Missing_genes.txt).

vii. Single gene transposition identified if any (Single_gene_transposition.txt).

2.17 Discussion

Genome rearrangements can be identified by comparing the genomes at the level of gene orders. Several approaches have been proposed to identify the genome rearrangements. These approaches beside being useful have certain limitations. I have developed an algorithmic approach to address some of these limitations. Table 2.3 shows the comparison of my approach with the others.

Table 2.3: Comparison of the tools

Tools	Multiple genome comparison	Set of genes	Reference genome	Orientation of genes	Type of genomes
Mauve [68]	Multiple, but the comparison of >30 genomes is difficult.	Whole genome sequences	Required	No	Bacteria
SPRING [173]	Pairwise only	Fully conserved genes	Required	Yes	Bacteria
GRIMM [174]	Pairwise only	Fully conserved genes	Required	Yes	Multi-chromosomes
This work	Multiple	Genes conserved in $\geq$ 85% of the genomes	Not required. (Consensus gene ordering only)	No	Bacteria

To demonstrate the use of my algorithm, I have tested it on the genomic dataset of the *Helicobacter pylori (H. pylori*) strains. The reason for selecting the *H. pylori* strains as a test data is explained in the next chapter. The next chapter includes the brief introduction of the *Helicobacter pylori* and the results of the genome rearrangements that were identified using the developed algorithm.

Chapter 3

Identification of Rearrangements and the Underlying Genomic Drivers

I have used *Helicobacter pylori* (*H. pylori*) species genomic data in order to test my algorithm. The section I of this chapter includes the brief description of *H. pylori*, describing about its general characteristics and the reason for taking this species as a model to test my algorithm. Later, section II of this chapter includes the detail of identifying the underlying genomic drivers of the genome rearrangements identified in *H. pylori* strains.

3.1 Section I

3.1.1 Overview

Helicobacter pylori is a Gram-negative, microaerophilic and spiral shaped bacterium [108]. It belongs to the class Epsilonproteobacteria [109] and is a member of family Helicobacteraceae. The *H. pylori* species are 0.5-1.0 μm wide having a length of 2.5-5.0 μm [110] and possess several flagella ranging from four to six which are required for its mobility [111]. Two Australian scientists, Marshall and Warren in 1980s made a great discovery by identifying this bacterium and elucidating its role in gastric diseases like peptic ulcer and gastritis [112]. This bacterium inhabits the human stomach [113] and is considered to have infected more than half of the human population [114]. Human being has been infected with *H. pylori* since its origin [115]. Its infection usually starts during a person's childhood and remains for their lifetime [116]. The mode of inheritance is still unclear, but *H. pylori* is considered to have co-evolved with *Homo sapiens* since its original migration "out of Africa" [117,118]. It can cause a wide range of diseases from mild gastritis to gastric cancer [108, 119].

3.1.1.1 General Characteristics

The whole genome sequencing of one of the *H. pylori* isolate 26695, made it the first bacterial species for which the genome was completely sequenced [120]. The person from which this

strain was isolated suffered from chronic gastritis. This strain has a genomic size of 1.67 Mbp and the average GC content of approximately 39 percent [109]. After two years another *H. pylori* strain J99, isolated from a duodenal ulcer patient was sequenced [121]. The genomic size of the strain J99 is 1.64 Mbp, a bit smaller compared to the strain 26695 [109]. Later, the genomes of two other strains of *H. pylori*, strain HPAG1 from chronic atrophic gastritis patient [122] and strain G27 were completely sequenced having the genomic size of 1.60 Mbp and 1.65 Mbp, respectively [123]. Large number of *H. pylori* strains have been sequenced to date, having an average genomic size of approximately 1.6 Mbp and ~40 % GC content. On average a genome encodes ~1500 genes along with 16S, 23S and 15S rRNA genes having more than one copy [121, 124].

3.1.1.2 Genomic Diversity

Genetic variability is one of the characteristics of *Helicobacter pylori* [125]. *H. pylori* is considered to be one of the most variable bacterial pathogens due to high mutation and recombination events [126]. The rate of mutation and recombination of *H. pylori* is one of the fastest among bacteria, possibly to enable its flexible host adaptation [126,127]. It has an open pan genome [128], and comparison of the two strains J99 and 26695 showed that they share around 1400 core genes with rearrangements [129], and that 6 to 7% of their genes are strain specific with gene gains and losses [130, 131]. The rearrangements observed by Alm *et.al*; [121] in the two strains of *Helicobacter pylori*, 26695 and J99 might be the cause of difference in the order of genes in these two strains [132]. The genetic diversity of *H. pylori* is related with the history of human migration [133] as it has been associated with humans for a long time, and has managed to survive in the challenging environment (human's stomach) which might have altered its genomic structure [109].

3.1.1.3 *Helicobacter pylori*, a Good Model

Greater diversity in the genomic structure and composition is observed in the *Helicobacter pylori* species [134]. This diversity is considered to be helpful for its survival and adaptation in different human populations [135, 136]. Compared to frequent genetic variations (mutations, insertions, or deletions), genome rearrangements (inversions and translocations) are rarer markers for delineating co-evolution. Genomic rearrangements keep the genetic repertoire intact without gene gain or loss [10] and theoretically do not alter *H. pylori's* survival fitness within the host. Flanking genes may be inserted or deleted in association with (or after)

rearrangements, but the evidence of large rearrangements is harder to erase from the genome than any other small-scale genetic variations.

To demonstrate the use of my algorithm described in Chapter 2, I have used *Helicobacter pylori* strains, as the species shows a diverse genomic structure and is a good model to study human migration across continents. The identification of genome rearrangements in *Helicobacter pylori* in the following analysis sheds light on not only the history of *H. pylori* but also of human beings after out-of-Africa.

3.1.2 Materials and Methods

3.1.2.1 Genome Sequences

Genome sequences of 73 *H. pylori* strains were obtained from NCBI/ENA/DDBJ repository. The strains belong to 8 different geographical locations:

1) East Asia annotated as: NY40, F30, ML3, ML1, UM299, UM298, UM032, UM037, UM066, F32, oki128, XZ274, OK310, 52, F16, oki673, oki154, oki828, oki898, oki112, oki102, oki422, F57, 26695-1CH, 26695-1CL, 26695–1, Hp238, OK113

2) South America annotated as: Sat464, Shi112, Shi169, Shi417, Cuz20, PeCan18, PeCan4, Puno120, Puno135, SJM180, v225d

3) North America annotated as: 7C, 29CaP, Aklavik117, Aklavik86, 26695–1, 26695-1MET, J166, J99, ELS37

4) Europe annotated as: B38, B8, HUP-B14, Rif1, Rif2, 26695, P12, 26695, G27, Lithuania75, 2017, 2018, 908

5) Africa annotated as: SouthAfrica20, SouthAfrica7, Gambia94/24

6) India annotated as: India7, Santal49

7) Australia annotated as: BM013A, BM013B, BM012A, BM012B, BM012S

8) others of unknown location annotated as 83 and 35A.

Detailed information regarding the strains is available (Table A.1 and Figure A.1 of Appendix). One strain was registered twice: strain 26695 by TIGR and strain 26695–1 by Oita University.

3.1.2.2 Orthologous Gene Clustering

Protein BLAST (version 2.2.29+, e-value<1e-5) was applied for 73 *H. pylori* strains and results were used to obtain the orthologous gene clusters through the bidirectional best-hits criterion as described by Tada *et al* [107]. For each gene cluster, its genomic position was recorded and represented as a gene table. Each column in the gene table represent the strain and the rows contained the gene cluster information. Each cell holds a genomic position of a particular orthologous gene present in a strain whereas if the orthologous gene is absent in a particular strain then it is represented by "-".

3.1.2.3 Phylogenetic Analysis using Core Genes

Phylogenetic analysis was performed using 900 core genes of 73 *H. pylori* strains obtained from the clustering result. Core genes were aligned using MAFFT (version 7.313) [137], alignments were trimmed using trimAl [138] with default parameters, which were later concatenated and phylogenetic tree was obtained using standard- RAxML-master with the parameters: -T 11, −N 1000, −m PROTCATBLOSUM62 [139].

3.1.2.4 Gene Order Identification

As reported by Tada *et al*; strain Aklavik86 was very different from other *H. pylori* strains, maybe because of sequencing anomalies [107]. This strain was excluded from the genome rearrangement analysis. For the remaining 72 strains, gene orders were identified using the gene clusters information. The table generated with the genomic positions for each gene cluster was used as an input. Out of all the gene clusters, 'almost conserved' clusters were considered. Here, almost-conserved indicates clusters that were present in all the strains except one

First, all genes in the P12 strain (used as an initial reference because analysis by Furuta *et al*. [59] reported no inversions in the P12 strain) were numbered from 1 to *n* in the order of their genomic positions where *n* represents the total number of genes. Genes absent in the reference strain obtained serial numbers larger than *n*. The gene order of the P12 strain was then used to obtain gene orders in other strains. In order to place '1' at the start and '*n*' at the end of all the strains, gene orders of some strains were rotated and flipped. After rotation and flipping, gene 1 was located at the start and gene *n* (last gene) at the end. For more details about the process of identifying the gene orders, rotation and flipping see the previous chapter (Chapter 2 of this dissertation).

3.1.2.5 Rearrangement Identification

Genome rearrangements were identified for the 72 *H. pylori* using the algorithm that I have developed. The algorithm is described in detail in Chapter 2, here I briefly describe how the rearrangements were identified. Rearrangements were identified as follows. 1) Creation of the consensus ordering: for each gene: the most common upstream and downstream gene are identified and the consensus gene ordering for all almost-conserved genes was created by the majority rule. Renumber all genes according to this consensus ordering. 2) Identification of breakpoints: in each strain, locations where gene numbers are gapped more than 2 are identified. Gain or loss of a single gene is not considered a breakpoint in this study. 3) Detection of rare reversals: find reversals that are observed only in a single strain and fix them. When multiple strains share the same gene ordering as a result of this process, then merge them. 4) Iteration of the merger: repeat the Step 3 (fixing and merging process) until all remaining reversals are shared. Later, the complex reversals were resolved as described in Chapter 2.

3.1.2.6 Rearrangement Based Phylogeny

The inversions identified by the program were manually curated to obtain the phylogenetic tree reflecting the inversion history of *H. pylori*. Rearrangement based phylogeny was manually created by referencing the program output of the rearrangement identification. Figure 3.1 describes the workflow of the process.

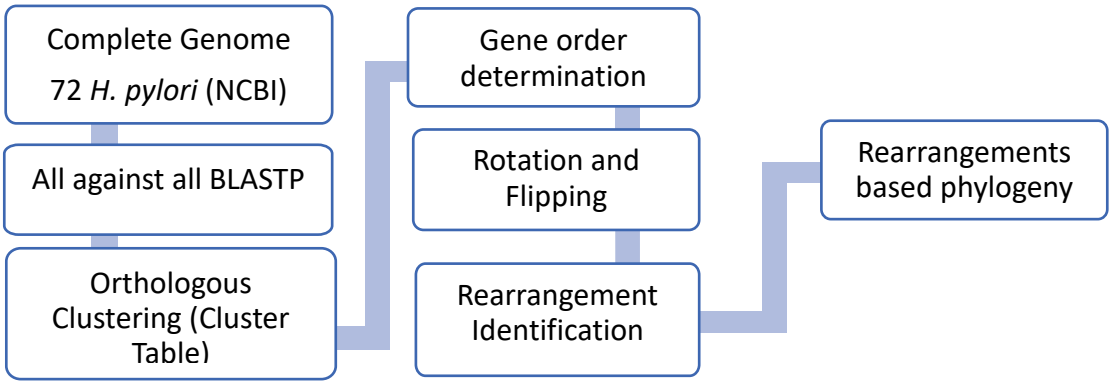


Figure 3.1: Workflow describing the methodology used for the analysis of the 72 *Helicobacter pylori* genomes obtained from the NCBI database.

3.1.3 Results

3.1.3.1 Orthologous Clusters and Gene Orders

For 72 *H. pylori* strains (excluding Aklavik86 strain, see Methods), 1856 orthologous gene clusters were obtained. Among these 749 clusters were fully conserved (core genes) and 972 were almost-conserved gene clusters (see Methods; Figure. 3.2). Taking the P12 strain as a reference, gene order data for the almost conserved gene clusters of 72 strains were identified. In this gene ordering, 15 strains did not possess the gene 1 at the start and gene *n* at the end. Among these 15 strains, gene order of 12 strains were rotated and flipped whereas gene order of 3 strains required flipping to align their gene orders. Information of strains whose gene order were rotated and flipped is given in Table 3.1.

Table 3.1: Information of the operation on gene order of the 15 strains [171].

Strain	Operation on gene order	Geographical region
B8	Rotation and flipping	Europe
35A	Flipping	Not known
UM032	Rotation and flipping	East Asia
UM299	Rotation and flipping	East Asia
UM037	Flipping	East Asia
UM066	Rotation and flipping	East Asia
UM298	Rotation and flipping	East Asia
NY40	Flipping	East Asia
ML1	Rotation and flipping	East Asia
ML3	Rotation and flipping	East Asia
oki128	Rotation and flipping	East Asia
oki154	Rotation and flipping	East Asia
oki673	Rotation and flipping	East Asia
oki828	Rotation and flipping	East Asia
J99	Rotation and flipping	North America

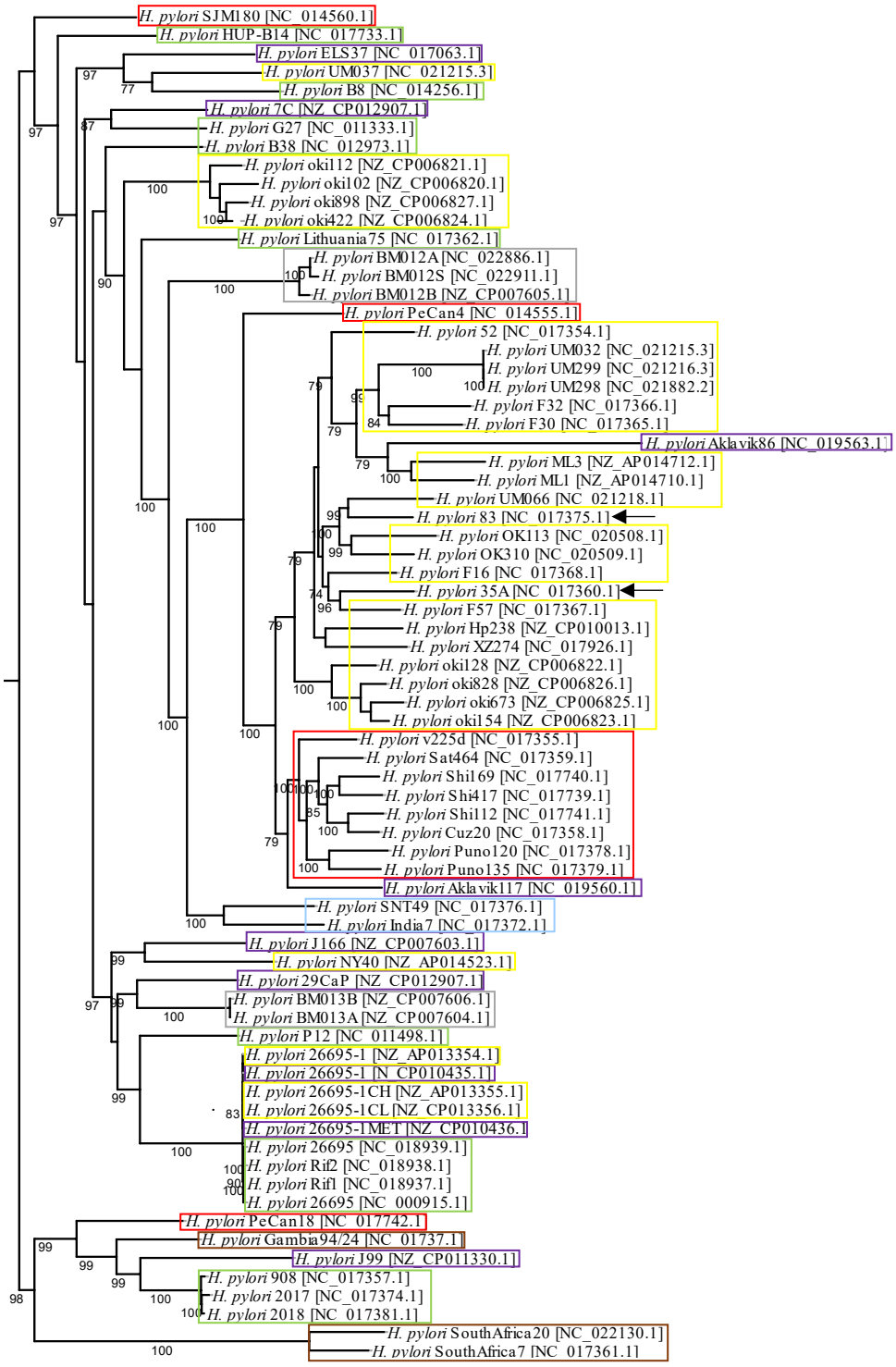


Figure 3.2: Phylogenetic tree based on the core genes of 73 *H. pylori* strains. Colored boxes represent the geographical region of the strains (Yellow: East Asia, Red: South America, Purple: North America, Green: Europe, Brown: Africa, Light Blue: India, Grey: Australia). Black arrows indicate strains with no geographical information [171].

3.1.3.2 Rearrangement Analysis

Gene order data of 72 *H. pylori* strains was used as the input. Identification of the consensus gene order was important in finding the average ordering. Renumbering of all genes using the consensus ordering revealed the positional differences of orthologous genes, which correspond to the rearrangement events. The number of breakpoints in each strain ranged from 0 to 10 (Table 3.2).

Table 3.2: Number of breakpoints identified in each strain [171].

No. of Strains	Breakpoints	Strains annotation
6	0	P12, Shi417, Shi169, Puno135, Cuz20, Lithuania75
1	1	Aklavik117
9	2	G27, PeCan4, SJM180, Sat464, Santal49, Puno120, Shi112, BM013A, BM013B
4	3	v225d, oki154, oki673, oki828
10	4	B38, 908, F30, 2017, OK113, NY40, ML3, J99, 7C, 29CaP
8	5	B8, Gambia94/24, 2018, oki102, oki112, oki128, oki422, oki898
9	6	ELS37, 52, F57, HUP-B14, PeCan18, SouthAfrica20, ML1, J166, Hp238
2	7	SouthAfrica7, India7
17	8	26695, 35A, F16, 83, XZ274, Rif1, Rif2, 26695, OK310, UM032, UM299, UM298, 26695-1, 26695-1CH, 26695-1CL, 26695-1, 26695-1MET
0	9	-
6	10	F32, UM037, UM066, BM012A, BM012S, BM012B

Total 41 inversions were identified, which included strain specific as well as shared inversions. Number of inversions in each strain ranged from 0 to 6. It was assumed that the strains with no inversion are closest to the tree root (not necessarily ancestral) and that the strains with 6 inversions are the farthest from the root (Table 3.3).

Table 3.3: Number of reversals (inversions) identified in each strain [171].

No. of Strains	No. of Reversals	Strains
7	0	Lithuania75, P12, Aklavik117*, Shi417, Shi169, Puno135, Cuz20
12	1	BM013A, BM013B, G27, oki154*, oki673*, oki828*, PeCan4, Shi112, SNT49, Puno120, Sat464, SJM180
15	2	29CaP, B38, ML3, oki128*, OK113, F30, v225d, 908, Gambia94/24*, 2017, 2018*, SouthAfrica20***, NY40, J99, 7C
13	3	B8, 52, Hp238, ML1, oki102, oki112, oki422, oki898, F57, ELS37, SouthAfrica7**, HUP-B14, PeCan18
9	4	OK310, UM032, UM299, UM298, XZ274, 83, 35A, F16, India7
12	5	26695, 26695-1, 26695-1MET, 26695-1, 26695-1CH, 26695-1CL, Rif1, Rif2, 26695, J166, UM066, F32
4	6	BM012A, BM012S, BM012B, UM037

* ignoring single gene transposition

** ignoring single gene transposition, 2 gene inverse transposition

*** ignoring single gene transposition, 2 gene inverse transposition and 3 gene deletions

Of total 41 inversions, 18 were found strain specific whereas 23 were shared (Table 3.4). Among all the inversions R17-R21, R22 and R23 were observed in strains from the same geographical locations such as Australia, East Asia and Africa, respectively. These inversions are called *region-specific* in this analysis. Figure 3.3a, illustrates the distribution of the inversions in each geographical location and Figure 3.3b describes the shared, strain-specific and region-specific inversion.

Table 3.4: Strain specific, shared and region-specific inversions. Dark grey: Strain specific inversion, Light grey: Shared inversions, Light Blue: Region-specific inversions [171].

Inversion No.	Inversion Block*	No. of strains	Geographical location of strains**
R1	237-632	9	EU, EA, NA
R2	302-587	9	EU, EA, NA
R3	934-947	18	EU, EA, NA, IN, SA, AU
R4	845-924	9	EU, EA, NA
R5	845-933	11	EU, EA, NA, AF
R6	313-530	27	EU, EA, SA, UN, IN, AU, NA
R7	30-825	5	EU, EA
R8	30-847	5	EU, EA
R9	391-428	9	EA, SA, UN
R10	417-420	8	EU, EA, NA, AF, SA
R11	855-858	8	EU, EA, NA, SA
R12	30-817	16	EA, UN
R13	14-844	17	EA, UN
R14	49-74	7	EA, UN
R15	325-328	2	EA, AF
R16	766-773	2	EA, EU
R17	229-587	3	AU
R18	669-858	3	AU
R19	30-826	3	AU
R20	740-825	3	AU
R21	740-847	3	AU
R22	30-933	4	EA
R23	717-818	2	AF
R24	846-847	1	NA
R25	826-845	1	NA
R26	588-825	1	NA
R27	588-847	1	NA
R28	112-669	1	NA
R29	925-933	1	NA
R30	213-619	1	EA
R31	112-134	1	IN
R32	134-699	1	IN
R33	848-933	1	IN
R34	837-933	1	SA
R35	173-645	1	EA
R36	870-924	1	EA
R37	329-468	1	EA
R38	329-560	1	EA
R39	127-924	1	EA
R40	5-742	1	EA
R41	859-933	1	SA

* Inversion column has the start and end gene number of inversions assigned by program.
** Geographical regions are abbreviated as, EA: East Asia, SA: South America, NA: North America, EU: Europe, AF: Africa, IN: India, AU: Australia, UN: Unknown

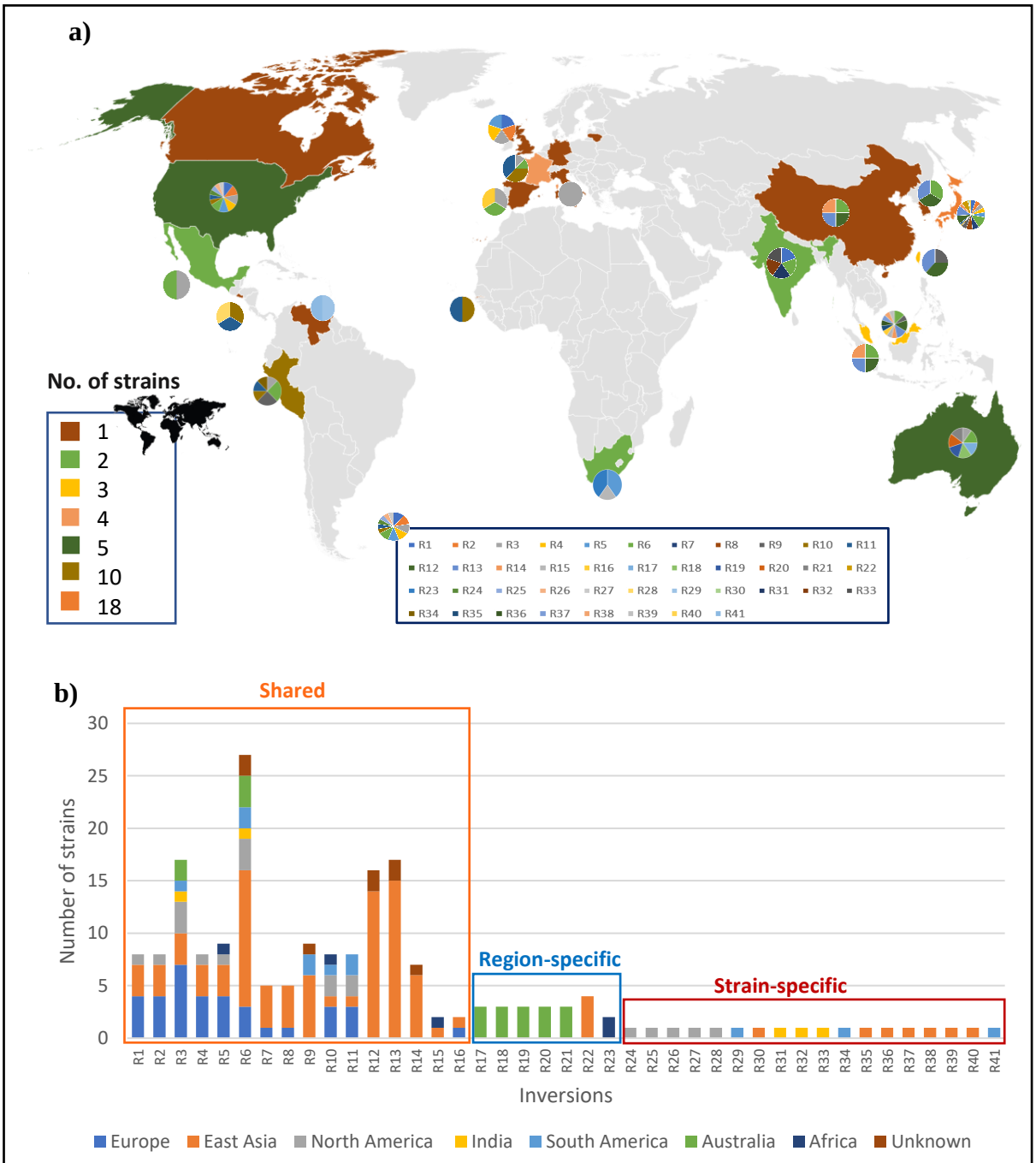


Figure 3.3: Distribution of inversions. **a)** Different color of regions in the map correspond to the number of strains included in this analysis. Pie chart along each region shows the distribution of the inversions in strains of that region. **b)** R1-R23 and R24-R41 were identified as shared and strain-specific inversions, respectively. Among the shared inversions, R17-R23 were region-specific [171].

Strains from Europe and East Asia shared as many as 11 inversions (R1-R8, R10, R11 and R16). Out of these 11 inversions, R7, R8 and R16 were found within them only and R1, R2 and R4 were in common with the strains from North America. Inversions R3, R5, R6, R9-R11 were shared with strains from other geographical areas (Figure 3.3b). The identified inversions were of different sizes. The three large inversions (R22, R13, R8) were identified in the East Asian strains. The largest inversion (R22) was found in 4 East Asian strains from Okinawa Japan.

Furuta *et al*. identified inversions in 10 *H. pylori* strains and proposed a mechanism of DNA duplication linked to the chromosomal inversions [59]. Our analysis also included seven of these strains (26695, G27, P12, F16, F30, F32 and F57), and 10 inversions (R1-R4, R6, R9, R12-R14 and R30) identified in these strains were similar to those reported by Furuta *et al*. (Table 3.5) [59]. Since we did not perform analysis at the DNA sequence resolution, true identity of inversion requires further sequence-level analysis. For 29 strains, inversion breakpoints were examined to identify the possible cause of the rearrangements. 19 strains possessed insertion sequences (IS), 10 possessed integrated conjugative elements (ICEs), and 7 possessed virulence related genes and pathogenicity island proteins around their inversion's breakpoints.

Table 3.5: Rearrangements that were in common with the previous study [171].

Strain	Inversion Label (my study)	Inversion (label) identified by Furuta *et. al* [59]
F16	R9, R12, R6, R13	A, C2, F, C1
F32	R9, R12, R6, R13, R30	A, C2, F, C1, G
F30	R13, R14	C1, D1
F57	R12, R13, R14	C2, C1, D1
G27	R3	I^e
26695	R3, R4, R2, R1	I^e, $M1^e$, $M3^e$, $M4^e$

3.1.3.2.1 Rearrangement Hotspots

Some regions were frequently involved in rearrangements and called *'rearrangement hotspots'* [128, 140]. Three such regions were identified in the analyzed strains. Breakpoints within these

regions were found to have IS, ICE, repeats, virulence related genes and restriction modification system proteins. Even if two inversions share a common breakpoint, however, the mobile elements around them were sometimes different or strain specific (Figure A.2 of Appendix).

3.1.3.2.2 Phylogenetic Tree Based on Inversions

Information of inversions that occur during the evolution was used to create a phylogenetic tree. First, the matrix representing the presence or absence of all inversions in each strain was constructed (Table A.2 of Appendix). Then the tree was created to reflect the evolution of *H. pylori* strains from different geographical locations (Figure 3.4). Some of the inversions (R3, R6, R9, R12, R13 and R14) occurred more frequently and were present in multiple strains. R3 and R6 were found in strains from all geographical locations except for Africa. R9 was found in strains from South America, East Asia and Africa. R12, R13, R14 occurred in strains from East Asia and in strains with no geographical information. R10 and R11 occurred less frequently and were present in strains from all geographical locations except for India.

3.1.3.2.3 Classification of Inversions

Inversions can be classified into two types: shared and specific inversions. The frequent inversions are regarded as shared, and the less frequent, specific. Strains from East Asia mostly showed the shared rearrangements whereas few strains had both types. Strains from South America and Europe mostly showed the shared rearrangements with few exceptions: PeCan18 strain (from South America) and B8, 26695, Rif1, Rif2, 26695 strains (from Europa) had both shared and specific rearrangements; v225d strain from South America had only specific rearrangements. Three strains from North America had the shared rearrangements only whereas four strains had both types. One strain (UM037) from East Asia and three strains (BM012A, BM012B, BM012S) from Australia had greater number of strain specific inversions compared to other strains. These strains were the most rearranged with six inversions: two or one shared and four or five specific inversions, respectively.

For some of the 72 *H. pylori* strains, information regarding the disease states of isolated patients was available: 8 from duodenal ulcer, 4 from gastritis, 4 from MALT lymphoma, 5 from gastric atrophy, 4 from peptic ulcer and 8 from gastric cancer (Table 3.6). East Asian group included strains isolated from the patients having almost all of the mentioned disease states, from duodenal ulcer to gastric cancer. Of the eight strains isolated from cancer patients, two strains (PeCan18, ELS37) had two (R10, R11) shared and one (PeCan18: R34, ELS37: R28) specific

inversion, whereas four other strains (2017, 2018, 908, J99) isolated from duodenal ulcer patients possessed the similar shared inversions (R10, R11) but no specific inversion. From this, we can infer that the specific inversion in strains PeCan18 and ELS37 might be associated with cancer. The remaining six cancer strains (F32, XZ274, F57, PeCan4, 7C, 29CaP) shared inversions except for F32 which had only one specific inversion (R30). The list of shared inversions in each strain were: F32: [R6, R9, R12, R13], XZ274: [R6, R12, R13, R14], F57: [R12, R13, R14], PeCan4: [R9], 7C: [R3, R6] and 29CaP: [R3, R6]. Although several strains shared same inversions, these inversions may be historically independent. More detailed sequence-level analysis is necessary to confirm the identity of inversions.

Table 3.6: Number of strains from different disease state individuals in various regions.

Region	No. of strains	DU	GU	ML	GA	GC	GS	UN
East Asia	28	4	3	3	4	3	1	10
South America	11	-	-	-	-	2	2	7
North America	8	1	-	-	-	3	-	4
Europe	13	4	1	1	-	-	1	6
Africa	3	-	-	-	-	-	-	3
India	2	-	-	-	-	-	-	2
Australia	5	-	-	-	-	-	-	5
Unknown region	2	-	-	-	-	-	-	2

* Disease states in columns are abbreviated as, DU: Duodenal Ulcer, GU: Gastric Ulcer, ML: MALT Lymphoma, GA: Gastric Atrophy, GC: Gastric Cancer, GS: Gastritis, UN: Unknown

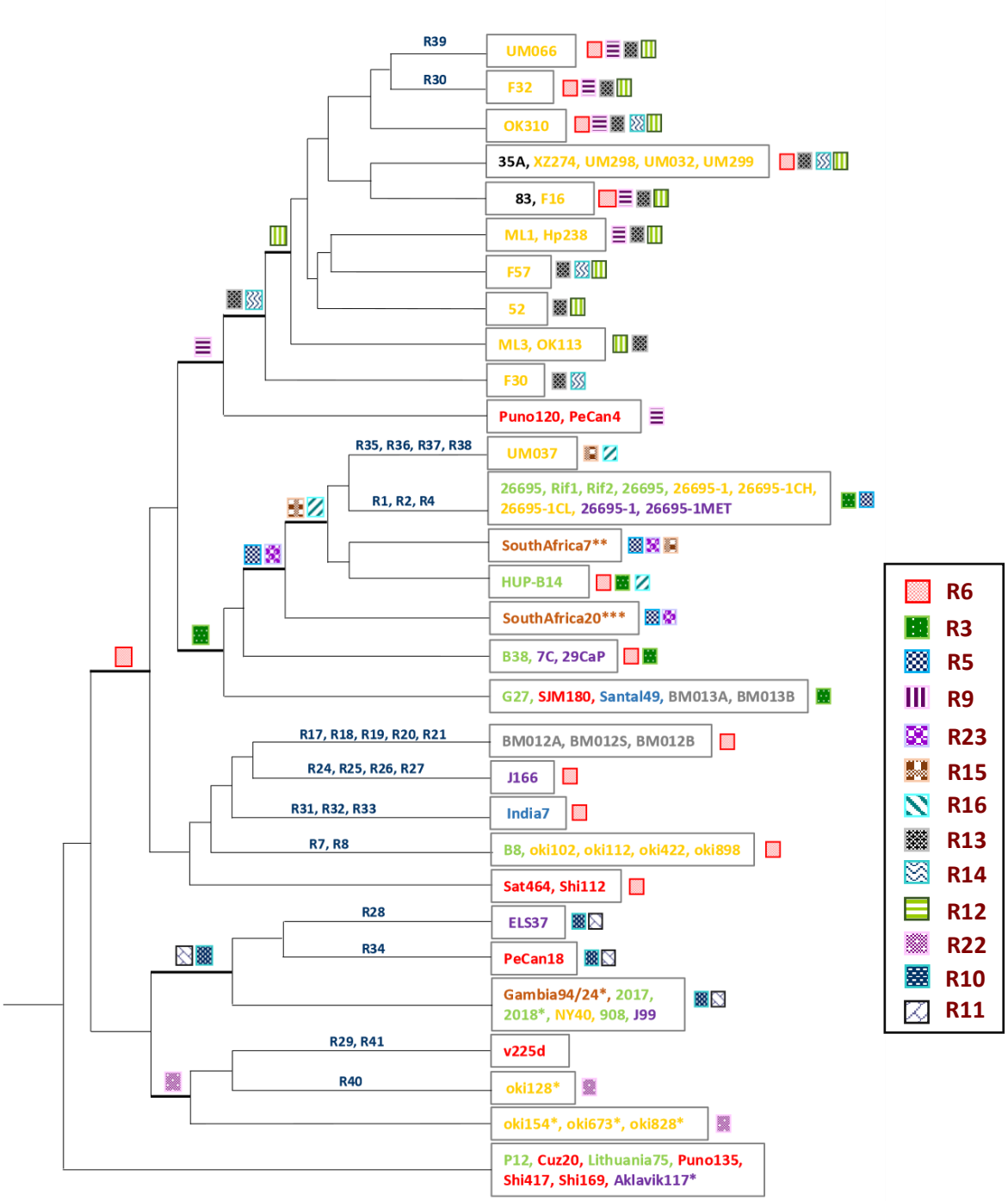


Figure 3.4: Inversion-based phylogeny. Labels beside the branches represent the inversions occurred in the strains (Table 3.3). Strains names are colored representing the geographical location (same as Figure 3.2). Strains name in black color show the strains with no geographical information. Legend on the right side indicate the reversals shared among multiple strains. * ignoring single gene transposition, ** ignoring single gene transposition and 2 gene inverse transposition, *** ignoring single gene transposition, 2 gene inverse transposition and 3 gene deletion [171].

