UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Genomes of *Helicobacter pylori* prophages

Ana Filipa Ferreira do Vale

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Dissertação orientada por:
Philippe Lehours
Francisco Dionísio

2017

Acknowledgments

I few years ago I wasn't expecting writing another thesis and another acknowledgements, but here I am again, refocusing my formation and expertise in a new field, full of data, sequences, numbers and tests, trying to get information out of it. I'd like to thank all of those who helped me in this re-orientation process. To my supervisor and friend Professor Philippe Lehours for his continuous support and help, with whom I intend continuing the fruitful work we have been developing. To my co-supervisor Professor Francisco Dionísio, thank you for answering so quickly to my e-mails and doubts and for scientific supervision. To my colleagues of the master degree, with whom I have the opportunity of experience again the stress of getting a work done in a short period of time. To my professors who taught me how to program, thank you Professor António Branco. To my colleagues at work for their understanding when I was doing my homework assignments. Finally, to my family for being always there, especially to my sister Isabel and my nice Maria who heard me countless times about continuing, or not continuing study, to go one way or another way, and always supported me in my decisions.

Abstract

Helicobacter pylori is a Gram-negative bacterium that infects over half of the human population. Most of the colonized people remain asymptomatic but near 20% of them develop serious gastroduodenal diseases like peptic ulcer or gastric cancer. *H. pylori* shares a co-evolutionary history with the human host presenting similar a phylogeographic structure. *H. pylori* is characterize by its high genome diversity attributed to high mutation and recombination rates. Mobile genomic elements also contribute for the genetic diversity of *H. pylori*. In the present study prophages, the least characterized mobile elements of *H. pylori* are studied using next generation sequencing and bioinformatics tools for genome assembly, comparison and phylogenetic analysis. The full genomic sequences, insertion sites and phylogenetic analysis of 28 prophages found in *H. pylori* isolates from patients of distinct disease types, ranging from gastritis to gastric cancer, and geographic origins, covering most continents are presented. The genome sizes of these prophages range from 22.6-33.0 Kbp, consisting of 27-39 open reading frames. A 36.6% GC was found in prophages in contrast to 39% in *H. pylori* genome. Remarkably a conserved integration site was found in over 50% of the cases. One phage was present as an episome, suggesting a pseudolysogenc life cycle. Nearly 40% of the prophages harbored insertion sequences (IS) previously described in *H. pylori*. Tandem repeats were frequently found in the intergenic region between the prophage at the 3'end and the bacterial gene. The pangenome of the described known phages of *H. pylori* consists in 55 genes. Furthermore, prophage genomes present a robust phylogeographic pattern, revealing four distinct clusters: one African, one Asian and two European prophage populations. Evidence of recombination was detected within the genome of some prophages, resulting in genome mosaics composed by different populations, which may yield additional *H. pylori* phenotypes.

Part of this study was published this year in Scientific Reports: Vale FF, Nunes A, Oleastro M, Gomes JP, Sampaio DA, Rocha R, Vítor JM, Engstrand L, Pascoe B, Berthenet E, Sheppard SK, Hitchings MD, Mégraud F, Vadivelu J, Lehours P. 2017. Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. Sci Rep. 7:42471. doi: 10.1038/srep42471.

Keywords: *Helicobacter pylori*, prophage, co-evolution, phylogeography, phage–host interactions

Resumo

Helicobacter pylori é uma bactéria Gram-negativa que infecta o estômago de mais de metade da população humana, pelo que é uma das infeções mais frequentes no mundo. A principal patologia associada a *H. pylori* é a gastrite, que pode permanecer assintomática. Outras patologias associadas a esta bactéria incluem a úlcera péptica, que atinge cerca de 20% dos infectados, e o cancro gástrico, desenvolvido em menos de 1% dos casos. A bactéria *H. pylori* partilha com o hospedeiro humano a mesma história evolutiva, estimando-se que esta associação date de há mais de 100.000 anos, tendo desde então coevoluído com o hospedeiro humano. De facto, *H. pylori* caracteriza-se pela sua estrutura filogeográfica estar dividida em populações que refectem as migrações humanas desde a sua diáspora a partir de África. Esta bactéria tem uma grande variabilidade genética, devido a uma taxa de mutação elevada, mas sobretudo pela sua alta frequência de recombinação, que torna *H. pylori* uma das bactérias mais recombinogénicas. Para a variabilidade do seu genoma contribuem também os elementos com mobilidade genética. No presente trabalho foram estudados os fagos de *H. pylori* que são, entre os elementos com mobilidade genética de *H. pylori*, aqueles com menor caracterização. Os fagos, ou bacteriófagos, são vírus que infectam bactérias e podem apresentar um ciclo lítico, caraterizado pela infecção, replicação do fago e lise da bactéria; um ciclo lisogénico, onde o genoma do fago se integra no genoma da bactéria hospedeira, constituindo um profago; ou, ainda, um ciclo pseudolisogénico, em que o fago não se replica nem se integra, mantendo-se como um episssoma. Os genomas dos fagos de *H. pylori* foram estudados recorrendo à sequenciação de nova geração e a ferramentas bioinformáticas, primeiramente usadas para obter os seus genomas e, posteriormente, para compará-los e proceder às análises filogenética e de recombinação. Neste trabalho é apresentada a sequência genómica de 28 profagos de estirpes de *H. pylori* isoladas de doentes de diversas origens geográficas e com quadros clínicos distintos, desde a gastrite ao cancro gástrico. Para estes fagos é também detalhado o seu local de inserção no genoma de *H. pylori*, caracterizado o seu pangenoma e genoma core e discutidas as evidências que apontam para uma coevolução com *H. pylori*.

Considerou-se como profago intacto a presença de um genoma com mais de 20 Kpb. Entre os 28 genomas de profagos descritos, 23 foram considerados intactos e cinco incompletos, ou remanescentes, podendo estes últimos resultar de um processo de deleções sucessivas, fruto da interacção prolongada entre o profago e a bactéria. Os genomas dos profagos sequenciados têm um tamanho entre 22,6 e 33,0 Kpb e apresentam entre 27 e 39 *open reading frames*. A percentagem G+C dos genomas dos fagos é inferior àquela encontrada no genoma da bactéria hospedeira, sendo de 36,6% e 39,0% respectivamente, o que remete para a transferência horizontal génica.

Do ponto de vista filogeográfico, os profagos encontrados pertencem a quatro populações distintas: hpAfrica1, hpEastAsia, hpNEurope e hpSWEurope. Em mais de metade dos casos os profagos encontravam-se inseridos no mesmo local do genoma da bactéria hospedeira. Esta observação foi especialmente evidente para os profagos que pertencem aos grupos hpAfrica1 e hpNEurope, pois nestes casos encontravam-se sempre inseridos entre os mesmos genes bacterianos, a saber o gene S-adenosilmetionina sintetase (síntese de S-adenosilmetionina) e o gene UDP-3-O-[3-hidroximiristoil] glucosamina N-aciltransferase (via metabólica do lípido A). Importa salientar que estes dois genes são habitualmente contíguos no genoma de *H. pylori*. Os profagos pertencentes à população hpEastAsia foram detectados entre os genes bacterianos que codificam a proteína de competência ComGF e uma proteína da membrana externa. Os fagos da população hpSWEurope parecem estar inseridos em locais aleatórios do genoma bacteriano. Um

dos genomas encontra-se como um episssoma, o que aponta, neste caso, para um ciclo de vida pseudolisogénico. Evolutivamente, os fagos na forma episssomal parecem ser instáveis e constituírem uma forma complexa de interacção entre o fago e a bactéria.

Cerca de 40% dos genomas dos fagos contêm sequências de inserção que tinham anteriormente sido descritas em *H. pylori*. As sequências de inserção detectadas não foram sempre as mesmas, nem o seu local de inserção no genoma do profago constante. Contudo, os genomas dos profagos de estirpes isoladas no mesmo país tendem a apresentar a mesma sequência de inserção no mesmo contexto genómico. As sequências de inserção podem eventualmente ser transferidas do genoma do profago para o da bactéria, ou vice-versa, contribuindo para a plasticidade de *H. pylori*, ou para a inactivação dos profagos. Foram encontradas sequências repetitivas na região intergénica entre a extremidade 3' do profago e o gene bacteriano seguinte. As sequências de DNA repetitivas, ou DNA satélite, são sequências repetidas de nucleótidos inter ou intragénicos que surgem duas ou mais vezes. Estas sequências são particularmente sensíveis, quer ao deslizamento da DNA polimerase durante a replicação, quer à recombinação, o que pode fazer variar o seu número de cópias, pelo que são consideradas regiões hipermutáveis. As sequências repetitivas podem modular reversivelmente a expressão génica, permitindo a adaptação a novos ambientes, sem aumento da taxa de mutação. A alteração da expressão génica pelas sequências repetitivas pode ocorrer pela alteração da afinidade da ligação de proteínas reguladoras, ou pela alteração da distância a que o promotor se encontra.

O pangenoma dos fagos de *H. pylori* é constituído por 55 genes, que correspondem ao conjunto dos genes encontrados no grupo dos genomas dos profagos intactos do presente estudo e doutros genomas anteriormente sequenciados. O genoma core consiste em 9 genes, a maioria dos quais com função desconhecida, presentes em todos os genomas dos profagos. O genoma core caracteriza-se por desempenhar frequentemente funções vitais para a sobrevivência, enquanto que o genoma acessório está associado a fenótipos específicos. Tal como referido, os genomas dos profagos apresentam um padrão filogeográfico robusto, constituído por quatro grupos distintos, um Africano, um Asiático e dois Europeus, um do sudoeste e outro do norte da Europa. De facto, a análise filogenética corrobora a presença de grupos que coincidem com a estrutura populacional dos profagos, obtida através de um método de tipagem em que se utilizaram apenas dois genes fágicos, os genes da integrase e da holina. A existência de grupos filogeográficos está de acordo com um modelo de coevolução entre o fago e a bactéria. Mais ainda, a integração do profago no genoma bacteriano é consistente dentro da maioria dos grupos filogeográficos, sugerindo a transmissão vertical dos fagos em alternativa à inserção aleatória. Para a evolução dos fagos contribui a troca horizontal de módulos funcionais entre fagos mais ou menos relacionados, o que é conseguido através da recombinação do DNA. A recombinação explica o mosaicismo genómico encontrado em fagos e é um factor que contribui para a rápida variabilidade de *H. pylori*, que é aliás o agente patogénico com maior grau de recombinação conhecido. Também nos genomas dos profagos foram encontrados mosaicos, na maioria da vezes envolvendo trocas de sequências entre as populações hpAfrica1 e hpSWEurope. Estas populações são muitas vezes encontradas na mesma área geográfica, o que pode justificar a recombinação frequente detectada entre ambas.

Neste estudo, foram ainda encontrados profagos remanescentes, sendo que estes também já tinham sido detectados noutras espécies do género *Helicobacter*, o que sugere uma progressiva perda de genes do fago em resultado de uma interacção complexa e prolongada entre estes e *H. pylori*, compatível com a existência de coevolução. Contudo, um modelo em que as linhagens bacterianas são infectadas com profagos de origem geográfica distinta é também possível, mas menos provável devido à elevada sintenia encontrada em fagos de origens geográficas diferentes. Em síntese, a integração no mesmo local do genoma, a sintenia e a existência do

mesmo reportório de genes sugere uma transmissão vertical e uma domesticação dos profagos pela bactéria hospedeira, que pode mediar a adaptação bacteriana.

Parte deste trabalho encontra-se publicado no jornal *Scientific Reports*: Vale FF, Nunes A, Oleastro M, Gomes JP, Sampaio DA, Rocha R, Vítor JM, Engstrand L, Pascoe B, Berthenet E, Sheppard SK, Hitchings MD, Mégraud F, Vadivelu J, Lehours P. 2017. Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. Sci Rep. 7:42471. doi: 10.1038/srep42471.

Palavras-chave: *Helicobacter pylori*, profago, coevolução, filogeografia, interacção fago-bactéria hospedeira

Table of contents

List of Figures

List of Tables

1. Introduction

1.1 *Helicobacter pylori* overview

The introductory section is subdivided in the biologic basis and the bioinformatics approaches used for studding the *Helicobacter pylori* phage biology from their genomes. This introduction is intended to provide information concerning *H. pylori* biology, the gastric bacterium that colonizes the human stomach, its prophages biology, evolution and phylogeography towards the motivations, aims and description of the studies conducted.

1.1.1 *H. pylori*: diseases, diagnosis and therapy

H. pylori is a Gram-negative bacterium that colonizes about half of the human population and is associated with several gastrointestinal diseases, such as gastritis (all cases), peptic ulcer (20% of the infected individuals), and in rare cases gastric cancer (1%) and gastric MALT (Mucosa Associated Lymphoid Tissue) lymphoma (<1%). Considering that 1% of half the human population will develop a severe complication of *H. pylori* infection, it's expectable to have 600,000 new cases of gastric cancer each year (1). Indeed, *H. pylori* infection is the strongest risk for gastric cancer and in 1994 the International Agency for Research on cancer classified *H. pylori* as type I carcinogen (2). Thus, *H. pylori* infection is a major public health issue. This infection is mainly acquired in childhood after a short acute infection and if left untreated is long lasting through adulthood, although the mechanism of transmission remains unknown. Presently, the human stomach is the only recognized and accepted reservoir, nonetheless other extra-gastric reservoirs have been suggested. Although the principal transmission route of *H. pylori* is not clearly defined there are two main models. First, the vertical transmission from parents to children, by direct person-to-person contact, most likely by gastro–oral or faecal–oral route. Second, the horizontal transmission, by ingestion of contaminated food or water, or via intensive contact between infants and non-parental caretakers. Most likely in urban (developed) areas the vertical transmission represents the main form of transmission, while in rural (non-developed) areas the transmission appear to be much more complex, with the horizontal form appearing to play a major role, but not excluding the first. In fact, horizontal transmission also includes person-to-person transmission, but does not exclude ingestion of contaminated water and food (3).

H. pylori discovery and association with gastritis and peptic ulcer disease was made by Barry Marshal and Robin Warren in 1984 (4), both of whom were awarded the Nobel Prize in Medicine in 2005.

The diagnosis of *H. pylori* is made using invasive and non-invasive tests. There is a broad spectrum of diagnostic methods, but only highly accurate tests should be used in clinical practice, *i.e.* the sensitivity and specificity of an adequate test should exceed 90%. The choice of the diagnosis method to be used depends on clinical circumstances, the likelihood ratio of positive and negative tests, the cost-effectiveness of the testing strategy and of the availability of the tests (5). Currently, the urea-breath test is the best recommended non-invasive test in the context of a 'test-and-treat strategy' (6).

For the eradication of *H. pylori* a triple therapy, which applied two antibiotics, more often amoxicillin and clarithromycin and a proton pump inhibitor (PPI) or ranitidine bismuth, has been recommended for many years (6, 7). Antibiotic therapy fails in about 20% of the patients, mainly due to antibiotic resistance (7), which is considerable an unacceptable (< 85% success)

eradication therapy (8), making urgent the development of new drugs. In fact, *H. pylori* eradication is not easy to achieve, because the bacterium is located in the stomach lining in an acidic environment that is not favorable to antibiotic activity (9). Presently, PPI-clarithromycin-containing triple therapy without prior susceptibility testing is recommend to be abandoned when the clarithromycin resistance rate in the region is more than 15%. Thus in areas of low clarithromycin resistance, triple therapy is recommended as first-line empirical treatment. In areas where clarithromycin resistance is high (>15%), bismuth quadruple or non-bismuth quadruple, concomitant (PPI, amoxicillin, clarithromycin and a nitroimidazole administered concurrently) therapies during 14 days are recommended (6). The World Health Organization (WHO) published this year for the first time the list of antibiotic-resistant "priority pathogens". Indeed, *H. pylori* makes part of WHO priority pathogens list for research and development of new antibiotics (priority 2 - clarithromycin-resistant), highlighting the particular threat of Gram-negative bacteria that are resistant to multiple antibiotics (10).

1.1.2 *H. pylori* virulence mechanism and virulence factors

H. pylori expresses several virulence factors that promote this bacterium survival in the human stomach, allowing the survival in acidic conditions and establishing a persistent infection of the gastric mucosa (reviewed in (11)). In order to survive this bacterium penetrates the outer mucous gel layer of the stomach, where the external pH is approximately 5-6. The spiral shape and flagella of the bacterium may provide a fitness advantage by promoting the ability to penetrate the mucous layer (12). Furthermore, *H. pylori* expresses urease (catalyzes the hydrolysis of urea into ammonia and carbon dioxide) in response to acidity, which allows increasing the pH surrounding the bacterium, its cytosol and periplasm (13). Although near 80% of the cells remain free swimming, the remaining adhere to epeithelial cells (14) through interaction bacterial adhesins (all belonging to the major outer membrane family 1, such as BabA, SabA and OipA) and their cognate receptors on host epithelial cells (11, 15).
The most well known virulence factors are the cag pathogenicity island (cagPAI) and the vacuolating cytotoxin VacA. The cagPAI is acquired by horizontal transfer, has 40 kb and encodes for a bacterial type IV secretion system (T4SS) which translocates CagA and peptidoglycan, into gastric epithelial cells, triggering multiple intracellular signaling cascades (11). VacA is pore-forming toxin that causes cellular alterations through endocytic alterations, autophagy, disruction of cell-cell junctions, and cell death (16). Other emerging virulence factors have been proposed having roles in apoptosis (CtkA), inflammation (CtkA), adhesion (HopQ, HopZ), disruption of adherence junctions (HtrA), modulation of the immune system (Tipα), effect on neutrophils (NapA), among others (reviewed in (17)).

1.1.3 *H. pylori* population structure and human migrations

The study of seven housekeeping genes of *H. pylori* has been widely used to characterize the strains. The genes used for genes multilocus sequence typing (MLST) are *atp*A, *efp*, *mut*Y, *ppa*, *trp*C, *ure*A and *yph*C (18). The Bayesian clustering of these seven housekeeping genes applied to hundreds of strains from distinct geographic regions (19, 20) revealed the presence of seven modern populations of *H. pylori* that clusters according to the geographic origin of the bacterium and its host (reviewed in (21-23)). The seven modern populations of *H. pylori*, hpAfrica2, hpAfrica1, hpNEAfrica, hpSahul, hpAsia2, hpEurope and hpEastAsia (Figure 1.1), evidences that *H. pylori* and man co-evolved together, since they went 'out of Africa' (19, 20, 24). Each of these populations may be divided into subpopulations. For example hpAfrica1 is

currently divided in hspSAfrica, hspWAfrica and hspCAfrica. The structured population provided strong evidence of ancient ancestry in Africa and of co-evolution with the human host since then. The original Human migration from Africa to the Middle East is estimated to have occurred ~60,000-150,000 years ago and then independently to Europe and Asia (25-27).

H. pylori is a highly recombinogenic species (28). Considering that recombination requires physical exchange of genomic DNA, recombination is more evident within populations than among populations (29). While mutations are passed vertically to the offspring, recombination occurs between unrelated organisms which can create homoplasies, *i.e.*, a similar sequence acquired from an unrelated lineage. This form of convergent evolution may biases the reconstructions of clonal phylogenies. This effect is observable in Figure 1.1.A. where branches separating strains are much longer than the ones separating populations (29). Figure 1.1.B. shows the resulting population assigned using the number of bacterial populations, K=7, using the program STRUCTURE, that uses a Bayesian approach. This program is run for several values of K and in each run, for each K, a Markov Chain Monte Carlo simulation of thousand of iterations approximate the posterior probability of K. The number of populations (K) that best clusters the data presents simultaneity higher posterior probability and is biologically interesting, *i.e.*, correspond to real populations (30).

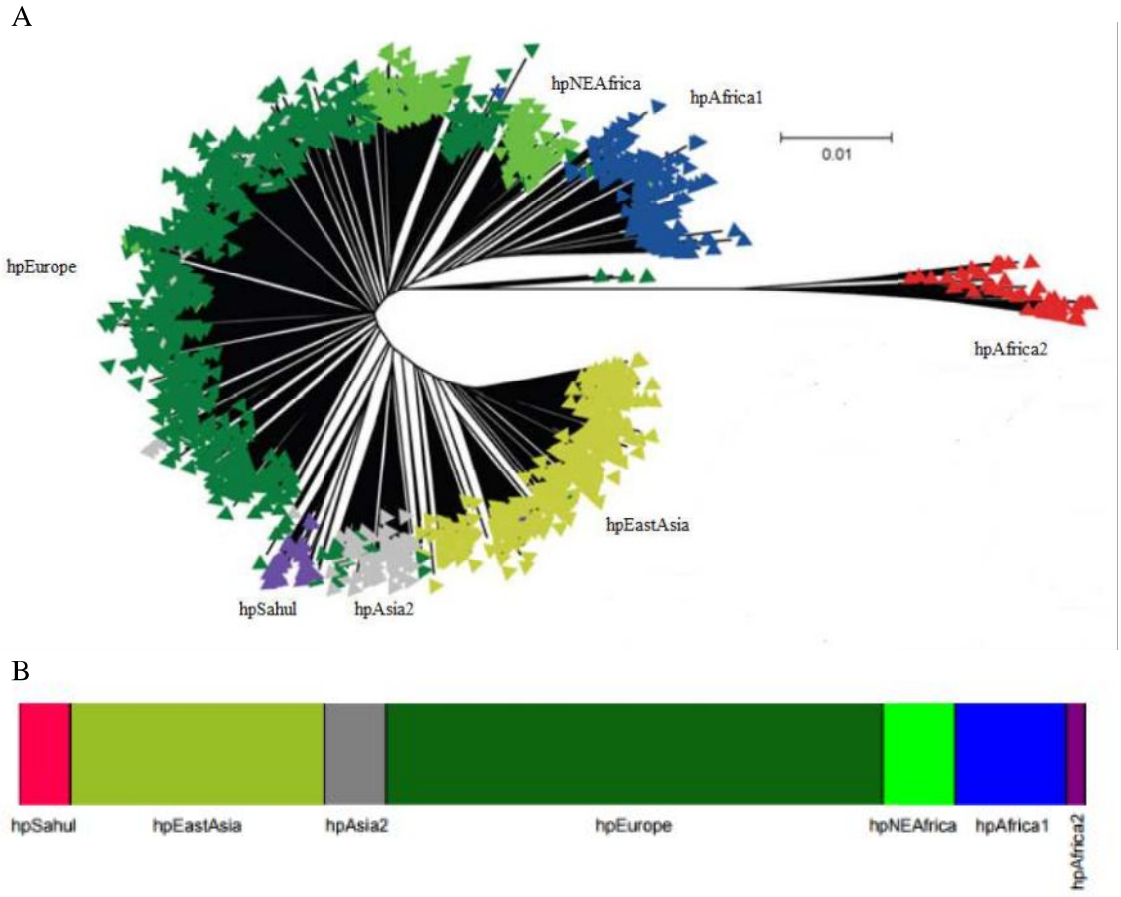


Figure 1.1 Worldwide population structure of *H. pylori*. A) Neighbor-joining tree using Kimura two-parameter model of the concatenated *H. pylori* housekeeping genes (adapted from (29)). B. DISTRUCT plot of the Bayesian assignment of *H. pylori* to populations using STRUCTURE V2.0 with no admixture model, where each isolate is represented by a thin line that is color coded according to the population assignment (adapted from (24).

The increasing number of *H. pylori* genomes available provides a mean to obtain more information about its phylogeny. This is the case of the overcome of the difficulty in inferring the population structure due to high recombination rate found in *H. pylori*. Briefly, a method

called chromosome painting in silico (31) was used to detect the transfer of DNA sequence chunks between genomes through homologous recombination throughout the genome (32). A co-ancestry matrix is generated showing the expected number of chromosome chunks imported from a donor to a recipient genome. The matrix is then used to assign each strain to a subgroup using fineSTRUCTURE clustering algorithm (31). This method revealed a finer population structure (Figure 1.2) than the one based on MLST genes (32). However, the strains analyzed by this method were mainly from East Asia and a deep understanding of the worldwide population fine structure of *H. pylori* is still missing.

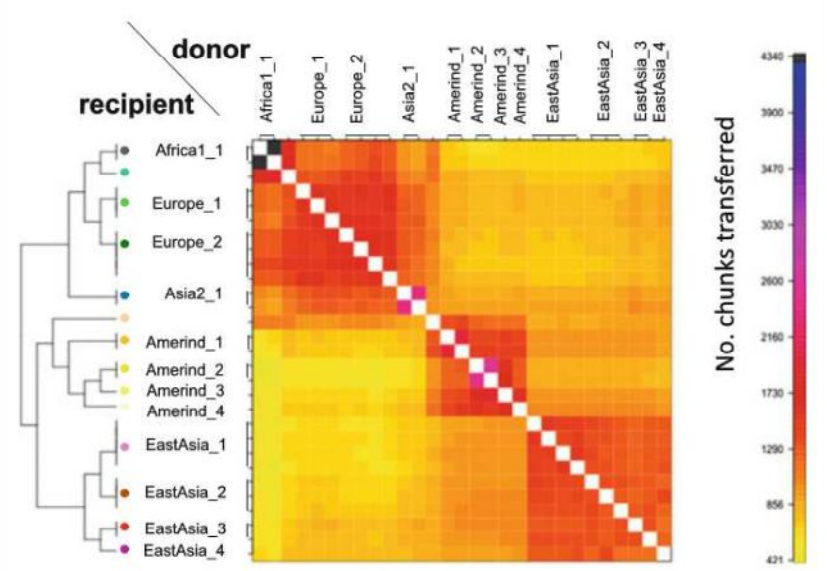


Figure 1.2. Co-ancestry matrix with population structure and genetic flux for *H. pylori* genomes evidencing signs of population admixture in Africa, Europe and Asia. The color of each cell of the matrix indicates the expected number of chunks imported from a donor genome to a recipient genome (adapted from (33)).

1.1.4 Ancestral *H. pylori* populations

The STRUCTURE software has three model options, the "no admixture model", the "admixture model" and the "linkage model". The selection of the most appropriate model depends on the user's data and study objectives. The "no admixture model" is the simplest case where each individual is assumed to have originated in a single population, whereas when there is prior knowledge about the origin of the populations under study and there is no reason to consider each population as completely discrete, the "admixture model" is appropriate. The "linkage model" is like the admixture model, but linked loci are more likely to come from the same population. The linkage model relies on linkage disequilibrium — the nonrandom association of alleles at different loci — that is a sensitive indicator of the population genetic forces that structure a genome (30). There are currently six ancestral or precursor populations inferred to *H. pylori* using the linkage model of STRUCTURE (Figure 1.3) to analyze the seven housekeeping genes used for MLST. These are Ancestral Sahul, Ancestral EastAsia, Ancestral Europe 1(AE1), Ancestral Europe 2 (AE2), Ancestral Africa1 and Ancestral Africa2 (19, 29, 34). Modern populations were produced by admixture of ancient populations.

