SELECTIVE BARRIERS
TO
HORIZONTAL GENE TRANSFER

by

Hande Acar

December, 2016

A thesis presented to the
Graduate School
of the
Institute of Science and Technology Austria, Klosterneuburg, Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

IST AUSTRIA
Institute of Science and Technology

The dissertation of Hande Acar, titled *Selective Barriers to Horizontal Gene Transfer*, is approved by:

Supervisor: Jonathan P. Bollback, PhD, IST Austria, Klosterneuburg, Austria

Signature: ______________________________

Committee Member: Călin C. Guet, PhD, IST Austria, Klosterneuburg, Austria

Signature: ______________________________

Committee Member: John F. Baines, PhD, Christian-Albrechts-University of Kiel, Kiel, Germany & MPI for Evol. Biol. Plön, Plön, Germany

Signature: ______________________________

Exam Chair: Herbert Edelsbrunner, PhD, IST Austria, Klosterneuburg, Austria

Signature: ______________________________

I hereby declare that this dissertation is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: ____________________

Hande Acar

December 30, 2016

Abstract

Horizontal gene transfer (HGT), the lateral acquisition of genes across existing species boundaries, is a major evolutionary force shaping microbial genomes that facilitates adaptation to new environments as well as resistance to antimicrobial drugs. As such, understanding the mechanisms and constraints that determine the outcomes of HGT events is crucial to understand the dynamics of HGT and to design better strategies to overcome the challenges that originate from it.

Following the insertion and expression of a newly transferred gene, the success of an HGT event will depend on the fitness effect it has on the recipient (host) cell. Therefore, predicting the impact of HGT on the genetic composition of a population critically depends on the distribution of fitness effects (DFE) of horizontally transferred genes. However, to date, we have little knowledge of the DFE of newly transferred genes, and hence little is known about the shape and scale of this distribution.

It is particularly important to better understand the selective barriers that determine the fitness effects of newly transferred genes. In spite of substantial bioinformatics efforts to identify horizontally transferred genes and selective barriers, a systematic experimental approach to elucidate the roles of different selective barriers in defining the fate of a transfer event has largely been absent. Similarly, although the fact that environment might alter the fitness effect of a horizontally transferred gene may seem obvious, little attention has been given to it in a systematic experimental manner.

In this study, we developed a systematic experimental approach that consists of transferring 44 arbitrarily selected *Salmonella typhimurium* orthologous genes into an *Escherichia coli* host, and estimating the fitness effects of these transferred genes at a constant expression level by performing competition assays against the wild type.

In chapter 2, we performed one-to-one competition assays between a mutant strain carrying a transferred gene and the wild type strain. By using flow cytometry we estimated selection coefficients for the transferred genes with a precision level of 10^{-3},

and obtained the DFE of horizontally transferred genes. We then investigated if these fitness effects could be predicted by any of the intrinsic properties of the genes, namely, functional category, degree of complexity (protein-protein interactions), GC content, codon usage and length. Our analyses revealed that the functional category and length of the genes act as potential selective barriers. Finally, using the same procedure with the endogenous *E. coli* orthologs of these 44 genes, we demonstrated that gene dosage is the most prominent selective barrier to HGT.

In chapter 3, using the same set of genes we investigated the role of environment on the success of HGT events. Under six different environments with different levels of stress we performed more complex competition assays, where we mixed all 44 mutant strains carrying transferred genes with the wild type strain. To estimate the fitness effects of genes relative to wild type we used next generation sequencing. We found that the DFEs of horizontally transferred genes are highly dependent on the environment, with abundant gene–by-environment interactions. Furthermore, we demonstrated a relationship between average fitness effect of a gene across all environments and its environmental variance, and thus its predictability. Finally, in spite of the fitness effects of genes being highly environment-dependent, we still observed a common shape of DFEs across all tested environments.

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

I would like to first thank my advisor, Jonathan Paul Bollback for providing guidance in all aspects of my life, encouragement, sound advice, and good teaching over the last six years.

I would also like to thank the members of my dissertation committee – Călin C. Guet and John F. Baines – not only for their time and guidance, but for their intellectual contributions to my development as a scientist.

I would like to thank Flavia Gama and Rodrigo Redondo who have taught me all the skills in the laboratory with their graciousness and friendship. Also special thanks to Bollback group for their support and for providing a stimulating and fun environment: Isabella Tomanek, Fabienne Jesse, Claudia Igler, and Pavel Payne.

Jerneja Beslagic is not only an amazing assistant, she also has a smile brighter and warmer than the sunshine, bringing happiness to every moment. Always keep your light Neja, I will miss our invaluable chatters a lot.

They are the secret super power behind every experimental biologist at IST, but for me they are much more than that, they are my angels, Mersija Smailagic, Renate Srsek, and Renate Eric. They bring their hugs together with the freshly autoclaved LB agar bottles, but actually those hugs are warmer than those bottles. I felt home whenever you gave me a hug my dear ladies.

Throughout the whole 'different environments' marathon, my dear friend Katharina Pöcher was always with me, supporting me, helping me, working with me… I cannot thank enough to you.

I met him in my first week at IST and since then he was my best friend, but also sometimes my brother, sometimes my mentor… Arjun Radhakrishna, thank you for being yourself.

Cezara Drăgoi is a very special friend for me as well, giving me strength at times, teaching me how to enjoy life, having fun together, tons of laughter… I always feel very lucky that I get to meet you.

Mato Lagator entered into my life with a fierce competition, not with me luckily. He is an amazing person, with an extraordinary taste of humor, and a huge heart. I cannot explain how much I love you Mato, but maybe saying this would mean something, because of you I almost regret that my PhD is over. He was supporting me in all circumstances. In fact, you deserve a dedication of this thesis because of your contributions…

She was my local mum in Vienna, Aslihan Karabiber. I am grateful for the chain of events that made us meet each other. And Eda Ertugrul, thank you not only for taking care of Kaymak, but also being part of my family.

Throughout the last twelve years, they would be impossible without you, Tuğba Keskin, my best friend, was with me all the time. Thank you my dear for helping me to find strength to get through the difficult times, and for all the emotional support, camaraderie, entertainment, and caring. I love you…

Lastly, and most importantly, I wish to thank my family, to whom I dedicate this thesis, my mother Asuman Acar, she is like a source of strength and morale in my desperate moments; my sisters Handan Acar, you were more than a sister for me, more than a friend, more than many things… and Aydan Acar, she never unclasps my hand. They supported me, raised me, taught me, and loved me...

This study was supported by European Research Council ERC CoG 2014 – EVOLHGT, under the grant number 648440.

About the Author

Hande Acar studied at the Middle East Technical University (METU), in Ankara, Turkey, working under the supervision of Inci Togan, PhD, on the national project TURKHAYGEN-I to investigate the origins of domestication of livestock animals in the Fertile Crescent by estimating the genetic diversity and admixture of Turkish native sheep breeds. She earned both her B.Sc. and M.Sc. degrees from METU.

For her PhD she worked in the field of experimental evolution with Jonathan P. Bollback, PhD, at IST Austria. Her main research interests include better understanding bacterial adaptation and innovation mechanisms, and bacterial genome evolution. During the first year of her PhD, she collaborated with external scientists on finding a consistent and reliable way of analyzing bacterial growth curves and has published these results in the high-impact journal *Molecular Biology and Evolution* (MBE 2014, 31:1).

In summer 2014 Hande was accepted to the Cold Spring Harbor Laboratory Advanced Bacterial Genetics course. She has also presented her research results in the *SMBE-LGT* satellite conference in Kiel, Germany in 2014 and *EVOLUTION* conference in Texas, USA in 2016.

List of Publications

Hall, B. G., **Acar, H.**, Nandipati, A., & Barlow, M. (2014). Growth rates made easy. *Molecular Biology and Evolution*, *31*(1), 232–238.

Table of Contents

List of Figures

List of Tables

List of Symbols/Abbreviations

AMP	Ampicillin
AMPR	Ampicillin Resistance
ATc	Anhydrotetracycline
C	Celsius Degree
$CaCl_2$	Calcium Chloride
CAA	Casein Amino Acid
CAM	Chloramphenicol
CDS	Coding Sequence
CFP	Cyan Fluorescence Protein
DFE	Distribution of Fitness Effects
DMSO	Dimethyl Sulfoxide
FOP	Frequency of Optimal Codons
GFP	Green Fluorescence Protein
GTA	Gene Transfer Agents
HGT	Horizontal Gene Transfer
IP	Interaction Partners
KAN	Kanamycin
KNR	Kanamycin Resistance
LB	Lennox Broth
LGT	Lateral Gene Transfer
$MgSO_4$	Magnesium Sulfate
NaCl	Sodium Chloride
NOX	Anaerobic Condition
PPI	Protein-Protein Interaction
RMS	Restriction Modification System
s	Selection Coefficient
SPR	Spectinomycin Resistance
TET	Tetracycline
TMP	Trimethoprim
TPM	Transcript per Million
YFP/VENUS	Yellow Fluorescence Protein

1 Introduction

1.1 Novel Genes and the Role of Horizontal Gene Transfer

What is the source of evolutionary novelty? Mutations, gene duplications, or horizontal gene transfer? Mutations can give rise to new functions and phenotypes but at the potential loss of previous functions and phenotypes. Gene duplications, on the other hand, can undergo neo-functionalization or sub-functionalization. One copy of a gene can acquire a new function through the accumulation of mutations while the old copy preserves the original function of the gene. However, the transiently non-functional copy must escape stochastic loss for sufficient time to acquire the new function. In the end, the organism could have two genes with similar coding sequences but different functions resulting in a novel phenotype (Innan & Kondrashov 2010). New gene function and phenotypes, can also arrive through **horizontal gene transfer** (HGT), defined as the transfer of genetic material from the genome of one organism to that of another, without parental relationship between the two organisms. HGT is pervasive across the tree of life, occurring between both closely and distantly related species and even between different kingdoms (Keeling & Palmer 2008). For instance, the red-green color polymorphism of the pea aphid, *Acyrthosiphon pisum*, results from carotenoid biosynthesis genes that were acquired from fungi (Moran & Jarvik 2010). Another example is the mannanase enzyme that exists in the genome of coffee berry borer beetle, *Hypothenemus hampei*. This enzyme, transferred from bacteria, enables the beetle to digest the complex sugars in coffee beans, thus turning it into an industrially relevant pest (Acuña et al. 2012). Despite the numerous examples of HGT between different kingdoms, HGT occurs most frequently within and among the Archaea and Eubacteria, which are the vast majority of the world's biomass and biological diversity (Ochman et al. 2000). In this thesis I focus on horizontal gene transfer and the evolutionary barriers that restrict the potential origin of novel functions and phenotypes.

In microbial populations, HGT is the primary source of novel genetic material, introducing new genes at rates far greater than that of gene duplication (Treangen & Rocha 2011), thus accelerating the appearance of novel metabolic capacities and phenotypes at rates far greater than that of mutation and gene duplication alone (Lawrence 2002). In fact, HGT is so prevalent that some have proposed that we consider diversity of microbial populations as a dynamic gene pool, the so-called *horizontal gene pool*, that creates novel genetic combinations and partially compensates for the cost of asexual reproduction, while offering the opportunity to exploit novel ecological niches (Barkay & Smets 2005).

The great evolutionary advantage of HGT for microbes, however, creates a challenge for human health and agriculture: drug resistance or virulence mechanisms spread quickly among parasites as a result of the substantial fitness benefits they receive from these transfer events (Andam et al. 2015).

1.2 Extent of Horizontal Gene Transfer in Microbes

Horizontal gene transfer appears to be extremely common as supported by numerous studies. For instance, Dagan and Martin (2007) have estimated that each "gene family" has experienced at least one successful HGT event during its evolutionary past. Other work exploring the whole genome sequences of 61 *Escherichia coli* strains showed that for any given *E. coli* genome, only about 20% of its genes are common to all other *E. coli* genomes (Lukjancenko et al. 2010). However, the extent of HGT is hard to estimate and still not precisely known (Zhaxybayeva & Doolittle 2011).

In the late 1990s, several studies attempted to quantify the extent of HGT in different organisms. Although the amount of data was sparse at the time, it was perceived as simply the tip of the horizontal transfer iceberg (Kurland 2005; Lawrence & Ochman 1997). This conjecture led some scientists to suggest that HGT occurs so frequently that it required a change in the standard model for the inheritance of genes in microbes. Some authors even went so far as to suggest that we should change the representation of phylogenetic relationship of microbial organisms from a "tree of life" metaphor to a "web of life" metaphor (Doolittle 1999b). In the early 2000s, technological developments allowed us to move beyond the limited datasets and sample sizes of the past, and begin more systematic phylogenetic analyses of whole genomes. These studies showed that although HGT is an important mechanism in shaping the evolution of microbial genomes, the inferred rate of HGT was not consistent with the idea of "rampant HGT as the essence of the phylogenetic process" (Kunin 2003; Kurland 2005). However, the debate over a tree of life versus a web of life still rages (Kurland 2005; Zhaxybayeva & Doolittle 2011).

1.3 Mechanisms of Horizontal Gene Transfer

Several mechanisms of HGT have been identified between bacteria: transformation, conjugation, transduction, gene transfer agents, bacterial vesicles, and bacterial nanotubes (Figure 1) (Popa & Dagan 2011). For the first three mechanisms, a particularly detailed understanding exists. However, the role and extent of the last three have yet to be determined (Soucy et al. 2015). Therefore, I will focus on the best understood briefly below.

Transformation is the bacterial uptake of free DNA in the surrounding environment (Figure 1-a), when bacteria are competent (i.e., the state in which bacteria are able to naturally acquire extracellular DNA). Competency, as we currently understand it, is a response to altered growth conditions or triggered by quorum sensing. In natural transformation, a group of specialized bacterial proteins are responsible for the translocation of the extracellular DNA into the cytoplasm (Seitz & Blokesch 2013). The exact mechanisms of natural transformation vary among organisms, but successful transformation requires, at least, two steps: (i) available DNA in the environment, and (ii) integration into the genome via some form of recombination (Chen & Dubnau 2004; Thomas & Nielsen 2005).

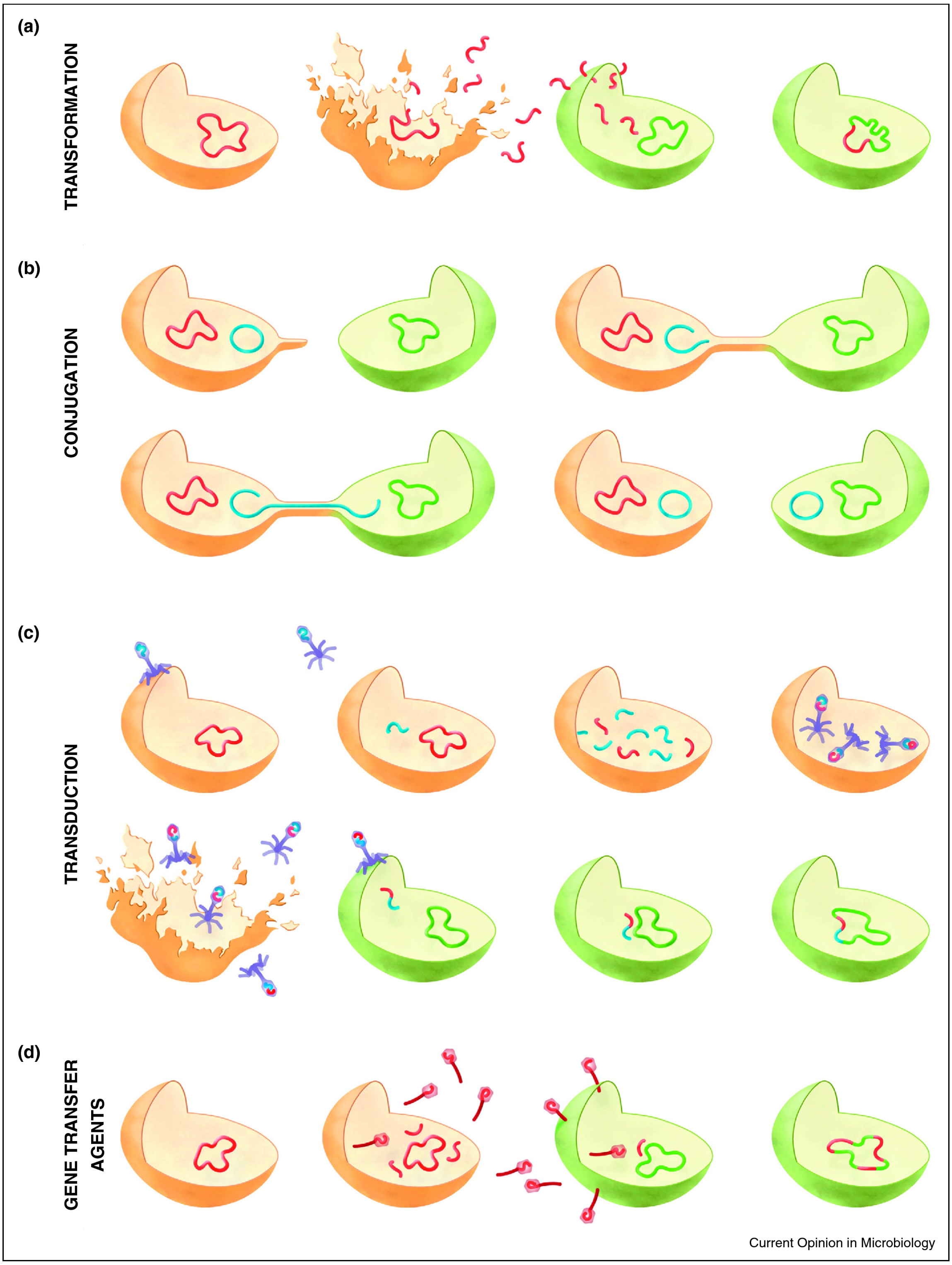

Figure 1. Schematic representation of several mechanisms of HGT. Image is reproduced from Popa and Dagan 2011 with permission.

Conjugation is a mode of gene transfer that occurs via mobile plasmids (Figure 1-b). For conjugation to occur, donors and recipient cells require a physical cell-to-cell connection, which is mediated by specialized genes carried on the plasmid and may be linked to quorum sensing. For example, a hair-like surface appendage (i.e., the pilus) provides a cell-to-cell bridge in Gram-negative bacteria. Once this connection is established, a signaling event activates a group of proteins to transfer a newly replicated copy of the plasmid DNA from the donor to the recipient cell through the pilus (Frost et al. 2005). The transfer of genetic material via plasmids can happen either by the stable integration of the plasmid in the host or by the mobilization and integration of the transferred gene via mobile elements present on the plasmid or in the host.

Transduction is a mode of gene transfer in which certain types of bacteriophage can transfer DNA fragments from the genome of one host to that of another (Frost et al. 2005). Since DNA is protected within the phage capsid, this mechanism may offer HGT between more physically distant donor and recipient bacteria than transformation and conjugation. Bacteriophages attach to the cell surface of their microbial host and insert their genetic material into the cell. If the phage is temperate the phage genomic material may become integrated into the host chromosome and remain dormant while replicating along with the host – a bacteriophage lifestyle choice called lysogeny. This state of dormancy continues until the induction of the lysogenized phage, usually through bacterial DNA damage or stress conditions experienced by the host. Once induced, the phage enters the lytic cycle resulting in lysis of the host cell and release of the new phage particles. During mobilization of the phage DNA from the host chromosome, bacterial DNA fragments may incidentally be packaged within the phage genome. These bacterial DNA fragments, then, may be transferred to the new hosts through another round of transduction, and even integrated into the recipient cell's chromosome through recombination or genomic integration (Canchaya et al. 2003). This role of transduction in the horizontal transfer of antimicrobial resistance genes came into the spotlight with the recent discovery that phages are capable of transferring antimicrobial resistance in chicken meat (Shousha et al. 2015).

Gene transfer agents (GTAs) constitute another mechanism of transfer, seemingly similar to transfer by bacteriophage (Figure 1-d). GTA particles are derived from a bacteriophage infection in which the physical structure of the phages fails to encapsulate their genomic DNA/RNA but rather carry host DNA fragments. These particles have a head and tail of phages and are able to adsorb and infect a new host. Some host cell interaction seems to be necessary suggesting that gene packaging is actively carried out by the donor cell, and the genes coding for the GTA particles are conserved throughout the genomes of the species with GTA production capacity. (Lang et al. 2012).

Extracellular vesicles released by bacteria have been implicated in signal transmission (quorum sensing), phage defense, and HGT. For instance, they contain genetic material that facilitates the transfer of virulence genes (Kolling & Matthews 1999); and in liquid environments vesicle release impedes the infection rate of bacteriophages by inactivating them (Biller et al. 2014).

Nanotubes, derived from bacterial membranes were recently discovered in both Gram-positive and Gram-negative bacteria, such as, *Bacillus subtilis*, *Staphylococcus aureus*, *Acinetobacter baylyi*, and *E. coli* (Dubey & Ben-Yehuda 2011; Pande et al. 2015). These cell-to-cell connections are able to mediate the exchange of not only proteins, amino acids and metabolites, but also mRNA molecules and non-conjugative plasmids. The extent of the role of these tubular conduits in HGT is yet to be fully studied and understood.

1.4 Methods for Detecting HGT

Over the past few decades, several computational techniques have been developed to identify horizontally transferred genes: analyses based on compositional properties of the genomes, detection of phylogenetic inconsistencies, network analysis, and comparative genomics (Kuo & Ochman 2009).

Compositional analyses rely on several characteristic features of genomes to detect horizontal gene transfer, such as: GC content, codon usage, and nucleotide motifs (Sharp et al. 2010; Raghavan et al. 2012). Genes within an organism's genome tend to have similar nucleotide compositions and codon usage, yet they often vary considerably between species (Lawrence 2002; Kuo & Ochman 2009). By comparing features of each gene with the average values of the whole genome, or some reference set of genes, researchers can detect outliers and infer them as "horizontally transferred genes". These methods, however, suffer from some limitations. First, following a successful HGT event, the compositional difference between the transferred gene and the recipient genome decreases with the accumulation of mutations in the transferred gene over time, in a process called amelioration. This process prevents the detection of transferred genes that are older than the time required for amelioration. A second problem with these methods occurs when the genomic background of the donor and the recipient have similar compositional features (Soucy et al. 2015). Therefore, we cannot detect HGT events between closely related species. Both of these problems give rise to an underestimation of HGT events and shed little light on barriers to HGT.