3.1.4 Discussion

Degree of genome rearrangements increases with time as point mutations accumulate, both reflecting the evolutionary history of genomes. The number of inversions in *H. pylori* genomes was far less than the number of strain-specific genes, not to say of point mutations. Inversions therefore tells evolutional history in a longer timescale.

Among the 41 identified rearrangements, many were specific and few were geographic region-related. Although the investigated number of *H. pylori* genomes was too small to grasp the human migration, many rearrangements were not shared within regions partly because insertion sequences or virulence genes induce similar inversions. This also suggests that some inversions are associated with disease states irrespective of geography (or human migration), and certain inversions were linked with gastric cancer in our analysis. The pattern of inversions was most diverse in Japan (Figure 3.3a) probably because of the larger number of sampling. The North American region also had the diverse inversion pattern (Fig. 3.3a) even though the number of samples was much smaller compared to Japan. This diversity occurred maybe because of human migration. Since my analysis is based on orthologs and not the entire genomic region, verification needs more in-depth analysis using the whole genome sequences.

The obvious benefit of my algorithm is scalability: whole genome comparison is difficult for many genomes using previous approaches comparing two genomes. My algorithm can handle hundreds of strains at the level of gene orders. In terms of methodology, my simple approach previously did not resolve some complex rearrangements automatically, and they were later resolved manually. I have automated the process of resolving the complex rearrangements. The rearrangements can be visualized as the heatmap along with the clustering based on the rearrangements present in the genomes.

3.1.5 Conclusion

Gene orders can be used as a measure to study the evolutionary relationship of species. Previous studies considered only fully conserved genes in the pairwise comparison. My approach considers conserved gene clusters in a large number of genomes and identifies their rearrangements. Many inversions in *H. pylori* strains were shared across geographic regions, and only few were found to be geographic region-specific. Some inversions were associated with disease states such as cancer, so analyzing *H. pylori* genomes on a larger scale more in details can help us to understand the disease mechanism. Since *H. pylori* has evolved with the

global human migration, studying inversions may reveal the migration pattern although few rearrangements were geography related.

3.2 Section II

3.2.1 Overview

Several factors can contribute to the genomic diversity of an organisms which include various factors such as mobile elements [141], insertion sequences [28], prophages [142] and restriction modification system genes [143]. Molecular mechanisms[8] causing rearrangements have been explained with several genetic factors such as repeat and insertion sequences [26, 33]. Repeat sequences are the cause of genetic recombination, and the average repeat size is 53 and 100 base pairs (bp) for *Methanococcus jannaschii* and *H. pylori*, respectively [37].

Insertion sequences (ISs), also called IS elements, are short transposable DNA fragments. ISs have been found extensively in bacterial genomes [144], often around large inversions [145,146]. In *H. pylori*, total five ISs, from IS605 to IS609, have been documented in detail [28, 147–149]. The IS605 was the first to be reported, as the element splitting the *H. pylori's* virulence region (cag pathogenicity island) in the rearranged strain NCTC11638 [150]. It was found in one third of *H. pylori* strains and contains two open reading frames (ORFs), orfA and orfB [147].

IS606 is similar to IS605, and the amino acid identity of two ORFs with those of IS605 is approximately 25% [147]. Similarly, IS607 and IS608 carry two ORFs, but they contain the overlap for 27 bp and 30 bp, respectively [148,28]. Finally, IS609 carries four ORFs (orf1, orf2, orfA, orfB). The gene products of the orfA in the five ISs are grouped into two subfamilies, whether encoding serine recombinases (IS607, IS609) or not (IS605, IS606, IS608) [149]. For the orfB gene, IS606, IS606, IS607, and IS608 form a large group of unknown function and only IS609 is separate.

In the section I of this chapter, I have reported 41 non-trivial genome inversions in 72 publicly available *H. pylori* strains. Among these inversions, 18 were strain-specific and 23 were shared (Table 3.4). The shared inversions were numbered from R1 to R23 throughout this work. Among these inversions, R1–R16 were shared in different geographical locations, and R17–R23 were region-specific. For example, the reference strain 26695 and eight related strains (26695-1CL, 26695- 1CH, 26695-1, 26695-1MET, 26695, Rif1, Rif2, and 26695-1) contained five inversions (R1– R5), two of which (R1 and R2) were nested. Seven strains (P12, Shi417,

8. For details, see Chapter 1 of this dissertation

Shi169, Puno135, Cuz20, Lithuania75, and Aklavik117) were devoid of shared inversions. This section (Section II) of Chapter 3, provide a detailed analysis on the relationship between molecular markers with the identified rearrangements and discuss their chronological ordering and the possible relation to the *H. pylori* pathogenicity.

3.2.2 Materials and Methods

3.2.2.1 Sequence Materials and Identification of Rearrangements

The similar set of genomes mentioned in the section I of this chapter was used for this analysis. The genome rearrangements identified using the algorithm explained in the chapter 2 were investigated to look for the genetic markers that can be the possible cause of these rearrangements.

GenBank accession numbers for insertion sequences (IS605, IS606, IS607, IS608, and IS609) are U60177, U95957, AF189015, AF357224, and AY639112, respectively. Identification of these sequences was performed using Blastn (Match/Mismatch scores of 1, −2 with linear gap cost; Word size 28).

3.2.2.2 Identification of Sequence Repeats

Direct and inverted repeats were identified using the Unipro UGENE software version 1.29.0 [151]. Parameters for the Find repeats utility were as follows: window size: 25 bp, minimum identity per window 100%, minimum distance between repeats 0 bp, and maximum distance between repeats 1,000,000 bp. The relative location of repeat sequences and the rearrangements were investigated manually.

3.2.2.3 Genomic Islands

IsalndViewer4 webserver was used to obtain the information regarding the presence of genomic islands (GIs) in *H. pylori* strains [152]. This webserver had the precomputed results for several genomes. GI information of all the *H. pylori* strains in this study were obtained from the precomputed results. The relative location of genomic islands and the rearrangements were investigated manually.

3.2.3 Results and Discussion

3.2.3.1 Genome Rearrangements

Some inversions occurred more frequently compared to others. The inversions R3, R5, R6, R12, and R13 were present in more than 10 strains from different geographical locations. The genomic regions around these inversions can be called rearrangement hotspots. For example, the reference strain 26695 possessed two nested inversions (R1 and R2) in comparison with Aklavik117, a strain from North America. The inner inversion R2 was associated with GIs with inverted IS605 repeat as its possible cause (Figure 3.5). Two African strains (SouthAfrica20 and SouthAfrica7) without the R2 inversion also lacked IS605 in their GIs. The conserved existence of the GIs indicated their early formation, followed by the uptake of the ISs and the R2 inversion event.

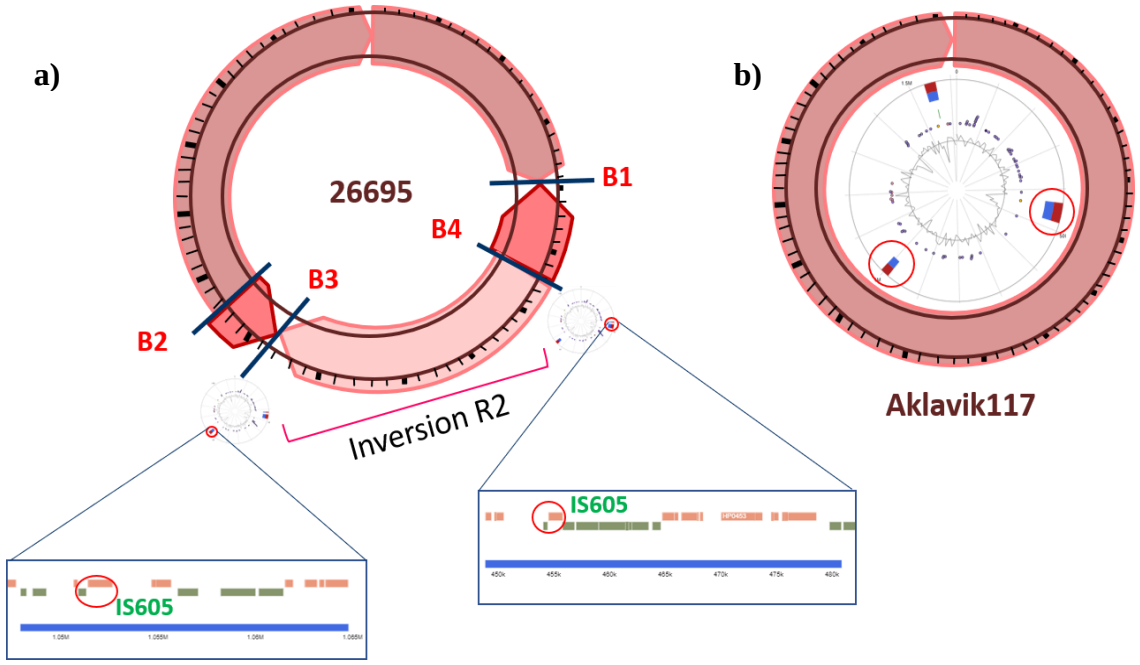


Figure 3.5: Genome rearrangements. **a)** Graphical representation of the *Helicobacter pylori* 26695 strain. Four breakpoints are indicated by crossing lines and the corresponding labels represent the breakpoint number (B1–B4). Two genomic islands (GIs) were identified in this strain that are present at the location of two breakpoints B3 and B4. Within these GIs, IS605 was present as an inverted repeat. **b)** Graphical representation of *H. pylori* Aklavik117 strain. This strain possessed the two GIs almost at the same location as in (**a**), but it lacked the insertion sequence (IS) elements in these GIs and the inversion R2 was absent [172].

However, not all the insertion sequences were associated with GIs (Figure 3.6). The type and the number of insertion sequences varied among strains, and 24 strains were devoid of intact IS elements. In strains with many IS elements, around half of them were associated with GIs, but the number of GIs also did not correlate with the number of IS elements. In order to discuss the relationship in more detail, I introduce the notion of breakpoints.

a)

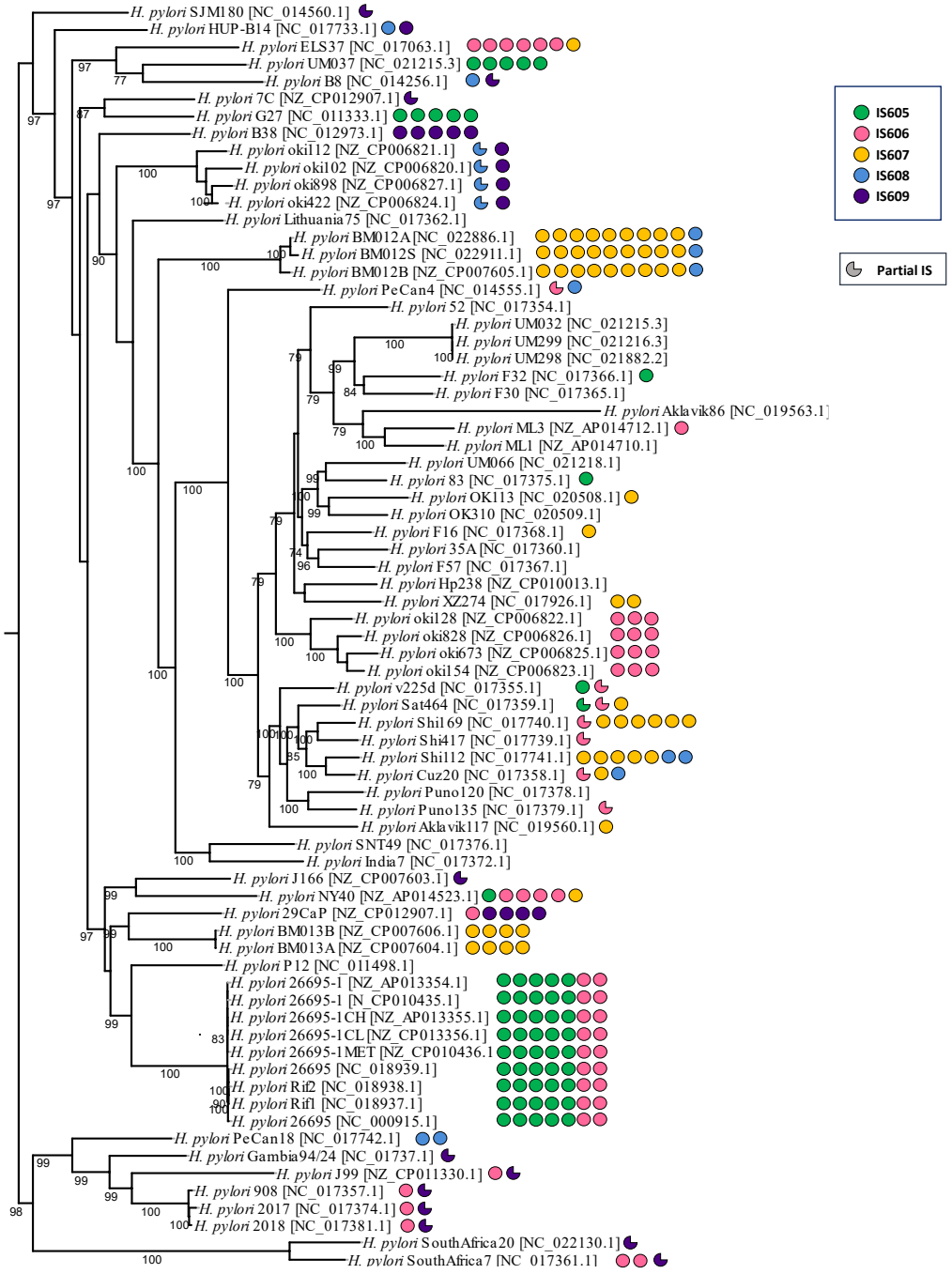


Figure 3.6: Distribution of IS elements and GIs. **a)** Core genes phylogenetic tree along with the distribution of the five insertion sequences (IS) shown in colored circle. Different colors represent the various IS elements as shown in the top right legend [172].

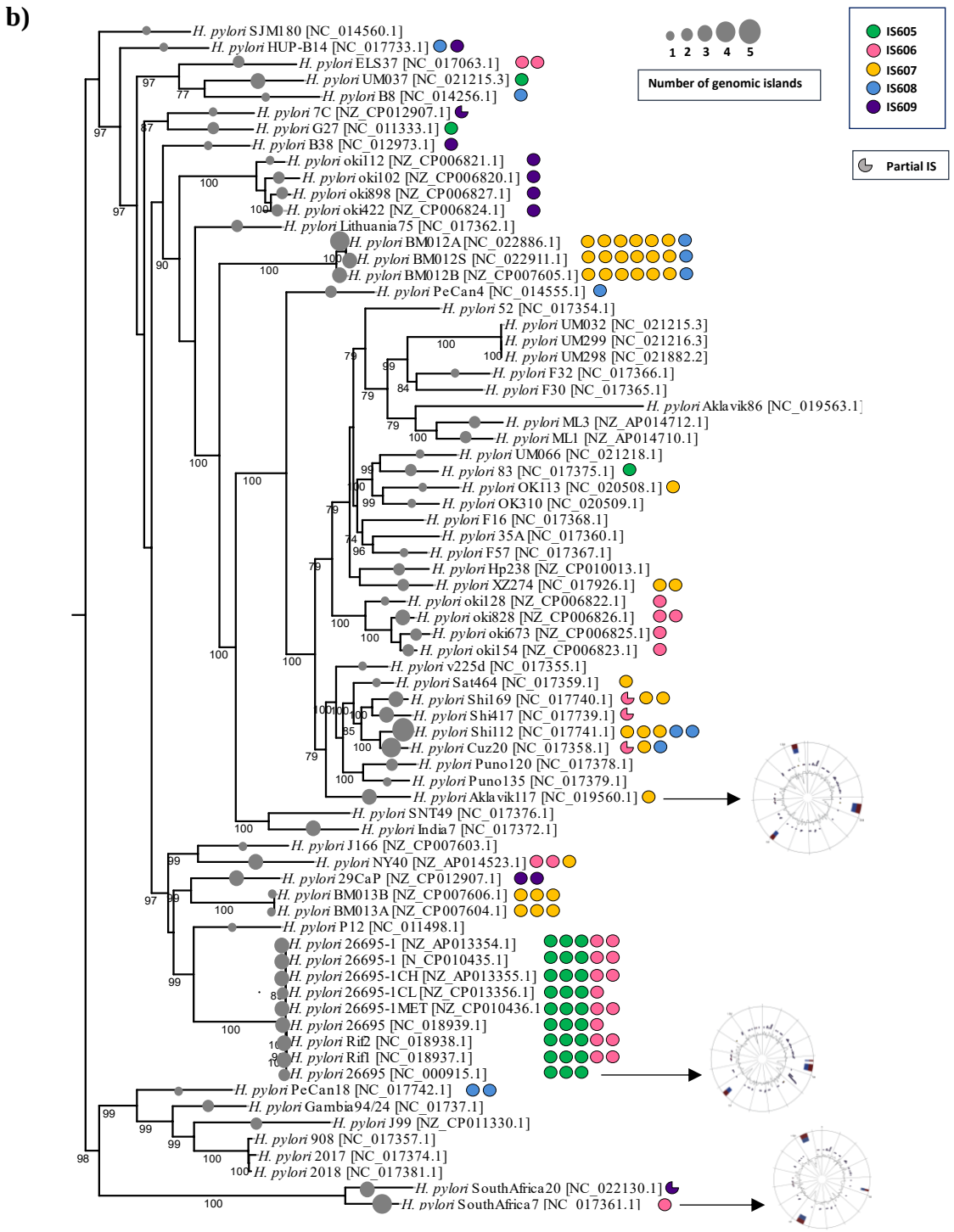


Figure 3.6: b) Core genes phylogenetic tree along with the distribution of the IS elements present in the genomic islands (GIs) identified in each strain. Gray colored circles along the branches indicate the number of GIs present in a particular strain. The plot showing the location of GIs is shown for the three strains. Different colors represent the various IS elements as shown in the top right legend [172].

3.2.3.2 Inversion Breakpoints

Two terminals of an inversion are referred to as breakpoints. Seventy-one breakpoints, designated as B1–B71, were identified in the analyzed strains, corresponding to the 41 inversions. The number of breakpoints did not match the doubled number of inversions because of their reuse: 13 breakpoints were involved in more than one inversion. Among the 71 breakpoints, B1–B30 were shared among the strains from different geographical locations whereas B31–B44 were shared among the strains from the same geographical location (region-specific) and B45–B71 were strain-specific. Among the shared breakpoints (B1–B30), B22–B27 were observed in large number of East Asian strains along with a few strains with unknown geographical location; we called them East-Asia-specific breakpoints. Similarly, some breakpoints were observed only in strains from particular geographical locations. Figure 3.7 illustrates the distribution of shared breakpoints among strains from different geographical locations. The largest number of breakpoints was 10 in strains from East Asia and Australia. Detailed information about the inversion and their corresponding breakpoints is shown in Table A.3 of Appendix.

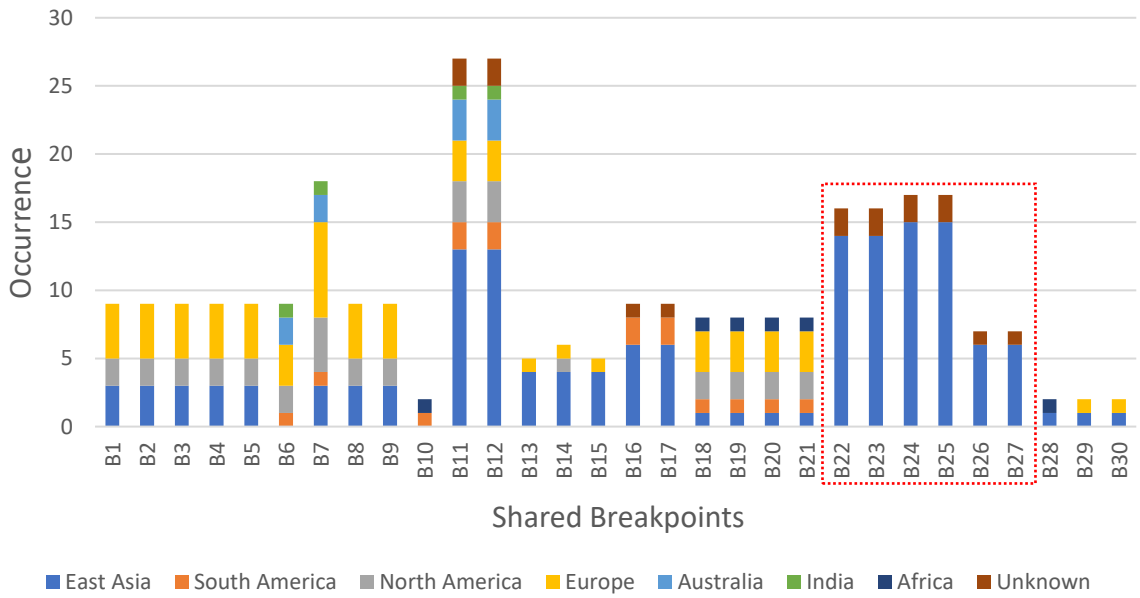


Figure 3.7: Distribution of shared breakpoints among strains from different geographical locations. Breakpoints are designated as B1-B30. B22-B27 can be regarded as East Asia specific breakpoints [172].

3.2.3.3 Repeat Sequences and Their Associated Inversions

In most prokaryotes, a repeat sequence of length > 25 is assumed to involve in homologous recombination with statistical significance [11,26,27]. We investigated all direct and inverted repeats of length >25 nucleotides with 100% sequence identity in all strains (Figure A.3 of Appendix). Among the 41 inversions, 20 inversions were associated with repeats. For example, the inversion R6 was observed in 27 strains, among which 20 were associated with inverted repeats around its two breakpoints. Exceptions were four strains from Okinawa (Japan) that possessed no element at one breakpoint (B11) and a direct repeat at the other (B12) and three strains from Australia that possessed a direct repeat at one breakpoint (B11) and an inverted repeat at the other (B12).

Table 3.7 shows the number of associated inverted and direct repeats with inversions. The ratio of inverted versus direct repeats (IR/DR) was less than 1 (Figure 3.8) and the total number (and their total length) of direct and inverted repeats was proportional to the genome size (Figure 3.9) [153].

Table 3.7: Number of inverted and direct repeats associated with different inversions [172].

Inversion Type	Total inversions	Number of IR associated Inversions	Number of DR associated Inversions
World-wide	16	5	2
Region-specific	7	4	1
Strain-specific	18	5	3

The correlation between the number of repeats and that of inversions was weak. This suggested that the occurrence of repeats was not the direct cause of inversions. Their relative position, especially the relation with GIs, seemed important for homologous recombination.

A larger number of direct and inverted repeats were found in South American and African strains (Figure 3.10). The longest direct and inverted repeats of length 8,041 bp,10,305 bp were observed in strains SouthAfrica7 (Africa) and F16 (East Asia) (Table A.4 of Appendix). The average size of longest repeats in each region is shown in Table 3.8. The least number of direct and inverted repeats was observed in the strains 2018 and F57 from Europe and East Asia,

respectively. The largest number of direct and inverted repeats was found in UM037, an East Asian strain. This strain contained six inversions, among which three were associated with inverted repeats (R16, R37, and R38).

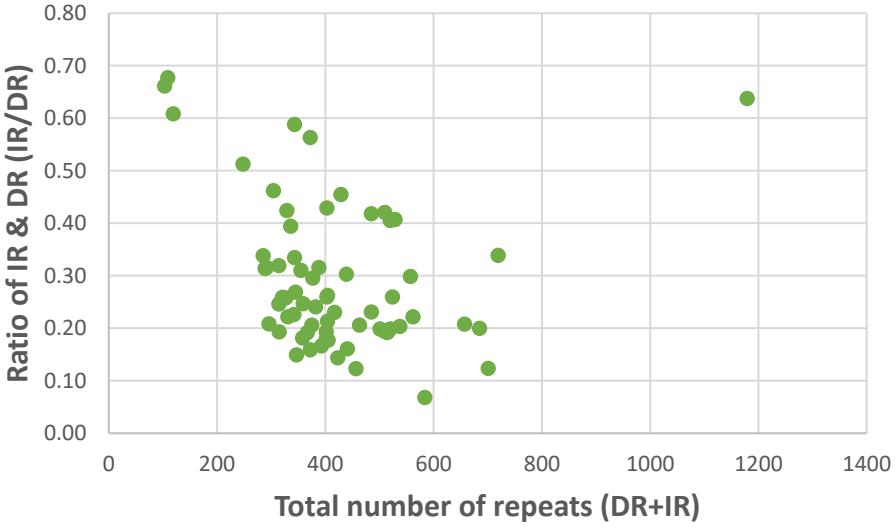


Figure 3.8: Distribution of the ratio of inverted repeats (IR) over direct repeats (DR). This ratio (IR/DR) less than 1 indicates the underrepresentation of inverted repeats [172].

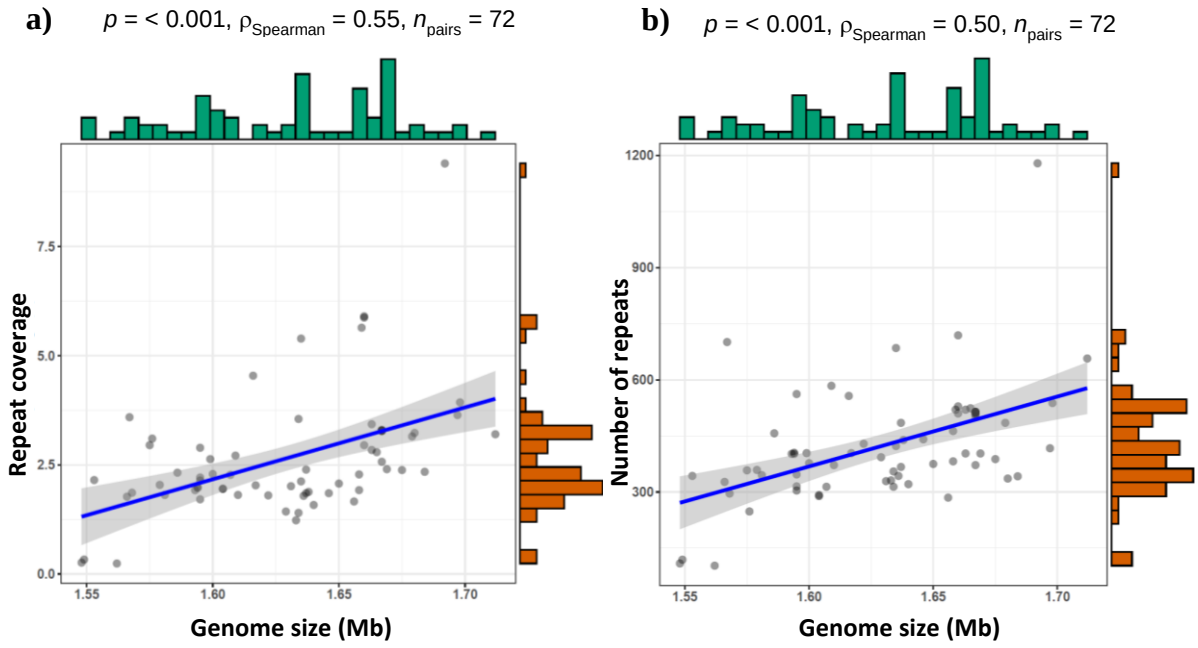


Figure 3.9: a) Association between genome size and repeat coverage. **b)** Association between genome size and number of repeats. Positive correlation was observed for both a and b [172].

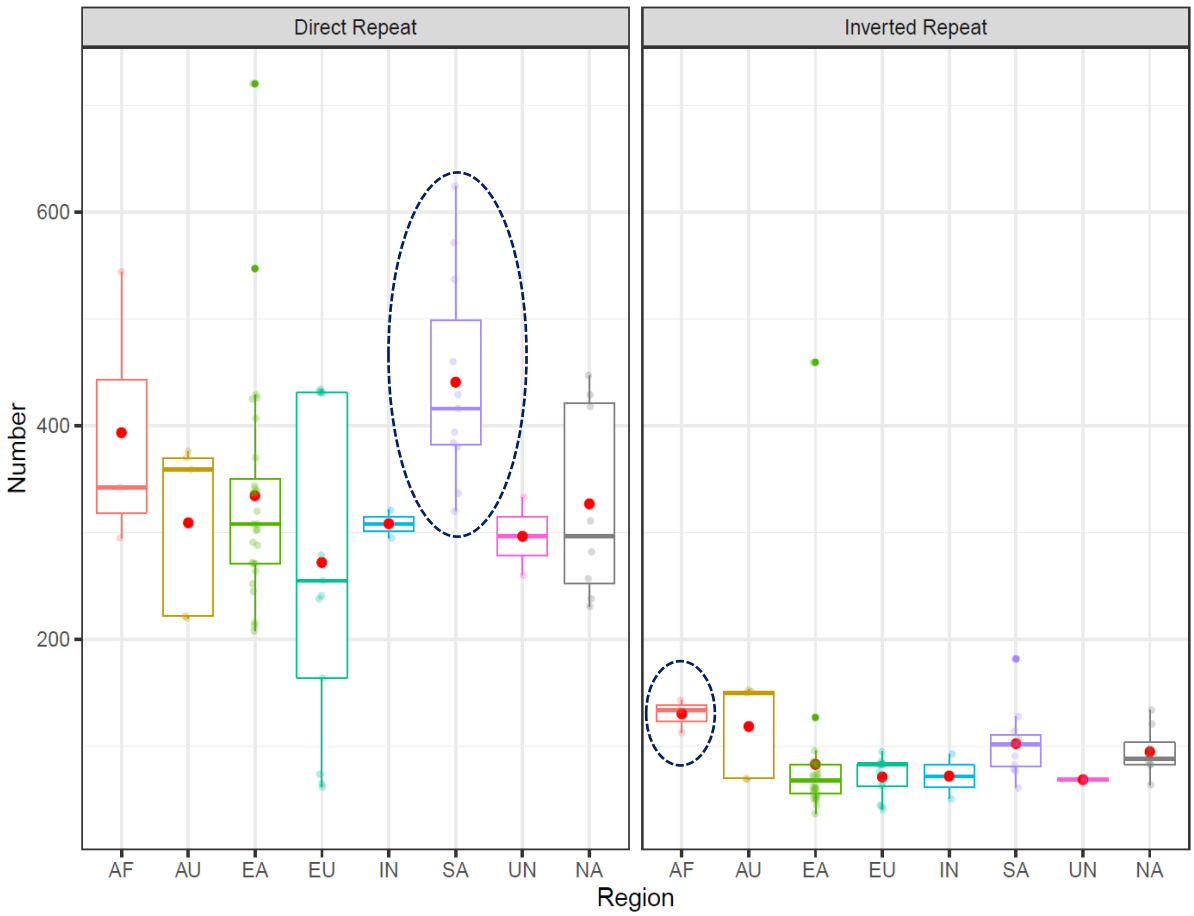


Figure 3.10: Distribution of direct and inverted repeats in different geographical regions. Region names are abbreviated as: (AF: Africa, AU: Australia, EA: East Asia, EU: Europe, IN: India, SA: South America, UN: Region not known, NA: North America). Red dot represents the average number of repeats identified in each region. Region with large number of repeats on average are encircled in both the panels [172].

Table 3.8: Average size of longest repeats observed in each geographical region [172].

Region	Average size of Longest Inverted Repeat	Average size of Longest Direct Repeat
East Asia	2181	3425
South America	1631	4495
North America	2145	3745
Europe	2268	2756
Africa	4587	5057
India	1772	4084
Australia	2315	3033
Unknown	1357	3154

Among the different types of inversions, five world-wide, five region-specific, and seven strains-specific inversions possessed the inverted repeat around their breakpoints. Larger inversions (in terms of the number of inverted genes) possessed larger repeats. A significant positive correlation was observed between the inversion size (number of inverted genes) and the average size of repeat found around those inversions (Figure 3.11).

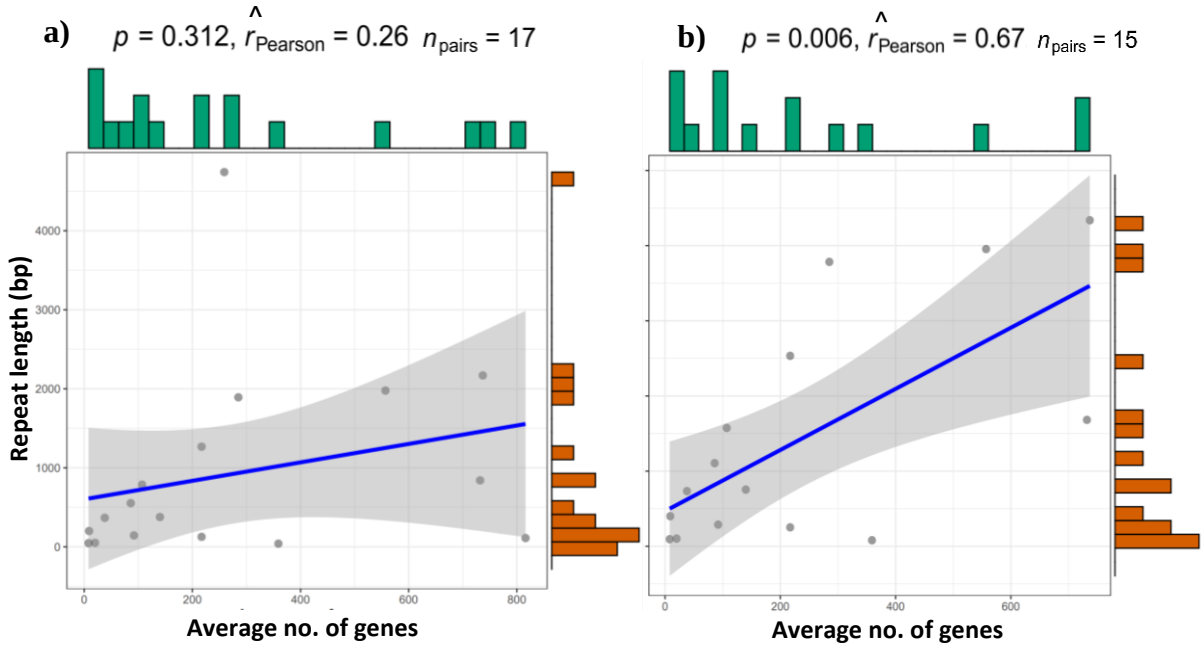


Figure 3.11: Correlation between inversion size and repeat size. **a)** Association between average number of genes in an inversion and length of repeat present around its breakpoints. **b)** Association between average number of genes in an inversion and length of repeat present around its breakpoints after removing two outliers [the inversion R7 (the inverse transposition of 22 genes when dealt as an inversion had 816 genes while the repeat was 111 bp in length) and inversion R26 (strain-specific inversion)]. A significant positive correlation was observed [172].

3.2.3.4 Presence of Genomic Islands around Inversion Breakpoints

GIs represent regions acquired by horizontal gene transfer [39]. A varying number of genomic islands was present in the analyzed strains. Six GIs were the largest and were found in two strains: Shi112 (South America) and J99 (North America). The average number of the identified GIs was two. Most region-specific and strain-specific breakpoints were observed in the neighborhood of GIs (Table 3.9). In three Australian strains, four GIs were located in the neighborhood of Australia-specific breakpoints.

Table 3.9: Strains having genomic island(s) associated with breakpoints [172].