The case of hpEurope is particularly interesting, as this population is a recombinant of mainly AE1 and AE2 (19). AE1 probably entered Europe via central or southern Asia, while AE2 entered Europe via Northeast Africa or Southern Europe (19, 35). The strains from India

assigned to hpEurope revealed residual evidence of AE2, but presented a higher influence of ancestral EastAsia (Figure 1.3). This influence of ancestral EastAsia was still observed, even that in a small scale, for countries where AE1 is more predominant, favoring the entering of AE1 through Asia. Southern European countries presented a higher proportion of AE2 (Figure 1.3). Interestingly, Iberian countries also present influence of the recombination with ancestral Africa1, that is even higher than AE1 in a few hpEurope strains from African Portuguese speaking countries (35).

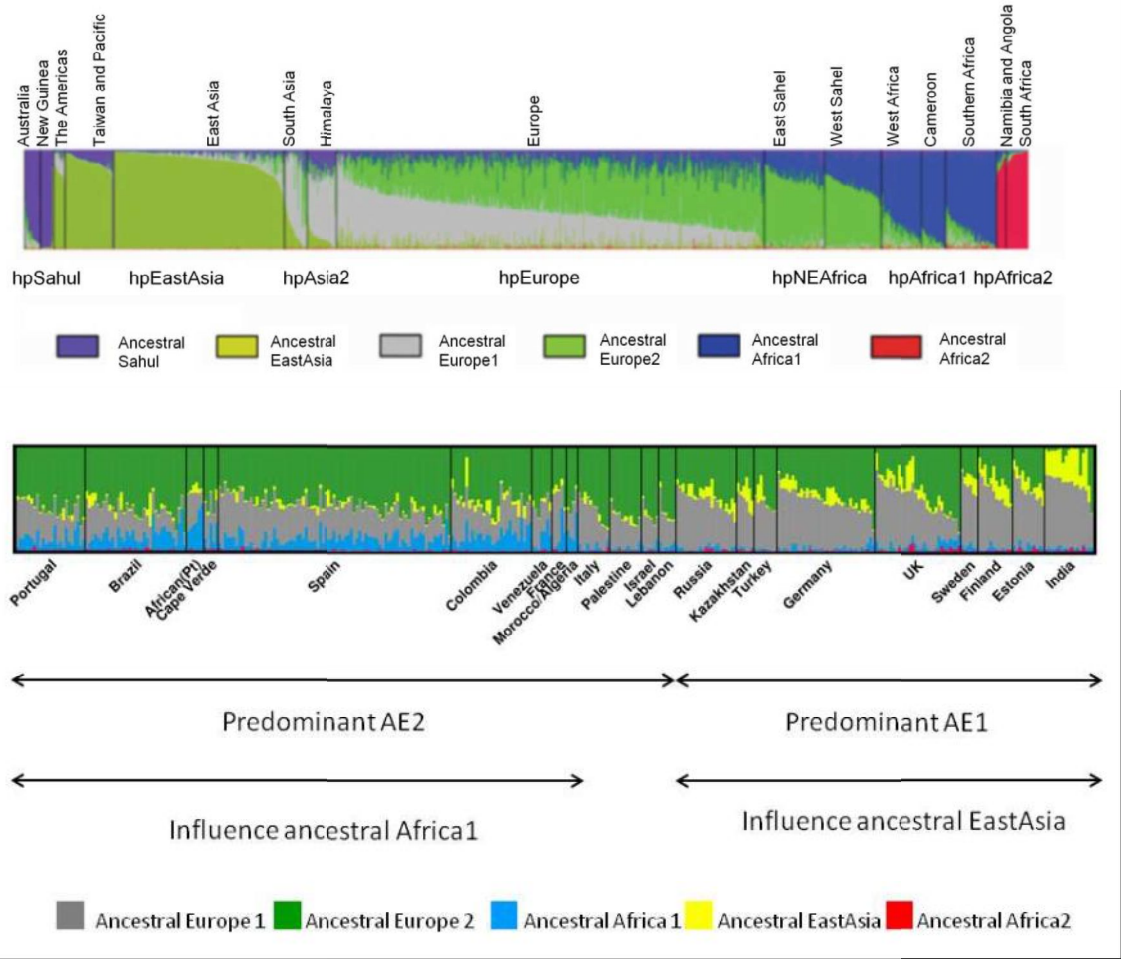


Figure 1.3. Ancestral populations of *H. pylori*. A. General 6 ancestral populations of *H. pylori* (adapted from (29)). B. Detailed ancestral populations found in Europe (adapted from (35)).

The spread of AE2 to Europe may have occurred during the Paleolithic population expansion from the "Atlantic zone" (southwestern Europe) 10,000-15,000 years ago, after the Last Glacial Maximum (36). It is also feasible that a second wave of migration from Africa to Iberia during the Arab Empire (711-1249) introduced ancestral Africa1. During the Arab empire the Iberia peninsula colonizers were mainly Berbers from North Africa, and not Arabs, which is in agreement with ancestral proportions of each population found in Iberian countries and northern Africa (AE1, AE2 and ancestral Africa1) and middle east (AE1 and AE2, but not ancestral Africa1). Before this period there was a commercial trade between Iberia peninsula and Mediterranean nations, which also may explain the influence of ancestral Africa1 (35). The exact way AE1 and AE2 recombination occurred is controversial, but may arose latter than previously expected, since the 5000 years old Iceman mummy found in Italian border presented only AE1 ancestry (37).

1.1.5 Bacteriophages

Bacteriophages (phages) are viruses that infect bacteria. Lytic phages have the property to lyse the bacterial cells and release the phage progeny, while lysogenic or temperate phages either go through a lytic cycle or the phage genome is integrated in the bacterial genome, constituting a prophage. Temperate phages contribute to the evolution of most bacteria, by promoting the transduction of various genes involved in virulence, fitness, and antibiotic resistance (38). Phage genome excision and integration are crucial steps for lytic and lysogenic cycles, respectively. These events are mediated by phage-encoded DNA recombinases, like integrases and excisionases, and take place at a specific attachment site in the bacterial genome (*attB*), which is identical to an attachment site (*attP*) in the phage genome (39). Another less frequent phage life cycle is pseudolysogeny, described as an unstable situation in which the phage genome fails to replicate (lytic cycle) or become established as a prophage (lysogenic cycle). Pseudolysogeny is associated with nutrient-deprived conditions, that impairs DNA replication or protein synthesis, during which the phage genome remains for an extended period of time as a non-integrated preprophage, resembling an episome. When the nutritional status is restored the phage enters either a lysogenic or a lytic life cycle (Figure 1.4) (39).

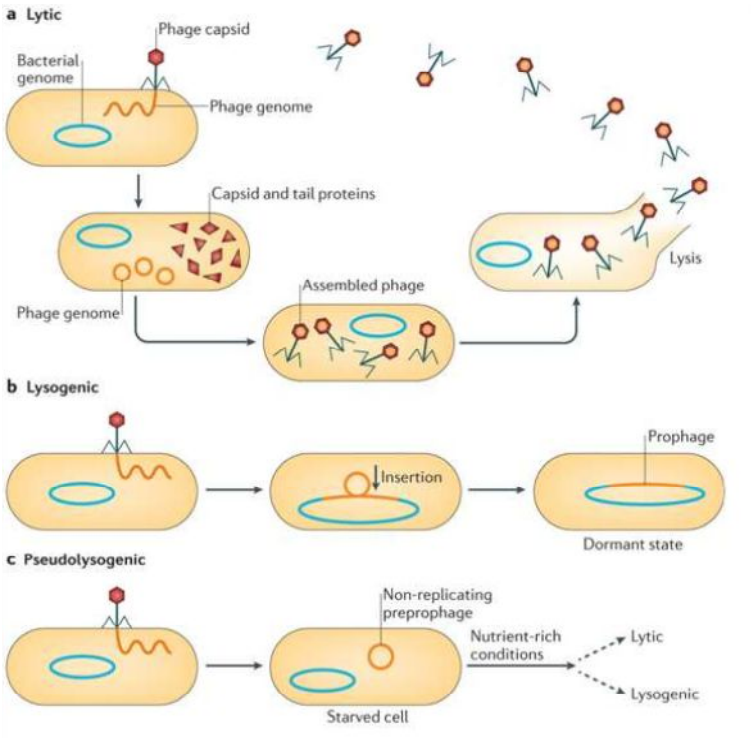


Figure 1.4. Phage replication cycles. A) Lytic cycle - phages immediately enter a productive cycle. B) Lysogenic cycle - phage genome is integrated into bacterial genome. C) Pseudolysogenic cycle - phage genome remains as non-integrated preprophage until nutritional status is restored (adapted from (39)).

Temperate phages contribute to the evolution of most bacteria, by promoting the transduction of various genes involved in virulence, fitness, and antibiotic resistance (38). Despite the putative bacterium–phage evolutionary conflict, phages profit from promoting the survival and proliferation of their hosts (39). Likewise, prophages may harbor cargo genes, or "morons", which while are not essential for the phage, benefits the host. Some very well known lysogenic phages carry genes that enhance the virulence of the bacterial host (40). In addition, the deletion of prophages from *E. coli* revealed that prophages improved the surviving under adverse

environmental conditions, including acid stress or early biofilm formation (41). Prophages may therefore work as gene reservoirs, many of which benefit pathogens, in ways which are only just beginning to be determined (42). In a hostile environment like the human stomach, any metabolic advantage or resistance/tolerance mechanism provided by prophages should be important in improving bacterial host competitiveness. Prophage induction may also be used as a weapon for colonizing new niches (43), displacing native strains, although this strategy may be rarely used, first by the creation of lysogens in the susceptible population, second by the cost of cell lysis in a fraction of the population, and third due to the purifying selection of prophages (44). Taken together, these properties may explain why prophages are more frequent in pathogenic bacteria (45). Host-prophage driven selection and genetic flux occurs even for prophage genes that do not effect host physiology (44). Thus, the role of prophages in disease establishment is being progressively acknowledged.

1.1.6 *H. pylori* phages and prophages

One of the remarkable characteristics of *H. pylori* is the extensive genetic diversity between different strains (20, 46, 47). This diversity has been attributed to an elevated high mutation rate, impaired DNA repair, lateral DNA transfer and frequent recombination events (48). Horizontal gene transfer takes place by direct transfer between two bacteria temporarily in physical contact (conjugation), transfer of a naked DNA fragment (transformation) and transport of bacterial DNA by phages (transduction). Conjugation (48, 49) and transformation (50, 51) have already been described for *H. pylori*, but not transduction. There are about 10^{31} phages on the planet, with phages exceeding bacteria in number by tenfold, but less than an estimated 1% have been described (52). Bacteriophage description in *H. pylori* is brief in the literature. The first descriptions of *H. pylori* phages came from the observation of micrographs where particles compatible with phages were observed (Figure 1.5) (53-57).

The development of the genomic studies, especially using high-throughput genome sequencing led to the first reports of prophages, some remnant (58), others apparently complete and capable of going through a lytic cycle (59-63). A screening for prophages in public available genomes of *H. pylori* revealed the presence of prophage sequences ranging from 5.5Kb to 39.3 Kb (64). Strains carrying prophages do not appear to have a higher pathogenicity or association with particular disease patterns (59, 65), but it has been suggested that the presence of phage orthologous genes correlates with the presence of *cagA* and/or *vacA* virulence genes (66). Despite the putative bacterium–phage evolutionary conflict, phages profit from promoting the survival and proliferation of their hosts (39).

Prophages and bacteria are linked by a long history of co-evolution, but the genetic dimension of this co-evolution cannot be defined at present (52). Indeed, a phylogenetic analysis of the integrase gene sequences present in *H. pylori* prophages revealed a strong phylogeographic signal within this phage gene, which was in agreement with a model of co-evolution between the virus and its bacterial host. The presence of prophages in other non-pylori *Helicobacter* species, such as *Helicobacter acinonychis* (67), *Helicobacter felis* (68), or *Helicobacter bizzozeronii* (69) points to a prophage acquisition before speciation. The presence of remnant prophages (prophage fragments) in *H. pylori* strains (70, 71) and in non-pylori Helicobacters (72), indicates a prophage decay during the complex interaction between *H. pylori* and the prophage. However, a model in which *H. pylori* strains from different geographical regions may have been infected by distinct phage lineages after the geographic separation of the bacterial host is also feasible, but less probable (65).

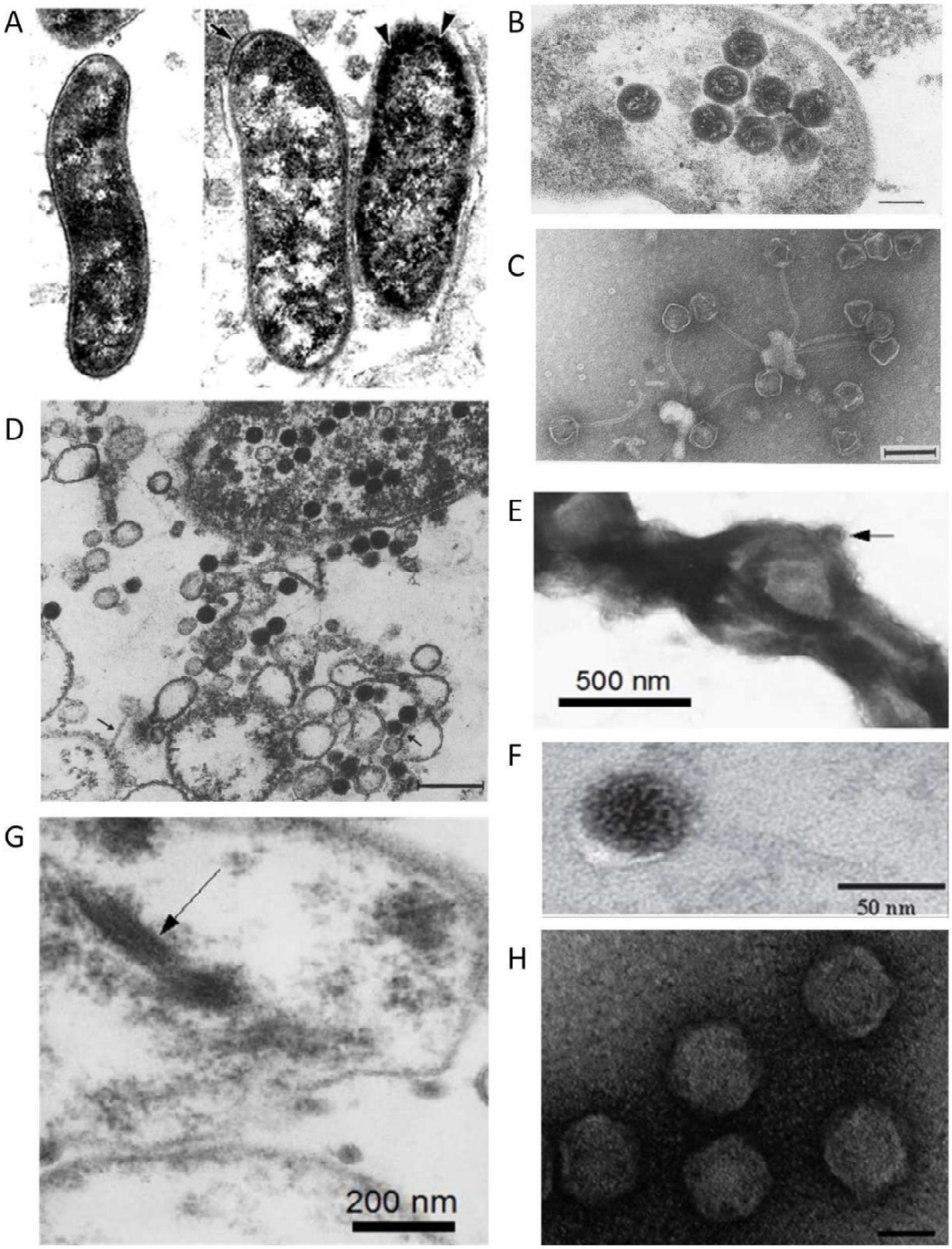


Figure 1.5. Electron microphages of *H. pylori* phages. A. Phage-like particles (arrow heads) within *H. pylori* (adapted from (53)) found in antral tissue. B. Thin section of *H. pylori* carrying phage particles. Bar = 100 nm (adapted from (54)). C. Thin section of *H. pylori* evidencing cell with empty and filled phage heads. Phage head diameter of about 50 nm and tail with 170 nm and a diameter of 9.5 nm. Arrow points phage tail in extracellular phage. Bar = 200 nm (adapted from (56)). D. Negative staining of *H. pylori* phage HP1. Bar = 100 nm (adapted from (57)). E. Negative staining of *H. pylori* cell showing a polyhedral phage-like particle (adapted from (55)). F. Thin section of *H. pylori* phage phiHP33G, presenting a total length of 150 nm, a phage head diameter of 62.5 nm (± 7.3 nm), and a tail with 92.4 nm (± 2.97 nm) long and 5 to 6 nm in diameter (adapted from (59)). G. Hypothetical 15 nm phage filamentous in *H. pylori* cell (adapted from (55)). H. Negative staining of intact *H. pylori* KHP30 phage particles with no tail and head diameter of 68.8 nm (± 2.3 nm). Bar = 50 nm (adapted from (62)).

Interestingly, like their host, *H. pylori* prophages also present a phylogeographic distribution. The population to which prophages belong is determined by prophage sequence typing (PST), which targets two prophage genes (integrase and holin) of *H. pylori* and applies a Bayesian

clustering analysis for the identification of distinct genetic populations. The prophage genes used by the PST method are the integrase (responsible for the integration of the phage genome into the bacterial chromosome) and holin (involved in cell lysis when a lytic cycle occurs). Currently there are 4 prophage populations described (Figure 1.6), hpAfrica1, hpEastAsia, hpNEurope and hpSWEurope (65).

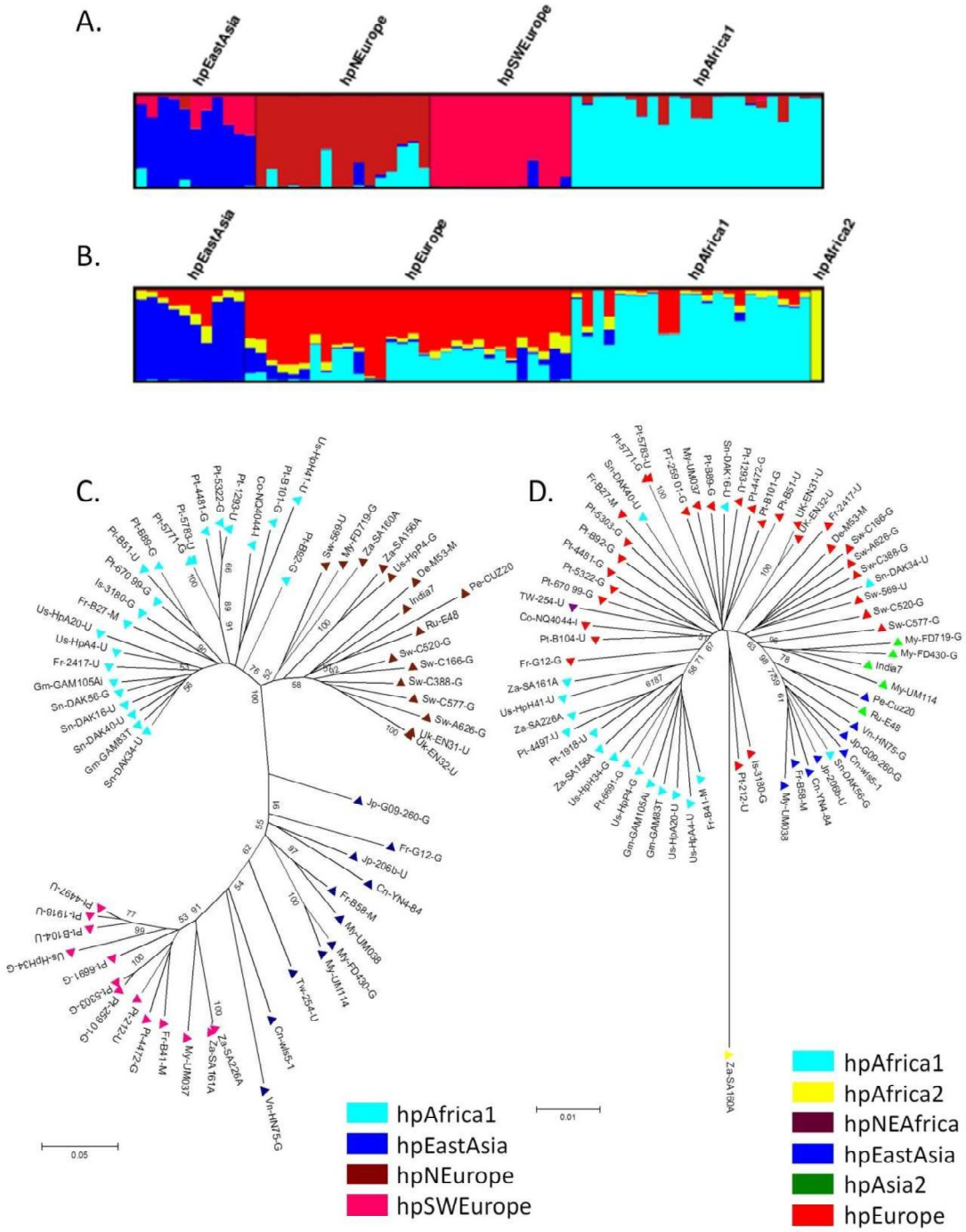


Figure 1.6. Population structure of *H. pylori* prophages. A. DISTRUCT plot of Bayesian population assignments using STRUCTURE and an admixture model (K= 4) for prophages genes (PST). B. Same methodology used for the sequences of seven housekeeping genes (MLST). Each bacterial isolate is depicted by a thin vertical line, which is divided into K colored segments representing the membership coefficients in each cluster. C. Neighbour-joining tree (Kimura 2-parameter) of concatenated prophage sequences. D. Neighbour-joining tree (Kimura 2-parameter) of concatenated sequences of MLST genes. In C and D the strains are colour-coded according to the population assignment by STRUCTURE using PST and MLST genes, respectively (adapted from (65)).

Importantly, the European *H. pylori* population (hpEurope), could not be discriminated using the MLST method, was separated into two different populations (hpNEurope and hpSWEurope) using these two prophage genes. A phylogenetic tree using the neighbour-joining method and the Kimura 2-parameter for the concatenated prophage genes showed that strains clustered according to their population assigned by STRUCTURE software (65).

1.2 Bioinformatics for Next Generation Sequencing

1.2.1 DNA sequencing chemistry

Next generation sequencing (NGS) technology is transforming molecular biology, enabling unprecedented parallelization of sequencing reactions with relatively rapid turnaround time and decreasing costs (73, 74). DNA sequencing is classified in first-, second- and third-generation sequencing. First generation sequencing is Sanger sequencing, where the terminator 2',3'-dideoxynucleotides (ddNTP) are used (75), the other two methods are referred as NGS.
Second-generation sequencing involves massively parallel sequencing of number of templates of same sample in a single run, producing a massive volume of short read length. Third-generation produces longer read length at low cost, preventing amplification artifacts and bias (73).
This literature review focuses only in the sequencing technology used in this work, *i.e.* Illumina sequencing platform. The other technologies are not covered, although interested readers may find further information in review papers (73, 76, 77).
The Illumina sequencing workflow is composed by sample preparation, cluster generation, sequencing and data analysis (Details on the sequencing chemistry were retried from Illumica, https://www.illumina.com/techniques/sequencing/dna-sequencing.html, consulted March 2017).
Sample preparations begins with tagmentation in which transposomes simultaneous fragment and covalently tags the target DNA with adaptors. Once adaptors have been added reduced cycle amplification adds additional motifs, including sequencing primer binding sites (one at each fragment end), indexes and regions that are complemented to the flow cell oligonucleotides. The flow cell is a glass side with lanes. Each lane is a channel coded with a lawn composed of two types of oligonucleotides. Clustering is a process were each fragment molecule is isothermally amplified. Hybridization occurs when the first of two types of oligonucleotides on the surface of the flow cell anneal to the complementary adapter region in one of the fragment strands. A polymerase creates a complement of the hybridized fragment and this double stranded molecule is denaturated and the original template is washed way. The strands are subjected to clonal amplification by bridge PCR. In this process the strands fold over and the adapter region hybridizes with the second type of oligonucleotide on the flow cell. Polymerases generate the complementary strand forming a double stranded DNA bridge. This bridge is then denaturated, resulting in two single stranded copies attached to the flow cell. The process is then repeated and occurs simultaneously for millions of clusters resulting in clonal amplification of all fragments. After bridge amplification the reverse strands are cleaved and wash off, leaving only the forward strands. The 3' ends are blocked to prevent unwanted priming. Sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle fluorescent tagged nucleotides protected at 3′-OH groups (2-cyanoethyl) compete for addition to the growing chain. Natural competition between all the four nucleotides present during each sequencing cycle reduces the inherent bias as compared to pyrosequencing (used by Roche 454) where only one type of nucleotide is made available at a

time for pairing (73, 74). This sequencing by reversible termination consists of three steps: incorporation of the complementary nucleotide by mutant DNA polymerase to the DNA strand attached with flow cell, exciting each a light source and detection of the different fluorescence signal for the four bases, and restoration of free 3′OH group by cleaving the terminating moiety and reporter molecule. This proprietary process is called sequencing by synthesis. The number of cycles determines the length of the read. For a given cluster all identical strands are read simultaneously. Hundreds of millions of clusters are sequenced in a massive parallel process. The emission wavelength along with the signal intensity determines the nucleotide base assignment base call, referred to as base calling. Base calling is scanned for error probability by PHRED software. PHRED reads DNA sequence chromatogram files and assigns PHRED quality scores by examining the peaks around each base call. PHRED score varies from 4 to 60 with better quality and low error probabilities associated with higher values (73, 78). The Illumina 1.3+ encodes PHRED scores with an ASCII offset of 64, and so can hold PHRED scores from 0 to 62 (ASCII 64–126), although currently raw Illumina data quality scores are only expected in the range 0–40 (79). Only Illumina NGS platforms are capable of paired-end sequencing as the clonal amplification here is done by bridge PCR. Illumina HiSeq and MiSeq have a four–channel sequencing system wherein each base is detected by individual image. On the other hand, Illumina NextSeq 500 has only two-channel SBS technology that requires only two images to determine all four base calls, reducing the image capturing time, cost of sequencing and time required for data processing (80).
The pitfalls of second-generation sequencing are first the short read length that need to be assembled with the help of various bioinformatics tools/pipelines into original length template and second to PCR bias introduced by clonal amplification, for detection of base incorporation signal (73, 80). The error rate of Illumina sequencing equipments varies from 0.26% (Hiseq) to 0.8% (MiniSeq, MiSeq, NextSeq) (73).

1.2.2 Bioinformatics importance for genomic research

High-throughput genomics is joining the big-data club as massive amounts of sequencing data keep to be produced. From 1982 the number of bases in Genbank double approximately every 18 months (Figure 1.7).