Phylogenetic analyses identify horizontally transferred genes by comparing the "gene" tree and the "species" tree. Discordance between the "gene" tree and the "species" tree are inferred to be the result of HGT. In this method, the "species" tree is the reference tree for the comparison, constructed typically from genes considered to be resistant to HGT (e.g., rRNA and core housekeeping genes). One limitation of this method is that the query gene should have homologs in all other species to be able to construct the tree. Moreover, computational complexity (e.g., difficulty in analyzing many whole genome sequences) and artifacts of phylogenetic analyses (e.g., long-branch attraction) could introduce several biases to this method (Zhaxybayeva 2009).

Comparative genomics uses the distribution of protein families on a phylogenetic tree. If a gene is present in all members of a clade, we can conclude that the gene is inherited from the ancestor of that clade. Or, if it is missing in only a few members then its absence is more likely a gene loss event. Conversely, if it is present only in a few closely related members, it may be an HGT event. However, the patchy distribution of a gene family through distantly related species might indicate several HGT events or a combination of gene duplication and differential deletion events. Therefore, likelihood estimations are used to decide among these different evolutionary scenarios given the distribution of genes. One caveat of this method is that the lack of genomic information from only one member of the clade can change the likelihood estimation and thus the inferences made (Kunin 2003).

As can be seen, each of these methods has limitations in detecting HGT. Thus, to correctly identify the horizontally transferred genes, researchers typically use several methods and assign genes as candidates according to a threshold metric of reliability.

1.5 Phases of HGT

The journey of a gene from the genome of a donor to that of recipient during an HGT event consists of three stages, as shown in Figure 2. The boundaries of these phases are blurry since in some cases it is difficult to assign an event to a particular phase.

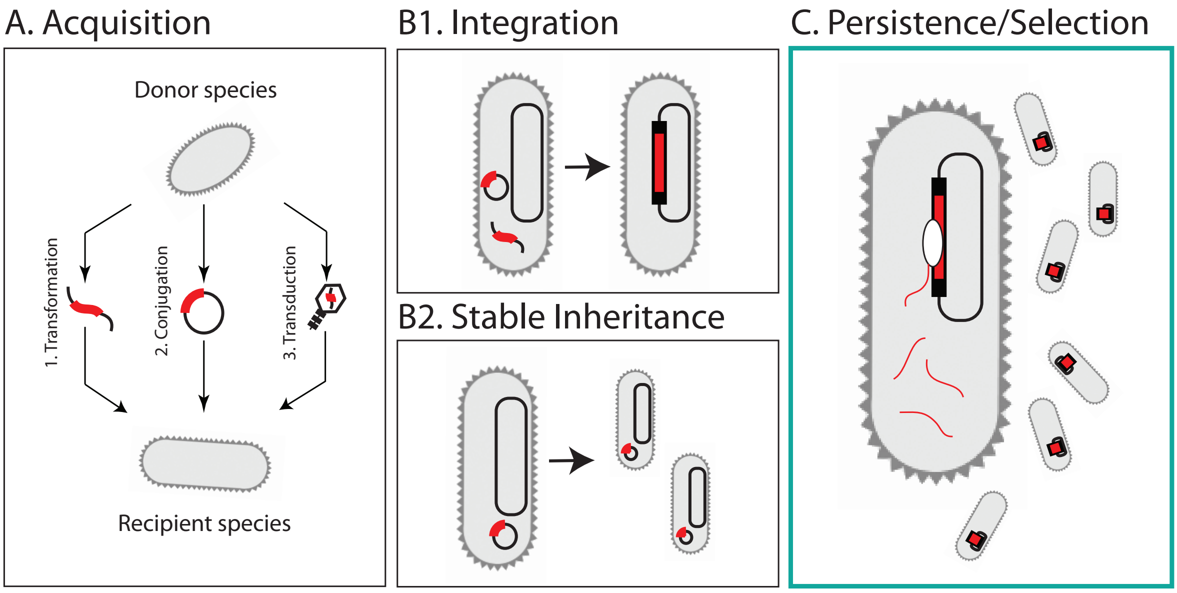

Figure 2. Different phases of HGT

Acquisition, the first phase, covers the period from when the DNA fragment leaves the donor cell until it enters the recipient cell. This acquisition takes place through the mechanisms described above in section 1.3. **Integration**, the second phase consists of the period after the gene enters into the recipient cell, until it successfully integrates itself into the genome, usually by recombination or complementary-strand synthesis.

Alternatively, the acquired gene may avoid the integration if it is carried on an autonomously replicating genetic element, like plasmids. **Persistence**, the final phase, occurs when a newly transferred gene becomes expressed. During this phase, the probability of fixation of the gene within species is exclusively determined by its selective effect and stochastic processes, such as genetic drift. The research reported in this thesis focuses on the persistence phase, since it is the most pertinent phase to the bacterial evolution.

1.6 Barriers to HGT

Compared to the vertical transmission of genes from parent to offspring, genes must overcome a number of barriers to be successfully transferred from one species to another. Each of the aforementioned phases of HGT comes with its own set of barriers.

1.6.1 *Barriers in the Acquisition Phase*

The barriers in the acquisition phase are those that prevent the entrance of the foreign gene(s) into the recipient cell. Any of the transfer mechanisms implies specific barriers to HGT. During transformation, the stability of the extracellular DNA in the environment and the encounter rate of an extracellular DNA fragment with a suitable host cell can limit the rate of HGT (Moscoso & Claverys 2004; Nielsen et al. 2007). Furthermore, some cells are selective in the DNA they transport across the outer membrane by means of sequence-specific selective trans-membrane proteins (Levine et al. 2007). During conjugation, plasmids use mating-pair recognition to choose recipients specifically (Beaber et al. 2002). Additionally, surface exclusion exerted by the already existing plasmid or bacteriophage in the host cell blocks the transfer of an incompatible plasmid or phage (Garcillán-Barcia & la Cruz 2008). Such mechanisms limit the number of possible recipient cells and therefore the range of HGT via these vectors (Thomas & Nielsen 2005). In addition, plasmids are restricted in their host range by their ability to replicate and segregate in the newly acquired host (Zhong et al. 2005). Host specificity also restricts HGT via certain temperate bacteriophages (Koskella & Meaden 2013).

Once inside the cell, the newly acquired gene encounters another set of barriers. The innate restriction modification (RM) systems that are ubiquitous in eubacteria operate as a protection mechanism from phages and plasmids by targeting and degrading foreign DNA (Tock & Dryden 2005). Additionally, a rather sophisticated strategy against phages and plasmids has been recently uncovered called clustered regularly interspaced short palindromic repeats (CRISPRs) (Marraffini & Sontheimer 2008; Bikard et al. 2012). CRISPR-based systems contain a set of spacer sequences originating from past unsuccessful phage or plasmid infections. Small RNAs produced from these spacers are therefore complementary to these invading mobile genetic elements (MGEs) and they guide the CRISPR-associated (Cas) proteins to recognize and degrade them.

1.6.2 *Barriers in the Integration or Stable Inheritance Phase*

The acquired gene that survives the previous barriers normally lingers in the host cytoplasm transiently. In order to establish itself, it has to be integrated into the bacterial host chromosome or exist on a stable autonomous replicon (Thomas & Nielsen 2005). Integration usually takes place through recombination (e.g., homologous recombination, illegitimate recombination, additive integration) and mechanistic details of different types of recombination can potentially be barriers to the integration of the transferred gene. Homologous recombination requires tracks of high sequence similarity. In fact, the recombination rate is inversely proportional to the sequence divergence between two DNA segments (Shen & Huang 1986). Other mechanisms like additive integration or illegitimate recombination reduce the strength of the *homology barrier*, however, with the cost of much lower efficiency (Brigulla & Wackernagel 2010). Moreover, if the gene is inserted in a transposable element or has flanking insertion sequences similar to the ones that are active in the recipient cell, the need for high sequence similarity may be reduced or eliminated (McGrath & Pembroke 2004). However, during integration, if the transferred gene is inserted into a coding sequence, its effect is usually deleterious. Therefore, the rate of HGT through random insertion of DNA fragments by transposable elements is expected to be low (Kurland 2005).

1.6.3 *Barriers in the Persistence Phase*

After an acquired gene is integrated into the host chromosome or an autonomously replicating plasmid its persistence in the population depends on how its expression affects the fitness of the recipient cell. If the gene product is (i) beneficial, it may be retained in the population with a relatively high probability, (ii) effectively neutral, meaning little or no effect on fitness, its persistence will depend on stochastic forces, like genetic drift or genetic draft, and (iii) deleterious, it will ultimately be removed from population by selection (Lawrence 2002; Soucy et al. 2015). Increasing number of bacterial genome sequences and comparative genomics have shown that, despite the high rate of HGT, genome sizes of species stay more or less constant over time, which supports the idea that if a gene fails to contribute to the fitness of the organism it is removed from the population (Mira et al. 2001). Therefore, crucial importance for bacterial evolution is attached to factors that determine the fitness effect of a gene on the recipient cell. Therefore, collectively we call these factors **selective barriers** to HGT and understanding these barriers can enable us to better understand and predict the outcomes of HGT events (González-Candelas 2012).

Over the last two decades, developments in the computational methods in detecting successful HGT events resulted in a number of different types of potential selective barriers. However, we still have a very limited understanding of how these potential barriers act together and which properties of newly transferred genes they are mainly composed of.

Lake's group used computational approaches to investigate HGT and they were first to claim the existence of two main functional categories with statistically different rates of successful HGT: information processing genes – those responsible for DNA replication, transcription, and translation – appear to have been transferred less often than the operational genes – those involved in cellular processes like metabolism and biosynthesis (Rivera et al. 1998; Jain et al. 1999). For the mechanism of this relationship between the function of a gene and likelihood of its transfer they proposed that products of informational genes generally require higher number of protein interactions to function properly, which makes the transfer of a foreign gene more difficult, compared to less connected products of operational genes. They called this the 'complexity hypothesis', which was revisited by Cohen et al. (Cohen et al. 2011) and redefined as connectivity of proteins playing the major role and function has rather an indirect relationship. Since then, this hypothesis had been tested by several studies both bioinformatically and experimentally – as case studies – the results of which, however, failed to support one position or the other (Pál et al. 2005; Wellner et al. 2007; Wellner & Gophna 2008; Omer et al. 2010; Gophna & Ofran 2011). Apart from these, a more detailed examination suggested that the biological functions of horizontally transferred genes, except mobile element genes, are biased to three categories: cell surface, DNA binding and pathogenicity-related functions (Nakamura et al. 2004).

Related to the connectivity of the gene products, the transferred gene should also be functional within recipient organism's existing gene regulatory networks (Kuo & Ochman 2009). This demonstrates the importance of the genomic background of the host cell on the transferred gene's persistence. Therefore, if there are groups of genes encoding for related functions (e.g., lactose metabolism), they will have a higher chance to persist after transfer if they are transferred together as operons that can be fully functional on arrival (Pál et al. 2005).

In addition, sequence characteristics of genes, such as GC content and codon usage, have been proposed as selective barriers as they are shown to affect gene expression. Specific mechanisms are identified as being related to this class of selective barriers. For instance, H-NS has been shown to repress the expression by binding to the AT rich motifs on the coding or regulatory region of foreign genes (Lucchini et al. 2006; Navarre 2016). Differences between codon usage of the foreign genes and tRNA pool of the recipient cells may result in compromised foreign gene expression, toxic protein configurations (Drummond & Wilke 2009), ribosomal sequestration (Shah et al. 2013; Roller et al. 2016), and a general metabolic cost (Shachrai et al. 2010; Tuller et al. 2011; Baltrus 2013).

Finally, gene dosage has been shown to affect fitness through the additional expression of the newly acquired copy which creates an imbalance in the stoichiometry of protein levels in the cell (Papp et al. 2003). Supporting the potential detrimental effect of dosage, Sorek et al. (2007) identified the class of universally single-copy genes such that an increase in dosage of these genes results in toxicity. In addition, highly expressed genes appear to exhibit a lower rate of successful HGT, while low expression may make transfer more permissive (Park & Zhang 2012).

1.6.4 *Role of the Environment*

The role of the environment on the fitness cost of a gene is poorly understood and mostly inferred from case studies of single gene(s). The best examples of this kind are studies on antibiotic resistance genes or mutations, which usually confer a fitness cost in the absence of antibiotic in the environment, but enable survival in the presence of the antibiotic (Melnyk et al. 2015; Roux et al. 2015). Similarly, metabolic enzymes utilizing specific carbon sources may be beneficial only in the presence of the carbon source (Eames & Kortemme 2012).

Only systematic studies investigating the role of environment on the distribution of fitness effects (DFEs) of some set of mutations are done either on a set of chemically induced point mutations (Kishony & Leibler 2003) or on a library of random transposon mutations (Remold & Lenski 2001), where mutations were not identified in both cases.

More broadly, the increase in the competency of several bacterial species upon entrance into the stationary phase or abrupt starvation implies that change in environment might boost the rate of HGT (Seitz & Blokesch, 2013). Similarly, higher rates of HGT in the spermosphere compared to the rhizosphere suggests that environmental heterogeneity could provide greater opportunities for transferred genes (Sengeløv et al. 2001).

1.7 Distribution of Fitness Effects

Distributions of fitness effects (DFEs) are fundamental in evolutionary biology as they allow us to draw inferences about the stability of molecular clocks, the average effect of a mutation, and the maintenance of genetic variation (Eyre-Walker & Keightley 2007).

Frequency of different categories of mutations (lethal, deleterious, neutral, or beneficial) helps us to predict the potential rate of adaptation and complexity of adaptive traits. Although DFE has been the subject of much research, to date, such a distribution for fitness effects of horizontally transferred genes is not available. Therefore, obtaining the DFEs of transferred genes with a systematic experimental study can shed light on the effect of HGT on microbial evolution.

The only study that investigated the DFEs of horizontally transferred genes obtained a DFE for about 100 random DNA fragments integrated into *Salmonella* chromosome. The inserted fragments contained none to several coding sequences from *Bacteroides fragilis, Proteus mirabilis*, and a human intestinal phage. The expression levels of the coding sequences were unknown and DFE of these fragments showed that a major fraction of the inserts exhibited only minor fitness effects on the recipient cells (Knöppel et al. 2014).

1.8 Motivation of the Study

There is an extensive literature on HGT, however, a great deal of that work relies on indirect bioinformatic inference. Although bioinformatics provides valuable insight into the extent of HGT through comparative genomics it has several limitations in drawing inferences from detected HGT event. First, they are sensitive to the assumptions of the specific analysis. For example, any inference has to be made only from successfully transferred genes, some of which may have survived by means of stochastic events rather than conferring selective advantage as models usually assume. Second, they are unable to provide information about the selective effects of newly transferred genes nor the environment in which the transfer happened. Third, their conclusions are sensitive to the amount of data available at the time of analyses. This latter point may explain why conflicting conclusions are common between comparative genomics studies. Another big contribution to the HGT literature was generated by experimental case studies where transfer of a single gene is studied. This situation results in divergent conclusions since findings are usually specific to that gene.

What we lack is a systematic experimental analysis to disentangle how and which of the different selective barriers interact to determine the outcome of an HGT event. To this end, we transferred 44 arbitrarily selected *Salmonella typhimurium* orthologs into an *Escherichia coli* host. By performing competition assays against the wild type we estimated the fitness effects of these transferred genes with constant expression during exponential growth.

In chapter 2, we performed one-to-one competition assays between mutant strain carrying the transferred gene and the wild type strain. By using flow cytometry we estimated selection coefficients for the transferred genes with a precision level of 10^{-3}, and obtained the DFE of horizontally transferred genes. We then investigated if these fitness effects could be predicted by any of the intrinsic properties of the genes, namely, functional category, degree of complexity (protein-protein interactions), GC content, codon usage and length. Finally, by exerting same procedure with the endogenous orthologs of these 44 genes, we identified the role of dosage in determining the fitness effects of transferred genes.

In chapter 3, using the same set of genes we investigated the role of environment on the success of HGT events. Under six different environments with different levels of stress we performed more complex competition assays, where we mixed all 44 mutant strains carrying transferred genes with the wild type strain. To estimate the fitness effects of genes relative to wild type we used next generation sequencing. As such, we addressed the question of whether the likelihood of HGT is primarily determined by some intrinsic genetic properties of the introduced genes, or if it is opportunistic, i.e., determined largely by gene-by-environment interactions.

2 The Role of Protein-Protein Interactions, Functional Categories, and Gene Dosage as Selective Barriers to HGT

2.1 Abstract

Horizontal gene transfer (HGT), the lateral acquisition of genes across existing species boundaries, facilitates bacterial adaptation and the origin of novel phenotypes. Selective barriers determine the probability of a successful HGT event. However, our understanding of how and which of the potential selective barriers interact to determine the outcome of an HGT event remains limited. Here we developed a systematic experimental approach to estimate the fitness effects of transferred genes with a precision of 10^{-3}. By analyzing a set of genes, we obtained distribution of fitness effects of newly transferred genes, and found that most of gene transfers exhibit a significant fitness cost on the host. We identified functional category and length of the genes as potential selective barriers, and gene dosage as the most prominent selective barrier to HGT. However, contrary to general expectations, the level of protein-protein interactions was not a good predictor of the fitness effects of transferred genes. In our work we have begun to build a systematic experimental understanding of the role of different selective factors in horizontal gene transfer.

2.2 Introduction

Horizontal gene transfer (HGT), the lateral transfer of genetic material between different species, is a major evolutionary force shaping microbial genomes (Koonin et al. 2001; Doolittle 1999b; Ochman et al. 2000). The horizontal gene pool facilitates adaptation to new environments as well as evolution of antibiotic resistance (Popa & Dagan 2011; Shapiro et al. 2012; Polz et al. 2013). As such, understanding the mechanisms and constraints that determine the outcomes of HGT events can help tackle the growing resistance problem, while also shedding light on the evolutionary origins of bacterial genomes.

Of particular importance is to understand the factors that impact the success of an HGT event once a gene has been transferred. Collectively we call these factors as selective barriers to HGT, as they may adversely affect the fitness of the host cell. For example, if a newly transferred protein adopts a toxic fold in the new host it will have detrimental effects. Selective barriers result in three distinct outcomes. If a gene is deleterious, it will ultimately be lost from the population. If it is effectively neutral, its survival will be determined by genetic drift. And if it confers a selective advantage to the cell, it will have the potential to get fixed in the population (Soucy et al. 2015). However, to date, we have little knowledge of the distribution of fitness effects (DFE) of newly transferred genes, and hence little is known about the shape and scale of this distribution.

In a previous study, Knöppel *et al.* (2014) obtained a DFE for about 100 random DNA fragments integrated into *Salmonella* chromosome showing that a major fraction of the inserts exhibited only minor fitness effects on the recipient cells. The inserted fragments contained none to several coding sequences derived from different donors. In addition, the expression levels of the coding sequences were not determined. The authors sought to identify casual factors behind the observed fitness costs of the inserted fragments, but failed to identify any sequence characteristics as strong predictors of their fitness effects.

It still remains unclear which, if any, properties of newly transferred genes may act as selective barriers adversely affecting the recipient cell fitness following HGT. To date, bioinformatics approaches have suggested a number of potential selective barriers to HGT. First, gene function has been implicated in restricting HGT: information processing genes - those responsible for DNA replication, transcription, and translation – appear to have been transferred less often than the operational genes (Rivera et al. 1998; Jain et al. 1999; Nakamura et al. 2004). Second, while not universally accepted, it has been suggested that the connectivity of a gene might act as a potential barrier, such that genes with high number of interaction partners (e.g., protein-protein or regulatory interactions) tend to have lower likelihood of successfully experiencing a HGT event (Wellner et al. 2007; Cohen et al. 2011; Wellner & Gophna 2008; Omer et al. 2010; Gophna & Ofran 2011). Third, differences in GC content and codon usage might affect the rate of HGT by adversely affecting translation, giving rise to toxic protein configurations (Drummond & Wilke 2009), ribosomal sequestration (Shah et al. 2013; Roller et al. 2016), being more likely to be targeted by anti-HGT systems (e.g., Cas/CRISPR, H-NS, etc.) (Lucchini et al. 2006; Labrie et al. 2010; Navarre 2016), and through the sequestration of cellular machinery diverting it away from performing critical housekeeping processes (Tuller et al. 2011; Baltrus 2013). Lastly, gene dosage has been proposed as a barrier, as the increase in the relative protein level arising from the presence of a second orthologous copy may influence fitness by causing an imbalance in the stoichiometry in the cell (Papp et al. 2003).

In spite of substantial bioinformatics efforts to identify horizontally transferred genes and selective barriers, a systematic experimental approach to elucidate the roles of different selective barriers in defining the fate of a transfer event has largely been absent. In this study, we employ an experimental framework to systematically disentangle and estimate the importance of different selective barriers to HGT. We transferred the coding sequences of 44 arbitrarily selected *Salmonella typhimurium* orthologs into an *Escherichia coli* host, and expressed them at a constant level. We performed competition assays to estimate the fitness effects of these transferred genes at a precision level of 10^{-3} and tested which, if any, of the selective barriers contributed to the observed fitness effects. *S. typhimurium* and *E. coli* are genetically and ecologically similar (Winfield & Groisman 2003; Mugnai et al. 2015). As we are interested in the role of protein-protein interactions and functional categories this similarity ensures the majority of gene products transferred from *S. typhimurium* are both functional in the *E. coli* genetic background, and that their functional partners exist and may be expressed. In addition, the two species are sufficiently divergent as to allow us to systematically test the effects of several intrinsic factors of the introduced genes. More specifically, we sought answers to the following general questions: What is the DFE of newly transferred genes? What are the effects of different functional gene categories on fitness? Are proteins with a high degree of complexity — many protein-protein interactions — more or less likely to be transferred? And more generally, what are the sources of deleterious fitness effects?

