Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi) to transfer and detect genes of interest in *Escherichia coli* and unravel the recombination patterns

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Arsh Shrikant Chavan

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

Date: April, 2023

Under the guidance of

Supervisor: Dr Olivier Tenaillon

Directeur de Recherche Inserm

Head of team: Quantitative Evolutionary Microbiology

Associate Professor at Ecole Polytechnique Biology Department

From May 2022 to March 2023

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

Certificate

This is to certify that this dissertation entitled "Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi) to transfer and detect genes of interest in *Escherichia coli* and unravel the recombination patterns" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by "Arsh Shrikant Chavan at IAME, Inserm and Institut Cochin" under the supervision of "Dr Olivier Tenaillon, Directeur de Recherche Inserm, Head of team : Quantitative Evolutionary Microbiology, Associate Professor at Ecole Polytechnique Biology Department during the academic year 2022-2023

Dr Olivier Tenaillon

Committee:

Supervisor: Dr Olivier Tenaillon 

TAC Expert: Dr Nishad Matange

This thesis is dedicated to my mother Dr Anjali Shrikant Chavan

Declaration

I hereby declare that the matter embodied in the report entitled "Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi) to transfer and detect genes of interest in *Escherichia coli* and unravel the recombination patterns" are the results of the work carried out by me at the Department of Microbiology, IAME, Inserm and Institut Cochin, under the supervision of Dr Olivier Tenaillon and the same has not been submitted elsewhere for any other degree.

Name: Arsh Shrikant Chavan

Date: 10 April 2023

Table of contents

List of Figures

List of tables

Abstract

Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi) is a technique developed to enable the exchange of genetic material between bacterial strains, in a manner that simulates the process occurring in the wild. CoMBacGeMi aims to elucidate the molecular basis of functional diversity observed in medically significant bacterial species, such as *Escherichia coli*. The method employs a library of strains, each harboring a conjugative plasmid integrated at a random position in the genome.

In this study, we employed CoMBacGeMi to investigate the effect of a mutant *matP* recipient on the bias in recombination frequency observed at the terminus. Analysis of the pools revealed that MatP does not significantly regulate the recombination pattern.

To go beyond a single locus and to investigate the recombination pattern across the entire genome, we utilized the Hfr library and the recipient *tse2* library. Our results revealed that the recombination frequencies across the genome were non-uniform, with regions such as the Ter macrodomain exhibiting low recombination frequencies. Additionally, we observed that the conjugation efficiencies of recipient clones were related to the position of the loci of selection in the genome.

Lastly, we tested the ability of CoMBacGeMi to identify specific targets of selection. Analysis of the pools enabled us to identify, with precision, the three distinct insertion positions of *tse2* in the three tested clones. Our results demonstrate that CoMBacGeMi is a simple, precise, and accurate method with the potential to identify key genetic loci and provide insight into the complexities of genetic traits in diverse bacterial populations.

Acknowledgements

I would like to express my sincere gratitude to all the people and organizations who have contributed to the successful completion of my thesis:

First and foremost, I would like to thank my thesis supervisor Dr Olivier Tenaillon for his guidance, encouragement, and support throughout the research process.

I am also deeply grateful to Dr Nishad Matange, my thesis committee expert, for his insightful comments and feedback, which helped me improve my work.

I am immensely grateful to Juliette Bellengier, who has helped me throughout my thesis and significantly contributed to this project. I would also like to thank Dr Ivan Matic, Marie Florence, Mélanie Magnan, Alaksh Choudhury, José Ibarra, Zoya Dixit, Flavia Hasenauer, Maureen Micaletto, Sébastien Fleurier, Arnauld Gutierrez, Chantal Lotton, Lucile Vigué and all the other lab members for teaching and helping me during my thesis and making my stay in Paris a memorable experience.

I would like to acknowledge IISER Pune, IAME INSERM and Institut Cochin for providing me with the resources, facilities, and opportunities to pursue my academic goals.

I would like to thank INSERM Dr Paris IDF Centre Nord and DST INSPIRE fellowship for the funding.

I am grateful to my friends: Vignesh, Nitin, Ritika, Sarang, Monali, Rishabh, Vedant, Siddhi, Tanaya, and Durvesh who have always been there for me during my good and bad days.

My sincere thanks to my mother, brother and loved ones, who stood by me through the highs and lows of my academic journey. Their unwavering support and encouragement kept me motivated and inspired throughout.

Contributions

Contributor name	Contributor role
Olivier Tenaillon, Alaksh Choudhary, Arsh Shrikant Chavan	Conceptualization Ideas
Thibault Corneloup, Alaksh Choudhury, Arsh Shrikant Chavan	Methodology
-	Software
Arsh Shrikant Chavan, Mélanie Magnan	Sequencing
Olivier Tenaillon	Validation
Arsh Shrikant Chavan, Juliette Bellengier	Formal analysis
Arsh Shrikant Chavan	Investigation
Olivier Tenaillon, Ivan Matic	Resources
Arsh Shrikant Chavan	Data Curation
Arsh Shrikant Chavan	Writing - original draft preparation
Olivier Tenaillon, Arsh Shrikant Chavan	Writing - review and editing
Arsh Shrikant Chavan, Juliette Bellengier	Visualization
Olivier Tenaillon	Supervision
Olivier Tenaillon	Project administration
Olivier Tenaillon	Funding acquisition

1. Introduction

1.1. Escherichia coli

Escherichia coli (*E. coli*) belongs to the Enterobacteriaceae family of bacteria. Enterobacteriaceae group includes 100 species and 12 genera, including *Escherichia*, *Yersinia*, *Shigella*, *Salmonella* and *Citrobacter*. The genus *Escherichia* is composed of three species *albertii*, *fergusonni* and *coli*, as well as five clades I to V. These clades were only discovered recently (Walk *et al.*, 2009) as they are indistinguishable from *E. coli* using classical metabolic-based species identification methods. However, a significant divergence from *E. coli* can be revealed using genome analysis.

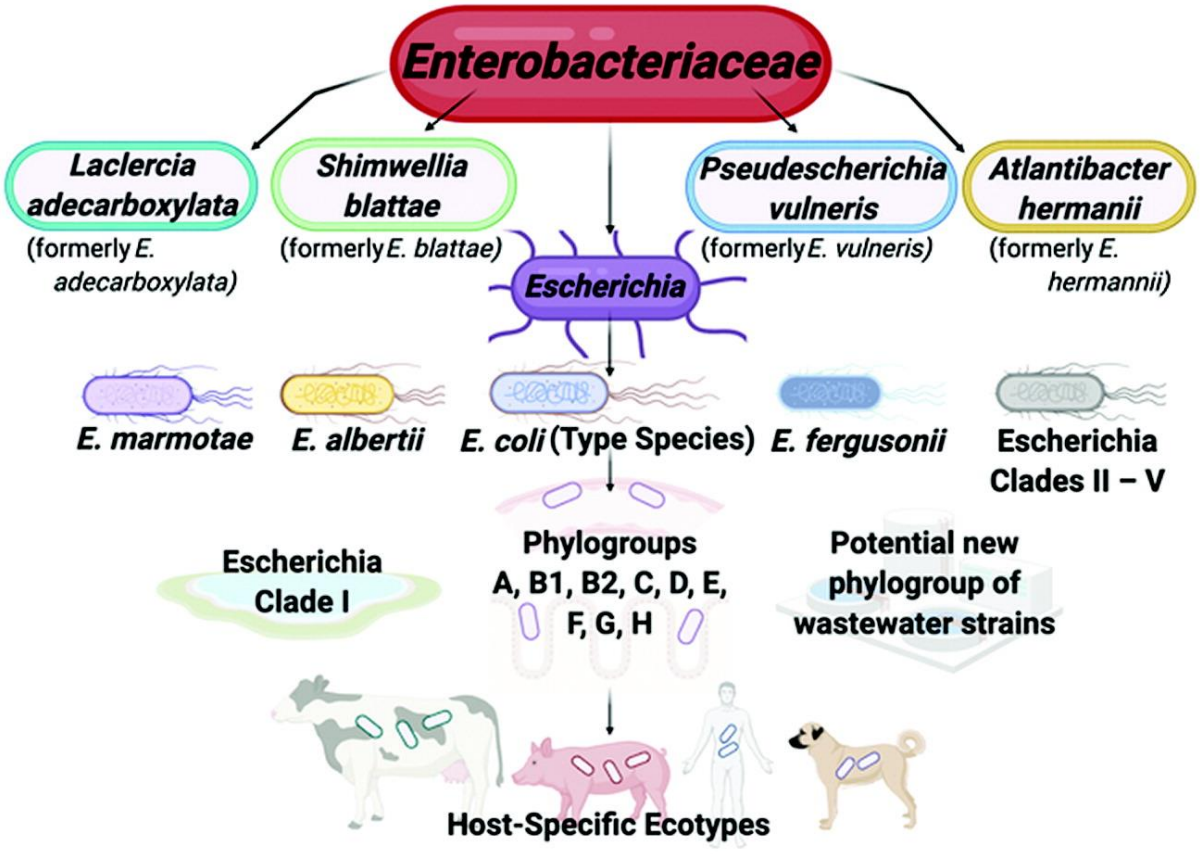


Figure 1.1. Taxonomic structure of the genus *Escherichia* and the different species like *E. coli* (Yu *et al.*, 2021).

E. coli are rod-shaped gram-negative bacteria, i.e., they possess an outer membrane apart from the peptidoglycan layer. *E. coli* have a facultative aerobic metabolism, i.e., they can grow with or without the presence of oxygen. Their primary habitat is the gastrointestinal tract of warm-blooded animals (Jang *et al.*, 2017), but they are also found in secondary habitats such as skin, urinary tract, fresh water and soil (Tenaillon

et al., 2010). *E. coli* are easy to culture and have a fast generation time varying from 20 to 40 minutes with an average length between 1.6 and 3.1 µm (Volkmer and Heinemann, 2011).

1.2. *E. coli:* friend or foe

E. coli primarily have a commensal relationship with the host where one of the two organisms gains from the other's presence while the other is not particularly injured or benefitted. The host provides *E. coli* with a consistent supply of nutrients, a stable environment, some degree of protection from stressors, as well as transport and dissemination (Conway *et al.*, 2004). But *E. coli* microbiota can also benefit the host by preventing pathogen colonization in the gut by producing bacteriocins and other compounds. They have a prevalence of about 90% in humans, 56% in non-domesticated mammals, 23% in birds and 10 % in reptiles (Tenaillon *et al.*, 2010).

However, *E. coli* can also be pathogenic and cause a variety of diseases. They can be obligate or opportunistic pathogens. Some intestinal pathogenic strains like entero-haemorrhagic *E. coli* (EHEC) are always pathogenic in humans but can be found as commensals in other animals like cattle. Pathogenic *E. coli* have virulence factors like toxins and adhesins. These virulence factors are usually encoded by a cluster of genes known as pathogenicity islands (PAIs). PAIs are large mobile genetic elements, usually 10 kb – 200 kb long, those harbour virulence genes that encode for toxins, adhesins and secretion systems (Hallstrom and McCormick, 2014). PAIs are often found adjacent to tRNA genes and are generally flanked with direct repeats (Karaolis, 2001) and are thought to be acquired by means of horizontal gene transfer. They can be found on the chromosome as well as on plasmids (Karaolis, 2001).

Pathogenic *E. coli* are broadly divided into intestinal pathogens and extraintestinal pathogens. Intestinal pathogens can be further divided into six groups: enteropathogenic *E. coli* (EPEC), entero-haemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), diffusely adherent *E. coli* (DAEC), enteroaggregative *E. coli* (EAEC), entero-invasive *E. coli* (EIEC) and enteroaggregative *E. coli* (EAEC) (Kaper *et al.*, 2004). There are several extraintestinal pathogenic *E. coli* like uro-pathogenic *E. coli* (UPEC) that cause urinary tract infections (UTIs), meningitis-

associated *E. coli* (MNEC) that cause meningitis and avian pathogenic *E. coli* (APEC) that causes respiratory infections in chickens and turkeys (Kaper *et al.*, 2004).