Strain	Accession Number	No. of GIs	GIs associated with breakpoints
UM037	NC_021217.3	3	1
F32	NC_017366.1	1	1
26695-1CL	NZ_AP013356.1	2	2
26695-1CH	NZ_AP013355.1	3	2
26695-1	NZ_AP013354.1	3	2
Aklavik117	NC_019560.1	4	2
26695-1	NZ_CP010435.1	3	2
26695-1MET	NZ_CP010436.1	3	2
ELS37	NC_017063.1	2	1
Rif2	NC_018938.1	3	2
Rif1	NC_018937.1	3	2
26695	NC_018939.1	3	2
26695	NC_000915.1	2	2
Gambia94/24	NC_017371.1	5	2
SouthAfrica20	NC_022130.1	4	1
India7	NC_017372.1	4	1
BM012A	NC_022886.1	5	4
BM012S	NC_022911.1	4	4
BM012B	NZ_CP007605.1	4	4

The most frequent global inversions, R3 and R6, were distant from any GIs but neighbored by repeat sequences. Compared to such global breakpoints, region- and strain-specific breakpoints were often neighbored with GIs. These local breakpoints seemed to have formed after the global breakpoints were established.

3.2.3.5 Distribution of Insertion Sequences and Their Association with Inversions

Different types of insertion sequences (IS605-IS609) have been reported in *H. pylori* [28, 147–149]. I performed detailed analyses of these five elements around inversions (Table 3.10 and

Figure A.4 of Appendix). Association between insertion sequences and breakpoints is summarized in Table 3.11.

Table 3.10: Number of copies of each IS element (IS605-IS609) in all the strains. Fraction indicates an incomplete IS element. (See also Figure 3.6) [172].

Strains	Region	IS605	IS606	IS607	IS608	IS609
NY40	East Asia	1	4	1	0	0
ML3	East Asia	0	1	0	0	0
UM032, UM298, UM299, F30, F57, ML1, UM066, OK310, 52, Hp238	East Asia	0	0	0	0	0
UM037	East Asia	5	0	0	0	0
F32	East Asia	1	0	0	0	0
XZ274	East Asia	0	0	2	0	0
F16, OK113	East Asia	0	0	1	0	0
oki128, oki154, oki673, oki828	East Asia	0	3	0	0	0
oki102, oki112, oki422, oki898	East Asia	0	0	0	0.5	1
26695-1CL, 26695-1CH, 26695-1	East Asia	5	2	0	0	0
Shi112	South America	0	0	5	2	0
Sat464	South America	0.5	0.5	1	0	0
Cuz20	South America	0	0.5	1	1	0
PeCan4	South America	0	0.5	0	1	0
PeCan18	South America	0	0	0	2	0
Puno120	South America	0	0	0	0	0
Shi169	South America	0	0.5	6	0	0
SJM180	South America	0	0	0	0	1
Puno135, Shi417	South America	0	0.5	0	0	0
v225d	South America	1	0.5	0	0	0
7C, J166	North America	0	0	0	0	0.5
29CaP	North America	0	1	0	0	4
Aklavik117	North America	0	0.5	1	0	0
26695-1, 26695-1MET	North America	5	2	0	0	0
J99	North America	0	1	0	0	0.5
ELS37	North America	0	6	1	0	0
B38	Europe	0	0	0	0	5
HUP-B14	Europe	0	0	0	1	1
Rif1, Rif2, 26695	Europe	5	2	0	0	0
B8	Europe	0	0	0	1	0.5
G27	Europe	5	0	0	0	0
Lithuania75, P12	Europe	0	0	0	0	0
2017, 2018, 908	Europe	0	1	0	0	0.5

Continued Table 3.10

SouthAfrica7	Africa	0	2	0	0	0.5
Gambia94/24, SouthAfrica20	Africa	0	0	0	0	0.5
India7, Santal49	India	0	0	0	0	0
BM012A, BM012B, BM012S	Australia	0	0	9	0.5	0
BM013A, BM013B	Australia	0	0	4	0	0
83	Unknown	1	0	0	0	0
35A	Unknown	0	0	0	0	0

Table 3.11: Number of insertion sequence (IS) present around different types of inversion breakpoints (BPs) [172].

IS	World-wide BPs	Region-specific BPs	Strain-specific BPs
IS605	4	0	1
IS606	3	1	0
IS607	0	3	0
IS608	0	1	0
IS609	0	0	0

Both IS605 and IS606 were found in multiple geographical regions around the widely shared breakpoints of inversion R2 (Figure 3.5a) and R28 respectively, with inverted repeats. IS605 was found in 16 strains and 13 of them carried two standard ORFs (orfA and orfB). Anomalies were one strain from South America (Sat464) lacking orfA and two strains (v225d from South America and 83 from Unknown) with nonsense mutations in orfB (pseudo gene). Of note, 26695 related strains possessed five copies of IS605, and the same number of IS605 were retained in distant strains of G27 (European) and UM037 (East Asia).

IS606 was present in 30 strains worldwide. It was observed in African strains. One strain (ELS37 from North America) possessed six copies, but all others possessed up to three. Eight strains in the same clade (Cuz20, Shi417, PeCan4, Shi169, Puno135, Sat464, and v225d from South America and Aklavik117 from North America) possessed orfB only; this observation indicated that the deletion of orfA occurred before the diversification of strains in America. These strains, however, possessed different numbers of IS607 and GIs. In addition, some IS606 were found within GIs whereas others were not. Therefore, the possibility of recombination between strains also remained. In two strains, Sat464 and v225d, the orfB contained a nonsense mutation.

IS607 was region-specific in South America and Australia. It was present in 15 strains, including all Australian strains. All strains except one had both orfA and orfB having an overlap of 27 bp between them [148]. In two strains, orfB contained nonsense mutation. In one East Asian strain (F16) its orfA was pseudo gene and orfB was split into two genes. IS608 was also region specific, mainly in South America. It was present in 13 strains, including four Peruvian strains: two from gastric cancer (PeCan4, PeCan18) and two from unknown disease state (Shi112, Cuz20) [28]. In Asia, only strains from Okinawa possessed this sequence with orfB only. In Australia, three strains possessed this sequence, but its orfB was dysfunctional. Finally, IS609 was found in Europe and North America but not in Asia, Australia, and South America. SJM180 was classified as South America, but its phylogenetic clade showed its closeness to European strains. Complete IS609 (all four ORFs) were found in few strains only: one European and one American strains (B38 and 29CaP). Four Okinawa strains were exceptions because they possessed the complete copy of IS609 and their phylogenetic clade was closer to European strains. Partially deleted IS elements were more likely to be outside of GIs. This indicated that IS elements were still active and transferred in/out of GIs (Figure 3.6).

3.2.3.6 Other Molecular Elements Related to Inversions

In addition to the repeats, ISs, and GIs, other elements like DNA methyltransferases, restriction modification (RM) system, and virulence related genes were also searched in the neighborhood of the identified breakpoints (Figure 3.12). Type II RM genes were more abundant than Type I and Type III RM genes. The strains sharing the same inversion breakpoints tended to possess similar elements (Table A.5 of Appendix). Since the number of analyzed strains was small, finding the specificity of these elements with any of the disease states requires analysis on a larger scale.

3.2.4 Conclusions

Analysis of genome rearrangements in association with insertion sequences and repeats can reveal genome evolution in a finer scale. I have compared the strains from different geographical locations to identify the association of several genomic elements with the inversions. Most of the shared inversions possessed similar IS elements with a few exceptions. This suggests that these elements are well-conserved irrespective of the different geographical region. Restricted distributions of IS607 and IS608 indicated their relatively recent proliferation compared to IS605 and IS606, and isolation of partial IS elements from GIs indicated the important roles of GIs in distributing IS elements. My analysis was limited to the

publicly available strains. A larger scale analysis can help us to understand the geographical distribution and association of disease with different genomic elements. Since *H. pylori* can cause different diseases, analysis of various rearrangements can lead us to identify the underlying possible causes, thus facilitating a better understanding of disease mechanisms.

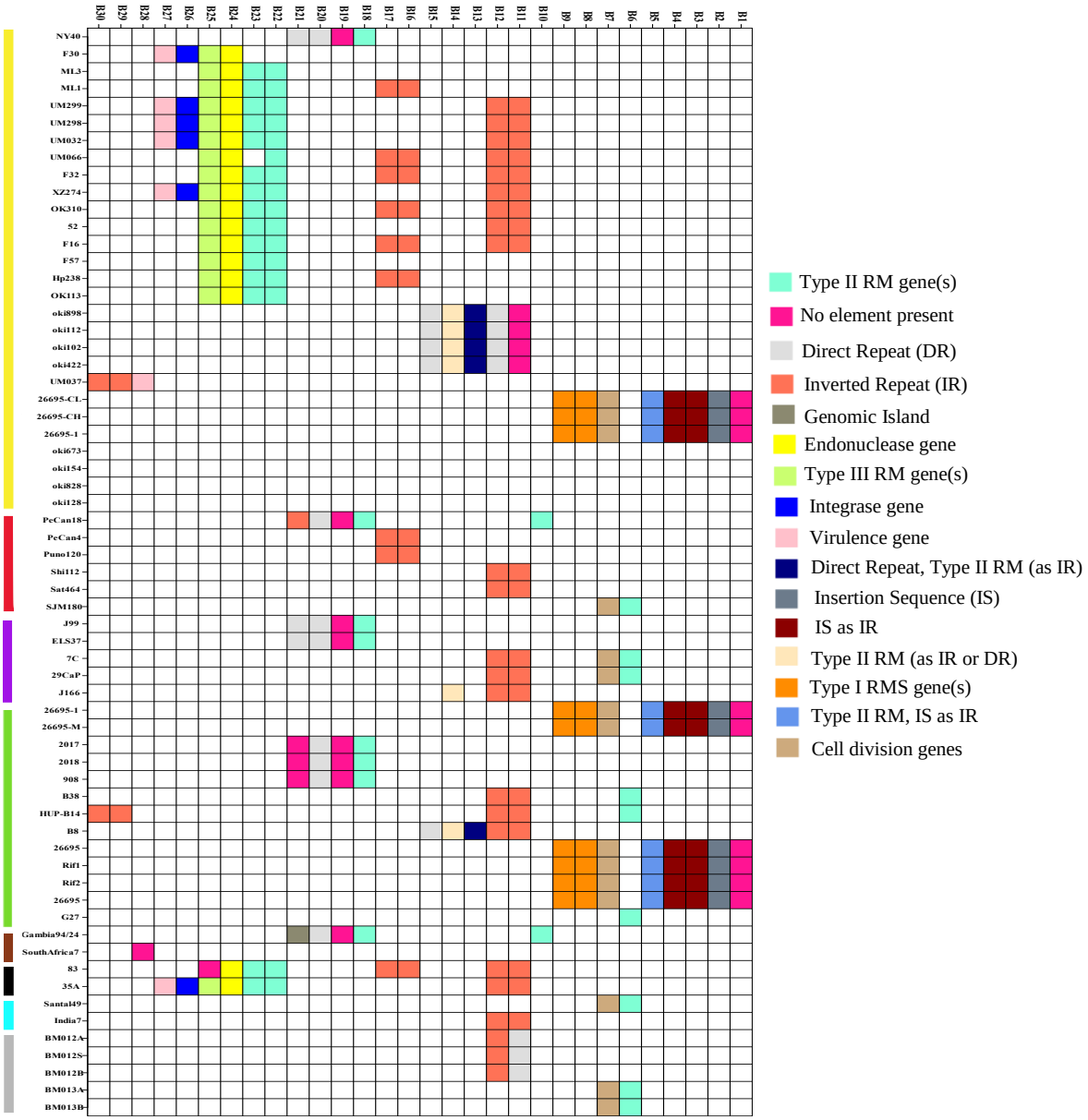


Figure 3.12: Presence of different elements around shared breakpoints at the strain level. Each column indicates one shared breakpoint whereas each row represents one strain. Different color in the cells represents different element. The white cell indicates the absence of breakpoint [172].

The next chapter includes the comparative genome scale analysis of a larger dataset of *Helicobacter pylori* strains obtained from the NCBI database. It discusses the differences in various genomic features and their possible association with a particular disease state.

Chapter 4

Comprehensive Analysis of Genomic Diversity: Identifying the Association of Rearrangements with the Disease State

In the previous chapter, I have used *Helicobacter pylori* (*H. pylori*) genomes to demonstrate the use of my algorithm for identifying the genome rearrangements. In that study inversions associated with a disease state were identified. As the dataset was small in the previous study, the analysis at the larger scale is required to investigate the association of inversion with the disease state. This chapter includes the in-depth analysis of more than 120 *Helicobacter pylori* genomes to understand the genomic diversity of this bacterium.

4.1 Overview

The infection of *H. pylori* is considered as an important risk factor and has been associated with various clinical outcomes such as peptic ulcer, gastritis, mucosa-associated lymphoid tissue (MALT) lymphoma and gastric cancer [154, 116]. Since 1994, the bacterium has been categorized as a type I carcinogen by the World Health Organization [155]. The prevalence of infection is higher in developing countries compared to the developed ones [156]. The different outcomes as a result of the infection with this bacterium depends on several factors such as environment, host, diet and certain bacterial factors [153, 157].

It has been suggested by several studies that the marked genetic variability of the *Helicobacter pylori* plays a role in the different clinical outcomes among the infected individuals [126, 158]. The mechanism of pathogenesis is complicated in *H. pylori*, the most important among them includes the expression of certain virulence related genes [159]. Several virulence factors have been shown to be the indicators of the critical outcomes of the infection [160, 161]. Other factors that help the *H. pylori* in its adaptation and survival include several flagellar genes and the outer membrane proteins [162, 163].

The possible reason for the various disease outcomes as a result of *H. pylori* infection include the virulence factors of the *H. pylori* strains along with the other environmental factors [164]. One of the most extensively studied virulence factor of *H. pylori* is a cytotoxin-associated gene A (*cagA*), which encodes a highly immunogenic protein (CagA) [165]. The different strains might carry the complete and intact cagPAI, the incomplete cagPAI with the missing genes or the cagPAI that has been affected by the genome rearrangements [166]. In addition to the *cagA* gene another well studied virulence factor is vacuolating toxin A (VacA) [164]. The *vacA* gene is present in all the *H. pylori* strains, but it may not be functional in all of them because of the allelic diversity in the three regions [167, 168]. In addition to this, the two strains of *H. pylori* from different individuals may differ in their genomic content along with its organization [169].

The *H. pylori* genomes are also quite diverse in terms of the genome organization. This diversity might be the result of the genome rearrangements that occurred during the course of evolution. These genome rearrangements might result in gene gain or loss and alter the expression of some genes. Some of the genome rearrangements might be associated with a particular disease outcome or the geographical location as reported in the previous chapter of this dissertation. In order to investigate the association of the genome rearrangements and other genomic features with the clinical outcomes more in detail, I have performed the comparative analysis of the publicly available *H. pylori* genomes from NCBI. This dataset includes the *H. pylori* genomes obtained from the individuals of various disease states such as: atrophic gastritis, gastritis, chronic active gastritis, gastric atrophy, peptic ulcer, duodenal ulcer, gastric ulcer, gastric cancer and MALT lymphoma. This study will provide insight into the dynamic nature of *H. pylori* genomes and the role of the genetic diversity in the pathogenesis.

4.2 Materials and Methods

4.2.1 Genome Sequences

Genome sequences of 123 *H. pylori* strains were obtained from NCBI/ENA/DDBJ repository. The strains were obtained from individuals with 9 different disease outcomes. The genomes were grouped according to each disease outcome and genomes with no information of the disease state were grouped as unknown. The other groups were named as: Atrophic gastritis, Gastric ulcer, Gastric cancer, Chronic active gastritis, Duodenal ulcer, Gastric atrophy, Gastritis, MALT lymphoma, Peptic ulcer. Detailed information regarding the strains is available (Table A.6 of Appendix).

4.2.2 Average Nucleotide Identity

Average nucleotide identity (ANI) for 123 *Helicobacter pylori* genomes was identified using the python script. The result was visualized using heatmap function in R.

4.2.3 Orthologous Gene Clustering

Orthologous gene clustering was performed using the GET_HOMOLOGUES software package [170] (cutoff: E-value 1.0 × 10−5, Minimum coverage percentage: 75%) and the OrthoMCL algorithm [106] was used to identify the gene clusters. Gene clusters were assigned the Clusters of Orthologous Group (COG) functional annotations using the Reverse Position-Specific BLAST search against the NCBI Conserved Domain Database (NCBI-CDD) and the Perl script "cdd2cog" (https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog)

4.2.4 Phylogenetic Analysis

The phylogenetic analysis was performed using the seven housekeeping genes: *atpA, efp, mutY, ppa, trpC, ureI, and yphC* of 123 *H. pylori* genomes. The sequences of these seven genes of 26695 *H. pylori* strain were obtained from PubMLST database (https://pubmlst.org/). The BLAST search was performed to obtain the sequence of the seven housekeeping genes for the other strains using the 26695 gene sequences as a reference. The sequences were aligned using MAFFT (version 7.313) [137], alignments were trimmed using trimAl [138] with default parameters, which were later concatenated and phylogenetic tree was obtained using standard-RAxML-master with the parameters: -T 11, −N 1000, −m PROTCATBLOSUM62 [139]. The phylogenetic analysis was also performed using the *vacA* gene following the same pipeline used for the housekeeping genes.

4.2.5 Identification of Restriction Modification Genes, CagPAI and other Virulence Genes

The information of the restriction modification system genes for all the *H. pylori* genomes was obtained from the REBASE database (http://rebase.neb.com). The cag pathogenicity island genes were identified using the BLAST search against the 26695 strain's cag genes. Later, the cag pathogenicity island genes and other virulence genes information for all the genomes was also obtained from the Virulence Factor database (http://www.mgc.ac.cn/VFs/). The distribution of these genes among the strains was visualized using color2D.matplot function in the plotrix package in R.

4.2.6 Identification of Repeat and Insertion Sequences

The Find repeats utility in the Unipro UGENE software version 1.29.0 [151] was used to identify the direct and inverted repeats. The parameters used were as follows: window size: 25 bp, minimum identity per window 100%, minimum distance between repeats 0 bp, and maximum distance between repeats 1,000,000 bp. The sequences of the insertion elements (IS605, IS606, IS607, IS608, and IS609) were obtained from NCBI database under the GenBank accession numbers (U60177, U95957, AF189015, AF357224, and AY639112). Theses sequences were used for the identification of the IS elements in the *H. pylori* genomes using the Blastn.

4.2.7 Rearrangement Analysis

The rearrangements were identified using the algorithm that I have developed and is described in detail in Chapter 2 of this dissertation. First, the orthologous gene clusters were obtained by protein blast search using the bidirectional best-hits criterion. The gene cluster that were present in $\geq$ 90% of the genomes were selected for the downstream analysis. These gene cluster were used to identify the gene orders in the genomes. Some of the genomes were rotated and flipped in order to have gene 1 at the start and gene *n* at the end. The gene order information was given as an input to the rearrangement identification algorithm which identifies the consensus gene ordering. Later, it reorders the gene orders in all the genomes using the consensus gene ordering. The breakpoints are identified and the rare reversals are fixed. For details see chapter 2 of this dissertation. The clustering of genomes on the basis of the identified inversions was performed using the heatmap2 function and the phylogram was created using the *ape* and *phangorn* packages in R.

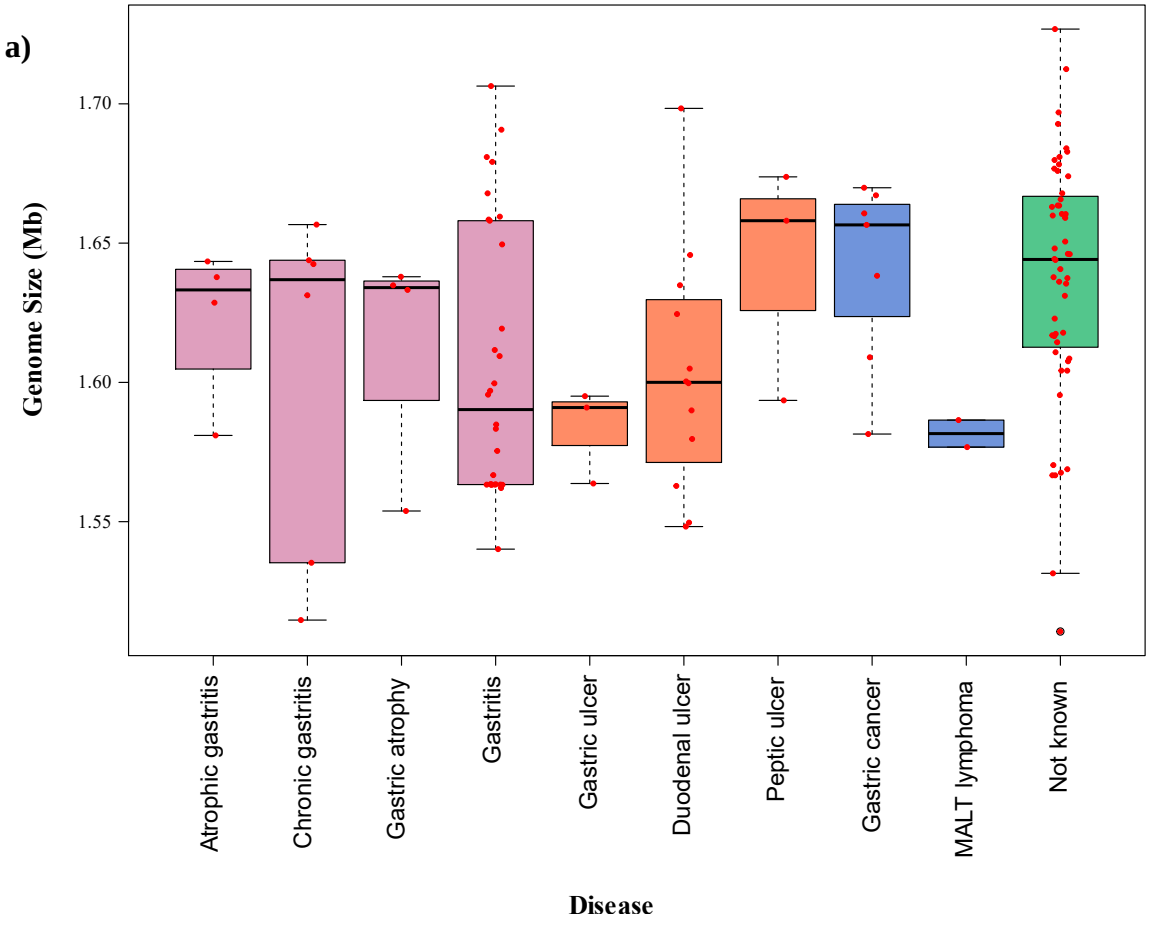
4.3 Results and Discussion

4.3.1 General Genomic Features

The 123 *H. pylori* genomes obtained from NCBI database were classified in 10 groups, 9 according to the disease outcome and one representing those genomes with no information of the disease outcome. The number of genomes in each group is shown in Table 4.1. The genomic size of the 123 *H. pylori* genomes ranged from 1.51~1.73 Mb and the GC content varied from 38.43% ~ 39.30%. Figure 4.1 shows the distribution of genomic size and the GC content among the different groups in which the genomes were classified.

Table 4.1: Classification of the 123 *H. pylori* genomes into ten groups

Disease state	Number of strains
Atrophic gastritis (AT)	4
Chronic active gastritis (CG)	6
Gastric atrophy (GA)	4
Gastritis (GS)	30
Gastric ulcer (GU)	3
Duodenal ulcer (DU)	12
Peptic ulcer (PU)	3
Gastric cancer (GC)	7
MALT lymphoma (ML)	2
Unknown (UN)	52



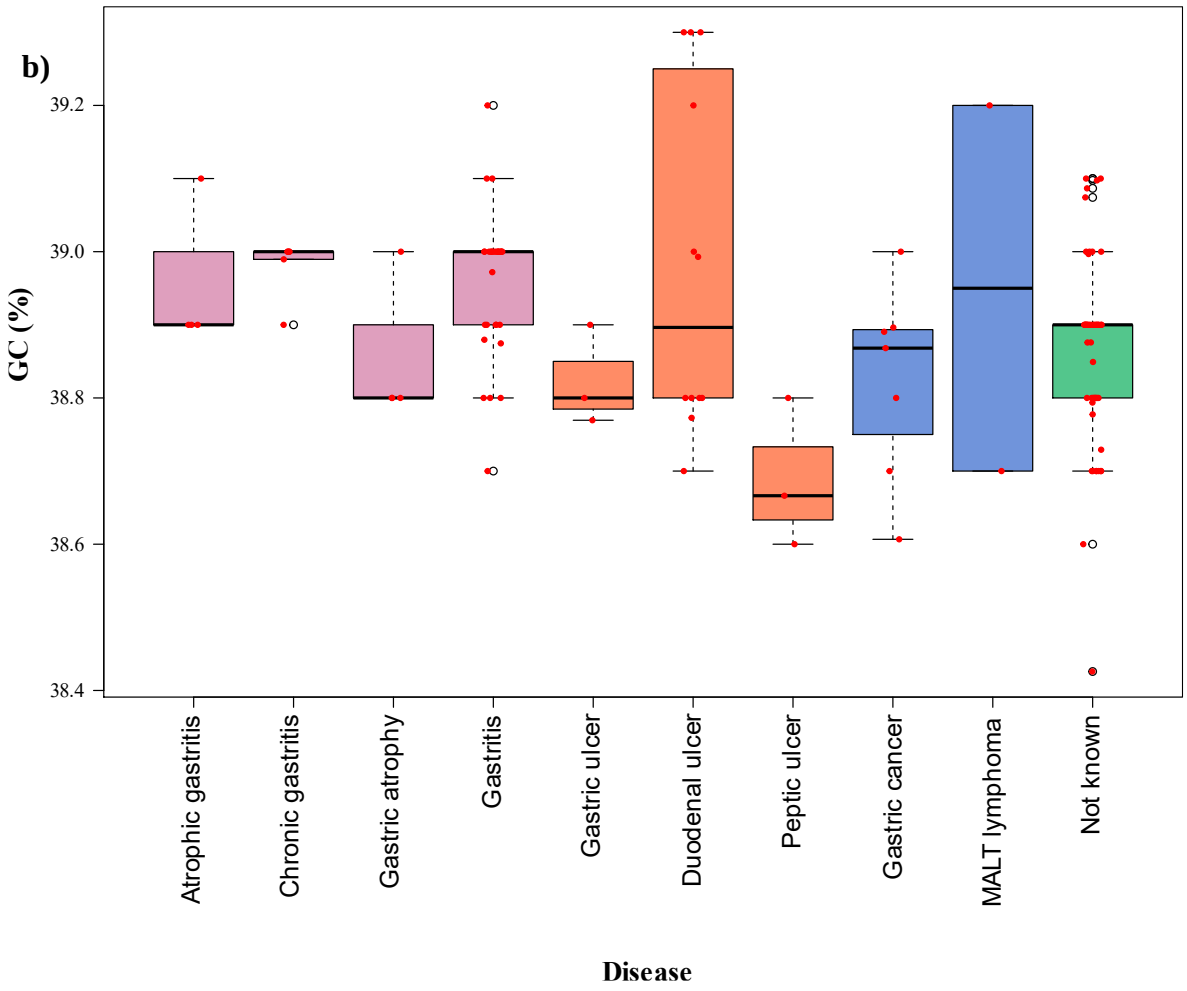


Figure 4.1: a) Distribution of the genomic size of the 123 *H. pylori* genomes. **b)** GC content variation of the genomes classified into ten groups.

The genomic size of the gastritis and duodenal ulcer groups was distributed widely compared to the other disease groups, whereas peptic ulcer and gastric cancer shows the similar distribution. The distribution of the GC content in the duodenal ulcer group was wide, whereas all the genomes in the chronic gastritis group have approximately the same GC content of 39 %. The two genomes in the MALT lymphoma group have the extremely different GC content, one having the 38.7% and the other having 39.2%.

The genomes were grouped into three large clusters that were formed on the basis of the average nucleotide identify (ANI) calculated for the 123 *H. pylori* genomes. Figure 4.2 shows the ANI based clustering of the genomes.

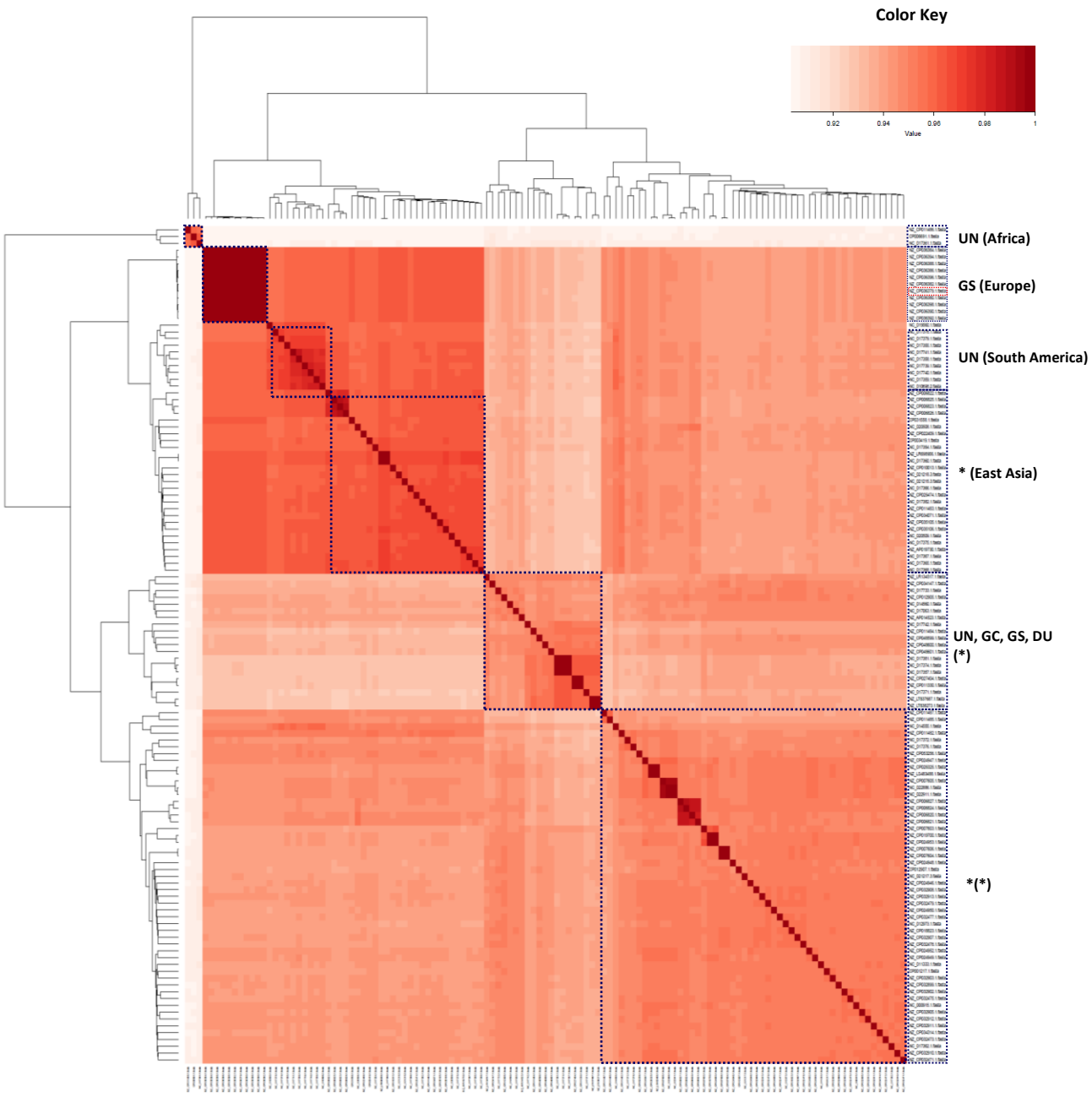


Figure 4.2: Average nucleotide identity. Clustering of the 123 genomes on the basis of average nucleotide identity (ANI). Various clusters are highlighted with the dotted line boxes. The group names based on the disease outcome are written next to the dotted boxes and geographical region is written in the round brackets (). * in the brackets or outside indicates that the strains belong to multiple geographical region and disease groups respectively.

Genomes from same geographical location seem to have the greater nucleotide identity compared to the disease outcome. Strains from East Asia, South America, Europe and Africa are clustered together.

4.3.2 Pan and Core Genome Analysis

For 123 *H. pylori* strains, a total (pangenome) of 4048 orthologous clusters were obtained. Among these the number of genes that formed the core genome was 636. Beside the core genes, some of the group specific genes were also identified as shown in Figure 4.3. For the distribution of the accessory genes see Figure A.5 of appendix. The pangenome was assigned the COGs functional annotation and the distribution of the different COG categories was observed among the groups. Half of the genes in the pangenome were assigned no functional category. In addition to this, approximately 10% (413) of the genes were classified as poorly categorized. The distribution of the pangenome into major COG categories is shown in Table 4.2. No significant difference in the distribution of COG categories was observed as it might be affected by number of genomes in each group (Figure 4.4).