2.3 Materials and Methods

2.3.1 *Chromosomal Modifications in Host Strain*

To differentiate two cell types with flow cytometry, we inserted the fluorescent markers *cfp* and *venus-yfp* into the phage p21 attachment site on the *Escherichia coli* MG1655 (DSM18039) chromosome by using the plasmid pAH95 from CRIM system, where *pstS** gene and its promoter were replaced by each fluorescent protein, respectively, and the constitutive Lambda phage right promoter, $P_{\lambda R}$ (Haldimann & Wanner 2001). Briefly, *E. coli* cells containing the helper plasmid pAH121 were electroporated with the pAH95 plasmid. Following electroporation, cells were suspended in SOC without ampicillin, incubated at 37°C for 1 h and at 42°C for 30 min, and then spread onto LB agar plates with kanamycin 10 μg/mL and incubated over night at 37°C. Colonies were streaked to purify once non-selectively and then tested for antibiotic resistance for stable integration and loss of the helper plasmid and by PCR for single integration of the fluorescent marker cassette. Sequences of the insertions were verified by double-stranded Sanger sequencing. The *cfp* gene was derived from the *E. coli* MC4100 chromosome (Elowitz et al. 2002), and *venus-yfp* gene was derived from the plasmid pZS123 (Cox et al. 2010).

We integrated the repressor protein gene *tetR* that controls the expression of transferred genes under control of the constitutive promoter P_{N25} into the lambda phage attachment site on the *E. coli* MG1655 att-p21::(CFP/Venus-Kn^R) chromosome by using the plasmids and protocol given in the (Lutz & Bujard 1997) (see Figure 3). The original pZS4Int plasmid contained the unneeded *lacI* gene, so first we removed it and its promoter from the plasmid prior to integration. Modified pZS4Int plasmid contained *tetR* gene under P_{N25} promoter, spectinomycin resistance gene, and origin of replication pSC101. Integration of this plasmid was carried out as described in (Lutz & Bujard 1997). Briefly, the origin of replication was cut out of the plasmid pZS4Int and ligated back. *E. coli* cells, containing the thermo-sensitive helper plasmid pLDR8 encoding the lambda integrase (Diederich et al. 1992), were then electroporated with this ligated

DNA. Cells were incubated first at 42°C for 2 hours and then at 37°C overnight on agar plates supplemented with spectinomycin 50 µg/mL to select resistant clones. Colonies were streaked to purify once non-selectively and then tested for antibiotic resistance for stable integration and loss of the helper plasmid and by PCR for single integration of the *tetR* cassette. Sequences of the insertions were verified by double-stranded Sanger sequencing. Plasmids, and plasmid sequences, generated in this study will be deposited with Addgene (www.addgene.org).

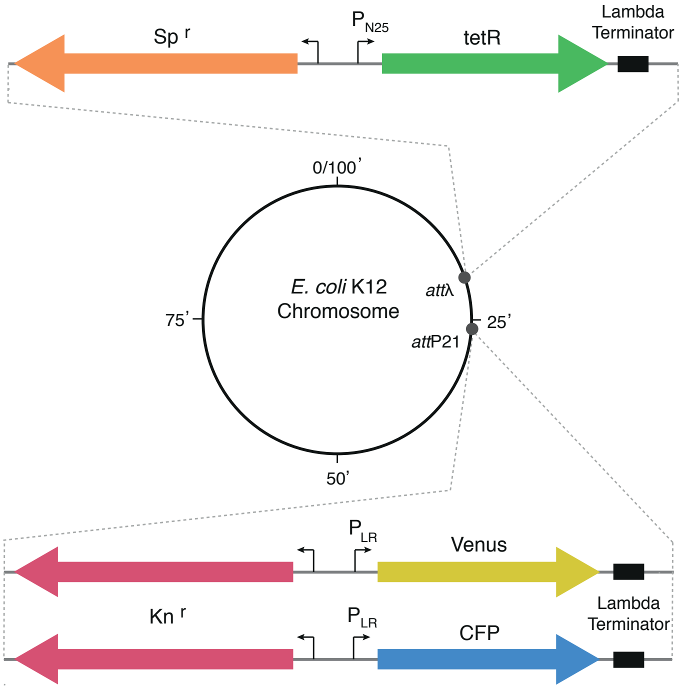

Figure 3. Schematic representation of the E. coli MG1655 att-λ::(tetR-Sp^R) att-p21::(CFP/Venus-Kn^R). Recipient strain for the transferred genes used in the competition assays.

2.3.2 *Selection of Genes*

Salmonella enterica serovar Typhimurium LT2 (DSM18522, Genbank AE006468.1, McClelland et al. 2001) was used as the gene donor. We excluded genes that are parasite related such as phage proteins, transposable elements or insertion sequences, as well as ribosomal and transfer RNAs. We selected 44 genes arbitrarily. As a random selection of genes would be biased towards large functional modules, and we expected specific functions of the genes to have an effect on fitness, we ensured that the sampled genes were from different functional modules (Hu et al. 2009). In addition, we ensured that the sampling included the widest possible range of protein-protein interactions (PPIs) — we sampled uniformly from a range of 1 to 40 physical PPI that were reported by Hu et al. (2009). Lastly, we selected genes only if their interactions had been previously experimentally validated with LCMS (liquid chromatography tandem mass spectrometry) and MALDI (matrix-assisted laser desorption/ionization mass spectrometry) after SPA (sequential peptide affinity) tagging of proteins in that same study.

One factor we wished to address was the role of gene dosage acting as a selective barrier. Dosage can be experimentally addressed in a couple of ways. First, the level of expression of the introduced gene can be modulated. Alternatively, we can express the endogenous *E. coli* copy at the same levels as the introduced *Salmonella* copy. While both have advantages and disadvantages we chose to introduce these 44 *E. coli* genes, using our system and methods (see below), as any observed fitness cost can unequivocally only arise to a dosage imbalance — as the sequences are identical to the endogenous copy it cannot be due to inadequate interactions with other proteins, protein function, and intrinsic aspects of the sequence (e.g., GC content). An extended table with all the relevant information is attached in Appendix 1.

2.3.3 *Cloning of Selected Genes*

Genes selected above were introduced into the recipient *E. coli* cells by transformation of a modified version of the pZS* class of plasmids (Lutz & Bujard 1997) (Figure 4). This plasmid is maintained in the host at approximately 3-4 copies allowing us to reduce the amount of stochastic variation in expression of the introduced gene due to the variation in copy number, and thus to reduce the variance in the fitness measurements. The coding regions of the selected 44 *S. typhimurium* genes (or the 44 endogenous *E. coli* orthologs) were cloned into the pZS*-HGT plasmids under the control of the hybrid promoter $P_{LtetO-1}$ (Lutz & Bujard 1997). Each gene was cloned at the *Avr*II site at 5'-end to ensure that start codon was located at the exact position relative to promoter and ribosomal binding site. For the 3'-end we used either *Hind*III or *Pst*I sites based on the gene sequences, followed by a T1 terminator.

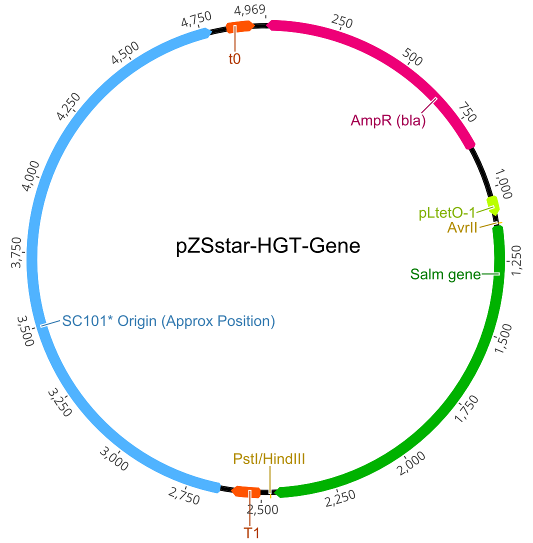

Figure 4. Diagram of the expression plasmid used in the competition assays.

The plasmids were then transferred into *E. coli* MG1655 att-λ::(tetR-Sp^R) att-p21::(CFP/Venus-Kn^R) cells by electroporation and successful transformants were selected by plating cells on LB agar plates supplemented with ampicillin 50 μg/mL. After two rounds of streak purification on 'rich M9 medium' (1x M9 salts, 1% CAA, 0.4% glucose, 2mM $MgSO_4$, 0.1mM $CaCl_2$) agar plates supplemented with ampicillin 50 μg/mL, single colonies were grown overnight in liquid rich M9 medium supplemented with ampicillin 50 μg/mL and stored at -80'C with 15% glycerol. All the cloned genes had both DNA strands sequenced to verify no mutations were introduced in the process.

2.3.4 *Competition Assays*

We performed competition assays using *E. coli* MG1655 att-λ::(tetR-Sp^R) att-p21::(CFP/Venus-Kn^R) strains, CFP strain carrying the plasmid pZS*-HGT with the transferred gene (referred to as the 'mutant' in this study) while the Venus strain carrying the same plasmid without an insert (referred to as the 'wild type' in this study). In total 32 replicate competitions were performed across 4 different days for each gene.

All competition assays were done in 'rich M9 medium' (1x M9 salts, 1% CAA, 0.4% glucose, 2mM $MgSO_4$, 0.1mM $CaCl_2$) supplemented with ampicillin 50 µg/mL. On the first day frozen stocks were streaked on rich M9 agar plates. On the second day a colony was picked and grown in rich M9 medium for 16 hours. And on the third day, overnight cultures were diluted 1000x and grown initially for 60 minutes, followed by the addition of 5ng/mL anhydrotetracycline (ATc, Sigma-Aldrich, Cat no. 37919) to initiate the induction of inserted genes, and then grown for another 60 minutes. After that, the two cell types (wild type and mutant) were mixed at equal ratios and competed with each other for 120 minutes (~3 generations) in 96 well plates (Figure 5). An initial sample (t_0) was taken at the beginning of the competition and three more samples (t_1, t_2, t_3) were taken after each generation (doubling time of the wild type was estimated as ~40 mins under our growth conditions). The ratio of the mutant to wild type was then determined by counting 50,000 cells at each sampling point using the high throughput sampler option of BD FACSCanto II flow cytometer (Figure 5). Using these four ratios, the fitness costs of selected genes (*s*) were estimated by using the regression model $\ln(1+s) = (\ln R_t - \ln R_0)/t$, where R is the ratio of mutant to wild type and t is the number of generations (Elena et al. 1998).

By conducting competition assays during the deterministic exponential phase of growth, and by using time-series data from flow cytometry, we were able to detect very small differences in selection coefficients of the transferred genes efficiently ($\Delta s \approx 0.002$). This estimation of precision comes from a power analysis (Figure 6), for which we used the variance that came from preliminary assays to account for experimental error in our measurements.

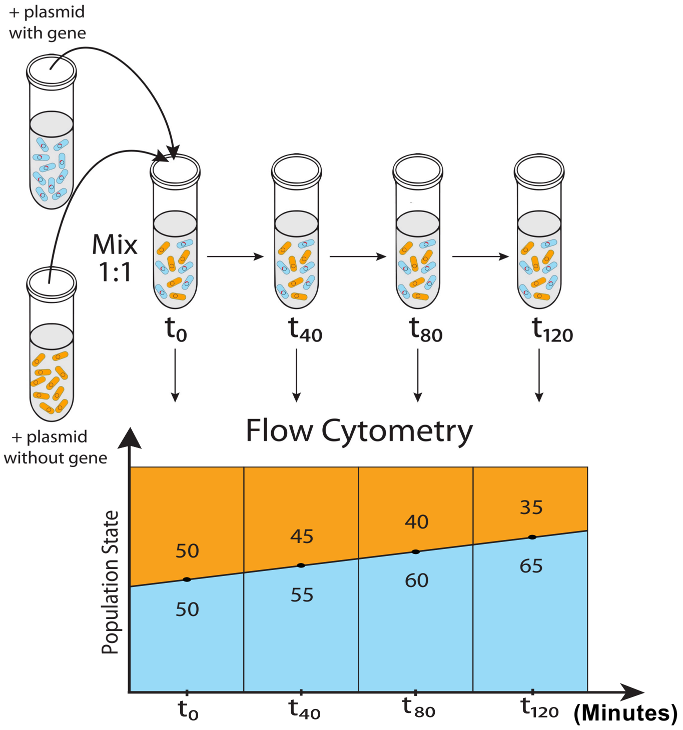

Figure 5. A schematic representation of the competition assay. Blue cells depict the 'mutant' strain that carries the pZS-HGT plasmid containing the introduced gene, whereas orange cells depict the 'wild type' strain that carries the empty pZS*-HGT plasmid. Two strains were mixed at equal frequencies and grown together for 120 minutes, during which samples were taken at 40 minute intervals and the frequency of the two strains were measured by flow cytometry. The plot illustrates an example where the fitness effect of the gene is beneficial, resulting in an increase in the frequency of blue cells over time. Numbers inside the segments represents the frequency of the type of the cell with same color.*

As we wished to control for any fitness differences of the two 'wild type' strains, i.e., strains carrying *cfp* vs *venus-yfp* (both with empty pZS*-HGT plasmids), that might be the result of introducing two different fluorescent markers we compared their fitness using the protocol described above. We detected a small but significant difference between the fitness of CFP strain and Venus strain ($s_{CFP>Venus}$ = 0.004, SD = 0.010, t(314)= 7.118, p<.001). Therefore, we accounted for this difference in the estimation of selection coefficients of introduced genes during the competition assays. We did that by running a set of competitions between these two 'wild type' cells during every competition assay in parallel as a control. Since we did the competition assays in the deterministic phase of the growth under pure haploid selection, this difference in the fitness costs of different fluorescent markers is a constant that we subtracted from the estimated selection coefficient of transferred gene. Such that, each estimation of selection coefficient was corrected for with the fitness difference of the two 'wild types' in the control wells of corresponding experiments.

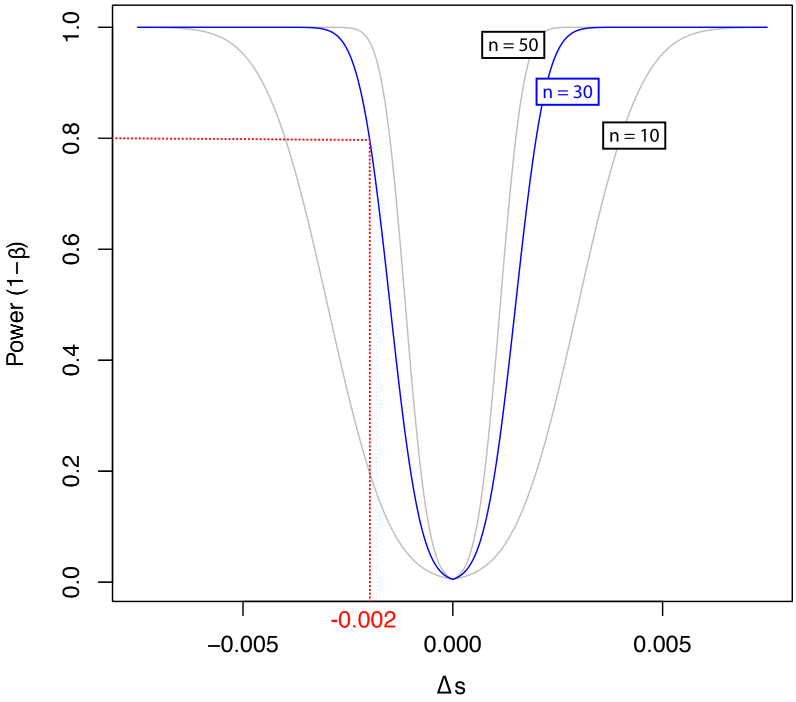

Figure 6. Power analysis performed to estimate the sensitivity of our selection coefficients (s) measurements during competition assays. Power is calculated for α=0.01 and sd=0.003 based on preliminary data.

To determine whether the introduced genes might show different fitness effects on the different fluorescent backgrounds (CFP strain and Venus strain) we did a reciprocal introduction by cloning a subset of 8 randomly selected genes out of our 44 *Salmonella* genes into the Venus strain and repeated the competition assays. The regression between the selection coefficients of genes (mean of 32 replicates) in CFP strain and Venus strain was highly significant ($F_{1,6}$=117, p<.001, r^2 =0.943, slope = 1.007), indicating different fluorescent backgrounds do not interact with this subset of genes, and all measured effects are solely due to the introduced genes.

2.3.5 *RNA-seq: Sample Preparation*

To estimate the expression level of transferred genes, we cloned *mCherry* gene into an empty pZS*-HGT plasmid as an expression control and used RNA-seq to measure relative expression. Cultures grown overnight in rich M9 medium were diluted 1000x and handled under the same conditions as the competition assays described above. When the OD of the cultures reached to ~0.12, growth was stopped by adding Qiagen

RNA protect Bacteria Reagent (cat no. 76506) to 20mL cultures (~6x10^8 cells). Total RNA was purified with Qiagen RNeasy Mini Kit (cat no. 74104). Quality and integrity of the total RNA samples were checked in Agilent 2100 Bioanalyzer and Agilent RNA 6000 Nano Kit (reorder number 5067-1511). Library preparation (RiboZero, NEB), further quality checks and next-generation sequencing (HiSeq2500-v4, SR100 mode) were performed at the VBCF NGS Unit (www.vbcf.ac.at). The data will be deposited in Dryad Digital Depository.

2.3.6 *RNA-seq: Data Processing*

Sequence reads with an average read quality of >= 34 were retained for further analysis. After quality controls, fastq files were aligned to the *E. coli* MG1655 genome (Genbank U00096.3) using the Bowtie2 aligner using RSEM's (Langmead & Salzberg 2012) default settings. The reference genome was modified *in silico* to contain the chromosomal modifications of *tetR* and fluorescent protein gene cassettes. Expected counts were calculated by using the defaults in RSEM (B. Li & Dewey 2011). After between-sample normalization of the counts with DESeq package of the R statistical software (Anders & Huber 2010), TPM (transcript per million) values for each gene were calculated and used in further analyses (B. Li & Dewey 2011). The RNA-seq pipeline is available upon request.

Expression level of whole transcriptome under experimental conditions, together with that of selected genes and induction level of transferred genes during the competition assays (~3300 TPM, or 0.33% of the transcriptome) are given in Figure 7.

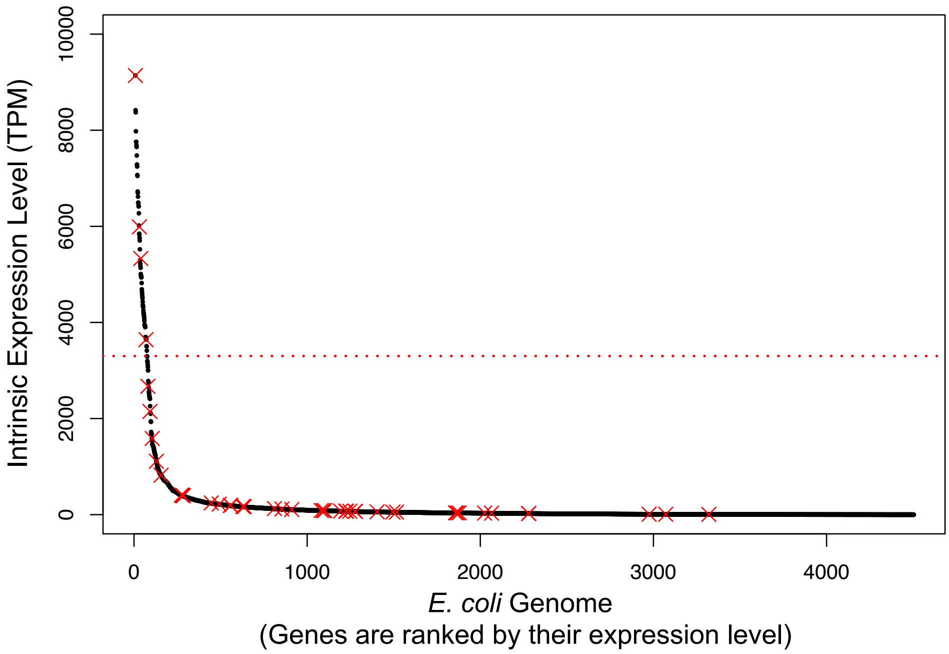

Figure 7. Distribution of expression levels for the E. coli transcriptome, as measured with RNA-seq under experimental conditions (TPM: Transcript per Million). Red crosses show where the native expression of the selected 44 genes under the experimental conditions fall on that distribution. Dotted line shows the induction level of introduced genes during competition assays, ~3300 TPM. Seven very highly expressed genes (with expression levels of 10743, 11650, 13027, 20007, 34683, 46188, and 96807) are excluded from the plot for the clarity of the figure.

In addition, we used RNA-seq results to correct the PPI level of our genes' partners if they were not expressed under our experimental conditions. After obtaining the expression levels for the whole transcriptome, we decided for a threshold level of expression below which a gene would have been eliminated from further analyses. To this end, we inspected the expression levels of genes that are known as being repressed under our experimental conditions, i.e., lactose operon, arabinose regulon, and flagellar regulon genes. Expression level of these genes ranged from 0.5 – 50 TPM in our RNA-seq data. To estimate the effect of different thresholds we corrected the PPI levels of our genes by eliminating genes below 10, 25, 50, 75, and 100 TPM expressions, and repeated all the analyses. The choice of threshold had only a negligible effect, therefore we conducted our analyses with PPI corrections using the threshold of 50 TPM.

2.3.7 *Statistical Analysis*

To determine if the genes were neutral or not, one-tailed one-sample t-tests were done for the 32 replicates of each gene, with $\mu_0>0$ or $\mu_0<0$. $\alpha=0.05$ was used as the significance level after false discovery rate (FDR) corrections for multiple testing (Benjamini & Hochberg 1995).

