The *lac* operon in the wild

by

Fabienne Jesse

August, 2017

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

Institute of Science and Technology

The thesis of [Student Name], titled *[Thesis Title]*, is approved by:

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

Signature: _____

Committee Member: Nick Barton, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Erik van Nimwegen, University of Basel, Basel, Switzerland

Signature: _____

Committee Member: Calin Guet, IST Austria, Klosterneuburg, Austria

Signature: _____

Defense Chair: Defense Chair Name, IST Austria, Klosterneuburg, Austria

Signature: _____

I hereby declare that this thesis 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: _____

Fabienne Jesse
August, 2017

Abstract

The *lac* operon is a classic model system for bacterial gene regulation, and has been studied extensively in *E. coli*, a classic model organism. However, not much is known about *E. coli's* ecology and life outside the laboratory, in particular in soil and water environments. The natural diversity of the *lac* operon outside the laboratory, its role in the ecology of *E. coli* and the selection pressures it is exposed to, are similarly unknown.

In Chapter Two of this thesis, I explore the genetic diversity, phylogenetic history and signatures of selection of the *lac* operon across 20 natural isolates of *E. coli* and divergent clades of *Escherichia*. I found that complete *lac* operons were present in all isolates examined, which in all but one case were functional. The *lac* operon phylogeny conformed to the whole-genome phylogeny of the divergent *Escherichia* clades, which excludes horizontal gene transfer as an explanation for the presence of functional *lac* operons in these clades. All *lac* operon genes showed a signature of purifying selection; this signature was strongest for the *lacY* gene. *Lac* operon genes of human and environmental isolates showed similar signatures of selection, except the *lacZ* gene, which showed a stronger signature of selection in environmental isolates.

In Chapter Three, I try to identify the natural genetic variation relevant for phenotype and fitness in the *lac* operon, comparing growth rate on lactose and LacZ activity of the *lac* operons of these wild isolates in a common genetic background. Sequence variation in the *lac* promoter region, upstream of the -10 and -35 RNA polymerase binding motif, predicted variation in LacZ activity at full induction, using a thermodynamic model of polymerase binding (Tugrul, 2016). However, neither variation in LacZ activity, nor RNA polymerase binding predicted by the model correlated with variation in growth rate. *Lac* operons of human and environmental isolates did not differ systematically in either growth rate on lactose or LacZ protein activity, suggesting that these *lac* operons have been exposed to similar selection pressures. We thus have no evidence that the phenotypic variation we measured is relevant for fitness.

To start assessing the effect of genomic background on the growth phenotype conferred by the *lac* operon, I compared growth on minimal medium with lactose between *lac* operon

constructs and the corresponding original isolates, I found that maximal growth rate was determined by genomic background, with almost all backgrounds conferring higher growth rates than lab strain K12 MG1655. However, I found no evidence that the lactose concentration at which growth was half maximal depended on genomic background.

Acknowledgments

Thanks to Jon Bollback for giving me the chance to do this work, for sharing the ideas that lay at the basis of this work, for his honesty and openness, showing himself to me as a person and not just as a boss.
Thanks to Nick Barton for his guidance at the last stage, reading and commenting extensively on several versions of this manuscript, and for his encouragement; thanks to both Jon and Nick for their kindness and patience.
Thanks to Erik van Nimwegen and Calin Guet for their time and willingness to be in my thesis committee, and to Erik van Nimwegen especially for agreeing to enter my thesis committee at the last moment, and for his very sharp, helpful and relevant comments during and after the defense.

Thanks to my collaborators and discussion partners: Anne Kupczok, for her guidance, ideas and discussions during the construction of the manuscript of Chapter Two, and her comments on the manuscript; Georg Rieckh for making me aware of the issue of parameter identifiability, suggesting how to solve it, and for his unfortunate idea to start the plasmid enterprise in the first place; Murat Tugrul for sharing his model, for his enthusiasm, and his comments on Chapter Three; Srdjan Sarikas for his collaboration on the Monod model fitting, fast forwarding the analysis to turbo speed and making beautiful figures, and making the discussion fun on top of it all; Vanessa Barone for her last minute comments, especially on Chapter Three, providing a sharp and very helpful experimentalist perspective at the last moment; Maros Pleska and Marjon de Vos for their comments on the manuscript of Chapter Two; Gasper Tkacik for his crucial input on the relation between growth rate and lactose concentration; Bor Kavcic for his input on growth rate modeling and error propagation.

Thanks to the Bollback, Bollenbach, Barton, Guet and Tkacik group members for both providing an inspiring and supportive scientific environment to work in, as well as a lot of warmth and colour to everyday life.

And thanks to the friends I found here, to the people who were there for me and to the people who changed my life, making it stranger and more beautiful than I could have imagined,

Maros, Vanessa, Tade, Suzi, Andrej, Peter, Tiago, Kristof, Karin, Irene, Misha, Mato, Guillaume and Zanin.

About the Author

Fabienne Jesse completed a propedeuse in Philosophy at the University of Leiden, a BSc in Biology at the Free University Amsterdam and a MSc in Animal Biology at the University of Leiden, before joining IST Austria in September 2011. She has done her bachelor's research project on visual attention in humans at the Netherlands Institute for Neuroscience, supervised by Matthew Self, and two master's research projects at the University of Leiden, one investigating whether adaptation limits diversification in *P. fluorescens* in the Evolutionary Biology and Microbiology groups, supervised by Hubertus Beaumont, and one investigating diurnal patterns of singing and song preference in the zebra finch in the Behavioural Biology group, supervised by Katharina Riebel. She has published her findings of the latter project with Katharina Riebel in the journal Behavioural Processes in 2012, under the title *Social facilitation of male song by male and female conspecifics in the zebra finch, Taeniopygia guttata*.

List of Publications

Table of Contents

List of Tables

List of Figures

1 General Introduction

The discovery of the *lac* operon, in a series of experiments over half a century ago (Jacob and Monod, 1961b) opened up the study of gene regulation. Since then, the *lac* operon has become a classic model system.

The *lac* operon consists of three genes located next to each other on the chromosome, encoding the proteins enabling lactose metabolism. The first gene, *lacZ*, encodes the protein β-galactosidase, the enzyme that splits lactose into glucose and galactose. The second gene, *lacY*, encodes the permease which transports lactose into the cell. The third gene, *lacA*, encodes galactoside acetyltransferase (thiogalactoside transacetylase), a protein which transfers an acetyl group to galactosides. The exact benefit that this protein confers is not known; it has been suggested to have a role in detoxification (Andrews and Lin, 1976; Roderick, 2005; Marbach and Bettenbrock, 2012). Transcription of the *lac* operon genes is regulated by a repressor protein, *lacI*, which is encoded directly upstream from the *lac* operon and its promoter region. In the absence of lactose, the *lac* repressor binds to the operators in the *lac* promoter region, preventing RNA polymerase from binding to the DNA and transcribing the genes of the operon (Oehler *et al.*, 1990; Gilbert and Müller-Hill, 1967). When lactose is present, its metabolite allolactose binds to the repressor, rendering it unable to bind to the DNA (Jobe *et al.*, 1972), leading to expression of the *lac* operon (induction). Allolactose is itself produced from lactose by beta-galactosidase, which means that a low level of basal or stochastic expression of the *lac* operon is required to enable its induction with lactose.

Due to its status as a model system, the *lac* operon has been studied extensively from many different angles: its complex regulation, its bistable response to lactose (Ozbudak *et al.*, 2004), the structure and function of its proteins (Jacobson *et al.*, 1994; Juers *et al.*, 2012; Lewis *et al.*, 1996; Abramson *et al.*, 2003), its response to experimental evolution (Quan *et al.*, 2012), and the costs and benefits of its expression under different conditions (Dekel and Alon, 2005; Stoebel *et al.*, 2008; Eames and Kortemme, 2012), to name a few.

However, with some exceptions (see e.g. (Dean, 1995)) most studies have focused on the

lac operon of standard laboratory strains of *E. coli*. Not much is known about the *lac* operon as it exists in nature: its natural diversity, both on a genetic and phenotypic level, its phylogenetic and evolutionary history, and the selection pressures that have shaped it to become the way it is, are still largely unknown.

In this thesis, I try to put the *lac* operon as we know it into a context, investigating its natural diversity and phylogenetic history across wild isolates of *E. coli* and divergent clades of *Escherichia*. In addition, I explore how this natural variation in the *lac* operon of these isolates maps to phenotypic variation.

With this work, I have several aims. First of all, studying the natural diversity of the *lac* operon can be taken as a proxy for studying the natural diversity of *E. coli*. In fact, about the life of *E. coli* outside the laboratory about as little is known as about the *lac* operon in nature. Discovered in the late 19th century by the German microbiologist Theodor Escherich (Blount, 2015), *E. coli*'s popularity took flight when strain K12, originally isolated in 1922 from a stool sample of a diphtheria patient in Palo Alto, California (Neidhardt and Curtiss, 1996)[1], was established as a lab model organism early in the 20th century. Since then, a lot has been discovered about its physiology, and *E. coli* has been the focus of numerous experimental evolution studies (see e.g. (Travisano and Lenski, 1996; Barrick *et al.*, 2009)). The study of the ecology of *E. coli*, however, has understandably been driven mainly by the motivation to understand pathogenic varieties, and recently also by the emerging interest in the human microbiome. In comparison, very little is known about the life of non-pathogenic strains of *E. coli* outside the human gut. Strains of *E. coli* have on occasion been isolated from soil and water environments (see e.g. (Ishii and Sadowsky, 2008)). Yet, it is not known whether these isolates represent strains that are adapted to grow in soil and water environments, or whether all *E. coli* have the gut as their principal habitat, and pass through the environment only in transit from one gut to another, without reproducing outside the gut. Studying the natural diversity of *E. coli* outside the human gut will yield a more complete picture of the ecology of *E. coli* as a species, potentially enhancing understanding of its pathogenic side as well.

For one who wants to acquire a broader understanding of the ecology and evolution of *E. coli*, the *lac* operon is a good place to start. Since lactose only occurs in mammalian milk (Savageau, 1983) it is a substance typical for the mammal, especially human gut, and rare in soil and water environments. Thus, adaptation to growth on lactose can be taken as a proxy for adaptation to the mammalian gut. Studying the conservation and genetic and phenotypic

[1] as quoted in https://www.genome.wisc.edu/resources/strains.htm

diversity of the *lac* operon across natural isolates of *E. coli* then can address the question to what extent environmentally isolated strains of *E. coli* still cycle through the mammal gut.

Another question I address in this work is where in the *lac* operon the genetic basis for natural phenotypic variation lies. In particular, I try to track whether natural variation in growth on lactose is caused primarily by variation in coding, or rather by variation in one of its several regulatory regions, and if the latter, whether cis- or trans- variation is more important. The *lac* operon is a suitable model system to investigate this, since it is a complete genetic module of manageable size, and therefore comparatively easy to work with in the lab and to design specific phenotypic measurements. In addition, the phenotypes encoded by the *lac* operon are relevant for fitness in a well defined environment. Thus, it is possible to get an idea whether the phenotypic variation we find may have been under selection in the wild, or is evolutionarily neutral. In addition, I tried to get a first insight into the role of genetic background in shaping the phenotype of growth on lactose.

Finally, a question in the background of this work is that of the origin and evolutionary history of the *lac* operon. While studying the natural diversity of the *lac* operon will not yield a definitive answer to this question, it may still contribute some perspective on this subject.

The question of the origin of the *lac* operon has two components. In the first place, one may ask how the *lac* operon as we know it ended up in *E. coli*. Answers to this question might involve horizontal transfer events, as was argued by e.g. (Ochman *et al*., 2000) but not supported by work by (Stoebel, 2005). Indeed, whether *E. coli* obtained its *lac* operon by horizontal transfer from another species is currently not known. In addition, there is the question of the 'ultimate' origin of the *lac* operon. Irrespective of the species in which it originated, one may ask how its genes became part of this co-regulated genetic module. In addition, more generally, one may ask what events and selection pressures explain the prevalence of operons in bacteria.

Tracing the signatures of selection on the genes of the *lac* operon will tell us something about the selection pressures the *lac* operon is exposed to in its natural environment. This in turn might give us an idea of the selection pressures that have once shaped the *lac* operon. In addition, getting to know the phylogenetic history of the *lac* operon and its prevalence across naturally occurring strains of *E. coli* gives a view on the recent evolutionary history of the *lac* operon, and the extent of horizontal gene transfer and gene loss it has been exposed to. Although it is hard to draw unequivocal conclusions from such distribution patterns, knowing more about them could be useful to fuel speculations about the origin of the *lac* operon.

In Chapter Two, I study the phylogenetic history and diversity of the *lac* operon across natural isolates of *E. coli* and divergent clades of *Escherichia*. I compared the phylogeny of the *lac* operon with the whole-genome phylogeny of the divergent *Escherichia* clades, to detect if horizontal transfer of complete *lac* operons has occurred, and performed a phylogenetic bootstrap test to confirm a signature of homologous recombination of part of the *lac* operon between *Escherichia* clades. I also estimated the ratios of non-synonymous to synonymous substitutions for the different genes of the *lac* operon, comparing the signatures of selection between human and environmental lineages.

In Chapter Three, I investigate the genetic basis of natural phenotypic variation in the *lac* operon. To this end, I compared phenotypes conferred by the *lac* operons of human and environmental isolates of *E. coli* and divergent clades of *Escherichia* in a common genetic background. Hoping to find the variation that is relevant for fitness, I looked for systematic differences between *lac* operons of strains of human and environmental origin. In addition, I tried to get a first view on the effect of genetic background on the growth phenotype, comparing growth on lactose of *lac* operon constructs with that of the same operons in their original genetic backgrounds.

2 Chapter Two: Recent history, diversity and selection pressures

Introduction

The *lac* operon in *E. coli* is the classic model system for the study of bacterial gene regulation. Yet, little is known about its phylogenetic history, and the selection pressures it is exposed to in nature.

The *lac* operon was discovered in *E. coli* and has been extensively investigated in that species (Pardee *et al.*, 1959; Müller-Hill, 1996), and lactose metabolism is used as one of the defining characteristics of *E. coli* in some phenotypic tests. However, the *lac* operon is present, completely or partially, in other Enterobacteriaceae as well. Complete *lac* operons have been found in *E. coli*/*Shigella* and some strains of *Citrobacter*, *Salmonella* and *Klebsiella* (see **Appendix 1, Table A.1** for an overview of *lac* operon genes which can be found on Genbank). In yet other species of Enterobacteriaceae, less complete versions of the *lac* operon have been found, often consisting of the beta-galactosidase gene in combination with the repressor and/or the permease (see (Stoebel, 2005) and **Table A.1**). In addition, the *lacZ* and *lacY* genes have been found on plasmids (Guiso and Ullmann, 1976; Cornelis, 1981).

Apart from the classification bias introduced by phenotypic tests for lactose metabolism, the conservation of the *lac* operon across strains of *E. coli* should depend on the selection pressures these strains are exposed to across habitats. Yet, surprisingly many open questions still remain about the ecology of *E. coli* (Blount, 2015).

E. coli is best known as a member of the gut microbiome of humans as well as other vertebrates. Originally, it was postulated that the colon of warm-blooded animals constitutes the main habitat of *E. coli*, and that *E. coli* cycles between hosts through periods of survival without reproduction in water and soil (Savageau, 1983). However, in more recent years, the isolation of “naturalized” *E. coli* from soil and water has been reported (Ishii *et al.*, 2006), and the assumption that *E. coli* does not replicate in the outside environment has been challenged (Ishii *et al.*, 2006; Ishii and Sadowsky, 2008).

Yet, while there is evidence suggesting that strains of *E. coli* subsist in the environment across seasons and years (Ishii *et al.*, 2006), it is not known to what extent these strains replicate in soil in the wild. Under laboratory conditions, environmental isolates of *E. coli* have been shown to replicate in soil at temperatures above 30°C, and to survive without replicating in soil at temperatures below 30°C at varying water contents (Ishii *et al.*, 2006). Since temperate soils rarely reach temperatures over 30°C in nature, this suggests that *E. coli* subsists rather than grows in these environments. However, these results are based on data from only three strains; moreover, no results were reported on growth in surface water, leaving it an open question whether some strains of *E. coli* do reproduce in temperate soils and surface water.

In addition to natural isolates of *E. coli*, genetically more divergent clades of *Escherichia* have been isolated from different environments across the world, which form distinct genetic clusters in multilocus sequence typing (MLST) analysis (Walk *et al.*, 2009). The level of genetic divergence of these clades from *E. coli* is high, with the common ancestor of these clades and *E. coli* being estimated to have lived between 48 and 75 million years ago, and the younger lineages (*E. coli*, CI, CIII and CIV) sharing their last common ancestor between 19 and 31 million years ago (Walk *et al.*, 2009). For this reason, these clades have not been designated *E. coli*, even though phenotypic profiles were found to be highly similar to indistinguishable to those of *E. coli* (Walk *et al.*, 2009).

Based on their locations of isolation, several of these divergent *Escherichia* clades appear to be overrepresented in water and soil (Walk *et al.*, 2009). Comparing the genomes of these environmental clades to other natural isolates, a large set of genes was found to be specific to or highly overrepresented in the environmental group. A similarly large set of genes was found to be specific to or highly overrepresented in human isolates (Luo *et al.*, 2011), suggesting that the two groups are to some extent specialized to the different niches. While the putative environmental clades score for the most part identical to *E. coli* on standard phenotypic tests (Walk *et al.*, 2009), they have been shown to form biofilms more readily, and to replicate at lower temperatures than regular *E. coli* (Ingle *et al.*, 2011). Yet, their optimal temperature for growth is the same as that of *E. coli*, suggesting they still cycle regularly through the human or animal gut (Ingle *et al.*, 2011). This being said, very little is known about what constitutes the main habitat of these clades, and how often they cycle through other habitats.

The *lac* operon is typically assumed to have its main function in the metabolism of lactose, a substance occurring only in mammalian milk (Savageau, 1983). If this is true, the *lac* operon would experience selection only in the mammalian gut. For this reason, we studied the diversity, conservation and phylogeny of the *lac* operon across naturally occurring strains of *Escherichia*.

With this, we aimed to find clues about the habitats these strains cycle through, and thus to raise insight into the ecology of *E. coli* and the genus *Escherichia*. In addition, we aimed to gain insight into the recent evolutionary history of the *lac* operon and the selection pressures that are acting on it in the wild, and to get some clues about its origin and the selection pressures that may have shaped it.

With regard to the 'proximate' question of how *E. coli* acquired its *lac* operon, it has been argued that *E. coli* acquired the *lac* operon through horizontal transfer from another bacterial species (Lawrence and Ochman, 1998), based on the atypical GC-content of the *lac* operon compared to the rest of the *E. coli* genome. However, a later study, comparing phylogenies of *lac* operon genes in 14 Enterobacteriaceae to those of two housekeeping genes, failed to find statistical support for such a transfer into *E. coli*, although transfer events between other species were detected (Stoebel, 2005).

In addition, one may ask what are the selection pressures that have led to the formation of the *lac* operon, and more generally, that have made operons a prevalent mode of gene organization. With respect to the latter question, several hypotheses have been brought forward; most notably, the co-regulation hypothesis (Jacob and Monod, 1961b) and the selfish operon hypothesis (Lawrence and Roth, 1996; Lawrence, 1999). The co-regulation hypothesis states that the operon structure is selected for because it enables genes of which the products are needed under the same conditions to be regulated by the same regulatory element. The selfish operon hypothesis, on the other hand, postulates that physical proximity of functionally related genes is selected for, because it increases the probability of these genes to be horizontally transferred together.

The selfish operon predicts that especially genes encoding non-essential metabolic functions will be often found in operons, since these are likely to be lost or lose fitness by drift during periods of no selection, after which replacement by horizontal transfer can become beneficial once the substance in question is encountered again. If this metabolic function is encoded by multiple genes, which each individually do not confer a benefit, then horizontal transfer of the genes in question only confers a fitness benefit when all necessary genes are transferred together. Assuming that genes in closer proximity are more likely to be transferred together, this would result in a selection pressure for operon structure, since this would maximize proximity and thus the likelihood of successful transfer events. The co-regulation hypothesis on the other hand predicts that essential genes are found in operons more often, simply because co-regulation of essential genes would be expected to be more crucial than for non-essential genes, which would result in stronger selection pressure on co-regulation of essential genes.

Bioinformatic results appear to favor the co-regulation hypothesis over the selfish operon hypothesis. For example, a larger percentage of essential *E. coli* genes was found in operons than of non-essential genes (Pal and Hurst, 2004), and functionally related essential genes have been found to cluster closer together on the chromosome than functionally related non-essential genes (Pal and Hurst, 2004). Moreover, horizontally transferred genes were not particularly likely to be in operons or in newly formed operons (Price and Arkin, 2005). Yet, multiple factors may contribute to selection for operon organization, and not all factors must have played a role for every single operon. Therefore, since the *lac* operon does encode what could be a peripheral metabolic function (since many mammals are likely exposed to lactose only for a short period after birth), it is imaginable that a scenario such as the one described in the selfish operon hypothesis did play a role in its formation.

We explored the diversity and phylogeny of the *lac* operon across 20 natural isolates of *E. coli* and divergent *Escherichia* clades. In particular, we investigated whether the *lac* operon structure and function are conserved across natural isolates of *E. coli* and *Escherichia spp.*; whether the *lac* operon has undergone horizontal transfer across isolates, either partially or in its entirety; and finally, whether selection pressures have been different for the different genes of the *lac* operon, as well as between isolates from different habitats.