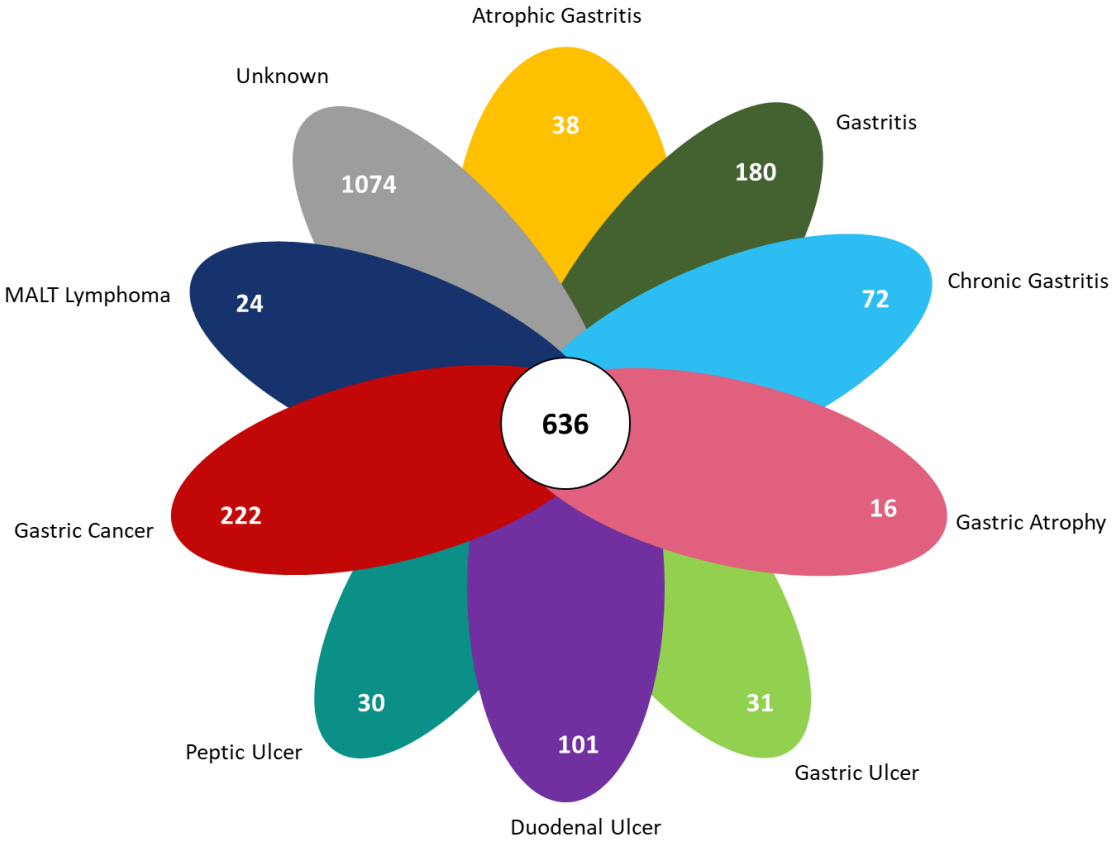
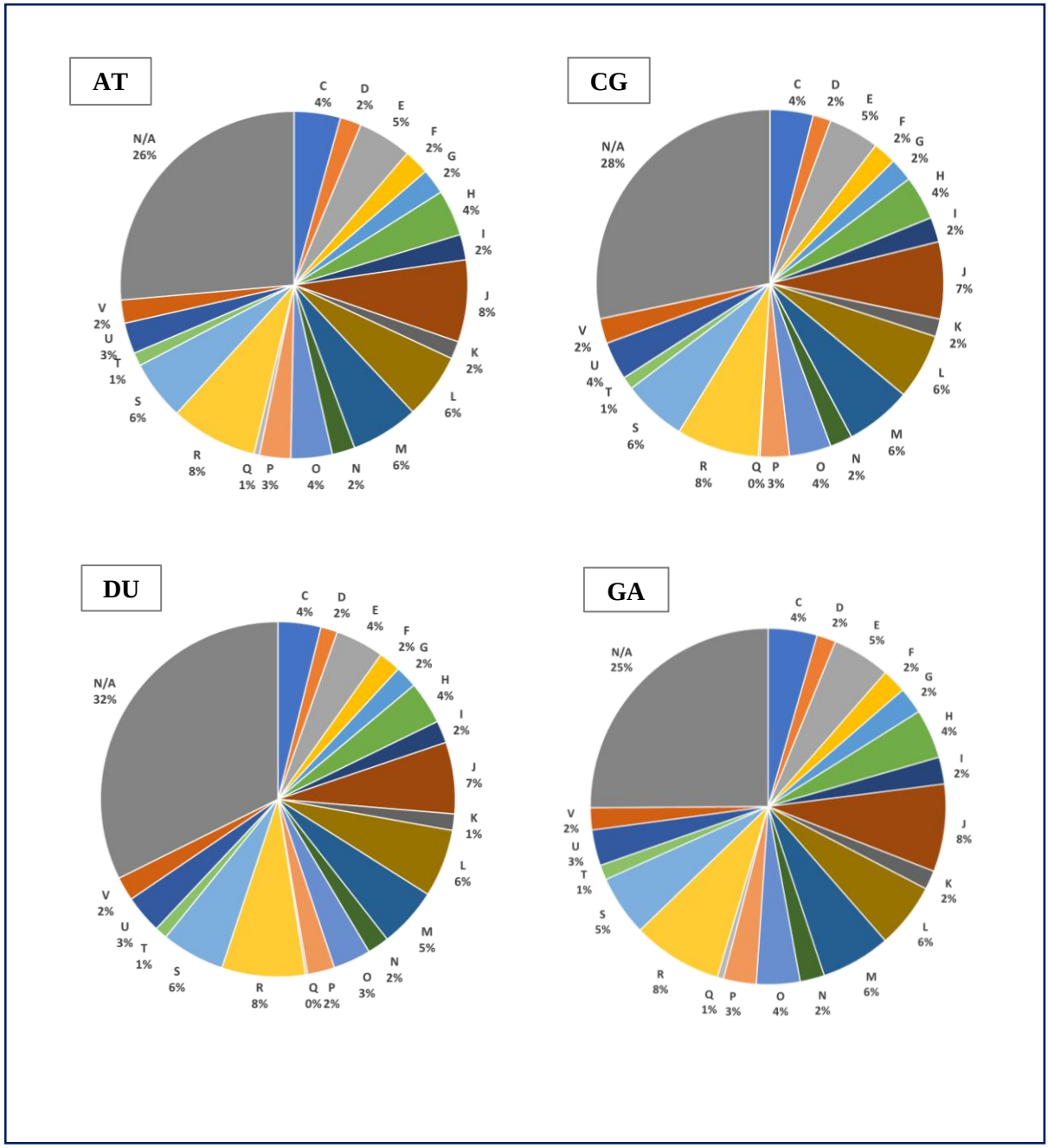


Figure 4.3: Core and group specific genes. Venn diagram showing the number of core genes and the number of specific genes observed for each group. The inner circle shows the number of core genes whereas the number at the edges of the eclipses show the number of specific genes observed for each group.

Table 4.2: Distribution of genes into major COG categories

Major COG Categories	Percentage of genes
Metabolism	13
Cellular processes and signaling	15
Information storage and processing	10
Poorly characterized	10



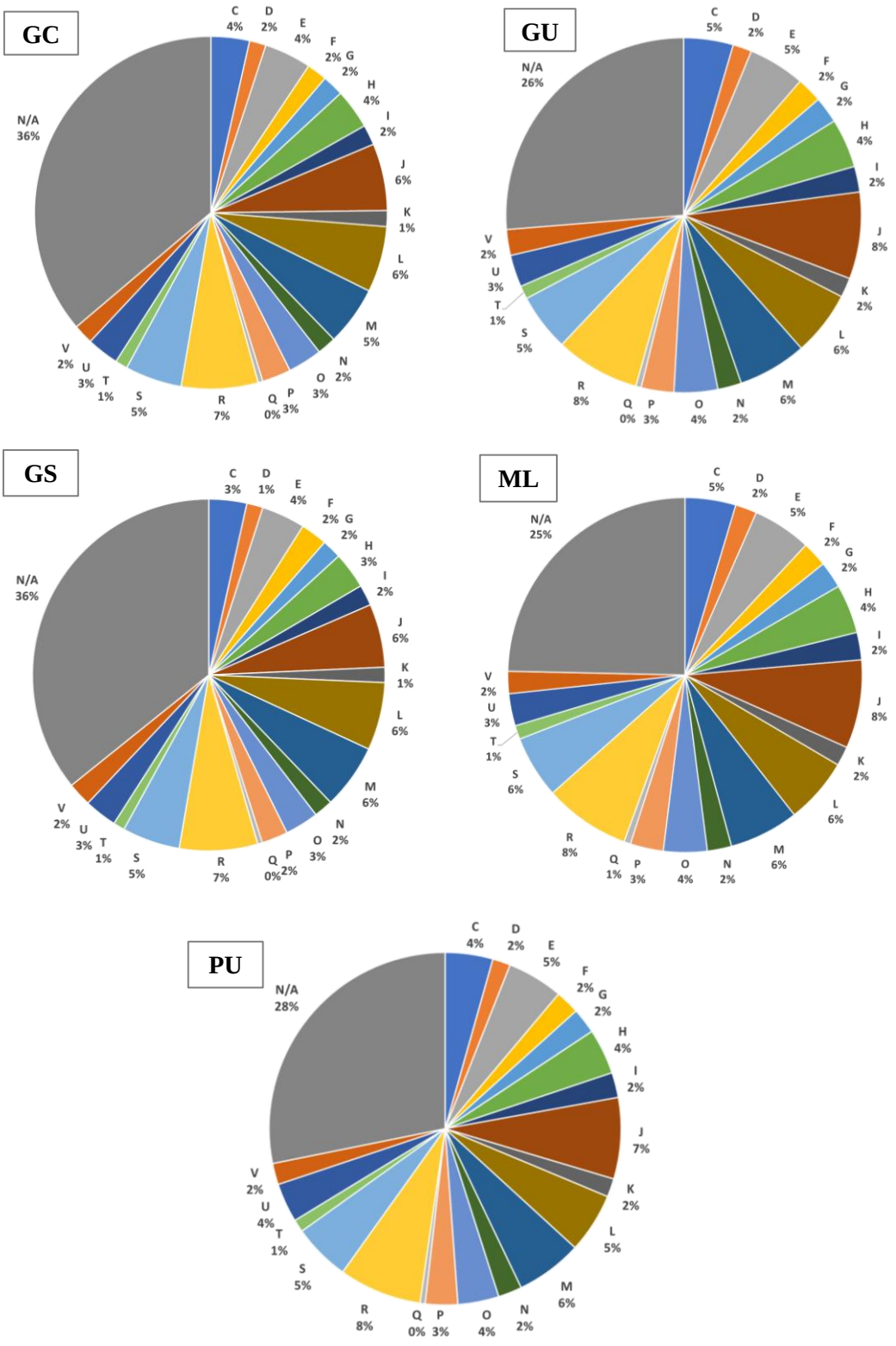


Figure 4.4: Distribution of the COG categories in the nine groups defined on the basis of disease outcomes. Names are abbreviated as: AT: Atrophic gastritis, CG: Chronic gastritis, DU: Duodenal ulcer, GA: Gastric atrophy, GC: Gastric cancer, ML: MALT lymphoma, GS: Gastritis, PU: Peptic ulcer, GU: Gastric ulcer

4.3.3 Shared and Group-specific Genes

The distribution of the total number of genes present in the strains of the different groups is shown in Figure 4.5a. The pangenome was investigated to identify the number of genes shared among the strains of each group (Figure 4.5b). About 90% of the strains for each group share the similar number of genes. For details of each group see Figure A.6 in appendix. In the pangenome, among the nine groups defined on the basis of the disease outcome, varying number of group-specific genes were observed. Large number of group specific genes were observed for the gastric cancer and gastritis groups compared to the other groups. The group-specific genes were further investigated to identify the strain-specific genes and were assigned the COG categories (Figure 4.6).

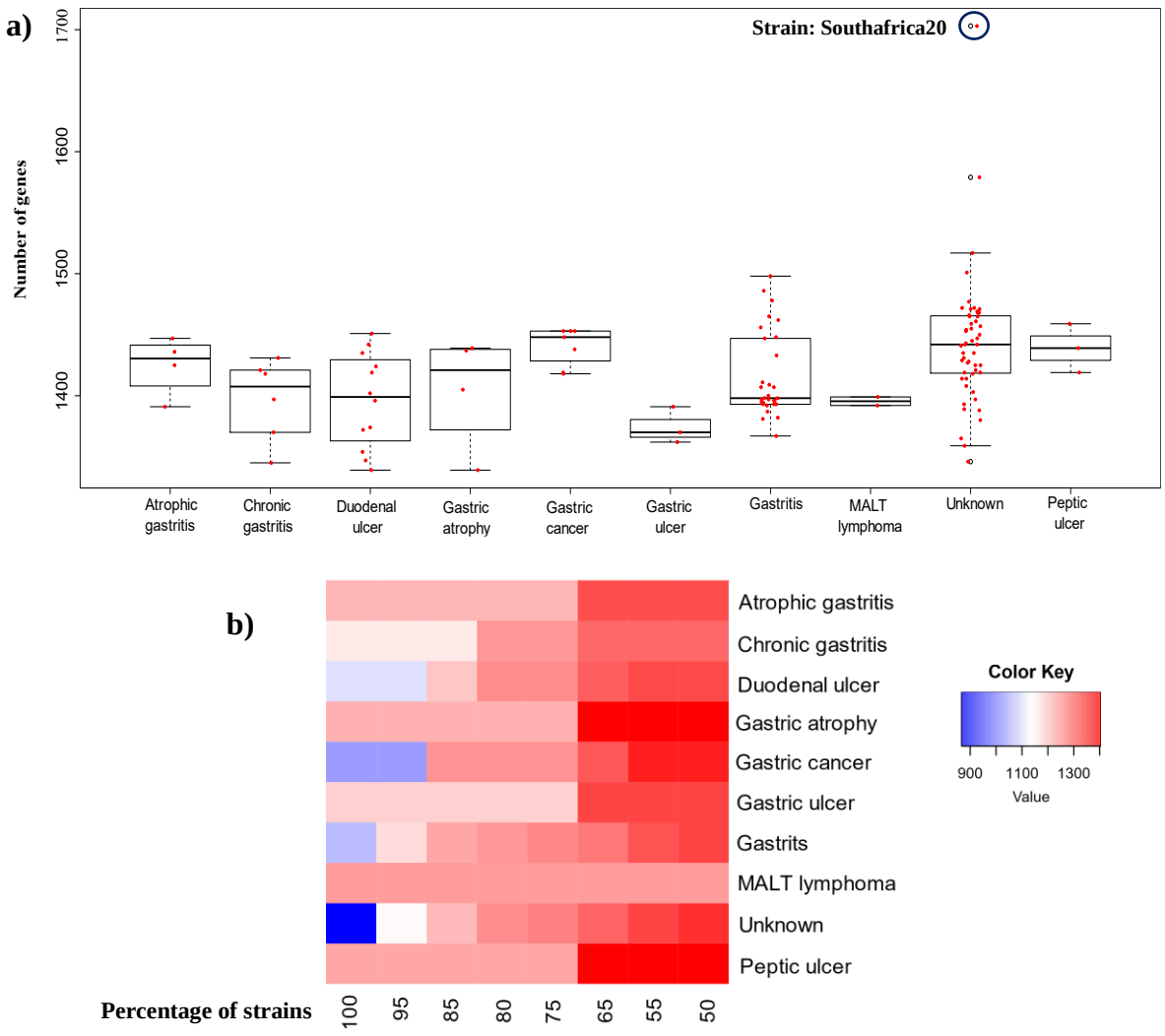
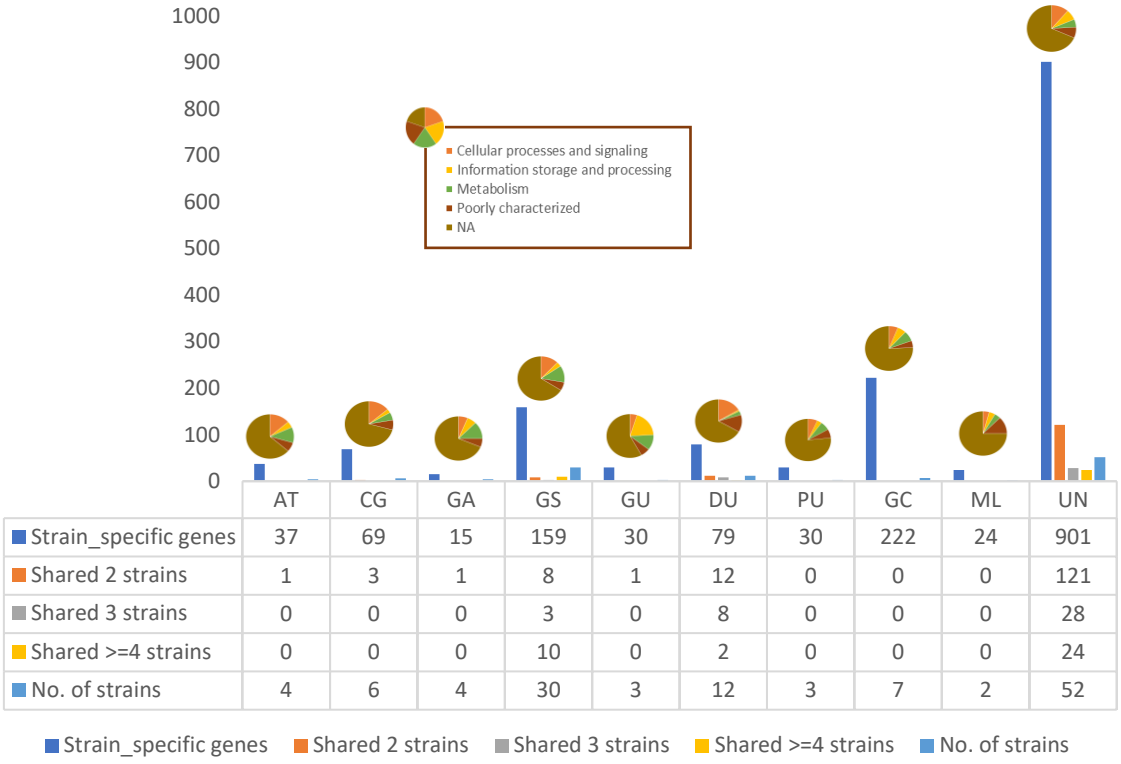


Figure 4.5: a) Distribution of number of genes in the strains of the different groups. **b)** Number of shared genes among different percentage of the strains in each group.



	AT	CG	GA	GS	GU	DU	PU	GC	ML	UN
Strain_specific genes	37	69	15	159	30	79	30	222	24	901
Shared 2 strains	1	3	1	8	1	12	0	0	0	121
Shared 3 strains	0	0	0	3	0	8	0	0	0	28
Shared >=4 strains	0	0	0	10	0	2	0	0	0	24
No. of strains	4	6	4	30	3	12	3	7	2	52

Figure 4.6: Distribution of group-specific genes. The table along the figure indicates the number of the strain-specific genes as well as the genes shared among the varying number of strains for each group. Group names are abbreviated as mentioned in the legend of Figure 4.4. The pie along the bar chart indicates the COG categories assigned to the group specific genes of each group.

Large number of group-specific genes observed in the gastritis group might be because of the large number of strains in this group. However, this does not hold for the gastric cancer group that has the significantly small number of strains but has the large number of group-specific genes compared to the gastritis group. Further, the group-specific genes of the gastric cancer group were classified as strain-specific genes as they were found in only one strain (XZ274) of the gastric cancer group.

4.3.4 Phylogenetic Analysis

The seven housekeeping genes of the 123 *H. pylori* strains were used for this analysis in order to identify how the strains with the various disease outcomes are distributed phylogenetically. No clear distribution of strains on the basis of the disease outcome was observed. Only few

clusters for the gastritis, duodenal ulcer and atrophic gastritis groups were observed (Figure 4.7a).

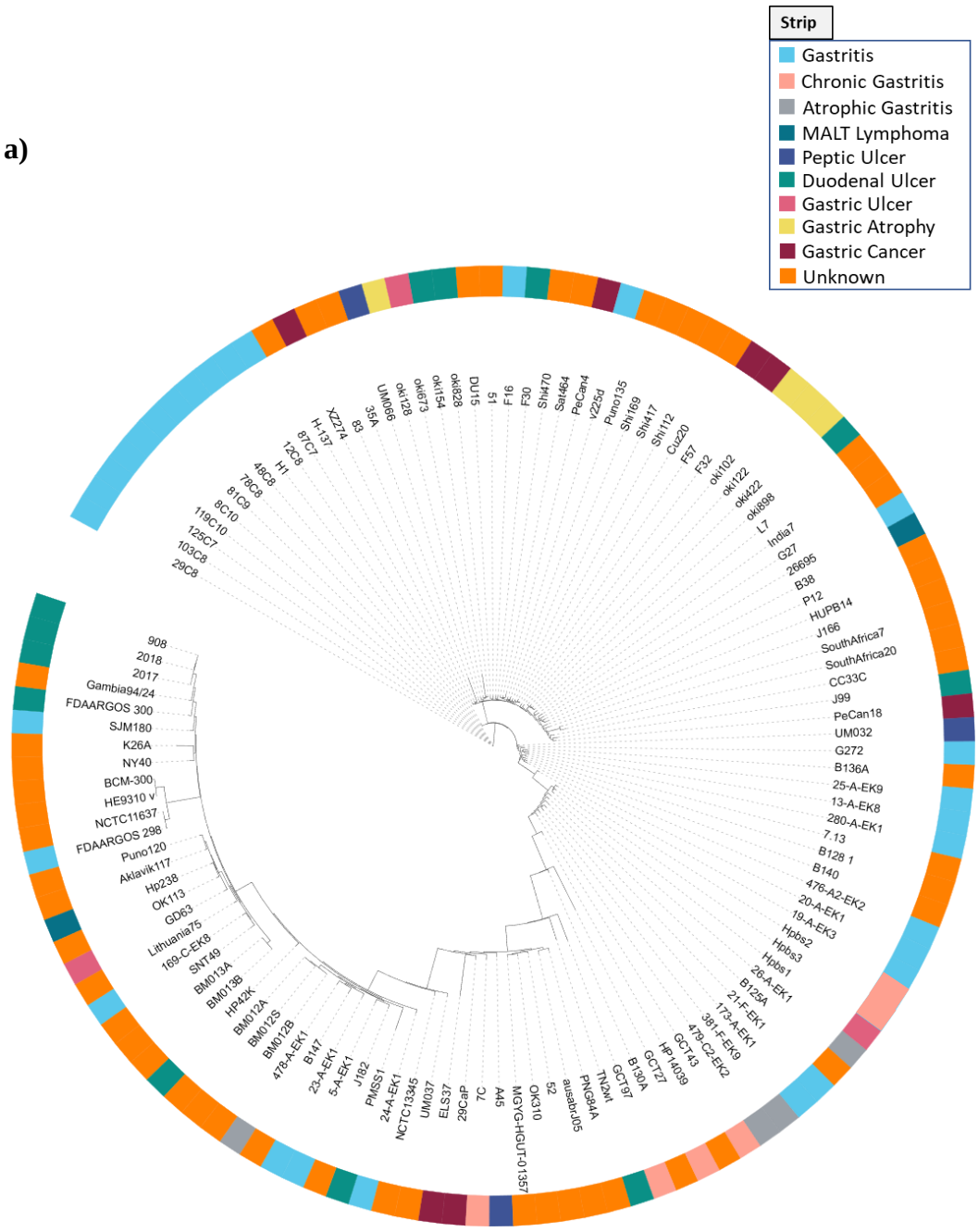


Figure 4.7: a) Phylogenetic tree of 123 *H. pylori* strains on the basis of the seven housekeeping genes. The colored strip indicates the group to which a particular strain belongs. Only few strains belonging to the gastritis, duodenal ulcer and atrophic gastritis group are clustered together. The distribution of the similar strains on the basis of regions is shown in Figure A.7a of appendix.

The distribution of the strains on the basis of the disease outcome was also not that clear in the phylogenetic analysis done on the basis of the *vacA* gene (Figure 4.7b). This analysis was done using the sequence of *vacA* gene of 105 *H. pylori* strains, as rest of the strains didn't have the *vacA* gene.

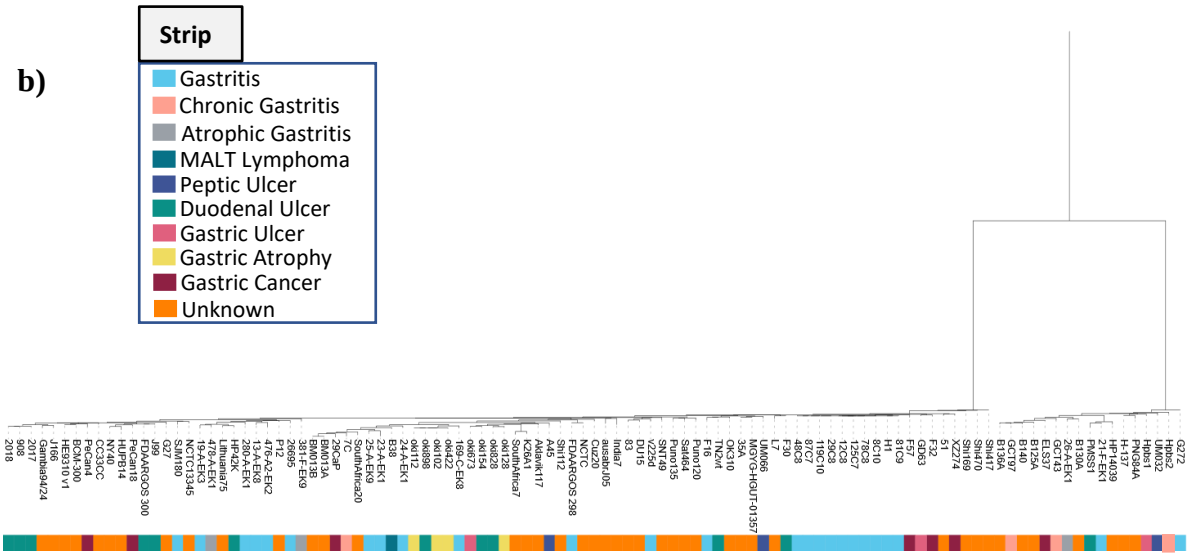
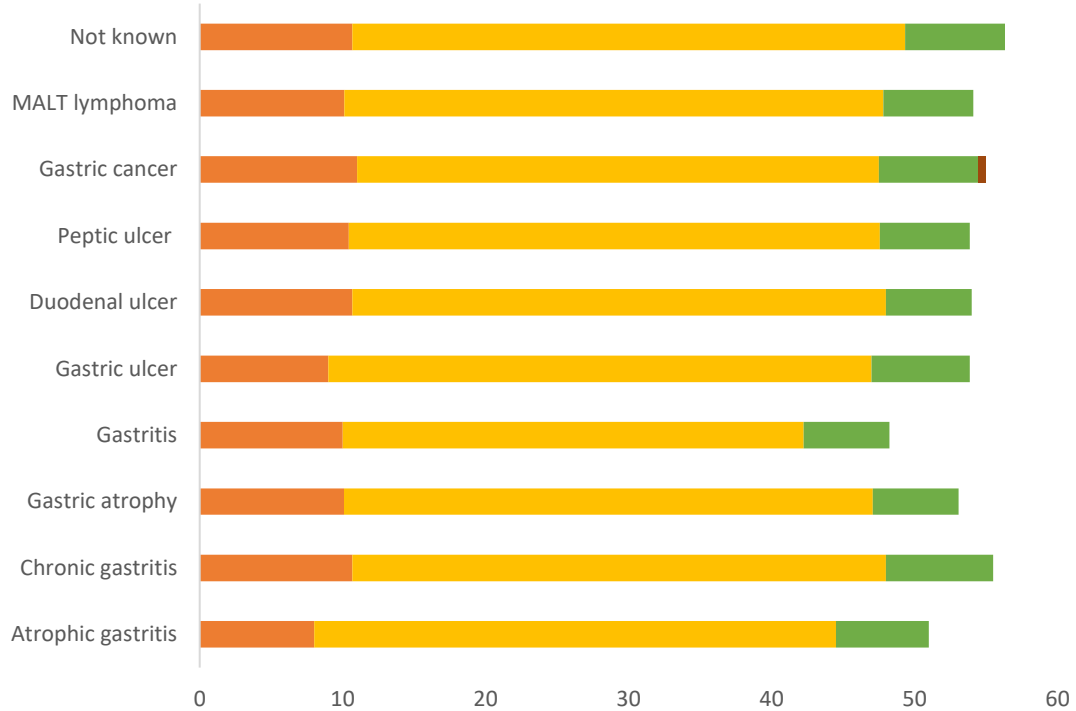


Figure 4.7: b) Phylogenetic tree of 105 *H. pylori* strains on the basis of the *vacA* gene. The colored strip indicates the group to which a particular strain belongs. Only few strains belonging to the gastritis and duodenal ulcer group are clustered together. The distribution of the similar strains on the basis of regions is shown in Figure A.7b of appendix. The legend on the left indicates the color assigned to each group.

4.3.5 Distribution of Restriction Modification Genes

The distribution of the three types of restriction modification (RM) genes named as: Type I, Type II, Type III and Type IV was analyzed for all the strains. Type II RM genes were the most commonly occurring and were observed in higher number in all the strains compared to the other RM genes. The average number of Type I, Type II and Type III RM genes found in all the groups was 10, 37 and 6, respectively. The Type IV RM genes were the least commonly occurring and were found in only a few strains (Figure 4.8). The largest number of Type IV RM genes in any strain was three and was observed in the strain of the gastritis group. The difference in the distribution of the RM genes in the different groups based on the disease outcome was not found to be statistically significant.



	Atrophic gastritis	Chronic gastritis	Gastric atrophy	Gastritis	Gastric ulcer	Duodenal ulcer	Peptic ulcer	Gastric cancer	MALT lymphoma	Not known
Type I (Avg no. of genes)	8	11	10	10	9	11	10	11	10	11
Type II (Avg no. of genes)	37	37	37	32	38	37	37	37	38	39
Type III (Avg no. of genes)	7	8	6	6	7	6	6	7	6	7
Type IV (Avg no. of genes)	0	0	0	0	0	0	0	1	0	0

Figure 4.8: Distribution of the four types of RM genes among the strains of the different groups.

4.3.6 Occurrence of cagPAI and other Virulence Genes

The information obtained from the VFDB has divided the genes into eight different virulence factor classes. The different classes are defined on the basis of the various virulence factors. These classes are named as: Acid resistance, Adherence, Immune evasion, Immune modulator, Motility, Secretion system, Toxin and one category was named as others. The number of genes in these classes were seven, ten, three, two, thirty-eight, twenty-seven, four, respectively. The number of genes in the class named others was four. The presence and absence of the cagPAI genes and the genes belonging to the other virulence factor classes is show in Figure 4.9. It was observed that strains mainly differ in the occurrence of cagPAI genes compared to the other virulence genes.

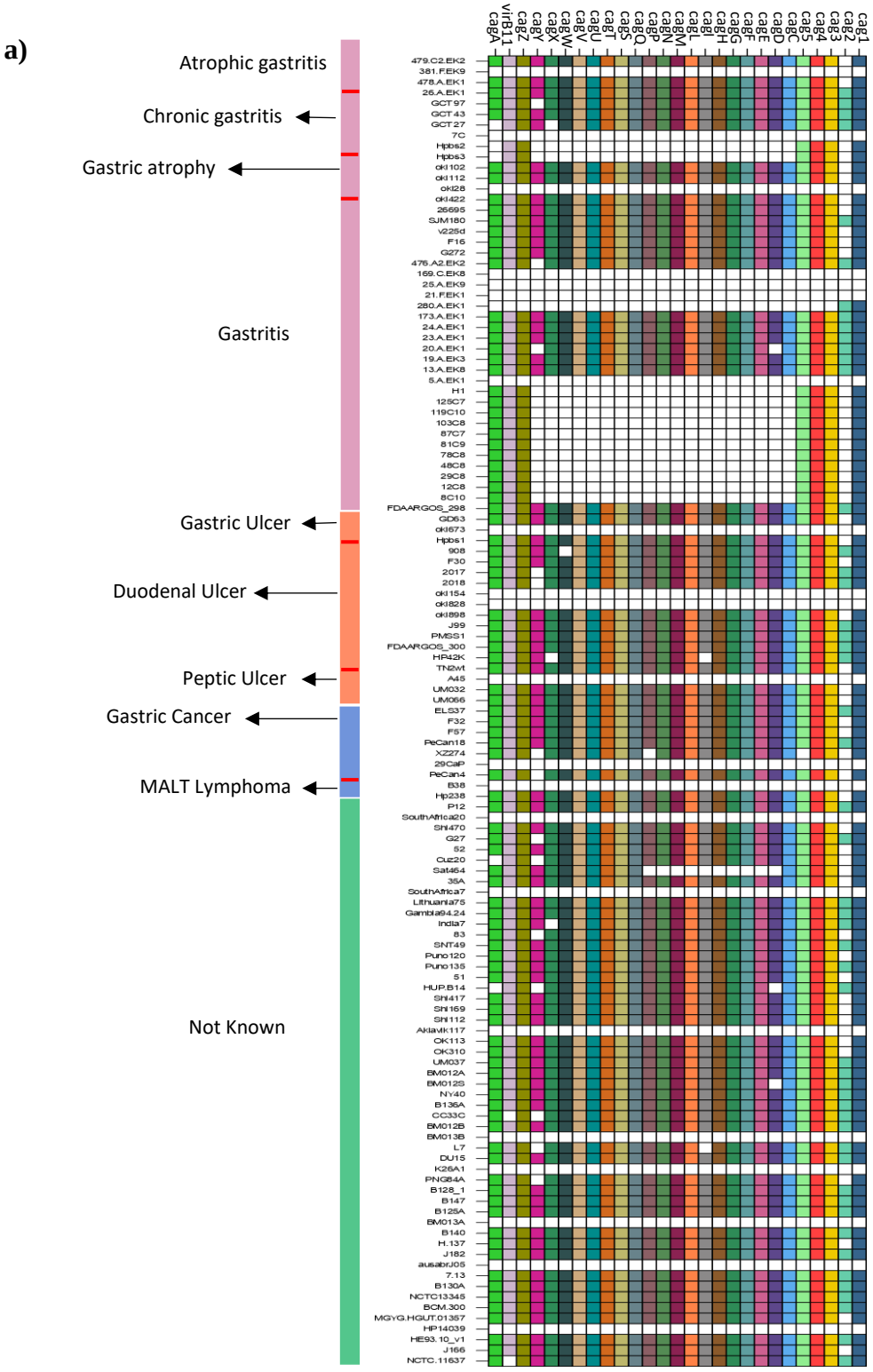


Figure 4.9: a) Presence and absence of 27 cag PAI genes in 123 *H. pylori* strains. The presence of each gene is indicated by a different color, whereas the white cell indicates that the gene is absent. Each row represents a strain whereas each column represents one of the cagPAI genes.

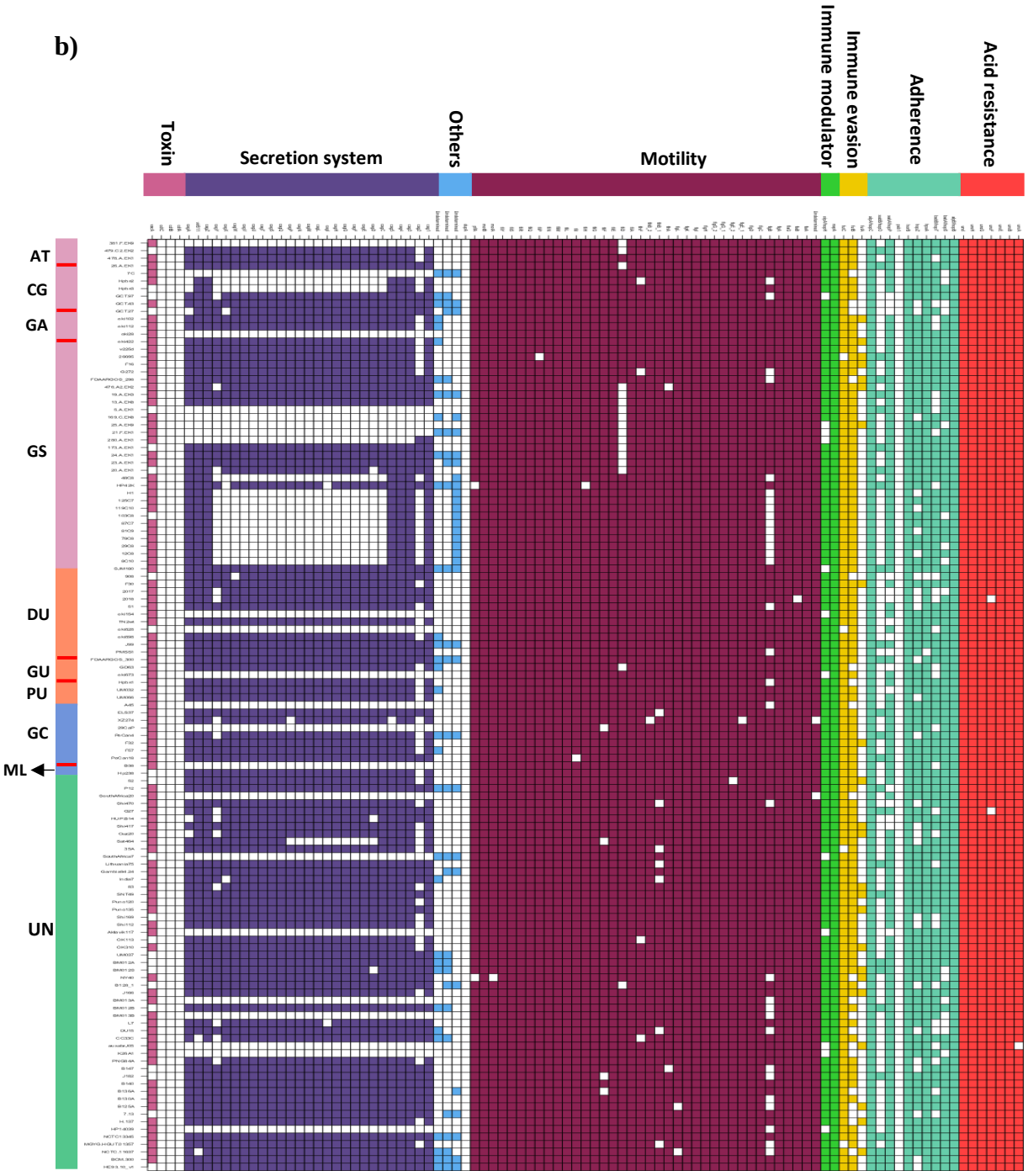


Figure 4.9: b) Presence and absence of genes belonging to the eight different virulence factor classes as defined by VFDB. The presence of genes of each class is indicated by a different color, whereas the white cell indicates that the gene is absent. The colored bar on the left indicates the different group of strains. Group names are abbreviated as mentioned in the legend of Figure 4.4. Each row represents a strain whereas each column represents one of the virulence genes.