After dividing genes into two according to their functional categories, two-sided Wilcoxon rank sum test (Mann-Whitney U test) was used to decide if the fitness effects of the two categories were different from each other. Analysis was done on the mean selection coefficients of genes for the 32 replicate measurements.

We investigated a number of intrinsic genetic properties — GC content, codon usage, and gene length. GC content was calculated as the absolute deviation between the introduced *Salmonella* gene and the *E. coli* ortholog. Codon usage was calculated as the absolute deviation of the frequency of optimal (FOP) usage in the introduced *Salmonella* sequence using the *E. coli* FOP. Gene length was quantified as the number of base pairs from the start to stop codon of the *Salmonella* gene (i.e., cds). To investigate the effect of these intrinsic factors we employed multiple linear regression. After investigating interactions and more complicated models, we used the following model: *Salmonella* selection coefficients ~ Protein - Protein Interaction levels + Functional Category (as dummy variable) + Deviation in GC% between orthologs + Deviation in codon usage between orthologs + Gene length in bp + Expression level of genes in TPM unit. The analysis was done on the mean selection coefficients of genes for the 32 replicate measurements.

Additional simple linear regressions were performed as follows:

- *Salmonella* selection coefficients ~ *E. coli* selection coefficients
- *E. coli* selection coefficients ~ expression level of genes in TPM unit

Finally, Fisher's exact test was performed to investigate the relationship between selective effect of transferred genes and the increase in their expression level in the cell relative to their native expression level. Highly deleterious fitness effects (selection coefficients lower than -0.1) were observed in 9 of the 31 genes (29 percent) which had more than 10-fold change in their expression level. Whereas none of the 13 genes with less than 10-fold change in their expression level showed such high fitness cost. Analysis was done on the mean selection coefficients of genes for the 32 replicate measurements.

All the statistical analyses were done using the R software package (version 3.1.1) and RStudio (Version 0.98.1062).

2.4 Results

2.4.1 *Distribution of Fitness Effects*

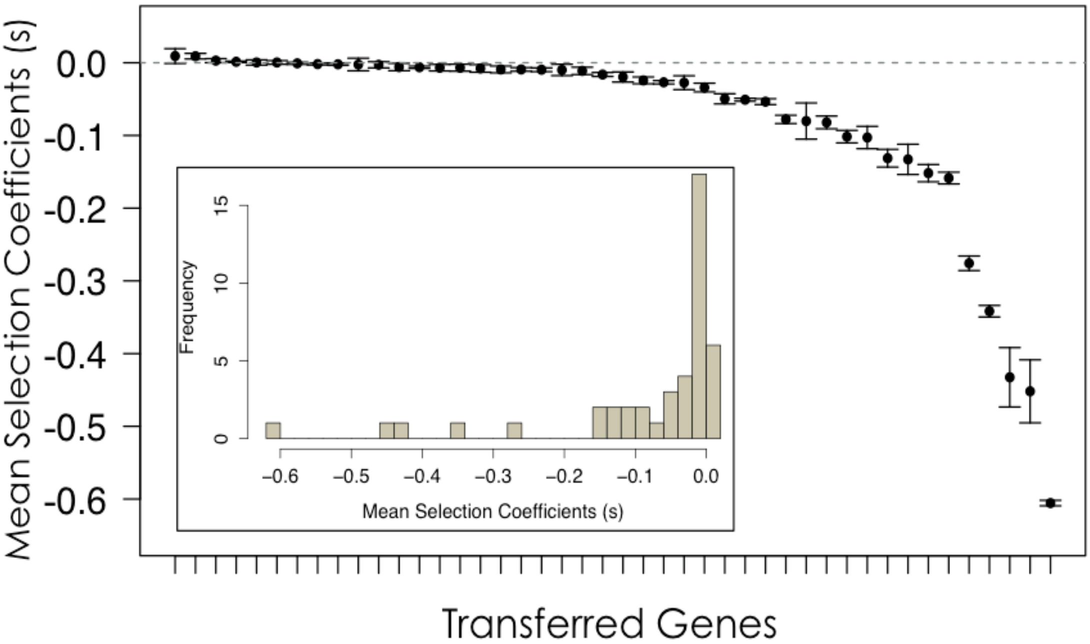

Figure 8. DFE of newly transferred genes. On the x-axis genes are sorted according to their selection coefficients. Error bars of the data are the 95% CI of the selection coefficients for the 32 replicate measurements of each gene. Embedded plot gives the classical histogram representation.

The distribution of fitness effects (DFEs) of new mutations plays a critical role in determining evolutionary outcomes (Eyre-Walker & Keightley 2007). DFEs for different types of mutations - random transposon insertions (Elena et al. 1998) and point mutations on coding sequences (Sanjuán et al. 2004), promoters (Kinney et al. 2010), and transcription factors (Shultzaberger et al. 2012) - have been experimentally determined. While DFEs might differ between species and genomic regions, they exhibit some general features: beneficial mutations are rare, and the effect of deleterious mutations can usually be well described by a log-normal distribution, often with an additional peak for the lethal mutations (Eyre-Walker & Keightley 2007). Here we confirm a similar distribution for 44 *S. typhimurium* ortholog genes in the *E. coli* host (Figure 8, embedded plot). The mean fitness over all mutants was -0.080 with a standard deviation of 0.137.

Table 1. Selection coefficients of S. typhimurium and E. coli orthologs.

Gene Name	STM Gene ID	STM Sel.Coef.	p-value	ECO Gene ID	ECO Sel.Coef.
lpxD	STM0226	0.00921	**<.001**	b0179	0.00959
ybhK	STM0801	0.00896	**<.001**	b0780	-0.03490
hfq	STM4361	0.00285	**.002**	b4172	0.00336
infC	STM1334	0.00125	.074	b1718	0.00266
pnp	STM3282	0.00037	.719	b3164	-0.03917
rlmL	STM1061	0.00035	.326	b0948	-0.00506
ydiI	STM1366	-0.00082	.102	b1686	-0.00670
sapF	STM1696	-0.00181	**.003**	b1290	-0.01514
yacL	STM0160	-0.00206	**.017**	b0119	-0.00860
exbB	STM3159	-0.00244	.077	b3006	-0.00302
hybG	STM3143	-0.00283	**.004**	b2990	-0.01440
rplI	STM4394	-0.00591	**<.001**	b4203	-0.00038
moaE	STM0806	-0.00624	**<.001**	b0785	-0.02008
cbpA	STM1112	-0.00678	**<.001**	b1000	-0.02457
uspG	STM0614	-0.00681	**<.001**	b0607	-0.00716
rimI	STM4558	-0.00776	**<.001**	b4373	-0.00356
ridA	STM4458	-0.00917	**<.001**	b4243	-0.00369
cspE	STM0629	-0.00923	**<.001**	b0623	0.00589
dps	STM0831	-0.00963	**<.001**	b0812	-0.00648
ibpB	STM3808	-0.00992	**<.001**	b3686	-0.00045
glyQ	STM3656	-0.01117	**<.001**	b3560	-0.41062
yibL	STM3689	-0.01623	**<.001**	b3602	-0.01602
dnaQ	STM0264	-0.01959	**<.001**	b0215	-0.04294
cspD	STM0943	-0.02427	**<.001**	b0880	-0.13406
iscS	STM2543	-0.02693	**<.001**	b2530	-0.06198
hupA	STM4170	-0.02744	**<.001**	b4000	-0.04459
kdpD	STM0703	-0.03411	**<.001**	b0695	-0.02397
clpA	STM0945	-0.04958	**<.001**	b0882	-0.02243
yqjI	STM3215	-0.05091	**<.001**	b3071	-0.01484
pstB	STM3854	-0.05355	**<.001**	b3725	-0.04937
acpP	STM1196	-0.07789	**<.001**	b1094	-0.06697
selB	STM3682	-0.08022	**<.001**	b3590	-0.10939
hupB	STM0451	-0.08211	**<.001**	b0440	-0.03703
lexA	STM4237	-0.10157	**<.001**	b4043	-0.07425
malP	STM3514	-0.10279	**<.001**	b3417	-0.13412
fadJ	STM2388	-0.13126	**<.001**	b2341	-0.02219
yadG	STM0172	-0.13294	**<.001**	b0127	-0.03153
thiI	STM0425	-0.15182	**<.001**	b0423	-0.21239
rne	STM1185	-0.15861	**<.001**	b1084	-0.22364
leuS	STM0648	-0.27576	**<.001**	b0642	-0.30644
srmB	STM2643	-0.34174	**<.001**	b2576	-0.03822
lolA	STM0961	-0.43262	**<.001**	b0891	-0.01427
topB	STM1298	-0.45193	**<.001**	b1763	-0.59428
uvrC	STM1946	-0.60555	**<.001**	b1913	-0.59745

Genes are sorted according to the selection coefficients of Salmonella orthologs. p-values are for the fitness effects of Salmonella orthologs being different from neutral. Fields indicated with ***bold*** *are significant values, α = 0.05 after corrections for multiple testing with FDR method.*

While most transferred genes were deleterious (37 of 44), none were lethal and only 11 were highly deleterious (s<-0.1), with majority having a relatively small negative effect on fitness (Figure 8 and Table 1). Out of 44 transferred genes, 2 were beneficial, and 5 were neutral – fitness not significantly different from zero.

2.4.2 *Factors Affecting DFE of Acquired Genes*

2.4.2.1 *Functional Gene Category*

One major hypothesis arising from bioinformatics analyses of HGT postulates that informational genes - those responsible from DNA replication, transcription, and translation - are less transferable than the operational genes - those involved in cellular processes like metabolism and biosynthesis (Rivera et al. 1998; Jain et al. 1999; Nakamura et al. 2004). In fact, while experimental work by Sorek et al. (2007) lends support to this for nearly lethal ribosomal genes, it is unclear if this is a general pattern or unique to the ribosomal genes in their collection. To test if functional category has an influence on the fitness effects of the transferred genes, we grouped our genes according to their COG annotations (Tatusov et al. 2000). We considered COG categories 'information - storage - processing' to be informational genes, constituting 18 of 44 genes. We did not find a significant difference in the mean fitness effects between the two groups (Figure 9, Wilcoxon rank sum test, $\mathrm{Mdn}_{\mathrm{Info}}$=-0.025, $\mathrm{Mdn}_{\mathrm{Oper}}$=-0.009, W=182, p=0.220, two-tailed), but even though not statistically significant, we do observe that informational genes show a greater variance in fitness effects compared to operational genes (σ^2_{Info}=0.032, σ^2_{Oper}=0.008, Levene's test, p= 0.062). This result is consistent with results from bioinformatics analyses suggesting informational genes are less likely to be transferred. While the fitness costs did not differ between the two groups, interestingly 4 of 5 nearly lethal genes (s<-0.25) were informational.

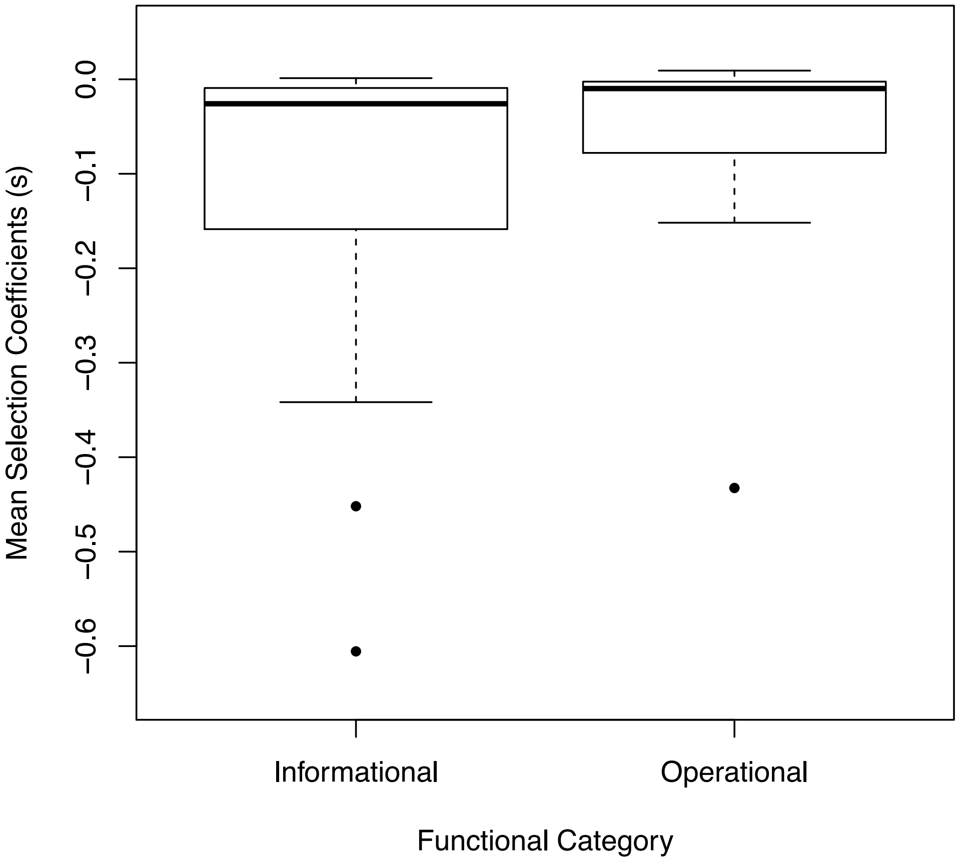

Figure 9. Boxplot representation of the selective effect of the transferred genes divided into two groups as Informational and Operational based on their functional categories. The observed higher variance of the informational genes is consistent with the expectation that informational genes have a higher probability of being highly deleterious.

2.4.2.2 *Number of Protein Interactions*

The second major barrier proposed by bioinformatics studies states that increasing number of protein-protein interactions (PPI) decreases the likelihood of HGT (Cohen et al. 2011). This relationship can occur for three reasons. First, a transferred orthologous gene with many potential interaction partners in the host cell may fail to interact with any of these partners and be effectively neutral and more prone to stochastic loss in a population (Wellner & Gophna 2008; Omer et al. 2010). Second, it may interact improperly with its partners interfering their functions and be selected against in a population. Third, it may interact properly with novel partners, however, decrease host fitness by disrupting the existing cellular stoichiometry (Papp et al. 2003). We experimentally tested this hypothesis by selecting 44 genes uniformly over a range of 1 to 40 physical PPI levels, which were obtained from Hu et al. (2009). We did not find support for PPI as a selective barrier to HGT, as the fitness costs of transferred genes could not be explained by their PPI levels (p=0.231). To make sure that all potentially interacting partners were expressed by the host cell under our experimental conditions we did RNA-seq and corrected the number of partners for each gene by eliminating the partners that were not expressed. Irrespectively, the PPI level did not explain the observed fitness effects (p=0.277, Figure 10).

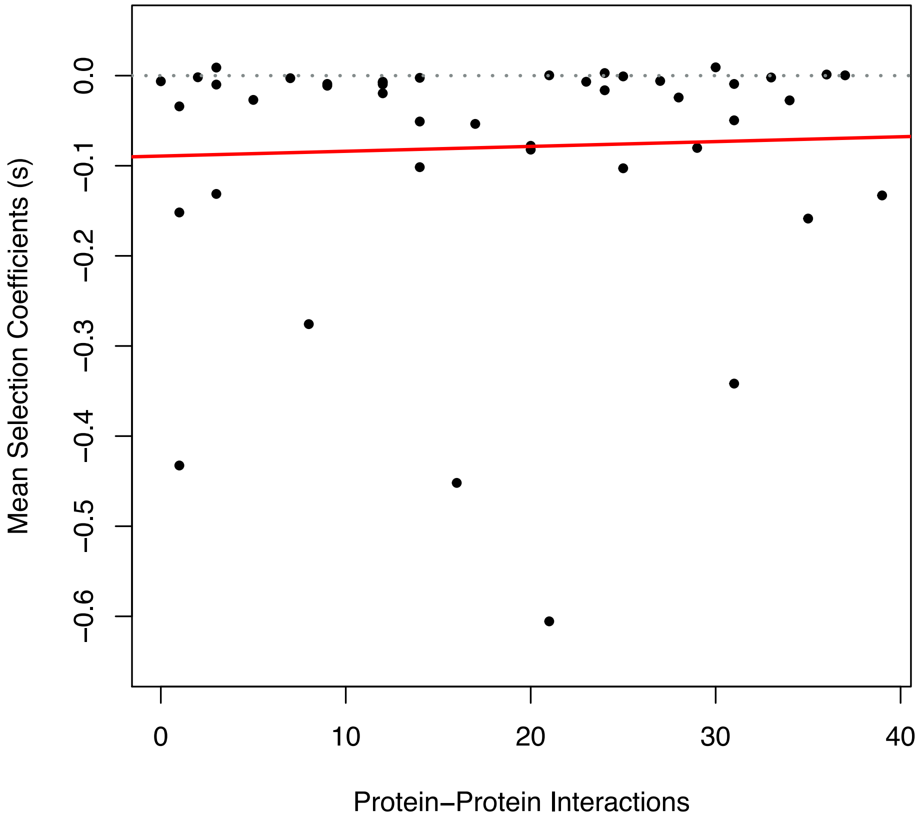

Figure 10. Selection coefficients of newly transferred genes plotted against the number of protein-protein interactions they have corrected by RNA-seq. Red line is the regression between the two variables, plotted only as representative. The relationship is not significant with p=0.277 (comes from the complex model of multiple regression analyses, see Materials and Methods section 2.3.7). Gray dashed line shows the zero line.

2.4.2.3 *GC Content and Codon Usage Bias*

Sequence specific signatures, such as GC content and codon usage bias, vary among species as well as among genes from the same genome; and mutations that lead to discrepancies can affect bacterial growth rate (Sharp et al. 2010; Bonomo & Gill 2005; Raghavan et al. 2012). First, one of the major protein components of the nucleoid structure in bacteria and a global repressor, the histone-like nucleoid-structuring protein (H-NS), may down-regulate gene expression by binding AT-rich regions of DNA. This can cause differential expression of AT-rich genes upon transfer to a new host resulting in inactivation (Navarre 2016). Secondly, differences in codon usage of the newly transferred gene and the tRNA pool of the recipient host may affect fitness, forming a selective barrier to HGT. This cost may arise from ribosomal sequestration resulting from stalled ribosomes during translation of HGT regions with different codon usage than the recipient host (Gingold & Pilpel 2011), as well as from an increase in the translational mutation rate resulting in toxic misfolded proteins (Drummond & Wilke 2009; Tuller et al. 2011).

Although GC content and codon usage bias are correlated, we examined their effects separately, while also accounting for the possible interactions between them. We examined whether the deviation in GC content between the transferred *S. typhimurium* copy and its *E. coli* orthologs correlated with the observed fitness effects and didn't find a significant interaction (Figure 11-a, p=0.268). Similarly, the absolute deviation in the frequency of optimal codon usage (FOP) between the two orthologs, which ranges from 0 to 12%, was not a significant predictor of the observed fitness effects (Figure 11-b, p=0.203). While in our dataset the differences in codon usage and GC content between *S. typhimurium* and *E. coli* orthologs are relatively modest, which may prevent us from detecting a small but evolutionarily significant effect as selective barriers to HGT, it is clear that among closely related species these intrinsic properties are not strong selective barriers.

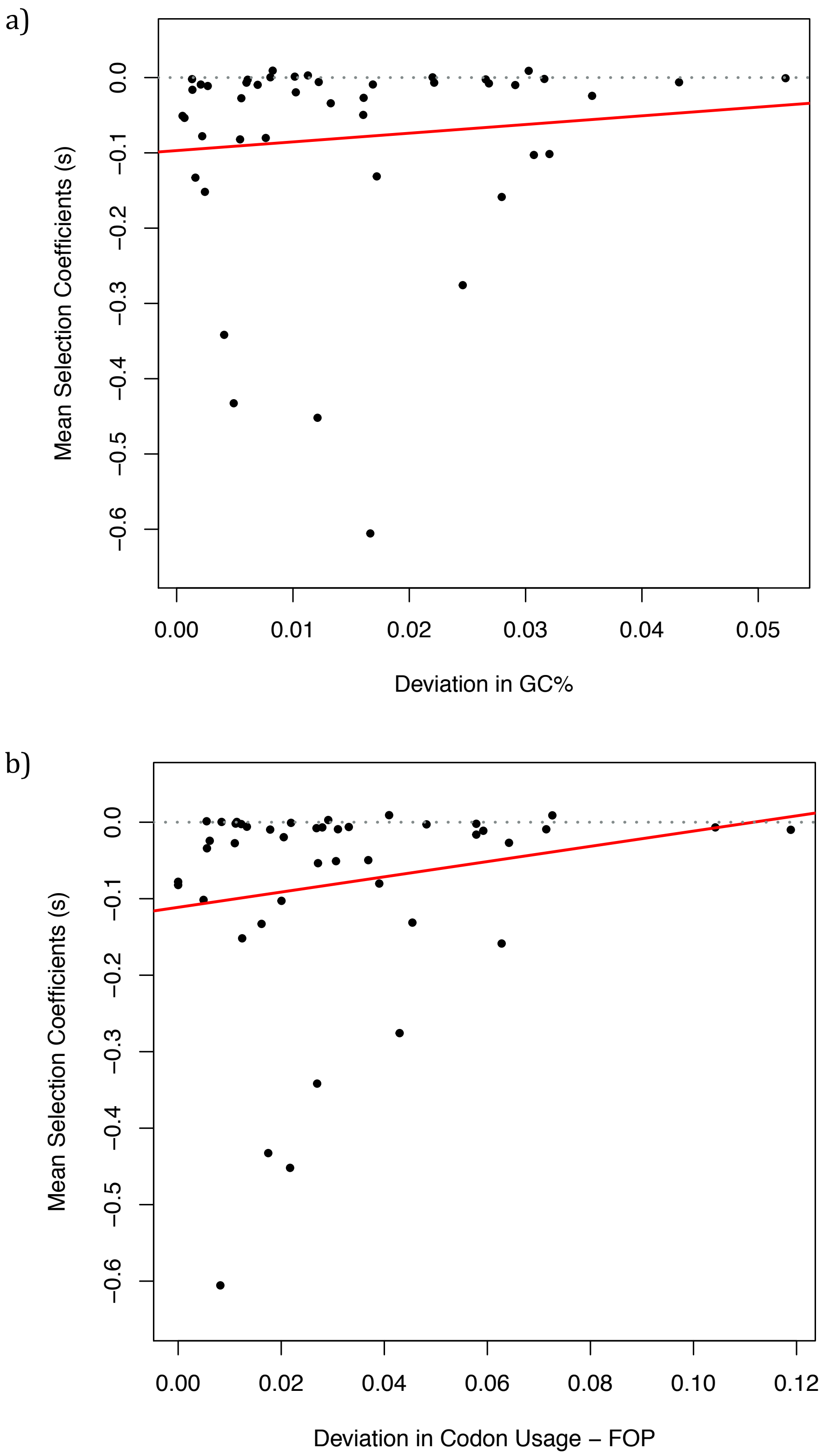

Figure 11. Selection coefficients of newly transferred genes plotted against, a) deviation in GC% between orthologs, b) deviation in FOP codon usage between orthologs. Red line is the regression between the two variables, plotted only as representative. The relationships are not significant with p=0.268 and p=0.203, respectively (comes from the complex model of multiple regression analyses, see Materials and Methods section 2.3.7). Gray dashed line shows the zero line.