Results and Discussion

Prevalence of functional *lac* operons across isolates

We compared the *lac* operons of 20 isolates of *Escherichia*, comprising human and environmental strains of *E. coli* and divergent clades of *Escherichia* (**Table 2.1**). All isolates in our study were found to possess a complete *lac* operon. With the exception of one infant strain (M5) carrying a frameshift mutation, LacZ and LacY of all these operons were functional, as verified experimentally by a phenol red assay (see **Table 2.1**) (Shuman and Silhavy, 2003). In addition, the adjacency, order and presence of the main regulatory elements of the genes of the operon were conserved in all isolates examined.

The *lacA* gene was found in all *E. coli* and *Escherichia spp.* under study. In the two *Escherichia* clade III strains we examined, TW09276 and TW09231, the sequence of this gene differs from the other strains at all positions from residue 197 onwards. These two strains contain a stop codon at residue 207, instead of at residue 204 like the other strains. In addition, the clade V strain, TW09308, has its stop codon in *lacA* two codons earlier than the other strains

Table 2.1: Overview of the strains used in this study.

Strain	**Species**	**Origin**	**Location of isolation**	**Lactose ferm w/acid**	**Lactose ferm w/gas**	***lac* operon sequence similarity to K12** (%)
K12 MG1655	*E. coli*	Human	USA	+	+	
M1	*E. coli*	Human	Mexico	+	-	98.7
M2	*E. coli*	Human	Mexico	+	+	97.4
M3	*E. coli*	Human	Mexico	+	-	99.1
M4	*E. coli*	Human	Mexico	+	+	97.6
M5	*E. coli*	Human	Mexico	-	-	99.7
M6	*E. coli*	Human	Mexico	+	+	98.8
M7	*E. coli*	Human	Mexico	+	+	98.9
SC1_A5	*E. coli*	Watershed soil	USA	+	+	98.8
SC1_F10	*E. coli*	Watershed soil	USA	+	+	98.3
SC1_G8	*E. coli*	Watershed soil	USA	+	+	97.8
SC1_G10	*E. coli*	Watershed soil	USA	+	+	98.4
SC1_H3	*E. coli*	Watershed soil	USA	+	+	98.6
TW10509	*Escherichia* clade I	Human	Guinea Bissau	+	+	94.6
TW15838	*Escherichia* clade I	Freshwater sediment	Australia	+	+	95.1
TW09231	*Escherichia* clade III	Freshwater beach	USA	+	+	91.6
TW09276	*Escherichia* clade III	Freshwater beach	USA	+	+	91.8
TW11588	*Escherichia* clade IV	Soil	Puerto Rico	+	+	92.8
TW14182	*Escherichia* clade IV	Freshwater beach	USA	+	+	92.7
TW15844	*Escherichia* clade IV	Human	Australia	+	+	92.7
TW09308	*Escherichia* clade V	Freshwater beach	USA	+	+	90.3

in our set. To our knowledge, to what extent these differences affect the functioning of the protein has not been investigated. Whether *lacA* is functionally fully conserved across our set of lac-positive strains is thus not clear; however, it is unclear how essential this gene is for lactose metabolism (Roderick, 2005; Marbach and Bettenbrock, 2012).

The fact that all but one of the isolates we examined possessed a complete and functional *lac* operon stands in contrast to other Enterobacteriaceae and even *E. albertii*, where partial or complete deletions of the operon are common and presence of individual *lac* operon genes differs across isolates (**Appendix 1, table A.1**). This pattern is however explained by the fact that beta-galactosidase activity is sometimes used as a criterion for classification as *E. coli* in phenotypic tests for bacterial species classification. The *Escherichia spp.* clades in this study were collected as part of different previous studies, and were reported to be phenotypically similar to *E. coli* based on their profiles on classification systems, some of which test for beta-galactosidase activity (Walk *et al.*, 2009; Luo *et al.*, 2011). Thus, it should come as no surprise that they resemble *E. coli* in their ability to metabolize lactose.

While a classification bias thus could explain why the environmental strains I examined were identified as *E. coli* or *Escherichia spp.*, it does not explain how these strains obtained or retained their functional *lac* operons.

To account for the presence of functional *lac* operons in these environmental isolates, one may invoke several hypotheses. First of all, the *lac* operon may have been under regular selection in these isolates, either because these strains still cycle regularly through the mammalian gut, or alternatively, because the *lac* operon is under selection in soil and/or water environments.

Alternatively, the functional *lac* operons of these strains could have been horizontally transferred from gut adapted strains of *E. coli*. To test whether this is the case, we compared the *lac* operon phylogeny of the 8 divergent *Escherichia spp.* strains to their published whole genome phylogenies (Luo *et al.*, 2011). The two phylogenies were identical (**Figure 2.1**); thus, there is no evidence that the *lac* operons of these strains have been horizontally transferred in their entirety. This suggests that the functional *lac* operons of these strains are the original *lac* operons of these clades, and are thus likely to have been under regular selection in the genetic background of these divergent clades.

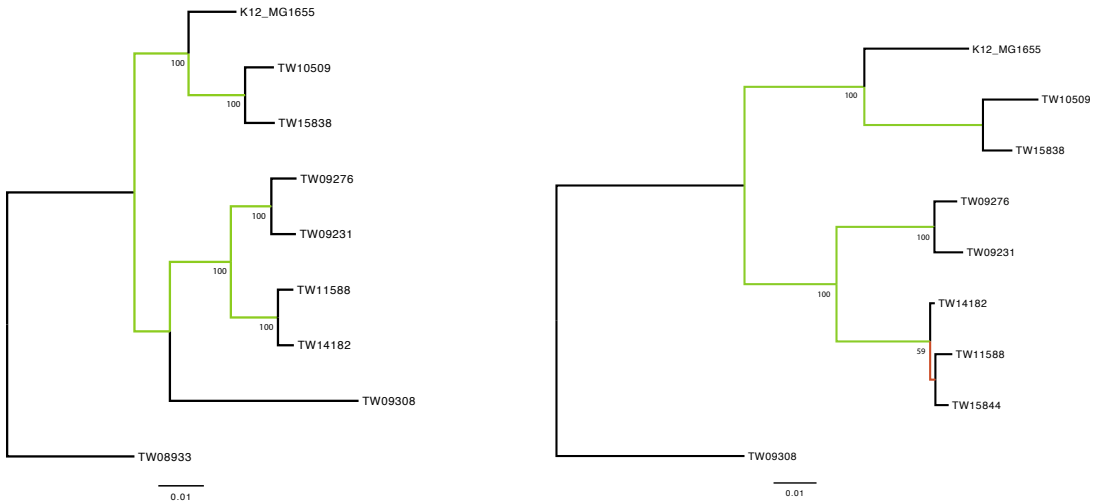


Figure 2.1: **ML trees for nucleotide alignments of the *lac* operon in comparison to genomic data.** (a) Nucleotide alignment of genomic data, of those strains for which these are available, with *E. albertii* strain TW08933 used as an outgroup; (b) Nucleotide alignment of *lac* operon sequences, of those strains for which genomic data are available. In (b), strain TW08933 is not present because it does not have a *lac* operon. Colors correspond to bootstrap support of branches, the percentage of times the nodes clustered together in 1000 repeated runs of the algorithm. Green indicates maximal, red minimal bootstrap support. At values over 50%, exact values are displayed. The scale bar indicates the expected number of substitutions per site per unit length.

There are two main ways this selection pressure on the divergent environmental *Escherichia* clades to retain a functional *lac* operon could be explained. First, assuming that the most common selection pressure on the *lac* operon occurs in the mammalian gut, our data suggest that all isolates under study, including environmental isolates and *Escherichia* clades I, III, IV and V, regularly pass through the mammalian gut environment. This is in line with previous results, showing that while clades I to V can grow at lower temperatures than *E. coli*, the optimal growth temperature of all these clades is still slightly higher than that of the human body (Ingle *et al.*, 2011). Similarly, it is conceivable that growth in soil conditions in the wild is very slow, rendering the number of replication cycles in the gut environment very high relative to those in soil environments, even for environmental strains. If this would be the case, the major part of the selection on the *lac* operon would occur in the gut, for all strains under consideration. Observations of these strains in a lab environment , however, suggest that environmental clades grow under relatively stringent conditions (e.g. at low temperatures) (Erik van Nimwegen, personal communication). This suggests that the first variant of this scenario is the more likely.

Alternatively, it is conceivable that in these clades, the main function of the *lac* operon does not lay in the metabolism of lactose, but of chemically similar sugars occurring outside the human gut. It has been suggested that the original substrate of the *lac* operon is not lactose but galactosyl-glycerol, a sugar which is released upon digestion of plant material in the guts of plant-eating animals (Egel, 1979; Boos, 1982; Egel, 1988). Natural isolates of *E. coli* have been shown to grow on galactosyl-glycerol, with a higher maximum growth rate and lower Michaelis constant than on lactose (Dean, 1995), suggesting that this sugar may indeed be a better substrate for the *lac* operon. In addition, galactosyl-glycerol is a natural inducer of the *lac* operon, unlike lactose, which needs to be converted to allolactose first (Boos, 1982). Since there are many more plant-eating animals than mammals, and except for humans, only infant mammals consume milk, galactosyl-glycerol might be a more prevalent substrate than lactose outside the human gut. Yet, to our knowledge, no data exist on the prevalence of galactosyl-glycerol or other lactose-like compounds in the natural environment. Galactosyl-glycerol does not rely exclusively on the lactose permease, since it can also be transported by a different galactose transport system encoded by the *mgl* genes (Boos, 1982). However, when BLASTing this transport system among the environmental clades, it appears not to be conserved, which might explain the conservation of *lacY* in these lineages.

While the environmental isolates of both *E. coli* and divergent clades of *Escherichia spp.* under study all have a functional *lac* operon, in published genomes of pathogenic *E. coli* and *Shigella* partial and full deletions of the *lac* operon can be found (**Table A.1**). Similarly, in *E.*

albertii and *E. fergusonii*, which are sister species to *E. coli* and have been reported to be frequently pathogenic (Ooka *et al.*, 2012), (Farmer *et al.*, 1985; Mahapatra A *et al.*, 2005), full or partial deletions of the *lac* operon are common. Here, the previously mentioned classification effect is likely to play a role: strains which do not ferment lactose and are pathogenic might be more likely to be classified as *E. albertii* or *E. fergusonii* (in fact, inability to ferment lactose has been reported to be a typical feature of *E. albertii* (Ooka *et al.*, 2012) as well as *E. fergusonii* (Farmer *et al.*, 1985)). In addition, given that pathogenicity comes with the ability to colonize new niches in the human body (Kaper, 2005), it is conceivable that pathogenicity tends to decrease the selection pressure to metabolize lactose, which would not be found outside the intestine. This would predict a negative correlation between the presence of a functional *lac* operon and pathogenicity. To ascertain whether such a correlation exists, more systematic data would be needed. While *Escherichia* clades I-V do exhibit some features associated with virulence, they were avirulent in a mouse model of septicemia and were rarely isolated from extraintestinal sites in humans. For these reasons, they have been hypothesized to be only opportunistic pathogens (Ingle *et al.*, 2011). The conservation of their ability to use lactose may thus indicate that their niche in host species is indeed primarily the commensal one.

Intra-operon recombination

While we did not find evidence for horizontal transfer of the *lac* operon in its entirety across the divergent clades, this does not exclude the possibility that parts of the operon have undergone horizontal transfer. To address this, we screened our alignment for recombination breakpoints using GARD (see Material and Methods). This analysis identified multiple breakpoints. To test whether there is statistical support for any intra-operon recombination having taken place, we performed a phylogenetic parametric bootstrap analysis on the first breakpoint identified by the GARD analysis, which was located 291 bp from the 5' end of *lacY* gene (**Figure 2.2**). This test was highly significant showing that estimating tree topologies independently for the two partitions of the alignment provided a significantly better model fit than when tree topology was constrained to be the same for the entire alignment (see Methods; **Figure 2.3**). This supports the hypothesis that the two parts of the operon sequence have different phylogenetic histories, suggesting that part of the *lac* operon sequence has undergone homologous recombination between different strains of *Escherichia*.

The tree topologies estimated for the alignment upstream and downstream of the breakpoint are shown in **Figure 2.4**. The phylogeny of the alignment upstream of the breakpoint, covering the *lac* repressor gene, *lacZ*, and part of *lacY*, conforms to the published phylogeny for the

Figure 2.2: **The *lac* operon and its repressor gene *lacI*.** *lacZ*, β-galactosidase; *lacY*, galactoside permease; *lacA*, thiogalactoside transacetylase. Gene lengths are drawn to scale. The triangle indicates the main recombination breakpoint, see Results.

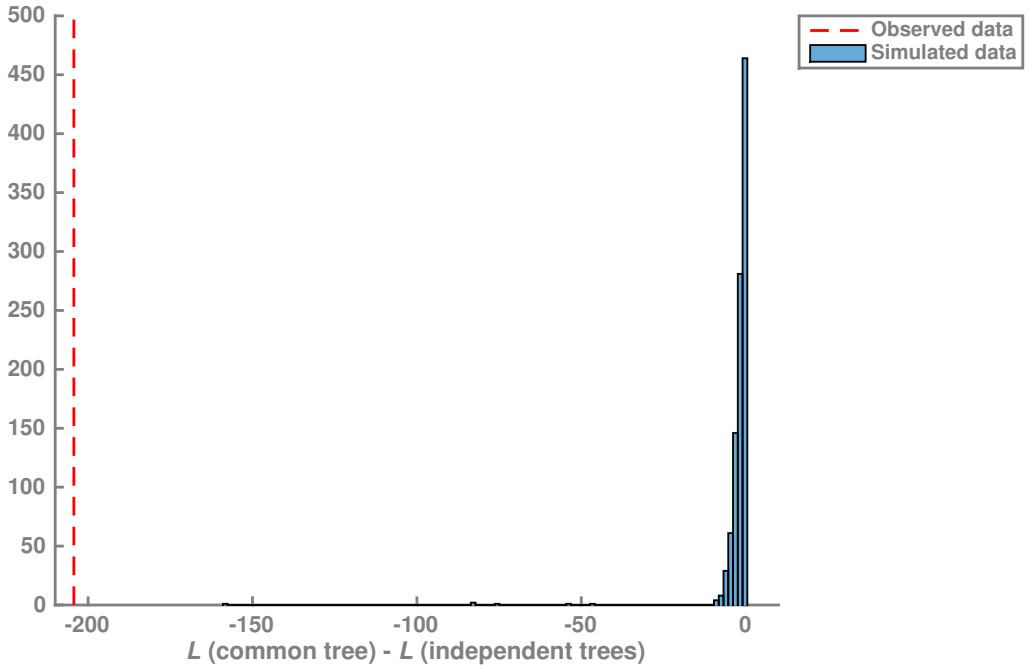
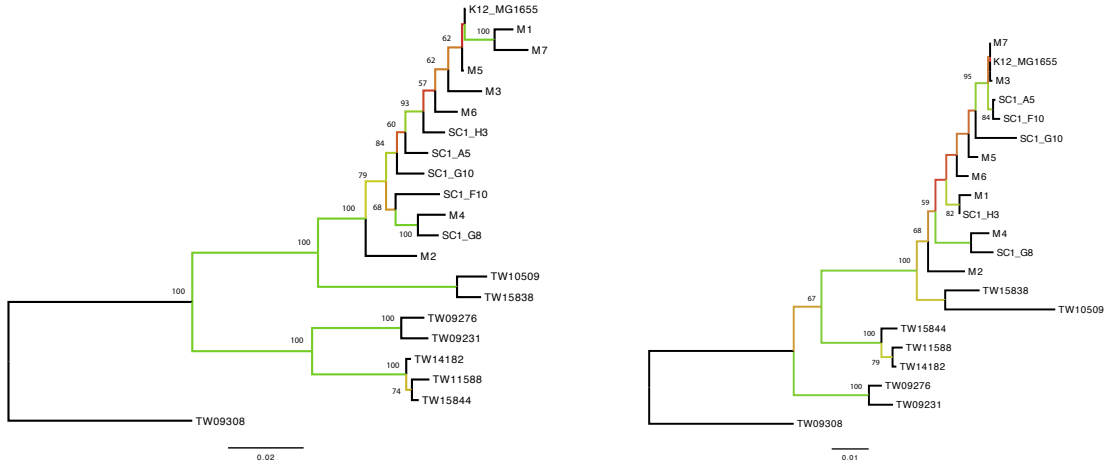


Figure 2.3: **Parametric bootstrap test for recombination.** A parametric bootstrap assesses to what extent a model with two independent tree topologies explains the sequence alignment better than a model assuming a single common tree topology. The log likelihood of our alignment of *lac* operon sequences was lower under a common tree topology estimated for the entire sequence, than under two independently estimated tree topologies on either side of the breakpoint. The difference between these likelihoods was much larger for the true sequence data (red dashed line) than for 1000 sequence alignments simulated under a common tree topology, with the remaining parameters set to the values estimated independently for either sub alignment. This suggests that differences in evolutionary rates (as manifested in the two independently estimated sets of branch lengths) are not enough to explain the apparent differences in phylogenetic history on either side of the breakpoint.

clade I to V strains. The phylogeny of the alignment downstream of the breakpoint, covering the remainder of *lacY* and *lacA*, differs in the placement of clade IV, suggesting that these genes have been horizontally transferred between clades. The horizontal transfer of part of the *lac* operon is congruent with previous work, showing that displacement of individual genes or arrays of genes within operons by horizontally transferred orthologs is not uncommon in bacteria (Omelchenko *et al.*, 2003).



(a) (b)

Figure 2.4: **ML trees for nucleotide alignments of the *lac* operon, upstream and downstream from identified recombination breakpoint.** (a) ML tree for alignment upstream from breakpoint (*lacI*, *lacZ* and part of *lacY*) (b) ML tree for alignment downstream from breakpoint (*lacY* and *lacA*). Colors correspond to bootstrap support of branches, the percentage of times the nodes clustered together in 1000 repeated runs of the algorithm. Green indicates maximal, red minimal bootstrap support. At values over 50%, exact values are displayed. The scale bar indicates the expected number of substitutions per site per unit length.

Signatures of selection

Since the *lac* operons in our study set appear to have diverse evolutionary histories, they might have been exposed to different selection pressures.

To address whether this is the case, we analyzed our alignment for signatures of selection, using the program PAML (Yang, 2007). This program estimates dN/dS ratios, which is the ratio of non-synonymous to synonymous substitutions, corrected for their probability of occurrence. Synonymous substitutions do not affect protein sequence, while non-synonymous substitutions do; thus, a dN/dS ratio below 1 implies that a smaller than expected proportion of substitutions affects protein sequence, which indicates purifying selection. A dN/dS ratio of 1 implies that ratios of synonymous and non-synonymous substitutions conform to their physical probability of occurrence, and thus indicates the absence of selection, whereas a dN/dS ratio significantly above 1 indicates diversifying selection.

The overall dN/dS ratio or set of dN/dS ratios for an alignment is estimated by the program using a maximum likelihood method. In this method, every observation (codon) has a certain likelihood under a model of a given dN/dS ratio or set of dN/dS ratios, at a given phylogeny. The complete sequence then has a likelihood under this model, which is the product of the likelihood of each individual column under that model; this is equivalent to the sum of the log likelihoods of all columns. The program thus searches for the dN/dS ratio(s) under which the

sequence data have the highest likelihood.

Since branch lengths of amino acid phylogenies differ markedly between genes of the operon (**Figure 2.6**), we analyzed each of the operon genes separately, using the amino acid phylogeny for that particular gene.

In addition to estimating one overall ratio of non-synonymous to synonymous substitutions (dN/dS ratio), we fitted two classes of models, site models and branch models (see Methods). In site models, different dN/dS ratios are estimated for the different sites (codons) in our alignment, but each site is assumed to have a single dN/dS ratio across strains. Models are specified based on the total number of distinct dN/dS ratios per alignment, and optionally, whether some dN/dS ratios are constrained to be at or above a certain value; the specific dN/dS ratios, and which sites would fall under which dN/dS ratio are estimated by maximum likelihood. In branch models, different dN/dS ratios are fit to different branches in the phylogeny, while all sites in the alignment are assumed to have the same dN/dS ratio within a branch. Here, models are specified using a branch labelling on the phylogeny.

Models are compared using a likelihood ratio test (see Methods), which can compare only models which are nested in one another. A model is nested in another, more complex model if it contains only a subset of the free parameters (in this case, dN/dS ratios) used in the more complex model. The other parameters are set to a defined value in the simpler model, which lies within the range from which they are estimated in the more complex model. In our case, this means that a lower number of distinct dN/dS ratios is estimated for a given phylogeny. Branch models are nested if no branches are labeled differently in the simpler model which are labeled the same in the more complex model, while the opposite can occur.

Selection pressures by environment

To assess whether whether selection has been acting differently on human and environmental strains, we estimated branch models fitting different dN/dS ratios to human and environmental branches in the phylogeny, first of all assigning the internal branches as human or environmental based on parsimony criteria (**Figure 2.5 (a) and (b)**). In these models, all sites in the alignment are assumed to have the same dN/dS ratio within a branch.

We found that for *lacY*, *lacI*, and *lacA*, branch models did not describe the data better than the null model, which assumes a uniform dN/dS ratio across sites in the sequence and branches in the phylogeny (data not shown). This suggests a uniform dN/dS ratio across branches.

For *lacZ*, branch models fitting different dN/dS ratios to different branches in the *lacZ* phylogeny did describe the data better than a model fitting a uniformly low dN/dS ratio across sites (**Table 2.2**). Whether this difference was significant depended on which internal branches in the tree were labeled as human and which as environmental. Since *E. coli* strains cycle between the gut and the environment, if strain habitat is taken as a trait, frequent trait reversions on the tree are likely. This leaves little confidence in the assignment of this trait to inner branches. For this reason we added a third category for ambiguous internal branches. Models including these three categories fit the data significantly better than the one-ratio model, as well as most of the two-ratio models (see **Table 2.2, Figure 2.5 (c) and (d)**).