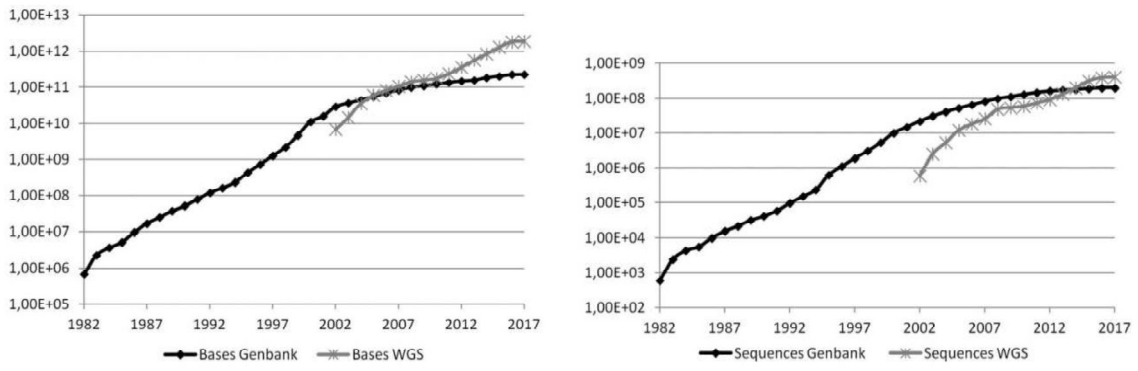


Figure 1.7. Growth of Genbank database from 1982 to 2017. The graphics were produced with the numbers available for the last month of each year, except for 2017 in which the data corresponds to the NCBI release of February (GenBank and WGS Statistics, https://www.ncbi.nlm.nih.gov/genbank/statistics/, consulted in March 2017). Left graphic evidences the growth of bases available in Genbank and whole-genome sequencing (WGS). Right graphic shows growing numbers for sequences from Genbank and WGS.

The number of bases available by WGS surpassed the number of bases available by Genbank in 2004, while the number of sequences exceeded the ones of Genbank in 2014. WGS projects typically based on NGS. For downloading purposes, the uncompressed GenBank Release 218.0 requires approximately 818 GB (Genbank realase 218, February 2017). The European Bioinformatics Institute (EMBL-EBI, www.ebi.ac.uk) continually expands its total disk capacity to keep pace with demand, storing 75 petabytes (1 petabyte is 10^{15} bytes) as of December 2015 (81).
Besides the problem of archiving big sequencing data, in order to obtain answers to scientific questions out off sequencing data bioinformatics tools are absolutely necessary. Software and web servers solutions for genomic analysis are constantly being developed. Choosing the appropriate ones will always depend on the answer that is pretended, the computing power available and the bioinformaticists skills.

1.2.3 Genome assembly

NGS is accelerating biological research in many areas such as genomics, transcriptomics, metagenomics, proteogenomics, gene expression analysis, noncoding RNA discovery, SNP detection, identification of protein binding sites, among others (82, 83). NGS produces reads that theoretically cover all the genome. The genome must then be reconstructed by joining these reads together, a process known by genome assembly. The reads are joined together to form longer contiguous reads known as contigs by the assembler software. The contigs are then joined together to form longer contigs known as scaffolds (composed by contigs and gaps). Genome assembly can be obtained by a comparative approach (reference-based assembly) against a reference genome from the same organism or a closely related species that is used as a map of the assembly process, or can be *de novo* assembled, *i.e.* without the use of any reference genome, which allows to reconstruct a genome that has not be previously sequenced (82, 84).
The input file for assembly algorithms has the sequence reads and their quality scores. The usual format is FASTQ, that has 4 lines for sequence. The first line starts with @ and contains the sequence description, the second line contains the sequence itself, the third line starts with + sign with optional description, and the fourth line has the quality score of each base in the sequence according to an alphanumeric code (79).
The four basic stages of NGS genome assembly are preprocessing filtering, graph construction process, graph simplification process, and postprocessing filtering (85). Before the beginning of the assembly the preprocessing filtering takes places, detecting and correcting erroneous reads. These errors are caused by the sequencing platforms and include substitutions (mismatch), indels (insertion/deletion), and ambiguous bases (i.e., N). Moreover, low-frequency reads are also checked by correction algorithms, being aligned with high-frequency reads that share substrings. Absence of uniform genome sampling and high-frequency genomic repeats difficult the analysis. The correction approaches are based in K-spectrum, Suffix Tree/Array, multiple sequence alignment (MSA) or hybrid approaches (Figure 1.8) (82). Velvet, the package used in the present work, removes erroneous reads (86, 87). After preprocessing, the graph construction process creates a graph model that is used to organize short-read sequences into a compact form and to create longer reads during assembly. A graph is a set of nodes (vertices) plus a set of edges (arcs) between the nodes. The reads are combined to form longer contiguous reads, referred to as contigs. These combined reads share nucleotides at their ends, i.e., merged reads share an overlap region (Figure 1.8). The graph construction can be classified as overlap-based (overlap graphs), k-mer–based (de Bruijn graphs), greedy-based (greedy graphs), and hybrid-based construction (hybrid graphs).

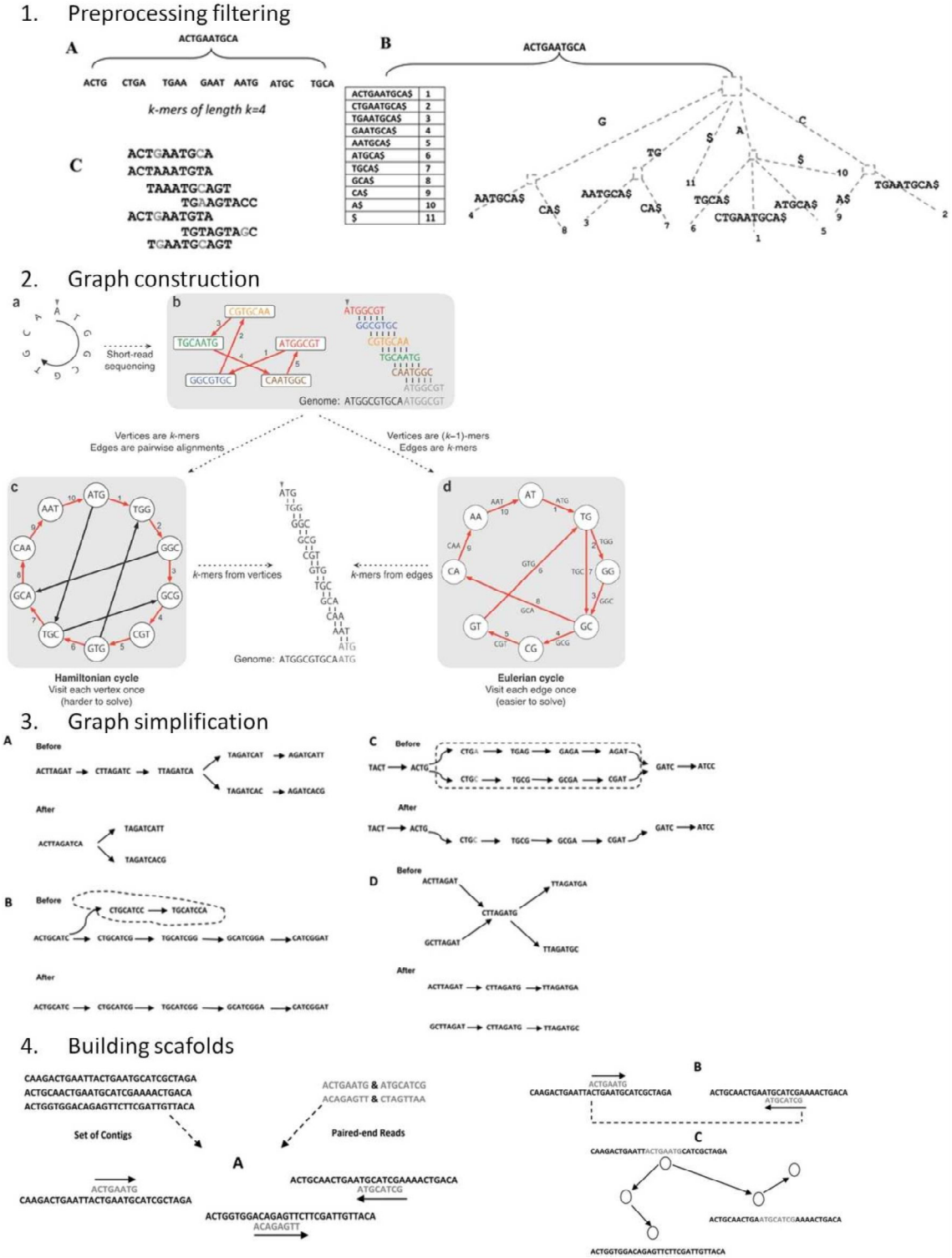


Figure 1.8. Genome assembly steps. 1) Preprocessing filtering. A. In K-spectrum approach a set of substrings of fixed length k are extracted from the read and ready to filter. B In suffix tree/array approach a set of substrings of different lengths of k (suffixes) are extracted from the read, represented in the suffix tree, and ready to filter. C. In multiple sequence alignment reads are aligned to each other to define consensus bases and correct erroneous ones (adapted from (87)). 2) Graph construction. A. Small example circular genome. B. For Sanger sequencing algorithms, reads were represented as nodes in a graph, and edges represented alignments between reads and following the edges in numerical order allows the reconstruction of the circular genome by combining alignments between successive reads (the repeated part of the sequence is grayed out in the alignment diagram). C. The read is split into all possible k-mers: with k = 3, ATGGCGT comprises ATG, TGG, GGC, GCG and CGT. Following a Hamiltonian cycle (indicated by red edges) allows the reconstruction of the genome by forming an alignment in which each successive k-mer (from successive nodes) is shifted by one position. D. Construct of de Bruijn graph representing all k-mer prefixes and suffixes as nodes and drawing edges that represent k-mers having a particular prefix and suffix. Finding an Eulerian cycle allows the reconstruction of the genome by forming an alignment in which each successive k-mer

(from successive edges) is shifted by one position, generating the same cyclic genome sequence without performing the computationally expensive task of finding a Hamiltonian cycle (adapted from (88)). 3) Graph simplification operations. A. Consecutive nodes are merged. B. Dead end (dotted circle) is removed. C. Bubble (dotted circle) is simplified removing low-coverage path. D. X-cut is simplified by splitting the connections into two parallel paths (adapted from (87)). 4) A. Paired-end reads are aligned to contigs and their orientations determined. B. The library insert size (dotted line) is determined between two pairs and compared with the one saved previously. C. Contig connectivity graph is constructed and filtered according to paired-end constraints.

For instance, Velvet uses a de Bruijn graph, a compact representation based on short words (k-mers) (86), while overlap graphs are used for *de novo* assembly, consisting in three stages: overlap, layout, and consensus (*i.e.*, OLC). From the Greek *mer* means part, a k-mer is a substring of length k (82). The first assemblies used the greedy algorithm, which has one basic operation consisting in adding one more read or contig to a given read or contig. The basic operation is repeated until no more operations are possible. In the case of false-positive overlaps unrelated sequences may be joined to either side of a repeat to produce chimera. OLC and de Bruijn graph are two robust approaches to assembly, both relying on a set of overlaps between the input reads, represented in a directed graph (84). The third step, the graph simplification process, is used to simplify the graph by reducing the number of graph nodes and edges, and removing erroneous ones, in order to overcome memory limitations and reduce computation time. Finally, postprocessing filtering builds contigs, detects misassembled ones, and extends them into scaffolds. Paired-end reads are used to order and orient contigs during the scaffolding process (Figure 1.8). The contigs are joined together depending on the positions of the paired-ends in the contigs, their orientation, and expected insert size. Unfiled gaps between contigs are filled with N characters, denoting unknown bases between them.

The quality of the assembly is evaluated by the size and accuracy of the contigs and scaffolds. Assembly accuracy is difficult to measure, but an alignment to a reference sequence is useful whenever trusted references exist. The statistics for assembly size include maximum length, average length, combined total length, and N50. The contig N50 (Figure 1.9) is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly, i.e., 50% of the genome is in contigs as large as the N50 value (84).

The main limitation of assembly software is that read length is much shorter than the genome length. The assembly of short reads requires high coverage to assure minimum detectable overlap criteria, which in return increases computational complexity. Shorter reads have less power to resolve genomic repeats but higher coverage increases the chance of spanning short repeats (84, 89). Unresolvable repeats are left out and break the assembly in fragments (contigs). The existence of repeat sequences in the target, especially if the repeats are longer than the reads, can turn the repeats impossible to differentiate. High coverage helps repeat separation, but may be confounded by high sequencing error. If the repeats that are inexact, high-stringency alignment can separate the repeat copies. Single reads that span a repeat with a sufficient unique sequence on either size of the repeat, known as spanners, improve repeat resolution. Thus, pairs that are on both sides of the repeat with each end in unique sequence, and pairs with exactly one end in the repeat contribute for resolution of the repeat sequence. Genome coverage is not uniform for all contigs. Coverage variation may be induced by chance, by bias of sequencing technologies, and by variation in cellular copy number between source DNA molecules, for instance the existence of plasmids of high copy number. Very low coverage induces gaps in assemblies (84, 89).

Genome assembly build on NGS are not perfect and choosing different programs may produce different genome sequence from the same departure reads. Genome assembly algorithms remain an active area of innovation.

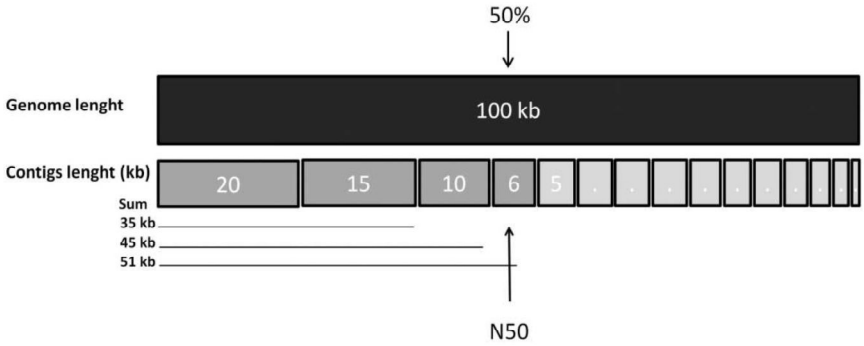


Figure 1.9. N50 statistics for genome assembly. For this example N50=6, meaning that at least 50% of the genome is in contigs as large as 6 kb.

1.2.4 Genome comparison

Comparative genomics is a broad field, larger than NGS alone, whose goal is to identify all differences among genomes, including detection and localization of functional genomic elements, identification of sequence differences responsible for phenotypic shifts in organisms, such as antibiotic resistance and virulence factors (90, 91). After genome assembly each genome is usually annotated. There are several annotation pipelines (92), such as RAST (93). Genome annotations starts by identifying genes, or more precisely open reading frames (ORF), *i.e.*, identifying start and stop positions in same frame of the prokaryote genome, along with function identification. Predicting ORF is done by softwares like Glimmer (94) or GeneMark (95). The next step consists in using these predictions and search databases, like Genbank (96) and SwissProt (97) using mainly BLAST (98), or other programs. The accuracy of this step depends of the annotation software and the quality of the annotations already in the reference database (92, 99). In order to efficiently decrease annotation error comparison of results from multiple annotation services should be performed, interchanging information between annotation services (99).

The comparison of genomes lead to development of the core and pangenome concepts, corresponding to the set of genes found in all genomes and the entire gene set of all genomes, respectively. The core genome tends to diminish as more genomes from the same species (or genus, etc.) are sequenced, while the remaining accessory genes tend to increase. Comparing the genes in a selection of genome sequences depends on a reliable annotation of orthologs, *i.e.*, genes of different organisms that arose from a single ancestral gene via speciation. The core genome usually represents the housekeeping functions require for survival, while the accessory genome is linked to strain-specific phenotypes. Species may have an open pangenome that continues to increase as new strains are sequenced, or otherwise a finite pangenome (91). For instance, the core genome of *H. pylori* have been determined and, depending of the studies was established in 1063 genes (100), 1281 genes (101), 1111 genes (102), or 1193 genes (103). In each *H. pylori* strain the core genome represents about 80% of the genes (103). Core and pan genomes studies provide insights in pathogenic traits and are also useful for phylogenetic diversity studies (104). Synteny conservation is another form for comparing closely related genomes. Another definition of core genome could be the existence of a set of genes within syntenically conserved blocks in different strains, likely corresponding to conserved genes in the common ancestor that were inherited mainly through vertical transfer. In opposition, many of the genes acquired by horizontal gene transfer (HGT) should be inserted into unpredictable positions (105).

Single nucleotide polymorphisms (SNP) and small insertions or deletions (indels) are known to inactivate genes involved in virulence, while the acquisition of novel functions has also been shown to be involved in bacterial pathogenicity (91). Analysis of specific core genes, identifying SNPs, have also been used for phylogenetic diversity studies (91), highlighting for instance a phylogeographic distribution in *H. pylori* (23). However, phylogenetic analysis based on gene content is affected by horizontal gene transfer and recombination (105).
Multiple sequence alignment allows comparison of multiple homologous genes to derive conserved segments and evolutionary trees. The notion of distance between two nucleotide sequences can be derived as the number of mismatches derived after pair-wise alignment of two sequences, or as a evolutionary distance between two microorganisms. The technique is based upon progressive pair-wise comparison to make intermediate alignments between nearest neighbors (106). Several multiple sequence alignments have been described, including Clustal (107), MAFFT (108) and others.

1.2.5 Phage detection

Identification of phage genes is similar to any other gene annotation process described above. There are, however specific packages and web servers for detecting and inferring prophage presence and completeness. Search for homologous may be done in relevant protein viral databases, like Phantome (http://www.phantome.org/), PHAST/PHASTER database (109), and VirSorter (110). These packages include the first ones developed, namely Prophage Finder (111), Phage_Finder (112) and Prophinder (113). Recently PHAST and PHASTER were developed performing 40 times faster and presenting results up to 15% more sensitive in comparison with the previous ones (109, 114). Other prophage identification tools have been described, such as PhiSpy (115) which analyzes several other sequence-based statistics to help identify novel phages (AT and GC skew, protein length and transcription strand directionality) that are not represented in existing phage databases, and VirSorter (110) that handles metagenomic data with improved performance for fragmented genomes. These tools may run locally like (PhiSpy, Phage_Finder, Prophage Finder), to access through a web-server (PHAST/PHASTER), or made available through cyberinfrastructure (iPlant Discovery Environment (116), a Web portal of iPlant's cyberinfrastructure that houses several apps for sequencing analysis and other data-intensive technologies) that provides a web-based user interface.

1.3 Aims

Despite the recent discovery of the importance of prophages in the diversity of *H. pylori* (59), they remain poorly characterized. The lack of information on bacteriophages of *H. pylori* prompted this study. Based on the presence of the prophage integrase gene we determined that an estimated 20% of *H. pylori* strains carry prophages (59, 65). Based on PCR screening, we compiled a collection of *H. pylori* strains carrying prophages (65). We therefore undertook a more holistic approach, using the NGS technology to study the full genome of strains from this collection as well as *H. pylori* strains presenting prophages found in public databases. This data is useful first for identification of phage sequences, and second for comparative genomics. These results contribute to increase the knowledge of *H. pylori* prophage genomic organization, insertion sites, phylogeography, and diversity. The detailed genomic structure of 28 prophages described here will provide in the future an important basis to identify the function of prophage genes and to verify if prophages provide advantageous phenotypes.

2. Material and methods

2.1 *H. pylori* strains

A total of 28 *H. pylori* strains carrying prophages were analyzed (Table 2.1). These included 15 strains isolated from patients with gastritis, nine from peptic ulcer patients, three from MALT patients and one from gastric cancer patient. The present study included strains from Portugal (n=14), France (n=6), Sweden (n=4), UK (n=2), Germany (n=1) and Israel (n=1). Prior to each assay, bacteria were grown in *H. pylori* selective medium (Biogerm, Portugal) at 37ºC in a microaerophilic environment (Anoxomat®, MART Microbiology BV, The Netherlands) for 24h to 48h. Total DNA was extracted using the QIAmp DNA Mini Kit (Qiagen, UK) according to the manufacturer's instructions.

Table 2.1. Description of the 28 *H. pylori* strains that harbor prophages.

Strain	Country of isolation	Disease associated
UK-EN31-U	United Kingdom	Peptic Ulcer
UK-EN32-U	United Kingdom	Peptic Ulcer
De-M53-M	Germany	MALT lymphoma
Sw-577-G	Sweden	Gastritis
Sw-A626-G	Sweden	Gastritis
Pt-B89-G	Portugal	Gastritis
Pt-1293-U	Portugal	Peptic Ulcer
Fr-ANT170-U	France	Peptic Ulcer
Fr-MEG235-U	France	Peptic Ulcer
Pt-5771-G	Portugal	Gastritis
Pt-5322-G	Portugal	Gastritis
Pt-228_99-G	Portugal	Gastritis
Pt-1846-U	Portugal	Peptic Ulcer
Pt-B92-G	Portugal	Gastritis
Pt-4481-G	Portugal	Gastritis
Fr-GC43-A	France	Gastric adenocarcinoma
Fr-G12-G	France	Gastritis
Fr-B58-M	France	MALT lymphoma
Pt-212-99R-U	Portugal	Peptic Ulcer
Pt-1918-U	Portugal	Peptic Ulcer
Pt-4497-U	Portugal	Peptic Ulcer
Pt-4472-G	Portugal	Gastritis
Fr-B41-M	France	Gastritis
Sw-C388-G	Sweden	Gastritis
Sw-C520-G	Sweden	Gastritis
Is-3180-G	Israel	Gastritis
Pt-259-G	Portugal	Gastritis
Pt-5303-G	Portugal	Gastritis

2.2 Whole-Genome Sequencing

Genomes were sequenced at the National Institute of Health, Lisbon, Portugal, with exception of four strains (Sw-577-G, Sw-A626-G, Sw-C388-G and Sw-C520-G) that were sequenced at Karolinska Institute, Stockholm, Sweden, and four strains (Fr-ANT170-U, Fr-MEG235-U, Fr-GC43-A and Fr-B41-M) that were sequenced at the Institute of Life Sciences, College of Medicine, Swansea, Wales, UK.

For genomes sequenced in Portugal and Sweden, the yield and integrity of the purified DNA were then assessed through a Qubit assay (Quanti-it dsDNA Assay Kit, Broad Range; Lifetechnologies, Paisley, CA, USA) and agarose gel electrophoresis (0.7% gel), respectively. High-quality DNA samples were then applied to prepare Nextera XT Illumina paired-end libraries. These were subsequently subjected to cluster generation and paired-end sequencing (2x250 bp, 2x150 bp and 2x100 bp) by using the Illumina MiSeq (Portugal) and HiSeq 2500 (Sweden) platforms (Illumina Inc., San Diego, CA, USA), according to the manufacturer's instructions.

The number of passing filter reads obtained per sample ranged from 0.6-2.7 million reads. The FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tools were applied to evaluate and improve the quality of the raw sequence data, respectively. Subsequently, high-quality reads were *de novo* assembled using Velvet (version 1.2.10) (86). Several assemblies using different k-mer sizes were run. The best assembly was assumed as the one with the best cumulative ranks for N50, number of contigs/scaffolds, and length of the largest contig/scaffold. The obtained mean depth of coverage ranged from 135- to 195-fold. The final contigs/scaffolds were visually inspected (using Tablet 1.14.04.10) (117) and corrected.

For genomes sequenced in UK, quantification of DNA was assessed after DNA extraction with a Nanodrop spectrophotometer, as well as the Quant-iT DNA Assay Kit (Life Technologies) prior to sequencing. High-throughput genome sequencing was performed using a HiSeq 2500 machine (Illumina Inc.), and the 100 bp short read paired-end data was *de novo* assembled using Velvet (version 1.2.08) (86). The VelvetOptimiser script (version 2.2.4) was run for all odd k-mer values from 21 to 99 (several assemblies using different k-mer sizes were run), with all program settings unchanged apart from a minimum output contig size set to 200 bp and the scaffolding option switched off.

All genomes were annotated using the RAST server (http://rast.nmpdr.org/) (118), the NCBI Prokaryotic Genomes Annotation Pipeline version 2.3. and PHAST web server (109). The respective trimmed reads were submitted to the Sequence Read Archive (SRA).

2.3 Determination of depth coverage

The theoretical coverage is the average number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length, assuming that reads are randomly distributed across the genome (119). The Velvet assembler provides all coverage values in k-mer coverage, *i.e.* how many times has a k-mer been seen among the reads. The relation between k-mer coverage Ck and standard (nucleotide-wise) coverage C is $Ck = C * (L - k + 1)/L$ where k is the hash length, and L the read length (86). To compute the depth coverage for each genome a script was developed in Python 2.7 (Appendix 1).

2.4 Assembly of prophage genomes

For prophage identification two strategies were taken. First, the PHAST web server (109) was used to identify putative prophages within contigs of each *H. pylori* genome. Second, MEGABLAST (98), with a word size of 28, was used to align the genome of *H. pylori* phage KHP30 or phiHP33 with the contigs of each sequenced *H. pylori* genome. PHAST analyses (http://phast.wishartlab.com/) applied over contigs allowed us to check homology, and to identify, annotate and graphically display prophage sequences, providing information on prophage completeness, categorized as either intact, incomplete, questionable or not detected. MEGABLAST was run using KHP30 or phiHP33 as reference since these prophages genomes were the most commonly found to be similar with the prophages detected by PHAST.
The MEGABLAST analysis results were particularly useful to determine which contigs were from phage origin and the order in which they probably appear. Based on this predicted contig order, primers flanking the contigs were designed, using primer3 v. 0.4.0 (120), to bridge gaps in the assembly in order to close the gaps (the gaps were of few bases to about five hundred bases). The PCR mix included Promega (Madison, WI, USA) buffer (1X), dNTPs (0.2 µM), primers (0.5 µM each), GoTaq polymerase (1.5 U), water to complete 25 µl and DNA sample (25 to 50 ng). The PCR cycle was composed of a first cycle at 95ºC for 4 min, 35 cycles at 95ºC for 30 sec, 59ºC for 30 sec and 72ºC for 1 or 2 min. A last cycle at 72ºC for 7 min was applied. The PCR products were purified using MicroSpin S-400 or S-300HR columns (GE Healthcare, Velizy-Villacoublay, France) and directly sequenced on both strands using an external sequencing service provider (Eurofins Genomics, Regensburg, Germany, and Stabvida, Lisbon, Portugal). A multiple sequence alignment (121) was carried out using flanking parts of the contigs and the PCR sequenced product after assembly of the forward and reverse sequences.
The insertion sequences of the prophages were identified whenever the prophage 5' and 3'ends were contiguously flanked by bacterial genes in a contig. The last bacterial gene before the prophage sequence and the first bacterial gene after the prophage were identified as well as the homologous locus_tag for the reference genome *H. pylori* J99 (122). The presence of repeated sequences at prophage insertion sites was verified using Tandem Repeat Finder (123) (available at https://tandem.bu.edu/trf/trf.basic.submit.html).

2.5 Comparative genomic analyses of prophages

2.5.1 Genome Annotation

The assembled prophages were analyzed using PHAST to provide a first annotation. The annotation of prophage genomes was carried out further using Phages v. 1.0 (http://www.phantome.org/PhageSeed/Phage.cgi?page=phast), and RAST (118). The annotation of coding sequences (CDS) found by the three different methods were compared.
The annotation of both *H. pylori* India7 (accession number CP002331) and Cuz20 (CP002076) prophages, as well as that of the *Helicobacter* 1961P (NC_019512.1), KHP30 (NC_019928.1), KHP40 (NC_019931.1), phiHP33 (NC_016568.1) phages, were used for comparative purposes.