Each year, there are around 1.7 billion episodes of diarrheal illness worldwide. Over 760,000 young children under the age of five die each year from diarrheal illnesses (Chowdhury *et al.*, 2015). Thankfully, environmental cleanliness and proper hygiene can prevent diarrhea brought on by pathogenic *E. coli*. Antibiotics and oral rehydration have been generally effective in treating diarrheal diseases (Yang *et al.*, 2017). Diarrhea associated with *E. coli* is now rare in developed nations, but UTIs and nosocomial infections caused by *E. coli* are on the rise.

Through the extraintestinal pathologies, *E. coli* has also become a public health concern. In 2019, it was involved in close to a million death worldwide, with 800,000 associated to antibiotic resistance and about 200,000 due to antibiotic resistance (Murray *et al.*, 2022). This makes *E. coli* the most challenging pathogen in term of antibiotic resistance.

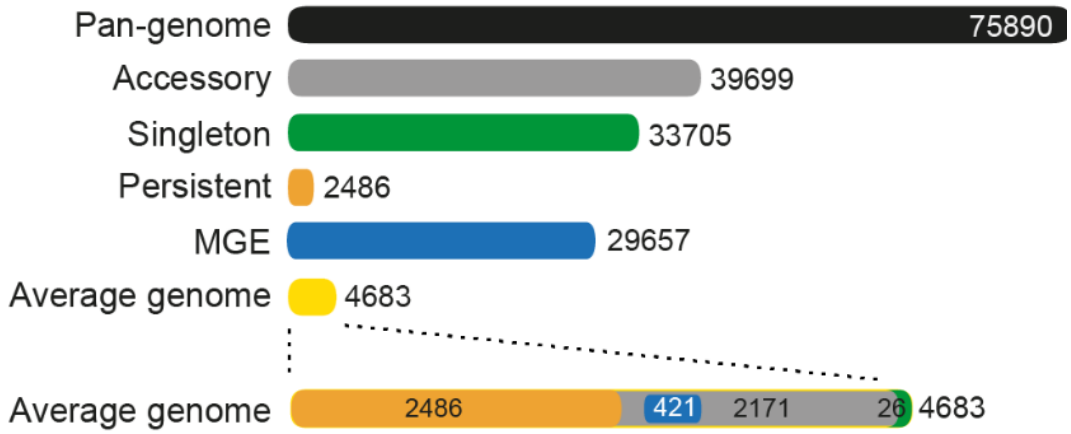
1.3. Diversity within *E. coli*

The circular chromosome that makes up the majority of the genetic material in *E. coli* is between 4 and 5.5 million base pairs (bp) long and can accommodate between 3900 and 5800 genes (Ochman and Bergthorsson, 1998; Touchon *et al.*, 2009). The average number of genes per genome is estimated to be 4700 genes. This is roughly one gene per 1000 bp DNA. In comparison, humans have a genome size of roughly 3.2 billion base pairs (>600 times the *E. coli* genome), but have only 4 or 5 times more coding genes than *E. coli*.

This variation in gene content among the several sequenced strains demonstrates that genomes in natural settings are always under an intense flow of gene acquisition and loss. Based on this, the genome can be grouped into the core genome and the accessory genome. The core genome is the set of genes that is common in all the sequenced strains whereas the set of genes that are shared by only one or some of the sequenced strains is referred to as the accessory genome. This can vary depending on the number of sequenced strains and the selection criteria used; genes present in 95%, 99% or 100% of the sequenced strains. The list of all the genes identified in all the sequenced strains is called the pangenome. From the early studies

on the *E. coli* genome it was found that the core genome is composed of around 2000 genes (Touchon *et al.*, 2009). As a result, the core genome was less than half the size of the average *E. coli* genome. As the number of sequenced strains increased from 20 strains to 1300 strains, it was observed that the number of genomes in the core genome decreased and the number of genes in the pangenome increased. The pangenome increased with the number of genomes studied from 15,000 genes to more than 75,000 genes (Touchon *et al.*, 2009, 2020; Yu *et al.*, 2021). In this now updated list of genes in the core and accessory genome, a given *E. coli* genome would only have about 20% of its genes conserved as part of the core genome, and the remaining 80% of genes would be from the variable pangenome across all other *E. coli*(Yu *et al.*, 2021).

Figure 1.2. Composition of the *E. coli* pangenome for 1294 genomes. It is composed



of 75890 gene families, of which 33705 are present in a single genome. The core genome consists of 2486 genes, present in at least 99% of genomes. 39% of the pan-genome is composed of mobile genetic elements. The average genome size is 4683 genes (Touchon *et al.*, 2020).

The acquisition of new genes can happen by horizontal transfer of genetic material between bacteria of the same or different species. This genetic material can either lead to the acquisition of new accessory genes or lead to the recombination of conserved parts of the genome. The newly acquired accessory genes can sometimes confer a selective advantage for adaptation to an ecological niche, the ability to colonize a specific host or via antibiotic resistance (Yang *et al.*, 2019). Homologous

recombination produces new combinations of alleles along the chromosomes, thus affecting both the accessory genome as well as the core genome.

Based on the core genome, the stains can be categorized into phylogenetic groups. *E. coli* strains can be clustered into eight phylogroups (Clermont *et al.*, 2013), including four main ones, A, B1, B2 and D and two secondary phylogroups, E, F, as well as some other marginal groups (Selander and Levin, 1980; Clermont *et al.*, 2013). They are divided into two clusters: (B2, G, F and D) and (A, B1, C and E) (Touchon *et al.*, 2020). The phylogroup C is relatively close to B1, while phylogroup F is closer to B2 and D. Moreover, there are differences in genome size among phylogroups. The genomes of strains from the A and B1 phylogroups are smaller than those from the E phylogroup (Touchon *et al.*, 2020).

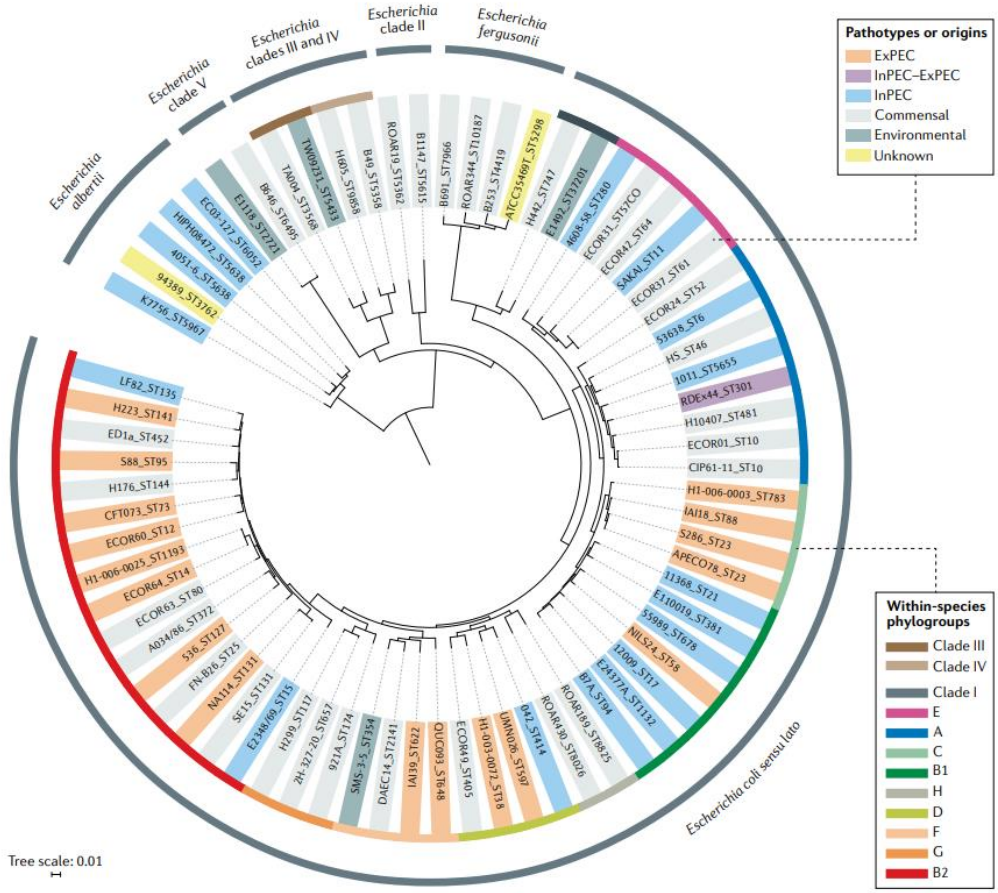


Figure 1.3. Phylogenetic history of *E. coli* based on the core genome of 72 representative strains. (Denamur *et al.*, 2021). Strain pathogenicity categories are listed by group phylogenetics.

E. coli appears to have a clonal structure, with clones displaying significant genetic variation at various scales. Hence, *E. coli* has various ecotypes and phylogroups tailored to distinct ecological niches or hosts. This diversification and clonal expansion certainly rely on the incorporation of various advantageous mutations. Not all clones have the same evolutionary success. Thus, some clones associated with particular phylogroups are more likely to be pathogenic or antibiotic-resistant. For instance the epidemic clone ST131 has spread worldwide in the early 2000's and is both multi-resistant to antibiotics and virulent (Kallonen *et al.*, 2017).

1.4. Genetic exchange and recombination

The majority of bacteria have an obligate asexual mode of reproduction and reproduce by binary fission, duplicating their DNA clonally. This clonality of bacterial lines is, however, affected by genetic exchanges, which in a generic way qualify as sex. Sex represents the acquisition of material not from the parent cell in the field of microbial genetics. Foreign DNA fragments can be introduced by cell-to-cell contact, by an infectious agent or acquired directly in the environment (Redfield, 2001; Bobay *et al.*, 2015). The sharing of genetic material between organisms that do not have a parent-offspring relationship is known as horizontal gene transfer (HGT) (Soucy *et al.*, 2015). HGT is a well-known adaptation process in bacteria and archaea (Garcia-Vallve *et al.*, 2000). HGT is frequently linked to microbial pathogenicity and antibiotic resistance and the transfer of gene clusters encoding biodegradative pathways (De la Cruz and Davies, 2000)

The exchange of genetic material between bacterial populations occurs according to three main horizontal transfer mechanisms; transformation, conjugation and transduction. Transformation is a process by which bacteria take up exogenous genetic material (naked DNA) directly from the environment. Natural transformation is not common in *E. coli* and occurs only under extreme conditions (Mandel and Higa, 1970). Conjugation is a mode of transfer of genetic material through cell contact with the help of a tube-like appendage called the sex pilus. Transduction is a method by which a bacteriophage transfers genetic material from one bacterium to another. This happens after a bacteriophage makes an encapsulation error in which it chooses

bacterial DNA rather than phage DNA, causing a genetic material transfer during the subsequent infection.

The acquired DNA fragment so obtained may follow several paths. It can be degraded, integrated into the chromosome or can stay as a vertically inheritable extrachromosomal element. If the acquired genetic material has some homology within the genome, then it can integrate into the chromosome through homologous recombination.

1.5. Mechanisms of recombination

Homologous recombination and site-specific recombination are the two basic categories into which genetic recombination may be classified. These pathways differ from one another in terms of their mechanism, recombination proteins, and substrate DNAs (Craig, 1988). Extensive sequence homology-based exchange between DNA segments is mediated by homologous recombination. Although this exchange is possible at any location between the homologous regions, the frequency of recombination can be influenced by specific DNA sequences. The rate of recombination decreases as the homology decreases (Rubnitz and Subramani, 1984). MEPS (minimal efficient processing segment), or the minimum sequence identity length that is necessary for pairing stability, usually lies between 20 to 151 bases (Vulić *et al.*, 1997; Fraser *et al.*, 2007).

On the other hand, site-specific recombination occurs between DNA strands that lack extensive homology by cleaving DNA and rejoining at specific positions without degradation or synthesis. Although it is critical to have at least very short regions of homology (Craig, 1988). Integration of DNA from phages and mobile genetic elements like transposons or conjugative plasmids takes place by this mode of recombination. No homology sequence is necessary for recombination by transposons. Several virulence factors, such as pathogenicity islands, are acquired by site-specific recombination. Hence site-specific recombination also significantly contributes to the evolutionary history of *E. coli* and its diversification. When the two sites are in the same DNA molecule, site-specific recombination can either result in excision (deletion) or inversion. But it leads to integration when recombination between sites on two district DNA molecules takes place, and at least one of them is circular. There is no loss or

synthesis of DNA during integration, and hence is a conservative mechanism (Proteau *et al.*, 1986).