Contrary to our expectations, the dN/dS ratio estimated for environmental branches was lower than of human branches. While this is a surprising result, results by Rocha et al. (Rocha *et al.*, 2006) suggest that caution is needed in the interpretation of dN/dS ratios: for closely related bacterial sequences, dN/dS ratios can be inflated due to selection not having had sufficient time to remove weakly deleterious nonsynonymous mutations. Since our phylogeny is composed of relatively closely related strains, particularly those of human origin, our dN/dS estimates might be inflated. In addition, the different levels of divergence between strains may be a factor confounding the dN/dS scores of the different branches in our phylogeny.

To get an idea whether this phenomenon might be playing a role in our data, we fit a free-ratios branch model to our data, estimating a separate dN/dS ratio for every single branch in the tree, and correlated the estimated values with the lengths of the corresponding branches. As predicted by Rocha et al. (Rocha *et al.*, 2006), a negative correlation between branch length and dN/dS ratio was found. However, this correlation was weak (rho=-0.14) and nonsignificant.

As an additional test, we repeated our analyses without the strains which are closely related to at least one other one in the dataset (the seven human infant strains, the five SC1 strains and strain TW11588), using the three category coloring (human, environmental and ambiguous inner branches, **Figure 2.5(c) and (d)**) as described above. Without the closely related strains, the results were still significant, although *p*-values were higher (data not shown).

A related issue pointed out by Rocha et al. (Rocha *et al.*, 2006) is that for bacterial isolates, it is often not possible to distinguish standing polymorphisms, where there is variability at a locus within a species, from substitutions, where all individuals of one species have one variant of a locus, and all individuals in another species have another variant. This is caused by the lack of a straightforward way to demarcate bacterial species, as well as by the fact that bacterial strains isolated from the wild are typically represented by a single sequenced clone, such that there is no picture of the variability within a species. In the latter case, a variant site might

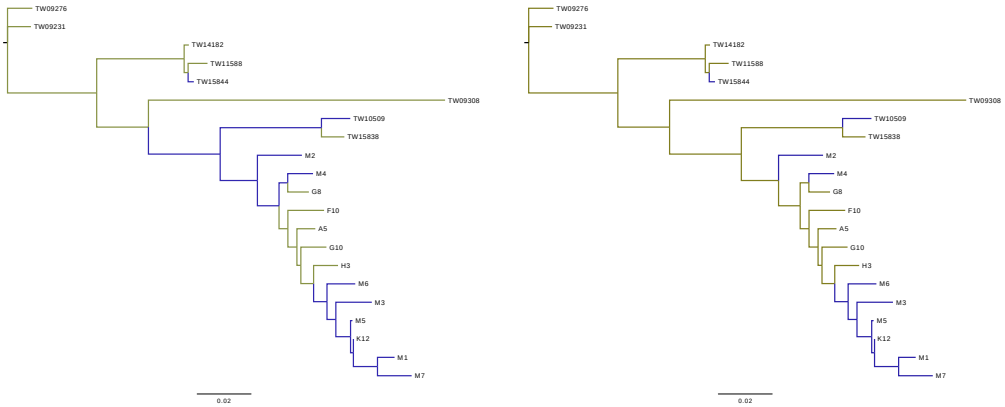
simply be a recent mutation, specific to that clone, which might be purged by selection in a few generations. Thus, there would be more non-synonymous polymorphisms than expected in an alignment, causing an inflated dN/dS ratio.

If using single clones did cause an additional inflation of the dN/dS ratio in our data, leaves (outer branches) in our tree should have a higher dN/dS ratio than inner branches. Indeed, a model in which separate dN/dS ratios were estimated for inner and outer branches (**Figure 2.5(e)**) explained the data significantly better than a model estimating a uniform dN/dS ratio for all branches ($p<0.001$), and this ratio was estimated to be lower for inner than for outer branches (outer branches, dN/dS = 0.20, inner branches dN/dS = 0.12); see **Table 2.2**. When outer branches then were differentiated according to human or environmental origin of the isolates, an additional improvement in fit was achieved (p=0.05). In this case, still, for human branches the highest dN/dS ratios were estimated (human leaves, dN/dS = 0.25; environmental leaves, dN/dS = 0.17; inner branches, dN/dS = 0.12). Thus, correcting for the inflation of dN/dS ratios due to the use of single bacterial clones improved the model fit, but retained the stronger signature of selection for environmental isolates.

In addition to the potential distortion of dN/dS ratios of human and environmental branches due to the different levels of relatedness of the strains in the dataset, the higher dN/dS ratios on human leaves may be a result of the population bottleneck lineages undergo upon colonizing the gut of a new individual. The resulting drop in population size would increase drift, leading to higher dN/dS ratios, since weakly deleterious mutations are removed less efficiently.

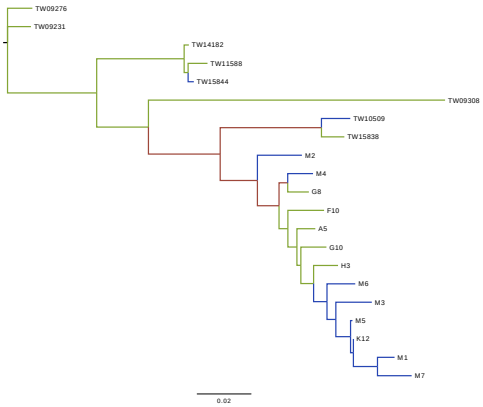
We are not aware of studies reporting on the extent of bottleneck effects in gut colonization of *E. coli*. However, phylogenetic signatures of a population bottleneck have been reported for *Bacterioides* in the gut flora of an infant (Vaishampayan *et al.*, 2010), making it plausible that similar effects exist in *E. coli*. In addition, it has been pointed out previously, based on mostly theoretical considerations, that periodic selection may also be acting as a bottleneck for *E. coli* (Levin, 1981).

Infant guts are sterile before birth (Palmer *et al.*, 2007). Aerobic species of bacteria, which include *E. coli*, tend to be among the first colonizers of the gut, with anearobic species following later (Palmer *et al.*, 2007; Favier *et al.*, 2002; Vaishampayan *et al.*, 2010); time of appearance of *E. coli* varied between studies. Total numbers of bacteria one day after birth, regardless of species, have been estimated to range from 10^4 to 10^{10} based on real-time PCR data (Palmer *et al.*, 2007). In an older study (Mata and Urrutia, 1971), *E. coli* was detected in half of 89 Native American infants on the first day after birth, and in all these infants on the second. In that study, total counts of *E. coli* alone were estimated to range between 10^8 to 10^{11} on the

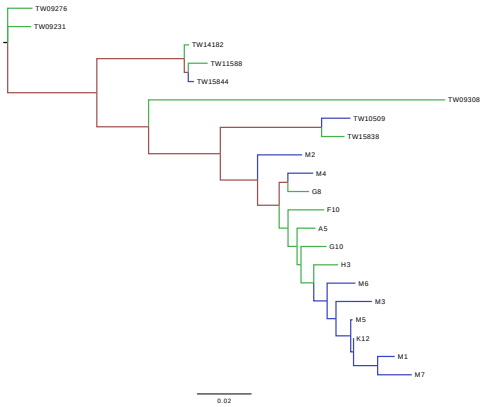


(a) Branch model 1a

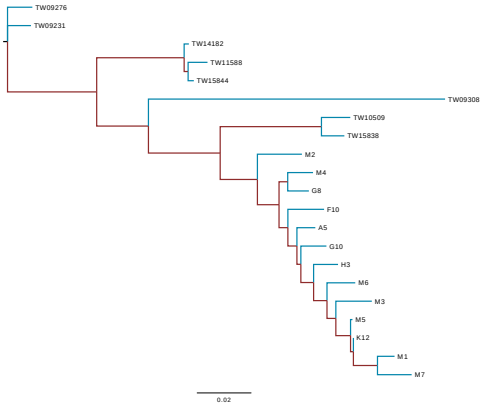
(b) Branch model 1b



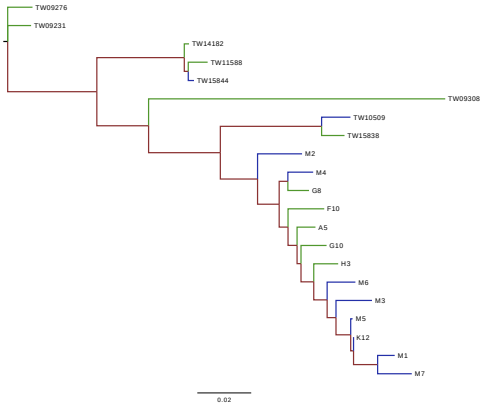
(c) Branch model 2a



(d) Branch model 2b



(e) Branch model 1c



(f) Branch model 2c

Figure 2.5: **Branch labeling as used for the different branch models on the *lacZ* gene sequence. a)** and **b)** Branch models 1a and 1b: 2 ω's are estimated, one for human, one for environmental branches.

c) and **d)** Branch models 2a and 2b: 3 ω's are estimated, one for human, one for environmental, one for internal branches of ambiguous classification.

e) Branch model 1c: 2 ω's are estimated, one for internal, one for outer branches.

f) Branch model 2c: 3 ω's are estimated, one for all internal branches, one for human outer branches, one for environmental outer branches.

Table 2.2: Branch model results on the *lacZ* gene sequence.

Model code	#p	df	*l*	Test statistic	*p*	Parameter estimates
M0	1		-10657.43			ω=0.16
B1a	2		-10656.67			ω(h)=0.17, ω(e)=0.15
B1b	2		-10651.65			ω(h)=0.23, ω(e)=0.14
B1c	2		-10650.56			ω(outer)=20, ω(inner)=0.12
B1a - M0		1		1.52	p=0.2	
B1b - M0		1		11.56	$p<0.001$	
B1c - M0		1		13.75	$p<0.001$	
B2a	3		-10650.46			ω(h)=0.23, ω(e)=0.15, ω(amb)=0.11
B2b	3		-10647.27			ω(h)=0.23, ω(e)=0.17, ω(amb)=0.11
B2c	3		-10648.67			ω(h)=0.25, ω(e)=0.17, ω(inner)=0.12
B2a - M0		2		13.94	$p<0.0001$	
B2a - B1a		1		12.42	$p<0.001$	
B2a - B1b		1		2.38	$p<0.1$	
B2b - M0		2		20.33	$p<0.0001$	
B2b - B1b		1		8.76	$p<0.01$	
B2c - M0		2		17.53	$p<0.001$	
B2c - B1c		1		3.79	p=0.05	

M0: null model, uniform ω for all sites on all branches.

B1a, B1b: 2 w's are estimated, one for human, one for environmental branches; see figs. 2.5 (a) and 2.5 (b).

B1c: 2 w's are estimated, one for inner, one for outer branches; fig. 2.5 (e).

B2a, B2b: 3 w's are estimated, one for human, one for environmental, one for internal branches of ambiguous classification; figs. 2.5 (c) and 2.5(d).

B2c: 3 w's are estimated, one for all internal branches, one for human outer branches, one for environmental outer branches; fig 2.5(f).

first, between 10^5 to 10^{11} on the second day, based on culturing. Yet, these data do not provide direct information on population sizes in the initial inoculum of particular species.

Selection pressures per gene

To be able to distinguish selection pressures across sites in the alignment, we fit different site models to the different genes of the operon, of which the amino acid trees are shown in **Figure 2.6**. In site models, different dN/dS ratios are estimated for the different sites in our alignment, while each site is assumed to have a single dN/dS ratio across strains.

For the permease gene (*lacY*), no site models fit the data better than the null model. The estimated dN/dS ratio was 0.046, implying that all sites in the *lacY* sequence are under purifying

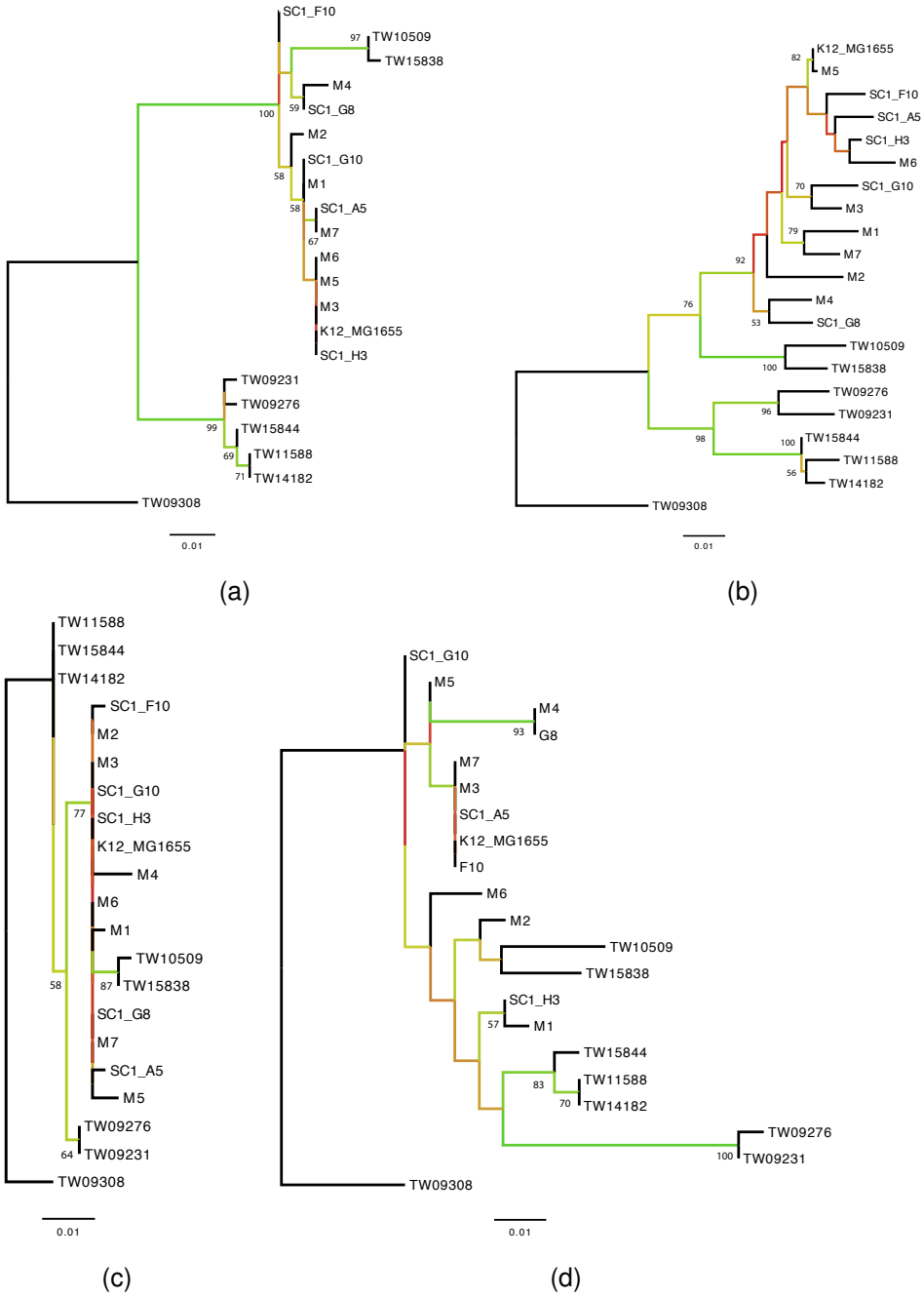


Figure 2.6: **ML trees for amino acid alignments per gene.** ML trees for amino acid alignment of (a) *lac* repressor *lacI*, (b) *lacZ*, (c) *lacY*, (d) *lacA*. Colors correspond to bootstrap support of branches, the percentage of times the nodes clustered together in 1000 repeated runs of the algorithm. Green indicates maximal, red minimal bootstrap support. At values over 50%, exact percentages are displayed. The scale bar indicates the expected number of substitutions per site per unit length.

selection (**Table 2.3**). This is a relatively, but not extremely strong level of conservation, taking into account that reportedly, for bacterial sequences that differ at around 2% of nucleotides and are assumed to be under stabilizing selection, dN/dS values of 0.04 to 0.2 are common (Rocha *et al.*, 2006).

Table 2.3: ***lacY* tests for selection.** Results of site-based tests for selection for the lactose permease gene.

Model code	#p	df	*l*	Test statistic	*p*	Parameter estimates
M0	1		-2986.34			ω=0.046
M1a	2		-2985.16			
M3	5		-2984.54			
M1a-M0		1		2.36	$p<0.1$	
M3-M0		4		3.6	$p<0.1$	

Model code indicates the code used to identify the different site models in the program codeml, see Methods. #p indicates the number of free parameters in the respective site models. df represents the degrees of freedom of model comparisons, obtained by subtracting the number of free parameters of the nested model from the number of free parameters of the more complex model. *l* represents the log likelihood of the alignment under the respective site models. The test statistic given is that of the likelihood ratio test, comparing the likelihood of the alignment under a more complex site model and under a simpler, nested model, see Methods. Parameter estimates are given only for those models which significantly improved the likelihood of the data with respect to the corresponding simpler model.

Model M0 fits a uniform dN/dS ratio (ω) to all sites in the sequence. In model M1a, a variable proportion of sites in the sequence is assumed to be evolutionarily neutral and thus have $\omega = 1$; for the rest of the sites, the dN/dS ratio is estimated. Model M2a also assumes a proportion of neutral sites, and in addition, a proportion of sites for which $\omega>1$ (positively selected sites). In model M3, three values for ω and their proportions are estimated like in M2; however, the values of these ratios are not constrained in any way.

For the repressor gene (*lacI*), as well as the *lacA* gene, site models including a proportion of neutrally evolving sites did describe the data significantly better than the null model for both these genes ($p<0.0001$; **Table 2.4 and 2.5**). This indicates that not all sites in the sequences of these genes are under strong purifying selection. More complex site models, assuming a proportion of positively selected sites, or three freely varying dN/dS rates, did not constitute an additional improvement in fit (**Table 2.4 and 2.5**). Thus, there is no evidence for positive selection on the *lacI* and *lacA* genes within this set of isolates. The estimated dN/dS ratio for the non-neutrally evolving sites was 0.043 for *lacI*; 95% of sites were estimated to fall into this class, with the rest evolving neutrally. For *lacA*, the dN/dS ratio for sites under purifying selection was estimated to be 0.042, and these were estimated to constitute 90% of all sites.

For the β-galactosidase gene, *lacZ*, site models assuming a proportion of positively selected sites in addition to sites under neutral and purifying selection, fit the data significantly better than models assuming only neutral and conserved sites ($p<0.0001$), which in turn fit the data better than the null model postulating a uniform dN/dS rate across sites ($p<0.0001$; Table 2.6). For the model estimating varying proportions of sites under positive (dN/dS >1),

Table 2.4: *lacI* tests for selection.

Model	#p	df	*l*	Test statistic	*p*	Parameter estimates
M0	1		-3014.22			ω=0.0842
M1a	2		-2997.81			ω0=0.0432, p0=0.949
M3	5		-2997.77			ω1=ω2=0.045, ω3=1.12
M2a	4		-2997.77			
M7	2		-2999.47			
M8	4		-2997.79			
M1a-M0		1		32.81	$p<0.0001$	
M3-M0		4		32.89	$p<0.0001$	
M2a-M1a		2		0.08	$p>0.95$	
M8-M7		2		3.35	$p>0.1$	

Results of site-based tests for selection for the *lac* repressor gene. See caption of Table 2.3 for details on models M0 to M3. Model M7 assumes that the value of ω for each site is drawn from a discrete beta distribution with 10 categories; this distribution is specified by two free parameters p and q, which are the only free parameters of this model. Model M7 is nested in model M8, which also estimates a beta distribution, but in addition assumes that a variable proportion of sites has $\omega \geq 1$.

Table 2.5: *lacA* tests for selection.

Model code	#p	df	*l*	Test statistic	*p*	Parameter estimates
M0	1		-1818.53			ω=0.120
M1a	2		-1805.75			ω0=0.0418, (ω1=1), p0=0.896
M2a	4		-1805.75			
M3	5		-1804.85			
M1a-M0		1		25.56	$p<0.0001$	
M2a-M1a		2		0	p=1	
M3-M1a		3		0.36	$p<$=0.95	

Results of site-based tests for selection for the *lacA* gene; see caption of Table 2.3 for details.

neutral (dN/dS =1) and purifying (dN/dS $<$1) selection, estimated dN/dS ratios were 0.044 for sites under purifying selection (90%) of sites), and 3.39 for sites under positive selection (1.9% of sites). Six sites were identified as being under positive selection (**Table 2.6; Figure 2.7**). Variability in these sites did not correlate with strain origin. Moreover, these sites had not previously been reported as important for protein function, specificity, or di- or tetramerization; all sites previously reported to be essential (Juers *et al.*, 2012) did not show any variation in the

sequences we investigated. However, variation in truly essential sites is likely to result in non-functional proteins, against which one would expect strong selection in all cases except those in which the proteins are not expressed or not conferring any benefit or disadvantage.

Table 2.6: ***lacZ* tests for selection.**

Model code	#p	df	*l*	Test statistic	*p*	Parameter estimates	Positively selected sites
M0	1		-10657.43			ω=0.160	
M1a	2		-10421.5			ω0=0.0375, (ω1=1), p0=0.892	
M2a	4		-10400.81			ω0=0.0437, p0=0.899, (ω1=1, p1=0.0813), ω2=3.39 p2=0.0191	59R, **85V**, 116T, 129V, **299K**, **530T**
M3	5		-10400.80			ω1=0.0417,	
						ω2=0.895,	
						ω3=3.17	
M1a-M0		1		471.92	$p<0.0001$		
M2a-M1a		2		41.32	$p<0.0001$		
M3a-M1a		3		41.37	$p<0.0001$		

Results of site-based tests for selection for the beta-galactosidase gene of the *lac* operon; see caption of Table 2.3 for details. Selected sites reported were identified by Bayes' Empirical Bayes analysis; normal font indicates >95% posterior probability, boldface indicates >99% posterior probability.