2.4.2.4 *Gene Length*

Interestingly, we observed a statistically significant negative relationship between gene length and the fitness effect of transferred genes in our dataset (p=0.037, Figure 12). The relationship between fitness cost and gene length can arise due to expenses at the genomic, transcriptional, and translational levels. The first two have been shown to be less relevant than the cost of protein synthesis, which has been estimated to be very small as well (Baltrus 2013; Lynch & Marinov 2015). Since they remain under the limits of our detection, it is unlikely that they explain the negative effect of gene length. The cost of protein synthesis, however, can be exacerbated by ribosomal sequestration, since long genes will acquire more ribosomes preventing them to perform other critical processes and giving rise to a reduction in growth, and thus our observed lower fitness in longer genes (Shah et al. 2013; Roller et al. 2016).

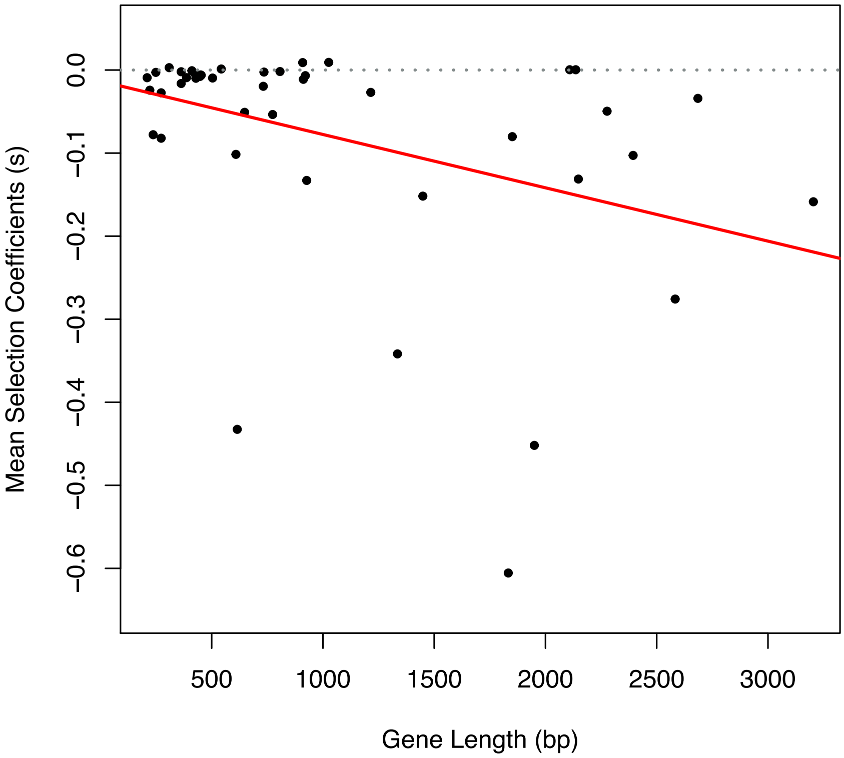

Figure 12. Selection coefficients of newly transferred genes plotted against gene length in bp. Red line is the regression between the two variables, plotted only as representative. The relationship is significant with p= 0.037 (comes from the complex model of multiple regression analyses, see Materials and Methods section 2.3.7). Gray dashed line shows the zero line.

2.4.2.5 *Gene Dosage*

Since *S. typhimurium* and *E. coli* are genetically similar to each other (90% median homology of orthologous CDS at amino acid level, McClelland et al. 2001), observed fitness effects may arise from an imbalance caused by an increase in the protein concentration within cells, i.e., from a dosage effect. Dosage has been suggested as a factor responsible from fitness decrease as a result of horizontally transferred gene expression (Papp et al. 2003; Sorek et al. 2007; Park & Zhang 2012; Bershtein et al. 2015). To elucidate the role of dosage as a selective barrier to HGT, we carried out same experiments with the native *E. coli* orthologs of our 44 selected genes. Since in this case the transferred gene is identical to the existing host copy, this scenario is equivalent to a gene duplication event and as such the only potential change in fitness arises from an increased concentration of the protein. We observed a significant correlation between the fitness effects of transferred *S. typhimurium* genes and the duplicated *E. coli* genes (Figure 13, p < .001).

By comparing the fitness effects of each ortholog pair, we can estimate when dosage is the primary factor driving fitness costs of horizontally transferred genes. We observed that *S. typhimurium* copy was significantly more deleterious than the *E. coli* copy for 18 genes, both orthologs had the same fitness cost for 6 ortholog pairs, and the *E. coli* copy was more deleterious than *S. typhimurium* copy for 20 genes (Figure 14). While the differences in the coding sequence between *S. typhimurium* and *E. coli* orthologs give rise to differences in their fitness effects, when those fitness effects are same or higher in the *E. coli* orthologs it implies that dosage is the most dominant selective barrier for those genes. Consequently, the fitness effects of 26 genes out of 44, or about 60% of the transferred *Salmonella* genes, can be explained solely by the dosage effect and not other factors, as observed. However, we did not observe an enrichment for the functional categories of informational or operational genes within the genes showing dosage effect (Fisher exact test, p=0.535, with 11 informational genes out of 26 genes showing dosage effect).

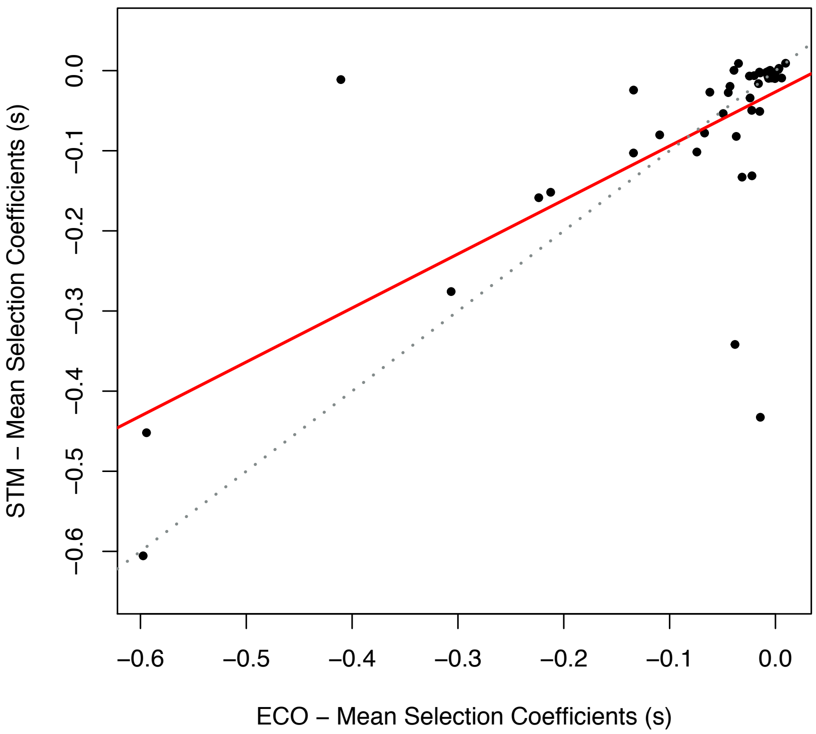

Figure 13. Selection coefficients of newly transferred genes from Salmonella plotted against those of their orthologs from E. coli. Red line is the regression between the two variables with p<.001. Gray dashed line corresponds to the y=x line.

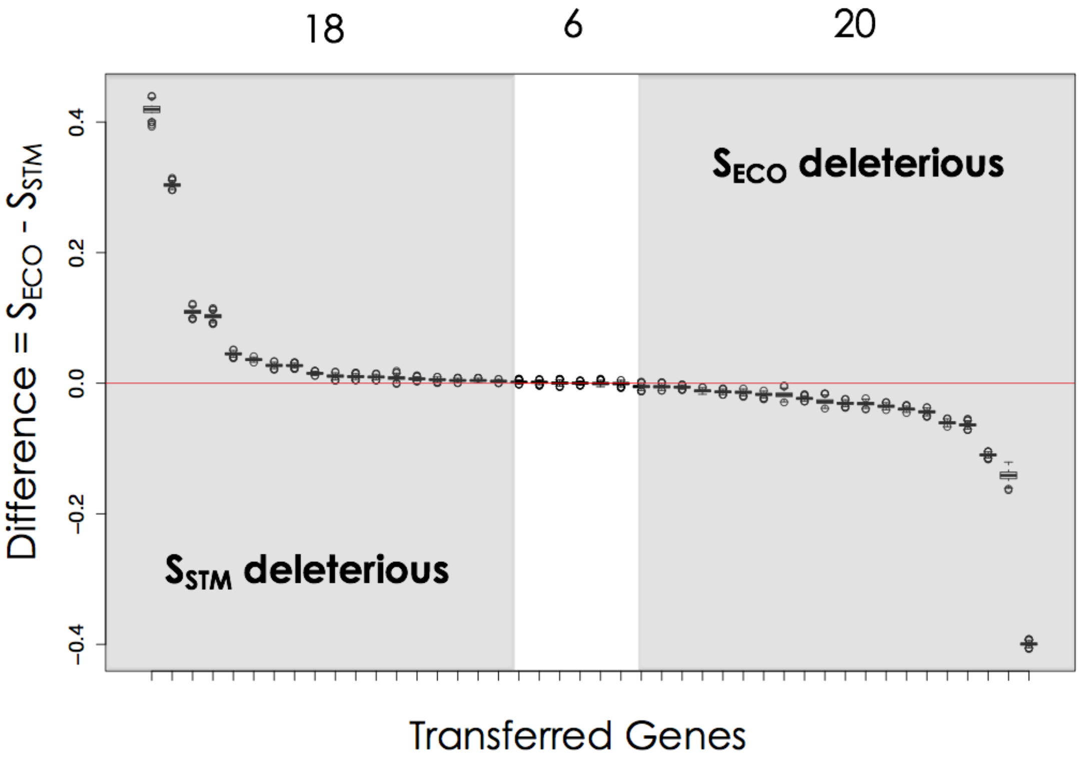

Figure 14. Difference between the selection coefficients of Salmonella and E. coli orthologs. On the x-axis, genes are sorted by the difference in their selection coefficients between orthologous pairs. Data points that fall in the gray shaded areas are for the gene pairs with fitness effects significantly different from each other.

We, then, explored if the dosage effect arises from the fold increase in intrinsic expression levels resulting from the addition of the second copy. While all experimentally transferred genes were induced at the same constant level, the fold-increase in their expression level depended on the intrinsic expression level of each gene in the cell (Fold Increase = Expression level of the transferred gene from the plasmid / Intrinsic expression of the endogenous copy from the chromosome). We used RNA-seq data to estimate the intrinsic expression level of each gene used in our study. Because fitness effects of genes transferred from *S. typhimurium* could result from either dosage or other factors, we focus only on the fitness effects of genes transferred from *E. coli*. Intrinsic expression levels of studied genes did not significantly correlate with their observed fitness effects (p=0.353). Interestingly, we observed high fitness costs (s<-0.1) only for those genes for which the additional copy resulted in at least a 10 fold-increase in expression level (9 out of 31 compared to none out of 13, mean selection coefficients -0.0195 vs -0.1032, p=0.028, Fisher's exact test, Figure 15).

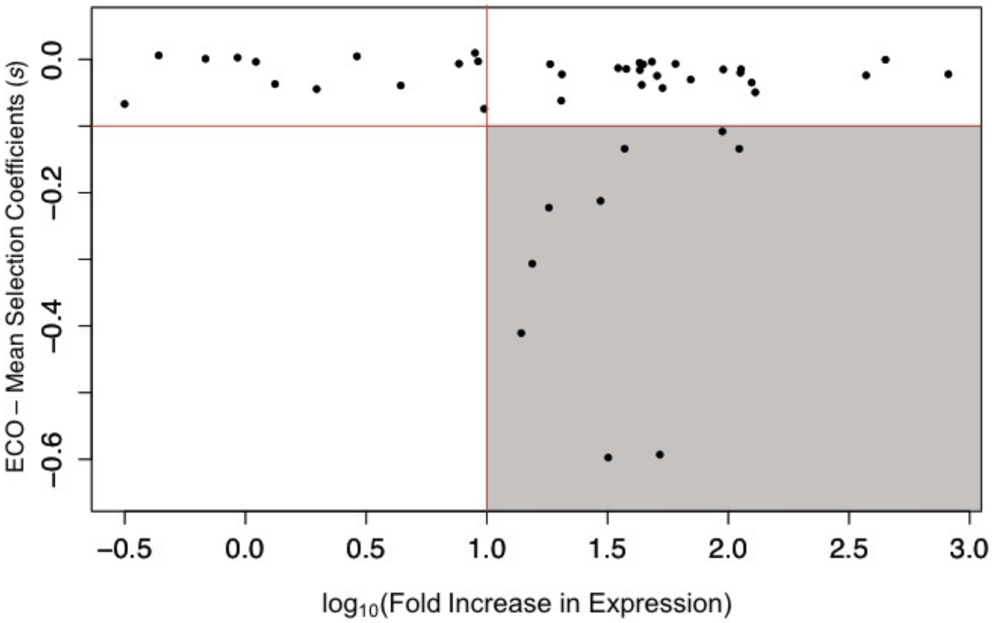

Figure 15. Selection coefficients of transferred E. coli orthologs are plotted against the fold change in their expression levels resulting from the induction of the expression plasmid during the competition assays. Shaded area is to emphasize the observation of highly deleterious genes only if the increase in their expression is larger than 10 fold-change.

2.5 Discussion

In this study, we explored the relative importance of several factors that may affect the probability of an HGT event, by transferring 44 orthologous genes from *S. typhimurium* into *E. coli*. Unlike the mostly neutral effects of transferred DNA fragments of random size, for which the expression state was not known (Knöppel et al. 2014), we find that, when expressed, most gene transfers impose a significant fitness costs on the host (37 out of 44 genes with mean selection coefficient of -0.080). This finding suggests that, if expressed, a large fraction of transferred genes are likely to be quickly eliminated by selection. Such an effect explains why gene silencing of transferred genes is common in microbes (Navarre et al. 2007; Navarre 2016). Yet, despite the tendency for most transfer events to result in negative fitness costs, HGT is thought to be one of the major sources of novel genetic material in microbes (Boto 2010; Soucy et al. 2015). The disparity between this observation and the DFE we report could be reconciled if the bacterial effective population sizes in nature are smaller than believed (Kimura 1968), due to being highly structured, experiencing recurrent bottlenecks, recurrent selective sweeps, or simply having carrying capacities that do not support large numbers. If these conditions hold, then deleterious transferred genes may persist long enough to be rescued by deactivating mutations or beneficial compensatory mutations. In addition, genes may enter the cell without being expressed, therefore hiding from selection allowing them to segregate neutrally in the population.

We identified a strong effect of dosage, which had previously found limited support in bioinformatics studies as a major barrier to HGT, in part because it is difficult to infer dosage from genomic data. However, the role of dosage as a barrier to HGT can be inferred from the observation that changing expression levels of proteins can dramatically reduce fitness (Papp et al. 2003). Furthermore, although host fitness was not estimated, dosage was identified as the main barrier along with toxicity in a data set of transferred genes that had lethal effects on their new host (Sorek et al. 2007). Here, we show that dosage effects play an important role as selective barrier of not only lethal, but all horizontally transferred genes.

In this study, we report a number of factors that affect the probability of horizontal gene transfer — gene dosage, gene length, and whether a gene is involved in information processing. Interestingly, we did not observe a significant effect of the factors of interaction level, GC content, and codon bias, which have been predicted by bioinformatics studies to be important selective barriers to HGT. We selected each gene from different functional modules, as we expect the specific function of the gene to have a large effect. It is possible that the identity of the gene itself is so prominent that it overshadows all other factors. In fact, we observe that the most deleterious genes – a DNA topoisomerase, an RNA helicase, an aminoacyl tRNA synthetase, and a DNA repair gene - are involved in essential cell functions. Moreover, transfers of different genes that are part of the same protein complex have been reported both as neutral and deleterious (Papp et al. 2003), indicating that genes have differential robustness to perturbations arising from HGT. In spite of the potential impact of unintended bias, the fact that we don't capture the effect of these factors at a sample size of 44 and with measurement accuracy of 10^{-3} suggests that the tested factors can act as only weak selective barriers to HGT, or are at least smaller than what we can practically measure.

The understanding of HGT relies primarily on comparative analyses of bacterial genomes, which can only study successful transfer events that have gone through the sieve of natural selection. Experimental approaches, which have been rare due to their labor-intensive nature and technological limitations, offer the potential to test the role of a single factor while controlling most other relevant parameters. At the same time, the limitations of experimental approaches, especially the relatively small sample sizes, prevent easy generalizations. As such, in the future these two approaches will continue to complement each other. While bioinformatics approaches can provide analyses of many genes across a large number of species, experimental approaches can be used to identify new selective barriers and disentangle their relative importance. In our work we have begun to build a systematic experimental understanding of the role of different selective factors in horizontal gene transfer.

2.6 Acknowledgements

We thank Marta Dravecká and Tobias Bergmiller for sharing the strain or the plasmid carrying the fluorescence protein genes CFP and Venus, respectively. We also thank Katharina Pöcher for her guidance during the RNA-seq analysis; and Mato Lagator for valuable discussions and inputs.

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: Hande Acar. Performed the experiments: Hande Acar. Wrote the paper: Hande Acar. Performed statistical analysis: Hande Acar. Performed RNA-seq analysis: Hande Acar.

3 The Role of the Environment in Horizontal Gene Transfer

3.1 *Abstract*

The rate of adaptation and complexity of adaptive traits under particular conditions can be predicted by the distributions of fitness effects (DFEs) of the transferred genes. Although we anticipate a substantial environmental dependency of fitness effects of genes from well documented studies of antibiotic resistance genes and metabolic enzymes, we lack a systematic study of the environment-dependence of DFEs. Here we addressed this question by measuring the fitness effects of newly transferred genes in six different environments with different types of cellular stress. We found that the DFEs of horizontally transferred genes are highly dependent on the environment, with abundant gene–by-environment interactions. Furthermore, we demonstrated a relationship between average fitness effect of a gene across all environments and its environmental variance, and thus its predictability. Finally, in spite of the fitness effects of genes being highly environment-dependent, we still observe a common shape of DFEs across all tested environments. In general, our study demonstrates the need for more realistic experiments that include fluctuating environments, heterogeneous environments, and spatially structured environments, such as the animal gut or soil.

3.2 *Introduction*

Horizontal gene transfer (HGT) is the transmission of genetic material between existing species without parental relatedness, and it is a major source of microbial genetic variation, on which natural selection can act (Doolittle 1999a; Ochman et al. 2000; Koonin et al. 2001). Following the insertion and expression of a newly transferred gene, the success of an HGT event will depend on the fitness effect it has on the recipient (host) cell. Deleterious genes are likely to be eliminated from the population, whereas the probability of fixation of effectively neutral and beneficial genes is determined by the interplay between genetic drift and selection (Soucy et al. 2015). As such, predicting the fate of a transferred gene, and more broadly the impact of HGT on the genetic composition of a population, critically depends on the distribution of fitness effects (DFE) of horizontally transferred genes.

While DFEs of transferred random DNA fragments of variable size (Knöppel et al. 2014) and of genes in a more systematically controlled experiment (see Chapter 2) have been previously described, the role of the environment is poorly understood and has been cobbled together from a mere handful of studies. The fact that environment might alter the fitness effect of a horizontally transferred gene seems obvious, however, little attention has been given to it in a systematic experimental manner. Implications that the environment plays a substantial role were derived from: (i) antibiotic resistance studies in which resistance is highly beneficial in the presence of antibiotics, but deleterious in the absence (Melnyk et al. 2015; Roux et al. 2015), (ii) observations that effects of other types of mutations, such as point mutations (Kishony & Leibler 2003) and random transposon insertions (Remold & Lenski 2001), can be environment-dependent, and (iii) competency of bacteria appears to be induced by nutrient scarcity (Seitz & Blokesch, 2013), (iv) higher rates of phage induction in response to pollutants in nature (Cochran et al. 1998). More broadly, the rate of HGT might be enhanced when environments vary as environmental heterogeneity could provide greater opportunities for transferred genes to be retained and then reach fixation (Sengeløv et al. 2001). In spite of these observations and predictions, a systematic study of the environment-dependence of DFEs has remained absent.

We addressed this question by transferring and expressing 44 genes obtained from *Salmonella typhimurium* in *Escherichia coli*, and estimating their relative fitness effects in six different environments using pooled genotype competition experiments and next-generation sequencing to track changes in genotype frequencies. With this experimental design we do not only obtain individual fitness effects of transferred genes, but also can describe the overall DFE and study its dependence on the environment. As such, we can address the question of whether the likelihood of HGT is primarily determined by some intrinsic genetic properties of the introduced genes, or if it is opportunistic, i.e., determined largely by a highly specific gene-by-environment interaction.