Some of the strains completely lack the cagPAI genes, irrespective of the disease outcome. However, some of the strains that have almost all of the cagPAI genes lacked the cag2 gene. Large number of strains from the gastritis group lacked most of the cagPAI genes. Among the 123 analyzed *H. pylori* strain only thirty-five strains carried all the twenty-seven cagPAI genes. Two strains (20-A-EK1, 13-A-EK8) of the gastritis group have almost all the cagPAI genes but all the genes were not intact. The cagPAI genes were divided into two parts, one having 13 and the other having 14 genes. Genes of the other virulence gene classes, adherence and the others have a varying distribution among the strains. Some of the genes of the virulence classes: adherence, others and toxin were not found in any of the analyzed *H. pylori* strains, but were present in the species closest to the *H. pylori* in the phylogeny.

4.3.7 Presence of Repeat and Insertion Sequences

The direct and inverted repeats of length $\geq$ 25 base pairs and 100% nucleotide identity were identified for all the strains. The largest number of direct and inverted repeats was found in the strain UM037 of the unknown group. After this one of the strains in the chronic gastritis group has the largest number of direct repeats but number of inverted repeats was not that high in comparison to the strain UM037. Two other strains of chronic gastritis also had the greater number of direct repeats in comparison to the other strains, but the number of inverted repeats was relatively less compared to the strain UM037. The ratio of inverted versus direct repeats (IR/DR) was less than 1 (Figure A.8). The largest number of direct and inverted repeats were found in the strains belonging to the chronic gastritis and atrophic gastritis groups, respectively (Figure 4.10a). The distribution of repeats for each strain is shown in Figure 4.10b.

All the strains were analyzed to identify the occurrence of the five different insertion sequences (IS605-IS609). Varying distribution of the five IS elements was observed in the strains of the different groups. The largest number of copies of any IS element was observed in one of the strains of gastritis group. The strains named 476-A2-EK2 was found to have twelve copies of IS606. Besides this, one of the strains of the gastric cancer group carried six copies of IS606, the largest among the strains of the same group. Some of the strains carried only one of the genes of the IS elements. Figure 4.11 shows the distribution of the different IS elements in the analyzed strains.

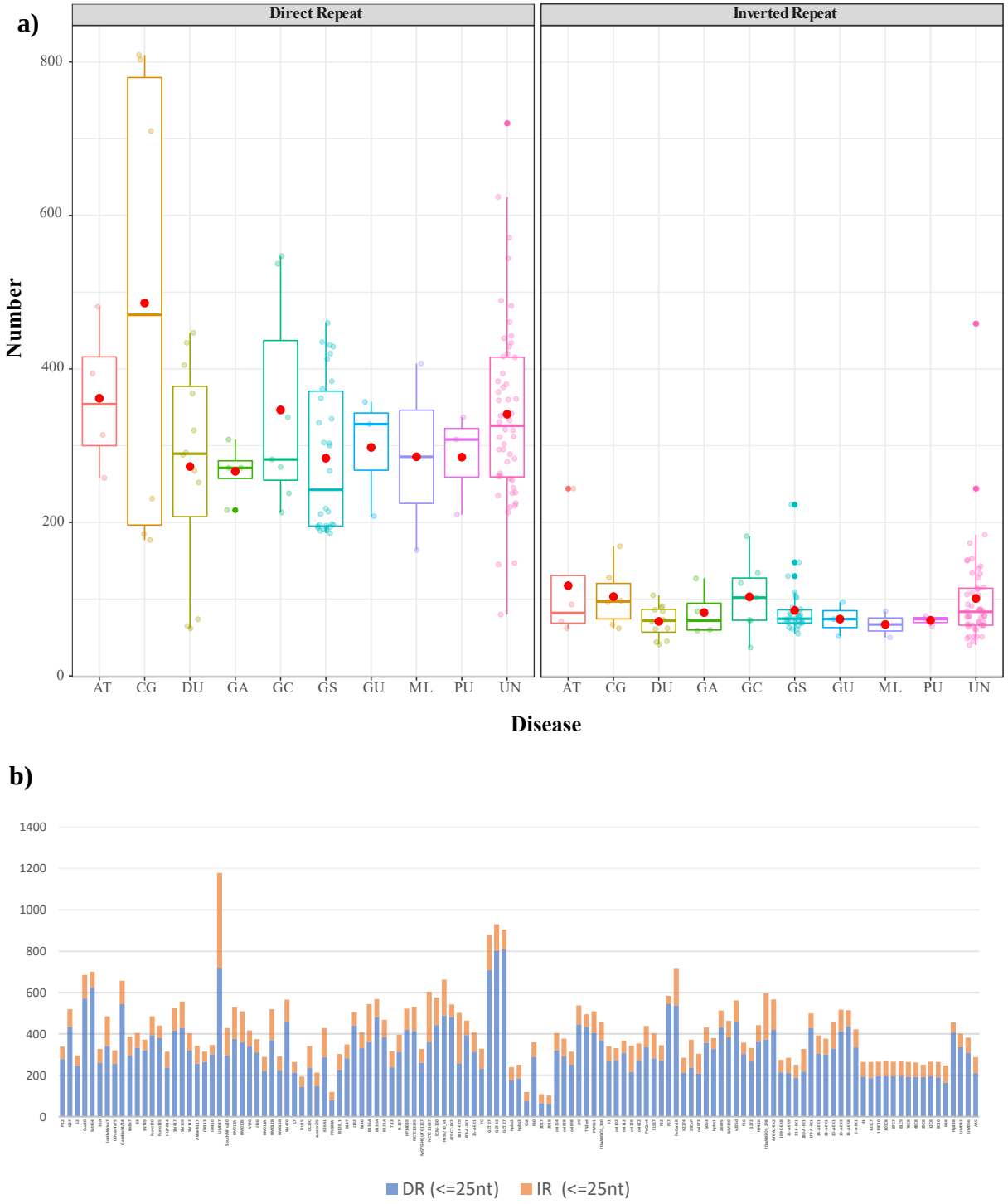


Figure 4.10: a) Distribution of direct and inverted repeats in different groups. Group names are abbreviated as mentioned in the legend of Figure 4.4. Red dot represents the average number of repeats identified in each region. **b)** Occurrence of the direct and inverted repeats in each strain. Direct and inverted repeats of length >=25 nucleotide and 100 % sequence identity are shown.

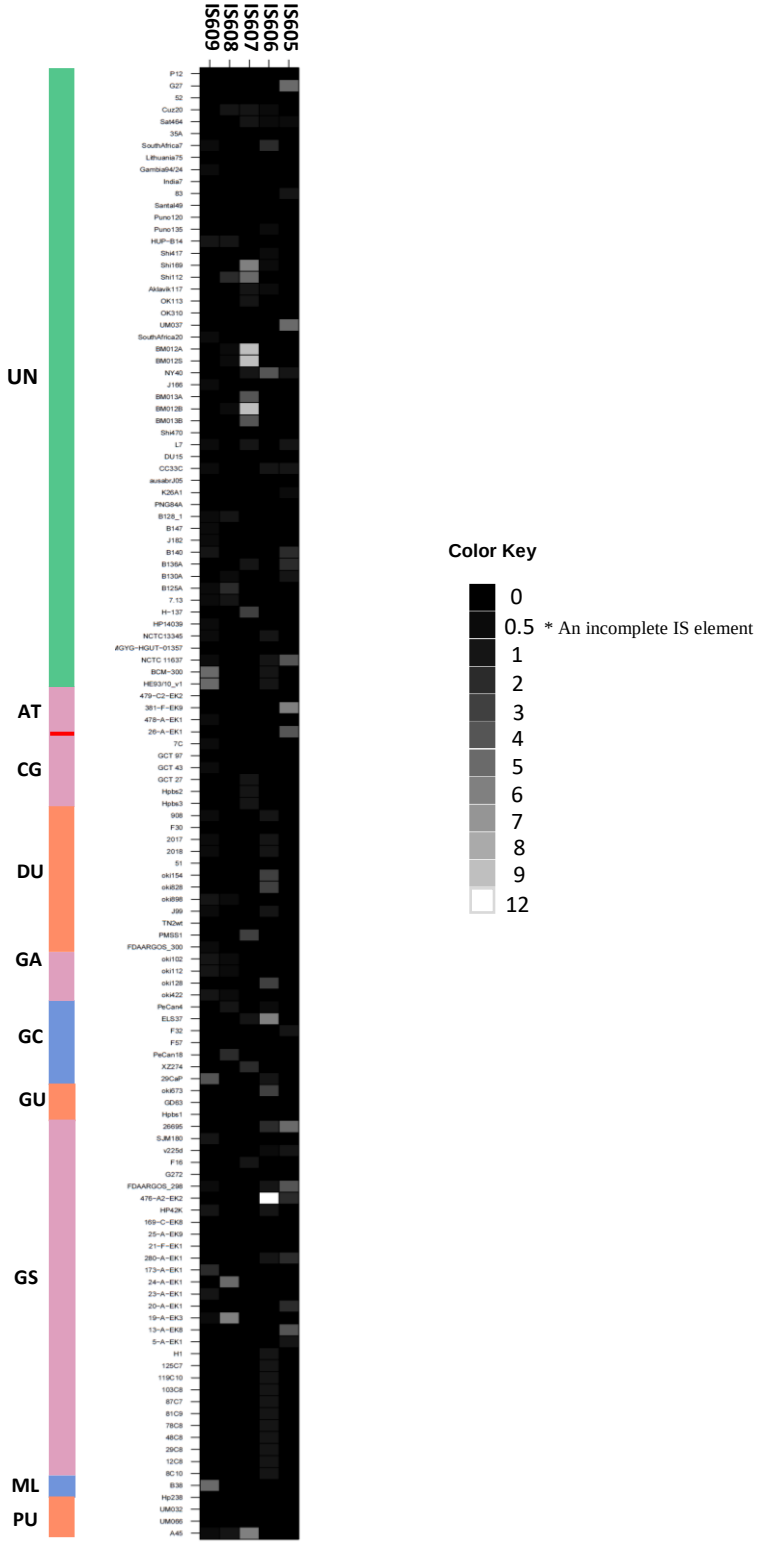


Figure 4.11: Distribution of IS elements (IS605- IS609) in different groups. Group names are abbreviated as mentioned in the legend of Figure 4.4. Row represents a strain and a column represents the IS element.

4.3.8 Gene Orders

In order to obtain gene order information of the 123 *H. pylori* strains, the clustering was performed by protein blast search using the bidirectional best-hits criterion to obtain the orthologous gene clusters. For the 123 *H. pylori* strains, 1379 gene clusters were obtained. Among these 833 gene clusters that were present in ≥ 90% of the strains were considered to identify the gene orders. Using P12 strain as a reference, gene orders were identified in all the other strains for the selected set of orthologous gene clusters. Gene order of sixty-six strains required the rotation and flipping as they didn't have the gene 1 at the start and the gene *n* at the end. Among these, gene orders of 22 strains were flipped, whereas the gene orders of 44 strains required both the rotation and flipping in order to align the gene orders. Table 4.3 shows the information of the strains that were rotated and flipped.

4.3.9 Rearrangement Analysis

The gene orders after rotation and flipping were used as an input to identify the inversions. The consensus gene order was identified using the majority rule. Later, gene orders were renumbered using the consensus gene order and the breakpoints were identified. The number of breakpoints ranged from 0 to 29 in the 123 analyzed strains (Table 4.4). Total 84 inversions were identified, 54 of which were shared, whereas 30 inversions were strain-specific. The number of inversions ranged from 0 to 26 (Table 4.5), giving rise to an assumption that the strains with no inversion are closest to the root and the one having 26 inversions being farthest from the root. One strains from the atrophic gastritis group and four strains from the unknown group had no inversions. Three strains (GCT43, GCT97, GCT27) from the chronic gastritis group had the largest number of inversions: 23, 24 and 26 inversions, respectively.

Among the 84 identified inversions, shared as well as strain-specific inversions were identified. Some of the shared inversions R14, R15, R38, R40-R42, R46 and R53 were found only in the strains of a particular disease group. These inversions can be called as *disease-specific* inversions as they were found in strains associated with only one particular disease outcome. The distribution of the identified inversions among the strains of the different groups is shown in the Figure 4.12. This figure describes the inversions that are shared, disease-specific and strain-specific.

Table 4.3: Information of operation on the gene order of 66 strains

Strain	Operation on gene order	Group
478-A-EK1	Flipping	Atrophic gastritis
479-C2-EK2	Rotation & Flipping	Atrophic gastritis
381-F-EK9	Rotation & Flipping	Atrophic gastritis
26-A-EK1	Rotation & Flipping	Atrophic gastritis
GCT27	Rotation & Flipping	Chronic gastritis
GCT97	Rotation & Flipping	Chronic gastritis
GCT43	Rotation & Flipping	Chronic gastritis
J99	Rotation & Flipping	Duodenal ulcer
FDAARGOS_300	Flipping	Duodenal ulcer
oki154	Rotation & Flipping	Duodenal ulcer
oki828	Rotation & Flipping	Duodenal ulcer
oki128	Rotation & Flipping	Gastric atrophy
Hpbs1	Rotation & Flipping	Gastric ulcer
GD63	Flipping	Gastric ulcer
oki673	Rotation & Flipping	Gastric ulcer
Hpbs3	Rotation & Flipping	Chronic gastritis
125C7	Rotation & Flipping	Gastritis
FDAARGOS_298	Flipping	Gastritis
24-A-EK1	Flipping	Gastritis
173-A-EK1	Rotation & Flipping	Gastritis
169-C-EK8	Flipping	Gastritis
20-A-EK1	Rotation & Flipping	Gastritis
25-A-EK9	Rotation & Flipping	Gastritis
8C10	Rotation & Flipping	Gastritis
29C8	Rotation & Flipping	Gastritis
476-A2-EK2	Rotation & Flipping	Gastritis
H1	Rotation & Flipping	Gastritis
87C7	Rotation & Flipping	Gastritis
81C9	Rotation & Flipping	Gastritis
280-A-EK1	Rotation & Flipping	Gastritis
19-A-EK3	Rotation & Flipping	Gastritis
G272	Rotation & Flipping	Gastritis

103C8	Rotation & Flipping	Gastritis
48C8	Rotation & Flipping	Gastritis
119C10	Rotation & Flipping	Gastritis
78C8	Rotation & Flipping	Gastritis
5-A-EK1	Flipping	Gastritis
13-A-EK8	Rotation & Flipping	Gastritis
23-A-EK1	Flipping	Gastritis
12C8	Rotation & Flipping	Gastritis
21-F-EK1	Rotation & Flipping	Gastritis
Hpbs2	Rotation & Flipping	Chronic gastritis
A45	Flipping	Peptic ulcer
UM066	Rotation & Flipping	Peptic ulcer
UM032	Rotation & Flipping	Peptic ulcer
B140	Rotation & Flipping	Unknown
PNG84A	Rotation & Flipping	Unknown
B128_1	Rotation & Flipping	Unknown
NCTC11637	Flipping	Unknown
CC33C	Flipping	Unknown
NY40	Flipping	Unknown
DU15	Flipping	Unknown
B125A	Rotation & Flipping	Unknown
35A	Flipping	Unknown
MGYG-HGUT-01357	Flipping	Unknown
B130A	Flipping	Unknown
B147	Flipping	Unknown
UM037	Flipping	Unknown
HE93/10_v1	Flipping	Unknown
7.13	Rotation & Flipping	Unknown
J182	Flipping	Unknown
H-137	Rotation & Flipping	Unknown
NCTC13345	Flipping	Unknown
L7	Flipping	Unknown
B136A	Rotation & Flipping	Unknown
HP14039	Rotation & Flipping	Unknown

Table 4.4: Number of breakpoints identified in each strain

No. of Breakpoints	No. of strains	Strains annotation
0	4	479-C2-EK2, Cuz20, Puno135, Shi417
1	1	Shi169
2	15	BM013A, BM013B, Puno120,173-A-EK1, Shi112, Sat464, B125A, SJM180, Shi470, oki154, Lithuania75, PeCan4, oki828, oki673, SNT49
3	2	P12, 21-F-EK1
4	15	G27, 169-C-EK8, 20-A-EK1, NY40, B38, J99, 29CaP, 7C, FDAARGOS_300, v225d, Gambia94/24, oki128, ausabrJ05, F30, Aklavik117
5	14	2018, PNG84A, DU15, OK113, 2017, oki422, B128_1,25-A-EK9, oki102, oki112, oki898, SouthAfrica7, SouthAfrica20, 7.13
6	18	908, 24-A-EK1, 478-A-EK1, Hp238, 51, 52, G272, ELS37, HUP-B14, J182, F57, 5-A-EK1,13-A-EK8, 23-A-EK1, PeCan18, 476-A2-EK2, 280-A-EK1, J166
7	9	Hpbs3, Hpbs1, B147, H-137, HP42K, Hpbs2, HP14039, India7, 19-A-EK3
8	11	CC33C, 83, 35A, MGYG-HGUT-01357, PMSS1, UM032, F16, OK310, 26695, 26-A-EK1, K26A1
9	2	B140, GD63
10	13	A45, NCTC11637, F32, UM066, HE93/10_v1, BCM-300, UM037, 381-F-EK9, L7, B140, BM012S, BM012B, BM012A, B136A
11	12	FDAARGOS_298, 8C10, 29C8, H1, 87C7, 81C9, B130A, 103C8, 48C8, 119C10, 78C8, 12C8
12	2	XZ274, 125C7
13	1	TN2wt
15	1	NCTC13345
27	1	GCT27
28	1	GCT97
29	1	GCT43

Table 4.5: Number of inversions identified in each strain

No. of Inversions	No. of strains	Strains annotation
0	5	479-C2-EK2, Cuz20, Puno135, Shi417, Shi169
1	17	P12, BM013A, BM013B, Puno120, 173-A-EK1, Shi112, Sat464, B125A, SJM180, G27, Shi470, oki154, Lithuania75, PeCan4, oki828, oki673, SNT49
2	20	2018, PNG84A, 169-C-EK8, 20-A-EK1, NY40, DU15, B38, J99, 908, FDAARGOS_300, 29CaP, v225d, Gambia94/24, OK113, 7C, oki128, 2017, ausabrJ05, 21-F-EK1, F30
3	32	oki422, Hpbs3, 24-A-EK1, 478-A-EK1, Hp238, B128_1, 25-A-EK9, oki102, oki112, oki898, CC33C, Hpbs1, SouthAfrica7, SouthAfrica20, 51, Aklavik117, B147, 52, G272, ELS37, HUP-B14, 7.13, J182, H-137, F57, 5-A-EK1, 13-A-EK8, 23-A-EK1, PeCan18, HP42K, Hpbs2, HP14039
4	13	XZ274, 83, 476-A2-EK2, 35A, MGYG-HGUT-01357, India7, 280-A-EK1, 19-A-EK3, PMSS1, UM032, GD63, F16, OK310
5	10	FDAARGOS_298, A45, NCTC11637, F32, UM066, HE93/10_v1, 26695, HE170/09, J166, BCM-300, 26-A-EK1
6	4	UM037, K26A1, 381-F-EK9, L7
7	2	B140, TN2wt
8	17	125C7, BM012S, 8C10, 29C8, BM012B, H1, 87C7, 81C9, B130A, 103C8, 48C8, 119C10, BM012A, 78C8, NCTC13345, 12C8, B136A
23	1	GCT43
24	1	GCT97
26	1	GCT27

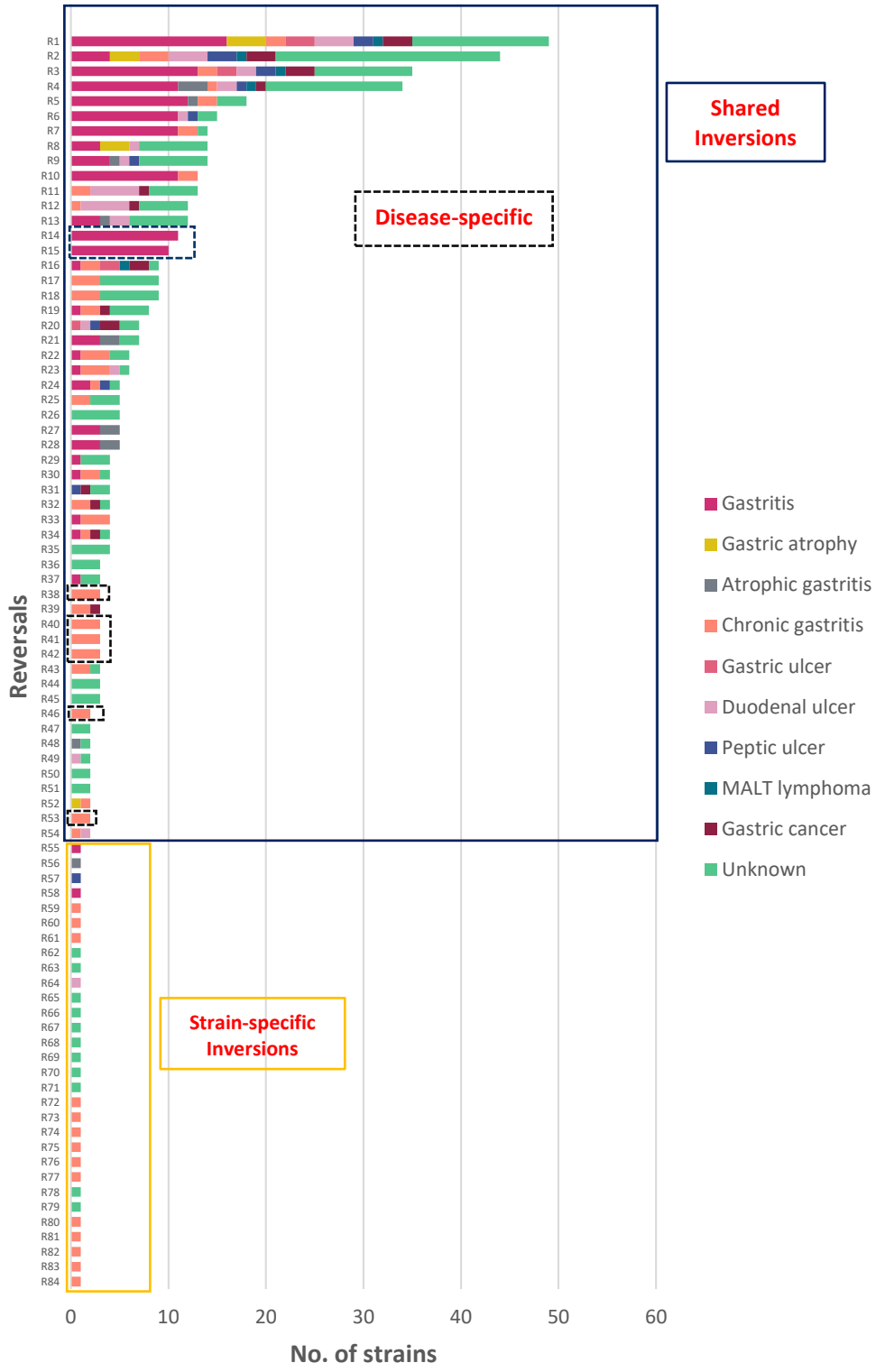


Figure 4.12: Distribution of inversions. R1-R54 and R55-R84 were identified as shared and strain-specific inversions, respectively. Among the shared inversions, disease-specific inversions are shown in dotted boxes.

4.3.10 Shared and Strain-specific Inversions

Among the shared inversions, the inversion R1 was found in the strains from all the groups except the strains from the atrophic gastritis group. Similarly, the inversion R2 was found in all the groups except the gastric ulcer and atrophic gastritis group. The strains from the chronic gastritis group showed all types of inversions: shared, disease-specific and the strain-specific inversions. The three out of six strains of the chronic gastritis group had the largest number of inversions compared to the others. Besides this, large number of repeat sequences was also observed in these strains (See Figure 4.10). The large number of inversions in these strains might be the result of homologous recombination between the repeat sequences. The three most rearranged strains of the chronic gastritis group were isolated from different individuals and the geographical location mentioned in the database is Colombia with a distinct region for each strain. The other strains of the chronic gastritis group had only two or three inversions and possessed very few repeat sequences compared to these three strains of this group. The geographical location of these strain mentioned in the database is Mexico and China. The differences observed in these strains compared to the other strains of the chronic gastritis group might be because of the distinct geographical location. As the human population world-wide is infected with *H. pylori* and its genomic diversity is thought to be the result of human migration.

Most of the strain-specific inversions were present in the strains of the chronic gastritis group. Only one strain-specific inversion was observed for the atrophic gastritis, peptic ulcer and duodenal ulcer groups. Two strain-specific inversions were observed for the gastritis group and the rest were found in the strains with unknown disease outcome. In my previous analysis described in Chapter 3 of this dissertation, I found two inversions that were strain-specific and were observed in the strains isolated from the patients with gastric cancer. I assumed that these inversions might be associated with disease state, however in this analysis no inversion was found to be associated with the gastric cancer. The two strain-specific inversions that were previously identified and were assumed to be associated with a disease state might be shared with the strains from other groups in the current analysis as the number of the analyzed strains in this study is large compared to the previous one. Figure 4.13a shows the clustering of the strains on the basis of presence and absence of the identified inversions and Figure 4.13b shows the phylogram.

a)

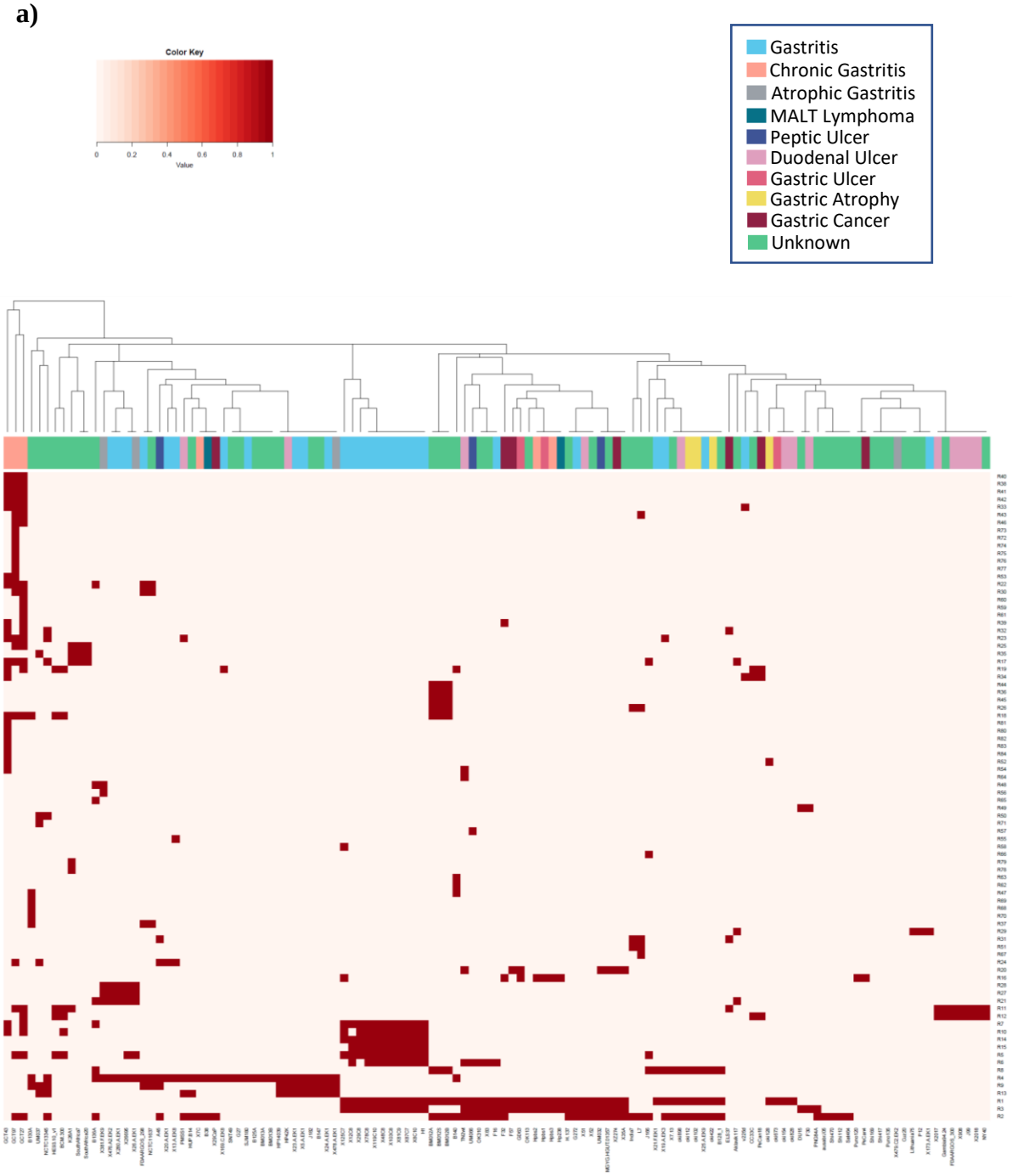


Figure 4.13: a) Hierarchical clustering of the strains based on the presence and absence of the 84 identified inversions. The colored bar indicates the group of each strain showed in the upper right legend. Row represents the inversion and the column represents the strain.

b)

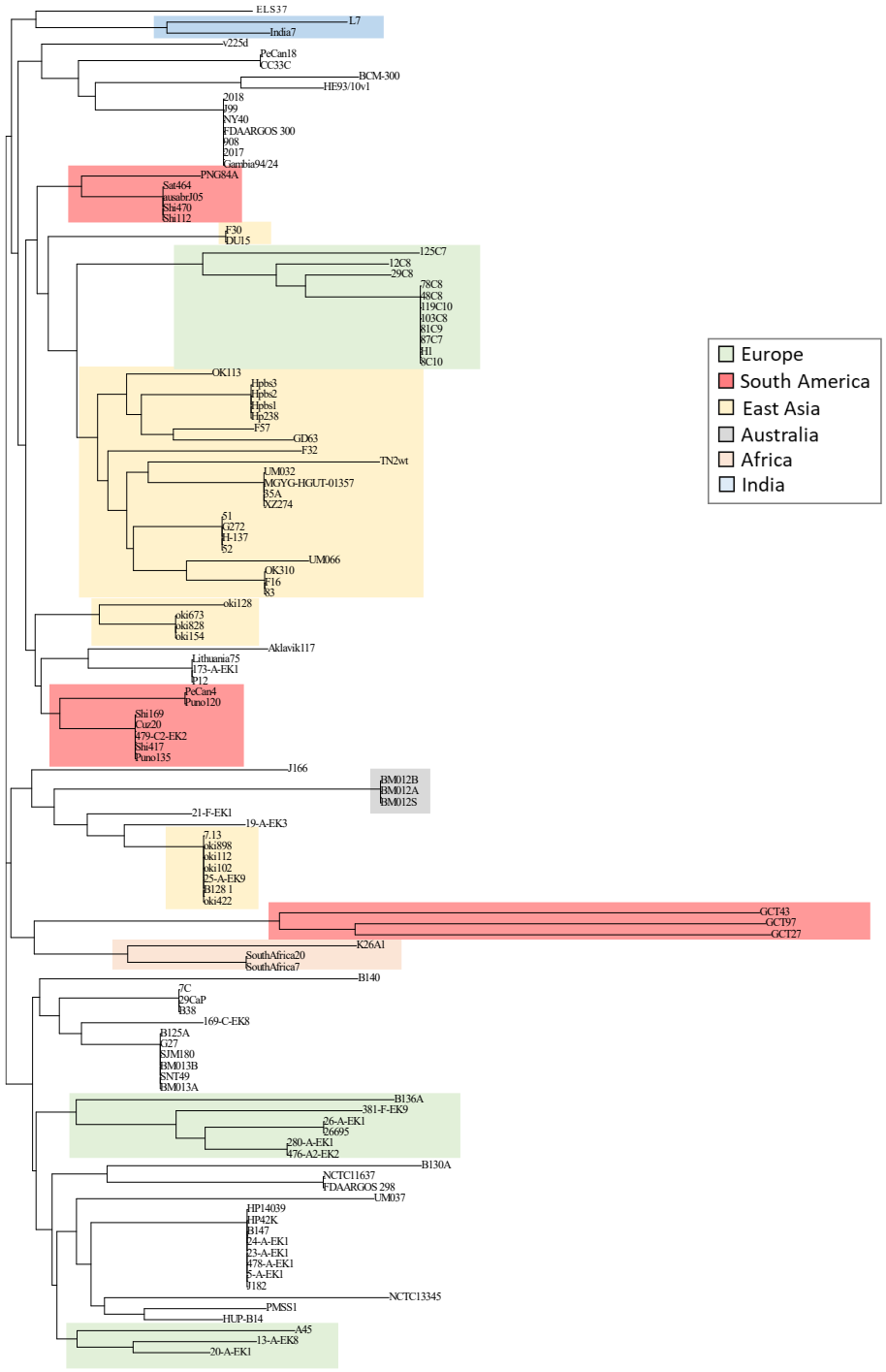


Figure 4.13: b) Neighbor joining phylogram of the strains based on the presence and absence of the 84 identified inversions. Here, the clades are colored based on the geographical location of the majority of the strains present in a particular clade. The legend shows the region to which each color corresponds.

4.4 Conclusions

Comparative genomics approach can help us to understand the genomic diversity of the species and identify the genomic features that are associated with the genetic variability. *Helicobacter pylori* being one of the diverse bacteria can be used as a model to investigate the factors that contribute to its diversity. As *H. pylori* can cause a wide range of disease from gastritis to gastric cancer, investigating the association of different genomic features with the disease outcome can help us understand the disease mechanism. Comparison of the strains of the different groups defined on the basis of the disease outcome revealed the presence of certain group-specific genes. The distribution of COG categories in the different groups didn't show any significant difference as it might be affected by the number of the strains. About 90% of the strains in each group share the similar number of genes. The distribution of the IS elements, cagPAI, virulence and restriction modification genes showed the varying distribution in the analyzed strains irrespective of the disease outcome. The large number of repeat sequences were observed in the three strains of the chronic gastritis group. These strains had the largest number of the inversions that might be the result of the homologous recombination between the repeat sequences. Most of the inversions were found to be shared among the strains from different groups, only few were found to be disease-specific. The analysis revealed that the strains were more related based on their geographical locations rather than the disease outcome (Figure 4.13). As the disease outcome of the *H. pylori* depends on other factors like environment, host and diet [153, 157], that might be the cause of weak association of the genomic features with the disease state. Besides this, large number of strains didn't have the information of the disease state which also made it difficult to find the association of the genomic features with the particular disease group. Analyzing the larger dataset along with the information about the disease state as well as geography can help us understand the disease mechanism and the genetic variability of the *H. pylori*. In addition, since *H. pylori* has been associated with humans and has evolved along with their migration, identifying the differences with respect to the geographical location can also reveal the human migration patterns.

Chapter 5

General Discussion and Conclusions

In this doctoral thesis, I reported an algorithm that I have developed for the identification of the genome rearrangements while comparing the multiple bacterial genomes. I have used the orthologous gene cluster data to obtain the order of genes in the multiple genomes. Gene order identification is important as it makes the multiple genome comparison easier. The algorithm takes the gene order data as an input and identify the genome rearrangements. My algorithm can also handle the gene order data with the missing genes. Initially, I considered only those gene clusters that were almost conserved (present in all except one) but later the algorithm was improved to handle the different set of genes that are conserved in ~85 to 100 percent of the genomes. The algorithm not only identifies the reversals that are shared by several genomes but also the ones that are specific for certain genomes. The obvious benefit of my algorithm is scalability: whole genome comparison is difficult for many genomes using previous approaches comparing two genomes. My algorithm can handle hundreds of strains at the level of gene orders. Besides this, it can also handle the large number of missing genes.