It must be noted that at the sites identified to be under positive selection, often more than two different amino acids were found at a single site. This would suggest to me that these are evolutionary neutral sites, rather than sites under selection, since it seems to me unlikely that directional selection would be operating in multiple different directions at a single site (however, opinions differ on this, and different amino acids can have comparable chemical properties).

From the results of the site models it can be concluded that while all three operon genes as well as the repressor protein show some divergence, the level of divergence is strikingly different between the different genes. In particular, the permease has been much more conserved than the other three genes, as is apparent when comparing amino acid alignments (see **Figure 2.6**). This suggests that the permease has been under stronger purifying selection than the other genes. This is in line with with previous work on natural variants of the *lac* operon, reporting that variation in permease activity has a larger impact on fitness than variation in beta-galactosidase activity in a chemostat environment (Dykhuizen *et al.*, 1987; Dean, 1989). The permease was inferred to be the limiting step in the flux of lactose metabolism, with small reductions in its activity leading to a lower growth rate, while the beta-galactosidase was inferred

Figure 2.7: **The LacZ protein (one monomer), with the 6 residues identified to be under positive selection drawn as spheres.** Colors indicate the different domains of the protein, following the color scheme of figure 3 of (Juers *et al.*, 2012). Blue: domain 1; green: domain 2; yellow: domain 3; cyan: domain 4; red: domain 5. The active loop is colored purple. Image created using Swiss-pdb Viewer (Guex and Peitsch, 1997), `http://www.expasy.org/spdbv/`

to be present in excess and therefore operating on a fitness plateau.[1] In addition, expression of a functional permease has been identified as the main cost of the *lac* operon in environments containing lactose, likely due to a reduction in the proton motive force across the membrane caused by its proton-lactose symport activity (Eames and Kortemme, 2012). This implies that not only a reduction, but also an increase in the activity of the permease might lead to lower fitness; irrespective of whether this increase is brought about by higher activity per molecule, or by an increase in expression level. This might have led to a narrow range of optimal transport activity, and thus a strong purifying selection on the permease.

[1] Permease activity in this study was inferred indirectly from total lactose flux, after correction for directly measured differences in LacZ activity; variation in this value could thus represent variation in expression as well as in molecular activity of both LacY and LacA in the investigated natural isolates.

2.0.1 Origin of the *lac* operon structure in *E. coli*

From the distribution of *lac* operon genes across *E. coli*, *Escherichia* and other Enterobacteriae, we can try to speculatively infer some clues about the origin of the *lac* operon structure in *E. coli*. Since the *lac* operon confers a metabolic function which is likely to not be continuously under selection, it might be an example of an operon for which the 'selfish operon' hypothesis holds true.

First of all, we found the *lac* operon genes in non-adjacent positions only in combination with traces of insertion sequences (see **Table A.1**). Thus, if separate locations on the chromosome were the ancestral state for the *lac* operon genes, and these genes gradually were selected to be in greater proximity to one another, there are no surviving intermediate forms to testify this.

Second, in other Enterobacteriaceae, frequently only a subset of the *lac* operon genes is found, and the *lac* operon is more patchily distributed (often being entirely absent) in other Enterobacterial species than within *E. coli* and *Escherichia* clades. This pattern does not match the prediction of the selfish operon hypothesis that operon genes only confer an advantage together. Yet, we did find operons with similar sequence to the *E. coli lac* operon, with a deletion at the same position in *lacA*, in several Enterobacterial species, which does suggest the *lac* operon is occasionally horizontally transferred across species. In addition, the signature of partial homologous recombination we found one may speculatively imagine to be a trace of the *lacY* of a strain losing fitness during a period without selection, after which the recombinant variant was fitter during a selection episode. Such gene displacement *in situ* was shown by (Omelchenko *et al.*, 2003) to be relatively common in operons, making it imaginable that such a rescue could happen more often in operons, and that operon organization might promote such events. However, at present this remains a speculation.

Conclusions

The presence of structurally and functionally conserved, but genetically divergent *lac* operons across environmental isolates of *E. coli* as well as divergent clades I, III, IV and V of *Escherichia* suggests that the *lac* operon is regularly under selection in all these lineages. Assuming that the main selection pressure on the *lac* operon occurs in the mammal gut, this corroborates the hypothesis that although overrepresented in the environment, divergent clades of *Escherichia* still regularly pass through the gut of humans or other mammals. However, we cannot exclude the alternative hypothesis that the *lac* operon confers a different advantage in the outside en-

vironment, e.g. metabolizing another substrate such as galactosyl-glycerol. Since very little is known about population sizes and replication times of *E. coli* and *Escherichia sp.* across different natural environments, it is not possible to estimate the time scale at which cycling through the mammal gut, or exposure to lactose-like substrates in the environment, would have to occur to maintain the *lac* operon by selection.

The phylogeny of the *lac* operon as a whole was similar to the phylogeny of the whole genome for the divergent *Escherichia* clades, confirming that the *lac* operons were not horizontally transferred into these clades from other *E. coli* strains. We did find evidence for a horizontal transfer event involving part of the *lac* operon. Taken together, our results thus suggest that homologous recombination happens, albeit infrequently, between *Escherichia* clades.

In their entirety, all genes of the *lac* operon were found to be under purifying selection, in all isolates examined. The signature of selection was strongest for the permease gene, of which all sites were found to be under purifying selection. Six sites in the β-galactosidase gene, *lacZ*, appeared to show a signature of positive selection. However, we suspect these sites to rather have been under relaxed selection.

Surprisingly, environmental isolates showed a signature of stronger purifying selection on the *lacZ* gene than strains of human origin. In addition to limitations of the method of comparing synonymous and non-synonymous substitutions in relatively closely related bacterial strains, this difference might reflect population bottlenecks occurring upon gut colonization.

Materials and Methods

Strains

Strains SC1_A5, SC1_F10, SC1_G8, SC1_G10, and SC1_H3, originally described in (Ishii *et al.*, 2006) were isolated from Lake Superior, Michigan, and kindly shared by Olin Silander. Strains M1, M2, M3, M4, M5, M6 and M7 were originally isolated from 6 month old infants in Morelos, Mexico, as part of a field study (Cravioto *et al.*, 1990) and kindly shared by Armando Navarro and Marjon de Vos. Isolation of these strains took place two weeks before the infants changed diet to solid food. The strains were classified as *E. coli* using the Vitek automated bacterial identification system from BioMérieux (Walk *et al.*, 2009). *Escherichia sp.* strains TW09231, TW09276, TW09308, TW10509, TW11588, TW14182, TW15838, and TW15844 were originally described in a study reporting the existence of divergent clades of the genus *Escherichia* (Walk *et al.*, 2009), isolated previously from different sources and locations. These strains were

ordered from the STEC Centre of the University of Michigan. An overview of the used strains, their origin and location of isolation can be found in **Table 2.1**.

Lactose fermentation test

To detect ability to ferment lactose, a Phenol Red Durham assay was used. Phenol Red medium was prepared consisting of proteose peptone 10 g/l, NaCl 5 g/l, lactose 5 g/l, beef extract 1 g/l, and phenol red 0.018 g/l (Atlas, 2010). This medium was distributed in glass tubes with inverted Durham tubes, and autoclaved for 15 minutes at 121°C. These tubes were inoculated with the strains to be tested from frozen stocks and grown overnight at 37°C. Lactose fermenting strains lower the pH of the Phenol Red broth, causing the medium to turn yellow; non-fermenters of lactose instead increase the pH, causing the broth to turn pink. In addition, gas production is assessed by the inverted glass tubes, in which produced gas bubbles are trapped.

Lac operon sequences

Of strains SC1_A5, SC1_F10, SC1_G8, SC1_G10, SC1_H3, and M1 to M5, DNA was extracted using a MasterPure Gram Positive DNA Purification kit (Epicentre, Madison, WI USA) and the *lac* operon region was amplified by PCR using GoTaq Long PCR Master Mix (Promega Corporation, Madison, WI USA) with the primers CTGGTATCAAACACTCGCCT and ACAACGGGTAGCAAAACAGA, which anneal in the genes *cynX* and *mhpR*, flanking the *lac* operon in strain K12 MG1655. The PCR products were Sanger sequenced by LGC Genomics (see supplement for a list of primers used); both strands were sequenced. Sequences were edited by eye using the Sequencher software (Sequencher version 5.1 sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA).

Lac operon sequences of TW10509 and TW15844 (H605) were downloaded from the antibiotic resistance database of the Broad institute (Escherichia coli Antibiotic Resistance Sequencing Project, Broad Institute of Harvard and MIT, `http://www.broadinstitute.org/`). *Lac* operon sequences of TW09231, TW09276, TW09308, TW11588, TW14182, and TW15838 were downloaded from NCBI.

All *lac* operon sequences were aligned with the *lac* operon sequence of strain K12 MG1655 using the Sequencher software (Sequencher version 5.1 sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA). Alignments were verified and adjusted by eye. LacI, lacZ, and lacY gene lengths were identical for all strains; in strains TW09276 and TW09231,

the intergenic region between lacY and lacA was only 29 bp long, while in all other strains it had a length of 65 bp. Gap placements were adjusted manually to favor longer consecutive gaps over frequent short gaps. After this manual adjustment, all gaps remaining were located in intergenic sequences, and gene boundaries were aligned. For *lacA*, which has a stop codon 2 bp upstream from the other strains in TW09308 and for which the last six amino acids differ from all other strains in TW09276 and TW09231, the gene length of the other strains was kept. Thus, six nucleotides following the stop codon were included in the sequence of TW09308, and the sequence of lacA in TW09276 and TW09231 does not contain a stop codon.
Amino acid sequence alignments were generated based upon the nucleotide sequence alignments using the Geneious software version 7.0.3 created by Biomatters.

Phylogenetic tree estimation

Phylogenetic analyses were performed using the development version of PhyML 3.0 released on October 16, 2013 (Guindon *et al.*, 2010). As a substitution model, HKY85+G (Hasegawa *et al.*, 1985) was used. Tree topology was optimized starting from a BioNJ tree (Gascuel, 1997), which was then stepwise refined using the SPR (subtree pruning and regrafting) search strategy (Guindon *et al.*, 2010). Bootstrap support was calculated for all trees using 1000 replicates.
Phylogenies of the translated alignments of the individual genes of the operon were estimated using PhyML version 3.0 released on April 12, 2012 (Guindon *et al.*, 2010) using the same parameters as for the nucleotide analysis and the LG model for amino acid substitution (Le and Gascuel, 2008). Trees were drawn using the program FigTree v1.4.2
(`http://tree.bio.ed.ac.uk/software/figtree/`).

Whole genome phylogeny

Genome data of *E. coli* strain K12 MG1655 and *Escherichia spp.* strains TW09231, TW09276, TW09308, TW10509, TW11588, TW14182, and TW15838, as well as *E. albertii* strains TW08933 and TW15818 were downloaded from NCBI.
The alignment of the genome data was made using the online tool REALPHY 1.10 (Bertels *et al.*, 2014), using the genomes of *E. coli* strain K12 MG1655 and *Escherichia spp.* strains TW10509, TW11588 and TW09308 as references. The whole genome alignment thus produced had a length of 1275218 nucleotides. This alignment was run in PhyML version 3.0 released on April 12, 2012 (Guindon *et al.*, 2010) using the same parameters as for the *lac*

operon analysis, to produce the whole-genome phylogeny including bootstrap support values for the tree branches.

Recombination tests

To probe whether the *lac* operon is likely to have been involved in horizontal gene transfer events, we compared whether a model including recombination explained our alignment better than a model without recombination. To this end, we first screened the sequence alignment of the operon and repressor gene for potential recombination breakpoints, using the GARD (Genetic Algorithm for Recombination Detection) analysis on the online server available at `www.datamonkey.org` (Kosakovsky Pond *et al.*, 2006). We used the most highly supported recombination breakpoint identified by this screen to partition the alignment into two subalignments, which were then used for a phylogenetic parametric bootstrap.

The aim of the parametric bootstrap test is to assess whether a model with two independent tree topologies (alternative model) explains the data significantly better than the null model assuming a single tree topology. In both the null and the alternative model, all parameters (branch lengths, gamma shape parameter, transition/transversion ratio, and equilibrium base frequencies) were estimated independently for either side of the breakpoint. Thus a partition model is applied in the case of the null model. The test statistic indicates how much better the alternative model explains the data, and is calculated as two times the difference between the log likelihoods of the data under the alternative model and the null model, respectively, as described in (Goldman, 1993). How to quantify the degrees of freedom of this comparison is not known (Goldman, 1993; Huelsenbeck and Rannala, 1997) and thus it is not possible to directly assign a significance level to this test statistic. The parametric bootstrap is a method to nevertheless assign a significance level, by estimating from a large number of simulated sequences how likely the observed data would be under the null model of no recombination (Goldman, 1993).

Using custom software written by J.P.B, 1000 new sequence alignments were simulated using the tree topology and parameters estimated under the null model. These simulated alignments were then subjected to the same maximum likelihood phylogenetic analysis as the original alignment, once under the null model of no recombination and once under the alternative model. The differences between the log likelihoods of each dataset under the null and alternative model together form a distribution of test statistics that are expected if the null model was true. The observed test statistic was compared to this distribution to assess the likelihood that both parts of the observed alignment have the same phylogenetic history.

Estimates of selection per gene

We used PAML version 4.9a (Yang, 2007), (`http://abacus.gene.ucl.ac.uk/software/paml.html`) and its GUI PamlX to estimate selection on the different proteins of the *lac* operon with site-based maximum likelihood methods. Different site models were fit to the nucleotide alignments of the *lacI*, *lacZ*, *lacY* and *lacA* genes, with stop codons deleted and alternative start codons, which occur in *lacI* and *lacA*, changed into the regular ATG. For *lacA*, the last 3 codons of the alignment were deleted since clade V strain TW09308 has a stop codon two codons before the other strains. As phylogenies, the amino acid phylogenies of each of the respective *lac* operon genes were used.

We compared site models M0 (Goldman and Yang, 1994; Yang and Nielsen, 1998), M1a (Nielsen and Yang, 1998; Yang *et al.*, 2005), M2a (Nielsen and Yang, 1998; Yang *et al.*, 2005), and M3 (Yang *et al.*, 2000), and M7 (Yang *et al.*, 2000) and M8 (Yang *et al.*, 2000). Model M0 fits a uniform dN/dS ratio (ω) to all sites in the sequence, and thus contains only one parameter. In model M1a, a variable proportion of sites in the sequence is assumed to be evolutionarily neutral and thus have $\omega = 1$; for the rest of the sites, the dN/dS ratio is estimated. This model thus contains two free parameters: the proportion of neutral sites, and the dN/dS ratio for the non-neutral sites. Model M2a also assumes a proportion of neutral sites, and in addition, a proportion of sites for which $\omega>1$ (positively selected sites). This model contains four free parameters: the proportions of neutral and positively selected sites, the value of $\omega<1$, and the value of $\omega>1$. Model M0 is nested in model M1a, which in turn is nested in model M2a. In model M3, three values for ω and their proportions are estimated like in M2; however, the values of these ratios are not constrained in any way. This model therefore contains five free parameters (the three values of $\omega = 1$, and the proportions of two of them) and models M2a, M1a and M0 are nested in it.

Model M7 assumes that the value of ω for each site is drawn from a discrete beta distribution with 10 categories; this distribution is specified by two free parameters p and q, which are the only free parameters of this model. Model M7 is nested in model M8, which also estimates a beta distribution, but in addition assumes that a variable proportion of sites has $\omega \geq 1$.

Model fits of nested models were compared using the likelihood ratio test, as described in (Goldman, 1993; Yang *et al.*, 2000). The test statistic for this test is calculated as two times the difference between the log likelihoods of the data under the simpler model and the more complex model, respectively. Test statistics were compared to the reference values in a χ^2 table, where the number of degrees of freedom is given by the difference in number of free parameters of the two models.

The comparison of models M7 and M8 and of M1a and M2a are alternative ways of performing a likelihood ratio test of positive selection. The M2-M1a comparison is more conservative than M8-M7 (see PAML documentation). Since our results were always qualitatively similar for both model comparisons, we only report the M2a-M1a values.

In addition to site models, branch models were fit to our alignment. For these models, different ω values are fit to different branches in the alignment. In branch models, different dN/dS ratios are fit to different branches in the phylogeny, while all sites in the alignment are assumed to have the same dN/dS ratio within a branch (strain). Branch models are specified in PAML by labelling branches in the tree file. We distinguished human and environmental branches, and compared several alternative labelings of inner branches (see for more details the results and discussion, **Figure 2.7 and Table 2.6**).

3 Chapter Three: A natural genotype-phenotype map

3.1 Introduction

Evolution by natural selection depends upon the presence of heritable phenotypic variation in nature. If there is no phenotypic variation, there is nothing for natural selection to select; and species can change over generations as a result of selection, only if the selected properties are heritable. This idea has been one of the fundamental components of the theory of evolution, ever since this theory was formulated by Darwin. Yet, Darwin formulated his theory without knowing how the information about heritable traits that is passed on to the next generation is encoded.

After discovery of proteins as the mediators of most fundamental processes in cell, and the gene as the basic unit of information encoding a protein, researchers initially focused on differences in protein coding sequence as the heritable basis of phenotypic variation that natural selection acts on. However, after the discovery of gene regulation by Jacob and Monod in the mid-20th century, this has changed (Jacob and Monod, 1961a). A few years later, Britten and Davidson proposed the theory (Britten and Davidson, 1969, 1971) that major evolutionary changes from simpler to more complex life forms are brought about by changes in the regulation of the expression of conserved genes, rather than by changes in the genes themselves. This idea became more widespread with the finding that human and chimpanzees differed only by 1% in their protein sequences. In an influential paper, King and Wilson (King and Wilson, 1975) suggested that this is too low a number to explain the (in their view) more substantial difference in phenotype between the two species. To explain this finding, they proposed that the differences between these species in morphology/physiology are driven by differences in gene regulation, rather than by differences in coding sequence.

Some authors have argued that in particular change in *cis*-regulatory elements is what underlies most evolution, as opposed to *trans*-regulatory elements or coding sequence (Carroll, 2000; Stern, 2000). The term '*trans*' is used to denote the sequence of transcription factors,

proteins which bind to the DNA and regulate the expression of genes encoding other proteins. *Cis* variation, on the other hand, denotes variation in the regions directly surrounding (or even within) genes that affect their expression, without encoding intermediary agents such as transcription factors (Stern and Orgogozo 2008); typically, these are sequences such as promoter and operator sequences, to which RNA polymerase and/or transcription factors can bind. Variation in these sequences affects the binding properties of RNA polymerase or transcription factors, thereby affecting gene expression.

While gene regulation was discovered in bacteria and initially studied in bacteria and bacteriophages, the question of the relative importance of changes in coding sequence and in gene expression/regulation, and more specifically *cis*- and *trans*- regulatory variation, in evolution has been studied mostly in eukaryotes. Most of the arguments have focused on morphological and developmental changes, which are specific to multicellular eukaryotes (see e.g. (Hoekstra and Coyne, 2007; Stern and Orgogozo, 2008)). In bacteria, the question of the relative importance of coding and regulatory changes in evolution is still more open. Since many of the arguments pertaining to development of morphological structures obviously do not apply, does this mean that (*cis*-)regulatory changes are less important in bacterial evolution?

The *lac* operon is a good model system to explore this question, first of all because it is a relatively independent genetic module, coding for what can be considered a single trait (lactose metabolism/ growth on lactose) which is relevant for fitness under defined conditions. Due to its modular structure, the *lac* operon contains both coding and regulatory sequences. Its regulatory sequences are of a rich variety, containing both *cis* (RNA polymerase binding site, operators, CRP binding site) and *trans* regulatory sequences (the repressor protein sequence). In addition, the *lac* operon was the first gene regulatory system to be discovered (Jacob and Monod, 1961a) and has been a beloved model system of bacterial genetics for more than half a century. For this reason, a wealth of literature is available about the *lac* operon and its components, which have been studied from molecular, biophysical and experimental evolution angles, to name a few (see e.g. (Juers *et al.*, 2012; Ozbudak *et al.*, 2004; Quan *et al.*, 2012)).

For these reasons, we decided to explore the natural diversity of the *lac* operon, on a genetic as well as a phenotypic level, across a set of natural isolates of *E. coli* and *Escherichia*. Comparing genetic and phenotypic variation of *lac* operons of *E. coli* and *Escherichia* clade I, III, IV and V isolates from different environments, we tried to find the genetic basis for phenotypic variation in the *lac* operon.

Finding genetic variation that underlies phenotypic variation, in itself, does not imply that this variation is the result of selection. First of all, the phenotypic variation could be evolutionarily

neutral, and not represent variation in fitness. Moreover, even if this genetic variation would correspond to differences in fitness under laboratory conditions, this does not imply that these fitness differences are relevant in nature: they could be too slight to outweigh the role of genetic drift, or they could be limited to specific conditions in the laboratory. In addition, variation in fitness that is relevant in the wild could be caused by neutral processes, such as relaxed selection and genetic drift.

Yet, one can get an idea which phenotypically relevant genetic variation is a result of selection by looking for variation that correlates with environment of isolation. If genetic variation correlates with environment of isolation, this is likely to be a result of differing selection pressures across environments. To look specifically for variation that is the result of selection, we compared *lac* operons of isolates of human and environmental origin, hypothesizing that these might have undergone systematically different selection pressures in the past.