3.3 *Material and Methods*

3.3.1 *Strains and Plasmids*

We used *E. coli* K12 MG1655 (DSM18039) strain with the chromosomal insertion of a *tetR* cassette (see Material and Methods section in Chapter 2). This cassette contains the gene for the repressor protein tetR that controls the expression of transferred genes under control of the constitutive promoter P_{N25}, and spectinomycin resistance. The cassette is inserted at the λ-att site of *E. coli* chromosome by using a modified version of pZS4Int1 plasmid and the helper plasmid pLDR8 carrying lambda integrase, as described in Lutz & Bujard (1997).

Genes were cloned under the inducible $P_{LtetO-1}$ promoter into the low copy number (3-4 plasmids/cell) pZS* plasmid backbone of Lutz & Bujard (1997) (also see Material and Methods section in Chapter 2).

3.3.2 *Culture Conditions and Environments*

We chose environments that are representative of some of the conditions that are potentially experienced by *S. typhimurium* and *E. coli* species. All 250 mL media cultures for sequencing were grown in 500 mL flasks at 37°C and 180 rpm (except “NOX” condition) in a water bath. We used rich M9 (1x M9 salts, 1% CAA, 0.4% glucose, 2mM $MgSO_4$, 0.1mM $CaCl_2$) supplemented with ampicillin 50µg/mL, at pH 7 with ATc 5ng/mL as the standard medium “M9”. The wild type has a doubling time (DT) of 40 min in this standard medium. Other tested growth conditions are described as: “CAM” - M9 rich medium supplemented with ampicillin 50µg/mL and chloramphenicol 1.2µg/mL and ATc 5ng/mL, DT_{CAM} = 80 min; “LB” - Lennox broth and ATc 12ng/mL, DT_{LB} = 24 min; “NOX” - M9 rich medium supplemented with ampicillin 50µg/mL and overlaid with paraffin oil to create an anaerobic condition and ATc 5ng/mL, DT_{NOX} = 60 min; “pH5” -

M9 rich medium at pH5 supplemented with ampicillin 50μg/mL and ATc 5ng/mL, DT_{pH5} = 60 min; “TMP”- M9 rich medium supplemented with ampicillin 50μg/mL, trimethoprim 0.3μg/mL and ATc 4ng/mL, DT_{TMP} = 80 min (Table 2).

To determine the concentration of the inducer, ATc, in different growth conditions, we first cloned *gfp* gene under $P_{LtetO-1}$ promoter and measured the fluorescence intensity of the cells with BD FACSCanto II flow cytometer with different concentrations of inducer in each environment. The lowest concentration of ATc that gives a coefficient of variance <40% for the fluorescence signal at the population level was picked at each condition.

3.3.3 *Selected Genes*

The set of genes, plasmid constructs, and bacterial strains used in this study are derived from a previous study (Chapter 2). Briefly, in total 44 genes from *Salmonella enterica serovar* Typhimurium LT2 genome (DSM18522, Genbank AE006468.1, McClelland et al. 2001) were chosen arbitrarily avoiding genes that are parasite related such as phage proteins, transposable elements or insertion sequences, as well as ribosomal and transfer RNAs. During selection a number of precautions were taken to avoid introducing biases (see Chapter 2). All *S. typhimurium* genes were cloned in the low copy number plasmid pZS*. In this study, an additional random fragment of the *tetA* gene (721 bp, the mean length of all inserted genes) was cloned into the pZS* plasmid without a promoter to be used as the ‘wild type’ in competition assays. We checked whether this strain confers a substantial fitness cost by conducting a competition assay using BD FACSCanto II flow cytometer as described before (Chapter 2). We found that the strain is nearly neutral (s = 0.0019, Wilcoxon signed rank test, p = 0.024), although it is significant, our formula to calculate the fitness effects of transferred genes corrects for this difference such that selection coefficients of genes are relative to wild type.

3.3.4 *Competition Assays and Sequencing*

An additional genotype, carrying a phenotypically neutral and unique sequence but otherwise identical to the host cell, is used as the ‘wild type’ during these competition assays. 1:1000 dilutions of separate overnight cultures of 45 *Escherichia coli* clones each carrying a different plasmid were grown in 20 mL M9 rich medium to OD_{600} 0.2 and then mixed at equal volumes. A concentrated stock culture was prepared in 1x M9 salts with 10% DMSO. Cell concentration was verified by CFU counting of the dilutions of frozen stock. Aliquots of this ‘mixed stock’ of 45 different clones was stored at -80°C.

For the competition assays, 250 mL of corresponding media was inoculated with 10^7 cells from mixed stock for each environment tested and grown until OD_{600} 0.4. Plasmid DNA was extracted from 200 mL of this culture using ZR Plasmid Miniprep™ Kit - Classic (Zymo research, www.zymoresearch.com). In total 6 replicate competitions were performed across 3 different days for each environment.

DNA was sheared with S220 AFA™ Focused-ultrasonicator (Covaris®) to obtain a fragmentation size of 200-800 bp. The DNA library of 6 different environments (6 biological replicates each) and the initial stock (2 technical replicates) were prepared and sequenced on Illumina HiSeq 2500 (100 bp SE) by the sequencing company VBCF NGS Unit (www.vbcf.ac.at, Vienna, Austria). In addition, quality checking and de-multiplexing of the raw data was also provided by the VBCF. Two of the libraries failed to produce sufficiently high quality sequences (one replicate from environments “M9” and “TMP”) and thus were eliminated from further analysis.

3.3.5 *Sequencing Data Processing and Calculation of Selection Coefficients*

Sequence reads with an average read quality of >= 34 were retained for further analysis. Sequencing reads were mapped with GMAP (version 2016-05-01) against a personalized reference containing fasta files of 44 *S. typhimurium* genes, *tetA* fragment and backbone of the plasmid, built with gmap_build function. Parameters were set to `--no-chimeras --nosplicing --nofails --npaths=0` to increase mapping accuracy (Wu & Watanabe 2005). Mapped sam files were converted to bam files, sorted and indexed with samtools (version: 1.3.1) (H. Li et al. 2009).

In order to obtain depth per gene, bedtools (version: 2.17.0) was used with the following parameters: `bedtools genomecov -d –ibam` (Quinlan & Hall 2010).

To compare replicates of a treatment, data were normalized to the mean depth of ampicillin resistance gene within treatment. The sequencing analysis pipeline is available upon request.

Fitness costs of selected genes (s) were estimated for each replicate by using the regression model $\ln(1+s) = (\ln R_t - \ln R_0)/t$, where R is the ratio of the frequencies of mutant (depth of gene) to wild type (depth of *tetA* fragment) and t is the number of generations (Elena et al. 1998). Initial frequencies were obtained from the mean depth of the two replicates of the mixed stock, which is used for inoculation of each competition. Time t is end of the competition assay, corresponding to 8.25 generations. According to this formula, fitness effects of genes are calculated relative to that of wild type. Selection coefficients of the transferred genes in our six different environments are given in Appendix 2.

3.3.6 *Comparison of Fitness Measurement Methods*

To test the reliability of the sequencing method in estimating the fitness effects of transferred genes we compared the selection coefficients of our 44 genes estimated here to the ones estimated earlier by our group with a different technique (Chapter 2). In that previous study we performed one-to-one competition assays between ‘wild type’ and the mutant strains and used flow cytometry to obtain the change in frequency of these two strains over time. As we see in Figure 16, two measurements gave very similar results ($p < .001$).

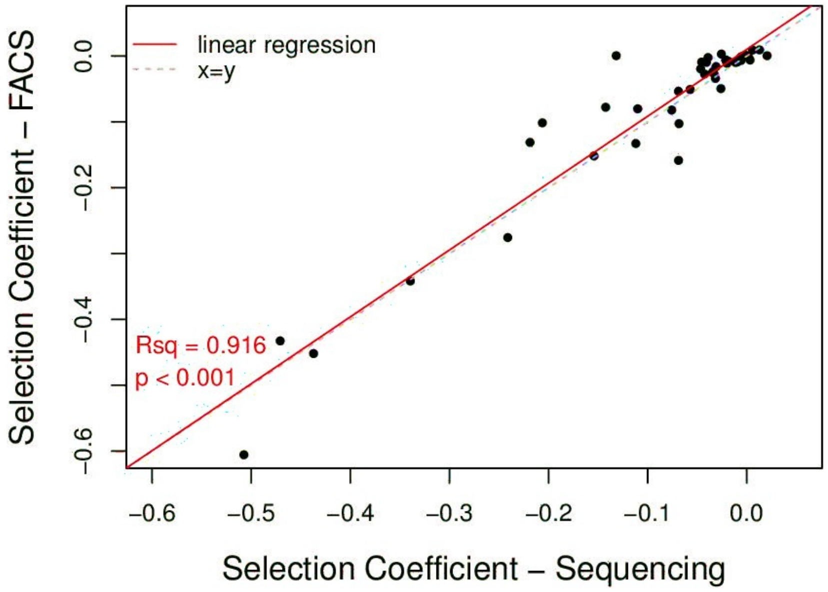

Figure 16. Comparison of the selection coefficients of transferred Salmonella orthologs measured with two different techniques.

3.3.7 *RNA-seq*

To precisely determine the expression level of the inserted genes in each environment and to estimate the expression of interaction partners in these environments a marker gene (*mCherry*) was cloned under the $P_{LtetO-1}$ promoter in the pZS* plasmid. Cells were grown under the same conditions as competition assays, where the starter stock (inoculation of 10^7 cells) only included mCherry containing strain instead of the mixed stock of 45 strains. Growth was stopped by adding Qiagen RNA protect Bacteria Reagent (cat no. 76506) onto 20 mL of cultures at OD_{600} 0.4. Total RNA preparation after this point was performed as described in Chapter 2 under RNA-seq section. Library preparation (RiboZero, NEB), further quality checks and next-generation sequencing (HiSeq2500-v4, SR100 mode) were performed at the VBCF NGS Unit (www.vbcf.ac.at). The data will be deposited in Dryad Digital Depository.

Similarly, RNA-seq data processing was performed together with the samples of the previous study in the same way as described in the Materials and Methods section 2.3.6

of Chapter 2 under RNA-seq data processing. We utilized this information during our investigation of the effect of intrinsic factors differentiating genes from each other into three groups of 'less deleterious genes', 'highly deleterious genes', and 'nearly lethal genes' (see description of ANOVA analysis below).

3.3.8 *Statistical Analysis*

In order to estimate differences between environments a paired and two-sided Wilcoxon signed-rank test was performed with $\alpha = 0.05$ on each pairwise comparison of six environments. In order to test if the shapes of distributions of environments were different, two-sample and two-sided Kolmogorov-Smirnov test was performed. In both of these tests, selection coefficients of the genes in each environment was represented by the mean of 5 or 6 biological replicates, and final p-values were corrected for multiple testing using FDR method (Benjamini & Hochberg 1995).

To investigate the interaction between environment and genes, a two-way analysis of variance (ANOVA) test was performed by using selection coefficients of the genes composed of 5 or 6 replicates for each environment, with the formula: Selection Coefficients ~ Environment * Gene + Error (replicates). We applied this test on each pairwise comparison of six environments, and final p-values were corrected for multiple testing using FDR method (Benjamini & Hochberg 1995).

Furthermore, the mean and standard deviation of the selection coefficients of each gene across all environments were calculated. We investigated the relationship between the mean and standard deviation of the selection coefficients of the transferred genes by performing linear regressions with the formulas: Standard Deviation ~ Mean and Standard Deviation ~ Mean + Mean^2.

Data was split into three groups: (i) 'less deleterious genes', with mean fitness effects across all environments of more than -0.1 and SD<0.05; (ii) 'highly deleterious genes', with SD>0.05; and (iii) 'nearly lethal genes', with a mean fitness effect between environments of less than -0.4 and SD<0.05. We inspected if these three groups could be separated from each other by the means of several intrinsic properties of the transferred genes (Table 6, number of interaction partners (PPI), length of the coding sequence, difference in the GC content between homologs, difference in the codon bias between homologs (FOP), and the change in the level of the expression of the endogenous copy of the gene over all conditions (TPM)). We performed separate one-way ANOVA tests for each of these properties with the formula: Gene property (dependent continuous variable) ~ three groups of genes (independent factorial variable). A statistically significant difference was considered at $p < 0.05$. An additional Fisher's exact test was performed to examine whether 'highly deleterious genes' were enriched for the functional category of the genes as Informational and Operational genes compared to the rest of the genes.

All statistical analyses were performed in the R software package (version 3.1.1) and RStudio (version 0.98.1062).

3.4 Results & Discussion

3.4.1 *Distribution of Fitness Effects*

The distribution of fitness effects is a critical parameter in that it tells us about the average effect of a mutation (in our case a newly transferred gene) and the frequency of different classes (e.g., deleterious, neutral, or beneficial) of mutations. And understanding of these distributions is crucial to our understanding of and ability to predict evolution. However, little attention has been given to the role of the environment in the probability of a successful horizontal gene transfer event, and therefore, in determining the DFE for horizontally transferred genes. To address this shortcoming, we estimated the DFEs of 44 genes transferred from *Salmonella* to *E. coli* in six environments: two standard laboratory growth conditions (M9 and LB media), and four stress conditions that represent ecological conditions commonly experienced by *S. typhimurium* and *E. coli* – chloramphenicol (CAM), trimethoprim (TMP), anaerobic (NOX), and low pH (pH5). To isolate the effects of the transferred genes from the potential global genomic stress of the new environments we competed them against the 'wild type' (carrying a phenotypically neutral DNA sequence). The only difference between the 'wild-type' and mutant types were the horizontally transferred gene. To this end, the relative frequencies of transferred genes before and after competition assays were determined by next generation sequencing, and the selection coefficients (s) were estimated from those frequencies (Figure 17). Each competition assay for an environment was replicated six times.

Each of the tested environment has a quite strong effect on the fitness of the recipient bacteria, shown as the doubling times of the wild type in these environments in Table 2.

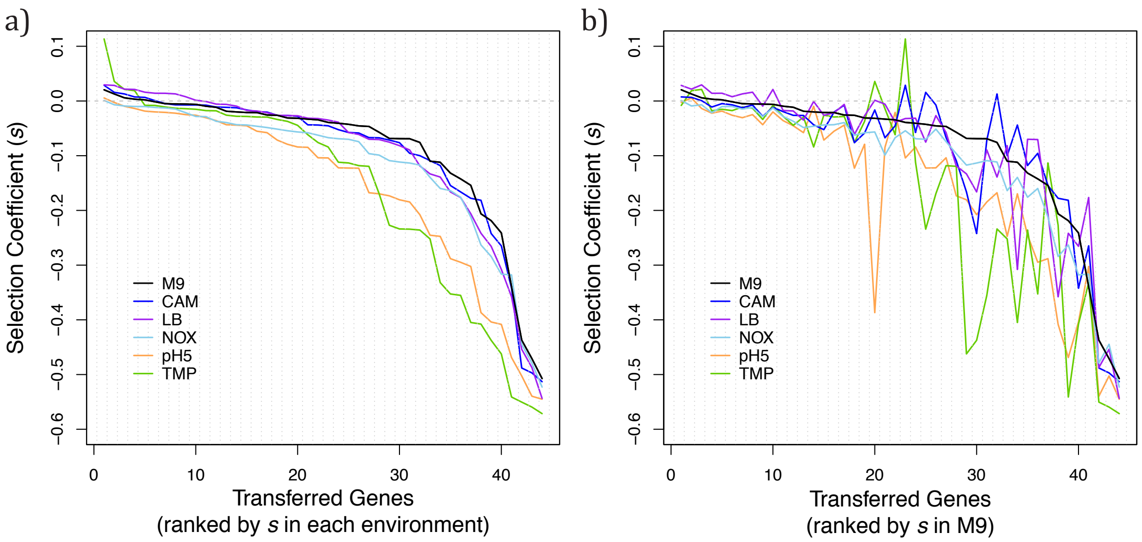

Figure 17. Selection coefficients of the transferred genes in six environments. Lines connect genes measured in the same environment. a) shows the overall shape of the DFEs. Genes are ranked according to their selection coefficients in the environment that they were measured in. b) shows how the fitness effects of individual genes vary between environments. Genes are ranked according to their selection coefficients only in standard medium of M9.

Table 2. Environments and growth rate of the 'wild type' in these environments.

Treatment	**Growth Media**	**Doubling Time [min]**	**DFE Medians**	**DFE Variance**
LB	Lennox broth	24	-0.033	0.021
M9	M9 rich medium pH7	40	-0.037	0.017
NOX	M9 rich medium overlaid with paraffin oil	60	-0.065	0.017
pH5	M9 rich medium pH5	60	-0.104	0.025
CAM	M9 rich medium 1.2 µg/mL Chloramphenicol	80	-0.045	0.019
TMP	M9 rich medium 0.3 µg/mL Trimethoprim	80	-0.080	0.036

3.4.2 *The Role of Environment on the DFEs*

The role of the environment on the DFEs of newly transferred genes can reveal itself in three different ways. First, the environment may not have any effect on the introduced gene. Since our mutants were competed against the 'wild type' under the same environmental conditions, this scenario would result in identical DFEs for each environment. Second, the environment affects the fitness of all mutants equally, resulting in an overall shift in the DFEs of the mutants while preserving their rank order. Finally, the specific environment affects the fitness of specific genes differentially, resulting in a strong gene-by-environment interaction. Under this last scenario, changes in the environment would make the fitness effect of a gene unpredictable. Inspection of Figure 17-b suggests the third scenario so we asked whether the environment altered the central tendency of the DFEs and whether the environment affected the shape and spread (variance) of the DFEs.

3.4.2.1 *Central Tendency*

The environment significantly altered the central tendency of some of the DFEs for the 44 transferred genes (Figure 17-a). Using a Wilcoxon signed rank test we detected significant differences in the median fitness effects between a number of pairs of DFEs (Figure 18, Table 3). However, some DFEs did not significantly differ. In particular, the DFEs for M9, LB, and CAM were not significantly different from each other, and neither were the DFEs for TMP and pH5. The lack of a difference between M9 and LB is not fully surprising as both are relatively rich media. A lack of a difference between the CAM environment and LB/M9 is very interesting and puzzling at the same time, as the selective effect of a gene in CAM and M9 can be dramatically different (see Figure 17-b). If the effect of the environment was simply the result of a common stress on the cell then we expect the medians of the DFEs to be identical as they are all relative to the 'wild-type'. Interestingly, our results suggest that the environment must be interacting with some genes in a way that is not simply additive by amplifying the cost of the introduced genes.

The scenario in which the environment affects the fitness effects of all genes in a similar fashion can be understood by the intuitive explanation that under more stressful conditions cells may be less tolerant to the additional stress of an acquired gene. This would suggest that the central tendency of DFEs to scale with the severity of the environmental stress relative to the standard medium M9 and therefore, we would be able to predict the fitness effect of a gene by knowing solely the growth rate of the recipient under that environment. From the growth rates of the cells in the absence of the genes we know that some environments are more deleterious than others (Table 2, $DT_{LB}<DT_{M9}<DT_{NOX}=DT_{pH5}<DT_{CAM}=DT_{TMP}$). Interesting, we did not see this relationship as DFEs for both CAM and pH5 violate this expectation (Figure 18).

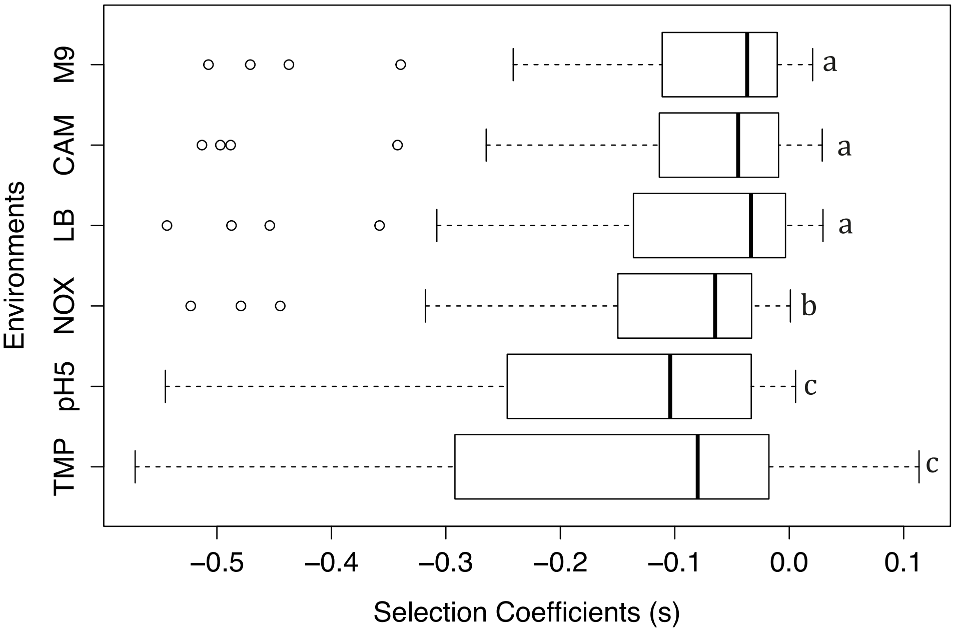

Figure 18. Boxplot representation of DFEs of transferred genes in six environments tested in this study.

Table 3. Wilcoxon signed rank tests of the pairwise comparisons of environments.

	M9	CAM	LB	NOX	pH5	TMP
M9	-	.282	.721	<.001	<.001	<.001
CAM	1.179	-	.121	.001	<.001	<.001
LB	0.432	1.669	-	.001	<.001	<.001
NOX	5.310	3.454	3.524	-	<.001	.027
pH5	5.520	5.042	5.287	4.750	-	.830
TMP	4.365	-3.653	4.353	-2.311	-0.222	-

p-values (upper diagonal) and Z statistics (lower diagonal) from pairwise comparisons of all environments with two-sided Wilcoxon signed rank tests. Shaded fields are significant with $\alpha = 0.05$, values are corrected for multiple testing by FDR.