To demonstrate the use of my algorithm, I have used the *Helicobacter pylori* genomes as this bacterium has a very diverse genomic structure. The analysis of the 72 *H. pylori* genomes revealed the presence of 41 inversions among which 18 were found strain specific whereas 23 were shared. Three regions were identified as rearrangement hotspots as they were found to be frequently involved in the rearrangements. The largest number of inversions in any strain was six and were found in three strains from Australia and one strain from East Asia. Some of the shared inversions were found in the strains having the same geographical locations and were called the *region-specific*. The region-specific inversions were observed for the strains from Australia, East Asia and Africa. Some inversions associated with the disease-state such as cancer were also identified in my analysis. It was also identified that the inversions were of variable sizes and the three largest inversions were found in the strains from East Asia. One of the largest inversions was found in the strains from Okinawa Japan. The pattern of inversions was most diverse in Japan probably because of the larger number of sampling. The North American region also had the diverse inversion pattern even though the number of samples was

much smaller compared to Japan. This diversity occurred maybe because of human migration. Genome rearrangements might be the result of various biological mechanisms. The analysis of the breakpoints of the inversions showed that most of the shared inversions possessed similar IS elements with a few exceptions. This suggests that these elements are well-conserved irrespective of the different geographical region. Some of the inversions were associated with the inverted or the direct repeats sequences.

The number of genomes investigated initially was small, so I performed a larger scale analysis to understand the association of genomic features more specifically the genome rearrangements with the disease outcome. Since *H. pylori* can cause different diseases, I identified the genome rearrangements to find their association with a particular disease outcome. Comparative analysis of the strains revealed the presence of certain group-specific genes. Besides this, no significant difference in the distribution of the IS elements, RM genes and repeat sequences was observed in the strains of the different groups defined on the basis of the disease outcome. Most of the inversions were found to be shared among the strains from different groups, only few were found to be disease-specific. The analysis revealed that the strains were more related based on their geographical locations rather than the disease outcome. Disease outcome of infection with *H. pylori* depends on several other factors so it might be the cause of weak association that was observed for the genomic features with the disease state. In addition, most of the strains didn't have the information of the disease state thus making it difficult to find the association.

This study provides the simple and scalable algorithmic approach that can be used for the identification of the genome rearrangements while comparing the multiple genomes. It describes how gene orders can be used to identify the rearrangement events that took place during the evolution of the organisms. It also provides insight into how the rearrangements events can be further analyzed to find the genomic elements that can be the possible drivers of these genome rearrangements. It also reports how to identify the association of genome rearrangements with a particular phenotype such as disease outcome in this study.

References

1. Martin Bader. Sorting by weighted transpositions and reversals. *University of Ulm, Faculty of Computer Science.* 2005
2. Zeira R, Shamir R. Genome rearrangement problems with single and multiple gene copies: a review. *Bioinformatics and Phylogenetics.* 2019:205-241.
3. https://basicbiology.net/biology-101/introduction-to-cells
4. Blanchette M. Evolutionary puzzles: An introduction to genome rearrangement. In *International Conference on Computational Science*. 2001(pp. 1003-1011). Springer, Berlin, Heidelberg.
5. Barış Ö, Karadayı M, Yanmış D, Güllüce M. Genomic Rearrangements and Evolution. In *Current Progress in Biological Research.* 2013 (pp. 19-39). IntechOpen.
6. Brown TA. Mutation, repair and recombination. In *Genomes.* 2nd edition 2002. Wiley-Liss.
7. Najafi MB, Pezeshki P. Bacterial mutation; types, mechanisms and mutant detection methods: a review. *European Scientific Journal.* 2013
8. Bader M, Abouelhoda MI, Ohlebusch E. A fast algorithm for the multiple genome rearrangement problem with weighted reversals and transpositions. *BMC bioinformatics.* 2008;9(1):1-13
9. Lara-Ramírez EE, Segura-Cabrera A, Guo X, Yu G, García-Pérez CA, Rodríguez-Pérez MA. New implications on genomic adaptation derived from the *Helicobacter pylori* genome comparison. *PloS one.* 2011;6(2): e17300.
10. Periwal V, Scaria V. Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics.* 2015;31(1):1-9.
11. Block DH, Hussein R, Liang LW, Lim HN. Regulatory consequences of gene translocation in bacteria. *Nucleic acids research.* 2012;40(18):8979-8992.
12. Wang EA, Mowry KL, Clegg DO, Koshland Jr DE. Tandem duplication and multiple functions of a receptor gene in bacterial chemotaxis. *Journal of Biological Chemistry.* 1982;257(9):4673-4676.
13. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nature Reviews Genetics.* 2009;10(8):551-564.
14. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences.* 1998;95(12):6578-6583.

15. Darmon E, Leach DR. Bacterial genome instability. *Microbiology and Molecular Biology Reviews.* 2014;78(1):1-39.
16. Krawiec ST, Riley MO. Organization of the bacterial chromosome. *Microbiology and Molecular Biology Reviews.* 1990;54(4):502-539.
17. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS genet.* 2009;5(1): e1000327.
18. Kresse AU, Dinesh SD, Larbig K, Römling U. Impact of large chromosomal inversions on the adaptation and evolution of Pseudomonas aeruginosa chronically colonizing cystic fibrosis lungs. *Molecular microbiology.* 2003;47(1):145-158.
19. Liang Y, Hou X, Wang Y, Cui Z, Zhang Z, Zhu X, Xia L, Shen X, Cai H, Wang J, Xu D. Genome rearrangements of completely sequenced strains of Yersinia pestis. *Journal of clinical microbiology.* 2010;48(5):1619-1623.
20. Skovgaard O, Bak M, Løbner-Olesen A, Tommerup N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome research.* 2011;21(8):1388-1393.
21. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical reviews in microbiology.* 2017;43(6):709-730.
22. Ohtsubo E, Sekine Y. Bacterial insertion sequences. *Transposable elements.* 1996:1-26.
23. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS microbiology reviews.* 2014;38(5):865-891.
24. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's guide to bacterial insertion sequences. *Mobile DNA III.* 2015:555-590.
25. Berg DE, Howe MM, Ajioka JW. Mobile DNA: American Society for Microbiology Washington. 1989
26. Mahillon J, Chandler M. Insertion sequences. *Microbiology and molecular biology reviews.* 1998;62(3):725-774.
27. Turlan C, Chandler M. Playing second fiddle: second-strand processing and liberation of transposable elements from donor DNA. *Trends in microbiology.* 200;8(6):268-274.
28. Kersulyte D, Velapatiño B, Dailide G, Mukhopadhyay AK, Ito Y, Cahuayme L, Parkinson AJ, Gilman RH, Berg DE. Transposable element ISHp608 of *Helicobacter*

pylori: nonrandom geographic distribution, functional organization, and insertion specificity. *Journal of bacteriology.* 2002;184(4):992-1002.

29. Spencer-Smith R, Varkey EM, Fielder MD, Snyder LA. Sequence features contributing to chromosomal rearrangements in Neisseria gonorrhoeae. *PLoS One.* 2012;7(9): e46023.
30. Song H, Hwang J, Yi H, Ulrich RL, Yu Y, Nierman WC, Kim HS. The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathog.* 2010;6(5): e1000922.
31. Haack KR, Roth JR. Recombination between chromosomal IS200 elements supports frequent duplication formation in *Salmonella typhimurium*. *Genetics.* 1995;141(4):1245-1252.
32. Daveran-Mingot ML, Campo N, Ritzenthaler P, Le Bourgeois P. A Natural Large Chromosomal Inversion in *Lactococcus lactis* Is Mediated by Homologous Recombination between Two Insertion Sequences. *Journal of Bacteriology.* 1998;180(18):4834-4842.
33. Treangen TJ, Abraham AL, Touchon M, Rocha EP. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS microbiology reviews.* 2009;33(3):539-571.
34. Aras RA, Kang J, Tschumi AI, Harasaki Y, Blaser MJ. Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proceedings of the National Academy of Sciences.* 2003;100(23):13579-13584.
35. Bao Z, Stodghill PV, Myers CR, Lam H, Wei HL, Chakravarthy S, Kvitko BH, Collmer A, Cartinhour SW, Schweitzer P, Swingle B. Genomic plasticity enables phenotypic variation of Pseudomonas syringae pv. tomato DC3000. *PloS one.* 2014;9(2): e86628.
36. Rocha EP. DNA repeats lead to the accelerated loss of gene order in bacteria. *TRENDS in Genetics.* 2003;19(11):600-603.
37. Rocha EP, Danchin A, Viari A. Functional and evolutionary roles of long repeats in prokaryotes. *Research in microbiology.* 1999;150(9-10):725-733.
38. Rocha EP, Danchin A, Viari A. Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. *Molecular biology and evolution.* 1999;16(9):1219-1230.
39. Shen P, Huang HV. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics.* 1986;112(3):441-457.
40. Achaz G, Coissac E, Netter P, Rocha EP. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics.* 2003;164(4):1279-1289.

41. Lovett ST. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular microbiology.* 2004;52(5):1243-1253.
42. Hughes D. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome biology.* 2000;1(6):1-8.
43. Roth JR, Benson NI, Galitski TI, Haack KE, Lawrence JG, Miesel LY. Rearrangements of the bacterial chromosome: formation and applications. *Escherichia coli and Salmonella: cellular and molecular biology.* 1996; 2:2256-2276.
44. Sun S. Dynamics and mechanisms of adaptive evolution in bacteria. *Doctoral dissertation, Uppsala University, Department of Medical Biochemistry and Microbiology*. 2012
45. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, Maddison M. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* 2007;3(2): e23.
46. Zivanovic Y, Lopez P, Philippe H, Forterre P. Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic acids research.* 2002;30(9):1902-1910.
47. Juhas M, Van Der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews.* 2009;33(2):376-393.
48. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology.* 2004;2(5):414-424.
49. Piel J, Höfer I, Hui D. Evidence for a symbiosis island involved in horizontal acquisition of pederin biosynthetic capabilities by the bacterial symbiont of *Paederus fuscipes* beetles. *Journal of bacteriology.* 2004;186(5):1280-1286.
50. Yan W, Wei S, Wang Q, Xiao X, Zeng Q, Jiao N, Zhang R. Genome rearrangement shapes Prochlorococcus ecological adaptation. *Applied and environmental microbiology.* 2018;84(17).
51. Lu B, Leong HW. Computational methods for predicting genomic islands in microbial genomes. *Computational and structural biotechnology journal.* 2016; 14:200-206.
52. Bader M, Ohlebusch E. Sorting by weighted reversals, transpositions, and inverted transpositions. *Journal of Computational Biology.* 2007;14(5):615-636.

53. Cui L, Neoh HM, Iwamoto A, Hiramatsu K. Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proceedings of the National Academy of Sciences.* 2012;109(25): E1647-1656.
54. Gaudriault S, Pages S, Lanois A, Laroui C, Teyssier C, Jumas-Bilak E, Givaudan A. Plastic architecture of bacterial genome revealed by comparative genomics of Photorhabdus variants. *Genome biology.* 2008;9(7):1-15.
55. Okinaka RT, Price EP, Wolken SR, Gruendike JM, Chung WK, Pearson T, Xie G, Munk C, Hill KK, Challacombe J, Ivins BE. An attenuated strain of Bacillus anthracis (CDC 684) has a large chromosomal inversion and altered growth kinetics. *BMC genomics.* 2011;12(1):1-13.
56. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics.* 2013;14(2):125-138.
57. Darling AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* 2008;4(7): e1000128.
58. Rocha EP. The organization of the bacterial genome. *Annual review of genetics.* 2008; 42:211-233.
59. Furuta Y, Kawai M, Yahara K, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, Uchiyama I. Birth and death of genes linked to chromosomal inversion. *Proceedings of the National Academy of Sciences.* 2011;108(4):1501-1506.
60. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics.* 2012;13(1):1-13.
61. Sun S, Ke R, Hughes D, Nilsson M, Andersson DI. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PloS one.* 2012;7(8): e42639.
62. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nature Reviews Genetics.* 2011;12(5):363-376.
63. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nature genetics.* 2004;36(9):949-951.
64. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM, Kohn S. High-resolution human genome structure by

single-molecule analysis. *Proceedings of the National Academy of Sciences.* 2010;107(24):10848-10853.

65. Das SK, Austin MD, Akana MC, Deshpande P, Cao H, Xiao M. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic acids research.* 2010;38(18): e177.
66. Beer NR, Hindson BJ, Wheeler EK, Hall SB, Rose KA, Kennedy IM, Colston BW. On-chip, real-time, single-copy polymerase chain reaction in picoliter droplets. *Analytical chemistry.* 2007;79(22):8471-8475.
67. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic acids research.* 2014;42(D1): D553-D559.
68. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research.* 2004;14(7):1394-1403.
69. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research.* 1994;22(22):4673-4680.
70. Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics (Oxford, England).* 1999;15(3):211-218.
71. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology.* 2000;302(1):205-217.
72. Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Reviews Genetics.* 2003;4(4):251-262.
73. T. Jiang, Y. Xu, and M. Q. Zhang. *Current Topics in Computational Molecular Biology.* MIT Press, 2002.
74. Belda E, Moya A, Silva FJ. Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. *Molecular biology and evolution.* 2005;22(6):1456-1467.
75. Li Z, Wang L, Zhang K. Algorithmic approaches for genome rearrangement: a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).* 2006;36(5):636-648.
76. Pevzner P. Computational molecular biology: an algorithmic approach. MIT press; 2000.

77. Hannenhalli S, Pevzner PA. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM (JACM)*. 1999;46(1):1-27.
78. Hannenhalli S, Chappey C, Koonin EV, Pevzner PA. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics*. 1995;30(2):299-311.
79. Boore JL. Comparative genomics: empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families, vol. 1., Computational biology series. *The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals*. 2000; 1:133-147.
80. O'Brien SJ. Genetic maps: locus maps of complex genomes. *Cold Spring Harbor Laboratory Press*; 1993.
81. Hannenhalli S, Chappey C, Koonin E, Pevzner P. Algorithms for genome rearrangements: herpesvirus evolution as a test case. In *Proc. of the 3rd International Conference on Bioinformatics and Complex Genome Analysis*; 1994.
82. Jones NC, Pevzner PA, Pevzner P. An introduction to bioinformatics algorithms. MIT press; 2004 (pp. 127).
83. Sankoff D, Cedergren R, Abel Y. [26] Genomic divergence through gene rearrangement; 1990: 428-438.
84. Sankoff D. Edit distance for genome comparison based on non-local operations. In *Annual Symposium on Combinatorial Pattern Matching*.1992; (pp. 121-135). Springer, Berlin, Heidelberg.
85. Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences*. 1992;89(14):6575-6579.
86. Watterson GA, Ewens WJ, Hall TE, Morgan A. The chromosome inversion problem. *Journal of Theoretical Biology*. 1982;99(1):1-7.
87. Nadeau JH, Taylor BA. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*. 1984;81(3):814-818.
88. Hayes B. Computing Science: Sorting Out the Genome. *American Scientist*. 2007;95(5):386-391.
89. Kececioglu J, Sankoff D. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*. 1995;13(1):180-210.

90. Ghaffarizadeh A, Ahmadi K, Flann NS. Sorting unsigned permutations by reversals using multi-objective evolutionary algorithms with variable size individuals. In *2011 IEEE Congress of Evolutionary Computation (CEC)* 2011; (pp. 292-295). IEEE.
91. Kececioglu J, Sankoff D. Exact and approximation algorithms for the inversion distance between two chromosomes. In *Annual Symposium on Combinatorial Pattern Matching.* 1993; (pp. 87-105). Springer, Berlin, Heidelberg.
92. Hannenhalli S, Pevzner P. To cut… or not to cut (applications of comparative physical maps in molecular evolution). In *Proceedings of the seventh annual ACM-SIAM symposium on Discrete algorithms.* 1996; (pp. 304-313).
93. Caprara A. Sorting by reversals is difficult. In *Proceedings of the first annual international conference on Computational molecular biology*. 1997; (pp. 75-83).
94. V. Bafna and P. A. Pevzner, "Genome rearrangements and sorting by reversals," *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, Palo Alto, CA, USA. 1993; pp. 148-157.
95. Bafna V, Pevzner PA. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing.* 1996;25(2):272-89.
96. Berman P, Hannenhalli S. Fast sorting by reversal. In *Annual Symposium on Combinatorial Pattern Matching*.1996;(pp. 168-185). Springer, Berlin, Heidelberg.
97. Kaplan H, Shamir R, Tarjan RE. A faster and simpler algorithm for sorting signed permutations by reversals. *In Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms.* 1997;344-351.
98. Bourque G, Pevzner PA. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome research*. 2002;12(1):26-36.
99. Sankoff D, Blanchette M. Multiple genome rearrangement and breakpoint phylogeny. *Journal of computational biology*. 1998;5(3):555-570.
100. Caprara A. Formulations and hardness of multiple sorting by reversals. In *Proceedings of the third annual international conference on Computational molecular biology*.1999; (pp. 84-93).
101. Blanchette M, Bourque G, Sankoff D. Breakpoint phylogenies. *Genome informatics.* 1997; 8:25-34.
102. Sankoff D, Blanchette M. The median problem for breakpoints in comparative genomics. In *International Computing and Combinatorics Conference*. 1997; (pp. 251-263). Springer, Berlin, Heidelberg.

103. Arjona-Medina JA. Algorithms and methods for large-scale genome rearrangements identification. 2017. https://www.bioinf.jku.at/people/arjona/tesis/thesis.pdf. Accessed 17 May 2019.
104. Srivatsan A, Han Y, Peng J, Tehranchi AK, Gibbs R, Wang JD, Chen R. High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet*. 2008;4(8): e1000139.
105. Fitch WM. Homology: a personal view on some of the problems. *Trends in genetics.* 2000;16(5):227-231.
106. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research.* 2003;13(9):2178-2189.
107. Tada I, Tanizawa Y, Arita M. Visualization of consensus genome structure without using a reference genome. *BMC genomics.* 2017;18(2):1-9.
108. Kalali B, Mejías-Luque R, Javaheri A, Gerhard M. *H. pylori* virulence factors: influence on immune system and pathology. *Mediators of inflammation*. 2014; doi: http://dx.doi.org/10.1155/2014/426309.
109. Dong QJ, Wang Q, Xin YN, Li N, Xuan SY. Comparative genomics of *Helicobacter pylori*. *World journal of gastroenterology*: WJG. 2009;15(32):3984.
110. Roesler BM, Rabelo-Gonçalves EM, Zeitune JM. Virulence factors of *Helicobacter pylori*: a review. *Clinical Medicine Insights: Gastroenterology.* 2014;7: CGast-S13760.
111. Goodwin CS, Armstrong JA. Microbiological aspects of *Helicobacter pylori* (Campylobacter pylori). *European Journal of Clinical Microbiology and Infectious Diseases.* 1990;9(1):1-13.
112. Ahmed N. 23 years of the discovery of *Helicobacter pylori*: Is the debate over? *Annals of Clinical Microbiology and Antimicrobials*. 2005;1-3.
113. Blaser MJ, Atherton JC. *Helicobacter pylori* persistence: biology and disease. *The Journal of clinical investigation.* 2004;113(3):321-333.
114. Lehours P, Yilmaz O. Epidemiology of *Helicobacter pylori* infection. *Helicobacter.* 2007; 12:1-3.
115. Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, Schlebusch CM, Bernhöft S, Hale J, Suerbaum S, Mugisha L, Van der Merwe SW. Age of the association between *Helicobacter pylori* and man. *PLoS pathog*. 2012;8(5):e1002693.
116. Kusters, J.G.; Van Vliet, A.H.; Kuipers, E.J. Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol. Rev.* 2006; 19,449–490.

117. Moodley, Y.; Linz, B.; Yamaoka, Y.; Windsor, H.M.; Breurec, S.; Wu, J.Y.; Maady, A.; Bernhöft, S.; Thiberge, J.M.; Phuanukoonnon, S.; et al. The peopling of the Pacific from a bacterial perspective. *Science*. 2009;323,527–530.
118. Linz, B.; Balloux, F.; Moodley, Y.; Manica, A.; Liu, H.; Roumagnac, P.; Falush, D.; Stamer, C.; Prugnolle, F.; van der Merwe, S.W.; et al. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*. 2007;445, 915–918.
119. Suerbaum S, Michetti P. *Helicobacter pylori* infection. *New England Journal of Medicine*. 2002;347(15):1175-1186.
120. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*. 1997;388(6642):539-547.
121. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, Dejonge BL, Carmel G. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*. 1999;397(6715):176-180.
122. Oh JD, Kling-Bäckhed H, Giannakis M, Xu J, Fulton RS, Fulton LA, Cordum HS, Wang C, Elliott G, Edwards J, Mardis ER. The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proceedings of the National Academy of Sciences*. 2006;103(26):9999-10004.
123. Baltrus DA, Amieva MR, Covacci A, Lowe TM, Merrell DS, Ottemann KM, Stein M, Salama NR, Guillemin K. The complete genome sequence of *Helicobacter pylori* strain G27. *Journal of bacteriology*. 2009;191(1):447-448.
124. Taneera J, Moran AP, Hynes SO, Nilsson HO, abu Al-Soud W, Wadström T. Influence of activated charcoal, porcine gastric mucin and β-cyclodextrin on the morphology and growth of intestinal and gastric Helicobacter spp. *Microbiology*. 2002;148(3):677-684.
125. Gu H. Role of Flagella in the Pathogenesis of *Helicobacter pylori*. *Current microbiology*. 2017;74(7):863-869.
126. Suerbaum S, Josenhans C. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nature reviews microbiology*. 2007;5(6):441-452.
127. Humbert O, Dorer MS, Salama NR. Characterization of *Helicobacter pylori* factors that control transformation frequency and integration length during inter-strain DNA recombination. *Molecular microbiology*. 2011;79(2):387-401.
128. Fischer W, Windhager L, Rohrer S, Zeiller M, Karnholz A, Hoffmann R, Zimmer R, Haas R. Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a

novel type IV secretion system and genomic island transfer. *Nucleic acids research*.2010;38(18):6089-6101.

129. Farnbacher M, Jahns T, Willrodt D, Daniel R, Haas R, Goesmann A, Kurtz S, Rieder G. Sequencing, annotation, and comparative genome analysis of the gerbil-adapted *Helicobacter pylori* strain B8. *BMC genomics*. 2010;11(1):1-22.
130. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, Suerbaum S, Achtman M. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet*. 2010;6(7): e1001036.
131. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D. *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences*. 2011;108(12):5033-5038.
132. Grigoriev A. Graphical genome comparison: rearrangements and replication origin of *Helicobacter pylori*. *Trends in Genetics*. 2000;16(9):376-378.
133. Kojima KK, Furuta Y, Yahara K, Fukuyo M, Shiwa Y, Nishiumi S, Yoshida M, Azuma T, Yoshikawa H, Kobayashi I. Population evolution of *Helicobacter pylori* through diversification in DNA methylation and interstrain sequence homogenization. *Molecular biology and evolution*. 2016;33(11):2848-2859.
134. Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, Handa N, Takahashi N, Yoshida M, Azuma T, Hattori M, Uchiyama I. Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC microbiology*. 2011;11(1):1-28.
135. Baltrus DA, Guillemin K, Phillips PC. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. *Evolution: International Journal of Organic Evolution*. 2008;62(1):39-49.
136. Baltrus DA, Blaser MJ, Guillemin K. *Helicobacter pylori* genome plasticity. *In Microbial Pathogenomics*. 2009 (Vol. 6, pp. 75-90). Karger Publishers.
137. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*. 2002;30(14):3059-3066.
138. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972-1973.

139. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-1313.
140. Kobayashi I. Genome Evolution: *Helicobacter pylori* as an extreme model. *In Helicobacter pylori Research*. 2016 (pp. 217-231). Springer, Tokyo.
141. Vale FF, Nunes A, Oleastro M, Gomes JP, Sampaio DA, Rocha R, Vítor JM, Engstrand L, Pascoe B, Berthenet E, Sheppard SK. Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. *Scientific reports*. 2017;7(1):1-12.
142. Lehours P, Vale FF, Bjursell MK, Melefors O, Advani R, Glavas S, Guegueniat J, Gontier E, Lacomme S, Matos AA, Menard A. Genome sequencing reveals a phage in *Helicobacter pylori. MBio*. 2011;2(6).
143. Kobayashi I. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic acids research*. 2001;29(18):3742-3756.
144. Ooka T, Ogura Y, Asadulghani M, Ohnishi M, Nakayama K, Terajima J, Watanabe H, Hayashi T. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. *Genome research*. 2009;19(10):1809-1816.
145. Rajaraman A, Tannier E, Chauve C. FPSAC: fast phylogenetic scaffolding of ancient contigs. *Bioinformatics.* 2013;29(23):2987-2994.
146. Wang D, Li S, Guo F, Ning K, Wang L. Core-genome scaffold comparison reveals the prevalence that inversion events are associated with pairs of inverted repeats. *BMC genomics*. 2017;18(1):1-13.
147. Kersulyte D, Akopyants NS, Clifton SW, Roe BA, Berg DE. Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of *Helicobacter pylori*. *Gene*. 1998;223(1-2):175-186.
148. Kersulyte D, Mukhopadhyay AK, Shirai M, Nakazawa T, Berg DE. Functional organization and insertion specificity of IS607, a chimeric element of *Helicobacter pylori*. *Journal of bacteriology*. 2000;182(19):5300-5308.
149. Kersulyte D, Kalia A, Zhang M, Lee HK, Subramaniam D, Kiuduliene L, Chalkauskas H, Berg DE. Sequence organization and insertion specificity of the novel chimeric ISHp609 transposable element of *Helicobacter pylori*. *Journal of bacteriology*. 2004;186(22):7521-7528.
150. Censini S, Lange C, Xiang Z, Crabtree JE, Ghiara P, Borodovsky M, Rappuoli R, Covacci A. cag, a pathogenicity island of *Helicobacter pylori,* encodes type I-specific

and disease-associated virulence factors. *Proceedings of the National Academy of Sciences*. 1996;93(25):14648-14653.

151. Okonechnikov K, Golosova O, Fursov M, Ugene Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. 2012;28(8):1166-1167.
152. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, Winsor GL, Brinkman FS. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic acids research*. 2017;45(W1): W30-35.
153. Romo-González C, Salama NR, Burgeno-Ferreira J, Ponce-Castaneda V, Lazcano-Ponce E, Camorlinga-Ponce M, Torres J. Differences in genome content among *Helicobacter pylori* isolates from patients with gastritis, duodenal ulcer, or gastric cancer reveal novel disease-associated genes. *Infection and immunity*. 2009;77(5):2201-2211
154. Crowe SE. *Helicobacter pylori* infection. *New England Journal of Medicine*. 2019;380(12):1158-1165.
155. International Agency for Research on Cancer (IARC). Schistosomes, Liver Flukes and *Helicobacter Pylori*, Monograph on the Evaluation of Carcinogenic Risks to Humans; IARC: Lyon, France, 1994; Volume 61.
156. Bauer B, Meyer TF. The human gastric pathogen *Helicobacter pylori* and its association with gastric cancer and ulcer disease. *Ulcers*. 2011.
157. Shiota S, Suzuki R, Matsuo Y, Miftahussurur M, Tran TT, Binh TT, Yamaoka Y. *Helicobacter pylori* from gastric cancer and duodenal ulcer show same phylogeographic origin in the Andean region in Colombia. *PloS one*. 2014;9(8):e105392.
158. Noto JM, Chopra A, Loh JT, Romero-Gallo J, Piazuelo MB, Watson M, Leary S, Beckett AC, Wilson KT, Cover TL, Mallal S. Pan-genomic analyses identify key *Helicobacter pylori* pathogenic loci modified by carcinogenic host microenvironments. Gut. 2018 Oct 1;67(10):1793-1804
159. Yang F, Zhang J, Wang S, Sun Z, Zhou J, Li F, Liu Y, Ding L, Liu Y, Chi W, Liu T, He Y, Xiang P, Bao Z, Olszewski MA, Zhao H, Zhang Y. Genomic population structure of *Helicobacter pylori* Shanghai isolates and identification of genomic features uniquely linked with pathogenicity. *Virulence*. 2021;12(1):1258-1270.
160. Yamaoka Y. Mechanisms of disease: *Helicobacter pylori* virulence factors. *Nature reviews Gastroenterology & hepatology*. 2010;7(11):629.
161. Shiota S, Suzuki R, Yamaoka Y. The significance of virulence factors in *Helicobacter pylori*. *Journal of digestive diseases*. 2013;14(7):341-349.

162. Yamaoka Y, Ojo O, Fujimoto S, Odenbreit S, Haas R, Gutierrez O, El-Zimaity HM, Reddy R, Arnqvist A, Graham DY. *Helicobacter pylori* outer membrane proteins and gastroduodenal disease. *Gut*. 2006;55(6):775-781.
163. Hathroubi S, Zerebinski J, Ottemann KM. *Helicobacter pylori* biofilm involves a multigene stress-biased response, including a structural role for flagella. *MBio*. 2018;9(5).
164. Matsunari O, Shiota S, Suzuki R, Watada M, Kinjo N, Murakami K, Fujioka T, Kinjo F, Yamaoka Y. Association between *Helicobacter pylori* virulence factors and gastroduodenal diseases in Okinawa, Japan. *Journal of clinical microbiology*. 2012;50(3):876-883.
165. Covacci A, Censini S, Bugnoli M, Petracca R, Burroni D, Macchia G, Massone A, Papini E, Xiang Z, Figura N. Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proceedings of the National Academy of Sciences.* 1993;90(12):5791-5795.
166. Olbermann P, Josenhans C, Moodley Y, Uhr M, Stamer C, Vauterin M, Suerbaum S, Achtman M, Linz B. A global overview of the genetic and functional diversity in the *Helicobacter pylori* cag pathogenicity island. *PLoS Genet*. 2010;6(8): e1001069.
167. Testerman TL, Morris J. Beyond the stomach: an updated view of *Helicobacter pylori* pathogenesis, diagnosis, and treatment. *World journal of gastroenterology: WJG*. 2014;20(36):12781.
168. Wroblenski LE, Peek RM Jr, Wilson KT. *Helicobacter pylori* and gastric cancer: factors that modulate disease risk. *Clin Microbiol Rev*. 2010;23(4):713–739.
169. Cover TL. *Helicobacter pylori* diversity and gastric cancer risk. *MBio*. 2016;7(1).
170. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Appl. Environ. Microbiol. 2013; 79, 7696–7701
171. Noureen M, Tada I, Kawashima T, Arita M. Rearrangement analysis of multiple bacterial genomes. *BMC bioinformatics*. 2019; 20(23):1-0.
172. Noureen M, Kawashima T, Arita M. Genetic Markers of Genome Rearrangements in *Helicobacter pylori*. *Microorganisms*. 2021; 9(3):621.
173. Lin YC, Lu CL, Liu YC, Tang CY. SPRING: a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic acids research*. 2006; W696-699.

174. Tesler G. GRIMM: genome rearrangements web server. *Bioinformatics*. 2002; 18(3):492-493.

Appendix

Table A.1: 72 *Helicobacter pylori* strains information [171].