2.5.2 Multiple Sequence Alignment

Sequence alignment provides mean of comparison that is useful to establish evolutionary relationships, identify conserved patterns and find similar domains, which could imply similar function. The annotated prophages were aligned using the progressive Mauve algorithm software (version 2.3.1) (124), to check the order of the CDS in the prophage genomes and the

existence of a consensus sequence. In order to infer the phylogenetic relationships among prophages, the intact genomes of the 23 prophages identified in the present study, were aligned using MAFFT version 7 (108) together with other six phage *Helicobacter* genomes available at public databases (1961P, KHP30, KHP40, phiHP33, *H. pylori* India7, and *H. pylori* Cuz20) as well as with the *H. acinonychis* (accession number NC_008229.1).

2.5.3 Prophage core and pan genome

In order to select prophage ORF found in all prophages genomes (core prophage genome) and all ORF found in this prophages a two step protocol was followed. The first step was to compile all information gathered from PHAST web server (109), *i.e.*, each ORF of each prophage presenting homology with ORFs of the phage KHP30 received its locus_tag identification (systematic, stable identifier for a gene and its associated features, used for tracking purposes) that for simplification purposes was numbered from 1 to 30 (phage KHP30 has 30 ORFs). Homologous ORFs are detected by PHAST (109) if the BLAST hit against the virus and prophage database for matched phage sequences has an e-values less than 10^{-4}. The ORFs left unnumbered were used in the next step. For this a database with all prophage sequences was build with the FASTA files of the prophages sequences and a local BLASTN using the prophage database was conducted for each unnumbered ORF. BLASTN hits found having a threshold limit of $<1e^{-6}$ and a query coverage >70% were given the same ORF number starting at 31. The process was conducted until all ORFs were numbered. To compute the prophage core and pangenome a script was developed in Python 2.7 (Appendix 2), that found common ORFs to all prophages belonging to a p group, as well as the total number of ORFs present in the p group. For instance, if the p group is formed by prophages A and B, the core genome of A and B is composed of orthologous ORFs that are simultaneously present in A and B, and the pan genome is composed of the sum of ORFs present in A and B. Repeated ORFs were only counted once. Adding more genomes produces an increasing of the pan genome and a reduction of the core genome. The selected p group were all the 30 complete prophage genomes, and the p groups for prophages belonging to SWEurope, NEurope, Africa1 and EastAsia. The standard deviation and trend curves showing the change in number of pan and core genome ORFs to the number of genomes were determined. A total of 100 random ordered p groups were used for each determination. The trend curve for the pan and core genome represented the mean number of ORFs presented for 100 possible combinations determinations and its standard deviation. A Venn diagram for the four phylogeographic groups were determined with the R package gplots (125), which allowed for a quick overview on the number of elements that multiple sets share.

2.5.4 Phylogenetic analysis

A nucleotide Neighbour-joining phylogenomic tree was constructed using the MEGA (Molecular Evolutionary Genetics Analysis) 6.0 software (126), with distances estimated using the Kimura two-parameter model (127), which takes into account transitional (interchanges of purines or of pyrimidines bases) and transversional (interchanges of purine for pyrimidine bases) substitution rates. Considering the huge genomic diversity observed among all prophage genomes as well as their different lengths, both complete and pairwise deletion options were used. While the former removes all sites containing missing data or alignment gaps before the distance estimations begin, in the pairwise-deletion, option sites are only removed during the analysis as the need arises. Branching significance was estimated using bootstrap confidence levels by randomly resampling the data 1,000 times with the referred evolutionary distance

model. Core alignments were also used to construct phylogenetic networks implemented in the SplitsTree program (128).

2.5.5 Population structure determination

To determine the population structure of prophages, we use prophage sequence typing (PST), as previously described (65). Briefly, the multi-fasta file with the alignment of integrase and holin gene sequences was converted to the STRUCTURE 2.3.4. (30, 129, 130) program input file using xmfa2structure by X. Didelot and D. Falush (http://www.xavierdidelot.xtreemhost.com/clonalframe.htm). STRUCTURE was used to study the number of K populations using the admixture, performing runs in duplicate. In each run, a Markov Chain Monte Carlo (MCMC) of 10,000 iterations and a burn-in period of 10,000 iterations were chosen. The highest mean value of ln likelihood was compared for multiple runs of $2 \leq K \leq 6$.

2.5.6 Detection of recombination

The existence of putative recombination phenomena within prophage genomes was first evaluated using the Recombination Detection Program version 4 (RDP4) (131) with default settings. RDP4 simultaneously applies different methods for detecting and characterizing individual recombination events that are evident within a sequence alignment without any need for predefined sets of non-recombinant reference sequences. SimPlot software (http://sray.med.som.jhmi.edu/SCRoftware/simplot/) was also used for characterizing with higher detail the genomic mosaicism of the identified recombinant prophages, as previously described for bacterial pathogens (132). The similarity estimations were performed by using the Kimura two-parameter model with sliding window and step sizes that varied according to each recombinant genome.

3. Results

3.1 Bacterial genome assembly

The Illumina FastQ format reads were first analyzed for their quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), in order to obtain a comprehensive quality check report. Usually the beginning and ending of the reads need to be trimmed before proceeding with the analysis of a sequence data. Figure 3.1 shows an example of the output obtained before and after trimming the sequences with FASTX. In Figure 3.1A. left panel it is showed per base sequence quality across all length on the read. In this case the reads are of 150 bases. For each position a box whisker plot is drawn. The red line is the medium value, the yellow box the inter-quartile ranging 25-75%, the upper and lower whiskers the 10% and 90% and the blue line the mean quality. The y-axis shows the quality scores for base call (PHRED quality score). The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read. Figure 3.1.A right plot, shows Per Base Sequence Content that is the proportion of each base position in a file for which each of the four normal DNA bases has been called. This module fail if the difference between A and T, or G and C is greater than 20% in any position, which occurred for the first and last bases of the read. After trimming with FASTX (http://hannonlab.cshl.edu/fastx_toolkit/), this was corrected (Figure 3.1B). Trimmed high-quality reads were *de novo* assembled using Velvet (version 1.2.10) (86) and were then annotated. A summary of *H. pylori* sequenced genomes can be found in Table 3.1.

Figure 3.1. Analysis of FastQC for reverse reads of strain Pt-B89-G. A) Analysis before trimming. Left panel show per-base sequence quality and right panel per base sequence content. B) Results after trimming. Equivalent panels as in A.

Table 3.1. General features of *H. pylori* sequenced genomes.

Strain	GC%	Genome size (Mbp)	number predicted proteins	Mean depth coverage (per base)	PHAST (109) Result#	Bioproject accession number
UK-EN31-U	39.0	1.61	1683	168	I	SRP071274
UK-EN32-U	38.9	1.65	1694	163	I	SRP071276
De-M53-M	38.8	1.65	1695	177	I	SRP064710
Sw-577-G	38.9	1.65	1707	188	Q	SRP071293
Sw-A626-G	38.8	1.67	1709	469	I	SRP071294
Pt-B89-G	39.0	1.62	1651	577	I	SRP071278
Pt-1293-U	39.0	1.63	1694	202	I	SRP071280
Fr-ANT170-U	39.0	1.64	1663	20	Q	SRP072438
Fr-MEG235-U	39.1	1.62	1644	20	Q	SRP072439
Pt-5771-G	39.0	1.67	1687	128	A	SRP064707
Pt-5322-G	39.1	1.58	1621	235	I	SRP071284
Pt-228_99-G	39.0	1.61	1652	39	Q	SRP071067
Pt-1846-U	39.0	1.54	1675	52	I	SRP071062
Pt-B92-G	38.8	1.63	1670	261	I	SRP071282
Pt-4481-G	39.0	1.59	1626	69	I	SRP071279
Fr-GC43-A	39.0	1.61	1634	20	U	SRP072440
Fr-G12-G	38.9	1.7	1711	831	I	SRP064708
Fr-B58-M	38.8	1.57	1630	219	I	SRP071277
Pt-212-99R-U	38.9	1.63	1655	225	I	SRP071292
Pt-1918-U	39.1	1.63	1663	212	I	SRP064706
Pt-4497-U	39.3	1.61	1639	242	I	SRP064709
Pt-4472-G	38.8	1.65	1680	343	I	SRP071271
Fr-B41-M	39.1	1.69	1729	20	I	SRP072441
Sw-C388-G	38.9	1.65	1729	145	Q	SRP071295
Sw-C520-G	38.8	1.67	1698	387	Q	SRP071296
Is-3180-G	39.1	1.57	1588	190	Q	SRP071289
Pt-259-G	39.0	1.59	1626	137	A	SRP071290
Pt-5303-G	38.9	1.65	1672	245	Q	SRP071291

I - Intact; Q - questionable; U - incomplete; A - absent prophage.

3.1 Prophage genome assembly

The contigs of each assembled genome were submitted to PHAST web server (109) to rapidly identify, annotate and graphically display prophage sequences within bacterial genomes or plasmids (Figure 3.2). For all phages listed above PHAST has able to identify prophages, except Pt-5771-G and Pt-259-G (Table 3.1). The former was shown to be a complete phage while the later was a remnant phage (Tables 3.2 and 3.3). Figure 3.2 shows a detail of the graphical output for Pt-B89-G. The sequence of phage KHP30 (NC_019928.1) were than MEGABLAST against the contigs of each *H. pylori* genome (Figure 3.3, table 3.2). All the genome assemblies showed significant alignment with the genome of KHP30. The query coverage was not high because

each contig just aligns will a small fraction of the genome of KHP30. As prophages are highly diverse (38), the identity found was considered quite reasonable.

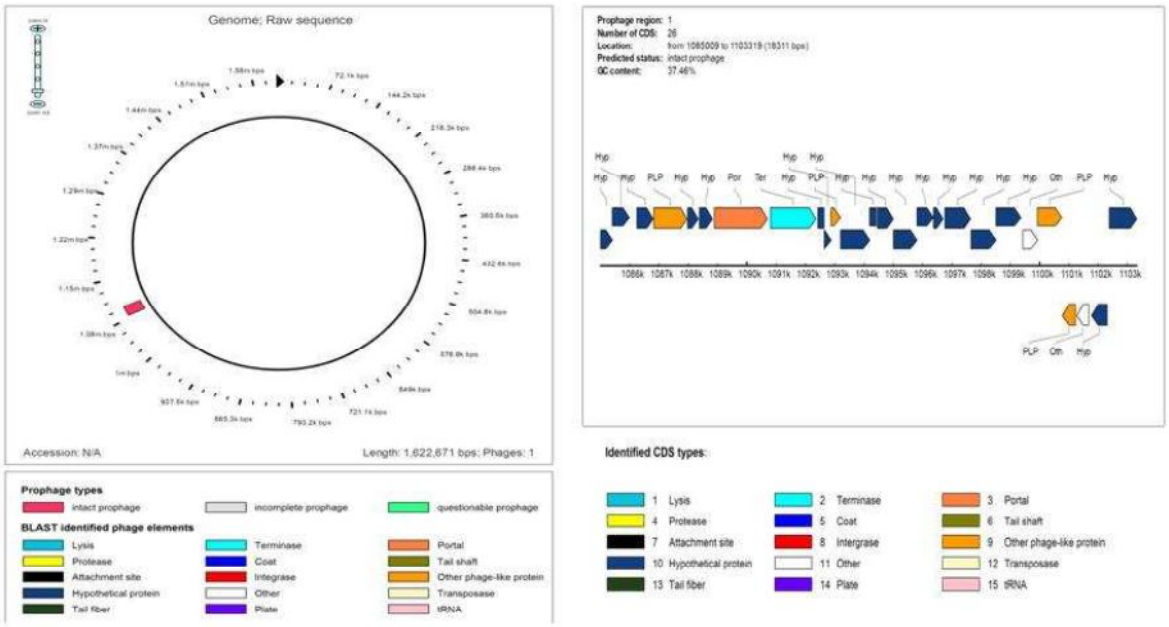


Figure 3.2. Graphical views including its circular and linear genome renderings as well PHAST's corresponding prophage annotation for the contigs of the strain Pt-B89-G. Note that the majority of the genes are from hypothetical proteins.

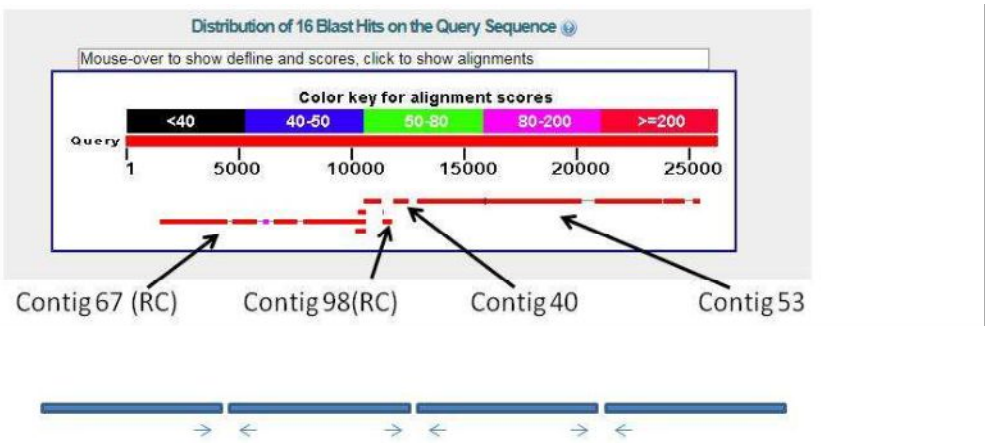


Figure 3.3. Alignment using MEGABLAST of phage KHP30 and contigs of Pt-B89-G. RC - reverse complement. The blue bars below represent the contigs (not in scale) and the arrows the primers to close the prophage sequence.

Table 3.2. MEGABLAST of alignment of the sequences of KHP30 and contigs of Pt-B89-G.

Description	Max Score	Total score	Query cover	E-value	Ident	Accession
NODE_40_length_607_cov_221.400330	793	793	0,02	0,0	0,88	Query_241272
NODE_53_length_14460_cov_66.832916	4490	10718	0,44	0,0	0,85	Query_241285
NODE_67_length_17993_cov_69.770912	3066	6932	0,31	0,0	0,87	Query_241295
NODE_98_length_467_cov_199.197006	488	488	0,01	8E-139	0,86	Query_241307

3.1 Prophage genome characteristics

We were able to close the physical gaps between contigs gaps in over 90% of the prophage genomes using PCR and Sanger sequencing. In most cases the prophage contigs were separated at insertion sequences, repetition zones and/or sequences showing homology with other bacterial genes. A prophage was considered intact if the size was larger than 20 Kb. According to this criterion, prophages were found to be intact in 23 of the 28 genomes (82%) (Table 3.2, first 23 genomes were considered intact after assembly). The other five genomes showed remnant prophages (Table 3.3) between 11.6 Kb and 19.8Kb. Intact prophages were initially divided in several contigs (min 1- max 7) and have an average of 34 predicted genes (min 24, max 39), 28.7Kb (min 22.6, max 33.0), and 36.7% GC, which is in line with other *H. pylori* prophages described (59, 60). The bacterial average GC percentage was 39.0%, suggesting horizontal gene transfer of the prophage region.

The gene content of intact prophages was similar to phage KHP30, a known complete phage with lytic cycle (62). The intact prophage genomes had a rather similar sequence (Figures 3.4 and 3.5) with a reasonably conserved gene order and in clear contrast with the host *H. pylori*, where the occurrence of genome rearrangement is well known (122). Genome annotation of prophage genes produced with either RAST (118) or PHAST (109) revealed that most of the open reading frames (ORF) corresponded to hypothetical proteins, disclosing the diversity of prophage genes and the consequent difficulty in the annotation process. The annotation with Phages 1.0 (http://www.phantome.org/PhageSeed/Phage.cgi?page=phast) did not add more information and was not further considered.

The similarity of prophage genomes was also quantified as a heat-map (Figure 3.6). This similarity matrix confirmed the percentages of bases which were identical. Only one prophage genome, strain Pt-4481-G, harbored a rearrangement (Figure 3.7), where the first segment of approximately 10.4 Kb appeared to be inverted. The second segment of about 15 Kb had the same gene order as all of the other prophages.

Regarding remnant prophages, different scenarios were observed: i) one phage (Sw-C388-G) has lost the putative DNA primase and helicase, among other proteins of unknown function placed in the first half of the prophage genome; ii) two phages (Sw-C520-G and Pt-259-G) most likely lost the second half of the prophage sequence; and iii) another two phages (Is-3180-G and Pt-5303-G) most likely lost specific ORFs. Among the later, only Is-3180-G could be assembled, yielding less than 20kb.

Table 3.2. Intact prophage genomes identified after whole genome sequencing.

Strain	Population		GC%		Insertion Site		Prophage				Accession number
	Phage - PST	MLST	bacteria	prophage	5'	3'	CDS * PHAST	CDS * PHAGES	CDS * RAST	Kb	
UK-EN31-U	hpNEurope	hpEurope	39.0	36.7	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	34	36	30.5	KX119174
UK-EN32-U	hpNEurope	hpEurope	38.9	36.7	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	34	35	29.9	KX119206
De-M53-M	hpNEurope	hpEurope	38.8	36.2	S-adenosylmethionine synthetase (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	33	32	33	28.1	KX119205
Sw-577-G	hpNEurope	hpEurope	38.9	36.3	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	30	29	32	26.9	KX119204
Sw-A626-G	hpNEurope	hpEurope	38.8	36.6	episome	episome	37	32	37	31.0	KX119177
Pt-B89-G	hpAfrica1	hpEurope	39.0	37.4	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	32	33	32	27.4	KX119203
Pt-1293-U	hpAfrica1	hpEurope	39.0	36.8	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	37	36	30.1	KX119202
Fr-ANT170-U	hpAfrica1	hpEurope	39.0	37.2	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	33	36	31.2	KX119201
Fr-MEG235-U	hpAfrica1	hpEurope	39.1	37.3	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	33	36	31.2	KX119200
Pt-5771-G	hpAfrica1	hpEurope	39.0	36.9	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	34	34	34	29.8	KX119199
Pt-5322-G	hpAfrica1	hpEurope	39.1	36.8	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	31	31	31	28.3	KX119198
Pt-228_99-G	hpAfrica1	hpEurope	39.0	37.2	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	36	38	30.1	KX119175

Pt-1846-U	hpAfrica1	hpEurope	39.0	37.0	GTP cyclohydrolase II/3,4-dihydroxy-2-butanone 4- phosphate synthase (jhp_0740)	ND	32	31	32	28.0	KX119176
Pt-B92-G	hpAfrica1	hpEurope	38.8	36.9	Membrane-associated phospholipid phosphatase (jhp_0787)	ND	39	36	38	30.5	KX119197
Pt-4481-G	hpAfrica1	hpEurope	39.0	36.8	ND	Ribosomal large subunit pseudouridine synthase B (EC 4.2.1.70) (jhp_1353)	32	31	32	25.4	KX119196
Fr-GC43-A	HpEastAsia	hpEurope	39.0	36.3	Competence protein ComGF (jhp_0650)	putative outer membrane protein HomA (jhp_0649)	38	37	39	33.0	KX119195
Fr-G12-G	hpEastAsia	hpEurope	38.9	36.3	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	36	35	36	28.6	KX119194
Fr-B58-M	hpEastAsia	hpEastAsia	38.8	36.0	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	26	24	26	22.6	KX119193
Pt-212-99R-U	hpAfrica1	hpEurope	38.9	37.1	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	24	24	24	23.0	KX119189
Pt-1918-U	hpSWEurope	hspWAfrica	39.1	36.2	Hypothetical protein (jhp_1347)	Putative outer membrane protein (jhp_1346)	34	33	34	28.7	KX119192
Pt-4497-U	hpSWEurope	hspWAfrica	39.3	36.2	hypothetical protein (jhp_0949)	Putative protein (jhp_0950)	35	34	36	29.4	KX119191
Pt-4472-G	hpSWEurope	hpEurope	38.8	36.6	hypothetical protein (jhp_0191)	hypothetical protein (jhp_0193)	32	30	32	27.6	KX119190
Fr-B41-M	hpSWEurope	hpWAfrica	39.1	35.5	Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2) (jph_0504)	hypothetical protein (jhp_0503)	35	35	36	29.4	KX119188

* Number of coding sequences (CDS) detected according to web service used; GC: guanine-cytosine; PST: prophage sequence typing; MLST: multilocus sequence typing.

Table 3.3. Remnant prophages identified after whole genome sequencing.

Strain	Population		GC%		Insertion Site		Prophage				Accession number
	Phage - PST	MLST	bacteria	prophage	5'	3'	CDS PHAST	CDS PHAGE	CDS RAST	Kb	
Sw-C388-G	hpNEurope	hpEurope	38.9	36.1	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	21	21	21	13.0	KX119183
Sw-C520-G[a]	hpNEurope	hpEurope	38.8	37.7	ND	ND	14	13	13	13.6	KX119184 KX119185 KX119186 KX119187
Is-3180-G	hpAfrica1	hpEurope	39.1	37.1	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	28	26	28	19.7	KX119182
Pt-259-G	hpSWEurope	hpEurope	39.0	36.2	hypothetical protein (jhp_0956)	hypothetical protein (jhp_0955)	12	10	11	11.6	KX119181
Pt-5303-G[b]	hpSWEurope	hpEurope	38.9	36.2	Competence protein ComGF (jhp_0650)	ND	20	21	21	19.8	KX119178 KX119179 KX119180

[a] Prophage in 4 contigs, complete genome sequence not available; [b] Prophage in 3 contigs, complete genome sequence not available. GC:guanine-cytosine; PST: prophage sequence typing; MLST: multilocus sequence typing.

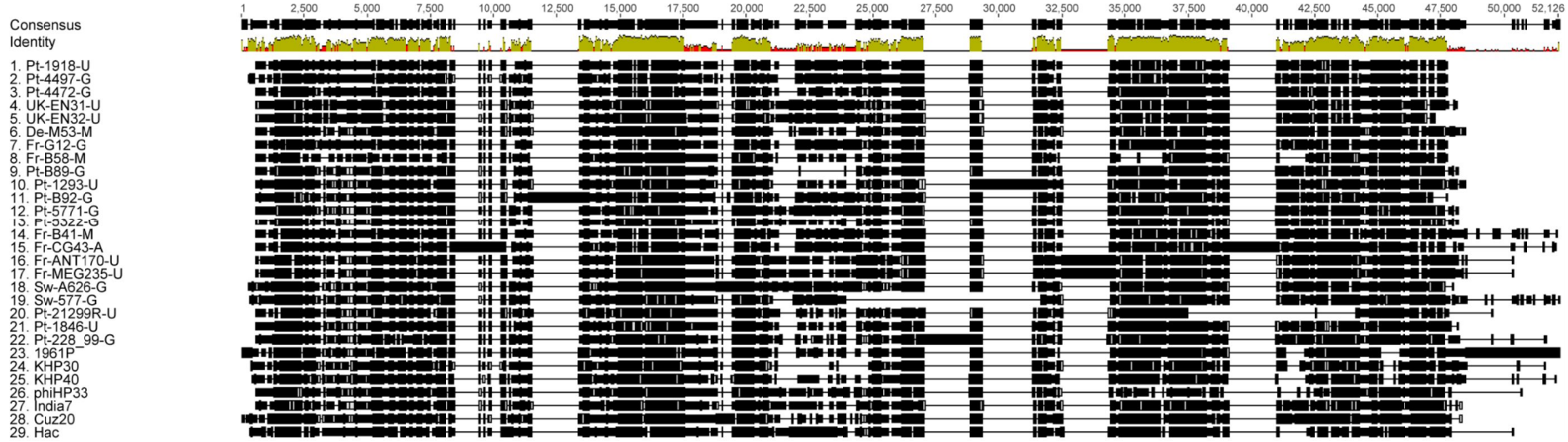
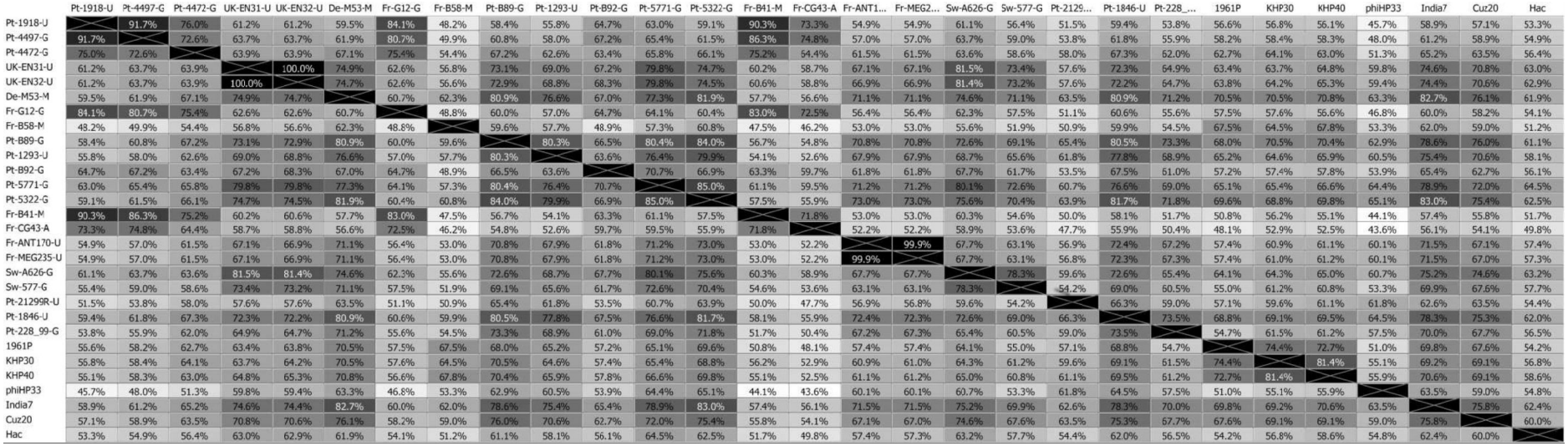


Figure 3.4. MAFFT alignment of prophage genomes showing consensus identity. Consensus identity along the multiple alignment of prophage genomes (bp) shows mean pairwise identity over all pairs in the column. The multiple sequence alignment is a basic tool in inferring the evolutionary history of biological sequences. For instance, the alignment of several homologs can be used to search for patterns of highly conserved regions. The detailed percent similarity between prophage genomes is presented in Figure 3.6. Green: 100% identity, Greenish-brown: at least 30% and under 100% identity, Red: below 30% identity.

Figure 3.5. Alignment of 29 complete prophages, using Mauve software (version 2.3.1).