1.5.1. Mechanism of homologous recombination

Homologous recombination mainly takes place with the help of the RecA protein. The *recA* gene was first identified by screening for the inability of mutagenized F^- colonies to produce recombinants post-conjugation with an (High Frequency of Recombination) Hfr strain (Clark and Margulies, 1965). The RecA protein catalyzes the matching of homologous DNA sequences in different forms: between single-stranded sequences, double-stranded (dsDNA) sequences and between double-stranded and single-stranded (ssDNA) sequences. When there is recombination between a dsDNA and an ssDNA, only one strand of the dsDNA takes part in the pairing. When two dsDNA are involved, a cross structure is formed (Shibata *et al.*, 1979; Cassuto *et al.*, 1980; Cox and Lehman, 1981). This is also called the Holliday junction.

The foreign DNA that enters the cells cannot be directly used by RecA. This DNA first needs to be prepared to be usable by RecA. The process of preparing the foreign DNA for RecA is carried out by the RecBCD protein complex. This complex has single and double-stranded helicase and exonuclease activity, as well as single-stranded endonuclease activity (Taylor and Smith, 1980). The Chi site is a specific 8-base sequence needed for the attenuation of the activity of RecBCD exonuclease and the RecABCD-mediated homologous recombination. This sequence is either 5′-GCTGGTGG-3′, or its complement (Dower and Stahl, 1981; Dixon and Kowalczykowski, 1993). The 3' end of the DNA required for RecA activity is then created after the imported DNA has been degraded and unrolled to the closest chi sequence by the RecBCD complex. In the absence of the RecBCD complex, this 3' end can be generated by RecE (double-stranded exonuclease), RecQ (single-stranded helicase), and RecJ (exonuclease in combination with RecQ) in different ways by degrading or by unrolling the DNA.

The homologous recombination by RecA takes place in three steps. The first step is homologous pairing. In order to achieve this, RecA must polymerize 5' to 3' along the DNA's single-stranded region. To pair with the homologous DNA fragment, RecA will form a complex with the ssDNA. The strand exchange reaction proceeds from a 3' end

on the pairing area (Register and Griffith, 1985). The Single-Strand binding protein (SSB) then slowly attaches to the ssDNA (Hobbs *et al.*, 2007). This complex subsequently allows faster polymerization. The search for the homologous sequence within the supercoiled chromosomal DNA will be guided by the ssDNA microfilament that is generated. Depending on the available space, the pairing is either carried out in a double helix or not. The Holliday junction is then created by the pairing. The second step is the extension of the heteroduplex. The paired complex created for both DNA segments by RecA is extended by strand migration. This process can cross heterologous DNA regions and is unidirectional from 3' to 5'. The final step involves the termination of the Holliday junction. The Holliday junction is cleaved by the RuvC protein. Depending on the DNA strand cleaved there can be two types of recombinant DNA.

1.5.2. Mechanism of site-specific recombination

Recombinases from one of two major families of related proteins, the serine recombinases and the tyrosine recombinases, catalyze conservative site-specific recombination (Grindley *et al.*, 2006). Their name corresponds to the name of the amino acid residue (serine or tyrosine) involved in the reaction. The recombinase alone or auxiliary proteins are sufficient to catalyze site-specific recombination. They can either be encoded by the host cell or the mobile element.

Even though these two families of recombinases have different mechanisms, we can broadly divide simple site-specific recombination into four steps. The first step involves the binding of the recombinase dimer with each of the recombination sites. The length of the dimer binding sites can range from 30bp, for simple sites, to more than 200bp for complex binding sites. Next, the two sites that are bound by the recombinase form a synaptic complex or synapse. This synaptic complex contains the recombinase tetramer (Craig, 1988). In the third step, the cleavage and the rejoining of the DNA strands are catalyzed by the recombinase. This cleavage and rejoining of the DNA take place at the core site or the crossover site, i.e., the area which binds the recombinase dimer. A DNA strand is cleaved by each of the recombinase subunits. They attack the DNA phosphodiester bond with the hydroxyl group of either the conserved serine or tyrosine side chain. This leads to the formation of a covalent

phosphoester bond between the DNA and the recombinase subunit (Grindley *et al.*, 2006). The breaks are usually 2bp to 8bp away from each other, and the breaks are staggered. During strand re-ligation, the covalent phosphoester bond between the DNA and the recombinase subunit is broken, and the DNA phosphodiester bond is restored. The final step involves the breakdown of the synaptic complex and the release of the recombinant products.

1.6. Barriers to genetic exchange

The *E. coli* genome has a high degree of plasticity, which leads to a wide variety of adaptation pathways. According to the recombination rate observed, a nucleotide is 100 times more likely to be engaged in genetic transfer than to undergo mutation (Touchon *et al.*, 2009). Despite its dynamic nature, the genome maintains its structure, and only a small number of genome rearrangements have been observed. In fact, it appears that the majority of horizontal gene transfers take place in certain integration hotspots (Oliveira *et al.*, 2017). This means that transfers are not homogeneous within the species, and there are certain barriers to genetic transfers.

It is necessary for different clones to be present in the same niche for genetic exchange to take place (Feil and Spratt, 2001). If there is an absence of contact between the different clones, then it is not possible to exchange genetic material. Even if there is no physical separation, it is also important that the donor and the recipient are compatible with each other. The presence of certain surface proteins can prevent the ability of a strain to receive genetic material. Once the genetic material has entered the cell, the transfer can still be prevented by restriction-modification systems (Vasu and Nagaraja, 2013). They recognize the foreign DNA and is cut into pieces by the nuclease. To prevent this, some plasmids have reduced the number of restriction sites and developed systems that limit the action of these restriction systems (Fomenkov *et al.*, 2020; Shaw *et al.*, 2022). The next control occurs at the stage of recombination. The degree of homology between the transferred DNA and the recipient genome modulates the amount of resistance faced. The more divergent the incoming DNA, the more likely it is that the mismatch repair enzymes intervene before recombination takes place and block it (Vulić *et al.*, 1997). It is also possible that the limitation of gene flow occurs after recombination has taken place through the action of selection.

1.7. Selection on recombinants

Within a niche, a beneficial mutation can arise spontaneously and confer a fitness advantage; if the selection is more rapid as compared to recombination, the genetic background in which the mutation appeared will outcompete other strains and become dominant in a population and eliminate genetic diversity (Cohan, 2002; Wiedenbeck and Cohan, 2011). *E. coli* can be found in various habitats like the gut, urinary tract, food, water and soil. The selective pressures under each condition and the plasticity of the *E. coli* genome suggest that some sets of genes or genomic islands would be favoured or selected for under certain environments and hence would only be found in a subset of strains (Tenaillon *et al.*, 2010) called ecotype. This also suggests that alleles beneficial in one niche may not be transmitted to strains adapted to a different niche. Consequently, recombination between strains from different ecotypes may lead to recombinants genotypes that are unfit in both niches defining the ecotypes.

1.7.1. The Hight pathogenicity island: an example of species-wide selection

The Hight Pathogenicity Island (HPI) is a genomic island. It is an iron capture system that gives an adaptive advantage in low-iron environments and increases colonization (Carniel, 2001). It is necessary for the mouse-virulence phenotype in *Yersinia pestis* (Schubert *et al.*, 2004). As opposed to most genomic islands, the HPI is a functional island that is widely disseminated among members of the family of Enterobacteriaceae (Karch *et al.*, 1999) and is spreading in *E. coli* species (Schubert *et al.*, 2004). It is indispensable for the pathogenicity of certain pathotypes of *E. coli* (Schubert *et al.*, 2004). It has long been established that HPI spreads within the species and that recombination is the mode of dissemination (Schubert *et al.*, 2009). As a result, there is a loss of variety in the areas surrounding the integration site, leading to a local phylogeny incongruent with that of the chromosome. The HPI is, therefore, the illustration that it is possible for a gene/allele to propagate at the scale of the species. Moreover, its distribution in the phylogenetic tree of the species and its link with virulence makes it one of the markers associated with virulence in studies from Genome-Wide-Association-Studies (Galardini *et al.*, 2020). As such, it is the perfect tool to study the fate of an HGT in *E. coli* species.

1.8. Mixing genomes in the lab

In nature, the transfer of genetic material in *E coli* occurs mainly by transduction and conjugation (Redfield, 2001). Comparative genomics has revealed recombinant fragments of some tens of thousands of base pairs (Dixit *et al.*, 2015; Sakoparnig *et al.*, 2021). This corresponds to the fragment size observed in laboratory transductions suggesting a dominant role of transductions as conjugation is supposed to transfer fragments of several hundred Kb (McKane and Milkman, 1995). The application of transduction has been primarily restricted to the transfer of selected markers between genetic backgrounds due to its low efficiency. In comparison, conjugation is far more efficient and has been frequently used for building genetic maps (Taylor and Thoman, 1964). However, genome sequencing led to the disuse of conjugation. But it has been revived over the last decade for high-throughput marker transfer (Typas *et al.*, 2008) or for assembling highly modified genome fragments (Ma *et al.*, 2014).

In the continuity of these approaches, a method to blend genomes was developed in our laboratory by Thibault Corneloup, with the objective of reproducing in the laboratory, horizontal transfers as they occur in nature using Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi). For this, a library of Hfr strains with different insertion sites was created. This would lead to the transfer of an unbiased sample from the donor. Using this approach, Thibault was able to produce a panel of recombinants incorporating a fraction of the donor genome to restore a metabolic capacity in the recipient strains. By the analysis of the pool of recombinants from these conjugations an asymmetry of the introgression signal was observed with a lack of recombination at the terminus.

The objectives of this thesis are threefold. Firstly, the aim is to test a mutant *matP* in order to investigate its effect on the bias observed at the terminus. Secondly, the objective is to test recombination efficiencies and to observe the recombination pattern across the whole genome using the K12 Hfr library and a recipient library with *tse2* counter selection marker. Lastly, to test the ability of the method (Conjugation Mediated Bacterial Genome Mixing) to specifically identify a target of selection.

2. Material & Methods

2.1. Datasets

Sixty thousand *E. coli* genomes were used in this study. These genomes were obtained from EnteroBase (Zhou *et al.*, 2020). The HPI sequence (12 genes) used in this study is from *E. coli* EC958 (Forde *et al.*, 2014). These 60,000 genomes of *E. coli* have roughly 5,000 genes each, a total number of ~300 million gene sequences. Lucile Vigué clustered these sequences based on similarity, where each cluster is composed of gene sequences sharing at least 90% identity and 80% length with the consensus sequence of the cluster to retrieve homologous genes. This gave rise to 440,000 clusters, each corresponding to a specific *E. coli* gene. Some clusters might also correspond to pseudogenes. MMseqs2 Version: 13.45111 (Steinegger and Söding, 2017), SQLite version: 3.38.5 2022-05-06 15:25:27 and Python 3.8.10 were used to check the prevalence of HPI and to understand the dynamics of HGT from distant species.

2.2. Strains

The conjugations were performed using the K12 BW25113 Hfr library as the donor, Rel606 *tse2* library and Rel606 Δ*matP* Δ*galK* strain as the recipients**.** The *tse* gene encodes for a toxin under the regulation of a rhamnose-inducible promoter and here it is used as a counter-selection marker. It is to be noted that when tse2 is integrated at different positions of the chromosome there is some escape observed in the *tse2* recipient library i.e., a *tse* inserted at certain positions in the chromosome is not able to kill the cells as efficiently as when integrated in other parts of the chromosome and hence some cells are able to survive in the presence of rhamnose. Therefore, the *tse2* library was scanned for efficiency (low escape).

2.3. Growth media

All growths were performed using LB media or M9 minimal media. LB was purchased from Sigma Aldrich with 5 g/litre yeast extract, 10 g/litre tryptone, and 10 g/litre NaCl; 15 g/litre agar was added as needed. In each 100 mL M9 media, we added M9 salts

(prepared using premixed powder from Sigma Aldrich), 0.4% Glucose or Galactose, 2 µM $MgS0_4$, 0.1 µM $CaCl_2$, and Thiamine 0.2 ug/ml. For plates, 1.6% select agar was added. Appropriate antibiotics were added to the medium as follows kanamycin (50ug/ml), spectinomycin (50ug/ml) and zeocin (50ug/ml).