We expected that environmental isolates, if they represent strains adapted to a lifestyle outside the human gut, might have had less exposure to lactose and thus experienced a reduced selection pressure on the *lac* operon. Alternatively, environmental strains could have been exposed to selection pressures of a different nature (e.g. to grow better at lower concentrations of lactose). If we would find systematic differences in phenotype between human and environmental *lac* operons, we could then try to trace back these differences to particular polymorphisms in the *lac* operon sequence.

Whether environmental isolates of *E. coli* and of divergent clades of *Escherichia* do typically represent strains adapted to an environmental lifestyle, is unclear. As discussed in the first chapter of this thesis, the extent of cycling of *E. coli* and of divergent clades of *Escherichia* through the environment is not known, nor are growth rates and population sizes of *E. coli* in the environment (see e.g. (Ishii *et al.*, 2006; Luo *et al.*, 2011)).

This is not a result of laziness or disinterest among researchers; in fact, it is not trivial to get an unbiased and complete overview of what bacteria are present in soil, and what substances they live on. To find out what bacterial species are present in soil, there are two main methods: culturing, and metagenomics approaches. Estimates are that only 0.1 to 1% of bacterial species are culturable using traditional methods (Daniel, 2005). Thus, culturing only gives a very limited picture of the species present in soil.

For this reason, many researches have turned to metagenomics approaches, which circumvent the culturing step and explore natural diversity on the DNA sequence level. Metagenomics can be applied to search for specific genes or functions, irrespective of what species these genes came from, or alternatively, to get an impression of the species diversity of a commu-

nity. However, these methods often do not give definitive information as to whether a particular species is present, rare or absent in a certain community, since it is hard to access the 'rare biosphere' (Lombard *et al.*, 2011); it has been estimated that to achieve substantial representation of the rarer (less than 1%) species in a soil, on the order of a hundred billion clones would be required (Daniel, 2005). In addition, screening for genes does not tell much about the species these genes came from, since genes can be exchanged between bacterial species by horizontal gene transfer, which plays an important role in gene acquisition in *E. coli* (Ishii and Sadowsky, 2008). Moreover, soil is a structured environment, so where and how is sampled determines which species are found, and the species composition of soil communities can be very different between sampling sites, as well as across time (Lombard *et al.*, 2011).

In this chapter, I aim to elucidate the genetic basis of natural variation in phenotype in the *lac* operon of *E. coli*. I compared growth rates and protein activities conferred by the *lac* operons of human and environmental isolates of *E. coli* and divergent clades of *Escherichia* in a common genetic background. To achieve this, *lac* operons of the strains under study were each cloned into a low copy number plasmid and introduced into a K12 MG1655 strain with a deleted *lac* operon (see Methods).

With this work I had three main aims:

First of all, I tried to trace the sources of genetic variation that underlie natural variation in growth rate on lactose and LacZ protein activity. In particular, I investigated the effect of natural sequence variation in the *lac* promoter region on LacZ protein expression and on growth rate on lactose as a sole carbon source, as well as the amount of covariation between the latter two phenotypes.

In addition, I compared phenotypes across *lac* operons, to assess whether *lac* operons of human and environmental origin differ in fitness. Finding out to what extent the *lac* operons of *E. coli* from different environments differ in phenotype could shed light on the ecology of *E. coli* and *Escherichia* spp. in nature.

Finally, I tried to assess the influence of genomic background on growth on lactose. To this end, I compared growth rates conferred by natural variants of the *lac* operon in a common genetic background to growth rates conferred by the same *lac* operons in their original genetic backgrounds.

3.2 Results

3.2.1 Natural sequence variation in the *lac* promoter region predicts variation in LacZ activity

While the canonical RNA polymerase binding region (the -10 and -35 elements of the *lac* promoter) was conserved across natural isolates, other positions in this region varied (**Figure 3.1**). I analyzed this variation using a thermodynamic model (Tugrul, 2016). This model uses the binding energy matrix inferred by (Kinney *et al.*, 2010), and in addition assumes that RNA polymerase can bind with some probability to not only its canonical binding site, but also with a lower affinity to other sites along a promoter region. The total binding probability is calculated as the sum of all binding probabilities along a region of interest, using a sliding window approach (Tugrul, 2016).

The predictions of RNA polymerase binding affinity of this thermodynamic model, based on sequence variation in the *lac* promoter region, correlated with LacZ protein activity measured by a beta-galactosidase assay at high inducer concentration (**Figure 3.2**). Additional predictions were generated for the CRP binding affinity of this promoter sequence and RNA polymerase binding affinity of the *lacI* promoter region. Neither of these correlated with variation in LacZ activity or growth rate; indeed, the single variable site in the CRP binding site was not predicted to affect affinity (Tugrul, 2016).

The beta-galactose assay was done by growing the cells overnight in LB medium with the chosen inducer concentration, and after two dilution steps in the following morning, when a desired OD is reached, halting growth and lysing the cells; subsequently, the compound ONPG was added, which turns yellow when broken down by the LacZ protein. The absorbance at the wavelength corresponding to this yellow color was then measured in the plate reader for the duration of one hour, and a straight line was fit to this increase in absorbance. The slope of this line corresponds to the total activity of all the LacZ molecules in the well; dividing by the initial OD and the volume of culture added corrects for the amount of cells added to the culture, so that one is left with a measure of the total ONPG metabolizing capacity of LacZ per cell. This capacity is influenced both by the amount of LacZ molecules, as well as by the activity of each individual molecule; thus, the LacZ assay reflects gene expression as well as molecular activity of the LacZ protein.

Since the variability of the data generated with this assay was very high, I repeated the assay at its highest inducer concentration (0.5 mM IPTG) as well as at an additional higher inducer concentration (1 mM) and without inducer, to test whether the model fit was reproducible

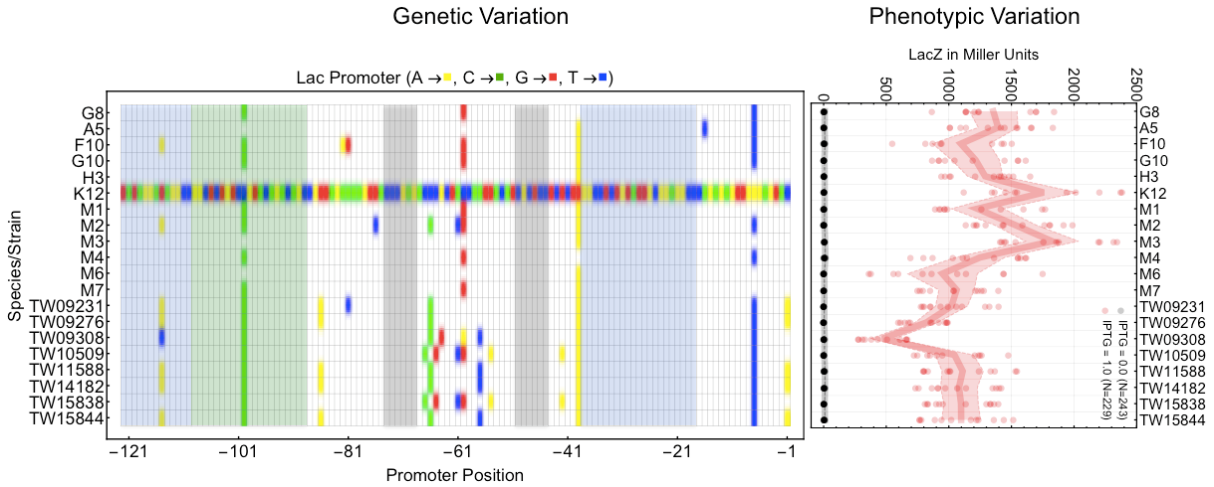


Figure 3.1: **Variation in *lac* promoter region and corresponding LacZ activity scores.** Figure by Murat Tugrul, Figure 3.2 from (Tugrul, 2016), shading added by me. This figure depicts the entire intergenic region between the coding sequences of the *lacI* and *lacZ* genes. Note that the numbering of positions in this figure counts back from the start codon, so the -10 and -35 positions in this figure have different numbers. Grey shading demarcates the -10 and -35 binding sites of the RNA polymerase, which as can be seen are conserved among the isolates I studied. Blue shading indicates operator O1 and part of operator O3 (the rest of this operator overlaps with the LacI coding sequence and is not shown in the figure). Green shading demarcates the CRP binding site. As can be seen, the CRP binding site contains one variable site among the operons I studied, which however was not predicted to affect CRP binding affinity (Tugrul, 2016), see text. Operator O1, the main operator of the *lac* operon, contains no variable sites among my set of isolates, while operator O3 contains three variable sites in total, two of which overlap with the LacI coding sequence. Operator O2 overlaps with the *lacZ* coding sequence and is not shown in this figure (this operator contains one variable site among the isolates I studied; see **Appendix 2, Table B.1**).

and perhaps even better at higher concentrations of inducer (**Figure 3.3 (b) and (d), 3.4 (e), (f) and (g)**). This reproduced the observed correlation at high inducer concentrations, increasing our confidence that this correlation is a result of genetic variation.

The log transformed LacZ activity data of both these runs of the experiment were analyzed together using a linear mixed effect model with IPTG concentration, strain origin (human or environmental) and predicted gene expression as fixed factors, and construct identity and whether the experiment was part of the first or second set of experiments as random factors. All factors except strain origin were found to be significant in this model ($p<0.001$ for predicted gene expression, $p<0.0001$ for the other factors, comparing models with and without each factor using an ANOVA).

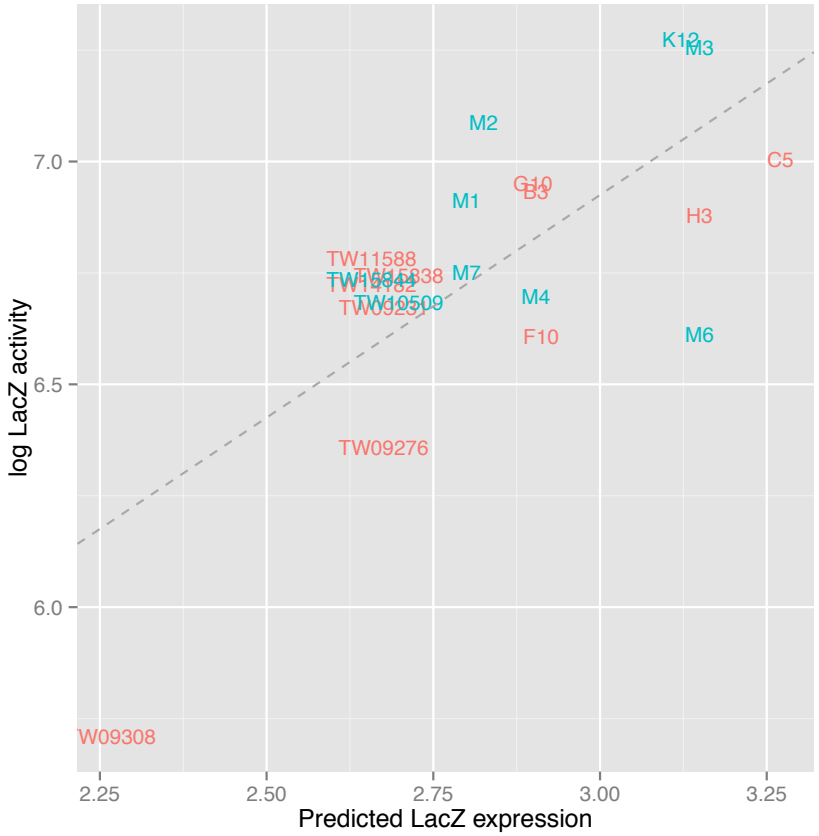


Figure 3.2: **RNA polymerase binding as predicted by a thermodynamic model (Tugrul, 2016) correlated with mean LacZ activity at full induction (0.5 and 1 mM IPTG), measured in a beta-galactosidase assay.** Model predictions (arbitrary units) are plotted on the x-axis; values on the y-axis represent log LacZ activity (log Miller units). Values in this figure represent pooled data of the first and second series of experiments (see text). Blue names denote *lac* operons of human isolates, red of environmental isolates. The dotted line represents the line of best fit of a linear regression of log LacZ activity at the depicted inducer concentrations on predicted gene expression, for illustrative purposes; predicted gene expression was significant as a fixed factor in a linear model including inducer concentration as another fixed factor, and first resp. second series of experiments and construct identity as random factors (see text).

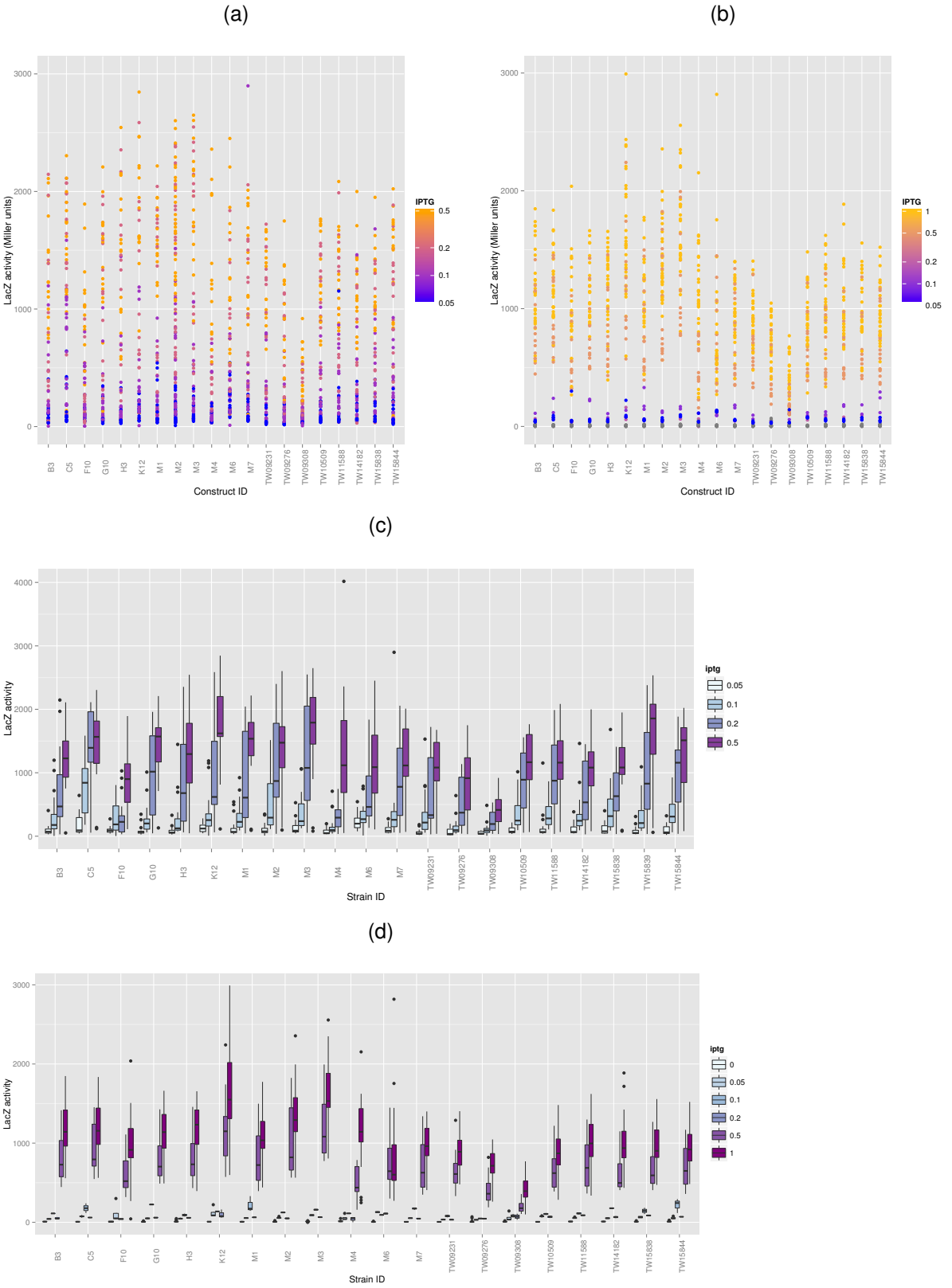


Figure 3.3: **(a) First round of LacZ activity assays. (b) Second round of LacZ activity assays.** Each dot represents one replicate assay. Gray dots represent assays performed without inducer. It can be seen that even though a higher IPTG concentration (1 mM) was included in the second series of experiments, on average lower values were measured at high inducer concentrations in the second, than in the first series of experiments. **(c) First round of LacZ activity assays. (d) Second round of LacZ activity assays.** Same data as in (a) and (b), depicted as box plots. Note the difference in scale between the two figures.

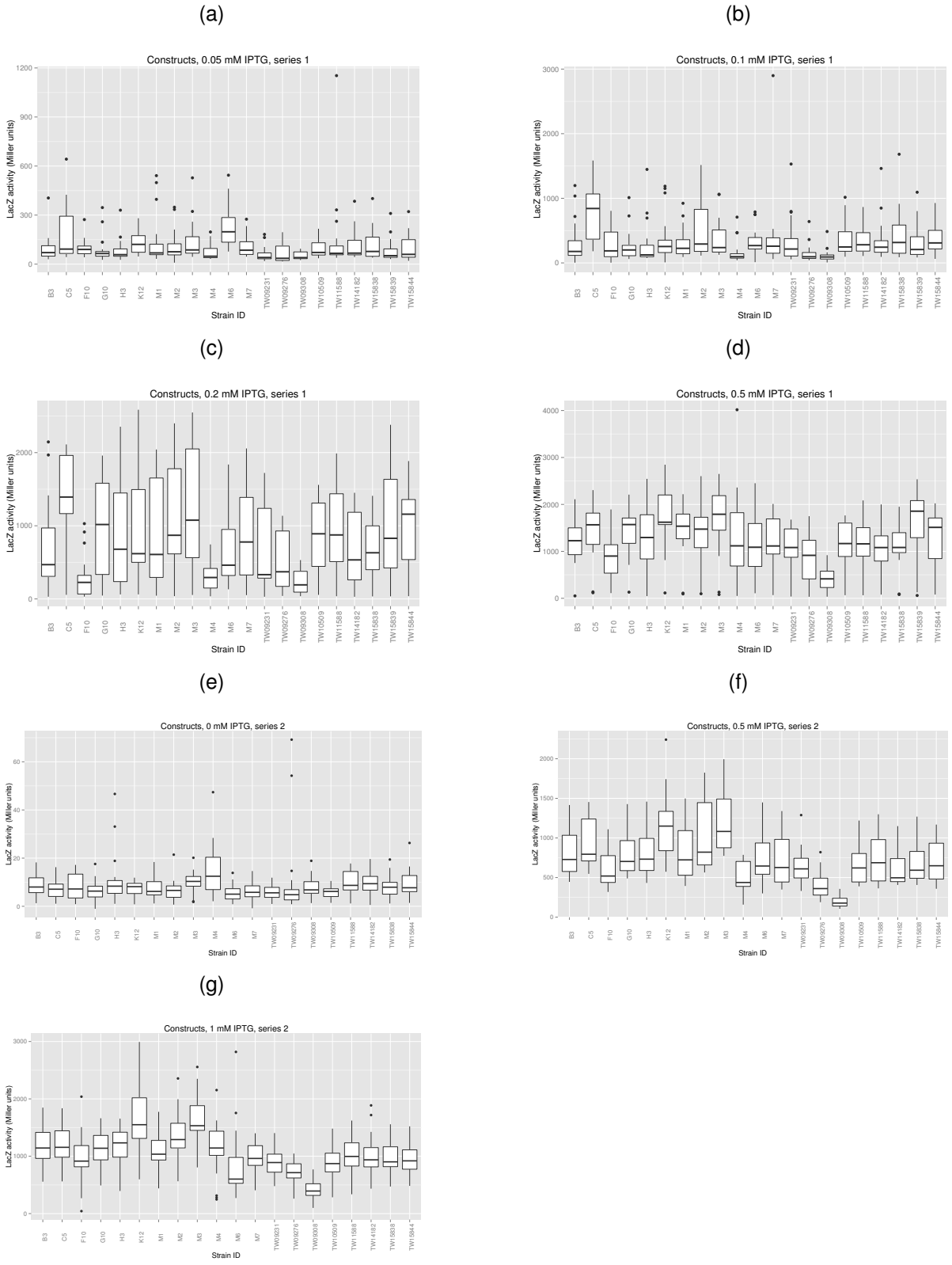


Figure 3.4: **LacZ activity for the *lac* operon constructs.** Each plot depicts LacZ at a different inducer concentration. These figures show the data of Figure 3.2 in separate plots.

3.2.2 Natural variation in LacZ activity does not predict variation in growth rate

Growth rate data of the *lac* operon constructs (**Figure 3.5; see also Figure 3.11**) were analyzed in a linear mixed model with construct identity as a random factor, and lactose concentration, strain origin and gene expression predicted by Murat Tugrul's thermodynamic model (Tugrul, 2016), resp. mean LacZ activity at full induction (in the first resp. second series of LacZ assays), as fixed factors.

Even though as discussed in the previous section, sequence variation in the *lac* promoter region predicted LacZ activity, neither gene expression predicted by the thermodynamic model, nor LacZ activity at high levels of induction predicted growth rates on lactose medium (see **Figure 3.6 and 3.7**), neither as a main effect nor in interaction with lactose concentration. Removing strain origin and predicted gene expression resp. mean LacZ activity as factors from these models did not significantly change model fit (according to model comparison using an ANOVA) and lowered the AIC, indicating that the simpler model is to be preferred. However, note that standard deviations of growth rate as well as LacZ activity measurements were high, precluding the detection of subtle effects.