	Pt-1918-U	Pt-4497-G	Pt-4472-G	UK-EN31-U	UK-EN32-U	De-M53-M	Fr-G12-G	Fr-B58-M	Pt-B89-G	Pt-1293-U	Pt-B92-G	Pt-5771-G	Pt-5322-G	Fr-B41-M	Fr-CG43-A	Fr-ANT1...	Fr-MEG2...	Sw-A626-G	Sw-577-G	Pt-2129...	Pt-1846-U	Pt-228_...	1961P	KHP30	KHP40	phiHP33	India7	Cuz20	Hac
Pt-1918-U		91.7%	76.0%	61.2%	61.2%	59.5%	84.1%	48.2%	58.4%	55.8%	64.7%	63.0%	59.1%	90.3%	73.3%	54.9%	54.9%	61.1%	56.4%	51.5%	59.4%	53.8%	56.6%	56.8%	56.1%	45.7%	58.9%	57.1%	53.3%
Pt-4497-G	91.7%		72.6%	63.7%	63.7%	61.9%	80.7%	49.9%	60.8%	58.0%	67.2%	65.4%	61.5%	86.3%	74.8%	57.0%	57.0%	63.7%	59.0%	53.8%	61.8%	55.9%	58.2%	58.4%	58.3%	48.0%	61.2%	58.9%	54.9%
Pt-4472-G	76.0%	72.6%		63.9%	63.9%	67.1%	75.4%	54.4%	67.2%	62.6%	63.4%	65.8%	66.1%	75.2%	54.4%	61.5%	61.5%	63.6%	58.6%	58.0%	67.3%	62.0%	62.7%	64.1%	63.0%	51.3%	65.2%	63.5%	56.4%
UK-EN31-U	61.2%	63.7%	63.9%		100.0%	74.9%	62.6%	56.8%	73.1%	69.0%	67.2%	79.8%	74.7%	60.2%	58.7%	67.1%	67.1%	81.5%	73.4%	57.6%	72.3%	64.9%	63.4%	63.7%	64.8%	59.8%	74.6%	70.8%	63.0%
UK-EN32-U	61.2%	63.7%	63.9%	100.0%		74.7%	62.6%	56.6%	72.9%	68.8%	68.3%	79.8%	74.5%	60.6%	58.8%	66.9%	66.9%	81.4%	73.2%	57.6%	72.2%	64.7%	63.8%	64.2%	65.3%	59.4%	74.4%	70.6%	62.9%
De-M53-M	59.5%	61.9%	67.1%	74.9%	74.7%		60.7%	62.3%	80.9%	76.6%	67.0%	77.3%	81.9%	57.7%	56.6%	71.1%	71.1%	74.6%	71.1%	63.5%	80.9%	71.2%	70.5%	70.5%	70.8%	63.3%	82.7%	76.1%	61.9%
Fr-G12-G	84.1%	80.7%	75.4%	62.6%	62.6%	60.7%		48.8%	60.0%	57.0%	64.7%	64.1%	60.4%	83.0%	72.5%	56.4%	56.4%	62.3%	57.5%	51.1%	60.6%	55.6%	57.5%	57.6%	56.6%	46.8%	60.0%	58.2%	54.1%
Fr-B58-M	48.2%	49.9%	54.4%	56.8%	56.6%	62.3%	48.8%		59.6%	57.7%	48.9%	57.3%	60.8%	47.5%	46.2%	53.0%	53.0%	55.6%	51.9%	50.9%	59.9%	54.5%	67.5%	64.5%	67.8%	53.3%	62.0%	59.0%	51.2%
Pt-B89-G	58.4%	60.8%	67.2%	73.1%	72.9%	80.9%	60.0%	59.6%		80.3%	66.5%	80.4%	84.0%	56.7%	54.8%	70.8%	70.8%	72.6%	69.1%	65.4%	80.5%	73.3%	68.0%	70.5%	70.4%	62.9%	78.6%	76.0%	61.1%
Pt-1293-U	55.8%	58.0%	62.6%	69.0%	68.8%	76.6%	57.0%	57.7%	80.3%		63.6%	76.4%	79.9%	54.1%	52.6%	67.9%	67.9%	68.7%	65.6%	61.8%	77.8%	68.9%	65.2%	64.6%	65.9%	60.5%	75.4%	70.6%	58.1%
Pt-B92-G	64.7%	67.2%	63.4%	67.2%	68.3%	67.0%	64.7%	48.9%	66.5%	63.6%		70.7%	66.9%	63.3%	59.7%	61.8%	61.8%	67.7%	61.7%	53.5%	67.5%	61.0%	57.2%	57.4%	57.8%	53.9%	65.4%	62.7%	56.1%
Pt-5771-G	63.0%	65.4%	65.8%	79.8%	79.8%	77.3%	64.1%	57.3%	80.4%	76.4%	70.7%		85.0%	61.1%	59.5%	71.2%	71.2%	80.1%	72.6%	60.7%	76.6%	69.0%	65.1%	65.4%	66.6%	64.4%	78.9%	72.0%	64.5%
Pt-5322-G	59.1%	61.5%	66.1%	74.7%	74.5%	81.9%	60.4%	60.8%	84.0%	79.9%	66.9%	85.0%		57.5%	55.9%	73.0%	73.0%	75.6%	70.4%	63.9%	81.7%	71.8%	69.6%	68.8%	69.8%	65.1%	83.0%	75.4%	62.5%
Fr-B41-M	90.3%	86.3%	75.2%	60.2%	60.6%	57.7%	83.0%	47.5%	56.7%	54.1%	63.3%	61.1%	57.5%		71.8%	53.0%	53.0%	60.3%	54.6%	50.0%	58.1%	51.7%	50.8%	56.2%	55.1%	44.1%	57.4%	55.8%	51.7%
Fr-CG43-A	73.3%	74.8%	64.4%	58.7%	58.8%	56.6%	72.5%	46.2%	54.8%	52.6%	59.7%	59.5%	55.9%	71.8%		52.2%	52.2%	58.9%	53.6%	47.7%	55.9%	50.4%	48.1%	52.9%	52.5%	43.6%	56.1%	54.1%	49.8%
Fr-ANT170-U	54.9%	57.0%	61.5%	67.1%	66.9%	71.1%	56.4%	53.0%	70.8%	67.9%	61.8%	71.2%	73.0%	53.0%	52.2%		99.9%	67.7%	63.1%	56.9%	72.4%	67.2%	57.4%	60.9%	61.1%	60.1%	71.5%	67.1%	57.4%
Fr-MEG235-U	54.9%	57.0%	61.5%	67.1%	66.9%	71.1%	56.4%	53.0%	70.8%	67.9%	61.8%	71.2%	73.0%	53.0%	52.2%	99.9%		67.7%	63.1%	56.8%	72.3%	67.3%	57.4%	61.0%	61.2%	60.1%	71.5%	67.0%	57.3%
Sw-A626-G	61.1%	63.7%	63.6%	81.5%	81.4%	74.6%	62.3%	55.6%	72.6%	68.7%	67.7%	80.1%	75.6%	60.3%	58.9%	67.7%	67.7%		78.3%	59.6%	72.6%	65.4%	64.1%	64.3%	65.0%	60.7%	75.2%	74.6%	63.2%
Sw-577-G	55.4%	59.0%	58.6%	73.4%	73.2%	71.1%	57.5%	51.9%	69.1%	65.6%	61.7%	72.6%	70.4%	54.6%	53.6%	63.1%	63.1%	78.3%		54.2%	69.0%	60.5%	55.0%	61.2%	60.8%	53.3%	69.9%	67.6%	57.7%
Pt-21299R-U	51.5%	53.8%	58.0%	57.6%	57.6%	63.5%	51.1%	50.9%	65.4%	61.8%	53.5%	60.7%	63.9%	50.0%	47.7%	56.9%	56.8%	59.6%	54.2%		66.3%	59.0%	57.1%	59.6%	61.1%	61.8%	62.6%	63.5%	54.4%
Pt-1846-U	59.4%	61.8%	67.3%	72.3%	72.2%	80.9%	60.6%	59.9%	80.5%	77.8%	67.5%	76.6%	81.7%	58.1%	55.9%	72.4%	72.3%	72.6%	69.0%	66.3%		73.5%	68.8%	69.1%	69.5%	64.5%	78.3%	75.3%	62.0%
Pt-228_99-G	53.8%	55.9%	62.0%	64.9%	64.7%	71.2%	55.6%	54.5%	73.3%	68.9%	61.0%	69.0%	71.8%	51.7%	50.4%	67.2%	67.3%	65.4%	60.5%	59.0%	73.5%		54.7%	61.5%	61.2%	57.5%	70.0%	67.7%	56.5%
1961P	56.6%	58.2%	62.7%	63.4%	63.8%	70.5%	57.5%	67.5%	68.0%	65.2%	57.2%	65.1%	69.6%	50.8%	48.1%	57.4%	57.4%	64.1%	55.0%	57.1%	68.8%	54.7%		74.4%	72.7%	51.0%	69.8%	67.6%	54.2%
KHP30	56.8%	58.4%	64.1%	63.7%	64.2%	70.5%	57.6%	64.5%	70.5%	64.6%	57.4%	65.4%	68.8%	56.2%	52.9%	60.9%	61.0%	64.3%	61.2%	59.6%	69.1%	61.5%	74.4%		81.4%	55.1%	69.2%	69.1%	56.8%
KHP40	56.1%	58.3%	63.0%	64.8%	65.3%	70.8%	56.6%	67.8%	70.4%	65.9%	57.8%	66.6%	69.8%	55.1%	52.5%	61.1%	61.2%	65.0%	60.8%	61.1%	69.5%	61.2%	72.7%	81.4%		55.9%	70.6%	69.1%	58.6%
phiHP33	45.7%	48.0%	51.3%	59.8%	59.4%	63.3%	46.8%	53.3%	62.9%	60.5%	53.9%	64.4%	65.1%	44.1%	43.6%	60.1%	60.1%	60.7%	53.3%	61.8%	64.5%	57.5%	51.0%	55.1%	55.9%		63.5%	59.0%	54.8%
India7	58.9%	61.2%	65.2%	74.6%	74.4%	82.7%	60.0%	62.0%	78.6%	75.4%	65.4%	78.9%	83.0%	57.4%	56.1%	71.5%	71.5%	75.2%	69.9%	62.6%	78.3%	70.0%	69.8%	69.2%	70.6%	63.5%		75.8%	62.4%
Cuz20	57.1%	58.9%	63.5%	70.8%	70.6%	76.1%	58.2%	59.0%	76.0%	70.6%	62.7%	72.0%	75.4%	55.8%	54.1%	67.1%	67.0%	74.6%	67.6%	63.5%	75.3%	67.7%	67.6%	69.1%	69.1%	59.0%	75.8%		60.0%
Hac	53.3%	54.9%	56.4%	63.0%	62.9%	61.9%	54.1%	51.2%	61.1%	58.1%	56.1%	64.5%	62.5%	51.7%	49.8%	57.4%	57.3%	63.2%	57.7%	54.4%	62.0%	56.5%	54.2%	56.8%	58.6%	54.8%	62.4%	60.0%	

Figure 3.6. Heat–map representing the phage similarity matrix between 29 phages genomes based on genome content. The heat-map provides a graphical representation of data where values in a matrix are represented as a grey scale: the darker the grey, the greater the similarity.

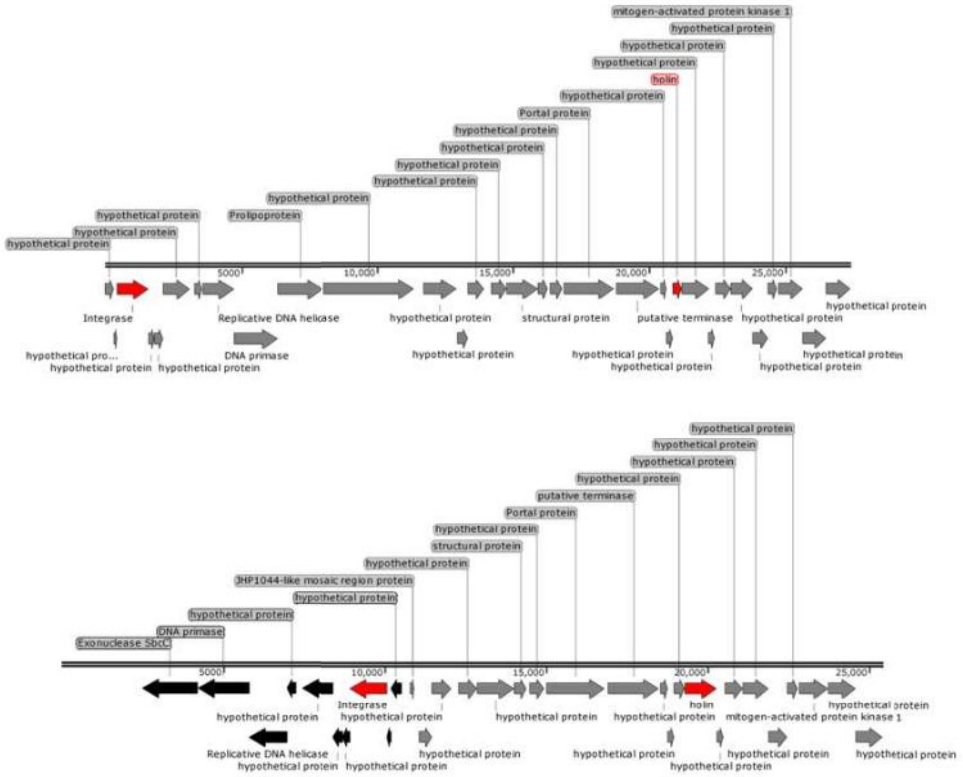


Figure 3.7. Genetic mapping of prophage genomes highlighting the most common positions of genes on top. Prophage open reading frame (ORF) organization for PtB89-G (on top) and for Pt-4481-G (on bottom). Red arrows show integrase and holin genes. Black arrows highlight an inversion of the prophage ORF for Pt-B89-G.

3.2 Insertion Sequences

Insertion sequences (IS), comprised of two ORFs inserted into prophage genomes were found in 39.1% (9/23) of complete prophages (Table 3.4) classified (according to PST typing) as hpNEurope (n=2), hpAfrica1 (n=5) and hpEastAsia (n=2), and in 50% (3/6) of remnant prophages classified as hpNEurope (n=1), hpAfrica1 (n=1) and hpSWEurope (n=1). The complete prophages Uk-EN31-U, Uk-EN32-U, Pt-B92-G and Fr-GC43-A had IS605 inserted once in the first three cases and twice in the last case. Interestingly, in Fr-GC43-A one copy of IS605 was inverted in relation to the other copy (Figure 3.8). IS605 was inverted in Uk-EN31-U, Uk-EN32-U and in one of the IS of Fr-GC43-A. The prophages Pt-228_99-G, Fr-ANT170-U and Fr-MEG235-U had two copies of ISHp608. The IS was inserted in a reverse order in relation to the other copy in Fr-ANT170-U, and twice with the same orientation in Pt-228_99-G. The third IS found was IS607 in genomes Pt-1293-U and Fr-B58-M.

Concerning remnant prophages, Sw-C388-G has the IS606 inserted at its 3' end and the second ORF is again truncated in two. Finally, Is-3180-G carries ISHp608. The remnant prophage Pt-5303-G could not be completely assembly but ISHp608 was also found in a separate contig. Despite all of our efforts, we were not able to determine if this IS was inserted into the prophage genome or not.

IS were not always found at the same position in the prophage genomes, but prophages from strains of the same country of origin tended to present the same IS at same genome context (Table 3.4). Nevertheless, IS were present in most cases (9/13, 69%) immediately before DNA

helicase (2/9), either before or after DNA primase (4/9), after structural protein (2/9), or after holin gene (1/9), which therefore could be considered as hotspots for IS in prophages.

Table 3.4. Insertion sequences (IS) within intact prophage genomes

Prophage	IS	Copies	Insertion sense*	5' Flanking gene	3' Flanking gene
UK-EN31-U	IS605	1	Inverse	hypothetical protein	DNA helicase
UK-EN32-U	IS605	1	Inverse	hypothetical protein	DNA helicase
Fr-GC43-G	IS605	2	Inverse	DNA primase	Exonuclease sbcC
			Direct	holin	hypothetical protein
Pt-B92-G	IS605	1	Direct	DNA primase	hypothetical protein
Fr-ANT170-U	ISHp608	2	Inverse	hypothetical protein	hypothetical protein
			Direct	hypothetical protein	hypothetical protein
Fr-MEG235-U	ISHp608	2	Inverse	hypothetical protein	hypothetical protein
			Direct	hypothetical protein	hypothetical protein
Pt-228_99-G	ISHp608	2	Direct	DNA helicase	DNA primase
			Direct	structural protein	structural protein
Pt-1293-U	IS607	1	Inverse	structural protein	Structural protein
Fr-B58-M	IS607	1	Inverse	Integrase	DNA primase

* Inverse or direct sense in relation to IS605 (accession number U60177), ISHp608 (accession number AF357224) and IS607 (accession number AB889602). Details on the location of IS sequences can be found in table S3.

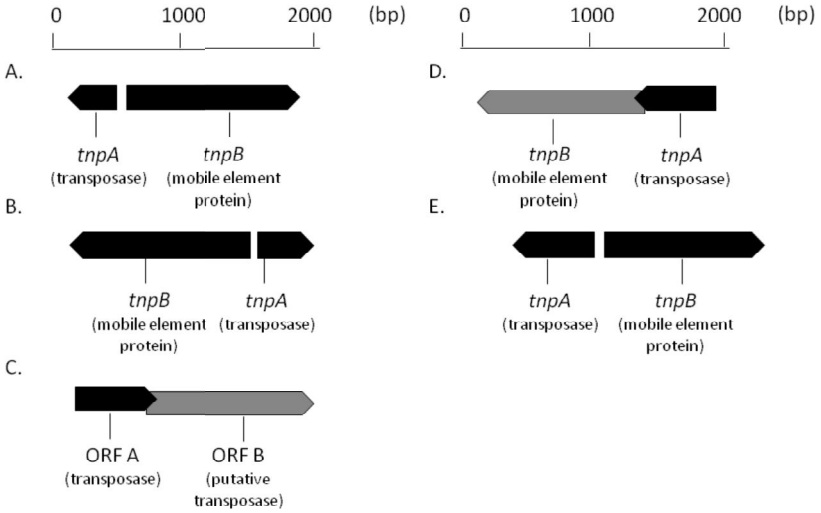


Figure 3.8. Genetic layout of the Insertion Sequences (IS) found in prophage genomes. A. IS605 (representation of accession number U60177) found in Pt-B92-G and in one case of Fr-GC43-A; B. IS605 found in UK-EN31-U, UK-EN32-U and in one case of Fr-GC43-A; C. ISHp608 (representation of accession number AF357224) found twice in Fr-ANT170-G and Fr-MEG235-G (copies in different sense) and once in Pt-228_99-G and Is-3180-G; D. IS607 (representation of accession number AB889602) found in Pt-1293-U and Fr-B58-M; E. IS606 (representation of accession number U95957) found in Sw-C388-G.

The transposase genes from IS605 were inserted near the lysis cassette, as described for Mu-like phages (133), DNA helicase and DNA primase. IS607 was located adjacent to DNA primase or a structural protein and ISHp608 near DNA primase, portal protein or structural protein. In a few cases IS were inserted into a coding sequence of a structural protein (Pt-1293-U and Pt-228_99-G) or a hypothetical protein (Pt-B92-G, Fr-ANT170-U and Fr-MEG235-U), which may impinge on transcription, and the prophage genes may be non-functional. Accordingly, IS do not appear to be randomly inserted into prophage genome. Our hypothesis is that the presence of IS within the prophage genome may inactivate the lytic cycle, benefiting the host.

3.3 Prophage insertion site

Knowledge of the insertion site of prophages provides clues about ancient acquisition and vertical heritage. Accordingly, prophages at similar loci in different genomes can derive from a single ancestral prophage (44). Furthermore, *H. pylori* prophage insertion sites have not been extensively studied before.

Prophage insertion site was mostly conserved among *H. pylori* PST populations. Interestingly, about 50% of the prophages enrolled in the present study and especially for the populations hpAfrica1 and hpNEurope are inserted between the same two genes, S-adenosylmethionine synthetase (synthesizes S-adenosylmethionine (AdoMet)), and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (metabolic pathway of lipid A). These two genes are usually contiguous in the *H. pylori* genome. Prophages classified as belonging to the hpEastAsia population, although represented in a very small number, appear to be inserted between genes coding for a competence protein ComGF and a putative outer membrane protein. Phages from hpSWEurope appear to be inserted at random locations (Tables 3.2 and 3.3).

The presence of tandem repeats at the 3' end of the prophage insertion site was often verified for prophages integrated between S-adenosylmethionine synthetase and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (Table 3.5).

Table 3.5. Sequences of tandem repeats at the 3' end of prophage genomes.

Prophage genome	Repetition sequence at 3' end Consensus pattern	Size (bp)	Copy number
De-M53-M	GGCGATGATAACGAAAGCGTGAGCGGTGTAGGTGTG	36	3.1
	GTGAAGCCCTATTACCACCATGCTGTAAAAAAATTTT	37	4.3
	ATATTTTAAATTTATCTTAC	20	2.4
Pt-1293-U	GCGCCTTGCCGTGCTATAAA	20	7.5
UK-EN31-U	AGCCTCCTTATAATAGCGGAAGCGTGGGCGGTGGCTATGATGTGAG CGAGTGA	53	4.9
UK-EN32-U	ATAGCGGAAGCGTGGGCGGTGGCTATGATGCGAGCGAGTGAAGCC TCCTTATA	53	5.7
Pt-B89-G	AGCGGTGGCTATGGTAGCGA	20	16.4
Pt-5771-G	GCTATAAAGCGCCTTGCCGT	20	5.8
Pt-5322-G	GCGCCTTGCCTTGCTATAAA	20	10.9
Is-3180-G	AATAAACCGCCTTGCCATTC	20	4.9
	AATGTGTGTTTTT	13	2.8
Fr-ANT170-U	AATGTTTTCA	10	6.8
	TTTTCAAATGTTTTTAAACT	20	3.0
Fr-MEG235-U	AATGTTTTCA	10	6.8
	TTTTCAAATGTTTTTAAACT	20	3.0
Pt-1918-U	AATTGTGACACA	12	2.1

The prophage Sw-A626-G appear to be present as an episome, in which the phage is present like a plasmid not integrated in the bacterial genome. This was found after sequencing the PCR product produced with a pair of primers in which the forward primer was in the 3'end of the prophage genome and the reverse primer was in the 5' end of the prophage genome. Moreover, this pair of primers produced two fragments with 58 bases difference. This difference occurs due to the presence of repeats in different number (Figure 3.9). Moreover, multiple sequences closing the prophage appear to coexist, since for the bigger fragment presented several overlapping bases pointing for existence of polymorphism, besides a higher number of repeats.

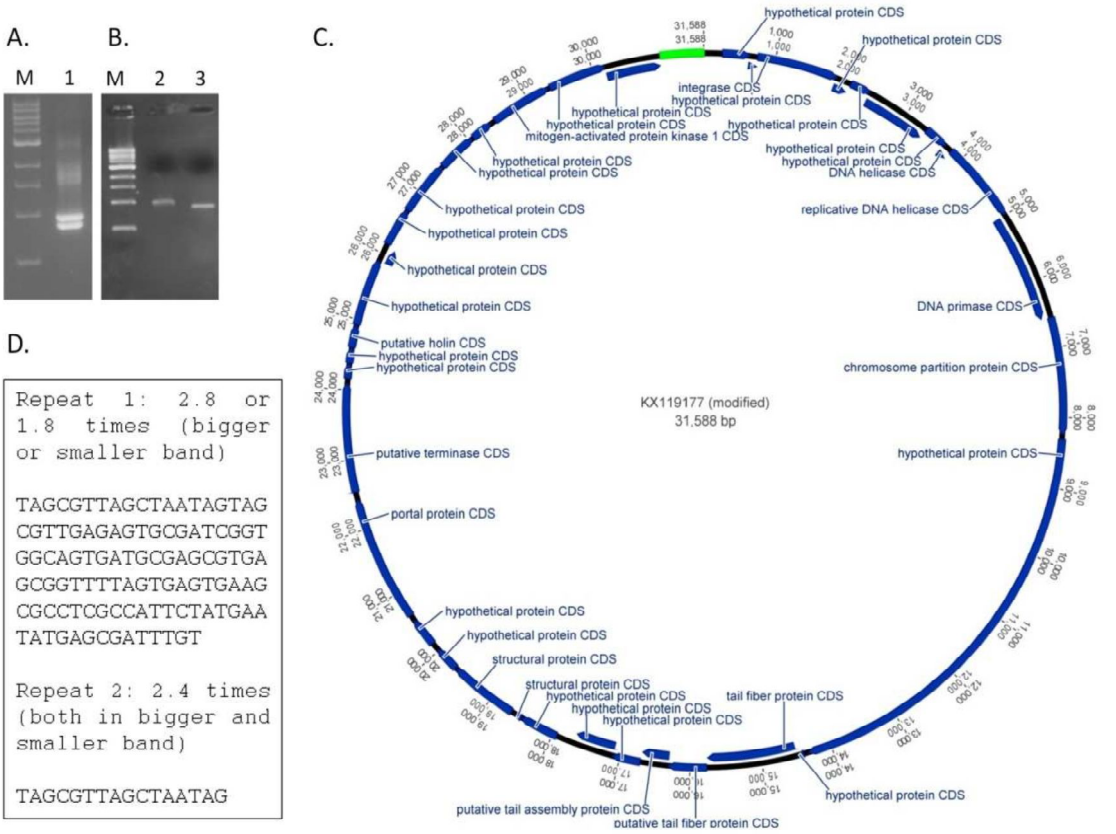


Figure 3.9. Episomal prophage Sw-A626-G. A) Electrophoresis of PCR product of prophage Sw-A626-G amplified with primers at 3'end and 5' end on 2% agarose gel. M is a 100 bp DNA ladder. B) Electrophoresis after gel band purification, lane 1 bigger band and lane 2 smaller band. C) Map of Sw-A626 episomal prophage evidencing smaller fragment closing the prophage in green. D) Repeats present in fragments closing the prophage.

3.4 Prophage core and pangenome

Given 30 *H. pylori* prophage genomes a pangenome analysis was performed. Results showed that the core genome is composed of nine ORF while the pangenome includes 55 ORFs (Figure 3.10). The core genome represents 16% of the pangenome and 27% of the average number of genes per genome (there is a mean number of 33 genes per genome).

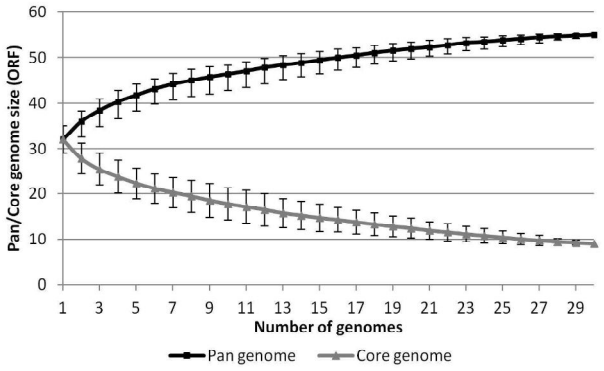


Figure 3.10. Curve for core and pangenome of 30 *H. pylori* prophages in light grey and black respectively using a Python script. Numbers of ORFs were estimated by performing 100 random different input orders of genomes. Solid line correspond to the average number and ORF and for each point the standard deviation of the mean is presented.