2.4. Conjugation

Overnight cultures were started for the donor strains and the recipient strains in 50 ml M9 glucose media with their respective antibiotics using 500 µl of glycerol stock as the starter. In the case where the recipient was a clone, we started 5ml cultures in M9 glucose with their respective antibiotics. The next day, fresh 10 ml cultures were launched for the donor and recipient in M9 glucose media with their respective antibiotics by a 1:100 dilution of the overnight cultures. When cultures reached late log phase OD_{600} of 0.5, 4ml of each culture was centrifuged at 13500g for 1 minute at room temperature. The supernatant was discarded, and the pellet was washed using 2 ml antibiotic-free M9 glucose. The washing step was repeated once. Finally, the cells were pelleted and resuspended in 100 µl of antibiotic-free M9 glucose. The optical cell density was normalized according to the OD_{600} for each tube. 10 µl of the culture was diluted in 990 µl of antibiotic-free M9 glucose in a spectrometer cuvette. The OD at 600 nm was measured, and the amount of media needed for normalization was calculated using the following equation: medium to add in µl = $(OD_0 \times V_0)/(OD_F)–V_0$, where OD_0 is the current OD_{600} value, V_0 is the current volume that the cells are suspended in, and OD_F is the desired OD_{600} value. The final OD of cells should be around 20. Then the donors and recipients were combined in a 4:1 ratio by mixing 90 µl donor and 22.5 µl of the recipient, respectively, in a microcentrifuge tube. This was mixed well and pipetted onto a prewarmed antibiotic-free LB agar plate as two 20-µl spots and six 10-µl spots. All agar plates were incubated at 37 °C for two hours. The cells were removed from each plate by rinsing 750 µl of $MgSO_4$ 10^{-2}M over the plate twice. The cells were removed from each plate, and the aspirated suspension was placed in a microcentrifuge tube. The cells were centrifuged at 13,500g for 1 min at room temperature to pellet the cells. The pellet was resuspended in 750 µl of $MgSO_4$ 10^{-2}M. The cells were serially diluted repeatedly up to 10^{-6} dilution, and the appropriate

dilutions were plated on M9 agar plated with the required selection carbohydrate source and antibiotics. The plates were incubated at 37 °C for roughly 42 hours.

2.5. Calculation of conjugation efficiency

Conjugation efficiencies were calculated by dividing CFUs/ml of the trans-conjugant cells to the CFU/ml of the recipient without selecting for the trans-conjugants. Similarly, the escape frequency was calculated by dividing the CFUs/ml of the recipient control cells with selection to the CFU/ml of the recipient without selection.

2.6. Competition

Overnight cultures were started for the recombinants and ancestral Rel606 in 10 ml antibiotic free LB and M9 glucose media. The next day, the recombinants and the ancestral Rel606 were mixed in a 3:1 ratio (750 µl of recombinants + 250 µl of ancestral Rel606). 10 µl of this mixture was used to start the competition in 10 ml antibiotic free LB and M9 glucose media. This was incubated at 37°C in a shaker incubator. The mixture (0.1%) was transferred to fresh 10 ml antibiotic free LB and M9 glucose media every 24 hours. The competition was continued for 10 days (~100 generations: 10 generations per growth cycle). The competitions were performed in duplicates. After each cycle, 100 µl of the culture was lysed by boiling for 10 minutes and stored at -20°C. Glycerol stocks were made after each growth cycle by mixing equal volumes of the culture and sterile glycerol (60%) and stored at -80°C.

2.7. Genome sequencing

WGS was performed using Illumina technology. DNA samples were extracted using the genomic DNA NucleoMag tissue kit from Macherey-Nagel. The whole genome sequencing libraries were prepared and indexed using the Illumina DNA Prep Tagmentation kit and IDT for Illumina DNA/RNA UD Indexes kit A/B/C/D. The paired-end sequencing was performed with Illumina MiniSeq to a read length of 2 by 150 bp. Genomes were sequenced at an average depth of 100× for the pooled libraries and 20x for individual clones.

2.8. Bioinformatics

To detect the fraction of the genome that was recombined, reads from the genome of a recombinant clone or from a pool of recombinants were mapped simultaneously against the donor and recipient genome with the mapper bwa to reveal the integration of transferred donor fragments to these recipient strains. The recipient genome included the plasmid. Reads were then sorted into three categories using a Perl script. The first category was Donor specific reads that refers to reads matching to the Donor genome specifically; that is to say, either no match is found in the Recipient genome, or the match on the recipient genome has a lower score. Similarly, the second category was Recipient specific reads. The third category was reads that matched with equal score to both Donor and Recipient genomes. For the specific reads, when existing, the position of the match in the alternative genome was recorded. Then using an R program, the positions of the reads specific to either Donor or Recipient genome that had a match in the alternative genome were sorted according to the position in the Recipient genome and associated a Recipient (0) or Donor (1) score. This table was used to feed a Hidden Markov Model (HMM) with three states Recipient (R), Donor (D) or a mixture of both (M). The objective of the HMM is to find the fraction of the genome that belongs to each of these states. Each state is assigned an emission probability that corresponds to the probability of generating a read with a 1 or a 0 signature along the genome and a probability of transition between states that was assigned at a very low level: 10^{-7}. D state generates 1 with probability 0.9, and 0 with 0.1, while R generates 0 with probability 0.9 and 1 with 0.1. M generates 0 and 1 with equal probability of 0.5. It corresponds to a state where the recombination has not been resolved in the colony. Using package HMM and the verbatim algorithm we could recover the recombined fragment.

3. Results

The results section is divided into two parts. The first part encompasses the analysis of 60000 *E. coli* genomes, unveiling the dynamics of HGT using the HPI as an example. The second part delves into the results of mixing genomes in the lab using CoMBacGeMi.

3.1. Dynamics of HGT in *E. coli*.

3.1.1. Prevalence of HPI

The clusters corresponding to High Pathogenicity Island (HPI) were obtained by employing MMSeqs2 with a minimum similarity threshold of 90%, ensuring that both query coverage and target coverage are also above this threshold (Table 3.1). The resulting clusters were used to identify the *E. coli* genomes that harbor HPI genes, and SQLite was employed for this purpose. Based on the analysis, it was determined that 23,064 E. coli genomes contain at least 10 HPI genes.

Further analysis was conducted on the genome CA3372AA, which was found to contain 11 HPI genes based on the MMSeq2 results. These genes were present on a single contig in a sequential order, as shown in Table 3.2.

Table 3.1. Clusters corresponding to HPI

HPI gene	Cluster	Similarity	Alignment length	Nmismatches	Ngap openings	Query start	Query end	Target start	Target end	E-value	Bit score
ybtS	G4371	1	434	0	0	1	434	1	434	8.49E-290	884
ybtX	G4376	0.997	426	1	0	1	426	1	426	3.14E-261	801
Irp2	G4454	0.999	2035	2	0	1	2035	1	2035	0.00E+00	4216
fyuA	G4360	0.998	673	1	0	1	673	1	673	0.00E+00	1374
ybtT	G4366	0.996	262	1	0	1	262	6	267	2.62E-182	563
ybtQ	G4369	0.996	600	2	0	1	600	1	600	0.00E+00	1157
ybtP	G4370	0.996	570	2	0	1	570	31	600	0.00E+00	1103
intB	G4997	0.992	420	3	0	1	420	1	420	7.88E-283	863
ybtU	G4379	0.997	366	1	0	1	366	1	366	1.50E-253	775
ybtE	G4367	0.998	525	1	0	1	525	1	525	0.00E+00	1052
ybtA	G4359	0.99	319	3	0	1	319	1	319	4.88E-213	655
irp1	G4422	0.998	3163	6	0	1	3163	1	3163	0.00E+00	6435

* MMSeq2 search conditions: similarity>= 90%, query coverage >= 90%, target coverage >= 90%

Table 3.2. HPI in the CA3372AA genome

Locus tag	Gene name	Product	Inference	Contig num	Contig length	Gene Nums	Gene above
00493	*mbtI*	Salicylate_synthase	UniProtKB:P9WFX1	45	51500	00485-00522	IntA
00494	*ampG_1*	Protein_AmpG	UniProtKB:P0AE16	45	51500	00485-00522	
00495	00495	Putative_multidrug_export_ATP-binding/permease_protein	UniProtKB:Q7A4T3	45	51500	00485-00522	
00496	*btuD_2*	Vitamin_B12_import_ATP-binding_protein_BtuD	HAMAP:MF_01005	45	51500	00485-00522	
00497	00497	hypothetical_protein	NA	45	51500	00485-00522	
00498	*dltA*	D-alanine--poly(phosphoribitol) ligase_subunit_1	HAMAP:MF_00593	45	51500	00485-00522	
00499	*ubiE_1*	Ubiquinone/_UbiE	HAMAP:MF_01813	45	51500	00485-00522	
00500	00500	hypothetical_protein	NA	45	51500	00485-00522	
00501	*pikAV*	Thioesterase_PikA5	UniProtKB:Q9ZGI1	45	51500	00485-00522	
00502	*dhbE*	2,3-dihydroxybenzoate-AMP_ligase	UniProtKB:P40871	45	51500	00485-00522	
00504	*fyuA_2*	Pesticin_receptor	UniProtKB:P46359	45	51500	00485-00522	fyuA_1

* Here, the gene name and product are based on Prokka annotation

The study observed that the presence of certain genes in the HPI can be represented by multiple clusters in the database, which may be attributed to differences in genome quality and sequence evolution (Table 3.3). The investigation revealed that several of these clusters corresponded to gene fragments which can be a result of frameshifts, transposable element insertions, or suboptimal assembly. This targeted analysis helped us define some criteria to identify a gene, functional or inactivated in the database.

Table 3.3. Clusters corresponding to HPI gene *fyuA*

HPI gene	Cluster	Similarity	Alignment length	Nmismatches	Ngap openings	Query start	Query end	Target start	Target end	E-value	Bit score
fyuA	G4360	0.998	673	1	0	1	673	1	673	0.00E+00	1374
fyuA	G174888	0.998	547	1	0	127	673	19	565	0.00E+00	1144
fyuA	G88484	0.997	445	1	0	229	673	1	445	4.59E-305	941
fyuA	G120345	1	452	0	0	1	452	1	452	4.71E-296	915
fyuA	G397798	0.997	350	1	0	1	350	1	350	9.83E-222	700
fyuA	G124299	0.996	311	1	0	363	673	1	311	1.55E-205	653
fyuA	G109314	0.995	234	1	0	440	673	1	234	3.74E-150	492
fyuA	G122515	1	250	0	0	1	250	1	250	4.00E-148	486
fyuA	G337296	0.988	168	2	0	506	673	1	168	1.75E-102	352
fyuA	G185841	0.978	185	4	0	1	185	1	185	7.06E-101	347
fyuA	G173132	0.992	125	1	0	549	673	1	125	1.70E-72	263
fyuA	G110174	1	121	0	0	1	121	1	121	4.20E-60	226
fyuA	G200800	1	84	0	0	590	673	1	84	5.28E-45	179
fyuA	G204309	1	58	0	0	1	58	1	58	4.42E-21	102
fyuA	G253601	1	45	0	0	629	673	1	45	3.82E-19	96

* MMSeq2 search conditions: similarity>= 90%, target coverage >= 90%

3.1.2. Dynamics of horizontal gene transfer from distant species

To go beyond the HPI, we then looked at the overall database for HGT. To find the gene sequences coming from distant species, an MMSeq2 search was run using the clusters as the query and the SwissProt database (Bateman et al., 2021) as the target. The best non-*E. coli* matches were analyzed further. It was interesting to observe matches from distant organisms like *Mus musculus* (14 genes), Influenza A virus (10 genes), *Zea mays* (4 genes), and *Homo sapiens* (3 genes). These need to be analyzed further to check whether they are transfers or artifacts. Figure 3.1 shows the organisms other than *E. coli* with at least five matches.