Correlating growth rate with LacZ activity separately across different concentrations of lactose resp. inducer, LacZ activity at some lower concentrations of inducer was found to correlate with growth rate on low lactose concentrations. This appeared to suggest that the *lac* operons we studied differed from each other in the level of repression they conferred. However, after removing the two extreme data points (M4 and TW09308), which represent two *lac* operons with reduced growth rate and LacZ activity, these correlations disappeared, suggesting this is not a general effect. I analyzed these data together in a linear mixed effect model with mean LacZ activity at low ($<$0.2 mM), mid (0.2 mM) and high $\geq$0.5 mM) IPTG concentration as fixed factors, together with lactose concentration, strain origin and predicted gene expression as fixed factors, and strain identity as a random factor. In this model, no factors except lactose concentration and strain identity were significant. Sequentially removing the non-significant predictors with highest *p*-values from the model resulted in a reduced model in which lactose concentration and LacZ activity at mid and high IPTG concentrations were significant predictors of growth rate, along with the random factor of strain identity. However, when considering the data without strain M4, mean LacZ activities were no longer significant factors.

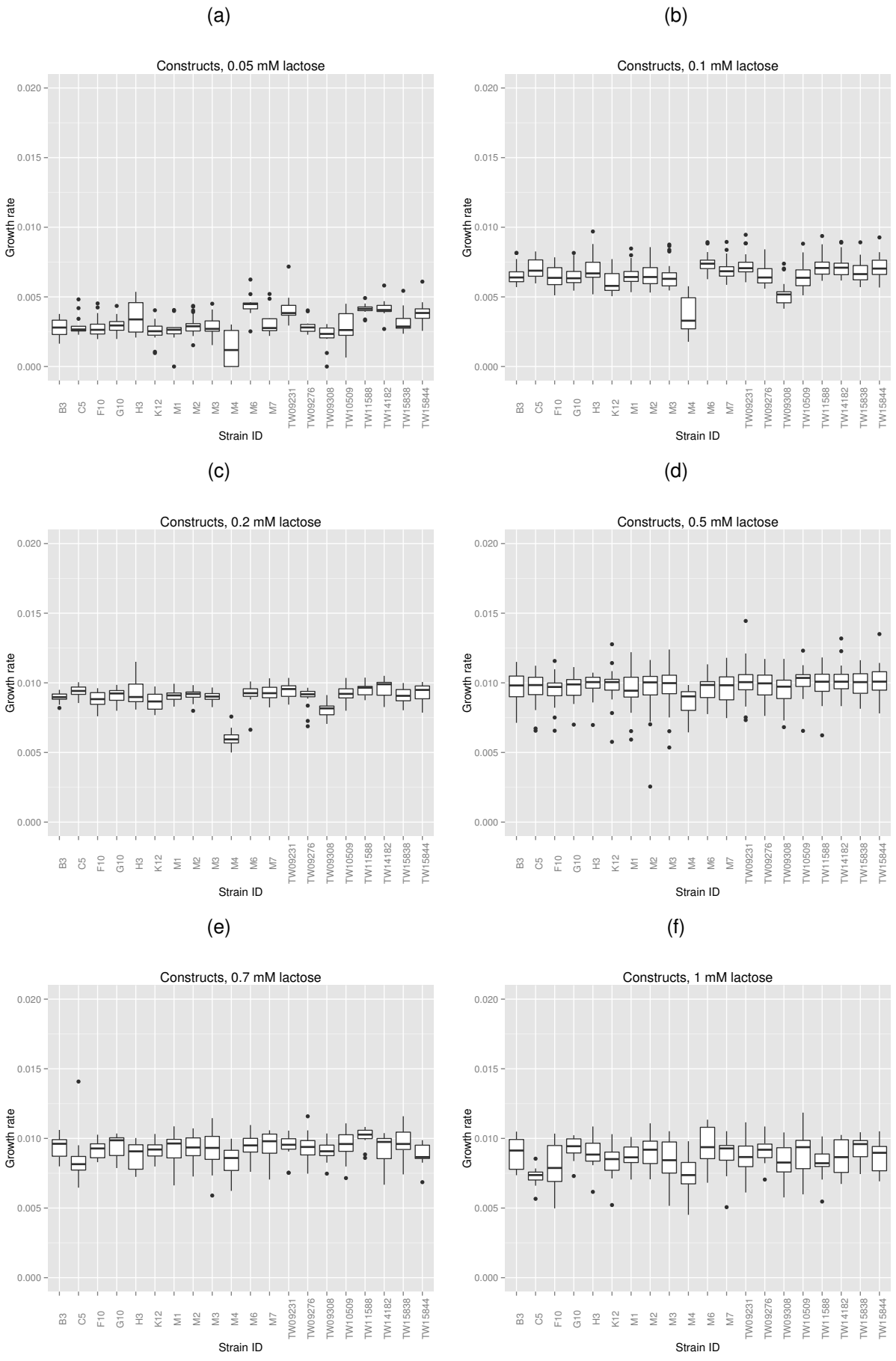


Figure 3.5: **Growth rates for the *lac* operon constructs, in a common genetic background, in M9 medium with lactose as sole carbon source.** Each plot depicts growth rates at a different lactose concentration. These figures show the data of Figure 3.5 in separate plots.

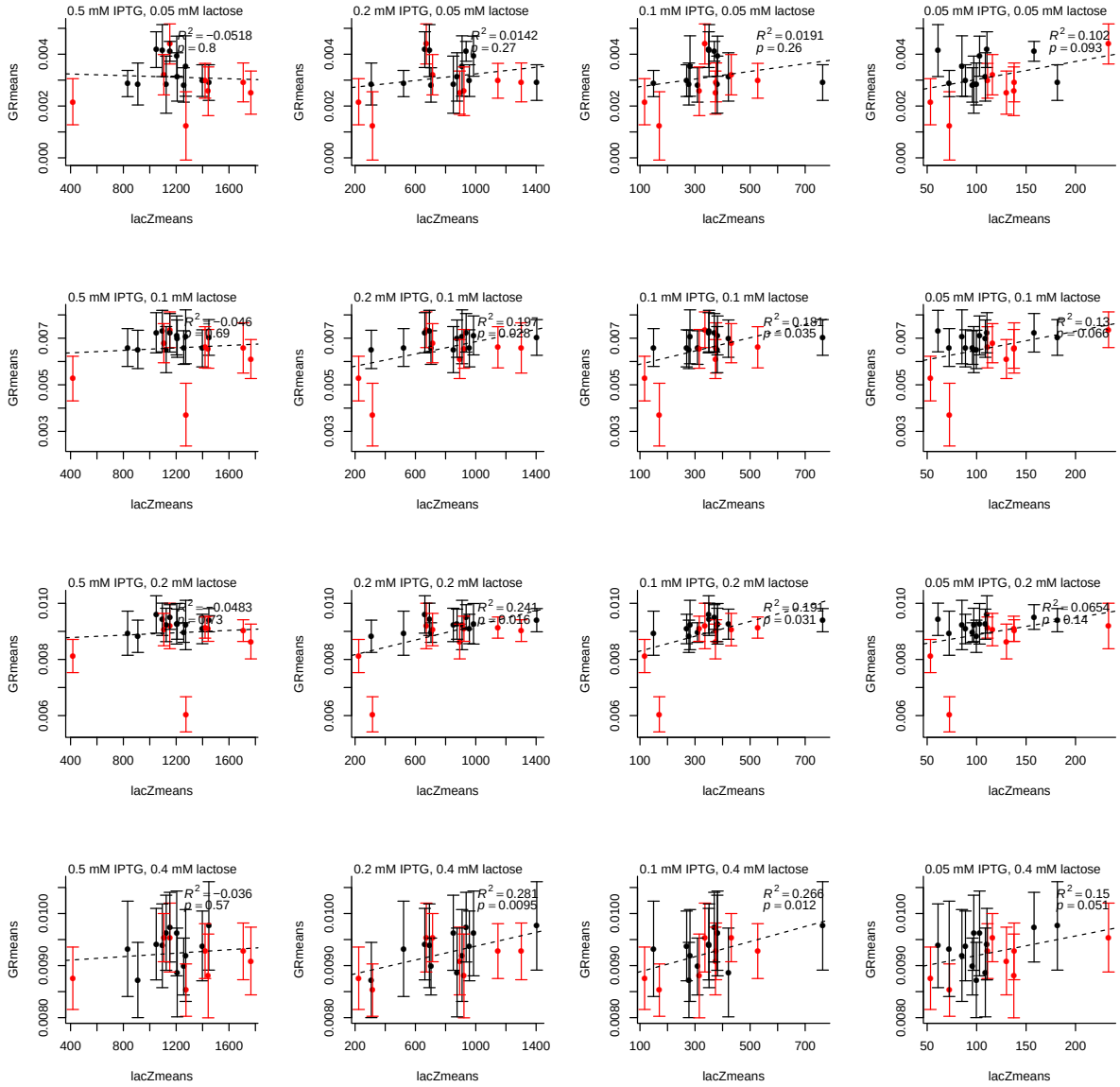


Figure 3.6: **Mean LacZ activity and mean growth rate on lactose medium of the different constructs, at different concentrations of lactose (for the growth rates) resp. inducer (IPTG, for the LacZ activity assays).** While a correlation between these values was observed at some concentrations of lactose and IPTG (middle panels), these correlations did not persist after removing the two most extreme data points of M4 and TW09308. Red dots denote operons of human isolates, black of environmental isolates. Note that the same values appear across multiple plots: each row of plots contains the mean growth rate values for one particular lactose concentration, each column the mean LacZ activity values for one particular IPTG concentration. Error bars represent standard deviations.

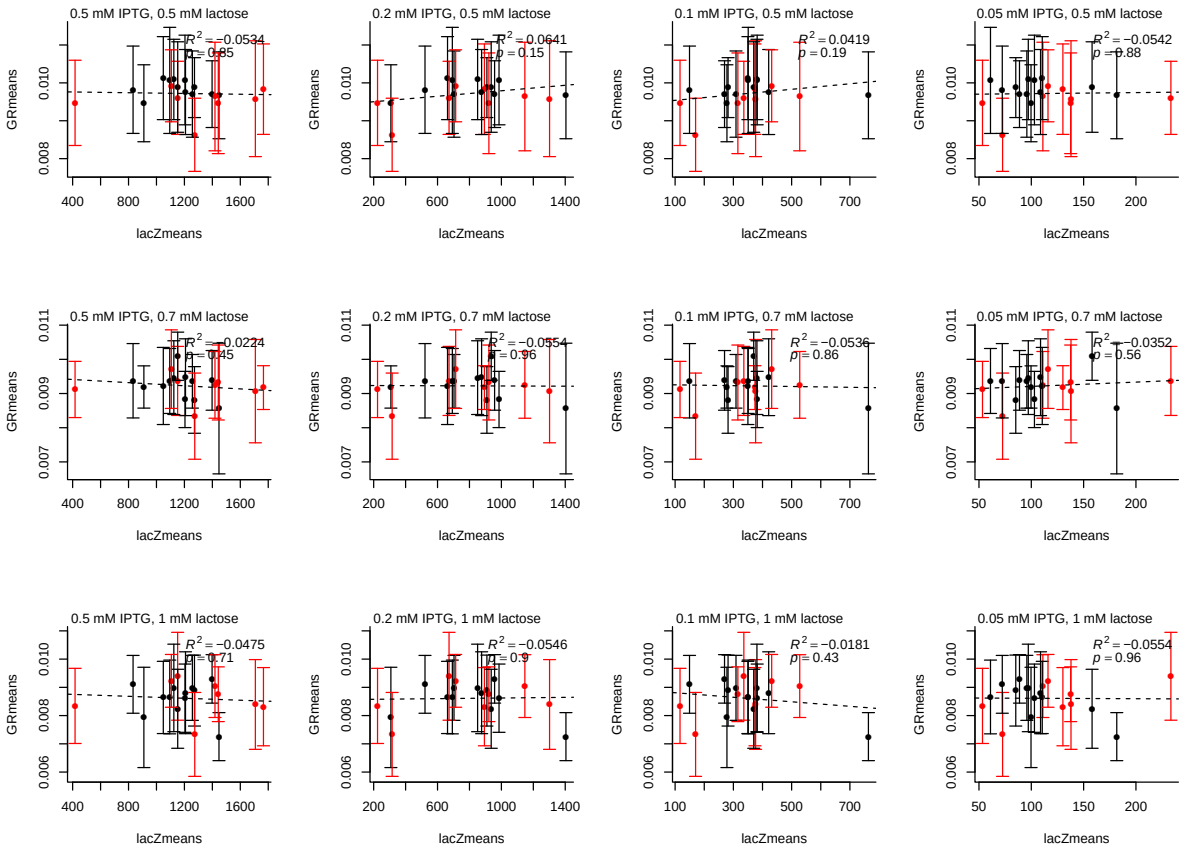


Figure 3.7: **Mean LacZ activity and mean growth rate on higher concentrations of lactose (for the growth rates), and different concentrations of inducer (IPTG, for the LacZ activity assays).** Red dots denote operons of human isolates, black of environmental isolates. Note that the same values appear across multiple plots: each row of plots contains the mean growth rate values for one particular lactose concentration, each column the mean LacZ activity values for one particular IPTG concentration. Error bars represent standard deviations.

3.2.3 Repression levels and growth rate

The level of repression can be calculated by dividing gene expression at full induction by uninduced levels of gene expression Razo-Mejia *et al.* (2014). When using the log transformed LacZ data, this corresponds to the difference between log LacZ activity at full induction resp. at zero induction. Growth rates at different levels of lactose are plotted against these repression values in **Figure 3.8**. While for some concentrations of lactose, a significant correlation between repression level and growth rate was found (**panel (a)**), this correlation was no longer significant upon removal of the data for construct M4 (**panel (b)**).

Correspondingly, when analyzing growth rate data in a linear mixed effect model with repression, origin, predicted gene expression, predicted repressor expression and lactose concentration as fixed factors and strain identity as a random factor, only repression and lactose concentration were significant predictors of growth rate ($p<0.0001$ and $p=0.005$, respectively), and strain identity was a significant random factor. In the simpler model with only lactose concentration and repression as fixed factors, repression was marginally nonsignificant ($p=0.068$). However, repression was no longer close to a significant predictor of growth rate upon removal of strain M4 from the dataset (e.g. $p=0.44$ in the simple model).

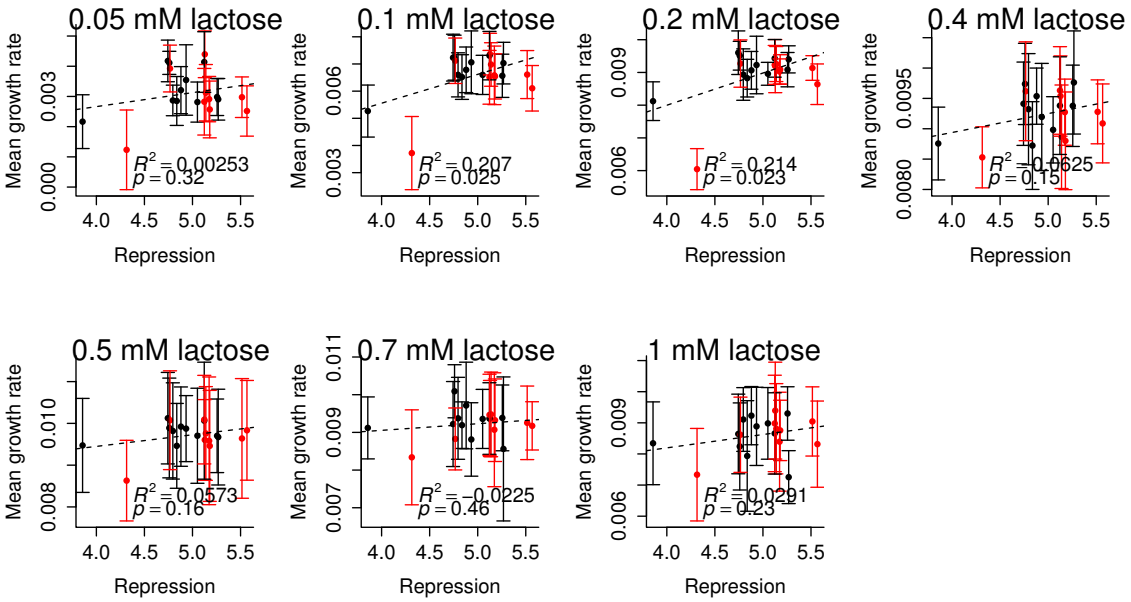
The thermodynamic model of Tugrul (2016) was also used to generate predictions for the expression of the repressor, which is constitutively expressed from a very weak promoter. This predicted repressor expression was no significant predictor of growth rate in any model I tested. Neither did calculated repression level correlate with this prediction (data not shown), although interestingly, construct TW09308 had the lowest predicted repressor expression as well as the lowest calculated repression level.

3.2.4 No systematic differences in phenotype between *lac* operons of human and environmental isolates

Throughout all figures, it can be noted that there are no systematic differences between phenotypes of *lac* operons of environmental isolates, as compared to those of human isolates. This is summarized in **Figures 3.9 and 3.10**. As mentioned at the end of **Section 3.2.1** and in **Section 3.2.2**, strain origin was not a significant predictor in linear mixed models of log LacZ activity, nor of growth rate.

Both for the models for growth rate and LacZ activity data discussed in the previous sections, removing strain origin as a factor did not result in a significantly worse fit and lowered the AIC.

(a)



(b)

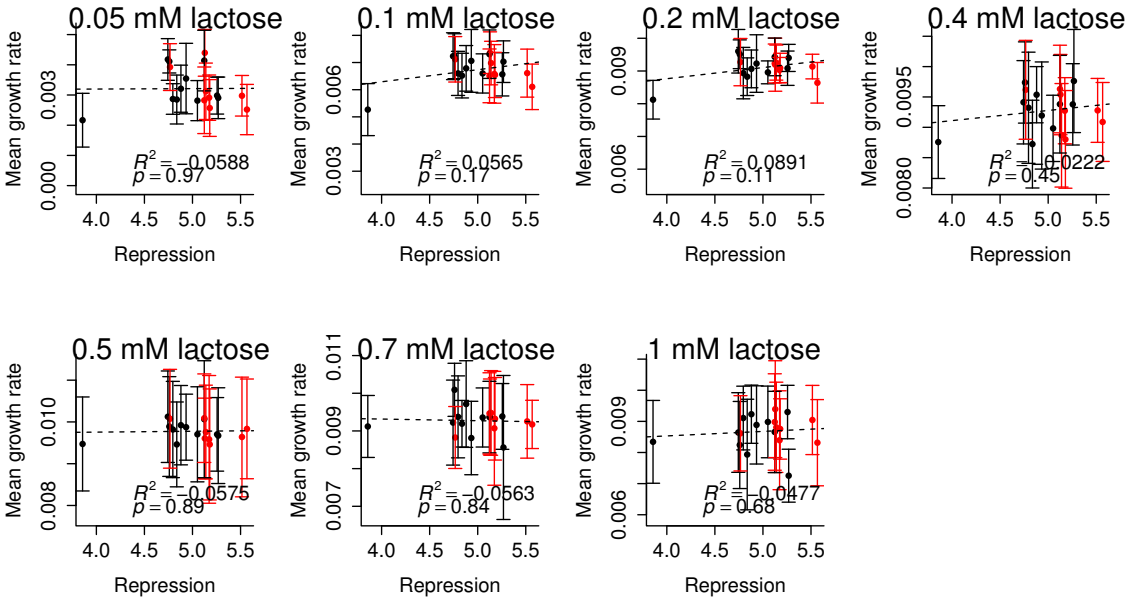


Figure 3.8: **(a) Mean growth rates per construct, at different concentrations of lactose, plotted against repression levels** as calculated from the LacZ activity assays, by subtracting log LacZ activity at zero induction from log LacZ activity at full induction (see text). **(b) Same data as in (a), but without the data point of construct M4.** Without the data for this construct, there are no significant correlations between repression level and growth rate at any lactose concentration.

Thus, there is no reason to believe that that *lac* operons of human isolates confer systematically different LacZ activities or growth rates compared to those of environmental isolates.

What can be noted in **Figure 3.4** is that all constructs of the divergent Escherichia clades (strain names starting with TW) have a lower predicted gene expression, and measured LacZ activity values that fall in the lower part of the range. Among these divergent clades, however, there are also two human isolates, which do not stand out from the other ones with regard to predicted gene expression or measured LacZ activity.

It can also be noted in **Figure 3.2** that the three constructs with highest LacZ activity scores are in fact human isolates (canonical lab strain K12, and infant strains M2 and M3). Whether this is a coincidence or part of a pattern, I do not have enough data points to assess.

3.2.5 Effect of genomic background

To assess whether genomic background affects growth rate on lactose, in addition to or in interaction with the effect of the *lac* operon, we compared the growth rate parameters of our *lac* operon constructs with those of the original isolates from which these *lac* operons were obtained. All *lac* operon constructs were maintained in the same genomic background; for each of the original isolates, the genomic background was different.

As can be seen in **Figure 3.11**, wild isolates tended to grow faster in minimal medium with lactose. (Note that the K12 construct and original isolate grow very similarly, suggesting that this does not have to do with the antibiotic in the medium to maintain constructs, nor with plasmid copy numbers.)

To quantify the pattern of growth on lactose beyond maximum growth rate, we fit Monod's model to our growth rate data.