The core genes correspond to ORF presented in table 3.6. Most core ORF code for hypothetical proteins. Only one ORF has an assigned function related with the lysogenic cycle that is the integrase gene. Interestingly there a group of other 9 ORFs (ORFs number 9, 17, 18, 19, 20, 21, 23, 24, 25) shared by three phylogeographic groups (SWEurope, NEurope and EastAsia) which code for DNA primase (ORF 9), portal protein (ORF 17), putative terminase (ORF 18) and holin (ORF 21) plus other 5 hyopthetical proteins.

The core and pangenome was also computed for each phylogenetic group of *H. pylori* prophages (Figure 3.11). For SWEurope prophages, 4 genomes were included and the core genome is made of 29 ORFs and the pangenome of 39 ORFs. The core genome represents 85% of the average number of genes per genome. For NEurope prophages, 7 genomes were included and the core genome is made of 23 ORFs and the pangenome of 42 ORFs. The core genome represents 70% of the average number of genes per genome (the mean number of ORFs per genome is 33). For EastAsia prophages, 6 genomes were included and the core genome is made of 22 ORFs and the pangenome of 44 ORFS. The core genome represents 69% of the average number of genes per genome (the mean number of ORFs per genome is 32). Finally, for Africa1 prophages, 12 genomes were included and the core genome is made of 17 ORFs and the pangenome of 45 ORFS. The core genome represents 36% of the average number of genes per genome (the mean number of ORFs per genome is 33). The trend lines evidence that the core genome tend to decrease with the number of added genomes, while the pangenome increases. It should be noted that for Africa1 prophages, if considering 6 genomes only the number of core genes is 26 (and the pangenome is 40), which is similar to the other phylogeographic phage groups. It is thus expectable that the number of core genes decreases as other prophages genomes are added, which of course should be confirmed with experimental data.

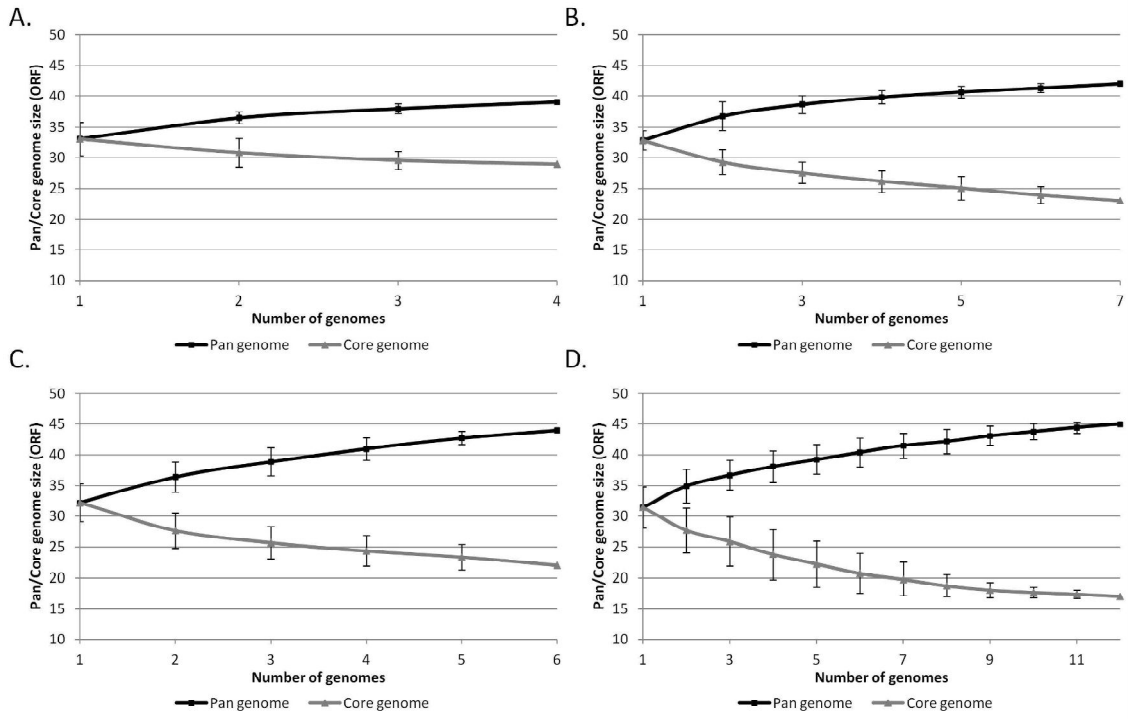


Figure 3.11. Curve for core and pangenome according to phylogeographic group. A. 4 *H. pylori* prophages from SWEurope; B. 7 *H. pylori* prophages from NEurope; C. 6 *H. pylori* prophages from EastAsia; and D. 12 *H. pylori* prophages from Africa1. Core and pangenome trend line in light grey and black respectively, using in-house Python script. Numbers of ORFs were estimated by performing 100 random different input orders of genomes. Solid line correspond to the average number and ORF and for each point the standard deviation of the mean is presented.

Table 3.6. Core genome of *H. pylori* prophages.

Core genome (ORF)	Putative function
1	Hypothetical protein
2	Integrase
10	Hypothetical protein
11	Hypothetical protein
15	Hypothetical protein
16	Hypothetical protein
26	Hypothetical protein
27	Hypothetical protein
29	Hypothetical protein

The Venn diagrams of the common gene pools of the analyzed genomes is presented in figure 3.12. Each prophage group was build with the their core genome. Every area in this Venn diagram represents a subset of the compared groups and is labeled with the number of ORFs in this subset.

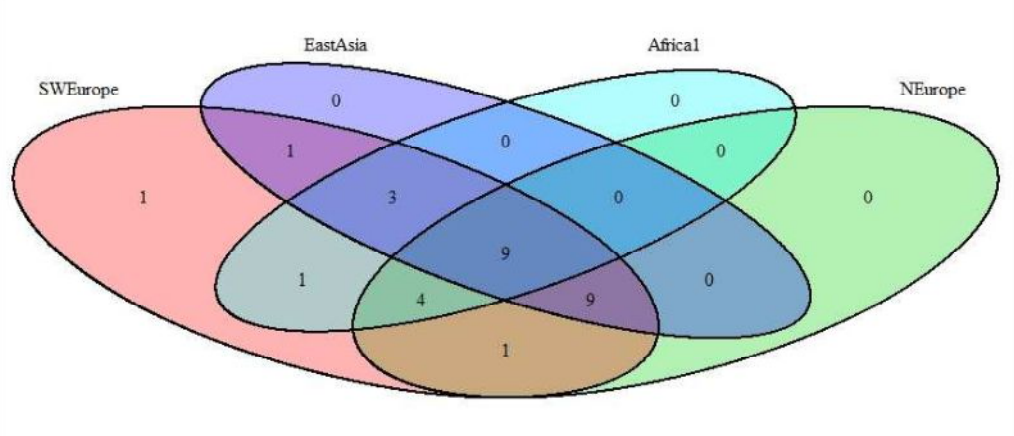


Figure 3.12. Venn diagram of the core-genome, variable-genome and group specific sizes for the four prophage phylogeographic groups. Values on diagram represent the numbers of shared ORFs for each group intersections.

3.5 Prophage phylogenetic relationships

To get insight into the genetic backbone of the identified prophages and to infer their phylogenetic relationships in the frame of the well-known *H. pylori* geographic distribution, all 23 intact genomic sequences (Table 3.2) as well as the publicly available complete genomes of six *Helicobacter* phages (India7, Cuz20, 1961P, KHP30, KHP40 and phiHP33) and the outgroup *H. acinonychis* prophage, were selected for increasing genetic diversity and were analyzed. Figure 3.13 shows the phylogenetic inferences found for the complete prophage genome and the concatenated integrase and holin prophage genes (PST).

We observed that the majority of the prophages gather by phylogeographic group, clustering accordingly to their population assigned by STRUCTURE (30, 130, 134), in a similar fashion to what we described previously for the concatenated integrase and holin genes only (65). However, evident exceptions were noted for some prophages, namely Pt-4472-G, Fr-G12-G, Fr-CG43-A, Pt-B92-G, Pt-21299R-U and Cuz20, which displayed a discrepant phylogeographic segregation from their PST classification, suggesting the existence of putative recombination events. For instance, Pt-4472-G prophage which, according to STRUCTURE analysis, belongs to hpSWEurope, appears to be a genomic mosaic composed of both hpSWEurope and hpAfrica1 populations. This is clearly evident in Figure 3.14A, where Pt-4472-G is >90% similar to the latter in the genome central region, whereas the similarity to the hpSWEurope

population reached values <50%. Curiously, the regions where the opposite is observed (i.e., >90% similarity to hpSWEurope) encompass both the integrase and holin genes that are used for PST classification.

Another clear example of prophage recombination is exhibited by Pt-B92-G, which was PST-classified as hpAfrica1. Although most of its genome appears to be inherited from a hpAfrica1 or hpNEurope population, it displays a small middle region where similarities to the hpSWEurope population reached >95% while is strikingly different from the remainder (Figure 3.14B). Although less evident, we would also like to highlight two other interesting cases involving mosaicism between hpSWEurope and hpAfrica1 populations, namely Fr-G12-G and Pt-21299R-U. Despite the fact that the former was PST-classified as hpEastAsia, most of its genome was clearly inherited from a hpSWEurope population with the exception of a small 3'-end region which is highly similar to an hpAfrica1 population (data not shown). To the contrary, most of the Pt-21299R-U genome is similar to hpAfrica1, except for its 3'-end which is highly similar (>95%) to an hpSWEurope population (similarity to hpAfrica1 is as low as 40%). Interestingly, the holin gene is absent in this prophage and, in the integrase-involved region, both hpAfrica1 and hpSWEurope populations are almost equally represented (data not shown). Considering the huge genomic diversity observed among all prophage genomes, a precise identification of the location of the breakpoint regions for all of the described recombination events was not possible.

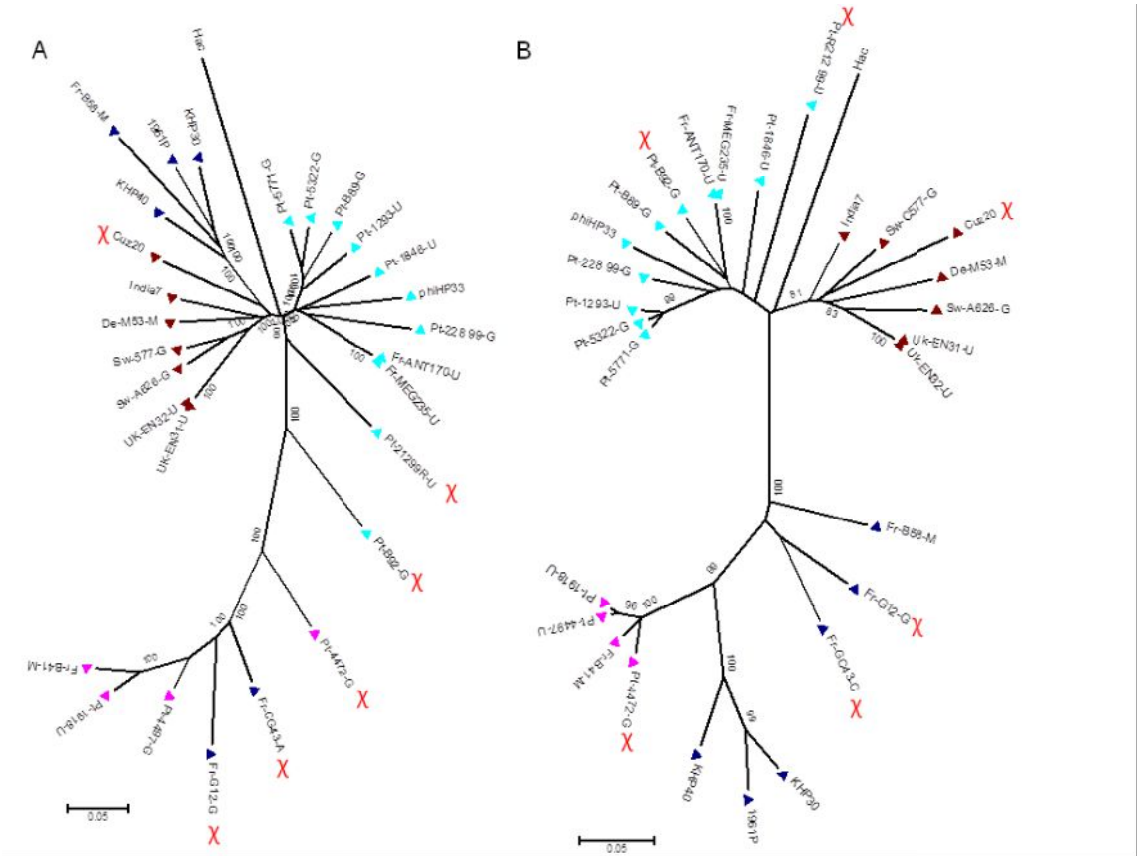


Figure 3.13. Phylogenetic trees based on A) prophage genomes and B) prophage sequence typing (PST). Neighbour-joining trees, Kimura two-parameter model, complete deletion option and 1000 resampling using MEGA 6.0 software. Phage population: brown triangles: hpNEurope; pink triangles: HpSWEurope; dark-blue triangles: HpEastAsia; light-blue triangles: hpAfrica1. Hac - *Helicobacter acinonychis* prophage. χ - highlights recombinogenic prophage genomes.

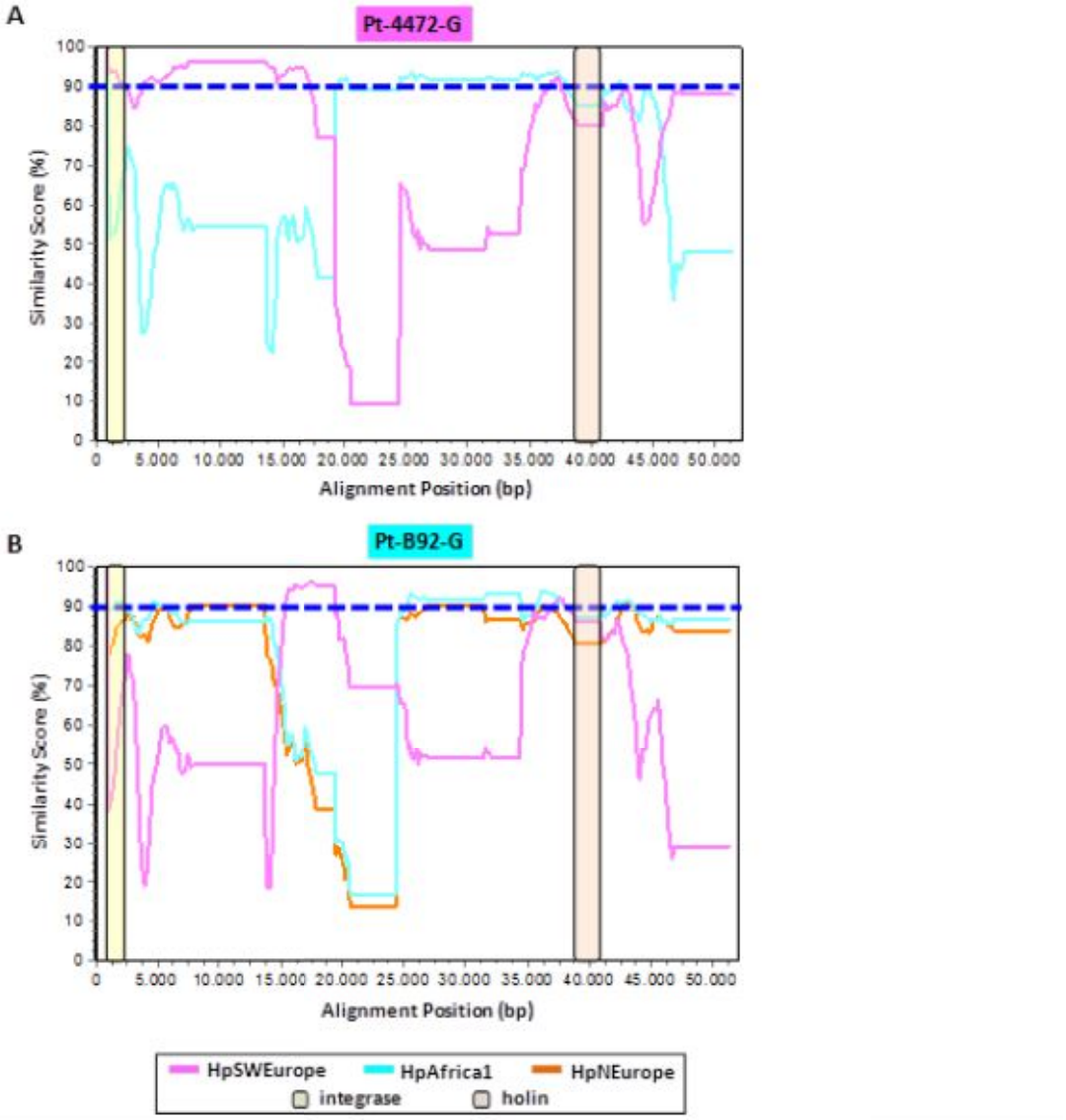


Figure 3.14. Genomic mosaicism of Pt-44772-G and Pt-B92-G prophages. (A) SimPlot showing the genetic similarity of PT-4472-G (PST-classified as HpSWEurope) to both the HpSWEurope and HpAfrica1 populations. (B) SimPlot showing the genetic similarity of Pt-B92-G (PST-classified as HpAfrica1) to HpSWEurope, HpAfrica1 and HpNEurope populations. For both plots, the Kimura 2-paramter model was used to calculate nucleotide similarities in a sliding-window of 1500bp and a step size of 150bp, with gap strip on. Cut-off of 90% similarity is shown in a blue dashed-line.

4. Discussion

This work provide an escalation of the number of sequenced *H. pylori* prophages, representing a fourfold increase made possible by next generation sequencing techniques. Comparative genomics evidenced a high genetic diversity, which is in line with early evolutionary origin, an intrusion of other mobile sequences within phage genomes reflecting cross horizontal genetic exchanges and a myriad of genetic sequences of unknown function that may play an important role in phage and host biology.

Prophage genome assembly using *de novo* assembly algorithms produced in most cases several contigs that needed to be further processed using KHP30 as a reference genome allowing to order phage contigs. Sanger sequencing was needed to close the genome gaps arriving to complete sequence for all intact phages (Table 3.2), that was not always possible to achieve in the case of remnant phages (Table 3.3). Sanger sequencing revealed that most contigs where not closed due to the presence of repeats larger than read length, creating gaps in the assembly.

Most phages identified in the present study, showed a remarkable genetic synteny among themselves (Figure 3.5, table 3.1). However, in comparison with phage KHP30, the synteny was punctuated by deletions of certain genes which were replaced by additional IS throughout the prophage genome. When prophages are present, the tendency in *H. pylori* is to have just one prophage per genome, which is in accordance with the small genome size of *H. pylori*, which is expected to have less neutral targets for prophage integration. The presence of one prophage only is generally observed in *H. pylori*, but there are exceptions, like *H. acinonychis* strain sheeba, which is considered as a *H. pylori* like organism, that harbors two prophages, apparently one intact and another remnant (67). Furthermore, *H. pylori* has slow bacterial growth, and a population at low density provides few resources for the production of virions, favoring lysogeny (45).

Prophage ORFs were typically found in the same direction, which was opposite to that of the bacterial flanking genes. Concerning annotation most ORFs have an unknown function, as described for other species phages (135). Although no known virulence gene was found in prophage genomes, the role of prophages in the virulence of *H. pylori* should not be immediately discarded. Frequently phages do not code for toxin genes, as they are not able of directly convert their host into a toxin producer (136), but they can, however, indirectly modulate toxin production, such as TcdA and TcdB in *Clostridium difficille* (137).

Considering the bacterium's ecological niche, *H. pylori*'s persistence might be associated with both its broad genetic variability (138) and its capability of biofilm developing (139, 140). In both cases the presence of extracellular DNA (eDNA) is important, either as a source of DNA taken up by the naturally competent *H. pylori*, promoting recombination or contributing to biofilm development (140). Apart from outer membrane vesicle shedding, cell lysis via spontaneous prophage induction might be a source of eDNA release, contributing to survival and to the wide genomic variability of *H. pylori*.

The IS found in the present study were previously described in *H. pylori* but outside a prophage context (141). IS were described to be present in about one-third of a set of 238 independent isolates of the gastric pathogen *H. pylori* (142). Bacterial IS of IS200/IS605 and IS607 family often encode a transposase (TnpA) and a protein of unknown function, TnpB, which were hypothesized to act as a methyltransferase (143); furthermore, *orfB* is also related to the *Salmonella* virulence gene *gipA*, a *Salmonella* prophage gene which enhances bacterial growth in Peyer's patches (144).

As IS found within prophage sequences showed robust homology with those found in the *H. pylori* genome, it can be hypothesized that prophages mediate the transfer of IS, further

contributing to the genome plasticity of *H. pylori*. In contrast, we cannot exclude that the transfer of IS otherwise from bacteria to prophages may also be feasible. Remarkably, IS have been described in other prophages, including cyanophage Ma-LMM01, specifically-infecting *Microcystis aeruginosa* and mediating the transfer of IS607 to the bacterial genome (145). Besides prophages, IS605 is also associated with the *cag* pathogenicity island, dividing this island into two parts called *cagI* and *cagII* by insertion of one or two copies of IS605, providing intermediate phenotypes (146). Prophage inactivation should be under selection because lytic cycle induction may kill the cell. Correspondingly, we find five remnant prophages that might result from these evolutionary dynamics, even though defective prophages can still provide an adaptive function to bacteria (44). Recombination with incoming phages can also imprint a signal for purifying selection. In addition, IS present in prophages have been postulated to play a role in the inactivation and immobilization of incoming phages (147).

We showed that the prophage insertion sites can be diverse in *H. pylori* genomes although with some common traits among *H. pylori* populations, as discussed below. All prophages from hpNEurope from the present study and from *H. pylori* Cuz20 and India7 genomes (available at the NCBI), as well as most prophages from hpAfrica1 populations, have the same genomic context, presenting the bacterial genes S-adenosylmethionine synthetase and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase at the 5'end and 3' end, respectively. Interestingly, the prophages genomes integrated between these two loci usually present tandem repeats at the 3' end, between the last prophage gene and the first bacterial gene after the prophage (Table 3.5), most often in noncoding regions. DNA tandem repeats or satellite DNA, are inter- or intragenic nucleotide sequences repeated two or more times in a head-to-tail manner. Because these repeat tracts are prone to strand-slippage replication and recombination events causing their copy number to increase or decrease, *loci* containing tandem repeats are hypermutable (148). Tandem repeats may reversibly shut down or modulate the function of specific genes, allowing them to adapt to changing environments on short evolutionary time scales without an increased overall mutation rate. The environmental adaptability in *H. pylori* depends primarily on tandem repeat variations, which may cause gene phase switching. DNA tandem repeats may modulate gene expression affecting transcription initiation by modifying binding affinity of regulatory proteins (upstream of -35 site) or altering the distance to promoter elements (between -35 and -10 sites), modifying the affinity of regulatory proteins or mRNA stability (between the transcriptional start and an ORF). The most frequent bacterial gene at the 5' end of prophage codes for S-adenosylmethionine synthetase, which catalyzes the synthesis of AdoMet. AdoMet is an essential metabolic intermediate involved in many biochemical processes, such as a donor of methyl groups that allows DNA methylation (reviewed in (149)). Once DNA is methylated it may switch genes (150).

All hpEastAsian prophages either described in the present study or found in the genomes of *H. pylori* YN4-84, UM038, FD430 and UM114 Asian strains (available at the NCBI) were inserted in the same genomic region, including the competence protein ComGF, which plays a role in transformation and DNA binding, at the 5' end and a putative outer membrane protein at the 3' end. The gene at the 5' end is important for genetic variability of *H. pylori* (151), while *H. pylori* outer membrane proteins are known to mediate adherence to gastric epithelium, and ultimately are associated with clinical outcome of the infection (152). All things considered, the prophage insertion site may not be neutral for *H. pylori* gene expression and further studies are needed to evaluate the impact of prophage insertion on gene expression.

The phage Sw-A626-G was present as an episome, a non-integrated preprophage genome. This was also observed for the *H. pylori* phage KHP30 (62). Besides the presence of lytic and lysogenic life cycle phages may undergo a pseudolysogenic life cycle without either phage

replication or lysogenization. Moreover, phage episomes appear to be evolutionarily unstable (153). Pseudolysogeny has been described as complex form of phage–bacterium interaction associated with starved host cells (40), representing a mechanism to avoid abortive replication or integration in nutrient-deprived conditions (154). However, the *H. pylori* strain A626 presenting phage SwA626 was not laboratory exposed to any nutrient depletion and showed a usual pattern of cell growth, seeming that in this strain the phage Sw-A626-G presented a pseudolysogenic life cycle preferably, as suggested for phage KHP30 (62). But we cannot exclude that a previous exposition to unfavorable growth conditions turned this phage pseudolysogenic. There was an heterogeneous population of Sw-A626-G phage concerning the sequence closing the genome ends. In fact, the episome formed presented several repeats in the sequence joining the 5' and 3' end which may point to multiple integrations and excisions in the host genome leading to different phage episomal genome assemblies. Whether Sw-A626-A is a non-integratative temperate phage or may switch between lysogeny and pseudolysogeny remains to be found. Either way, episomal phages may mediate horizontal gene transfer (153).

Within a bacterial genome prophages are considered part of the accessory genome and are not considered part of the bacterial species core genome (103). The core and pan genome analysis was determined for the prophages genomes, since there must be a minimal amount of genes that make possible to complete the phage (lytic, lysogenic, or pseudoysogenic) life cycle. Given 30 prophages genomes (23 from the present study and 7 from public databases) 9 genes were found in all phages, constituting the core genome and representing 27% of the average number of genes per genome. Comparison of the core genome sequences is useful to clarify evolutionary relationships within prophages, including tree-like clonal evolution and non-tree-like evolution, due to homologous recombination (105). There are two obvious limitations in this analysis. The first is that there are several genes coding for hypothetical proteins whose function is not determined, and the second is that this determination was based on BLAST and due to the high sequence diversity it is possible that other genes playing similar roles were not detected. Thus we cannot rule out that other genes make part of the core genome. We found that the core genome within each phylogeographic group was larger, which reinforces the phylogeographic groups identified and is in agreement with a vertical transmission of prophages and recombination essentially within each group. However, it should be noted that the absence of a gene in the core genome of a certain phylogeographic group does not mean its absence in all prophages genomes of that group, as a presumable result of a phage decay process, such as punctual gene deletion (38).

In general, the phylogenetic analysis of intact prophages presents clusters according to prophage population structure (exceptions are discussed below), confirming our previous results obtained by prophage sequence typing (65). The prophage genomes cluster in four groups corresponding to the hpSWEurope, hpNEurope, hpAfrica1 and hpEastAsia phage populations. The strong phylogeographic signal of prophage genomes is in agreement with a model of co-evolution between the virus and its bacterial host. Indeed, prophages and bacteria are linked by a long history of co-evolution, but the genetic dimension of this co-evolution cannot be defined at present (38). The phylogeographic clustering was in agreement with integration sites of prophages (discussed above). As suggested by others (155), this could be explained by a vertical transmission of the phage rather than by random insertions which are common to prophages.