Accession	Strain	Disease	Country	Geographical Region	Gene	Protein
NZ_AP014523.1	NY40	-	Japan	East Asia	1693	1479
NC_017365.1	F30	Duodenal ulcer	Japan	East Asia	1569	1427
NZ_AP014712.1	ML3	MALT lymphoma	Taiwan	East Asia	1617	1365
NZ_AP014710.1	ML1	MALT lymphoma	Taiwan	East Asia	1610	1385
NC_021216.3	UM299	-	Singapore	East Asia	1573	1441
NC_021882.2	UM298	-	Singapore	East Asia	1575	1442
NC_021215.3	UM032	Peptic ulcer disease	Malaysia	East Asia	1572	1441
NC_021217.3	UM037	-	Malaysia	East Asia	1675	1507
NC_021218.3	UM066	Peptic ulcer disease	Malaysia	East Asia	1614	1471
NC_017366.1	F32	Gastric cancer	Japan	East Asia	1570	1434
NZ_CP006822.1	oki128	Gastric atrophy	Japan	East Asia	1540	1357
NC_017926.1	XZ274	Gastric cancer	China	East Asia	1682	1438
NC_020509.1	OK310	-	Japan	East Asia	1575	1443
NC_017354.1	52	-	Korea	East Asia	1549	1390
NC_017368.1	F16	Gastritis	Japan	East Asia	1565	1409
NZ_CP006825.1	oki673	Gastric ulcer	Japan	East Asia	1573	1390
NZ_CP006823.1	oki154	Duodenal ulcer	Japan	East Asia	1582	1405
NZ_CP006826.1	oki828	Duodenal ulcer	Japan	East Asia	1586	1396
NZ_CP006827.1	oki898	Duodenal ulcer	Japan	East Asia	1599	1461
NZ_CP006821.1	oki112	Gastric atrophy	Japan	East Asia	1602	1451
NZ_CP006820.1	oki102	Gastric atrophy	Japan	East Asia	1594	1456
NZ_CP006824.1	oki422	Gastric atrophy	Japan	East Asia	1594	1430
NC_017367.1	F57	Gastric cancer	Japan	East Asia	1584	1439
NZ_AP013356.1	26695-1CL	-	Japan	East Asia	1643	1508
NZ_AP013355.1	26695-1CH	-	Japan	East Asia	1645	1510
NZ_AP013354.1	26695-1	-	Japan	East Asia	1644	1510
NZ_CP010013.1	Hp238	MALT lymphoma	Taiwan	East Asia	1569	1410
NC_020508.1	OK113	-	Japan	East Asia	1575	1443
NC_017741.1	Shi112	-	Peru	South America	1635	1493
NC_017359.1	Sat464	-	Peru	South America	1553	1419
NC_017358.1	Cuz20	-	Peru	South America	1616	1475

NC_017739.1	Shi417	-	Peru	South America	1622	1490
NC_014555.1	PeCan4	Gastric cancer	Peru	South America	1611	1462
NC_017742.1	PeCan18	Gastric cancer	Peru	South America	1622	1467
NC_017378.1	Puno120	-	Peru	South America	1585	1444
NC_017740.1	Shi169	-	Peru	South America	1595	1456
NC_014560.1	SJM180	Gastritis	Peru	South America	1617	1478
NC_017379.1	Puno135	-	Peru	South America	1606	1481
NC_017355.1	v225d	Gastritis	Venezuela	South America	1574	1430
NZ_CP012905.1	7C	Cancer	Mexico	North America	1593	1427
NZ_CP012907.1	29CaP	Gastric cancer	Mexico	North America	1656	1443
NC_019560.1	Aklavik117	-	Canada	North America	1590	1459
NZ_CP010435.1	26695-1	-	USA (Texas)	North America	1644	1510
NZ_CP010436.1	26695-1MET	-	USA (Texas)	North America	1645	1510
NZ_CP007603.1	J166	-	Nashville	North America	1608	1459
NZ_CP011330.1	J99	Duodenal ulcer	USA (Nashville)	North America	1645	1488
NC_017063.1	ELS37	Cancer	El Salvador	North America	1643	1485
NC_012973.1	B38	MALT lymphoma	France	Europe	1565	1414
NC_017733.1	HUP-B14	-	Spain	Europe	1574	1429
NC_018938.1	Rif2	-	-	Europe	1644	1506
NC_018937.1	Rif1	-	-	Europe	1643	1500
NC_018939.1	26695	Gastritis	-	Europe	1645	1507
NC_014256.1	B8	Gastric ulcer	USA	Europe	1637	1488
NC_011498.1	P12	Duodenal ulcer	Germany	Europe	1650	1468
NC_000915.1	26695	-	UK	Europe	1555	1445
NC_011333.1	G27	-	Italy	Europe	1619	1468
NC_017362.1	Lithuania75	-	Lithuania	Europe	1619	1445
NC_017374.1	2017	Duodenal ulcer	France	Europe	1548	1377
NC_017357.1	908	Duodenal ulcer	France	Europe	1548	1374
NC_017381.1	2018	Duodenal ulcer	France	Europe	1557	1387
NC_017361.1	SouthAfrica7	-	South Africa	Africa	1619	1461
NC_022130.1	SouthAfrica20	-	South Africa	Africa	1568	1343
NC_017371.1	Gambia94/24	-	Gambia	Africa	1682	1528
NC_017372.1	India7	-	India	India	1629	1470
NC_017376.1	Santal49	-	India	India	1579	1427
NZ_CP007605.1	BM012B	-	Australia	Australia	1661	1492
NC_022886.1	BM012A	-	Australia	Australia	1663	1493
NC_022911.1	BM012S	-	Australia	Australia	1662	1488
NZ_CP007606.1	BM013B	-	Australia	Australia	1573	1435
NZ_CP007604.1	BM013A	-	Australia	Australia	1571	1435
NC_017375.1	83	-	-	-	1599	1432
NC_017360.1	35A	-	-	-	1566	1411

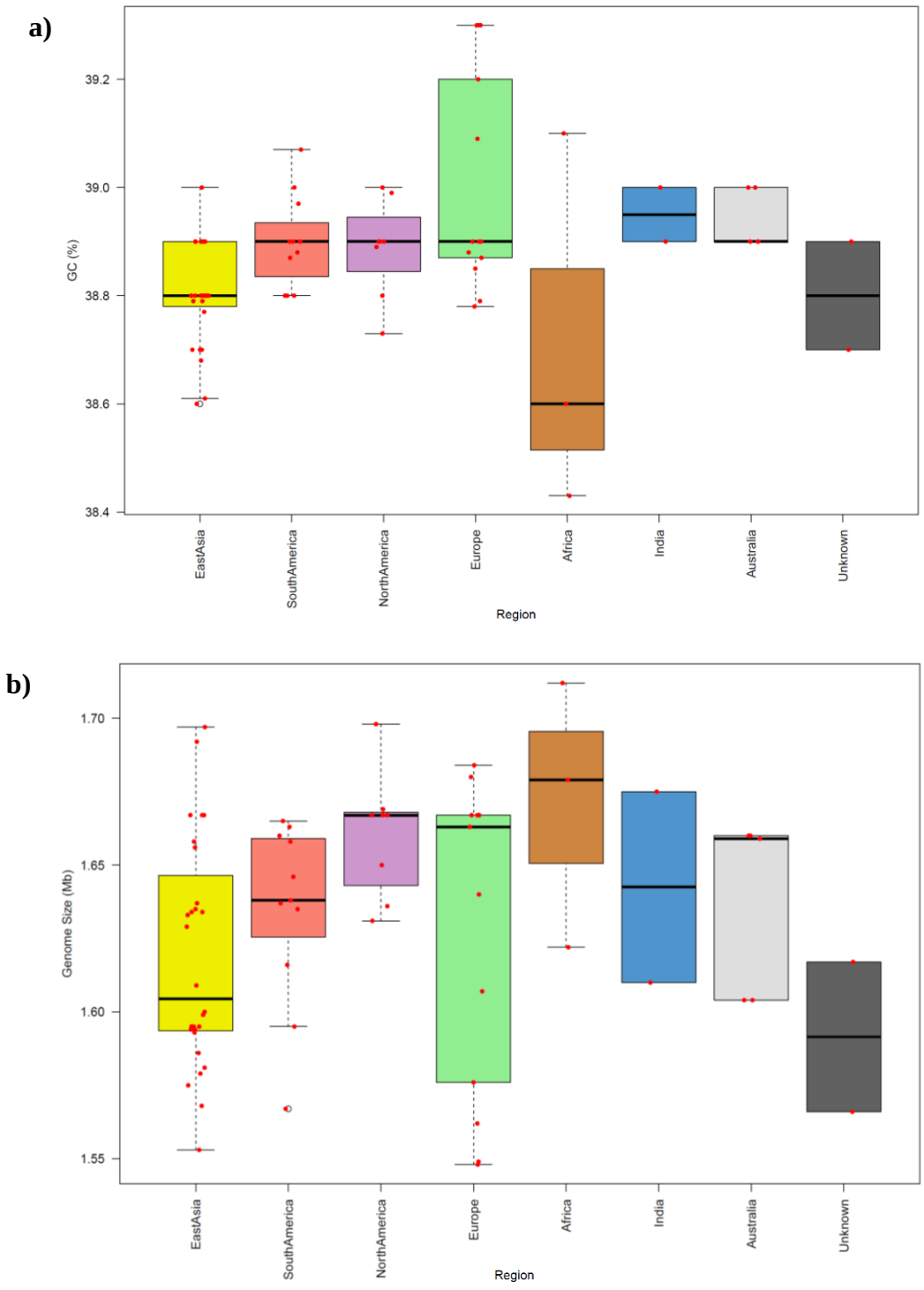
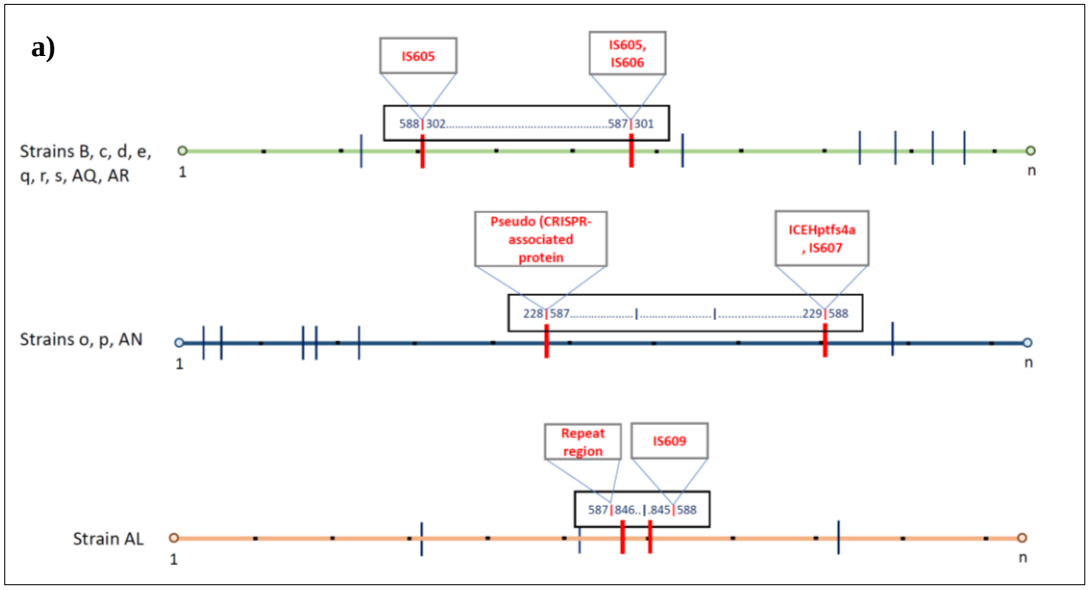
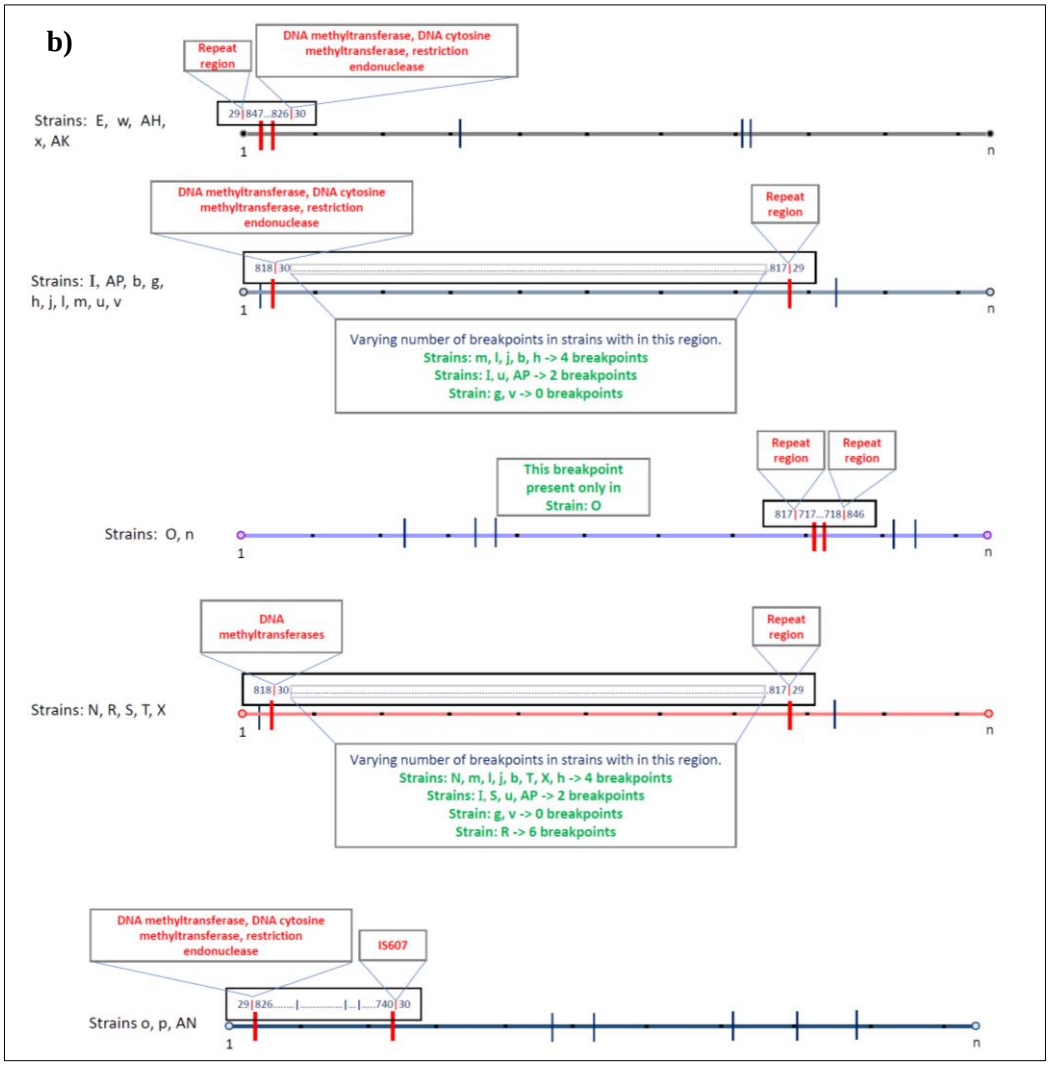


Figure A.1: a) GC content in strains from different geographical locations. **b)** Genome size variation in strains from different geographical locations.


a)
IS605
IS605, IS606
588|302
587|301
Strains B, c, d, e, q, r, s, AQ, AR
1
n
Pseudo (CRISPR-associated protein
ICEHptfs4a , IS607
228|587
229|588
Strains o, p, AN
1
n
Repeat region
IS609
587|846..|845|588
Strain AL
1
n


b)
Repeat region
DNA methyltransferase, DNA cytosine methyltransferase, restriction endonuclease
29|847..826|30
Strains: E, w, AH, x, AK
1
n
DNA methyltransferase, DNA cytosine methyltransferase, restriction endonuclease
Repeat region
818|30
817|29
Strains: I, AP, b, g, h, j, l, m, u, v
1
n
Varying number of breakpoints in strains with in this region.
Strains: m, l, j, b, h -> 4 breakpoints
Strains: I, u, AP -> 2 breakpoints
Strain: g, v -> 0 breakpoints
This breakpoint present only in Strain: O
Repeat region
Repeat region
817|717...718|846
Strains: O, n
1
n
DNA methyltransferases
Repeat region
818|30
817|29
Strains: N, R, S, T, X
1
n
Varying number of breakpoints in strains with in this region.
Strains: N, m, l, j, b, T, X, h -> 4 breakpoints
Strains: I, S, u, AP -> 2 breakpoints
Strain: g, v -> 0 breakpoints
Strain: R -> 6 breakpoints
DNA methyltransferase, DNA cytosine methyltransferase, restriction endonuclease
IS607
29|826.....|...|...|740|30
Strains o, p, AN
1
n

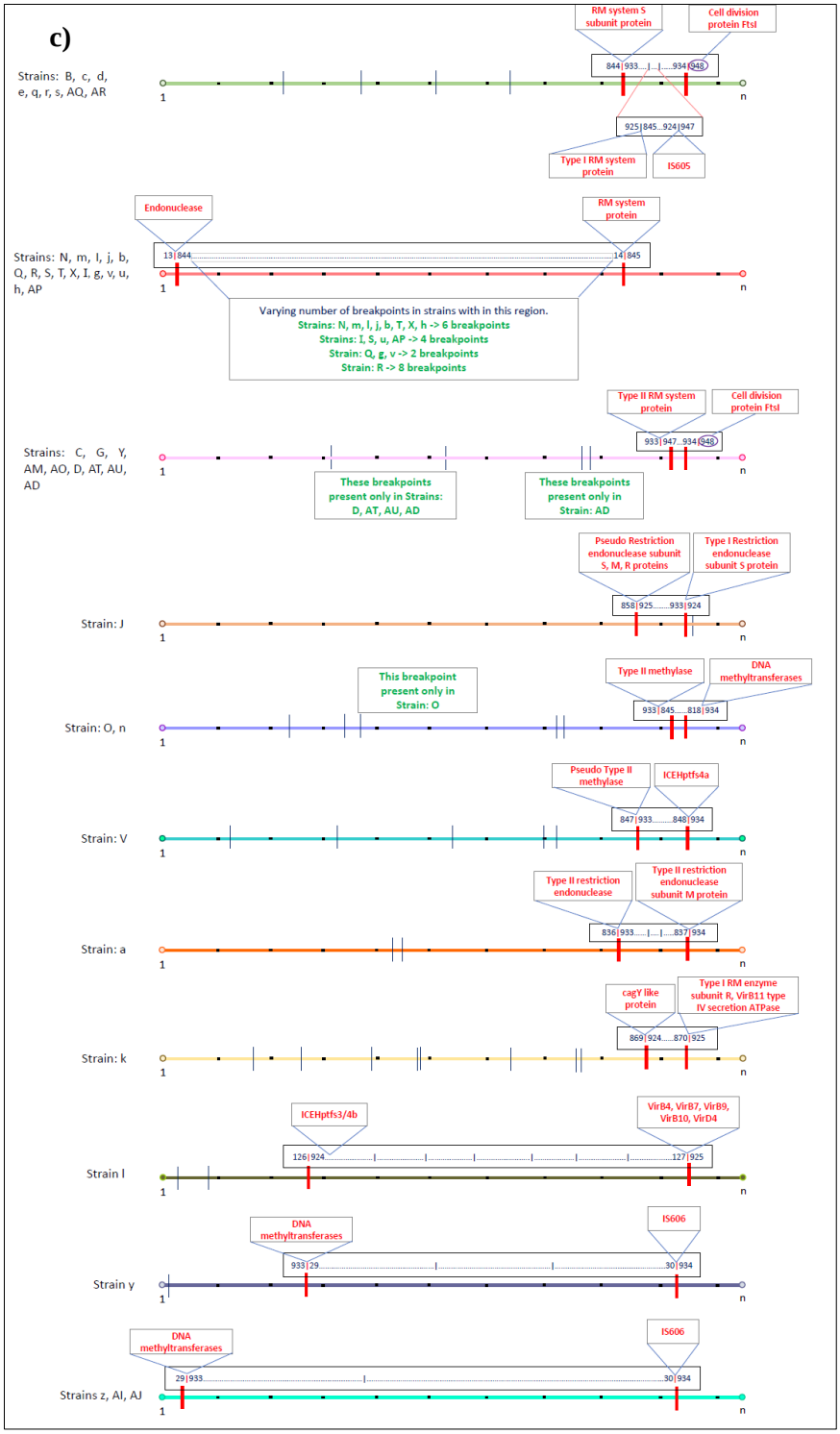


Figure A.2: Rearrangement hotspot. The breakpoints [588, 587] in **a**, [(844, 845), (933, 934), (924, 925)] in **b** and [(29, 30), (818)] in **c** reflect the region involved in different rearrangements. The blue vertical lines indicate the other breakpoints present in each strain. The red vertical lines indicate the region (breakpoint) called the hotspot. The boxes show the different elements present around these breakpoints [171].

Table A.2: Matrix representing the presence and absence of all the identified inversions. First column has the labels assigned to strains. Second column has the corresponding strain names. Column 3 to column 43 represent the inversions labeled as R1 to R41. The values in the cell of these columns represent the presence of inversion as 1 and absence as 0. Presence of all inversions is given a different color [171].

Label	Strain	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30	R31	R32	R33	R34	R35	R36	R37	R38	R39	R40	R41
A	P12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	26695	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	G27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	B38	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	B8	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	PeCan4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	SJM180	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	ELS37	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
I	52	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	v225d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
K	908	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	Cuz20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	Sat464	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	35A	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	SouthAfrica7	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	Lithuania75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	F30	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	F32	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
S	F57	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	F16	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	Gambia94/24	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	India7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
W	2017	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	83	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	SNT49	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Z	Puno120	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AB	Puno135	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AC	2018	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AD	HUP-B14	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AE	Shi417	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AF	Shi169	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AG	Shi112	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	PeCan18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
b	XZ274	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	Rif1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	Rif2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	26695	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	Aklavik117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	OK113	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	OK310	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	UM032	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
j	UM299	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	UM037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
l	UM066	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
m	UM298	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	SouthAfrica20	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	BM012A	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p	BM012S	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q	26695-1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	26695-1CH	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s	26695-1CL	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	NY40	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
u	ML1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v	ML3	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
w	oki102	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	oki112	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y	oki128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
z	oki154	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AH	oki422	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AI	oki673	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AJ	oki828	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AK	oki898	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AL	J166	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AM	BM013A	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AN	BM012B	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AO	BM013B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AP	Hp238	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AQ	26695-1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AR	26695-1MET	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AS	J99	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AT	7C	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AU	29CaP	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.3: Inversions along with their corresponding breakpoints [172].

Inversion	Corresponding Breakpoints (BrPs)	BrPs Labels
R1	236|632 237|633	B1 B2
R2	588|302 587|301	B3 B4
R3	924|947 OR 933|947 934|948	B5 or B6 B7
R4	924|947 925|845	B5 B8
R5	925|845 OR 933|845 844|932 OR 836|933	B8 or B31 B9 or B10
R6	312|530 313|531	B11 B12
R7	826|30 825|848	B13 B14
R8	826|30 29|847	B13 B15
R9	390|428 391|429	B16 B17
R10	416|420 417|421	B18 B19
R11	854|858 855|859	B20 B21
R12	818|30 817|29	B22 B23
R13	13|844 14|845	B24 B25
R14	49|75 48|74	B26 B27
R15	328|324 325|329 OR 560|325	B28 B32 or B45
R16	765|773 766|774	B29 B30
R17	228|587 229|588	B33 B34
R18	669|859 670|858	B35 B36
R19	29|826 740|30	B37 B38
R20	740|30 848|825	B38 B39
R21	740|30 847|739	B38 B40
R22	933|29 30|934	B41 B42

R23	817|717 818|934	B43 B44
R24	587|846 847|826	B46 B47
R25	847|826 845|588	B47 B48
R26	847|826 825|848	B47 B14
R27	587|846 845|588	B46 B48
R28	111|669 112|670	B49 B50
R29	858|925 933|924	B51 B52
R30	212|619 213|620	B53 B54
R31	111|134 112|700	B55 B56
R32	111|134 699|133	B55 B57
R33	847|933 848|934	B58 B59
R34	836|933 837|934	B10 B60
R35	173|646 172|645	B61 B62
R36	869|924 870|925	B63 B64
R37	561|468 329|469	B65 B66
R38	560|325 329|469	B45 B66
R39	126|924 127|925	B67 B68
R40	4|742 5|741	B69 B70
R41	933|924 859|934	B52 B71

*	740|57, 931|845	B72, B75
*	238|57	B73
**	718|846	B74

*	Single gene transposition
**	2 gene inverse transposition

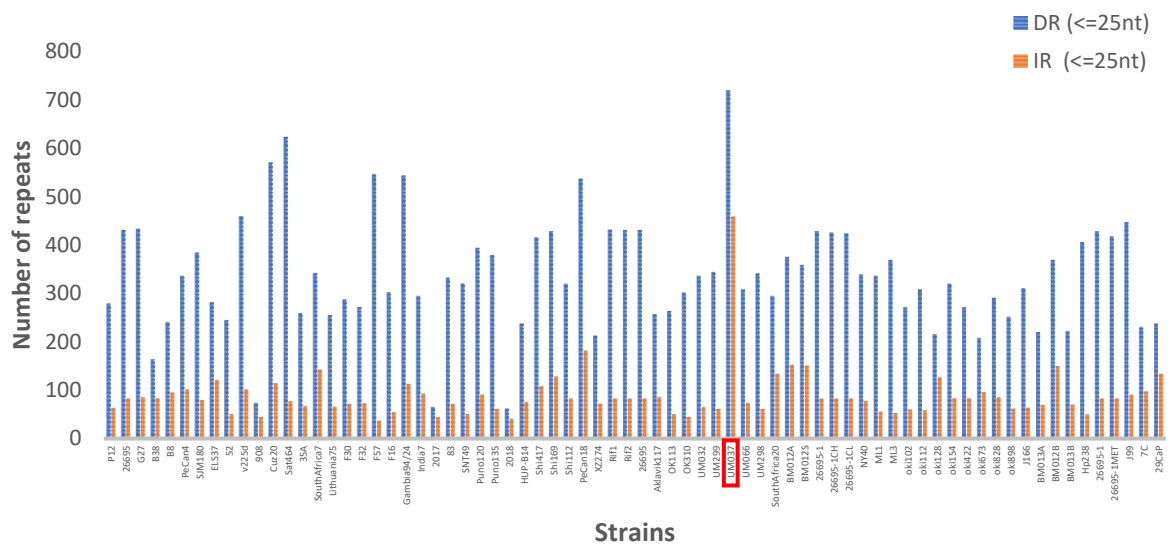


Figure A.3: Occurrence of the direct and inverted repeats in each strain. Direct and inverted repeats of length >=25 nucleotide and 100 % sequence identity are shown. Strain UM037 had the highest number of direct as well as the inverted repeats among all of the analyzed strains [172].

Table A.4: Number of the identified direct and inverted repeats in 72 *Helicobacter pylori* strains. The longest direct and inverted repeats identified in each strain are mentioned [172].

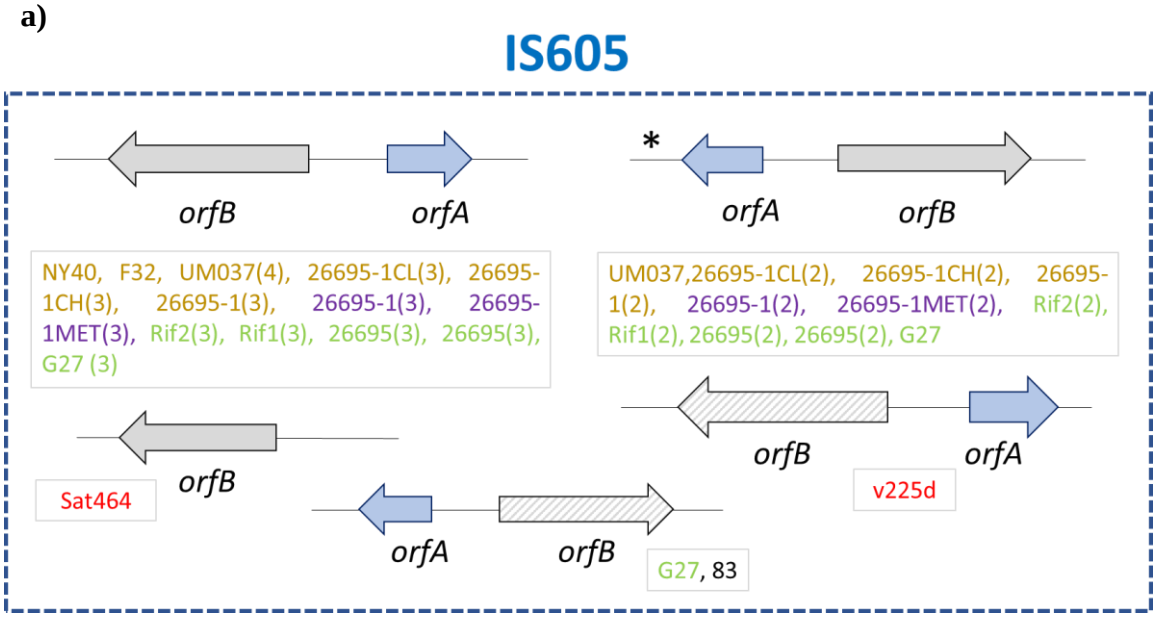
Accession	Strains	Number of inverted repeats	Maximum repeat length (bp)	Number of direct repeats	Maximum repeat length (bp)
NC_011498.1	P12	63	3331	279	3787
NC_000915.1	26695	83	3731	431	1890
NC_011333.1	G27	86	1953	434	4037
NC_012973.1	B38	84	2300	164	4134
NC_014256.1	B8	95	3237	241	3237
NC_014555.1	PeCan4	102	635	337	3329
NC_014560.1	SJM180	79	3113	384	2722
NC_017063.1	ELS37	121	1976	282	2851
NC_017354.1	52	51	2099	245	4254
NC_017355.1	v225d	102	2112	460	6183

NC_017357.1	908	45	122	74	182
NC_017358.1	Cuz20	114	490	571	4411
NC_017359.1	Sat464	77	786	624	4580
NC_017360.1	35A	67	1481	260	3618
NC_017361.1	SouthAfrica7	143	8041	342	3886
NC_017362.1	Lithuania75	66	2465	255	4901
NC_017365.1	F30	71	7642	288	755
NC_017366.1	F32	73	3984	272	2100
NC_017367.1	F57	37	1445	547	8033
NC_017368.1	F16	55	1330	303	10305
NC_017371.1	Gambia94/24	113	1644	544	8138
NC_017372.1	India7	93	2275	295	4004
NC_017374.1	2017	44	58	65	119
NC_017375.1	83	71	1232	333	2690
NC_017376.1	SNT49	51	1268	321	4163
NC_017378.1	Puno120	91	2121	394	7099
NC_017379.1	Puno135	61	269	380	4480
NC_017381.1	2018	41	54	62	115
NC_017733.1	HUP-B14	76	1042	238	7752
NC_017739.1	Shi417	108	981	416	6091
NC_017740.1	Shi169	128	2033	429	4603
NC_017741.1	Shi112	83	2032	320	3815
NC_017742.1	PeCan18	182	3371	537	2127
NC_017926.1	XZ274	72	1001	213	4834
NC_018937.1	Rif1	83	3731	432	1890
NC_018938.1	Rif2	83	3731	431	1890
NC_018939.1	26695	83	3731	431	1890
NC_019560.1	Aklavik117	86	1466	257	4181
NC_020508.1	OK113	51	2218	264	4171
NC_020509.1	OK310	45	1548	302	4013
NC_021215.3	UM032	65	1237	337	3819
NC_021216.3	UM299	61	1158	344	4002
NC_021217.3	UM037	459	2379	720	3364
NC_021218.3	UM066	74	7379	308	4042

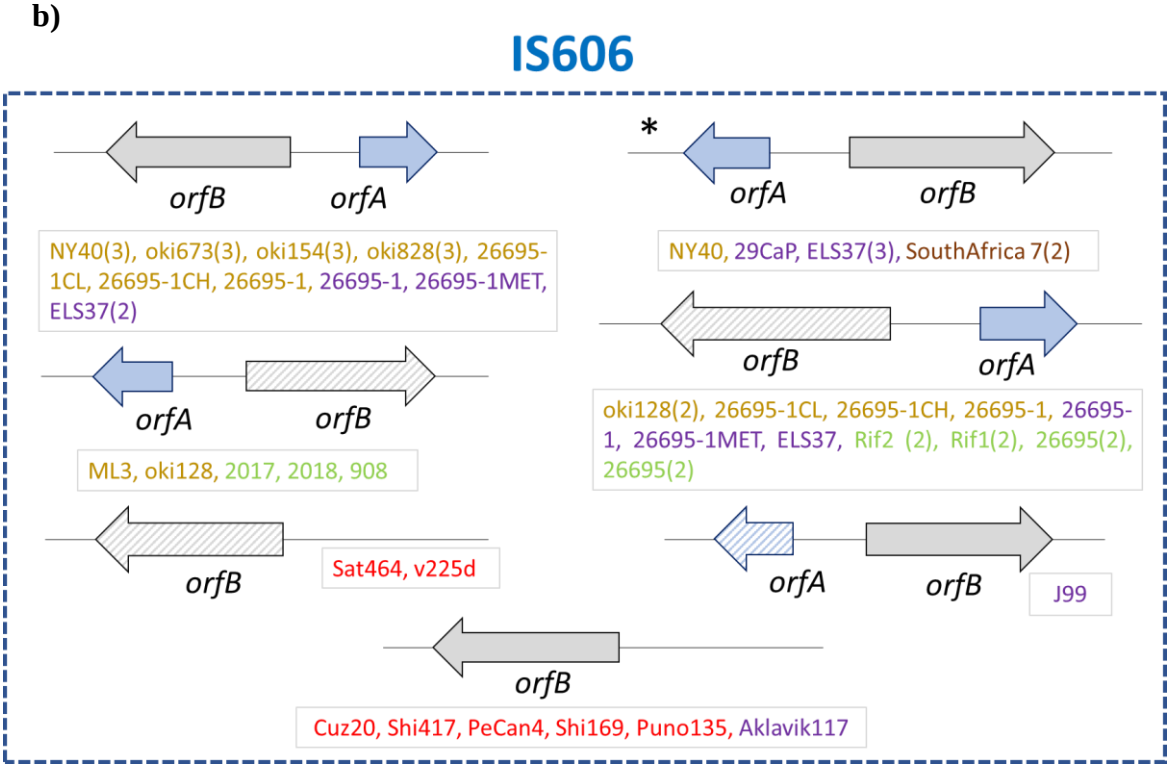
Continued Table A.4

NC_021882.2	UM298	61	1158	341	3819
NC_022130.1	SouthAfrica20	134	4076	295	3148
NC_022886.1	BM012A	153	3352	376	2033
NC_022911.1	BM012S	151	3352	359	2035
NZ_AP013354.1	26695-1	83	3731	429	1890
NZ_AP013355.1	26695-1CH	83	3731	426	1890
NZ_AP013356.1	26695-1CL	83	3731	425	1890
NZ_AP014523.1	NY40	78	1975	339	2299
NZ_AP014710.1	ML1	56	1057	337	1536
NZ_AP014712.1	ML3	53	2105	370	1272
NZ_CP006820.1	oki102	60	73	271	2230
NZ_CP006821.1	oki112	59	1006	308	2498
NZ_CP006822.1	oki128	127	3986	216	1912
NZ_CP006823.1	oki154	84	198	320	4909
NZ_CP006824.1	oki422	84	2230	271	2230
NZ_CP006825.1	oki673	96	170	208	4755
NZ_CP006826.1	oki828	86	205	291	4409
NZ_CP006827.1	oki898	62	1053	252	2230
NZ_CP007603.1	J166	64	2115	311	4681
NZ_CP007604.1	BM013A	69	1514	220	4531
NZ_CP007605.1	BM012B	150	2033	370	2035
NZ_CP007606.1	BM013B	70	1324	222	4531
NZ_CP010013.1	Hp238	50	1241	407	2432
NZ_CP010435.1	26695-1	83	3731	429	1890
NZ_CP010436.1	26695-1MET	83	3731	418	1890
NZ_CP011330.1	J99	91	2114	447	9135
NZ_CP012905.1	7C	98	812	231	2408
NZ_CP012907.1	29CaP	134	1218	238	2921

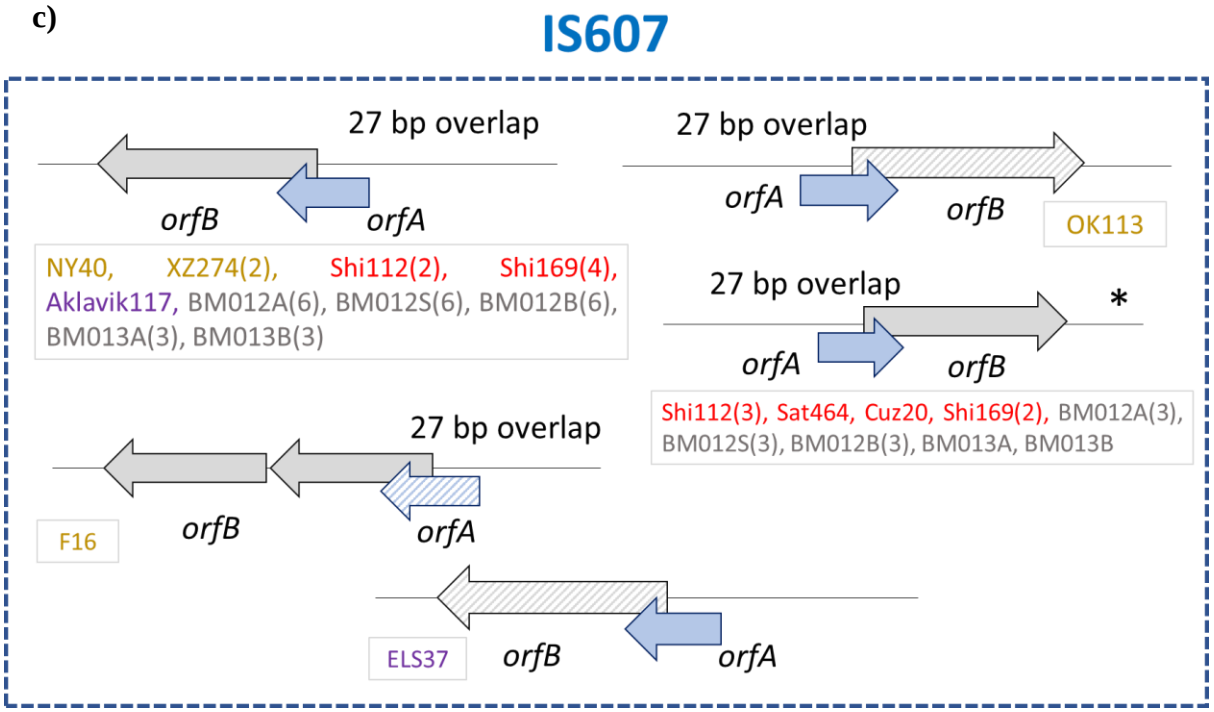
a)


IS605
orfB
orfA
NY40, F32, UM037(4), 26695-1CL(3), 26695-1CH(3), 26695-1(3), 26695-1(3), 26695-1MET(3), Rif2(3), Rif1(3), 26695(3), 26695(3), G27 (3)
*
orfA
orfB
UM037,26695-1CL(2), 26695-1CH(2), 26695-1(2), 26695-1(2), 26695-1MET(2), Rif2(2), Rif1(2), 26695(2), 26695(2), G27
orfB
Sat464
orfB
orfA
v225d
orfA
orfB
G27, 83