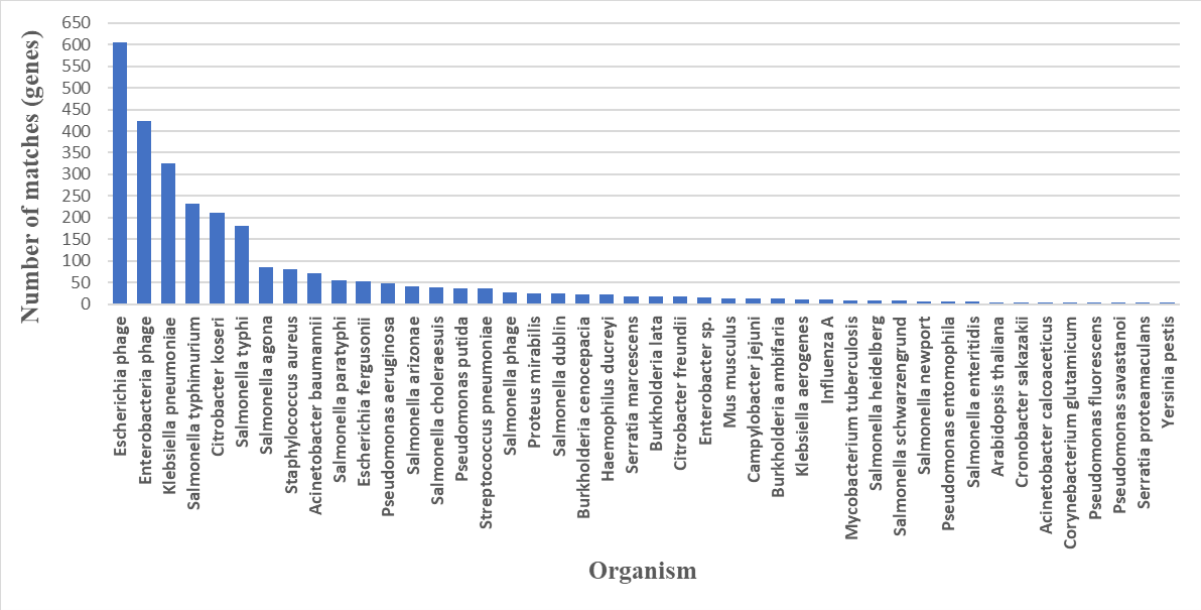


Figure 3.1. Best non-*E. coli* matches (>4 genes)

3.2. CoMBacGeMi: Examining Lab-based Genome Mixing

3.2.1. Conjugations & conjugation efficiencies

In this study, we performed conjugation experiments using the BW25113 Hfr library (donor) and three different sets of recipients: the Rel606 *tse2* library, Rel606 *tse2* clones (three clones picked from the Rel606 *tse2* library) and Rel606 Δ*matP* Δ*galK* (with Δ*galK* as control). The Rel606 *tse2* library harbors the *tse2* toxin gene under the regulation of a rhamnose-inducible promoter and displays resistance to kanamycin and zeocin. The Rel606 Δ*matP* Δ*galK* strains are resistant to spectinomycin. Following the mating with the Hfr library, selection for recombinants for the Δ*galK* strains was

performed by plating on minimal media with galactose and spectinomycin, while selection for recombinants for the *tse2* strains was carried out by plating on minimal media with rhamnose and zeocin.

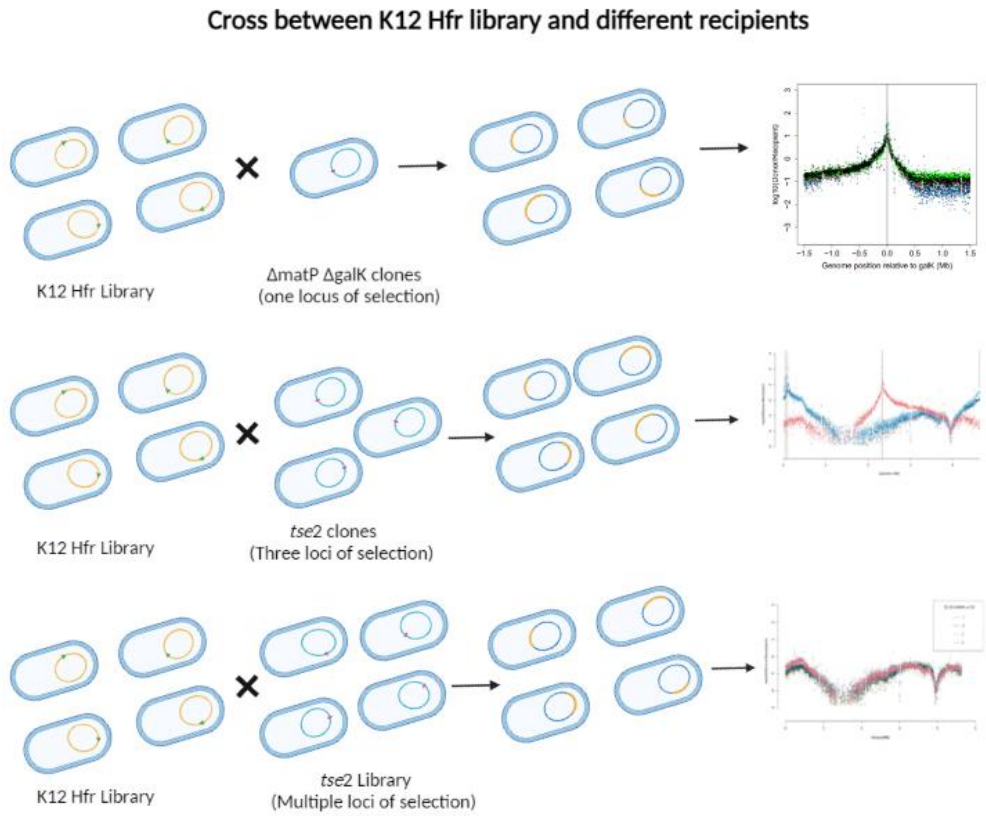


Figure 3.2. Schematic representing the various crosses between the K12 Hfr library and the different recipients **(**ΔmatP Δ*galK* clones, *tse2* clones and *tse*2 library). The green arrow represents the origin of transfer of the plasmid and for simplicity the integrated plasmid. The red line shows the position of the selection marker in the recipient.

The number of recombinants varied according to the recipient strains. There was a high variation in the conjugation efficiencies across the recipients. The Rel606 *tse2* library had the highest average recombination efficiency of 3.36% and a standard deviation (SD) of 1.54%. This was followed by the *tse2* clones with a mean recombination efficiency of 2.46% and a standard deviation of 2.03%. The ΔmatP Δ*galK* recipients showed an average recombination efficiency of 0.77% with a standard deviation of 0.44%. The Δ*galK* strain showed a recombination efficiency of 0.41%.

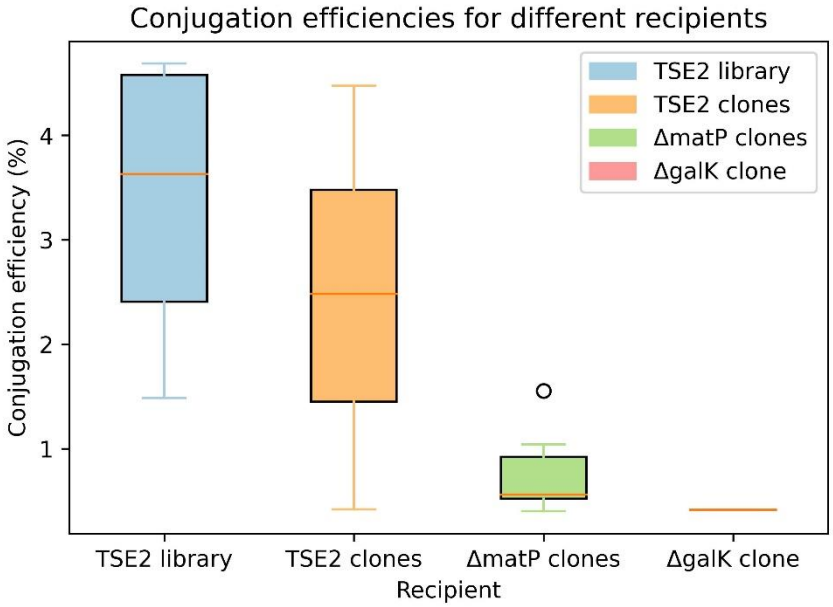


Figure 3.3. Conjugation efficiencies measured for Hfr library crossed with different recipients: Rel606 *tse2* library, Rel606 *tse2* clones, Rel606 ΔmatP Δ*galK* and Rel606 Δ*galK* (control for Rel606 ΔmatP Δ*galK*)

3.2.2. Role of MatP in the recombination pattern

Based on previous experiments involving conjugation, conducted by Thibault utilizing the BW25113 Hfr library and the Rel606 Δ*galK* recipient strain, it was observed that the presence of donor DNA exhibited an asymmetrical pattern of decay. Specifically, the decay extended over 1.5 Mb on the left end side of the chromosome relative to the selective marker. Surprisingly, it extended only over a distance of 0.4 Mb on the right end. Furthermore, it was found that the region marking the onset of the terminus of replication, macrodomain, delimited the right end side limit of the decay. This observation suggests that the probability of initiating or concluding recombination in the terminus macrodomain is limited. It was hypothesized that this limitation results from the difficulty in initiating or concluding recombination within the compactly organized terminus macrodomain, which has a limited number of copies in comparison to the origin.

It is known that the Macrodomain Ter Protein (MatP) accumulates on the chromosome around the Ter macrodomain. Through experiments involving the deactivation of *matP*, it has been discovered that this protein plays a critical role in

organizing the Ter macrodomain. When MatP is absent, the DNA within the macrodomain is less tightly packed, resulting in an increase in marker mobility. Furthermore, the Ter macrodomain undergoes segregation earlier in the cell cycle (Mercier *et al.*, 2008). These findings illustrate the significance of MatP in regulating the Ter macrodomain's organization, and demonstrate how the protein's presence is necessary for proper DNA compaction and segregation timing within the cell (Mercier *et al.*, 2008).

Conjugations were performed using the Hfr library with the recipient Rel606 ΔmatP Δ*galK* strain to understand the role of MatP in the recombination pattern. Three clones of the Rel606 ΔmatP Δ*galK* strain and one Rel606 Δ*galK* strain as control were used as the recipients. The pools from each of these crosses were analysed. All the pools had the same recombination pattern as the control. There was no change observed in the asymmetrical pattern of the decay.

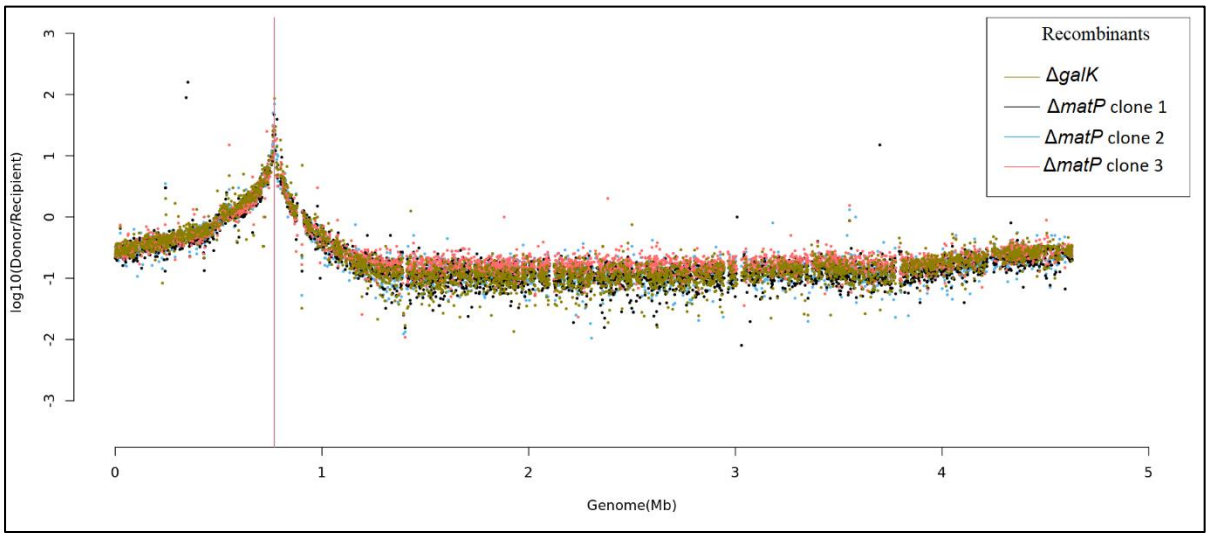


Figure 3.4. Log-ratio of donor to recipient coverage along the genome for the Δ*galK* and the three Δ*matP* clones.