3.4.2.2 *Shape and Spread*

Figure 17-a and 18 seem to indicate a similarity of shape and spread of DFEs in different environments, even though environment significantly alters the central tendency of DFEs of our set of genes. We tested this by employing the Kolmogorov-Smirnov (K-S) test, which is less sensitive to the changes in the median of the fitness effects than the Wilcoxon singed rank test and as such is more indicative of the shape and spread of our distributions (Lehmann & D'abrera 2006). We observe that none of the distributions are significantly different from each other in terms of their shape and spread (Table 4). Taken together with the observation in Figure 17-b, these results suggest that, even though we might not be able to predict how the fitness cost of a gene changes in a particular environment the nature of the shape and spread is an intrinsic property of HGT DFEs.

Table 4. Kolmogorov - Smirnov tests of the pairwise comparisons of environments.

	M9	**CAM**	**LB**	**NOX**	**pH5**	**TMP**
M9	-	0.943	0.737	0.228	0.228	0.274
CAM	0.114	-	0.872	0.274	0.228	0.388
LB	0.159	0.136	-	0.228	0.228	0.435
NOX	0.273	0.250	0.273	-	0.435	0.435
pH5	0.318	0.273	0.273	0.205	-	0.583
TMP	0.250	0.227	0.205	0.205	0.182	-

p-values (upper diagonal) and D statistics (lower diagonal) from pairwise comparisons of all environments with two-sided Kolmogorov - Smirnov tests. α = 0.05, values are corrected for multiple testing by FDR.

3.4.2.3 *Gene-by-Environment Interactions*

Given that the environment changes the central tendency of DFEs in our set of genes significantly, we investigated if environments affect the fitness effects of different genes differently and found a strong interaction between individual genes and the environment ($F_{215,\,1227}$ = 77.4, p<.001) (Figure 17-a and Table 5 – see Overall line). Interestingly, this interaction at the single gene level results in a completely unpredictable distribution of fitness effects (Figure 17-b). In fact, all possible pairwise comparisons of the gene-by-environment interaction between the six environments were significant (Table 5), demonstrating that fitness effects of individual genes depend on the specific environment in a seemingly unpredictable manner. This is the case even between those environments that have same median fitness effects, such as M9, CAM and LB; or pH5 and TMP. What's more, especially for genes whose fitness effects vary substantially between environments (Figure 17-b), the host cells become more tolerant to having these genes under specific conditions, so that a gene otherwise with a strong negative fitness effect can become beneficial or neutral in other environments. For such genes, the likelihood of HGT might increase in heterogeneous or fluctuating environments.

The reasons for why some genes in our dataset shows strong dependency to environment can be better understood from their specific functions (Figure 19). For example, gene *exbB* is known to be beneficial in the presence of antibiotics, as it is a subunit of a TonB-dependent energy transduction complex that provides energy to an efflux pump (*mtrCDE*) involved in multidrug resistance (Zhao et al. 1998; Toone 2011). Similarly, genes *hupA* and *hupB* are subunits of the same DNA binding protein and are known to provide fitness benefits in CAM (Kano et al. 1986; Kano et al. 1987). Furthermore, *kdpD* gene is part of a histidine kinase/response regulator system that senses K^+ limitation and induces the *kdpFABC* operon encoding a high-affinity K^+ uptake complex, which becomes vital under low pH conditions (Yan et al. 2011; Heermann et al. 2014). The *S. typhimurium* ortholog seems to interfere with this function and decreases the fitness of the host cell dramatically in pH5. While explanations for all genes are not nearly as clear as the above examples, it is clear that the probability of HGT is dominated by a gene-by-environment interaction determined by specific gene function — genes in similar functional categories may not behave in similar ways.

Table 5. Analysis of Variance of the pairwise comparisons of environments.

Pairwise comparisons		**Environment**	**Gene**	**Gene X Environment**
M9	**CAM**	$F_{1,391}$ = 133.5 p<.001	$F_{43,391}$ = 3590.5 p<.001	$F_{43,391}$ = 114.6 p<.001
M9	**LB**	$F_{1,391}$ = 59.5 p<.001	$F_{43,391}$ = 2042.8 p<.001	$F_{43,391}$ = 87.1 p<.001
M9	**NOX**	$F_{1,391}$ = 346.3 p<.001	$F_{43,391}$ = 1001.9 p<.001	$F_{43,391}$ = 6.5 p<.001
M9	**pH5**	$F_{1,391}$ = 4353.0 p<.001	$F_{43,391}$ = 1723.9 p<.001	$F_{43,391}$ = 114.6 p<.001
M9	**TMP**	$F_{1,348}$ = 2609.9 p<.001	$F_{43,348}$ = 978.9 p<.001	$F_{43,348}$ = 141.9 p<.001
CAM	**LB**	$F_{1,435}$ =2.9 p=.090	$F_{43,435}$ = 2191.9 p<.001	$F_{43,435}$ = 127.2 p<.001
CAM	**NOX**	$F_{1,435}$ = 247.9 p<.001	$F_{43,435}$ = 864.7 p<.001	$F_{43,435}$ = 25.3 p<.001
CAM	**pH5**	$F_{1,435}$ = 4226.3 p<.001	$F_{43,435}$ = 1771.8 p<.001	$F_{43,435}$ = 160.8 p<.001
CAM	**TMP**	$F_{1,391}$ = 2226.7 p<.001	$F_{43,391}$ = 1166.3 p<.001	$F_{43,391}$ = 160.5 p<.001
LB	**NOX**	$F_{1,435}$ = 317.1 p<.001	$F_{43,435}$ = 1345.7 p<.001	$F_{43,435}$ = 43.1 p<.001
LB	**pH5**	$F_{1,435}$ = 3264.8 p<.001	$F_{43,435}$ = 1521.7 p<.001	$F_{43,435}$ = 109.8 p<.001
LB	**TMP**	$F_{1,391}$ = 2453 p<.001	$F_{43,391}$ = 1339 p<.001	$F_{43,391}$ = 148 p<.001
NOX	**pH5**	$F_{1,435}$ = 1029 p<.001	$F_{43,435}$ = 932.5 p<.001	$F_{43,435}$ = 48.2 p<.001
NOX	**TMP**	$F_{1,391}$ = 1006.5 p<.001	$F_{43,391}$ = 881.9 p<.001	$F_{43,391}$ = 119.5 p<.001
pH5	**TMP**	$F_{1,391}$ = 4.6 p=.034	$F_{43,391}$ = 918.4 p<.001	$F_{43,391}$ = 110.3 p<.001
OVERALL		$F_{5,1227}$ = 1184.1 p<.001	$F_{43,1227}$ = 2805 p<.001	$F_{215, 1227}$= 77.4 p<.001

F statistics and p-values from pairwise comparisons of all environments with ANOVA tests. Shaded fields are significant at α = 0.05, values are corrected for multiple testing by FDR. 44 genes in each environment represented by 5 or 6 replicates, which are given to the model in error structure.

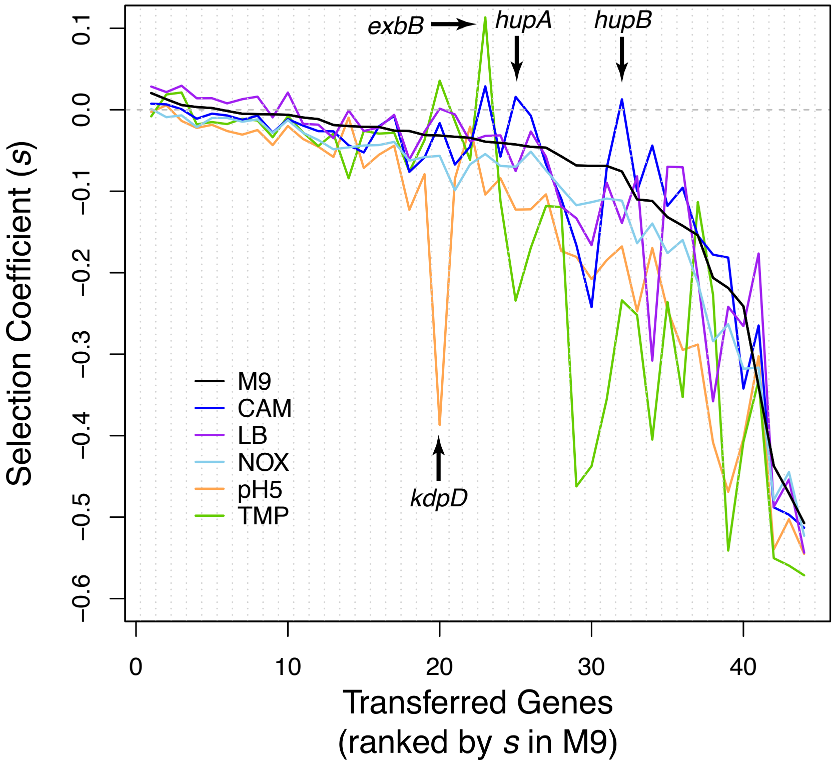

Figure 19. Selection coefficients of the transferred genes in six different environments, with few specific examples of genes whose effect changes conditionally.

3.4.3 *Predictability is Related to Fitness Effects*

Data shown in Figure 17-b seems to indicate that genes with higher average cost over all environments (selection coefficient, s<-0.1) exhibit more unpredictability – their fitness effect varies substantially between environments – except the nearly lethal genes (selection coefficient, s<-0.4). Indeed, we observe a significant linear relationship between the mean and standard deviation of fitness effects of genes across the six environments ($F_{1,42}$ = 7.869, p = .008, r^2 = 0.158). A quadratic model, however, explains the data significantly better ($F_{2,41}$ = 32.160, p < .001, r^2 = 0.611), suggesting a bell-shaped relationship between mean fitness effects of the genes and variance of them across all environments (Figure 20). This relationship is observed between fitness effects of genes in each environment and the variance of fitness effects of the genes even when the environments are analyzed separately suggesting this is a common property of HGT (all six quadratic models gave p values less than .01). Based on this observation we divided the tested genes into three categories: (i) 'less deleterious genes', with mean fitness effects across all environments of more than -0.1 and SD<0.05; (ii) 'highly deleterious genes', with SD>0.05; and (iii) 'nearly lethal genes', with mean fitness effects across all environments of less than -0.4 and SD<0.05. Interestingly, it is the middle 'highly deleterious genes' category that shows the strongest environmental dependence, and are therefore highly unpredictable (Figure 20). Whereas, categories (i) and (iii) are highly predictable and less influenced by the environment but rather solely by their intrinsic properties.

The existence of these three distinct groups can at least in part be explained by the specific functions of genes in question. The low environmental variability of fitness effects of 'nearly lethal genes' (*uvrC*, *lolA*, and *topB*) likely arises from the vital role these genes play in the cell. However, it should be noted that while 'nearly lethal genes' are statistically more predictable and less dependent on the environment they are not completely independent of their properties. Similarly, we can understand why some of the 'highly deleterious genes' are beneficial in some of the tested environments from their functions, such as the genes *expB*, *hupA* and *hupB* which contribute to antibiotic resistance. However, this is possible only for those genes that we understand the function in great detail, which constitutes a marginally small part of all *E. coli* genes. The reason for the low environmental variability of the 'less deleterious genes' (s>-0.1) is harder to interpret directly from their function.

We attempted to better understand the potential differences between the three groups of genes. Specifically, we examined if the genes in three groups differed based on the number of interaction partners, length of the coding sequence, level of divergence between homologs, difference in the codon bias between homologs, and the change in the level of the expression of the endogenous copy of the gene over all conditions. We focused on these factors to differentiate between the three groups of genes as they were previously suggested to impact the likelihood of successful HGT (see the review of (Baltrus 2013, and Chapter 2). The genes in the three tested groups did not significantly differ based on any of these factors (Table 6). Additionally, we did not observe an enrichment for the functional categories of informational or operational genes within this group of non-predictable genes (8 informational out of 17 unpredictable genes, whereas 10 informational out of 27 less deleterious + nearly lethal genes, Fisher exact test, p=0.544). Again, these findings are a strong indicator that the gene-by-environment interactions are largely determining the outcomes of HGT.

Table 6. Analysis of Variance for the intrinsic properties of transferred genes.

Factors	F ratio	p-Values
GC content	0.016	0.984
PPI *	1.877	0.498
FOP	0.162	0.984
Length	4.731	0.084
TPM	1.415	0.51

See Materials and Methods section 3.3.8 for the detailed description of the test. Factorial variable is composed of three groups: 'less deleterious genes', 'highly deleterious genes', and 'nearly lethal genes'. α = 0.05, values are corrected for multiple testing by FDR.
** Since difference in the PPI level among environments were insignificant mean number of PPI over all environments is used for the analysis.*

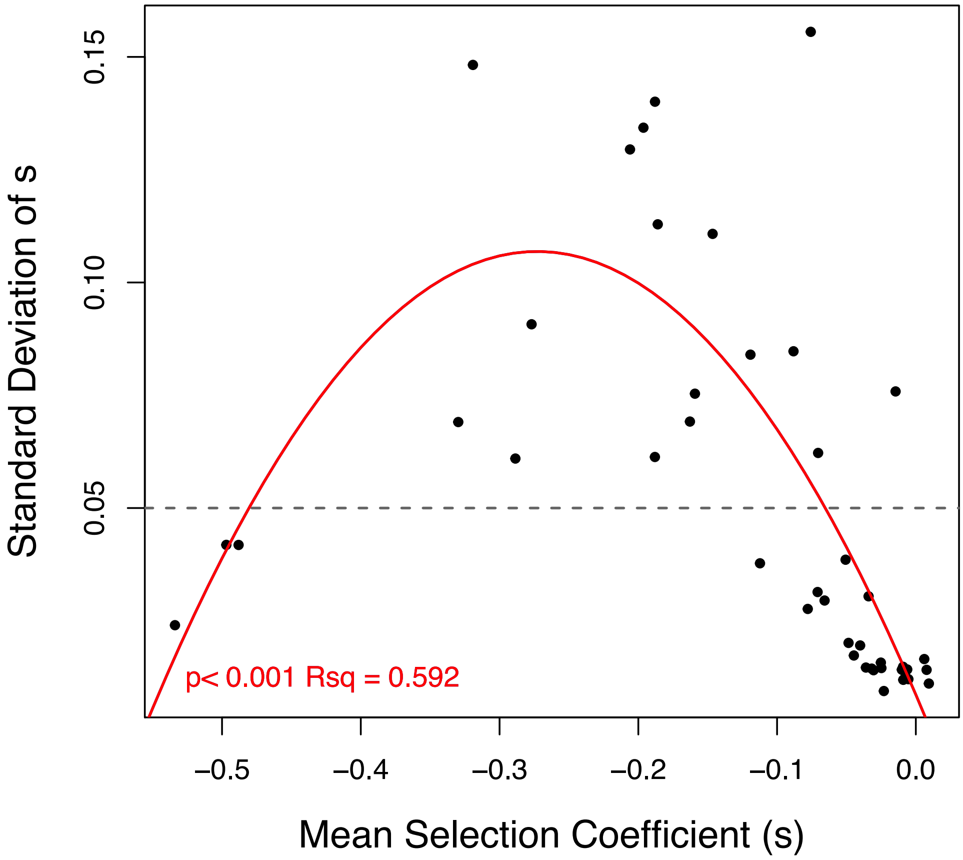

Figure 20. Relationship between mean and standard deviation of selection coefficients of newly transferred genes over all environments.

3.5 Conclusion

We found that the DFEs of horizontally transferred genes are highly dependent on the environment, with abundant gene–by-environment interactions indicating that in the long run success of an HGT event is determined by the heterogeneity of the environment in which the gene is transferred. Furthermore, we demonstrated a relationship between average fitness effect of a gene across all environments and its environmental variance, and thus its predictability. Nearly lethal genes seem to stay nearly lethal independent of the environment. Similarly, neutral genes do not have a strong effect on fitness in any environment. However, genes in the middle of these two extremes exhibit the highest dependency on the environment and they span whole spectrum of the fitness from beneficial to lethal in a completely unpredictable manner. For example, the *kdpD* gene is almost lethal in one environment while being neutral or even beneficial in other environments. Taken together, these findings indicate that these middle class genes may actually segregate in the population a lot longer than we predict through the models of population genetics, depending on how these genes interact with the environment and how quickly the environment fluctuates. In other words, although counterintuitive, the likelihood of a successful HGT event does not necessarily correlate with the average fitness cost of a gene.

Previously we showed that the fitness effect of a gene might be affected by the functional category, length or dosage of the transferred gene (Chapter 2). With this study we understand that trying to explain HGT only by those factors is oversimplification of the reality and the chance of an HGT event is largely determined by the environment.

Finally, in spite of the fitness effects of genes being highly environment-dependent, we still observe a common shape of DFEs across all tested environments. This finding enables more robust and well-founded modeling of HGT. In general, our study points to the potential caveat in experimentally observing HGT in only a single environment, and suggests that understanding the evolutionary likelihood of a successful horizontal transfer must be viewed across a range of biologically meaningful environments while still considering the mechanistic nature of the gene’s role in the recipient cell.

3.6 Acknowledgements

We thank Mato Lagator for valuable discussions and inputs.

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: Hande Acar. Performed the experiments: Hande Acar and Katharina Pöcher. Wrote the paper: Hande Acar and Katharina Pöcher. Performed statistical analysis: Hande Acar and Katharina Pöcher. Performed NGS and RNA-seq analysis: Katharina Pöcher.

4 Conclusions

In this study, we designed an experimental system by which we aimed at elucidating the selective barriers and their relative importance in a systematic way.

In chapter 2, we obtained a DFE for the newly transferred genes in the recipient host at a constant expression level. Analyzing this DFE revealed gene dosage as an important barrier to HGT, especially for genes for which the native expression level seems limited, more than 10-fold increase in their expression exhibited significantly higher probability of fitness loss. In addition, the functional category and the length of the genes emerged as potential selective barriers, while protein-protein interaction, GC content or codon usage did not show an effect on the host fitness in any predictable way.

In chapter 3, we investigated the role of environment on the DFEs of newly transferred genes by testing the same set of genes in six environments with different cellular stress levels on the recipient host. Our data confirmed a strong effect of environment with even stronger gene-by-environment interactions. This high rate of environmental dependency of the genes was related to the mean fitness effect of the genes overall environments, (i) nearly lethal genes remain similarly fatal among environments, (ii) mildly deleterious genes stay steady, and (iii) the genes between these two classes, the highly deleterious genes, exhibit an unpredictable change in their fitness in each environment. Despite this turmoil, the overall shape of the DFEs remains similar that suggest the average distribution of HGT DFEs are invariant to the specifics of the environment.

Overall, our study demonstrates the further need for more realistic experiments that include fluctuating environments, heterogeneous environments, and spatially structured environments, such as the animal gut or soil.

References

Acuña, R. et al., 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences of the United States of America*, 109(11), pp.4197–4202.

Andam, C.P., Carver, S.M. & Berthrong, S.T., 2015. Horizontal Gene Flow in Managed Ecosystems. *Annual Review of Ecology, Evolution, and Systematics*, 46(1), pp.121–143.

Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), p.R106.

Baltrus, D.A., 2013. Exploring the costs of horizontal gene transfer. *Trends in Ecology & Evolution*, pp.1–7.

Barkay, T. & Smets, B., 2005. Horizontal gene flow in microbial communities. *Asm News*, 71(9), pp.412–419.

Beaber, J.W. et al., 2002. Comparison of SXT and R391, two conjugative integrating elements: definition of a genetic backbone for the mobilization of resistance determinants. *Cellular and Molecular Life ...*, 59(12), pp.2065–2070.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57, pp.289–300.

Bershtein, S. et al., 2015. Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. M. Achtman, ed. *PLoS Genetics*, 11(10), p.e1005612.

Bikard, D. et al., 2012. CRISPR Interference Can Prevent Natural Transformation and Virulence Acquisition during In Vivo Bacterial Infection. *Cell Host and Microbe*, 12(2), pp.177–186.

Biller, S.J. et al., 2014. Bacterial Vesicles in Marine Ecosystems. *Science*, 343(6167), pp.183–186.

Bonomo, J. & Gill, R.T., 2005. Amino acid content of recombinant proteins influences the metabolic burden response. *Biotechnology and Bioengineering*, 90(1), pp.116–126.

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

Brigulla, M. & Wackernagel, W., 2010. Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regard to safety issues. *Applied microbiology and biotechnology*, 86(4), pp.1027–1041.

Canchaya, C. et al., 2003. Phage as agents of lateral gene transfer. *Current Opinion in Microbiology*, 6(4), pp.417–424.

Chen, I. & Dubnau, D., 2004. DNA uptake during bacterial transformation. *Nature Reviews Microbiology*, 2(3), pp.241–249.

Cochran, P.K., Kellogg, C.A. & Paul, J.H., 1998. Prophage induction of indigenous marine lysogenic bacteria by environmental pollutants. *Marine Ecology Progress Series*, 164, pp.125–133.

Cohen, O., Gophna, U. & Pupko, T., 2011. The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer. *Molecular biology and evolution*, 28(4), pp.1481–1489.

Cox, R.S., Dunlop, M.J. & Elowitz, M.B., 2010. A synthetic three-color scaffold for monitoring genetic regulation and noise. *Journal of biological engineering*, 4, p.10.

Dagan, T. & Martin, W., 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences*, 104(3), pp.870–875.

Diederich, L., Rasmussen, L.J. & Messer, W., 1992. New cloning vectors for integration in the lambda attachment site attB of the Escherichia coli chromosome. *Plasmid*, 28(1), pp.14–24.

Doolittle, W.F., 1999a. Lateral genomics. *Trends in genetics*, 15(12), pp.M5–M8.

Doolittle, W.F., 1999b. Phylogenetic Classification and the Universal Tree. *Science*, 284(5423), pp.2124–2128.

Drummond, D.A. & Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, 10(10), pp.715–724.

Dubey, G.P. & Ben-Yehuda, S., 2011. Intercellular Nanotubes Mediate Bacterial Communication. *Cell*, 144(4), pp.590–600.