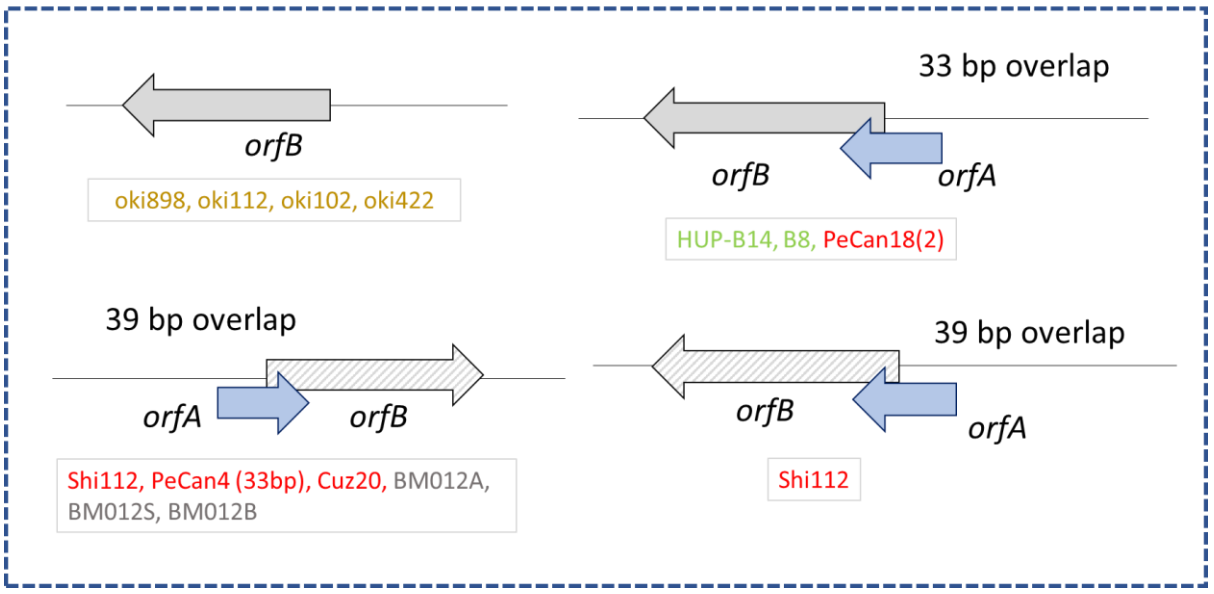
b)


IS606
orfB
orfA
NY40(3), oki673(3), oki154(3), oki828(3), 26695-1CL, 26695-1CH, 26695-1, 26695-1, 26695-1MET, ELS37(2)
*
orfA
orfB
NY40, 29CaP, ELS37(3), SouthAfrica 7(2)
orfA
orfB
ML3, oki128, 2017, 2018, 908
orfB
orfA
oki128(2), 26695-1CL, 26695-1CH, 26695-1, 26695-1, 26695-1MET, ELS37, Rif2 (2), Rif1(2), 26695(2), 26695(2)
orfB
Sat464, v225d
orfA
orfB
J99
orfB
Cuz20, Shi417, PeCan4, Shi169, Puno135, Aklavik117

c)


IS607
27 bp overlap
orfB
orfA
NY40, XZ274(2), Shi112(2), Shi169(4), Aklavik117, BM012A(6), BM012S(6), BM012B(6), BM013A(3), BM013B(3)
27 bp overlap
orfA
orfB
OK113
27 bp overlap
*
orfA
orfB
Shi112(3), Sat464, Cuz20, Shi169(2), BM012A(3), BM012S(3), BM012B(3), BM013A, BM013B
27 bp overlap
F16
orfB
orfA
ELS37
orfB
orfA

d)


IS608
orfB
oki898, oki112, oki102, oki422
33 bp overlap
orfB
orfA
HUP-B14, B8, PeCan18(2)
39 bp overlap
orfA
orfB
Shi112, PeCan4 (33bp), Cuz20, BM012A, BM012S, BM012B
39 bp overlap
orfB
orfA
Shi112

e)

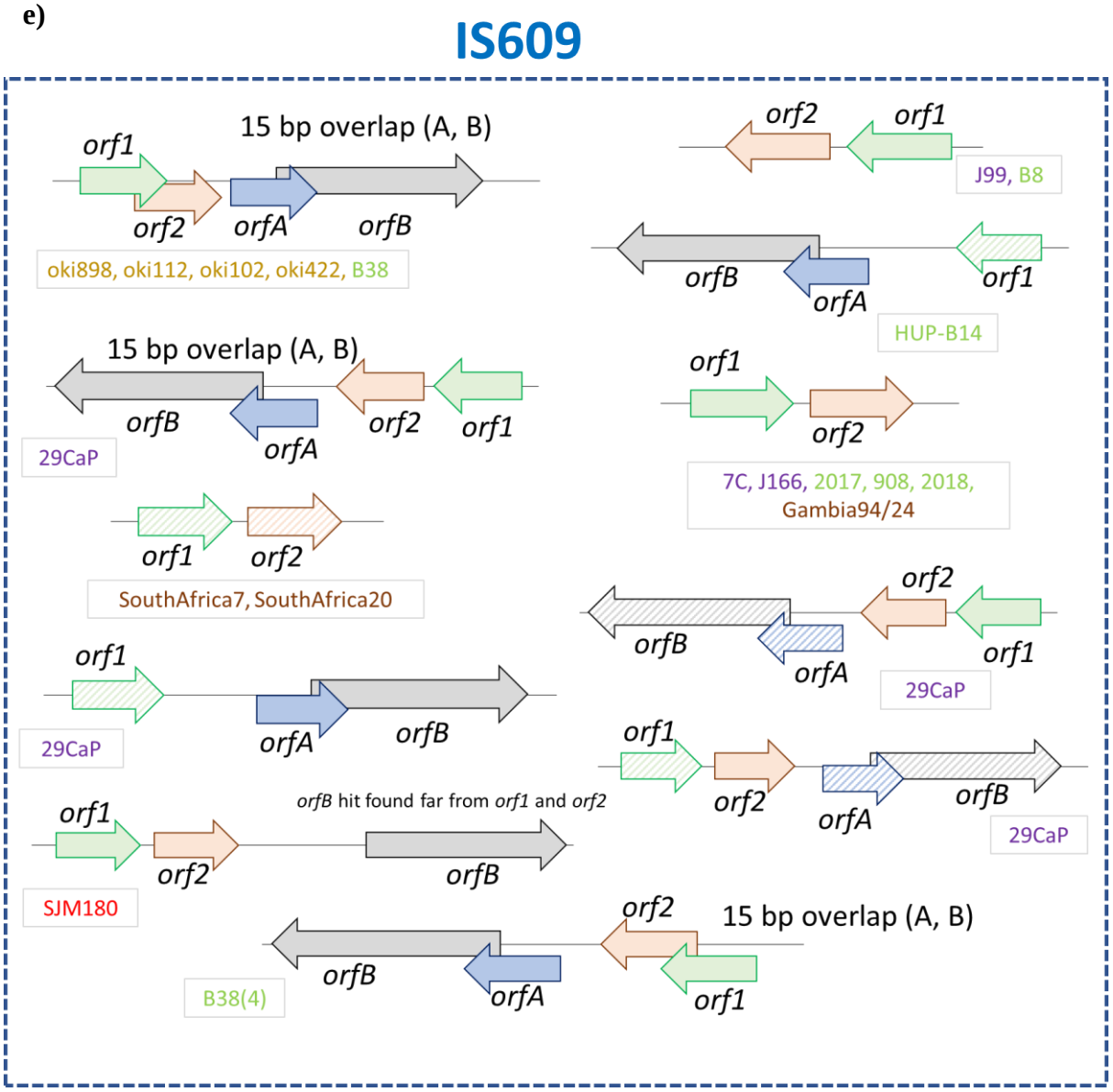


Figure A.4: Structure of each insertion sequence found in strains from different geographical locations designated by different colors (East Asia, South America, North America, Australia, Europe, Africa). The number written next to each strain name represents the number of copies of IS element. The stripped arrows indicate that it is a pseud gene. Asterisk (*) represent the orientation as in ref [147, 148]. **(a)** IS605; **(b)** IS606; **(c)** IS607; **(d)** IS608; **(e)** IS609 [172].

Table A.5: Elements present around the inversion breakpoints. IS: Insertion Sequence, IR: Inverted Repeat, GI: Genomic Island, DR: Direct Repeat, RMS: Restriction Modification System [172].

Inversion	Breakpoints [1st BrP], [2nd BrP]	Element around breakpoints (BrPs)
R1	[B1], [B2]	[-], [IS]
R2	[B3], [B4]	[IS {as IR}, GI], [IS {as IR}, GI]
R3	[B5 or B6], [B7]	[IS {as IR}, Type II RM genes OR Type II RM genes], [Cell division genes]
R4	[B5], [B8]	[IS {as IR}, Type II RM genes], [Type I RMS gene(s)]
R5	[B8 or B31], [B9 or B10]	[Type I RMS gene(s) OR Type II RM genes], [Type I S gene OR Type II R gene]
R6	[B11], [B12]	[IR or DR], [IR or DR]
R7	[B13], [B14]	[Type II RM gene {as IR}, DR], [Type II RM gene {as IR}, DR]
R8	[B13], [B15]	[Type II RM gene {as IR}, DR], [DR]
R9	[B16], [B17]	[IR], [IR]
R10	[B18], [B19]	[Type II RM gene(s)], [-]
R11	[B20], [B21]	[DR], [DR or IR or GI or -]
R12	[B22], [B23]	[Type II RM gene(s)], [Type II RM gene(s)]
R13	[B24], [B25]	[Type III RM gene(s)], [Endonuclease]
R14	[B26], [B27]	[Integrase gene], [Virulence gene]
R15	[B28], [B32 or B45]	[Virulence gene], [-]
R16	[B29], [B30]	[IR], [IR]
R17	[B33], [B34]	[IR], [IR, IS, GI]
R18	[B35], [B36]	[IR], [IR, IS, GI]
R19	[B37], [B38]	[DR], [DR, IS, GI]
R20	[B38], [B39]	[IR], [IR]
R21	[B38], [B40]	[IR], [IR]
R22	[B41], [B42]	[DNA Methyltransferase], [DNA Methyltransferase, IS]
R23	[B43], [B44]	[DNA Methyltransferase], [Restriction Endonuclease, GI]
R24	[B46], [B47]	-
R25	[B47], [B48]	[DR], [DR]
R26	[B47], [B14]	[DR, Type II RM gene], [DR, Type II R gene]
R27	[B46], [B48]	[DR], [DR]
R28	[B49], [B50]	[IR, IS, Type II M gene, GI], [IR, IS, Type II R gene]
R29	[B51], [B52]	[IR, Type I RMS gene], [IR, Type I S gene]
R30	[B53], [B54]	[IS], [GI]
R31	[B55], [B56]	[-], [Type II M gene]
R32	[B55], [B57]	-
R33	[B58], [B59]	[-], [GI]
R34	[B10], [B60]	[Type II R gene], [Type II M gene]
R35	[B61], [B62]	[DNA Methyltransferase], [DNA Methyltransferase]
R36	[B63], [B64]	[Virulence gene], [Virulence gene, GI]
R37	[B65], [B66]	[IR, Type I S gene], [IR, Type I S gene]
R38	[B45], [B66]	[IR], [IR]
R39	[B67], [B68]	[RM genes], [Virulence genes]
R40	[B69], [B70]	[IR], [IR]
R41	[B52], [B71]	[Type I S gene], [-]

Table A.6: 123 *Helicobacter pylori* strains information

Accession	Strain	Geographical location	Disease
NC_014560.1	SJM180	Peru	Gastritis
NC_017063.1	ELS37	El Salvador	Gastric cancer
NC_017354.1	52	Korea	Unknown
NC_017355.1	v225d	Venezuela	Gastritis
NC_017357.1	908	France	Duodenal ulcer
CP001217.1	P12	-	Unknown
CP003419.1	XZ274	China	Gastric cancer
CP006691.1	SouthAfrica20	SouthAfrica	Unknown
CP012907.1	29CaP	Mexico	Gastric cancer
CP031558.1	GD63	Viet Nam: Ho Chi Minh City	Gastric ulcer
NC_000915.1	26695	UK	Gastritis
NC_010698.2	Shi470	Peru: Shimaa (Amazonian region)	Unknown
NC_011333.1	G27	Italy	Unknown
NC_012973.1	B38	France	MALT lymphoma
NC_014555.1	PeCan4	Peru	Gastric cancer
NC_017733.1	HUP-B14	Spain	Unknown
NC_017739.1	Shi417	Peru	Unknown
NC_017358.1	Cuz20	Peru	Unknown
NC_017359.1	Sat464	Peru	Unknown
NC_017360.1	35A	-	Unknown
NC_017361.1	SouthAfrica7	SouthAfrica	Unknown
NC_017362.1	Lithuania75	Lithuania	Unknown
NC_017365.1	F30	Japan	Duodenal ulcer
NC_017366.1	F32	Japan	Gastric cancer
NC_017367.1	F57	Japan	Gastric cancer
NC_017368.1	F16	Japan	Gastritis
NC_017371.1	Gambia94/24	Gambia	Unknown
NC_017372.1	India7	India	Unknown
NC_017374.1	2017	France	Duodenal ulcer
NC_017375.1	83	-	Unknown
NC_017376.1	Santal49	India	Unknown
NC_017378.1	Puno120	Peru	Unknown
NC_017379.1	Puno135	Peru	Unknown
NC_017381.1	2018	France	Duodenal ulcer
NC_017382.1	51	-	Unknown
NZ_CP006822.1	oki128	Japan	Gastric atrophy
NZ_CP006823.1	oki154	Japan	Duodenal ulcer
NC_017740.1	Shi169	Peru	Unknown
NC_017741.1	Shi112	Peru	Unknown
NC_017742.1	PeCan18	Peru	Gastric cancer
NC_019560.1	Aklavik117	Canada	Unknown

NC_020508.1	OK113	Japan	Unknown
NC_020509.1	OK310	Japan	Unknown
NC_021215.3	UM032	Japan	Peptic ulcer disease
NC_021217.3	UM037	Malaysia	Unknown
NC_021218.3	UM066	Malaysia	Peptic ulcer disease
NC_022886.1	BM012A	Australia	Unknown
NC_022911.1	BM012S	Australia	Unknown
NZ_AP014523.1	NY40	Japan	Unknown
NZ_AP017633.1	ATCC 43504	-	Unknown
NZ_AP019730.1	TN2wt	Japan: Oita	Gastroduodenal disease (Duodenal ulcer)
NZ_CP006820.1	oki102	Japan	Gastric atrophy
NZ_CP006821.1	oki112	Japan	Gastric atrophy
NZ_CP019700.1	B128_1	USA: Tennessee	
NZ_CP022409.1	G272	China:Guizhou	Gastritis
NZ_CP006824.1	oki422	Japan	Gastric atrophy
NZ_CP006825.1	oki673	Japan	Gastric ulcer
NZ_CP006826.1	oki828	Japan	Duodenal ulcer
NZ_CP006827.1	oki898	Japan	Duodenal ulcer
NZ_CP007603.1	J166	Nashville	Unknown
NZ_CP007604.1	BM013A	Australia: Perth	Unknown
NZ_CP007605.1	BM012B	Australia: Perth	Unknown
NZ_CP007606.1	BM013B	Australia: Perth	Unknown
NZ_CP010013.1	Hp238	Taiwan	MALT lymphoma
NZ_CP011330.1	J99	USA (Nashville)	Duodenal ulcer
NZ_CP011482.1	L7	India: Ladakh	Unknown
NZ_CP011483.1	DU15	Korea: Seoul	Unknown
NZ_CP011484.1	CC33C	South Africa: Cape Town	Unknown
NZ_CP011485.1	ausabrJ05	Australia: Jigalong	Unknown
NZ_CP011486.1	K26A1	Angola	Unknown
NZ_CP011487.1	PNG84A	Papua New Guinea: Goroka	Unknown
NZ_CP012905.1	7C	Mexico	Chronic gastritis
NZ_CP018823.1	PMSS1	Australia: Sydney	Duodenal ulcer
NZ_CP032475.1	381-F-EK9	Germany: Magdeburg	Atrophic gastritis
NZ_CP024946.1	B147	-	Unknown
NZ_CP024947.1	J182	-	Unknown
NZ_CP024948.1	B140	-	Unknown
NZ_CP024949.1	B136A	-	Unknown
NZ_CP024950.1	B130A	-	Unknown
NZ_CP024952.1	B125A	-	Unknown
NZ_CP024953.1	7.13	-	Unknown
NZ_CP025474.1	H-137	South Korea: Seoul	Unknown
NZ_CP027404.1	FDAARGOS_300	USA:VA	Duodenitis
NZ_CP028325.1	FDAARGOS_298	Australia: Perth	Gastritis
NZ_CP032471.1	479-C2-EK2	Germany: Magdeburg	Atrophic gastritis
NZ_CP032473.1	476-A2-EK2	Germany: Magdeburg	Gastritis
NZ_CP032911.1	19-A-EK3	Germany: Magdeburg	Gastritis

NZ_CP032912.1	13-A-EK8	Germany: Magdeburg	Gastritis
NZ_CP032913.1	5-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP032477.1	169-C-EK8	Germany: Magdeburg	Gastritis
NZ_CP032478.1	25-A-EK9	Germany: Magdeburg	Gastritis
NZ_CP032479.1	21-F-EK1	Germany: Magdeburg	Gastritis
NZ_CP032899.1	478-A-EK1	Germany: Magdeburg	Atrophic gastritis
NZ_CP032902.1	280-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP032903.1	173-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP032905.1	26-A-EK1	Germany: Magdeburg	Atrophic gastritis
NZ_CP032907.1	24-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP032908.1	23-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP032910.1	20-A-EK1	Germany: Magdeburg	Gastritis
NZ_CP036392.1	48C8	Germany: Berlin	Gastritis
NZ_CP034071.1	Hpbs1	China: Baise	Gastric ulcer
NZ_CP034147.1	HP14039	Australia: Perth	Unknown
NZ_CP034314.1	HP42K	Belarus	Gastritis and duodenitis (Duodenal ulcer)
NZ_CP035105.1	Hpbs2	China: Baise	Chronic gastritis
NZ_CP035106.1	Hpbs3	China: Baise	Chronic gastritis
NZ_CP036379.1	H1	USA: Houston	Gastritis
NZ_CP036380.1	125C7	Germany: Berlin	Gastritis
NZ_CP036382.1	119C10	Germany: Berlin	Gastritis
NZ_CP036384.1	103C8	Germany: Berlin	Gastritis
NZ_CP036386.1	87C7	Germany: Berlin	Gastritis
NZ_CP036388.1	81C9	Germany: Berlin	Gastritis
NZ_CP036390.1	78C8	Germany: Berlin	Gastritis
NZ_CP048599.1	GCT 97	Colombia: Tolima	Chronic active gastritis
NZ_CP048600.1	GCT 43	Colombia: Risaralda	Chronic active gastritis
NZ_CP048601.1	GCT 27	Colombia: Valle del Cauca	Chronic active gastritis
NZ_CP053256.1	A45	Russia: Moscow	Peptic ulcer & chronic gastritis
NZ_LR134517.1	NCTC13345	Nigeria	
NZ_LR698956.1	MGYG-HGUT-01357	-	Unknown
NZ_LS483488.1	NCTC 11637	Australia	Unknown
NZ_LT837687.1	BCM-300	-	Unknown
NZ_LT838273.1	HE93/10_v1	-	Unknown
NZ_CP036394.1	29C8	Germany: Berlin	Gastritis
NZ_CP036396.1	12C8	Germany: Berlin	Gastritis
NZ_CP036398.1	8C10	Germany: Berlin	Gastritis

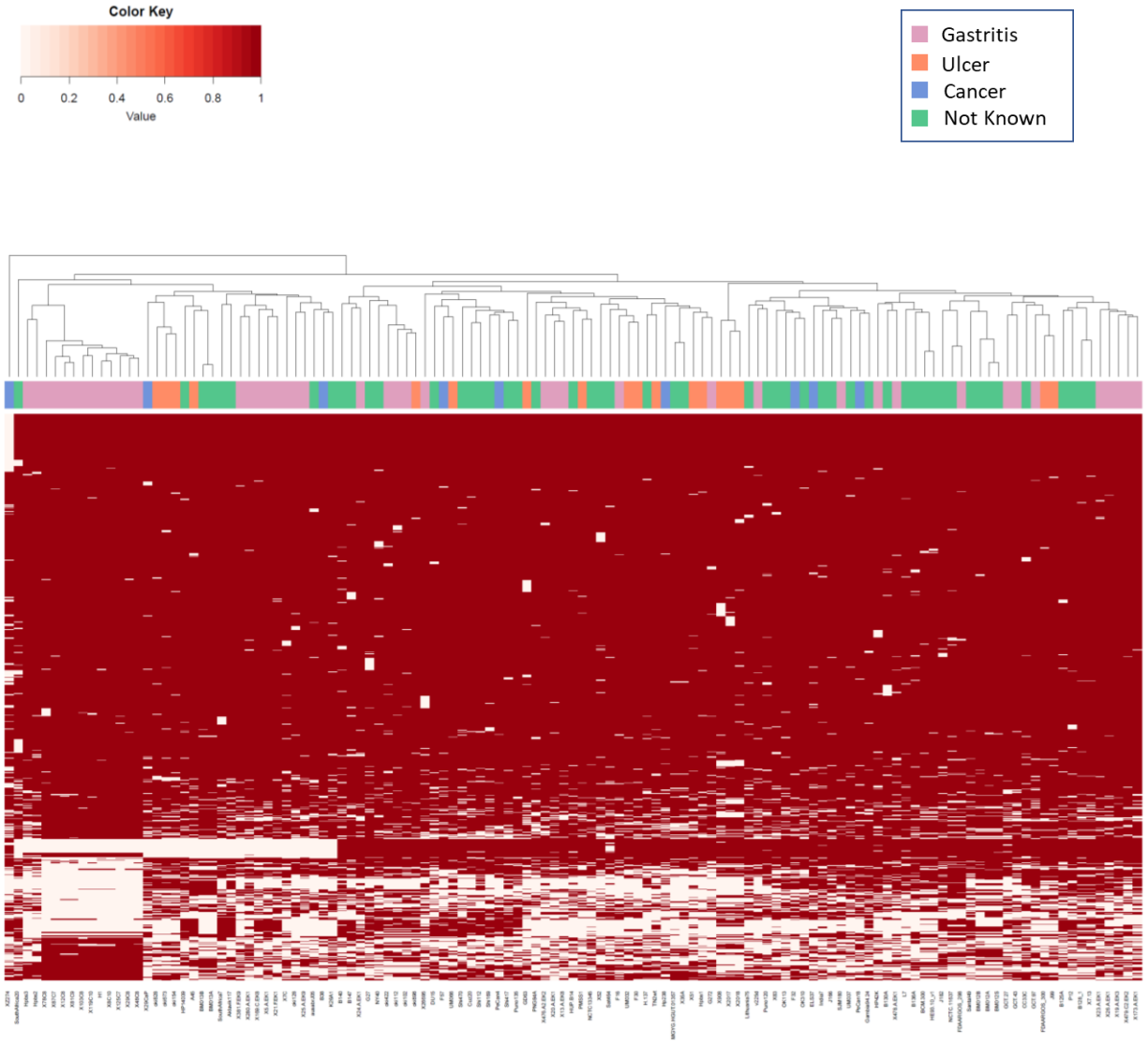


Figure A.5: Clustering of the genomes based on the presence and absence of the accessory genes. The genes present in more than 30% percent of the strains, excluding the core genes were used for this analysis. Here, the strains were divided into 4 groups as shown in the upper right legend. The strains of the atrophic gastritis, gastritis, chronic active gastritis and gastric atrophy are included in the Gastritis group. The group named ulcer includes the strains having gastric ulcer, duodenal ulcer and peptic ulcer as disease outcomes. The MALT lymphoma and gastric cancer strains are classified into cancer group. The group named Not known here represent the strains having no information of the disease outcome.

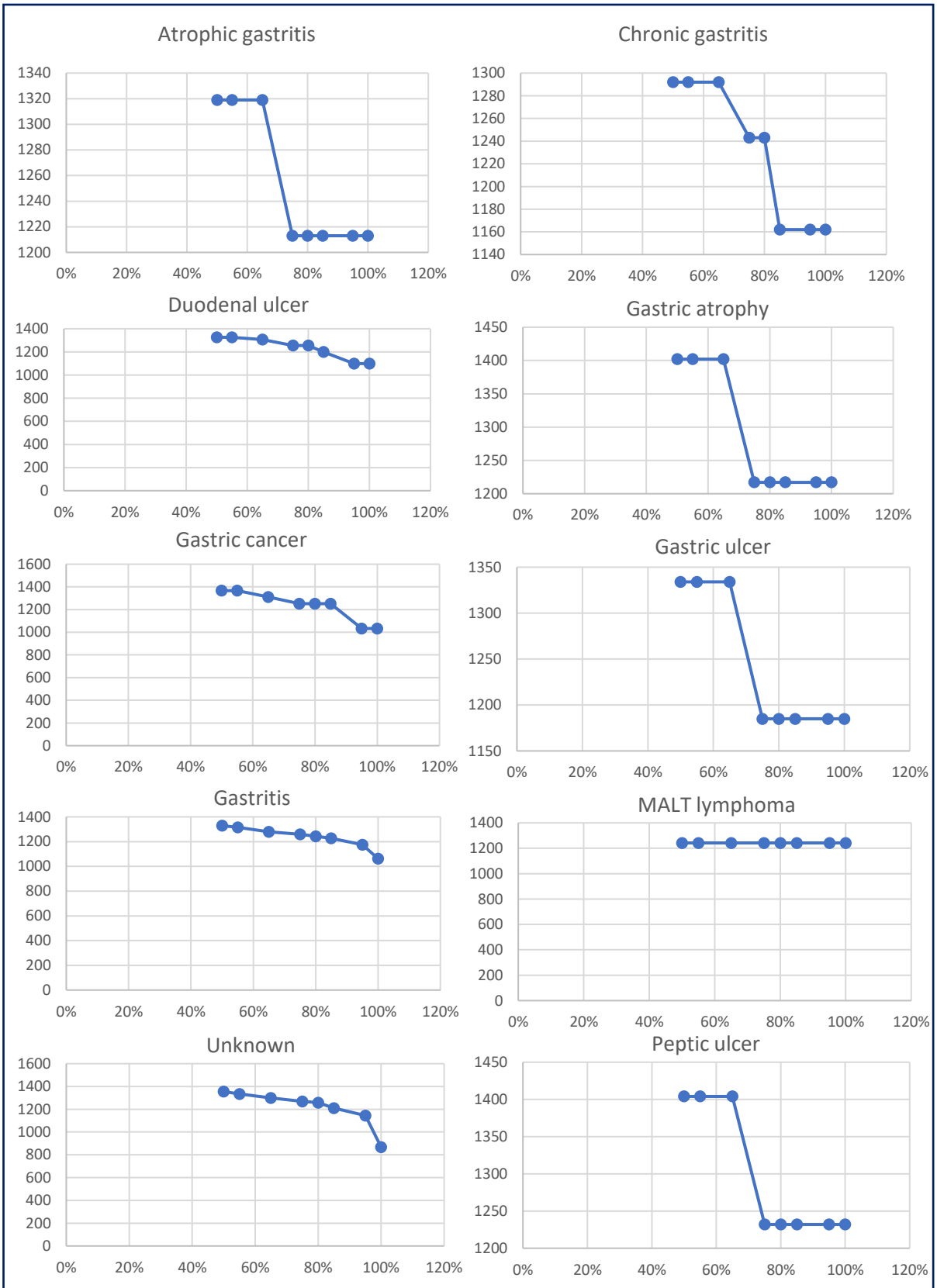


Figure A.6: Number of shared genes among different percentage of the strains in each group. Sharp increase in the number of genes as the percentage of strains is decreased in various groups is due to small number of strains in those groups. X-axis represents the percentage (50-100) of strains sharing the genes whereas, the y-axis represents the number of shared genes.

a)

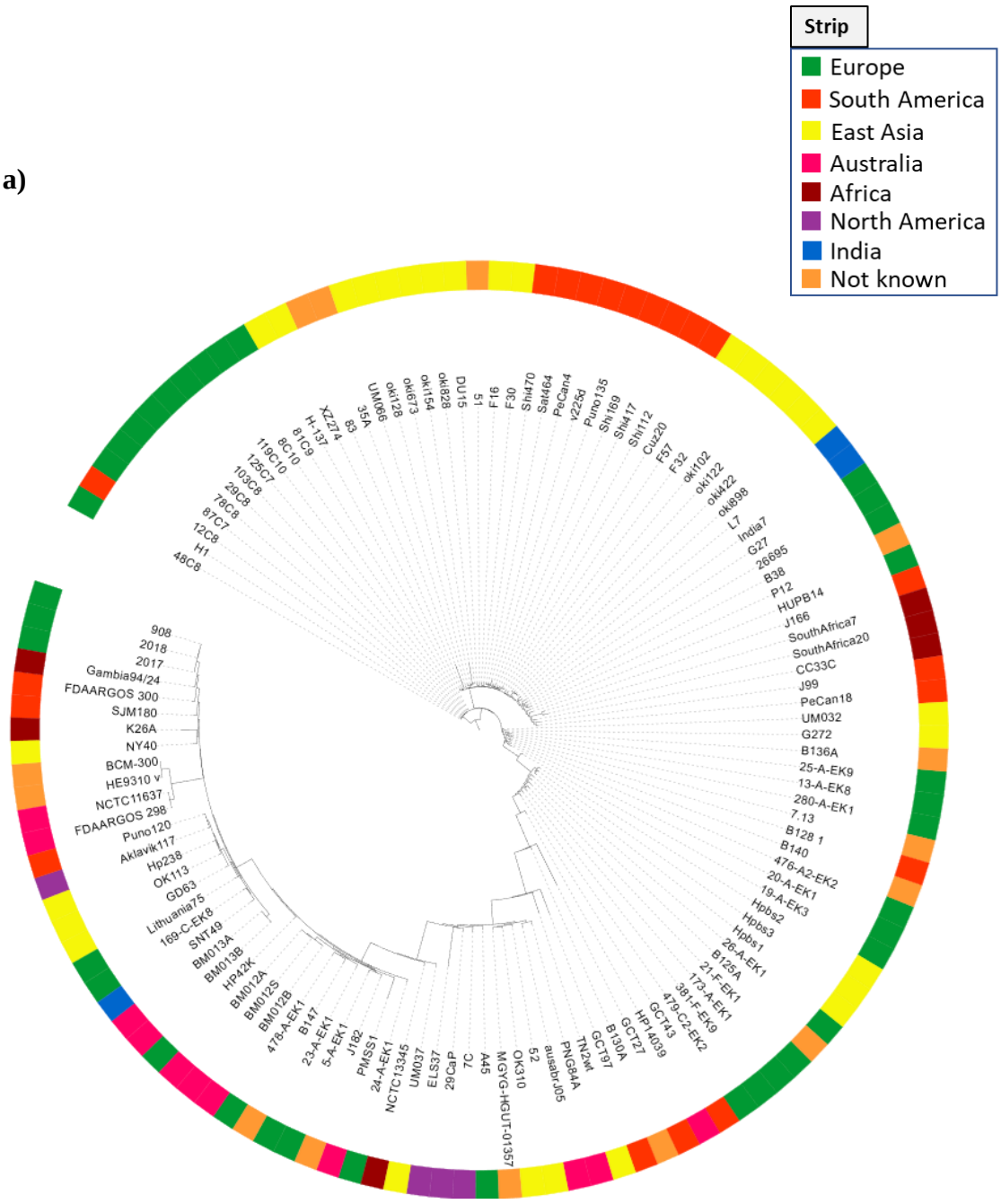


Figure A.7: a) Phylogenetic tree of 123 *H. pylori* strains on the basis of the seven housekeeping genes. The colored strip indicates the geographical location to which a particular strain belongs. Some of the strains belonging to the Europe, East Asia and South America are clustered together.

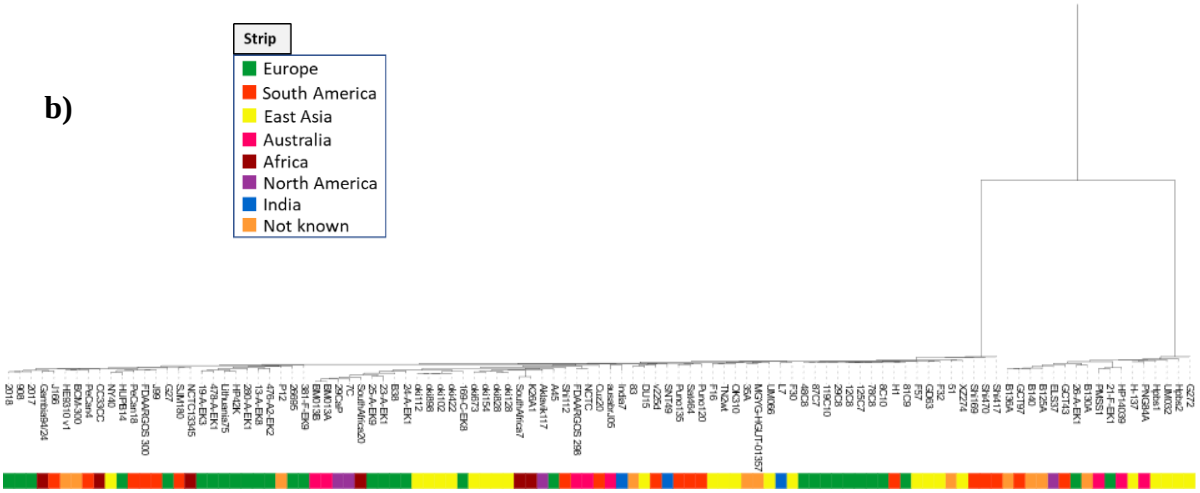


Figure A.7: b) Phylogenetic tree of 105 *H. pylori* strains on the basis of the *vacA* gene. The colored strip indicates the geographical location to which a particular strain belongs. Some of the strains belonging to the Europe are clustered together. Few strains from East Asia and South America are also clustered together.

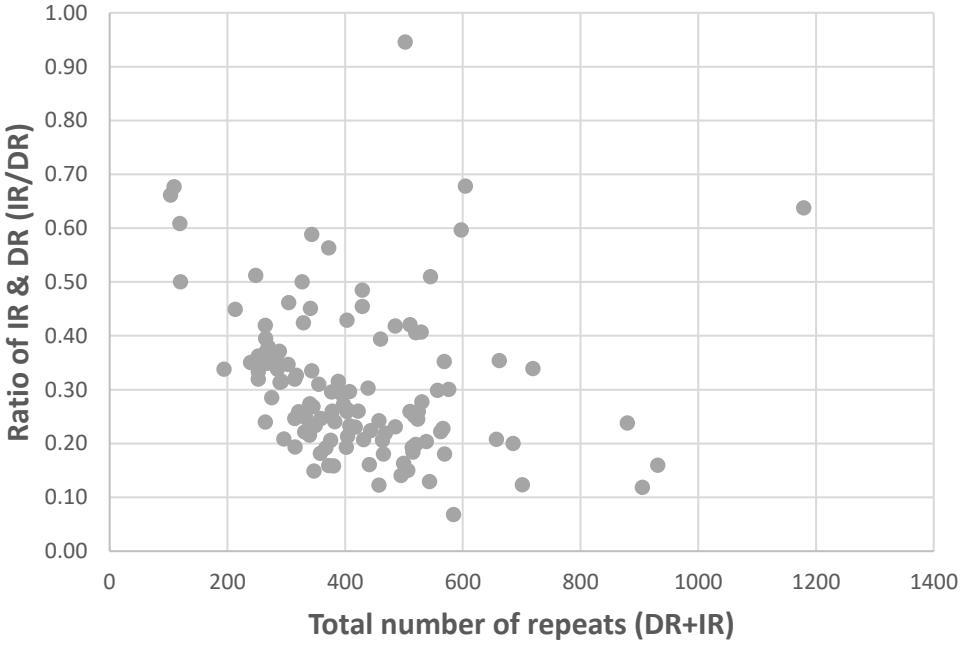


Figure A.8: Distribution of the ratio of inverted repeats (IR) over direct repeats (DR). This ratio (IR/DR) less than 1 indicates the underrepresentation of inverted repeats.