3.2.3. Recombination pattern across the whole genome

We performed high-throughput sequencing of the bulk of the *tse2* library recombinants to analyse the recombination pattern across the entire genome. This allowed us to investigate the distribution of donor and recipient-specific reads along the genome. Our analysis of all four sequenced pools revealed a consistent recombination pattern. Notably, the highest peak for the donor-specific reads was observed at approximately

0.5 Mb. Additionally, we observed a significant decrease in the number of reads within the Ter macrodomain. Furthermore, a second sharp and narrow decrease in reads was observed in close proximity to the location of the zeocin cassette.

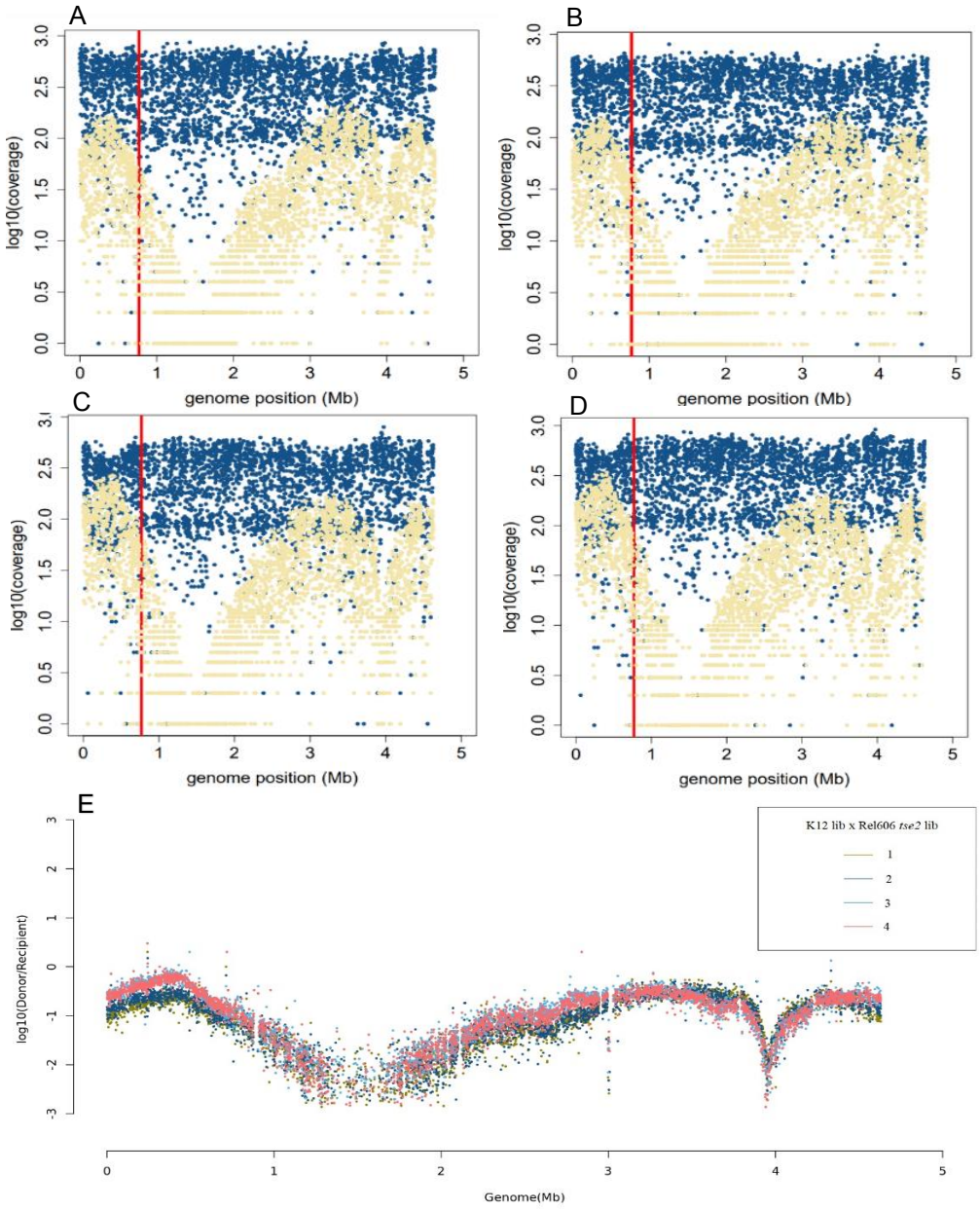


Figure 3.5. A, B, C, D. Recombination pattern across the whole genome using cross between K12 Hfr library and recipient *tse2* library for replicates 1, 2, 3 and 4 respectively. **E.** The recombination pattern shown in the form of the ration of donor to recipient specific reeds for all the four replicates.

3.2.4. Recombination pattern at different loci

To investigate the recombination patterns at different genetic loci, we selected three clones from the Rel606 *tse2* library and utilized them as recipients for conjugation. We observed distinct peaks in the density of donor-specific reads in all three recipient pools. Specifically, the peaks for recipient clone 1 and clone 2 were located on the left of the Ter macrodomain, while the peak for clone 3 was on the right side of the Ter macrodomain. Based on the donor-specific read peak, it first appeared that the *tse2* was inserted at the same position in recipient *tse2* clone 1 and clone 2.

However, upon analyzing the ratio of donor-specific reads to recipient-specific reads, we detected two distinct peaks for the recipient *tse2* clone 1 and clone 2. This finding suggested that examining the ratio of donor-specific reads to recipient-specific reads can provide valuable information about the location of the target of selection and might serve as an indicator of the insertion site of *tse2* in the recipients.

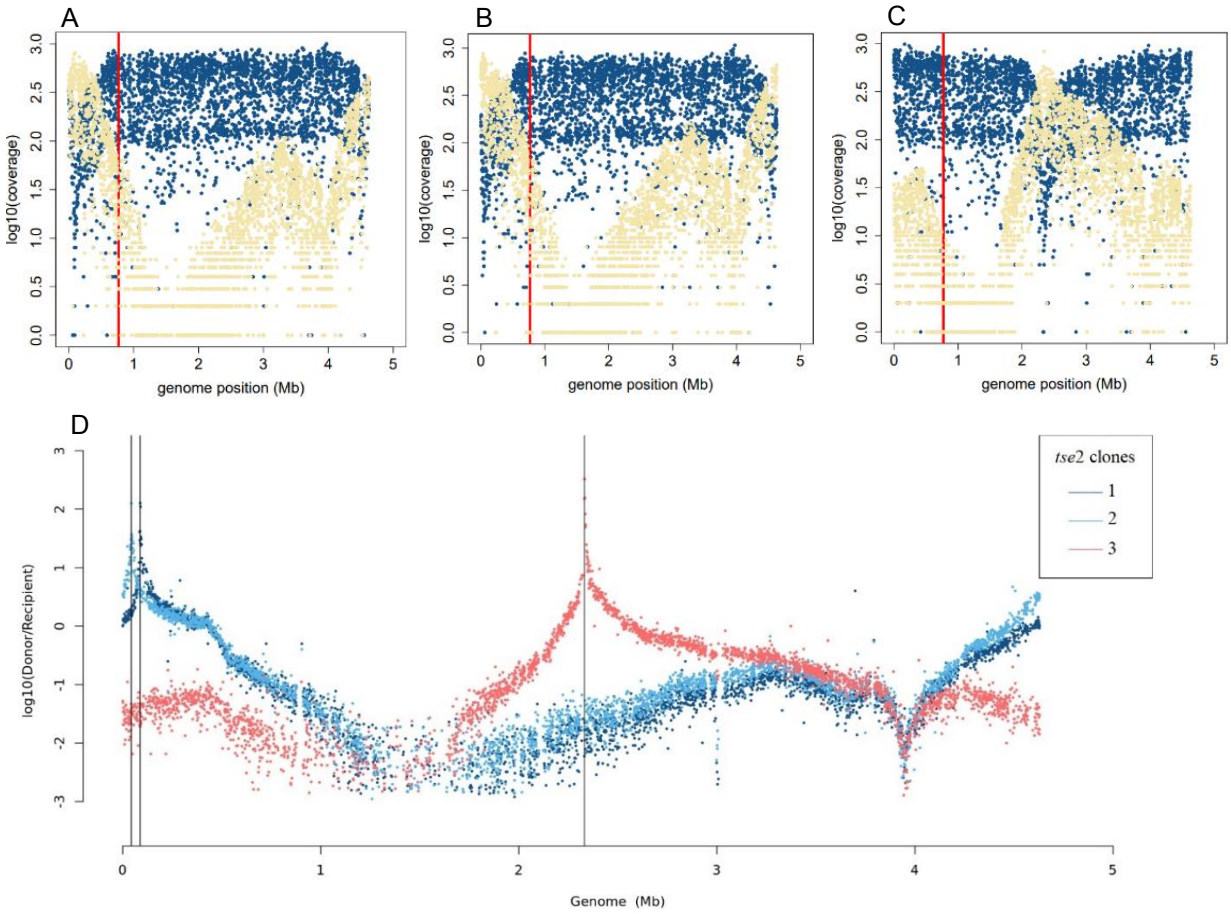


Figure 3.6. A, B, C. Recombination pattern observed for the *tse2* recipient clone 1, clone 2 and clone 3 respectively. **D.** Recombination pattern for all three recipient *tse2*

clones represented in the form of the ratio of donor to recipient specific reads. The vertical line represents the maxima of the ratio of donor to recipient specific reads.

3.2.5. Identifying the position of *tse2*

In order to determine the position of *tse2* in each of the three clones, we initially estimated the position of the *galK* loci by determining the maxima of the ratio of donor to recipient-specific reads in the recombinant pools of the Rel606 Δ*galK* and ΔmatP + Δ*galK* strains. The position of *galK* was already known, which allowed us to evaluate the accuracy of our estimates. All the four estimates were located only 1kb away from the actual position of the *galK* loci.

Table 3.4. The estimated and the actual position of *galK* locus on the genome

Sample	Estimated postion	Actual position	Difference
ΔgalK ΔmatP	768500	769500	1000
	770500	769500	-1000
	768500	769500	1000
ΔgalK	770500	769500	-1000

Next, we utilized the same approach to estimate the position of the different *tse2* insertions, with estimates positions of 88500, 43500, and 2331500 for the recipient *tse2* clones 1, 2, and 3, respectively. Primers were then designed for each for the clones at a distance of approximately 1 Kb on either side of the estimated position of *tse*2. The *tse2* cassette spans 2960 bp, indicating that if *tse*2 is present, the expected amplicon size after PCR should be roughly 5 Kb in length. Conversely, if *tse*2 is absent, the amplicon size would be 2 Kb. The resulting amplicon sizes were confirmed by gel electrophoresis.

For two clones, the shift in bands suggested that we had identified the *tse*2 insertion (Fig. 3.7). For the third clone we developed another set of primers to test a 2kb distance from the estimated position. The amplified products were sent for Sanger sequencing to confirm the insertion position of *tse*2. The insertion positions identified after Sanger sequencing are listed in Table 3.5.

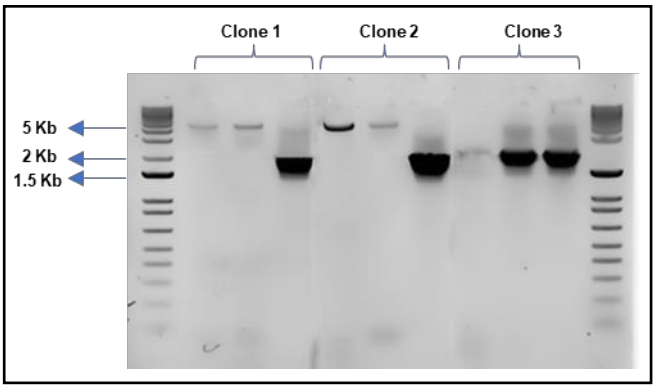


Figure 3.7. Agarose gel image showing the amplified products to check the presence of *tse2* in the clones 1, 2 & 3. Lanes 1 and 11 show the 1kb+ DNA ladder. Lanes 2-4: *tse2* clone 1 (replicate 1, 2 and the control respectively). Lanes 5-7: *tse2* clone 2 (replicate 1, 2 and the control respectively). Lanes 8-10: *tse2* clone 3 (replicate 1, 2 and the control respectively).