Eames, M. & Kortemme, T., 2012. Cost-benefit tradeoffs in engineered lac operons. *Science*, 336(6083), pp.911–915.

Elena, S.F. et al., 1998. Distribution of fitness effects caused by random insertion mutations in Escherichia coli. *Genetica*, 102-103(1-6), pp.349–358.

Elowitz, M.B. et al., 2002. Stochastic gene expression in a single cell. *Science*, 297(5584), pp.1183–1186.

Eyre-Walker, A. & Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8), pp.610–618.

Frost, L.S. et al., 2005. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9), pp.722–732.

Garcillán-Barcia, M.P. & la Cruz, de, F., 2008. Why is entry exclusion an essential feature of conjugative plasmids? *Plasmid*, 60(1), pp.1–18.

Gingold, H. & Pilpel, Y., 2011. Determinants of translation efficiency and accuracy. *Molecular Systems Biology*, 7, pp.1–13.

González-Candelas, F., 2012. Barriers to Horizontal Gene Transfer: Fuzzy and Evolvable Boundaries. *Horizontal Gene Transfer in ….*

Gophna, U. & Ofran, Y., 2011. Lateral acquisition of genes is affected by the friendliness of their products. *Proceedings of the National Academy of Sciences,* 108(1), pp.343-348.

Haldimann, A. & Wanner, B.L., 2001. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. *Journal of bacteriology*, 183(21), pp.6384–6393.

Heermann, R. et al., 2014. Dynamics of an Interactive Network Composed of a Bacterial Two-Component System, a Transporter and K+ as Mediator H. W. van Veen, ed. *PLoS ONE*, 9(2), p.e89671.

Hu, P. et al., 2009. Global Functional Atlas of Escherichia coli Encompassing Previously Uncharacterized Proteins A. Levchenko, ed. *PLoS Biology*, 7(4), p.e1000096.

Innan, H. & Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2), pp.97–108.

Jain, R., Rivera, M.C. & Lake, J.A., 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7), pp.3801–3806.

Kano, Y. et al., 1987. Cloning and sequencing of the HU-2 gene of Escherichia coli. *Molecular & general genetics : MGG*, 209(2), pp.408–410.

Kano, Y. et al., 1986. Genetic characterization of the gene hupB encoding the HU-1 protein of Escherichia coli. *Gene*, 45(1), pp.37–44.

Keeling, P.J. & Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8), pp.605–618.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129), pp.624–626.

Kinney, J.B. et al., 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20), pp.9158–9163.

Kishony, R. & Leibler, S., 2003. Environmental stresses can alleviate the average deleterious effect of mutations. *Journal of biology*, 2(2), p.14.

Knöppel, A. et al., 2014. Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Molecular biology and evolution*, 31(5), pp.1220–1227.

Kolling, G.L. & Matthews, K.R., 1999. Export of virulence genes and Shiga toxin by membrane vesicles of Escherichia coli O157:H7. *Applied and Environmental Microbiology*, 65(5), pp.1843–1848.

Koonin, E.V., Makarova, K.S. & Aravind, L., 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual review of microbiology*, 55, pp.709–742.

Koskella, B. & Meaden, S., 2013. Understanding bacteriophage specificity in natural microbial communities. *Viruses*, 5(3), pp.806–823.

Kunin, V., 2003. The Balance of Driving Forces During Genome Evolution in Prokaryotes. *Genome Research*, 13(7), pp.1589–1594.

Kuo, C.-H. & Ochman, H., 2009. The fate of new bacterial genes. *FEMS Microbiology Reviews*, 33(1), pp.38–43.

Kurland, C.G., 2005. What tangled web: barriers to rampant horizontal gene transfer. *BioEssays*, 27(7), pp.741–747.

Labrie, S.J., Samson, J.E. & Moineau, S., 2010. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, 8(5), pp.317–327.

Lang, A.S., Zhaxybayeva, O. & Beatty, J.T., 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nature Reviews Microbiology*, 10(7), pp.472–482.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Lawrence, J.G., 2002. The Dynamics of Bacterial Genomes. In M. Syvanen & C. I. Kado, eds. *Horizontal gene transfer*. Academic Pr, pp. 95–110.

Lawrence, J.G. & Ochman, H., 1997. Amelioration of Bacterial Genomes: Rates of Change and Exchange. *Journal of molecular evolution*, 44(4), pp.383–397.

Lehmann, V. & D'abrera, H.J., 2006. *Nonparametrics Statistical Methods Based on Ranks*, Springer.

Levine, S.M. et al., 2007. Plastic cells and populations: DNA substrate characteristics in Helicobacter pylori transformation define a flexible but conservative system for genomic variation. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 21(13), pp.3458–3467.

Li, B. & Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1), p.323.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Lucchini, S. et al., 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathogens*, 2(8), pp.e81–752.

Lukjancenko, O., Wassenaar, T.M. & Ussery, D.W., 2010. Comparison of 61 Sequenced

Escherichia coli Genomes. *Microbial Ecology*, 60(4), pp.708–720.

Lutz, R. & Bujard, H., 1997. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6), pp.1203–1210.

Lynch, M. & Marinov, G.K., 2015. The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America*, 112(51), pp.15690–15695.

Marraffini, L.A. & Sontheimer, E.J., 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, 322(5909), pp.1843–1845.

McClelland, M. et al., 2001. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature*, 413(6858), pp.852–856.

McGrath, B.M. & Pembroke, J.T., 2004. Detailed analysis of the insertion site of the mobile elements R997, pMERPH, R392, R705 and R391 in E. coliK12. *FEMS Microbiology Letters*, 237(1), pp.19–26.

Melnyk, A.H., Wong, A. & Kassen, R., 2015. The fitness costs of antibiotic resistance mutations. *Evolutionary Applications*, 8(3), pp.273–283.

Mira, A., Ochman, H. & Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. *Trends in genetics*, 17(10), pp.589–596.

Moran, N.A. & Jarvik, T., 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*, 328(5978), pp.624–627.

Moscoso, M. & Claverys, J.-P., 2004. Release of DNA into the medium by competent Streptococcus pneumoniae: kinetics, mechanism and stability of the liberated DNA. *Molecular microbiology*, 54(3), pp.783–794.

Mugnai, R. et al., 2015. A Survey of Escherichia coli and Salmonella in the Hyporheic Zone of a Subtropical Stream: Their Bacteriological, Physicochemical and Environmental Relationships. *PLoS ONE*, 10(6), p.e0129382.

Nakamura, Y. et al., 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, 36(7), pp.760–766.

Navarre, W.W., 2016. *Chapter Three - The Impact of Gene Silencing on Horizontal Gene Transfer and Bacterial Evolution*, Elsevier.

Navarre, W.W. et al., 2007. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes & development*, 21(12), pp.1456–1471.

Nielsen, K.M. et al., 2007. Release and persistence of extracellular DNA in the environment. *Environmental biosafety research*, 6(1-2), pp.37–53.

Ochman, H., Lawrence, J. & Groisman, E., 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405, pp.299–304.

Omer, S. et al., 2010. Integration of a Foreign Gene into a Native Complex Does Not Impair Fitness in an Experimental Model of Lateral Gene Transfer. *Molecular biology and evolution*, 27(11), pp.2441–2445.

Pande, S. et al., 2015. Metabolic cross-feeding via intercellular nanotubes among bacteria. *Nature Communications*, 6, p.6238.

Papp, B., Pál, C. & Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945), pp.194–197.

Park, C. & Zhang, J., 2012. High Expression Hampers Horizontal Gene Transfer. *Genome Biology and Evolution*, 4(4), pp.523–532.

Pál, C., Papp, B. & Lercher, M.J., 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12), pp.1372–1375.

Polz, M.F., Alm, E.J. & Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in genetics*, 29(3), pp.170–175.

Popa, O. & Dagan, T., 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*, 14(5), pp.615–623.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Raghavan, R., Kelkar, Y.D. & Ochman, H., 2012. A selective force favoring increased G+C content in bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36), pp.14504–14507.

Remold, S.K. & Lenski, R.E., 2001. Contribution of individual random mutations to genotype-by-environment interactions in Escherichia coli. *Proceedings of the National Academy of Sciences*, 98(20), pp.11388–11393.

Rivera, M.C. et al., 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.6239–6244.

Roller, B.R.K., Stoddard, S.F. & Schmidt, T.M., 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nature microbiology*, 1, p.16160.

Roux, D. et al., 2015. Fitness cost of antibiotic susceptibility during bacterial infection. *Science Translational Medicine*, 7(297), pp.–297ra114.

Sanjuán, R., Moya, A. & Elena, S.F., 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences*, 101(22), pp.8396–8401.

Seitz, P. & Blokesch, M., 2013. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. *FEMS Microbiology Reviews*, 37(3), pp.336–363.

Sengeløv, G. et al., 2001. Effect of Genomic Location on Horizontal Transfer of a Recombinant Gene Cassette Between Pseudomonas Strains in the Rhizosphere and Spermosphere of Barley Seedlings. *Current Microbiology*, 42(3), pp.160–167.

Shachrai, I. et al., 2010. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Molecular cell*, 38(5), pp.758–767.

Shah, P. et al., 2013. Rate-Limiting Steps in Yeast Protein Translation. *Cell*, 153(7), pp.1589–1601.

Shapiro, B.J. et al., 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336(6077), pp.48–51.

Sharp, P.M., Emery, L.R. & Zeng, K., 2010. Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544), pp.1203–1212.

Shen, P. & Huang, H.V., 1986. Homologous recombination in Escherichia coli: dependence on substrate length and homology. *Genetics*, 112(3), pp.441–457.

Shousha, A. et al., 2015. Bacteriophages Isolated from Chicken Meat and the Horizontal Transfer of Antimicrobial Resistance Genes. J. Björkroth, ed. *Applied and Environmental Microbiology*, 81(14), pp.4600–4606.

Shultzaberger, R.K. et al., 2012. Probing the Informational and Regulatory Plasticity of a Transcription Factor DNA–Binding Domain H. D. Madhani, ed. *PLoS Genetics*, 8(3), pp.e1002614–13.

Sorek, R. et al., 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, 318(5855), pp.1449–1452.

Soucy, S.M., Huang, J. & Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8), pp.472–482.

Tatusov, R.L. et al., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1), pp.33–36.

Thomas, C.M. & Nielsen, K.M., 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, 3(9), pp.711–721.

Tock, M.R. & Dryden, D.T., 2005. The biology of restriction and anti-restriction. *Current Opinion in Microbiology*, 8(4), pp.466–472.

Toone, E.J., 2011. *Advances in Enzymology and Related Areas of Molecular Biology*, John Wiley & Sons.

Treangen, T.J. & Rocha, E.P.C., 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics*, 7(1), p.e1001284.

Tuller, T. et al., 2011. Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Research*, 39(11), pp.4743–4755.

Wellner, A. & Gophna, U., 2008. Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. *Molecular biology and evolution*, 25(9), pp.1835–1840.

Wellner, A., Lurie, M.N. & Gophna, U., 2007. Complexity, connectivity, and duplicability as barriers to lateral gene transfer. *Genome Biology*, 8(8), p.R156.

Winfield, M.D. & Groisman, E.A., 2003. Role of nonhost environments in the lifestyles of Salmonella and Escherichia coli. *Applied and Environmental Microbiology*, 69.7(July), pp.3687–3694.

Wu, T.D. & Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), pp.1859–1875.

Yan, H. et al., 2011. Expression and Activity of Kdp under Acidic Conditions in Escherichia coli. *Biological and Pharmaceutical Bulletin*, 34(3), pp.426–429.

Zhao, Q. et al., 1998. Influence of the TonB energy-coupling protein on efflux-mediated multidrug resistance in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, 42(9), pp.2225–2231.

Zhaxybayeva, O., 2009. Detection and quantitative assessment of horizontal gene transfer. *Methods in molecular biology (Clifton, N.J.)*, 532(Chapter 11), pp.195–213.

Zhaxybayeva, O. & Doolittle, W.F., 2011. Lateral gene transfer. *Current Biology*, 21(7), pp.R242–R246.

Zhong, Z., Helinski, D. & Toukdarian, A., 2005. Plasmid host-range: restrictions to F replication in Pseudomonas. *Plasmid*, 54(1), pp.48–56.

A. Appendix 1

Data used in the multiple linear regression analysis in Chapter 2, related to the transferred *Salmonella* genes.

STM	Gene	STM		Functional	DEV	DEV	STM	
Gene ID	Name	Sel.Coef.	PPI	Category	GC	FOP	Length	TPM
STM0160	*yacL*	-0.00206	33	Operational	0.001	0.058	363	76.412
STM0172	*yadG*	-0.13294	39	Operational	0.002	0.016	927	48.556
STM0226	*lpxD*	0.00921	30	Operational	0.008	0.041	1026	379.717
STM0264	*dnaQ*	-0.01959	12	Informational	0.010	0.020	732	63.496
STM0425	*thiI*	-0.15182	1	Operational	0.002	0.012	1449	114.670
STM0451	*hupB*	-0.08211	20	Informational	0.005	0.000	273	2558.926
STM0614	*uspG*	-0.00681	12	Operational	0.022	0.028	429	185.550
STM0629	*cspE*	-0.00923	31	Informational	0.002	0.071	210	7758.413
STM0648	*leuS*	-0.27576	8	Informational	0.025	0.043	2583	220.013
STM0703	*kdpD*	-0.03411	1	Operational	0.013	0.006	2685	9.113
STM0801	*ybhK*	0.00896	3	Operational	0.030	0.073	909	27.137
STM0806	*moaE*	-0.00624	0	Operational	0.043	0.033	453	30.234
STM0831	*dps*	-0.00963	12	Operational	0.007	0.018	504	442.589
STM0943	*cspD*	-0.02427	28	Informational	0.036	0.006	222	91.136
STM0945	*clpA*	-0.04958	31	Operational	0.016	0.037	2277	165.892
STM0961	*lolA*	-0.43262	1	Operational	0.005	0.017	615	97.031
STM1061	*rlmL*	0.00035	21	Informational	0.008	0.011	2109	79.059
STM1112	*cbpA*	-0.00678	23	Operational	0.006	0.104	921	66.899
STM1185	*rne*	-0.15861	35	Informational	0.028	0.063	3204	187.729
STM1196	*acpP*	-0.07789	20	Operational	0.002	0.000	237	10743.340
STM1298	*topB*	-0.45193	16	Informational	0.012	0.022	1950	65.180
STM1334	*infC*	0.00125	36	Informational	0.010	0.006	543	3656.568
STM1366	*ydiI*	-0.00082	25	Operational	0.052	0.022	411	56.133
STM1696	*sapF*	-0.00181	2	Operational	0.032	0.011	807	35.585
STM1946	*uvrC*	-0.60555	21	Informational	0.017	0.008	1833	115.128
STM2388	*fadJ*	-0.13126	3	Operational	0.017	0.045	2148	4.159
STM2543	*iscS*	-0.02693	5	Operational	0.016	0.064	1215	166.841
STM2643	*srmB*	-0.34174	31	Informational	0.004	0.027	1335	77.441
STM3143	*hybG*	-0.00283	7	Operational	0.006	0.048	249	89.704
STM3159	*exbB*	-0.00244	14	Operational	0.027	0.012	735	368.523
STM3215	*yqjI*	-0.05091	14	Informational	0.001	0.031	648	30.034
STM3282	*pnp*	0.00037	37	Informational	0.022	0.008	2136	770.983
STM3514	*malP*	-0.10279	25	Operational	0.031	0.020	2394	30.570
STM3656	*glyQ*	-0.01117	9	Informational	0.003	0.059	912	244.495
STM3682	*selB*	-0.08022	29	Informational	0.008	0.039	1851	35.890
STM3689	*yibL*	-0.01623	24	Operational	0.001	0.058	363	78.815
STM3808	*ibpB*	-0.00992	3	Operational	0.029	0.119	429	7.576
STM3854	*pstB*	-0.05355	17	Operational	0.001	0.027	774	26.221
STM4170	*hupA*	-0.02744	34	Informational	0.006	0.011	273	1720.314
STM4237	*lexA*	-0.10157	14	Operational	0.032	0.005	609	348.214
STM4361	*hfq*	0.00285	24	Operational	0.011	0.029	309	1171.244
STM4394	*rplI*	-0.00591	27	Informational	0.012	0.013	450	4963.791
STM4458	*ridA*	-0.00917	9	Informational	0.017	0.031	387	3069.722
STM4558	*rimI*	-0.00776	12	Operational	0.027	0.027	447	70.275

B. Appendix 2

Data used in the analyses in Chapter 3, selection coefficients of the transferred *Salmonella* genes in six different environments.

STM	**Gene**								
Gene ID	**Name**	**M9**	**CAM**	**LB**	**NOX**	**pH5**	**TMP**	**Mean**	**SD**
STM0160	*yacL*	-0.006	-0.012	0.021	-0.013	-0.020	-0.008	-0.006	0.014
STM0172	*yadG*	-0.112	-0.044	-0.308	-0.140	-0.170	-0.405	-0.196	0.134
STM0226	*lpxD*	0.013	0.007	0.022	-0.009	0.006	0.019	0.009	0.011
STM0264	*dnaQ*	-0.047	-0.067	-0.058	-0.074	-0.104	-0.118	-0.078	0.028
STM0425	*thiI*	-0.154	-0.154	-0.206	-0.212	-0.288	-0.113	-0.188	0.061
STM0451	*hupB*	-0.076	0.013	-0.139	-0.112	-0.168	-0.234	-0.119	0.084
STM0614	*uspG*	-0.021	-0.052	-0.026	-0.044	-0.071	-0.026	-0.040	0.020
STM0629	*cspE*	-0.046	-0.007	-0.027	-0.051	-0.122	-0.169	-0.070	0.062
STM0648	*leuS*	-0.241	-0.342	-0.266	-0.318	-0.404	-0.408	-0.330	0.069
STM0703	*kdpD*	-0.032	-0.016	0.001	-0.057	-0.387	0.036	-0.076	0.156
STM0801	*ybhK*	0.006	0.001	0.029	-0.007	-0.014	0.021	0.006	0.017
STM0806	*moaE*	0.003	-0.011	0.014	-0.021	-0.022	-0.018	-0.009	0.015
STM0831	*dps*	-0.041	-0.058	-0.031	-0.069	-0.084	-0.112	-0.066	0.030
STM0943	*cspD*	-0.034	-0.045	-0.039	-0.067	-0.021	-0.062	-0.045	0.017
STM0945	*clpA*	-0.026	-0.076	-0.061	-0.062	-0.123	-0.076	-0.071	0.031
STM0961	*lolA*	-0.471	-0.497	-0.454	-0.445	-0.503	-0.559	-0.488	0.042
STM1061	*rlmL*	0.020	0.007	0.028	0.001	-0.003	-0.008	0.008	0.014
STM1112	*cbpA*	-0.006	-0.028	-0.009	-0.029	-0.043	-0.034	-0.025	0.015
STM1185	*rne*	-0.069	-0.242	-0.166	-0.113	-0.208	-0.437	-0.206	0.129
STM1196	*acpP*	-0.142	-0.096	-0.071	-0.160	-0.295	-0.353	-0.186	0.113
STM1298	*topB*	-0.437	-0.488	-0.487	-0.479	-0.540	-0.550	-0.497	0.042
STM1334	*infC*	-0.001	-0.007	0.008	-0.010	-0.026	-0.018	-0.009	0.012
STM1366	*ydiI*	-0.005	-0.012	0.013	-0.015	-0.031	-0.011	-0.010	0.014
STM1696	*sapF*	-0.005	-0.007	0.016	-0.012	-0.025	-0.013	-0.008	0.014
STM1946	*uvrC*	-0.507	-0.513	-0.544	-0.523	-0.545	-0.571	-0.534	0.024
STM2388	*fadJ*	-0.219	-0.181	-0.242	-0.263	-0.469	-0.541	-0.319	0.148
STM2543	*iscS*	-0.033	-0.067	-0.006	-0.099	-0.085	-0.014	-0.051	0.039
STM2643	*srmB*	-0.339	-0.265	-0.176	-0.317	-0.302	-0.332	-0.289	0.061
STM3143	*hybG*	0.002	-0.005	0.014	-0.010	-0.018	-0.015	-0.005	0.012
STM3159	*exbB*	-0.039	0.029	-0.032	-0.054	-0.104	0.113	-0.015	0.076
STM3215	*yqjI*	-0.057	-0.109	-0.119	-0.096	-0.173	-0.119	-0.112	0.038
STM3282	*pnp*	-0.132	-0.118	-0.070	-0.176	-0.245	-0.236	-0.163	0.069
STM3514	*malP*	-0.068	-0.166	-0.133	-0.117	-0.181	-0.462	-0.188	0.140
STM3656	*glyQ*	-0.020	-0.044	-0.001	-0.046	-0.010	-0.084	-0.034	0.030
STM3682	*selB*	-0.110	-0.100	-0.082	-0.164	-0.247	-0.252	-0.159	0.075
STM3689	*yibL*	-0.031	-0.058	-0.027	-0.058	-0.079	-0.037	-0.048	0.020
STM3808	*ibpB*	-0.011	-0.026	-0.018	-0.037	-0.045	-0.045	-0.030	0.014
STM3854	*pstB*	-0.069	-0.071	-0.090	-0.109	-0.185	-0.355	-0.146	0.111
STM4170	*hupA*	-0.043	0.016	-0.075	-0.070	-0.123	-0.234	-0.088	0.085
STM4237	*lexA*	-0.206	-0.178	-0.358	-0.284	-0.408	-0.227	-0.277	0.091
STM4361	*hfq*	-0.026	-0.008	-0.006	-0.039	-0.044	-0.028	-0.025	0.016
STM4394	*rplI*	-0.021	-0.020	-0.022	-0.043	-0.055	-0.029	-0.032	0.014
STM4458	*ridA*	-0.010	-0.019	-0.017	-0.028	-0.036	-0.028	-0.023	0.009
STM4558	*rimI*	-0.019	-0.026	-0.035	-0.049	-0.058	-0.029	-0.036	0.015

Mean is the mean selection coefficient of the transferred genes overall environments, and SD is the standard deviation of selection coefficient of the transferred genes overall environments.