Phage evolution is driven by a horizontal exchange of functional modules between more or less related phages, achieved by DNA recombination, explaining the genomic mosaicism among phages (156). Recombination is a factor of rapid variability in *H. pylori*, which is among the most recombinogenic known pathogens (157). In parallel, in the present study, phage genomes were shown to be prone to recombination events. Indeed several prophage genome mosaics

were detected, involving, for the vast majority of the cases, both hpAfrica1 and hpSWEurope populations. This is not surprising considering that both populations were detected in the same geographic area. Nevertheless, most phage ORFs are of unknown function, so no assumptions can be performed regarding a putative impact of these recombination events on pathogenicity. These mosaic structures also highlight the need for a prudent use of the PST-based classification. In fact, although an agreement is observed for most of the cases, for the studied mosaic structures, for some of the studied mosaic structures only the integrase and/or the holin genes appeared to support the PST-based classification.

The remnant prophages encountered in the present study as well as in other *H. pylori* strains (64, 70) and in non-*pylori Helicobacter* species (72) highlight an evolutionary scenario consistent with a prophage decay process during the complex interaction between *H. pylori* and the prophage. However, a model in which *H. pylori* strains from different geographical regions may have been infected by distinct phage lineages after the geographic separation of the bacterial host is also feasible (59), but less likely due to the high genetic synteny between prophages from different geographic areas. Altogether, the integration at the same locus and a gene repertoire relatedness points to a vertical transmission, suggesting the so called pervasive domestication of prophages by the bacterial host which may drive bacterial adaptation (44). Remarkably, the most divergent *H. pylori* prophage population (hpSWEurope), presented neither conserved loci for integration site nor IS.

5. Conclusion

This work not only provides a compendium of novel sequences, but also sets the stage for future studies aimed at better understanding the virus-host relationship. Results of the present study showed that prophages are more common in *H. pylori* than initially expected and that, in most cases, prophages appear to be intact, with a sequence size of over 20 Kb. Remarkably, we show for the first time that for phages classified as hpNEurope, hpAfrica and hpEastAsia, the insertion site appears to be preserved. Furthermore, the phylogenetic analysis for a vast majority of phage genomes is similar to the phylogenetic analysis previously presented by our team (65) using two phage genes (integrase and holin), confirming our previous findings and reinforcing the hypothesis of co-evolution between prophages and *H. pylori*. Some recombinant phages were found, suggesting additional genetic diversity that hypothetically may provide *H. pylori* with advantageous phenotypes. Major challenges at present are to identify the function of prophage genes, to understand if the insertion site is neutral for the host and whether prophage presence plays a role in the adaptation of *H. pylori* to its host, or if prophage genes belonging to the lysis cassette are useful for biomedical applications, namely phage therapy.

6. References

(1) Kusters JG, van Vliet AH, Kuipers EJ. Pathogenesis of *Helicobacter pylori* infection. Clin Microbiol Rev 2006 Jul;19(3):449-90.

(2) International Agency for Research on Cancer. IARC monographs on the evaluation of carcinogenic risks to humans. Lyon, France: International Agency for Research on Cancer, 1994.

(3) Vale FF, Vitor JM. Transmission pathway of *Helicobacter pylori*: does food play a role in rural and urban areas? Int J Food Microbiol 2010 Mar 31;138(1-2):1-12.

(4) Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 1984 Jun 16;1(8390):1311-5.

(5) Lopes AI, Vale FF, Oleastro M. *Helicobacter pylori* infection - recent developments in diagnosis. World J Gastroenterol 2014 Jul 28;20(28):9299-313.

(6) Malfertheiner P, Megraud F, O'Morain CA, Gisbert JP, Kuipers EJ, Axon AT, et al. Management of *Helicobacter pylori* infection-the Maastricht V/Florence Consensus Report. Gut 2017 Jan;66(1):6-30.

(7) Megraud F, Lehours P. *Helicobacter pylori* detection and antimicrobial susceptibility testing. Clin Microbiol Rev 2007 Apr;20(2):280-322.

(8) Graham DY, Lee YC, Wu MS. Rational *Helicobacter pylori* therapy: evidence-based medicine rather than medicine-based evidence. Clin Gastroenterol Hepatol 2014 Feb;12(2):177-86.

(9) Megraud F, Gisbert JP. Towards effective empirical treatment for *Helicobacter pylori* eradication. Lancet 2016 Nov 12;388(10058):2325-6.

(10) WHO. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. 2017.

(11) Praszkier J, Sutton P, Ferrero RL. Virulence mechanisms in *Helicobacter pylori*: an overview. In: Backert S, Yamaoka Y, editors. *Helicobacter pylori* Research From Bench to Bedside. Springer, 2016. p. 57-87.

(12) Nakazawa T. Growth cycle of *Helicobacter pylori* in gastric mucous layer. Keio J Med 2002 Dec;51 Suppl 2:15-9.

(13) Stingl K, Altendorf K, Bakker EP. Acid survival of *Helicobacter pylori*: how does urease activity trigger cytoplasmic pH homeostasis? Trends Microbiol 2002 Feb;10(2):70-4.

(14) Hessey SJ, Spencer J, Wyatt JI, Sobala G, Rathbone BJ, Axon AT, et al. Bacterial adhesion and disease activity in Helicobacter associated chronic gastritis. Gut 1990 Feb;31(2):134-8.

(15) Alm RA, Bina J, Andrews BM, Doig P, Hancock RE, Trust TJ. Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. Infect Immun 2000 Jul;68(7):4155-68.

(16) Cover TL, Holland RL, Blanke SR. *Helicobacter pylori* vacuolating toxin. In: Backert S, Yamaoka Y, editors. *Helicobacter pylori* Research From Bench to Bedside. Springer, 2016. p. 113-41.

(17) Wessler S. Emerging novel virulence factors of *Helicobacter pylori*. In: Backert S, Yamaoka Y, editors. *Helicobacter pylori* Research From Bench to Bedside. 2016. p. 165-88.

(18) Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, et al. Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. Mol Microbiol 1999 May;32(3):459-70.

(19) Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human migrations in *Helicobacter pylori* populations. Science 2003 Mar 7;299(5612):1582-5.

(20) Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, et al. An African origin for the intimate association between humans and *Helicobacter pylori*. Nature 2007 Feb 22;445(7130):915-8.

(21) Suerbaum S, Achtman M. *Helicobacter pylori*: recombination, population structure and human migrations. Int J Med Microbiol 2004 Sep;294(2-3):133-9.

(22) Suzuki R, Shiota S, Yamaoka Y. Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. Infect Genet Evol 2012 Mar;12(2):203-13.

(23) Megraud F, Lehours P, Vale FF. The history of *Helicobacter pylori*: from phylogeography to paleomicrobiology. Clin Microbiol Infect 2016 Jul 21.

(24) Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu JY, et al. The peopling of the Pacific from a bacterial perspective. Science 2009 Jan 23;323(5913):527-30.

(25) Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Abridged Paperback Edition. New Jersey: Princeton University Press, 1996.

(26) deMenocal PB, Stringer C. Human migration: Climate and the peopling of the world. Nature 2016 Oct 6;538(7623):49-50.

(27) Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, Schlebusch CM, et al. Age of the association between *Helicobacter pylori* and man. PLoS Pathog 2012;8(5):e1002693.

(28) Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E, et al. Free recombination within *Helicobacter pylori*. Proc Natl Acad Sci U S A 1998 Oct 13;95(21):12619-24.

(29) Moodley Y. *Helicobacter pylori*: genetics, recombination, population structure, and human migrations. In: Backert S, Yamaoka Y, editors. *Helicobacter pylori* Research From Bench to Bedside. Springer, 2016. p. 3-27.

(30) Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000 Jun;155(2):945-59.

(31) Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet 2012 Jan;8(1):e1002453.

(32) Yahara K, Furuta Y, Oshima K, Yoshida M, Azuma T, Hattori M, et al. Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol 2013 Jun;30(6):1454-64.

(33) Kobayashi I. Genome evolution: *Helicobacter pylori* as an extreme model. In: Backert S, Yamaoka Y, editors. *Helicobacter pylori* Research From Bench to Bedside. Springer, 2016. p. 217-31.

(34) Breurec S, Guillard B, Hem S, Brisse S, Dieye FB, Huerre M, et al. Evolutionary history of *Helicobacter pylori* sequences reflect past human migrations in Southeast Asia. PLoS ONE 2011;6(7):e22058.

(35) Oleastro M, Rocha R, Vale FF. Population genetic structure of *Helicobacter pylori* strains from Portuguese-speaking countries. Helicobacter 2017 Mar 8.

(36) Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, et al. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 1998 May;62(5):1137-52.

(37) Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, et al. The 5300-year-old *Helicobacter pylori* genome of the Iceman. Science 2016 Jan 8;351(6269):162-5.

(38) Brussow H, Canchaya C, Hardt WD. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 2004 Sep;68(3):560-602.

(39) Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. Nat Rev Microbiol 2015 Oct;13(10):641-50.

(40) Golais F, Holly J, Vitkovska J. Coevolution of bacteria and their viruses. Folia Microbiol (Praha) 2013 May;58(3):177-86.

(41) Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic prophages help bacteria cope with adverse environments. Nat Commun 2010;1:147.

(42) Wang X, Wood TK. Cryptic prophages as targets for drug development. Drug Resist Updat 2016 Jul;27:30-8.

(43) Gama JA, Reis AM, Domingues I, Mendes-Soares H, Matos AM, Dionisio F. Temperate bacterial viruses as double-edged swords in bacterial warfare. PLoS ONE 2013;8(3):e59043.

(44) Bobay LM, Touchon M, Rocha EP. Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci U S A 2014 Aug 19;111(33):12127-32.

(45) Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J 2016 Mar 25.

(46) Vale FF, Vitor JM. Genomic Methylation: a Tool for Typing *Helicobacter pylori* Isolates. Appl Environ Microbiol 2007 Jul;73(13):4243-9.

(47) Vale FF, Encarnacao P, Vitor JM. A new algorithm for cluster analysis of genomic methylation: the *Helicobacter pylori* case. Bioinformatics 2008 Feb 1;24(3):383-8.

(48) Backert S, Kwok T, Konig W. Conjugative plasmid DNA transfer in *Helicobacter pylori* mediated by chromosomally encoded relaxase and TraG-like proteins. Microbiology 2005 Nov;151(Pt 11):3493-503.

(49) Kuipers EJ, Israel DA, Kusters JG, Blaser MJ. Evidence for a conjugation-like mechanism of DNA transfer in *Helicobacter pylori*. J Bacteriol 1998 Jun;180(11):2901-5.

(50) Haas R, Meyer TF, van Putten JP. Aflagellated mutants of *Helicobacter pylori* generated by genetic transformation of naturally competent strains using transposon shuttle mutagenesis. Mol Microbiol 1993 May;8(4):753-60.

(51) Stingl K, Muller S, Scheidgen-Kleyboldt G, Clausen M, Maier B. Composite system mediates two-step DNA uptake into *Helicobacter pylori*. Proc Natl Acad Sci U S A 2010 Jan 19;107(3):1184-9.

(52) Brussow H, Kutter E. Phage ecology. In: Kutter E, Sulakvelidze A, editors. Bacteriophages biology and applications.London: CRC Press, 2005. p. 129-63.

(53) Marshall BJ, Armstrong JA, Francis GJ, Nokes NT, Wee SH. Antibacterial action of bismuth in relation to Campylobacter pyloridis colonization and gastritis. Digestion 1987;37 Suppl 2:16-30.

(54) Goodwin CS, Armstrong JA, Peters M. Microbiology of *C. pylori*. In: Blaser MJ, editor. *Campylobacter pylori* in gastritis and peptic ulcer disease.New York: MD.IGAKU-SHOIN, 1989. p. 25-49.

(55) Vale FF, Alves Matos AP, Carvalho P, Vitor JM. *Helicobacter pylori* phage screening. Microsc Microanal 2008;14 (supp 3):150-1.

(56) Schmid EN, von RG, Ansorg R. Bacteriophages in *Helicobacter (Campylobacter) pylori*. J Med Microbiol 1990 Jun;32(2):101-4.

(57) Heintschel von HE, Nalik HP, Schmid EN. Characterisation of a *Helicobacter pylori* phage (HP1). J Med Microbiol 1993 Apr;38(4):245-9.

(58) Thibergue JM, Lehours P, Bouchier C, Ma L, Creno S, Medigue C, et al. Sequence of the first *Helicobacter pylori* strains involved in low-grade Mucosa-Associated Lymphoid Tissue (MALT) Lymphoma. Helicobacter 2006;11:02.01.

(59) Lehours P, Vale FF, Bjursell MK, Melefors O, Advani R, Glavas S, et al. Genome sequencing reveals a phage in *Helicobacter pylori*. MBio 2011;2(6):pii: e00239-11.

(60) Luo CH, Chiou PY, Yang CY, Lin NT. Genome, integration and transduction of a novel temperate phage of *Helicobacter pylori*. J Virol 2012 Jun 13.

(61) Uchiyama J, Takeuchi H, Kato S, Takemura-Uchiyama I, Ujihara T, Daibata M, et al. Complete Genome Sequences of Two *Helicobacter pylori* Bacteriophages Isolated from Japanese Patients. J Virol 2012 Oct;86(20):11400-1.

(62) Uchiyama J, Takeuchi H, Kato S, Gamoh K, Takemura-Uchiyama I, Ujihara T, et al. Characterization of *Helicobacter pylori* bacteriophage KHP30. Appl Environ Microbiol 2013 May;79(10):3176-84.

(63) You Y, He L, Zhang M, Zhang J. Comparative genomics of a *Helicobacter pylori* isolate from a Chinese Yunnan Naxi ethnic aborigine suggests high genetic divergence and phage insertion. PLoS ONE 2015;10(3):e0120659.

(64) Fan X, Li Y, He R, Li Q, He W. Comparative analysis of prophage-like elements in Helicobacter sp. genomes. PeerJ 2016;4:e2012.

(65) Vale FF, Vadivelu J, Oleastro M, Breurec S, Engstrand L, Perets TT, et al. Dormant phages of *Helicobacter pylori* reveal distinct populations in Europe. Scientific Reports 2015;5:14333.

(66) Kyrillos A, Arora G, Murray B, Rosenwald AG. The Presence of Phage Orthologous Genes in *Helicobacter pylori* Correlates with the Presence of the Virulence Factors CagA and VacA. Helicobacter 2015 Nov 27.

(67) Eppinger M, Baar C, Linz B, Raddatz G, Lanz C, Keller H, et al. Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. PLoS Genet 2006 Jul;2(7):e120.

(68) Arnold IC, Zigova Z, Holden M, Lawley TD, Rad R, Dougan G, et al. Comparative whole genome sequence analysis of the carcinogenic bacterial model pathogen Helicobacter felis. Genome Biol Evol 2011;3:302-8.

(69) Schott T, Kondadi PK, Hanninen ML, Rossi M. Comparative genomics of *Helicobacter pylori* and the human-derived *Helicobacter bizzozeronii* CIII-1 strain reveal the molecular basis of the zoonotic nature of non-pylori gastric *Helicobacter* infections in humans. BMC Genomics 2011;12:534.

(70) Thiberge JM, Boursaux-Eude C, Lehours P, Dillies MA, Creno S, Coppee JY, et al. From array-based hybridization of *Helicobacter pylori* isolates to the complete genome sequence of an isolate associated with MALT lymphoma. BMC Genomics 2010;11:368.

(71) Uchiyama J, Takemura-Uchiyama I, Kato S, Takeuchi H, Sakaguchi Y, Ujihara T, et al. Screening of KHP30-like prophages among Japanese *Helicobacter pylori* strains, and genetic analysis of a defective KHP30-like prophage sequence integrated in the genome of the *H. pylori* strain NY40. FEMS Microbiol Lett 2016 Aug;363(16).

(72) Kersulyte D, Rossi M, Berg DE. Sequence divergence and conservation in genomes of *Helicobacter cetorum* strains from a dolphin and a whale. PLoS ONE 2013;8(12):e83177.

(73) Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol 2016 Dec;56(4):394-404.

(74) Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010 Jan;11(1):31-46.

(75) Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977 Dec;74(12):5463-7.

(76) Tipu HN, Shabbir A. Evolution of DNA sequencing. J Coll Physicians Surg Pak 2015 Mar;25(3):210-5.

(77) Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012;2012:251364.

(78) Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998 Mar;8(3):175-85.

(79) Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010 Apr;38(6):1767-71.

(80) Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell 2015 May 21;58(4):586-97.

(81) Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res 2016 Jan 4;44(D1):D20-D26.

(82) El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol 2013;9(12):e1003345.

(83) Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 2017 Feb 10;243:16-24.

(84) Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010 Jun;95(6):315-27.

(85) Guo H, Arambula D, Ghosh P, Miller JF. Diversity-generating Retroelements in Phage and Bacterial Genomes. Microbiol Spectr 2014 Dec;2(6).

(86) Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008 May;18(5):821-9.

(87) Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 2012 Mar;22(3):557-67.

(88) Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 2011 Nov 8;29(11):987-91.

(89) Chin FY, Leung HC, Yiu SM. Sequence assembly using next generation sequencing data--challenges and solutions. Sci China Life Sci 2014 Nov;57(11):1140-8.

(90) Lawrie DS, Petrov DA. Comparative population genomics: power and principles for the inference of functionality. Trends Genet 2014 Apr;30(4):133-9.

(91) Hu B, Xie G, Lo CC, Starkenburg SR, Chain PS. Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. Brief Funct Genomics 2011 Nov;10(6):322-33.

(92) Siezen RJ, van Hijum SA. Genome (re-)annotation and open-source annotation pipelines. Microb Biotechnol 2010 Jul;3(4):362-9.

(93) Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 2014 Jan;42(Database issue):D206-D214.

(94) Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999 Dec 1;27(23):4636-41.

(95) Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998 Feb 15;26(4):1107-15.

(96) Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res 2017 Jan 4;45(D1):D37-D42.

(97) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 2011 Jan;39(Database issue):D214-D219.

(98) Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997 Sep 1;25(17):3389-402.

(99) Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SA. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics 2013 Jul;12(4):366-80.

(100) Lu W, Wise MJ, Tay CY, Windsor HM, Marshall BJ, Peacock C, et al. Comparative analysis of the full genome of *Helicobacter pylori* isolate Sahul64 identifies genes of high divergence. J Bacteriol 2014 Mar;196(5):1073-83.

(101) Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc Natl Acad Sci U S A 2000 Dec 19;97(26):14668-73.

(102) Gressmann H, Linz B, Ghai R, Pleissner KP, Schlapbach R, Yamaoka Y, et al. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. PLoS Genet 2005 Oct;1(4):e43.

(103) Ali A, Naz A, Soares SC, Bakhtiar M, Tiwari S, Hassan SS, et al. Pan-genome analysis of human gastric pathogen *H. pylori*: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. Biomed Res Int 2015;2015:139580.

(104) Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 2015 Sep;7:72-85.

(105) Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to *Helicobacter pylori* Pan-Genome Analysis for Identification of Genomic Islands. PLoS ONE 2016;11(8):e0159419.

(106) Bansal AK. Bioinformatics in microbial biotechnology--a mini review. Microb Cell Fact 2005 Jun 28;4:19.

(107) Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003 Jul 1;31(13):3497-500.

(108) Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013 Apr;30(4):772-80.

(109) Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res 2011 Jul;39(Web Server issue):W347-W352.

(110) Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ 2015;3:e985.

(111) Bose M, Barber RD. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biol 2006;6(3):223-7.

(112) Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 2006;34(20):5839-51.

(113) Lima-Mendez G, Van HJ, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 2008 Mar 15;24(6):863-5.

(114) Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 2016 Jul 8;44(W1):W16-W21.

(115) Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res 2012 Sep;40(16):e126.

(116) Oliver SL, Lenards AJ, Barthelson RA, Merchant N, McKay SJ. Using the iPlant collaborative discovery environment. Curr Protoc Bioinformatics 2013 Jun;Chapter 1:Unit1.

(117) Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform 2013 Mar;14(2):193-202.

(118) Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 2008;9:75.

(119) Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 2014 Feb;15(2):121-32.

(120) Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res 2012 Aug;40(15):e115.

(121) Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 1988 Nov 25;16(22):10881-90.

(122) Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 1999 Jan 14;397(6715):176-80.

(123) Gelfand Y, Rodriguez A, Benson G. TRDB--the Tandem Repeats Database. Nucleic Acids Res 2007 Jan;35(Database issue):D80-D87.

(124) Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 2010;5(6):e11147.

(125) Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. Package 'gplots'. 2016.

(126) Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013 Dec;30(12):2725-9.

(127) Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980 Dec;16(2):111-20.

(128) Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 2006 Feb;23(2):254-67.

(129) Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003 Aug;164(4):1567-87.

(130) Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 2007 Jul 1;7(4):574-8.

(131) Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol 2015;1(1):vev003.

(132) Gomes JP, Bruno WJ, Nunes A, Santos N, Florindo C, Borrego MJ, et al. Evolution of Chlamydia trachomatis diversity occurs by widespread interstrain recombination involving hotspots. Genome Res 2007 Jan;17(1):50-60.

(133) Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. Prophage genomics. Microbiol Mol Biol Rev 2003 Jun;67(2):238-76, table.

(134) Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003 Aug;164(4):1567-87.

(135) Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Elife 2015;4:e06416.

(136) Fortier LC, Sekulovic O. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 2013 Jul 1;4(5):354-65.

(137) Goh S, Chang BJ, Riley TV. Effect of phage infection on toxin production by Clostridium difficile. J Med Microbiol 2005 Feb;54(Pt 2):129-35.

(138) Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. PLoS Genet 2010 Jul;6(7):e1001036.

(139) Grande R, Di Marcantonio MC, Robuffo I, Pompilio A, Celia C, Di ML, et al. *Helicobacter pylori* ATCC 43629/NCTC 11639 Outer Membrane Vesicles (OMVs) from Biofilm and Planktonic Phase Associated with Extracellular DNA (eDNA). Front Microbiol 2015;6:1369.

(140) Grande R, Di GM, Bessa LJ, Di CE, Baffoni M, Guarnieri S, et al. Extracellular DNA in *Helicobacter pylori* biofilm: a backstairs rumour. J Appl Microbiol 2011 Feb;110(2):490-8.

(141) Kalia A, Mukhopadhyay AK, Dailide G, Ito Y, Azuma T, Wong BC, et al. Evolutionary dynamics of insertion sequences in *Helicobacter pylori*. J Bacteriol 2004 Nov;186(22):7508-20.

(142) 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 Nov 26;223(1-2):175-86.

(143) Bao W, Jurka J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. Mob DNA 2013;4(1):12.

(144) Kersulyte D, Kalia A, Zhang M, Lee HK, Subramaniam D, Kiuduliene L, et al. Sequence organization and insertion specificity of the novel chimeric ISHp609 transposable element of *Helicobacter pylori*. J Bacteriol 2004 Nov;186(22):7521-8.

(145) Kuno S, Yoshida T, Kamikawa R, Hosoda N, Sako Y. The distribution of a phage-related insertion sequence element in the cyanobacterium, Microcystis aeruginosa. Microbes Environ 2010;25(4):295-301.

(146) Censini S, Lange C, Xiang Z, Crabtree JE, Ghiara P, Borodovsky M, et al. cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. Proc Natl Acad Sci U S A 1996 Dec 10;93(25):14648-53.

(147) Ooka T, Ogura Y, Asadulghani M, Ohnishi M, Nakayama K, Terajima J, et al. 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 Res 2009 Oct;19(10):1809-16.

(148) Zhou K, Aertsen A, Michiels CW. The role of variable DNA tandem repeats in bacterial adaptation. FEMS Microbiol Rev 2014 Jan;38(1):119-41.

(149) Kozbial PZ, Mushegian AR. Natural history of S-adenosylmethionine-binding proteins. BMC Struct Biol 2005;5:19.

(150) Vitoriano I, Vitor JM, Oleastro M, Roxo-Rosa M, Vale FF. Proteome variability among *Helicobacter pylori* isolates clustered according to genomic methylation. J Appl Microbiol 2013 Jun;114(6):1817-32.

(151) Baltrus DA, Guillemin K, Phillips PC. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. Evolution 2008 Jan;62(1):39-49.

(152) Oleastro M, Cordeiro R, Yamaoka Y, Queiroz D, Megraud F, Monteiro L, et al. Disease association with two *Helicobacter pylori* duplicate outer membrane protein genes, homB and homA. Gut Pathog 2009;1(1):12.

(153) Xue H, Cordero OX, Camas FM, Trimble W, Meyer F, Guglielmini J, et al. Eco-Evolutionary Dynamics of Episomes among Ecologically Cohesive Bacterial Populations. MBio 2015 May 5;6(3):e00552-15.

(154) Cenens W, Makumi A, Mebrhatu MT, Lavigne R, Aertsen A. Phage-host interactions during pseudolysogeny: Lessons from the Pid/dgo interaction. Bacteriophage 2013 Jan 1;3(1):e25029.

(155) Bobay LM, Rocha EP, Touchon M. The adaptation of temperate bacteriophages to their host genomes. Mol Biol Evol 2013 Apr;30(4):737-51.

(156) Desiere F, Lucchini S, Brussow H. Evolution of Streptococcus thermophilus bacteriophage genomes by modular exchanges followed by point mutations and small deletions and insertions. Virology 1998 Feb 15;241(2):345-56.

(157) Go MF, Kapur V, Graham DY, Musser JM. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. J Bacteriol 1996 Jul;178(13):3934-8.