Table 3.5. The estimated and the actual position of *tse2* in the genome

Sample	Estimated position	Actual position	Difference
tse2 Clone 1	88500	88069	431
tse2 Clone 2	43500	43319	181
tse2 Clone 3	2331500	2329869	1631

3.2.6. Effect of selection on the fate of the recombinants

To understand the effect of selection on the recombinants, competition experiments were setup using a mix of 75% recombinants and 25% of Rel606 ancestral stain that was used to create the *tse2* library. After evolving this mixture for 10 days (~100 generations) in antibiotic free LB and antibiotic free M9 glucose media. The pool of the mixtures at different time points was sequenced (results awaited).

4. Discussion

4.1. Advantages of using a donor Hfr library and a recipient library

The transfer of chromosomal fragments through conjugation is a process that can occur at high frequency when a conjugative plasmid is inserted in the chromosome (Hayes, 1953). The resulting strain, a High Frequency Recombination strain or Hfr, has an integrated plasmid that enables the transfer of genetic material located in close proximity to the integrated conjugative plasmid. Specifically, the genetic material located 5' downstream of the origin of transfer (oriT) is transferred from the Hfr to the recipient strain (Virolle *et al.*, 2020).

This biased transfer of genetic material from a single Hfr to a recipient strain may have limitations for discovering unknown loci that impart novel phenotypic traits in natural isolates of *E. coli* (Quandt *et al.*, 2014). Since the position of such loci is unknown, a biased transfer from a fixed position in the chromosome may restrict their discovery. To increase the likelihood of success in identifying such loci, transfers should be initiated from multiple positions along the chromosome. Therefore, to achieve a uniform and unbiased transfer of DNA from a donor strain, a library of Hfr strains can be used, with each strain having the conjugative plasmid integrated at a different site along the chromosome.

To achieve a comprehensive understanding of the recombination patterns across the entire genome, it is necessary to complement the uniform and unbiased transfer of DNA from a donor Hfr library with an unbiased recipient library containing loci of selection integrated at different positions of the genome. Restricting the recipients to a single locus of selection would limit the search to the immediate vicinity of the loci of selection. Thus, in order to achieve an unbiased and uniform transfer of DNA from a donor strain, and its unbiased integration across the whole genome, a donor Hfr library and an unbiased recipient library were employed.

This approach allows for a comprehensive and unbiased transfer of genetic material, enabling the discovery of unknown loci and corresponding phenotypic traits across the entire genome. The use of an Hfr library offers a means of efficient transfer of genetic material, while an unbiased recipient library ensures that the search for unknown loci is not restricted to specific regions of the genome. Combining these two libraries

makes it possible to investigate the recombination patterns across the entire genome in a comprehensive and unbiased manner. Thus, this approach has the potential to uncover novel loci and corresponding phenotypic traits, which may have important implications for the study of bacterial genetics and the development of new therapeutic strategies

4.2. Variability in conjugation efficiencies

A significant degree of variability in the efficiencies of conjugation was observed among the various recipients employed in this study. The underlying reasons for this variability can be multifactorial in nature. One of the principal determinants of this variance is the genetic makeup of the recipient strains. Notably, one of the recipients (Rel606 *tse2* library) represents a recipient library, while the remaining recipients represent recipient clones. This difference is significant because the selection/counter-selection marker is dispersed throughout the genome of the recipient library. Since the donor utilized in this study was an Hfr library, which contains multiple loci for recombination, the presence of multiple loci for recombination in the recipient library likely contributed to a higher degree of conjugation efficiency. Conversely, in cases where selection occurs at only one locus, donors carrying the origin of transfer (oriT) at a considerable distance from the selection locus may be less capable of transferring DNA efficiently, compared to donors carrying the oriT located proximal to the selection locus.

Having multiple loci is not the sole determining factor for these variations as we also see variation within the *tse2* clones and between *tse2* clones and the ΔmatP clones. This can be explained by the recombination pattern across the whole genome observed by analyzing the recombinant pool of the *tse2* library. It can be observed that the frequency of recombination varies based on the location of the recombination event within the genome. The maxima being around 0.5 Mb and the least recombination was observed at the Ter macrodomain. Therefore, when the selection is happening at a single locus, the position of that locus on the genome might play a role in the conjugation efficiency.

Another important factor to consider is the genetic makeup of the recipient strains. Although all recipients are in the Rel606 background, various genetic modifications

have been introduced, including different antibiotic-resistant genes, *tse2* toxin, and *matP* knockout. These modifications may influence the efficiency of mating and recombination events by altering the expression of specific genes or the structure of the genome.

4.3. The role of MatP in regulating the pattern of recombination

MatP is a DNA-binding protein that has been shown to play a role in the organization and segregation of chromosomes in bacteria. Thus, it was hypothesized that MatP may also play a role in regulating recombination events in *E. coli*.

Conjugation experiments were conducted using the Rel606 ΔmatP Δ*galK* strain as the recipient and compared with the control strain, Rel606 Δ*galK* to investigate the role of MatP in the pattern of recombination. The analysis of the pools obtained from each of these crosses revealed that the recombination pattern in the Rel606 ΔmatP Δ*galK* strain was similar to that observed in the control strain, indicating that MatP may not have a significant role in regulating the pattern of recombination in this bacterial system. However, it is essential to note that various factors, like environmental conditions, may influence the role of MatP in recombination. Therefore, the role of MatP in regulating recombination events may be context-dependent.

It has been shown that deleting MatP in *E. coli* MG1655 causes severe chromosome segregation and cell division defects, which results in a significant number of anucleate cells (7%) and filamentous cells with condensed nucleoids (12%) during the exponential phase when cultured in Lennox Rich Medium. Filamentous and anucleate cells are most abundant during the exponential phase and decrease during the stationary phase. In contrast, most ΔmatP cells display a wild-type phenotype in minimal medium at every step of the growth curve (Mercier *et al.*, 2008).

Notably, all the bacterial cultures (except for a two-hour incubation step on LB agar) for the conjugation experiment were done in minimal media (for details, refer to the methods section). Hence further conjugations can be performed using rich media to check if the role of MatP in regulating recombination events is media-dependent.

4.4. Asymmetry and the lack of recombination in the Ter macrodomain

This study has revealed a genome-wide pattern of recombination that displays an asymmetry in the introgression signal. It was observed that the direction of the asymmetry based on the relative position of the locus of selection from the Ter macrodomain.

Upon plotting the ratio of donor-specific reads to recipient-specific reads for all recombinants of the recipient clones (*tse2* clones, ΔmatP Δ*galK* clones, and Δ*galK* clone) with respect to the genome position relative to the Ter macrodomain, it was observed that the decay pattern for all Δ*galK* strains resembles the mirror image of the decay pattern of *tse2* clone 3. To verify this finding, we plotted the decay pattern of a Δ*galK* clone and *tse2* clone 3, overlapped at the locus of selection. Subsequently, we generated the same graph, utilizing the mirror image of one of the two recipients. Remarkably, we observe that these patterns are indeed mirror images of each other, and this asymmetry is contingent upon the relative position of the locus of selection from the Ter macrodomain.

Our findings indicate that the probability of initiating or completing recombination in the terminus macrodomain is restricted in all the recipient strains tested. This observation is consistent with prior research indicating a lower recombination rate in the terminus macrodomain than in the rest of the genome (Touchon *et al.*, 2009). The observed reduced ratio of recombination to mutation in that region was suggested to be linked to a local elevation of the mutation rate around the terminus (Bobay *et al.*, 2015). The results observed in this study suggest a more direct impact of the Ter macrodomain on recombination. The MatP deletions could however not provide any significant insights for the lower recombination rate at the Ter macrodomain.

This lower rate of recombination at the Ter macrodomain can be because of several possible reasons. The first is that there might be a bias in the Hfr library or the recipient *tse2* library i.e., either the integration of the conjugative plasmid or the *tse2* cassette is depleted at the terminus. This can be easily checked using transposon sequencing. But the bias in the recipient library will not be able to explain the lower recombination at the terminus when the recipient is a clone and not a library. This suggest that either there is a bias in the donor library or there is no bias but the donor DNA is not able to recombine at the Ter macrodomain. One possible reason might be due to the low copy

number of the Ter region as compared to the origin. But the copy number alone is not enough to explain the substantial reduction in recombination.

One potential approach to understand the recombination at the terminus would be by creating Hfrs by integrating the conjugative plasmid at the terminus domain (start, middle and end) and performing conjugations using these Hfrs with the recipient *tse*2 library and with two recipient clones, one with the locus of selection at the left of the Ter macrodomain (e.g., Δ*galK*) and the other with the locus of selection on the right side of the Ter macrodomain (e.g., *tse*2 clone3).

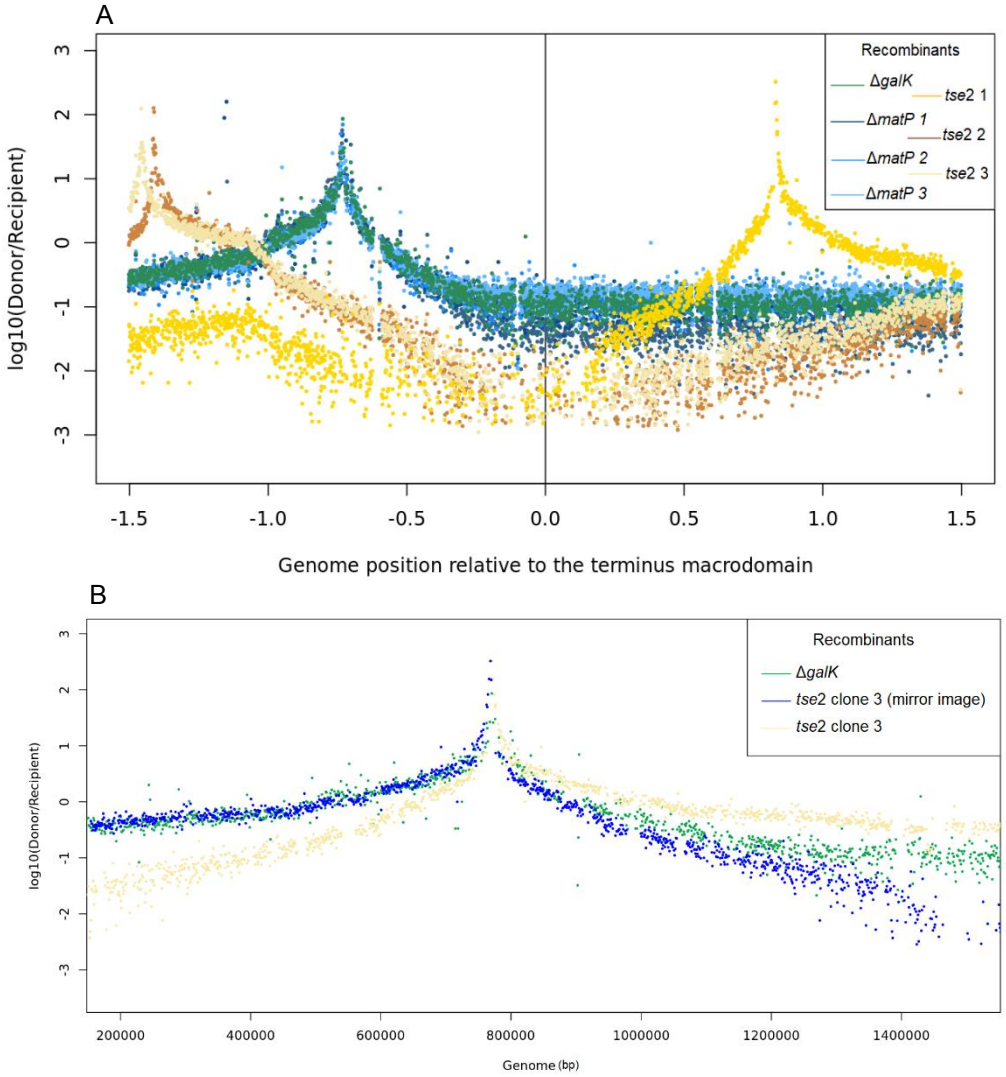


Figure 4.1. A. Ratio of donor to recipient reads for all the recipient clones (*tse2* clones, Δ*matP* clones and Δ*galK*) shown relative to the Ter macrodomain. **B.** The decay pattern of the Δ*galK* clone, tse2 clone 3, and the mirror image of the decay pattern of tse2 clone 3, overlapped at the locus of selection.