The performance of a *lac* operon on different concentrations of carbon source can be described, broadly, by the maximal growth rate it confers on abundant levels of carbon source, its ability to grow on small trace amounts of carbon source, and the shape of the relation between carbon source concentration and growth rate. Monod's model describes growth on different lactose concentrations using only two parameters, according to the formula $\mu = \mu_{max}\frac{S}{K_s+S}$. Here, μ_{max} represents maximal growth rate attained at non-limiting substrate concentrations, and K_s represents the substrate concentration at which half maximal growth is reached (see **Figure 3.12**).

To describe differences in growth rate between the different natural *lac* operons, we fitted the Monod equation to our growth rate data for different values of S up until 0.5 mM, independently

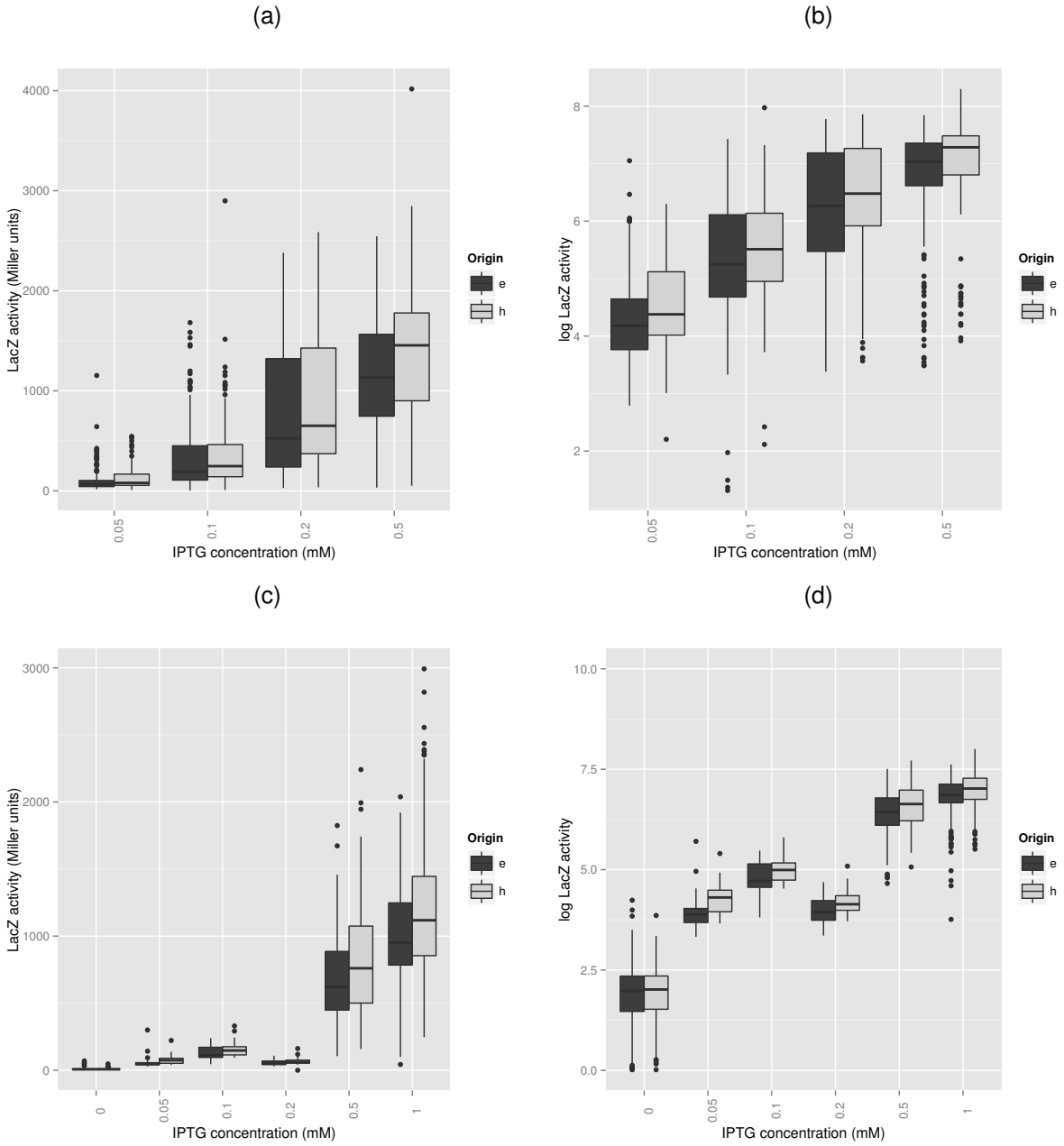


Figure 3.9: **LacZ activity by *lac* operon origin (human or environmental).** (a) First series of experiments. (b) First series of experiments, log transformed data. (c) Second series of experiments. (d) Second series of experiments, log transformed data.

estimating the μ_{max} and K_s parameters per construct. At higher concentrations of lactose, the Monod relation broke down as growth rates decreased slightly with increasing substrate concentration, a phenomenon which has been described before for growth on substrates that become toxic at high concentrations (Andrews, 1968). For this reason, we did not include growth rates for lactose concentrations over 0.5 mM in our model fitting [1].

[1] Andrews (Andrews, 1968) published a formula incorporating the drop in growth at higher substrate concentrations; however, this formula contained three parameters, which was too many to fit it on to the number of independent

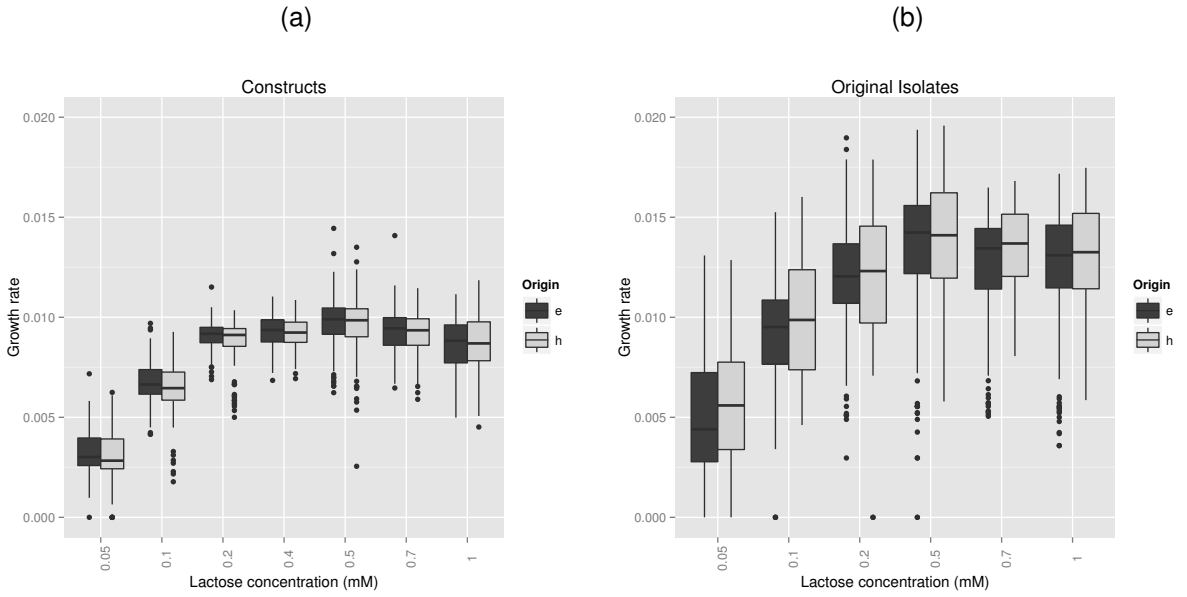


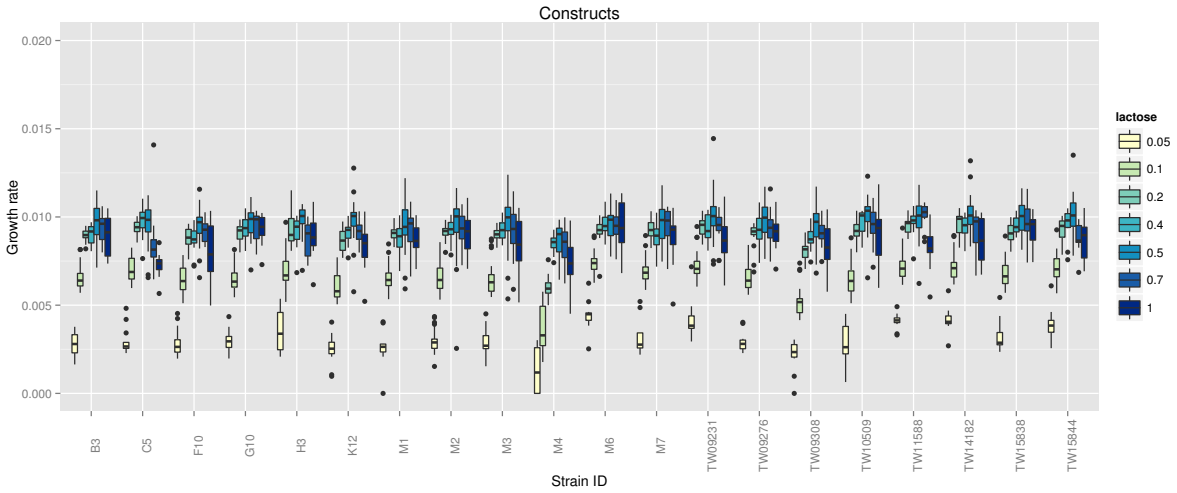
Figure 3.10: **Growth rate by *lac* operon origin (human or environmental).** (a) Growth rate data of *lac* operon constructs. (b) Growth rate data for the corresponding original isolates (i.e. the same *lac* operons, in their original genomic backgrounds).

Fitting Monod's model to my growth rate data yielded strongly correlated estimates of the μ_{max} and K_s parameters. While initially seen as a reason for excitement and thought to reflect a trade-off between growth on high and low substrate concentrations, this correlation was shown by Georg Rieckh to be an artefact of the model fitting. This was caused by a lack of parameter identifiability, meaning that unique estimates of the parameters could not be obtained due to nature of the function being fitted (see (Grady *et al.*, 1996; Robinson and Tiedje, 1983; Robinson, 1985)). This happens when the best fitting value for one parameter depends upon the value estimated for another parameter to be fitted, leading to correlated parameter estimates and fit errors. Following Georg Rieckh's suggestion, we tried to improve parameter identifiability by using a modified Monod model which included a Hill coefficient, n: $\mu = \mu_{max} \frac{S^n}{(K_s)^n + S^n}$. The remainder of this part of the project was done in collaboration with Srdjan Sarikas.

This three parameter model proved too complex compared to the number of independent measurements in my data set, making a direct fit impossible. To get around this, we assumed that the optimal Hill coefficient *n* would be the same for all constructs. Setting *n* to this optimal value would leave the model with two free parameters to fit again. To obtain this optimal value for *n*, I fit the model with different values of *n*, hoping to find a value for *n* for which the cross correlation between the two fit parameters μ_{max} and K_s (the off-diagonal element in the 2x2 correlation matrix) would be minimal (see **Figure 3.13**). By minimizing this cross correlation, we

data points I had.

(a)



(b)

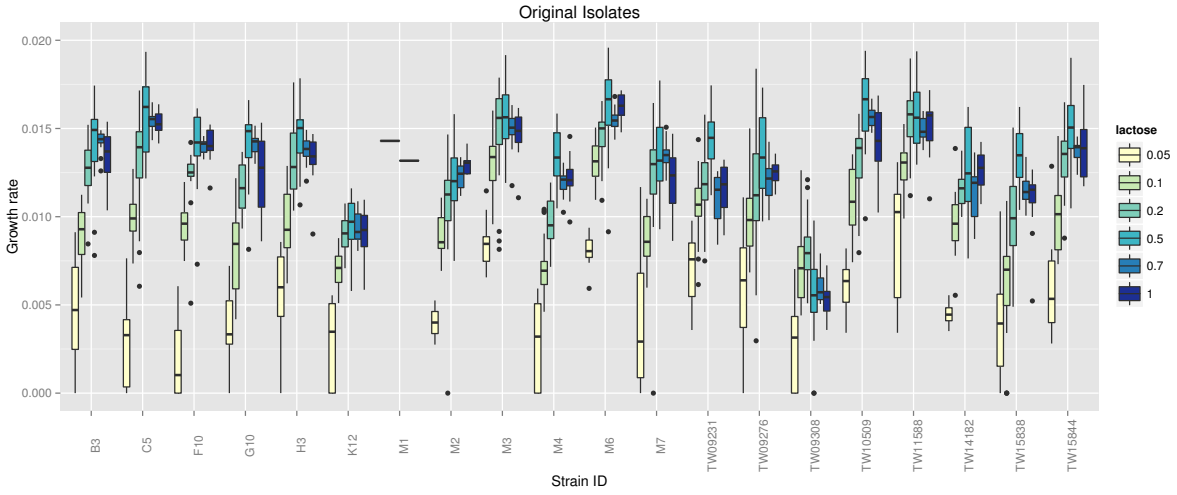


Figure 3.11: **(a) Growth rates for the *lac* operon constructs, in a common genetic background, in M9 medium with lactose as sole carbon source. (b) Growth rate in the same medium (minus antibiotic) for these same *lac* operons in their original genomic backgrounds.**

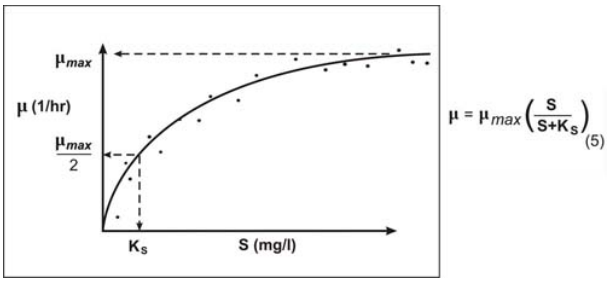


Figure 3.12: **Monod's relation between substrate concentration and growth rate.** Figure by A. Cunningham, Center for Biofilm Engineering, Montana State Univeristy, Bozeman, MT.

aimed to minimize the interdependence of the values for best fitting parameters. However, the correlation of fit parameters decreased monotonically with increasing Hill coefficient (**Figure 3.14**). Thus, aiming for a minimal cross correlation alone did not suffice as a criterion for choosing a Hill coefficient.

To find the optimal value for the Hill coefficient, we then calculated the sum of squared differences between predicted and measured values, χ^2, of the fits ($\sum_{i=1}^{n} \frac{(\mu_i^{measured} - \mu_i^{predicted})^2}{\tau_i^{measured}}$, where $\mu_i^{measured}$ is the measured growth rate at substrate concentration i, $\mu_i^{predicted}$ is the growth rate at that substrate concentration predicted by the model, and τ_i is the standard deviation of the measured growth rates at that particular substrate concentration) for a range of different Hill coefficients. Average χ^2 across strains and constructs was minimal for a Hill coefficient of 2, meaning this value minimized the error of the fits. For this reason, we decided to use a Hill coefficient of 2 for all our subsequent Monod model fitting.

Growth rate parameter estimates for the Monod model with the optimal Hill coefficient of 2 are plotted in **Figure 3.15** for constructs together with those of the corresponding original isolates, surrounded by error clouds representing 95% confidence intervals. Estimates of each of the μ_{max} and K_s parameters separately, corresponding to projections of the points and their error clouds on the respective axes, are plotted in the flanking panels.

In these plots, several things become apparent. First of all, parameter estimates for construct K12 in the genomic background of K12 Δ*lac* overlapped with parameter estimates for the unmodified K12 strain, both for the μ_{max} and K_s parameters. This is important as a sanity check, assuring us that parameter estimates are reproducible, and that any differences in copy number of the *lac* operon between constructs and original isolates are not large enough to skew our results.

Secondly, it can be noted that estimates for μ_{max} are almost universally higher for our *lac* operons in their original genomic backgrounds. Exceptions are infant *E. coli* strain M1, which

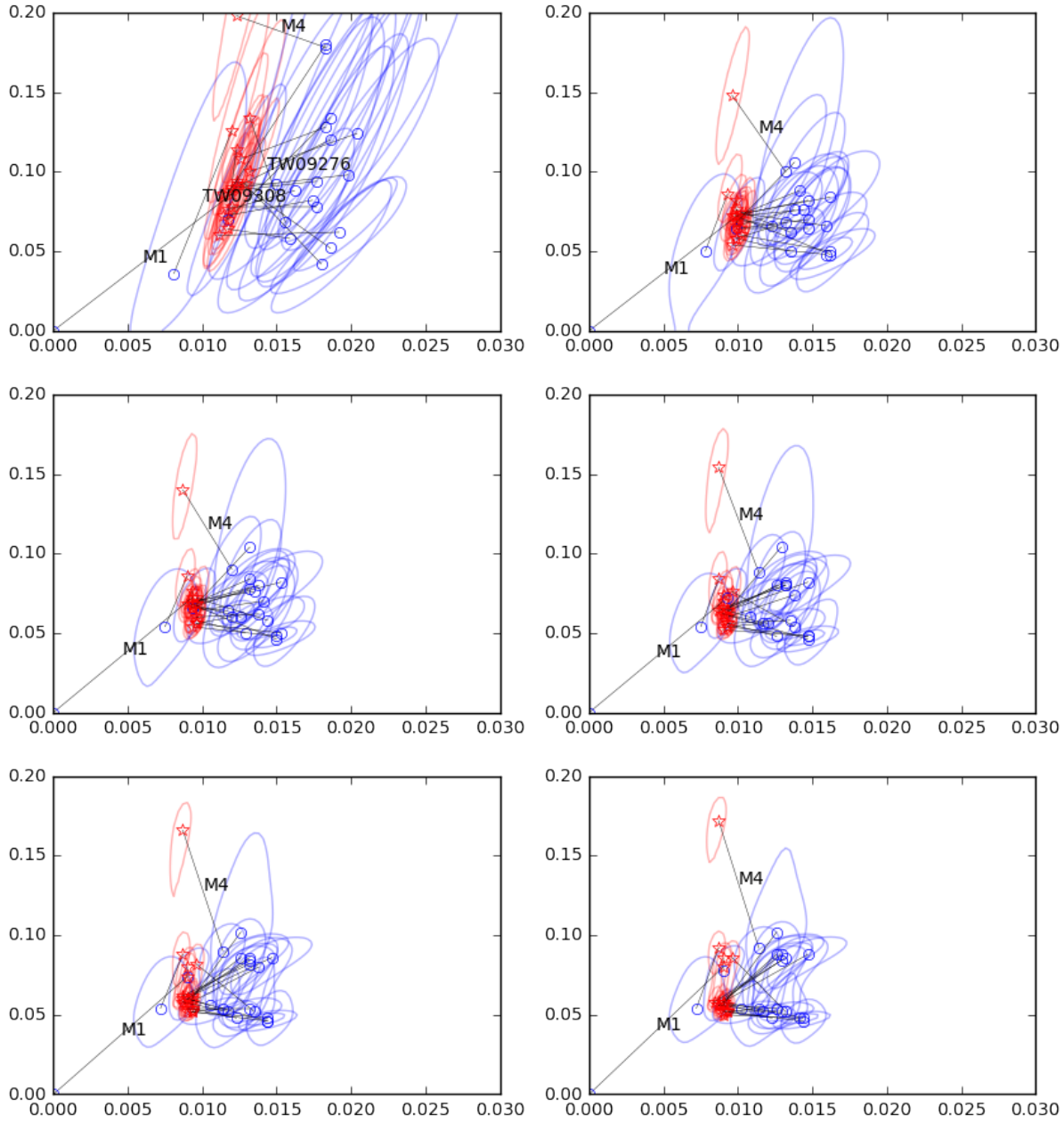


Figure 3.13: **Fit parameters for constructs (plotted in red) and original isolates (plotted in blue), with error clouds, for increasing values of the Hill coefficient** (top left panel: Hill coefficient of 1; top right: Hill=2, second left from the top: Hill=3, and so on). Fit parameter values for μ_{max} are plotted on the x axes, fit parameter values for K_s on the y axes. As can be seen, parameter identifiability is very low at a Hill coefficient of 1, which corresponds to the original Monod model, and improves with increasing values of the Hill coefficient. We chose a Hill coefficient of 2 to fit our growth rate data, because this minimized the fitting error (see Methods). Figure by Srdjan Sarikas.

did not grow in the minimal medium we used for our experiments, and *Escherichia* clade V strain TW09308, for which growth rate of the original isolate was on average lower than of its corresponding construct, although confidence intervals overlapped.