Appendix 1 - Phyton script to compute the depth coverage

```
#Script for coverage determination
#Opens for reading only Velvet output FASTA file with the genome assembly
#In this example the file is #C520_cov41_kmer71_cutoff20.fasta
fp=open('C520_cov41_kmer71_cutoff20.txt', 'r')

#Creates a new txt document for writing
out=open('C520.txt', 'w')

#Copies the title of each contig to the new document
for line in fp:
     if '>' in line:
        out.write(line+'\n')

# Closes all documents
fp.close()
out.close()

#The new document will have writen the contigs names from
#the first to the last contig. For example:
#>NODE_1_length_2156_cov_82.203156
#...
#>NODE_92_length_889_cov_12.386951

#Opens the new created document for reading
fp=open('C520.txt', 'r')

#Creates an empty list
lista=[]

#Split the string in text on delimiter: '_' and removes paragraphs
for line in fp:
    lista.append(line.split('_'))

for i in lista:
    if i==['\n']:
        lista.remove(i)
#Creates variable for genome size
genomeSize=0

#Calculates the genome size by summing each contig length
for i in lista:
    genomeSize=genomeSize +int(i[3])

#Receives variables for k-mer and read length
kmer=71
readL=100
cov=0
j=0

#Calculates the coverage according to Velvet formula:
#Ck = C * (L - k + 1)/L
for i in lista:
    Cbase=float(i[5])*readL/(readL-kmer+1)
    cov=cov+Cbase
    j=j+1

Coverage=cov/j
```

```
#Closes document
fp.close()

#Opens document for adding
fp=open('C520.txt', 'a')

#Writes genome size and coverage
fp.write('genomeSize=' + str(genomeSize) + '\n')
fp.write('Depth Coverage=' + str(Coverage))
#Closes document
fp.close()
```

>NODE_1_length_2156_cov_82.203156

>NODE_2_length_42453_cov_40.328693

>NODE_3_length_54344_cov_37.447849

>NODE_4_length_613_cov_37.659054

>NODE_6_length_11495_cov_45.267944

>NODE_7_length_32090_cov_45.287785

>NODE_8_length_3190_cov_1307.077148

>NODE_9_length_57057_cov_43.912086

>NODE_10_length_27198_cov_39.821274

>NODE_11_length_38123_cov_42.857620

>NODE_12_length_1960_cov_1767.798462

>NODE_13_length_33884_cov_41.335705

>NODE_14_length_24175_cov_37.459358

>NODE_15_length_11568_cov_35.202541

>NODE_16_length_61985_cov_38.662209

>NODE_17_length_17202_cov_45.539005

>NODE_18_length_13024_cov_36.348434

>NODE_19_length_86543_cov_37.923229

>NODE_20_length_31616_cov_45.186520

>NODE_22_length_46737_cov_39.065578

>NODE_23_length_1709_cov_161.066116

>NODE_24_length_115525_cov_46.314728

>NODE_25_length_22823_cov_42.639488

>NODE_26_length_962_cov_1170.755737

>NODE_27_length_1446_cov_405.703308

>NODE_28_length_2024_cov_39.708500

>NODE_29_length_32793_cov_35.923031

>NODE_30_length_24390_cov_40.916359

>NODE_31_length_1054_cov_66.569260

>NODE_32_length_95381_cov_47.993038

>NODE_33_length_58783_cov_38.828930

>NODE_34_length_49165_cov_40.940548

>NODE_35_length_3482_cov_41.577255

>NODE_36_length_25952_cov_40.584465

>NODE_37_length_4474_cov_92.938980

>NODE_38_length_11791_cov_48.518700

>NODE_39_length_46419_cov_40.621578

>NODE_40_length_14878_cov_41.463905

>NODE_41_length_4859_cov_51.414902

>NODE_42_length_98707_cov_36.883038

>NODE_44_length_942_cov_92.468155

>NODE_45_length_1666_cov_94.282715

>NODE_46_length_13942_cov_34.781452

>NODE_47_length_15059_cov_45.976292

>NODE_48_length_2175_cov_101.236320

>NODE_49_length_21918_cov_44.446209

>NODE_50_length_66441_cov_36.728527

>NODE_51_length_5397_cov_34.730034

>NODE_52_length_20303_cov_46.252720

>NODE_53_length_454_cov_94.718063

>NODE_54_length_5469_cov_37.054489

>NODE_55_length_1823_cov_143.103668

>NODE_56_length_491_cov_79.896133

>NODE_57_length_54581_cov_38.718365

>NODE_59_length_5828_cov_45.054394

>NODE_60_length_69113_cov_48.305660

>NODE_62_length_4510_cov_44.189137

>NODE_64_length_466_cov_20.896996

>NODE_65_length_622_cov_99.990356

>NODE_66_length_2071_cov_37.326412

>NODE_67_length_1335_cov_35.111610

>NODE_68_length_3196_cov_77.722466

>NODE_69_length_8971_cov_30.335526

>NODE_72_length_1000_cov_47.855000

>NODE_74_length_1157_cov_39.235954

>NODE_77_length_654_cov_104.415901

>NODE_80_length_607_cov_97.517296

>NODE_82_length_130337_cov_40.608700

>NODE_91_length_13558_cov_35.593746

>NODE_92_length_889_cov_12.386951

genomeSize=1669005
Depth Coverage=386.604036333

Appendix 2 - Phyton script to compute pangenome and core genome

```
##name of the document to be created
text='AfricaProphagesPanCore'

##ORF code for prophage genomes

UKEN31U = [1, 2, 3, 4, 31, 32, 7, 8, 9, 10, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
           15, 16, 17, 37, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
UKEN32U = [1, 2, 3, 4, 31, 32, 7, 8, 9, 10, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
           15, 16, 17, 37, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]
DeM53M = [1, 38, 2, 3, 4, 5, 6, 8, 9, 10, 11, 36, 12, 39, 13, 14, 15, 16, 17,\
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 29, 30]
Sw577G = [40, 1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 33, 34, 15, 16, 17, 18, 19,\
          20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 41]
SwA626G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 42, 33, 34, 35, 36, 12, 13,\
           14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtB89G = [1, 38, 2, 3, 4, 5, 6, 8, 8, 10, 11, 11, 36, 12, 13, 14, 15, 16, 17, \
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt1293U = [1, 38, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 11, 36, 12, 13, 14, 43, 44, \
           14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrANT170U = [1, 38, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 45, 46, 13, 14, 15, 16,\
             46, 45, 16, 17, 18, 47, 20, 21, 22, 23, 24, 25, 26, 27, 28, 28, 29,\
             30]
FrMEG235U = [1, 38, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 45, 46, 13, 14, 15, 16,\
             46, 45, 16, 17, 18, 47, 20, 21, 22, 23, 24, 25, 26, 27, 28, 28, 29,\
             30]
Pt5771G = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14, 15,\
           16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt5322G = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 18,\
           19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtB92G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 32, 31, 10, 11, 33, 34, 35, 36, 12,\
          13, 14, 15, 16, 17, 48, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4481G = [11, 10, 9, 8, 6, 5, 4, 3, 2, 38, 1, 11, 36, 12, 39, 13, 14, 15, 16, \
           17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrGC43G = [1, 2, 3, 4, 49, 5, 6, 8, 9, 31, 32, 10, 11, 33, 34, 35, 36, 12, 13, \
           14, 15, 16, 17, 18, 19, 20, 21, 32, 31, 22, 23, 24, 25, 26, 27, 28, 29, 30, 50]
FrG12G = [1, 2, 2, 3, 4, 49, 5, 6, 51, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
          15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrB58M = [1, 2, 43, 44, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,\
          23, 24, 25, 26, 27, 28, 29, 30]
Pt1918U = [1, 2, 3, 4, 49, 5, 6, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14, 15, \
           16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4497U = [52, 1, 2, 3, 4, 49, 5, 51, 8, 9, 10, 11, 33, 34, 35, 36, 12, 53, 13, \
           14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4472G = [1, 2, 3, 4, 5, 7, 8, 8, 9, 10, 11, 11, 36, 12, 13, 14, 15, 16, 17,\
           18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtR21299U = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 24, 18,\
             26, 27, 28, 29, 30]
FrB41M = [1, 2, 3, 4, 49, 5, 6, 8, 9, 10, 11, 33, 34, 35, 36, 12, 53, 13, 14, \
          15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 55]
Pt1846U = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, \
           18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt22899G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 46, 45, 9, 10, 11, 36, 12, 13, 14, 46, \
            45, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, \
            30, 54]
KHP30 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
         21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
KHP40 = [1, 2, 3, 4, 5, 6, 8, 9, 9, 10, 11, 12, 39, 13, 14, 15, 16, 17, 37, \
         18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 52]
```

```
India7 = [1, 38, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, \
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
phiHP33 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 11, 34, 36, 12, 13, 14, 15, 16, \
           46, 45, 25, 26, 27, 28, 29, 30]
P1961P = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11, 36, 12, 39, 13, 14, 15, 16,\
          17, 18, 19, 20, 21, 22, 23, 24, 25, 25, 26, 27, 29, 30]
Cuz20 = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 37, \
         18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Hac = [1, 10, 2, 18, 5, 5, 7, 8, 9, 10, 11, 33, 34, 36, 12, 13, 14, 15, 16, \
       17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30]


##ALL Phages
##p=[UKEN31U,UKEN32U,DeM53M,Sw577G,SwA626G,PtB89G,Pt1293U, \
##    FrANT170U,FrMEG235U,Pt5771G,Pt5322G,PtB92G,Pt4481G,\
##    FrGC43G,FrG12G,FrB58M,Pt1918U,Pt4497U,Pt4472G,PtR21299U, \
##    FrB41M,Pt1846U,Pt22899G,KHP30,KHP40,India7,phiHP33,P1961P, \
##    Cuz20,Hac]


##p list has the phages to be considered in the analysis
##SWEurope phages
##p=[Pt1918U,Pt4497U,Pt4472G,FrB41M]

##NEurope phages
##p=[UKEN31U,UKEN32U,DeM53M,Sw577G,SwA626G,India7,Cuz20]

##EastAsia phages
##p=[KHP30,KHP40,P1961P,FrB58M, FrGC43G,FrG12G]

##Africa1 phages
p=[PtB89G,Pt1293U,FrANT170U,FrMEG235U,Pt5771G,Pt5322G,PtB92G,Pt4481G,\
   PtR21299U, Pt1846U,Pt22899G,phiHP33]


##The sets module provides classes for constructing and manipulating unordered
##collections of unique elements. Common uses include membership testing,
##removing duplicates from a sequence, and computing standard math operations
##on sets such as intersection, union, difference, and symmetric difference.

##result is composed by the first element of the p list
result = set(p[0])
une = set ([])

##print result

##For the first element of p list makes a list with not repeated ORF
uniqueValues=[]
for i in p[0]:
    if i not in uniqueValues:
        uniqueValues.append(i)

##Finds commun genes to ALL phages in p list and
##counts regression of core genome by adding new genome
regressionCore=[]
regressionCore.append(len(uniqueValues))
for s in p[1:]:
    result.intersection_update(s)
```

```
    regressionCore.append(len(result))

print result
print regressionCore
##cheking step
print len(p)
print len(regressionCore)


##Finds all ORF - pan genome of phages and
##counts progression of core genome by adding new genome

test=set(p[0])

al=[]

progressionPan=[]

progressionPan.append(len(uniqueValues))

print progressionPan

for i in p[1:]:

    al=test.union(i)
    test=set(al)
    progressionPan.append(len(al))

print al
print progressionPan
##cheking step
print len(progressionPan)


##Calculates the number of ORF that are not in the p list
notInGroup=[]
first=1
AllORF=[]

while int(first)<56:
    AllORF.append(first)
    first =first + 1

##print AllORF

for i in AllORF:
    if i not in al:
        notInGroup.append(i)

print notInGroup

##Print the trend of the pangenome
for i in progressionPan:
    print i

print 'take a break'

for i in regressionCore:
```

```
    print i


##Creates a txt document summarizing the analysis results.
fp=open(text+'.txt', 'w')

fp.write('For ' + text + ' the number of genomes included is:'+'\n')
fp.write(str(len(p))+'\n')
fp.write('the core genome is:'+'\n')
fp.write(str(result)+'\n')
fp.write('the pan genome is:'+'\n')
fp.write(str(al)+'\n')
fp.write('absent ORFs are:' +'\n')
fp.write(str(notInGroup)+' corresponding to: '+ str(len(notInGroup))+'\n')
fp.write('the core genome size regression by adding a genome is:' +'\n')
fp.write(str(regressionCore)+'\n')
fp.write('the pan genome size progression by adding a genome is:' +'\n')
fp.write(str(progressionPan)+'\n')

fp.close()
```

```
For AfricaProphagesPanCore the number of genomes included is:
12
the core genome is:
set([1, 2, 3, 4, 5, 8, 10, 11, 13, 14, 15, 16, 26, 27, 28, 29, 30])
the pan genome is:
set([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 39, 43, 44, 45, 46, 47, 48, 54])
absent ORFs are:
[37, 40, 41, 42, 49, 50, 51, 52, 53, 55] corresponding to: 10
the core genome size regression by adding a genome is:
[30, 29, 26, 26, 26, 26, 26, 26, 19, 19, 19, 17]
the pan genome size progression by adding a genome is:
[30, 34, 37, 37, 40, 40, 43, 44, 44, 44, 45, 45]
```

```
from __future__ import division
import copy
from random import shuffle
import numpy
import matplotlib.pyplot as plt

##name of the document to be created
text='AllProphagesPanCoreAverage'

##ORF code for prophage genomes

UKEN31U = [1, 2, 3, 4, 31, 32, 7, 8, 9, 10, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
           15, 16, 17, 37, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
UKEN32U = [1, 2, 3, 4, 31, 32, 7, 8, 9, 10, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
           15, 16, 17, 37, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]
DeM53M = [1, 38, 2, 3, 4, 5, 6, 8, 9, 10, 11, 36, 12, 39, 13, 14, 15, 16, 17,\
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 29, 30]
Sw577G = [40, 1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 33, 34, 15, 16, 17, 18, 19,\
          20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 41]
SwA626G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 42, 33, 34, 35, 36, 12, 13,\
           14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtB89G = [1, 38, 2, 3, 4, 5, 6, 8, 8, 10, 11, 11, 36, 12, 13, 14, 15, 16, 17, \
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt1293U = [1, 38, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 11, 36, 12, 13, 14, 43, 44, \
           14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrANT170U = [1, 38, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 45, 46, 13, 14, 15, 16,\
             46, 45, 16, 17, 18, 47, 20, 21, 22, 23, 24, 25, 26, 27, 28, 28, 29,\
             30]
FrMEG235U = [1, 38, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 45, 46, 13, 14, 15, 16,\
             46, 45, 16, 17, 18, 47, 20, 21, 22, 23, 24, 25, 26, 27, 28, 28, 29,\
             30]
Pt5771G = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14, 15,\
           16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt5322G = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 18,\
           19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtB92G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 32, 31, 10, 11, 33, 34, 35, 36, 12,\
          13, 14, 15, 16, 17, 48, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4481G = [11, 10, 9, 8, 6, 5, 4, 3, 2, 38, 1, 11, 36, 12, 39, 13, 14, 15, 16, \
           17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrGC43G = [1, 2, 3, 4, 49, 5, 6, 8, 9, 31, 32, 10, 11, 33, 34, 35, 36, 12, 13, \
           14, 15, 16, 17, 18, 19, 20, 21, 32, 31, 22, 23, 24, 25, 26, 27, 28, 29, 30, 50]
FrG12G = [1, 2, 2, 3, 4, 49, 5, 6, 51, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14,\
          15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
FrB58M = [1, 2, 43, 44, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21,\
          23, 24, 25, 26, 27, 28, 29, 30]
Pt1918U = [1, 2, 3, 4, 49, 5, 6, 8, 9, 10, 11, 33, 34, 35, 36, 12, 13, 14, 15, \
           16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4497U = [52, 1, 2, 3, 4, 49, 5, 51, 8, 9, 10, 11, 33, 34, 35, 36, 12, 53, 13, \
           14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt4472G = [1, 2, 3, 4, 5, 7, 8, 8, 9, 10, 11, 11, 36, 12, 13, 14, 15, 16, 17,\
           18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
PtR21299U = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 24, 18,\
             26, 27, 28, 29, 30]
FrB41M = [1, 2, 3, 4, 49, 5, 6, 8, 9, 10, 11, 33, 34, 35, 36, 12, 53, 13, 14, \
          15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 55]
Pt1846U = [1, 38, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, \
           18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Pt22899G = [1, 38, 2, 3, 4, 5, 6, 7, 8, 46, 45, 9, 10, 11, 36, 12, 13, 14, 46, \
```

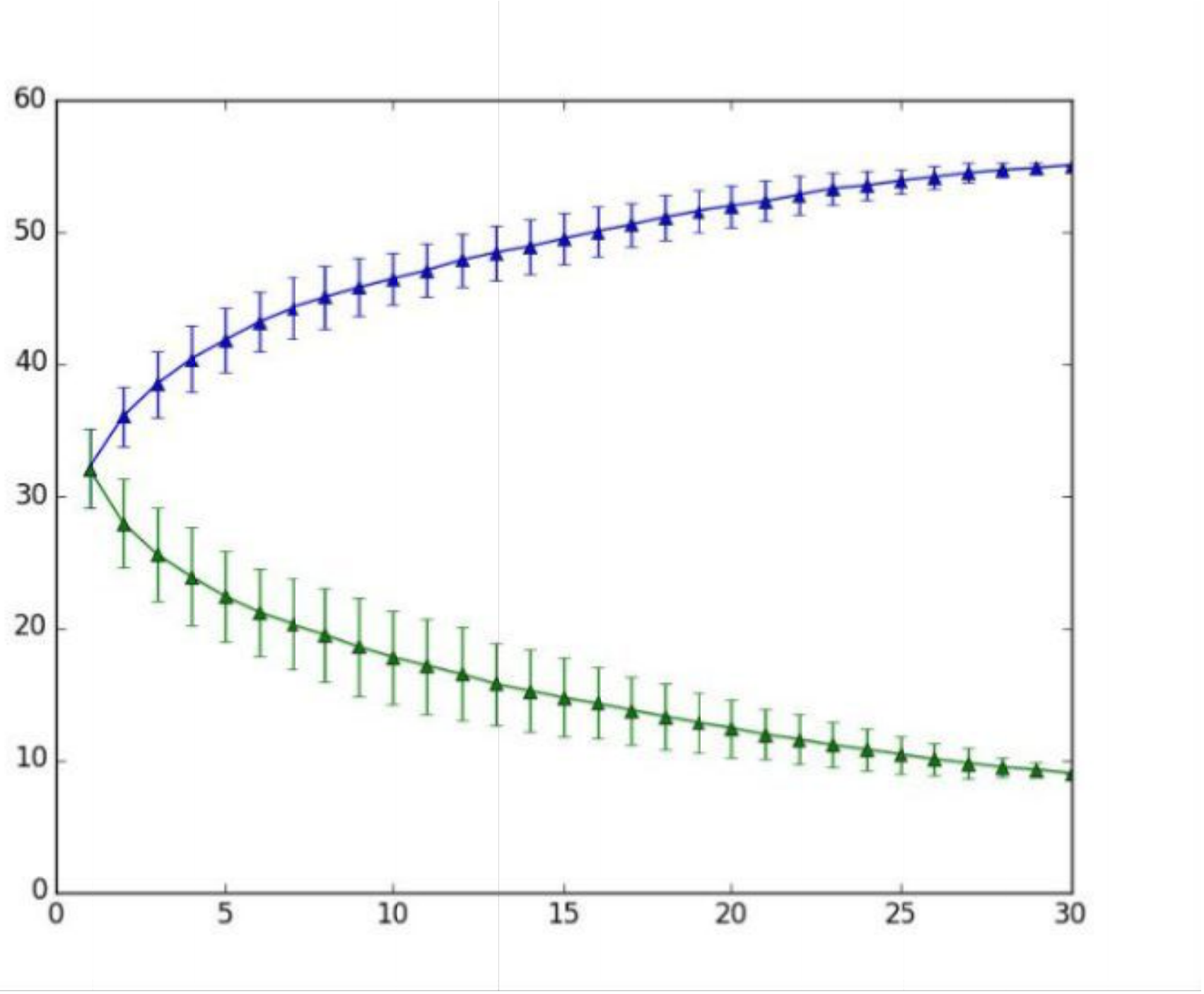
```
            45, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, \
            30, 54]
KHP30 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
         21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
KHP40 = [1, 2, 3, 4, 5, 6, 8, 9, 9, 10, 11, 12, 39, 13, 14, 15, 16, 17, 37, \
         18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 52]
India7 = [1, 38, 2, 3, 4, 5, 7, 8, 9, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, \
          18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
phiHP33 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 11, 34, 36, 12, 13, 14, 15, 16, \
           46, 45, 25, 26, 27, 28, 29, 30]
P1961P = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11, 36, 12, 39, 13, 14, 15, 16,\
          17, 18, 19, 20, 21, 22, 23, 24, 25, 25, 26, 27, 29, 30]
Cuz20 = [1, 38, 2, 3, 4, 5, 7, 8, 9, 10, 11, 36, 12, 13, 14, 15, 16, 17, 37, \
         18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]
Hac = [1, 10, 2, 18, 5, 5, 7, 8, 9, 10, 11, 33, 34, 36, 12, 13, 14, 15, 16, \
       17, 18, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30]


##ALL Phages
p=[UKEN31U,UKEN32U,DeM53M,Sw577G,SwA626G,PtB89G,Pt1293U, \
   FrANT170U,FrMEG235U,Pt5771G,Pt5322G,PtB92G,Pt4481G,\
   FrGC43G,FrG12G,FrB58M,Pt1918U,Pt4497U,Pt4472G,PtR21299U, \
   FrB41M,Pt1846U,Pt22899G,KHP30,KHP40,India7,phiHP33,P1961P, \
   Cuz20,Hac]


##p list has the phages to be considered in the analysis
##SWEurope phages
##p=[Pt1918U,Pt4497U,Pt4472G,FrB41M]

##NEurope phages
##p=[UKEN31U,UKEN32U,DeM53M,Sw577G,SwA626G,India7,Cuz20]

##EastAsia phages
##p=[KHP30,KHP40,P1961P,FrB58M, FrGC43G,FrG12G]

##Africa1 phages
##p=[PtB89G,Pt1293U,FrANT170U,FrMEG235U,Pt5771G,Pt5322G,PtB92G,Pt4481G,\
##   PtR21299U, Pt1846U,Pt22899G,phiHP33]

#Copy of p
#p2=copy.deepcopy(p)

#number of combinations to be tested
combination=100

##The sets module provides classes for constructing and manipulating unordered
##collections of unique elements. Common uses include membership testing,
##removing duplicates from a sequence, and computing standard math operations
##on sets such as intersection, union, difference, and symmetric difference.

coreDeterminations=[]
panDeterminations=[]

while combination !=0:
    shuffle(p)
```

```
##result is composed by the first element of the p list
    result = set(p[0])
    une = set ([])


##print result

##For the first element of p list makes a list with not repeated ORF
    uniqueValues=[]
    for i in p[0]:
        if i not in uniqueValues:
            uniqueValues.append(i)

##Finds common genes to ALL phages in p list and
##counts regression of core genome by adding new genome
    regressionCore=[]
    regressionCore.append(len(uniqueValues))
    for s in p[1:]:
        result.intersection_update(s)
        regressionCore.append(len(result))

    coreDeterminations.append(regressionCore)

##Finds all ORF - pan genome of phages and
##counts progression of core genome by adding new genome
    test=set(p[0])
    al=[]

    progressionPan=[]
    progressionPan.append(len(uniqueValues))

    for i in p[1:]:

        al=test.union(i)
        test=set(al)
        progressionPan.append(len(al))

    panDeterminations.append(progressionPan)

##print al
##print progressionPan
####checking step
##print len(progressionPan)

    combination=combination-1

##print coreDeterminations
##print regressionCore

##Calculates the mean of the core genome
meanCore=[sum(e)/len(e) for e in zip(*coreDeterminations)]

print meanCore

##Calculates the standard deviation of the core genome
sdCore=[numpy.std(e, ddof=1) for e in zip(*coreDeterminations)]

print sdCore
```

```
##print panDeterminations
##print progressionPan

##Calculates the mean of the pan genome
meanPan=[sum(e)/len(e) for e in zip(*panDeterminations)]

print meanPan

##Calculates the standard deviation of the core genome
sdPan=[numpy.std(e, ddof=1) for e in zip(*panDeterminations)]

print sdPan


####Calculates the number of ORF that are not in the p list
notInGroup=[]
first=1
AllORF=[]

while int(first)<56:
    AllORF.append(first)
    first =first + 1

##print AllORF

for i in AllORF:
    if i not in al:
        notInGroup.append(i)

print notInGroup

print 'PAN'
##Print the trend of the mean of the pan genome
for i in meanPan:
    print i

print 'take a break'

##Print the trend of the SD of the pan genome
for i in sdPan:
    print round(i, 2)

print 'CORE'
##Print the trend of the mean of the core genome
for i in meanCore:
    print i

print 'take a break 2'

##Print the trend of the SD of the core genome
for i in sdCore:
    print round(i, 2)


##Creates a txt document summarizing the analysis results.
fp=open(text+'.txt', 'w')
```

```
fp.write('For ' + text + ' the number of genomes included is:'+'\n')
fp.write(str(len(p))+'\n')

fp.write('absent ORFs are:' +'\n')
fp.write(str(notInGroup)+' corresponding to: '+ str(len(notInGroup))+'\n')
fp.write('the mean core genome size regression by adding a genome is:' +'\n')
fp.write(str(meanCore)+'\n')
fp.write('the mean standard deviation genome size regression by adding a genome is:' +'\n')
fp.write(str(sdCore)+'\n')
fp.write('the mean pan genome size progression by adding a genome is:' +'\n')
fp.write(str(meanPan)+'\n')
fp.write('the mean standard deviation genome size progression by adding a genome is:' +'\n')
fp.write(str(sdPan)+'\n')
fp.close()


##List from one to the number of included genomes
counting=1
phages=[]
while int(counting)<= len(p):
    phages.append(counting)
    counting=counting +1

##Code for the graphic of the pan and core genome including trend and SD
x = numpy.array(phages)
y = meanPan
e = numpy.array(sdPan)

y1=meanCore
e1=numpy.array(sdCore)

#plt.errorbar(x, y, e, linestyle='None', marker='^')
plt.errorbar(x, y, e, marker='^')
plt.errorbar(x, y1, e1, marker='^')


plt.show()
```

For AllProphagesPanCoreAverage the number of genomes included is:
30
absent ORFs are:
[] corresponding to: 0
the mean core genome size regression by adding a genome is:
[32.07, 27.92, 25.53, 23.86, 22.37, 21.16, 20.27, 19.46, 18.54, 17.77, 17.13, 16.49, 15.74, 15.24, 14.7, 14.28, 13.77, 13.29, 12.81, 12.43, 11.92, 11.55, 11.13, 10.78, 10.41, 10.06, 9.75, 9.44, 9.26, 9.0]
the mean standard deviation genome size regression by adding a genome is:
[2.951647374301718, 3.3414446763716628, 3.5489506202609391, 3.7064508330447032, 3.4572570431812353, 3.3324847404683706, 3.4285049897362647, 3.5460035834924666, 3.7454888691180046, 3.5358338779022427, 3.5975721330537982, 3.5547521188210331, 3.1610636235803975, 3.0554471973440442, 2.9865016188162228, 2.682076686736484, 2.5696558744569891, 2.4630399218375767, 2.2948999229956972, 2.20309699369691, 1.8839031050165265, 1.9141947168826408, 1.6796764960292594, 1.5607560975244006, 1.4077347661108508, 1.2457604873807775, 1.1315047637120359, 0.75638694602776702, 0.54346438278284182, 0.0]
the mean pan genome size progression by adding a genome is:
[32.07, 35.99, 38.46, 40.34, 41.78, 43.15, 44.21, 45.01, 45.76, 46.41, 47.05, 47.82, 48.38, 48.85, 49.43, 50.0, 50.49, 51.07, 51.54, 51.92, 52.28, 52.75, 53.24, 53.48, 53.82, 54.11, 54.42, 54.64, 54.79, 55.0]
the mean standard deviation genome size progression by adding a genome is:
[2.951647374301718, 2.2940194533885827, 2.4961384317996216, 2.4872198588509669, 2.4228708241057819, 2.2669117514559072, 2.3582864863105906, 2.4266824841309336, 2.197886955852614, 1.9801795653705525, 2.0269646910061998, 1.9816328335160092, 2.0878677925558216, 2.0319418984773847, 1.9501618684435693, 1.8802535827258875, 1.6544705277739846, 1.6832808272639306, 1.5467136878954788, 1.55492050529208, 1.5247453836127804, 1.4168894655461863, 1.1730991501836561, 1.0586822378825862, 0.92529001732517135, 0.88642967896927738, 0.74100668626599375, 0.59492805778463154, 0.43333333333333329, 0.0]


60
50
40
30
20
10
0
0
5
10
15
20
25
30