4.5. Detection of selection markers

The identification of genetic loci underlying strain-specific differences represents a fundamental challenge in the field of genetics. In this context, sequencing of a pool of recombinants has emerged as a powerful approach for precisely estimating the target of selection. This technique relies on the capacity to generate a large number of recombinants from crosses and requires only a single sequencing experiment.

Here, we report the successful application of the pool sequencing approach to the identification of the *galK* and *tse2* loci with an impressive accuracy of less than 1K b. Our results demonstrate the utility of this approach for the analysis of the molecular basis of several strains.

In addition, the pool sequencing approach can be extended to the investigation of complex traits that are influenced by multiple genetic loci. The capacity to accurately estimate the target of selection using this approach provides a powerful tool for studying such traits in diverse populations. Our results suggest that the pool sequencing approach has the potential to lead to a more comprehensive understanding of the genetic basis of complex traits and to the identification of novel targets for intervention. With its simplicity, accuracy, and precision, this approach has the capacity to facilitate the identification of key genetic loci and to deepen our understanding of complex genetic traits in diverse populations.

5. Conclusion

In this study, we employed Conjugation Mediated Bacterial Genome Mixing (CoMBacGeMi), a technique developed in the laboratory, to examine the influence of a mutant *matP* recipient on the recombination frequency bias that occurs at the terminus. Our findings from the analysis of the pools revealed that MatP was not a significant regulator of the recombination pattern. To obtain a comprehensive understanding of the recombination pattern across the entire genome, we expanded our investigation by using the Hfr library and a recipient library with *tse2* counter-selection marker. We observed that the recombination frequencies were not uniformly distributed across the genome, with the Ter macrodomain displaying extremely low recombination frequencies. Moreover, our study also found that the conjugation frequencies of recipient clones correlated with the recombination pattern observed across the whole genome, when crosses were performed using recipient clones with the selection marker at different loci. Lastly, we explored the ability of CoMBacGeMi to identify a target of selection, and we precisely identified the three different positions of insertion of *tse2* in the three tested clones from the analysis of the pools. Therefore, our study highlights that the accuracy, precision, and simplicity of CoMBacGeMi can be leveraged to identify key genetic loci and deepen our understanding of complex genetic traits in diverse populations.

6. References

Bobay, LM, Traverse, CC, and Ochman, H (2015). Impermanence of bacterial clones. Proc Natl Acad Sci U S A 112.

Carniel, E (2001). The Yersinia high-pathogenicity island: An iron-uptake island. Microbes Infect 3, 561–569.

Cassuto, E, West, SC, Mursalim, J, Conlon, S, and Howard-Flanders, P (1980). Initiation of genetic recombination: Homologous pairing between duplex DNA molecules promoted by recA protein. Proc Natl Acad Sci U S A 77.

Chowdhury, F et al. (2015). Diarrheal illness and healthcare seeking behavior among a population at high risk for diarrhea in Dhaka, Bangladesh. PLoS One 10.

Clark, AJ, and Margulies, AD (1965). ISOLATION AND CHARACTERIZATION OF RECOMBINATION-DEFICIENT MUTANTS OF. Proc Natl Acad Sci United States 53.

Clermont, O, Christenson, JK, Denamur, E, and Gordon, DM (2013). The Clermont Escherichia coli phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep 5.

Cohan, FM (2002). What are bacterial species? Annu Rev Microbiol 56.

Conway, T, Krogfelt, KA, and Cohen, PS (2004). The Life of Commensal Escherichia coli in the Mammalian Intestine . EcoSal Plus 1.

Cox, MM, and Lehman, IR (1981). recA protein of Escherichia coli promotes branch migration, a kinetically distinct phase of DNA strand exchange. Proc Natl Acad Sci U S A 78.

Craig, NL (1988). The mechanism of conservative site-specific recombination. Annu Rev Genet 22.

Denamur, E, Clermont, O, Bonacorsi, S, and Gordon, D (2021). The population genetics of pathogenic Escherichia coli. Nat Rev Microbiol 19.

Dixit, PD, Pang, TY, Studier, FW, and Maslov, S (2015). Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A 112.

Dixon, DA, and Kowalczykowski, SC (1993). The recombination hotspot χ is a

regulatory sequence that acts by attenuating the nuclease activity of the E. coli RecBCD enzyme. Cell 73.

Dower, NA, and Stahl, FW (1981). χ Activity during transduction-associated recombination. Proc Natl Acad Sci U S A 78.

Feil, EJ, and Spratt, BG (2001). Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 55.

Fomenkov, A, Sun, Z, Murray, IA, Ruse, C, McClung, C, Yamaichi, Y, Raleigh, EA, and Roberts, RJ (2020). Plasmid replication-associated single-strand-specific methyltransferases. Nucleic Acids Res 48.

Forde, BM, Ben Zakour, NL, Stanton-Cook, M, Phan, MD, Totsika, M, Peters, KM, Chan, KG, Schembri, MA, Upton, M, and Beatson, SA (2014). The complete genome sequence of Escherichia coli EC958: a high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone. PLoS One 9.

Fraser, C, Hanage, WP, and Spratt, BG (2007). Recombination and the nature of bacterial speciation. Science (80-) 315.

Galardini, M, Clermont, O, Baron, A, Busby, B, Dion, S, Schubert, S, Beltrao, P, and Denamur, E (2020). Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. PLoS Genet 16.

Garcia-Vallve, S, Romeu, A, and Palau, J (2000). Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. Genome Res 10, 1719.

Grindley, NDF, Whiteson, KL, and Rice, PA (2006). Mechanisms of site-specific recombination. Annu Rev Biochem 75.

Hallstrom, KN, and McCormick, BA (2014). Pathogenicity Islands: Origins, Structure, and Roles in Bacterial Pathogenesis, Elsevier Ltd.

Hayes, W (1953). The mechanism of genetic recombination in Escherichia coli. Cold Spring Harb Symp Quant Biol 18.

Hobbs, MD, Sakai, A, and Cox, MM (2007). SSB protein limits RecOR binding onto single-stranded DNA. J Biol Chem 282.

Jang, J, Hur, HG, Sadowsky, MJ, Byappanahalli, MN, Yan, T, and Ishii, S (2017). Environmental Escherichia coli: ecology and public health implications—a review. J Appl Microbiol 123.

Kallonen, T, Brodrick, HJ, Harris, SR, Corander, J, Brown, NM, Martin, V, Peacock, SJ, and Parkhill, J (2017). Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131.

Kaper, JB, Nataro, JP, and Mobley, HLT (2004). Pathogenic Escherichia coli. Nat Rev Microbiol 2, 123–140.

Karaolis, DKR (2001). Pathogenicity Islands. Encycl Genet, 1422–1424.

Karch, H, Schubert, S, Zhang, D, Zhang, W, Schmidt, H, Ölschläger, T, and Hacker, J (1999). A Genomic Island, Termed High-Pathogenicity Island, Is Present in Certain Non-O157 Shiga Toxin-Producing Escherichia coli Clonal Lineages. Infect Immun 67, 5994.

De la Cruz, F, and Davies, J (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol 8, 128–133.

Ma, NJ, Moonan, DW, and Isaacs, FJ (2014). Precise manipulation of bacterial chromosomes by conjugative assembly genome engineering. Nat Protoc 9.

Mandel, M, and Higa, A (1970). Calcium-dependent bacteriophage DNA infection. J Mol Biol 53.

McKane, M, and Milkman, R (1995). Transduction, restriction and recombination patterns in Escherichia coli. Genetics 139.

Mercier, R, Petit, MA, Schbath, S, Robin, S, El Karoui, M, Boccard, F, and Espéli, O (2008). The MatP/matS Site-Specific System Organizes the Terminus Region of the E. coli Chromosome into a Macrodomain. Cell 135.

Murray, CJ et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet (London, England) 399, 629.

Ochman, H, and Bergthorsson, U (1998). Rates and patterns of chromosome evolution in enteric bacteria. Curr Opin Microbiol 1.

Oliveira, PH, Touchon, M, Cury, J, and Rocha, EPC (2017). The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun 8.

Proteau, G, Sidenberg, D, and Sadowski, P (1986). The minimal duplex DNA sequence required for site-specific recombination promoted by the FLP protein of yeast in vitro. Nucleic Acids Res 14.

Quandt, EM, Deatherage, DE, Ellington, AD, Georgiou, G, and Barrick, JE (2014). Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in Escherichia coli. Proc Natl Acad Sci U S A 111.

Redfield, RJ (2001). Do bacteria have sex? Nat Rev Genet 2.

Register, JC, and Griffith, J (1985). The direction of RecA protein assembly onto single strand DNA is the same as the direction of strand assimilation during strand exchange. J Biol Chem 260.

Rubnitz, J, and Subramani, S (1984). The minimum amount of homology required for homologous recombination in mammalian cells. Mol Cell Biol 4.

Sakoparnig, T, Field, C, and van Nimwegen, E (2021). Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. Elife 10.

Schubert, S, Darlu, P, Clermont, O, Wieser, A, Magistro, G, Hoffmann, C, Weinert, K, Tenaillon, O, Matic, I, and Denamur, E (2009). Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog 5.

Schubert, S, Rakin, A, and Heesemann, J (2004). The Yersinia high-pathogenicity island (HPI): Evolutionary and functional aspects. Int J Med Microbiol 294, 83–94.

Selander, RK, and Levin, BR (1980). Genetic diversity and structure in Escherichia coli populations. Science (80-) 210.

Shaw, LP, Rocha, EPC, and MacLean, RC (2022). Restriction-modification systems have shaped the evolution and distribution of plasmids across bacteria. BioRxiv, 2022.12.15.520556.

Shibata, T, Cunningham, RP, DasGupta, C, and Radding, CM (1979). Homologous

pairing in genetic recombination: complexes of recA protein and DNA. Proc Natl Acad Sci U S A 76.

Soucy, SM, Huang, J, and Gogarten, JP (2015). Horizontal gene transfer: building the web of life. Nat Rev Genet 2015 168 16, 472–482.

Steinegger, M, and Söding, J (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 2017 3511 35, 1026–1028.

Taylor, A, and Smith, GR (1980). Unwinding and rewinding of DNA by the RecBC enzyme. Cell 22.

TAYLOR, AL, and THOMAN, MS (1964). THE GENETIC MAP OF ESCHERICHIA COLI K-12. Genetics 50.

Tenaillon, O, Skurnik, D, Picard, B, and Denamur, E (2010). The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8.

Touchon, M et al. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5.

Touchon, M, Perrin, A, De Sousa, JAM, Vangchhia, B, Burn, S, O'Brien, CL, Denamur, E, Gordon, D, and Rocha, EPC (2020). Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genet 16.

Typas, A et al. (2008). High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods 2008 59 5, 781–787.

Vasu, K, and Nagaraja, V (2013). Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. Microbiol Mol Biol Rev 77.

Virolle, C, Goldlust, K, Djermoun, S, Bigot, S, and Lesterlin, C (2020). Plasmid transfer by conjugation in gram-negative bacteria: From the cellular to the community level. Genes (Basel) 11.

Volkmer, B, and Heinemann, M (2011). Condition-Dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. PLoS One 6.

Vulić, M, Dionisio, F, Taddei, F, and Radman, M (1997). Molecular keys to

speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A 94.

Walk, ST, Alm, EW, Gordon, DM, Ram, JL, Toranzos, GA, Tiedje, JM, and Whittam, TS (2009). Cryptic lineages of the genus Escherichia. Appl Environ Microbiol 75.

Wiedenbeck, J, and Cohan, FM (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35.

Yang, SC, Lin, CH, Aljuffali, IA, and Fang, JY (2017). Current pathogenic Escherichia coli foodborne outbreak cases and therapy development. Arch Microbiol 199.

Yang, ZK, Luo, H, Zhang, Y, Wang, B, and Gao, F (2019). Pan-genomic analysis provides novel insights into the association of E.coli with human host and its minimal genome. Bioinformatics 35.

Yu, D, Banting, G, and Neumann, NF (2021). A review of the taxonomy, genetics, and biology of the genus escherichia and the type species escherichia coli. Can J Microbiol 67.

Zhou, Z, Alikhan, NF, Mohamed, K, Fan, Y, and Achtman, M (2020). The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 30, 138–152.