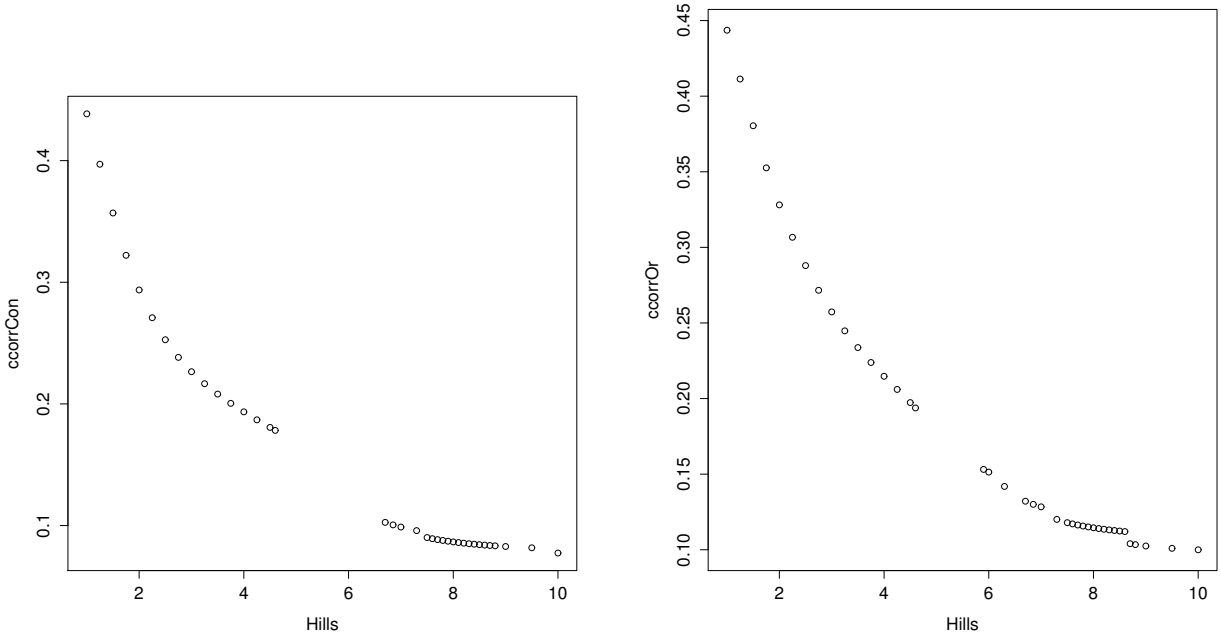


Figure 3.14: **Cross correlation of fit parameters for increasing values of the Hill coefficient, for growth rate data of *lac* operon constructs (panel a) and original isolates (panel b).** Gaps in the figure correspond to Hill coefficients for which the fitting did not converge (nonlinear least squares fitting using R).

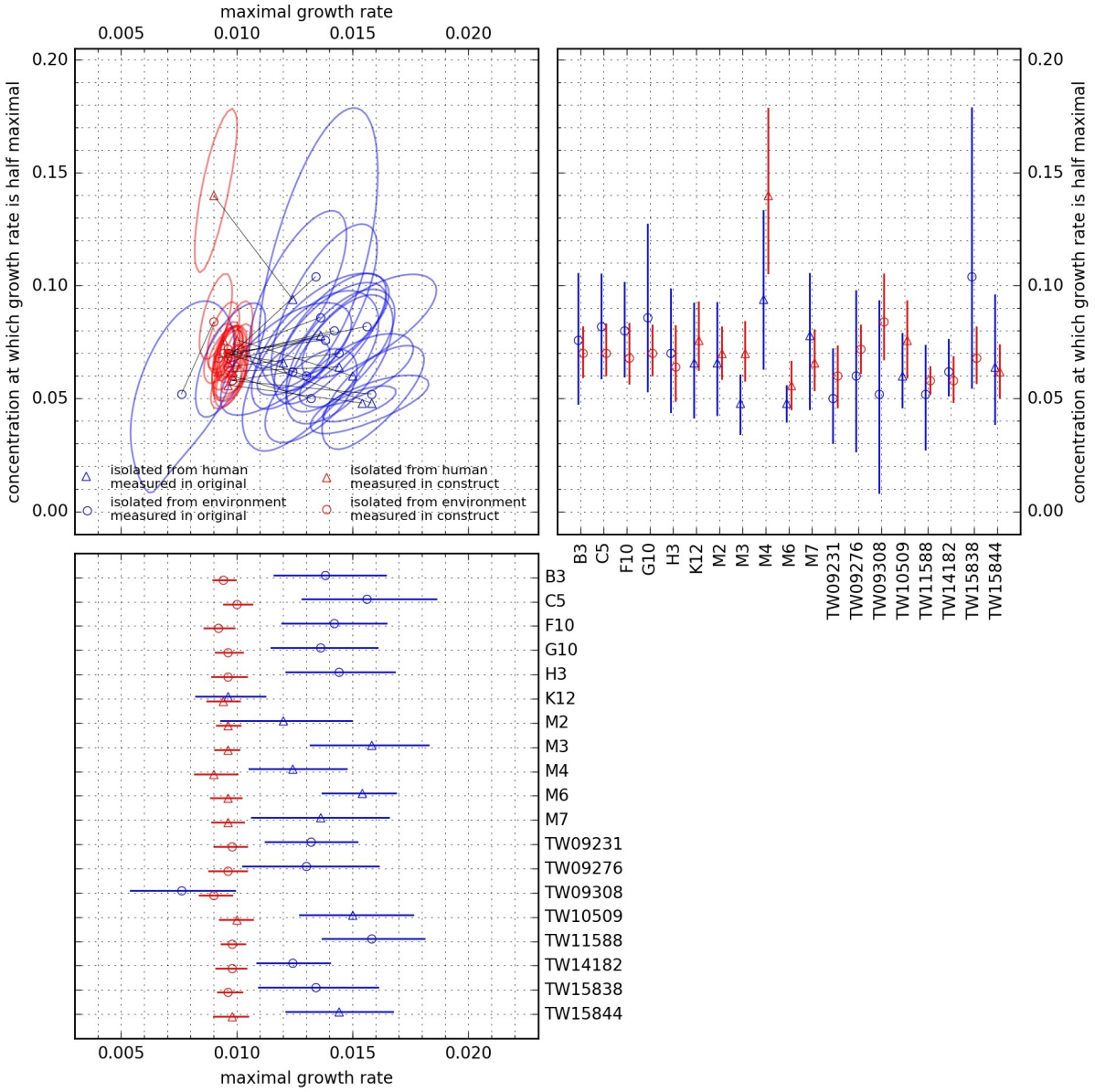


Figure 3.15: **Estimates of μ_{max} and K_s for a Monod model with Hill coefficient equal to 2, flanked by projections of the estimates for the μ_{max} and K parameters and their confidence intervals on their respective axes.** Figure by Srdjan Sarikas.

3.3 Discussion

3.3.1 Where in the *lac* operon does the natural variation relevant for phenotype lie?

Phenotypic variation in lactose metabolism could have different genetic sources, which can be broadly divided into coding and regulatory sources of variation. Here I explore the genetic basis of phenotypic variation among *E. coli* and *Escherichia* in nature, focusing on the phenotype of growth rate on minimal medium with lactose as sole carbon source. While this phenotype does not directly correspond to the situation *E. coli* find themselves in in nature, I use it as a first proxy for the more complex phenotypes related to lactose metabolism that are expected to contribute to fitness in the wild.

In addition, as a step in between DNA sequence and growth rate, I performed LacZ activity assays. This protein activity assay is a combined measure of gene expression (which affects the quantity of β-galactosidase (LacZ) molecules produced) and molecular activity of individual LacZ molecules. In addition to these two sets of data I collected, I have used a set of predictions generated by (Tugrul, 2016) of gene expression of the *lac* operon, based on thermodynamic model predictions of RNA polymerase binding probability to the *lac* promoter region.

Together, these gene expression predictions, protein activity values and growth rates provide information on three 'stops' on the path from genotype to phenotype on lactose. My first aim in this chapter is to explore what information I can extract from these data, about the relative contribution of particular regulatory and coding sources of variation on natural variation in phenotype on lactose.

The coding regions that might contribute to variation in growth on lactose are the three *lac* operon genes, lacZ, lacY, and lacA. Potential *cis*-regulatory sources of variation in the *lac* operon are located in the *lac* promoter region, which lies directly upstream of the three genes of the *lac* operon. This region can be divided into several subregions: the RNA polymerase binding site, the CRP binding site, and the operator regions, which are the binding site for the *lac* repressor. The *lac* repressor itself is another potential source of regulatory variation. Its coding sequence determines (together with the operator sequence it binds to) how strongly it can bind to the operator regions (*trans*-regulatory variation). In addition, the *lac* repressor gene has its own *cis*-regulatory region, the RNA polymerase binding site, which ensures a very weak constitutive expression of the repressor protein. The sequence of this region determines the number of repressor molecules that are on average present in a cell.

Natural sequence variation in *lac* promoter region predicts variation in LacZ activity

Correlating the thermodynamic model predictions of RNA polymerase binding probability with the LacZ activity data, we found that natural variation in the *lac* promoter sequence, upstream of the canonical -10 and -35 RNA polymerase binding site, predicted natural variation in LacZ protein activity measured by a beta-galactosidase assay at full induction.

The correlation of thermodynamic model predictions with the scores on this assay suggests that variation in these positions in the sequence causes variation in gene expression, and that the assay is sensitive to these differences in gene expression. Alternative explanations of this correlation are that it reflects coincidental correlation of genotypes with protein expression, or that it is an artifact of measurement variability. However, the fact that the correlation found again when the experiments were repeated suggests that it is not an artifact of noisy measurements. The other alternative explanation, that the correlation is caused by other differences in the genotypes which happen to co-occur with the variation in their promoter sequences, could be addressed by creating a separate set of *lac* operon constructs which vary only in their *lac* promoter sequence, and repeating our measurements with those. This is something we are considering to do in the future.

The positions in the *lac* promoter region that did vary among the set of *lac* operons I studied, and that predicted variation in gene expression, corresponded to a previously reported alternative RNAP binding site in the *lac* promoter ((Xiong *et al.*, 1991); see (Tugrul, 2016)). The hypothesis proposed by (Tugrul, 2016) is that RNA polymerase may bind to non-canonical binding sites in the vicinity of the canonical RNA polymerase binding site, and that binding to these non-canonical sites affects the overall binding probability of RNA polymerase to a promoter region, and with that gene expression. Whether the variation among these sites in the natural variants of the *lac* operon I studied is adaptive and reflects differences in past selection, or whether this variation represents neutral variation that is not accessible to selection because its effects on fitness are too slight, is a question the present work cannot answer. Since variation in phenotype did not correlate with environment of isolation, we have no positive clues that indicate that this variation is relevant for fitness. However, phenotypic variation and environment of isolation may well be uncorrelated due to extensive environmental cycling of all strains. If this is the case, then variations in phenotype would not correlate with environment of isolation, even if they do differ in fitness.

The fact that construct identity was a significant random factor predicting variation in LacZ activity in addition to predicted gene expression indicates that not all the differences in LacZ

activities between constructs are explained by differences in their predicted *lac* operon expression. This either could mean that the thermodynamic model does not capture the full variation in gene expression across promoter regions, or that in addition to gene expression, other factors such as differences in sequence of the LacZ protein underlie the different LacZ activity scores of the constructs. Note that these two explanations are not mutually exclusive.

Natural variation in LacZ activity does not predict variation in growth rate

Having shown that natural variation in the *lac* promoter sequence predicts natural variation in LacZ protein activity, the next step is to ask to what extent this variation in LacZ activity predicts growth rate. Assuming that the operon is fully induced when growing at the highest lactose concentrations of 0.5 to 1 mM, at which growth rate is (close to) maximal, finding a significant correlation between LacZ activity at the highest inducer concentrations and growth rate, at least at the highest lactose concentrations would indicate that variation RNA polymerase binding and/or molecular activity of the LacZ protein would underlie natural variation in growth on lactose. At lower concentrations of inducer and lactose, the *lac* operon is expected to be partially repressed, and thus variation in repression is an additional potential source of variation.

Mean LacZ activity at full induction did not predict growth rate at different lactose concentrations in a linear mixed effects model, neither as a main effect nor in interaction with lactose concentration. This suggests that neither molecular variation in LacZ, nor variation in polymerase binding underlies natural variation in growth rate. Depending on the combination of factors in the reduced model, LacZ activity at mid and high induction were significant predictors of growth rate; however, these effects vanished when strain M4 was removed from the dataset, which suggests they do not reflect a general trend in the data.

What can be said about the role of the repressor?

If variation in some aspect of repression, either the abundance of the repressor, or the affinity of the repressor molecule to the operator region, would be apparent in our phenotypic measures, one would expect LacZ activities at lower levels of induction to correlate with growth rate at lower concentrations of lactose. At low lactose concentrations, the repressor should be partly active and bound to the operator region, resulting in sub-maximal operon expression. However, at no combination of inducer and carbon source concentrations did LacZ activity correlate with growth rate; that is, not after removing the data points of constructs M4 and TW09308, two

lac operons with reduced growth rate and LacZ activity, which might have been under relaxed selection.

In addition, repression can be quantified as the difference between log LacZ activity at full induction ($\geq$ 0.5 mM IPTG) and log LacZ activity without inducer. While repression was a significant predictor of growth rate in a linear mixed effect model, again this was no longer the case upon removal of construct M4 from the dataset.

Thus, it appears that variation in repression does not underlie variation in growth rate for the whole set of constructs, but might do so for constructs M4 and Tw09308. The behavior of these two operons was idiosyncratic (TW09308 showing strongly reduced LacZ activity under all conditions together with a consistent but less pronounced reduction in growth rate, while M4 showed extremely reduced growth rates, with LacZ activity only being reduced at lower IPTG concentrations). Thus, there seems to be no systematic pattern of variation in repression underlying variation in growth rate at low lactose concentrations.

The thermodynamic model of [Tugrul, 2016] was also used to generate predictions for the expression of the repressor, which is constitutively expressed from a very weak promoter. These predictions did not predict growth rate in any model I tested. Thus, variation in repressor abundance does not seem to underlie variation in growth rate. Model predictions of repressor expression did not explain the variation in repression, although interestingly, construct TW09308 had the lowest predicted repressor expression as well as the lowest calculated repression level, suggesting this operon might indeed have a reduced repression level.

Future work could try to correlate variable sites in the operator regions with variation in repression. This would answer the question of whether variation in repressor binding sites affects its binding affinity, and with that gene regulation. In addition, using more sequences one could attempt, while correcting for the phylogeny, to investigate whether variable sites in the operator regions covary with variable sites in the repressor sequence. If they do, this could be an example of compensatory mutations if repression would not covary in turn, or of co-evolution if they would. However, given that I have found no clear indication that natural variation in repression is relevant for growth rate, such a line of work would not be too promising to yield interesting results.

Molecular variation in LacY might provide the 'missing link' explaining variation in growth rate on lactose

Excluding natural variation in RNA polymerase binding, in LacZ molecular activity and in repression as predictors for natural variation in our growth rate measurements, we hypothesize that the genetic basis for the natural variation in growth rate on lactose medium we found could mainly reflect variation in molecular activity of the lactose permease.

In Chapter 2 of this thesis, I have shown that the lactose permease gene is the most conserved *lac* operon gene among the isolates I studied. The variation that does exist in this protein, however (21 variable sites out of 416 sites in total, one to ten variable sites per pairwise comparison, see **Appendix 2, Table B.3**), might have a relatively large impact on fitness. Dykhuizen et al. (Dykhuizen *et al.*, 1987) have shown that variation in LacY activity has a much larger impact on lactose flux, and with that, on fitness in a lactose-limited environment, than variation in the activity of LacZ. It could thus be the case that LacY is under selection to retain a precisely controlled level of activity, and that the low amount of variation in this protein is a consequence of this selection pressure.

In their experiments, Dykhuizen et al. (Dykhuizen *et al.*, 1987) showed the wildtype variant of LacY to be at an intermediate level between high and low expression and activity.[2] Since fitness in the lactose-limited chemostat environment is directly proportional to lactose flux, this allele is of an intermediate fitness level in that environment. As shown by (Dykhuizen *et al.*, 1987), wildtype LacY is 'on the shoulder of the fitness surface': small (positive or negative) changes in the expression or activity of the protein have a large (positive resp. negative) impact on lactose flux, and with that on fitness. Changes in permease activity have been shown to have a much larger effect on lactose flux than changes in beta-galactosidase activity, since permease activity is the rate limiting step in the flux.[3]

If the permease activity were also the rate limiting step in nature (and we have no reason to assume otherwise), this would mean that in nature, too, small changes in permease activity have large effects on lactose flux. However, in the natural environment, fitness is not always

[2] As in my experiments, Dykhuizen et al. (Dykhuizen *et al.*, 1987) did not distinguish between molecular activity of the protein, and operon expression. Instead, they estimated the activity of LacY by using different levels of inducer and comparing relative fitness levels to a wildtype induction level; the relative contribution of LacY they inferred from comparing the impact of mutations in LacZ, affecting only beta-galactosidase activity, to that of changes in inducer level, which affect both LacY and LacZ concentration.

[3] One can imagine this as follows: as long as the permease is the rate limiting step, making the beta-galactosidase work faster will not have a large impact on the total flux of lactose, since the additional free beta-galactosidase molecules would mainly be 'waiting for' the permease to deliver them more lactose molecules to process.

directly proportional to flux. For example, at sudden pulses of abundant lactose, pumping in too much lactose in a short time can cause cell death (Dykhuizen and Hartl, 1978). In addition, in less extreme situations, it has been shown that the main energetic cost of expressing the *lac* operon in the presence of lactose is linked to permease activity (Eames and Kortemme, 2012). This is due to dissipation of the membrane potential that is used to drive the transport of lactose across the cell membrane . Thus, too high activity of LacY could be costly.

One might speculate that if under some conditions, increasing the flux of lactose can be harmful to cells, while at low concentrations of lactose, fitness is directly proportional to lactose flux, the LacY protein could be under strong selection to remain at an intermediate level of activity. The high level of conservation of the LacY protein, and the lack of systematic differences between strains from different environments (although there are too few variable sites, and too few divergent human isolates in my batch to say much about this), would then suggest that this optimal level of activity has been similar across the strains I studied. This in turn could imply that these strains, although isolated from different environments, have experienced similar conditions with regard to lactose availability in their recent selective history, which would speak in favor of extensive cycling through different environments. However, more data would be needed to support this interpretation.

Indeed, it should be noted that the genetic component of growth rate variation under our measuring conditions is very large for most constructs, as can be seen from **Figure 3.5**. While the operon of clade V strain TW09308 and infant isolate M4 showed systematically lower growth rates and LacZ activities, other constructs did not stand out so obviously. This also has to do with random variation in the growth rate and LacZ activity data, which remained relatively large in both types of measurements. Finally, it is conceivable that there is not one common factor underlying the variation across constructs, but rather that mutations in different regions affected the growth rate of individual operons in different ways.
While two operons mentioned above appeared to have lost fitness, perhaps due to relaxed selection or drift, our data suggest that the rest of the *lac* operons we studied are of roughly equivalent fitness, in so far as this can be determined without direct competition. This would support the hypothesis that they have quite similar selective histories, as suggested above.

3.3.2 No systematic differences in phenotype between *lac* operons of human and environmental isolates

If environmental isolates typically would represent strains that cycle less than average through the mammal gut, we would expect them to have been exposed to lactose less frequently than

human isolates. This we would have expected to result in relaxed selection on the *lac* operon. However, *lac* operons of environmental isolates did not perform systematically different from those of human isolates, neither w.r.t. growth rate on lactose medium nor w.r.t LacZ protein activity.

We thus have found no evidence that the *lac* operons of environmental isolates have been exposed to systematically different selection pressures than those of human isolates, either with regard to the strength, or the direction of selection.

As mentioned in Chapter 2, the presence of lactose-like compounds, such as galactosyl-glycerol, in the environment or the herbivore gut might result in selection pressure on the *lac* operon, even in the absence of extensive cycling through the mammalian gut (Egel, 1979; Boos, 1982; Egel, 1988). Naively, however, I would still expect this to result in a distinguishable phenotype on lactose. Yet, this would depend on the chemical similarity of these substances to lactose, and their prevalence in the environment. In future work, we could address this issue by measuring growth rates of our constructs on galactosyl-glycerol. In addition, our constructs might differ from each other along phenotypic dimensions which were not measured by our growth rate assays. Growth on combinations of different carbon sources or rapidly alternating carbon sources are obvious examples, which is another possible line of future work.

Alternatively, it could be the case that most *E. coli* and *Escherichia* clade I-V strains do in fact cycle regularly through the mammal gut environment. This was already hypothesized about the *Escherichia* clade I-V strains based on their optimal growth temperature (Ingle *et al.*, 2011), which is similar to human isolates of *E. coli*.

It may also well be possible that growth in temperate soil and water environments is so slow that there has not been enough time for systematic phenotypic differences to arise by selection, or for fitness to drop due to the absence of selection. In fact, as I mentioned in Chapter 2, environmental *E. coli* do not grow, but only survive, in soil at temperatures below 30°C (Ishii *et al.*, 2006). For the divergent *Escherichia* clades, it has been shown that in laboratory medium (phenol red broth), they can grow at lower temperatures than *E. coli* (minimum growth temperature 5°C, compared to 11°C for *E. coli*). Given the large difference in minimal growth temperature in soil and culture medium, it is not unlikely that divergent *Escherichia* clades, too, do not grow much if at all in soil in temperate regions. In surface water, however, things may again be different. While selection also can act on survival, and strains which survive only a more limited range of environmental conditions would have less of a chance for future replication in the mammal gut, it is hard to see how differences in the *lac* operon could affect this survival rate.

As I reported in Chapter 2, *lacZ* genes of environmental isolates showed a stronger signature of purifying selection than those of human isolates. As a possible explanation for this difference, I hypothesized that population bottlenecks may occur upon gut colonization, which might lead to genetic drift. This could explain why human isolate operons do not have a systematically higher fitness than those of environmental isolates. However, based on this result, one would expect *lac* operons of human isolates to show a reduced fitness, compared to *lac* operons of environmental isolates. The fact that this prediction is also not supported by my data might be a result of the above mentioned low control coefficient and flat fitness surface of the LacZ protein. This might also explain why genetic drift would have occurred more readily in the *lacZ* gene than in the other *lac* operon genes.

Finally, it is imaginable that several of the above mentioned factors are at play for *E. coli* and *Escherichia* clade I-V in the wild. Occasional population bottlenecks followed by selection in the human gut, interspersed with periods of slower growth or subsistence in the environment, during which there may or may not be selection on the *lac* operon, might interact in ways that are difficult to predict.

3.3.3 Effect of genomic background

Variation in growth rate on lactose medium is expected to be caused by two different factors: variation in the *lac* operon sequence, and variation in the genomic backgrounds of the different strains. It is not known to what extent genomic background influences growth rate on lactose, which aspects of growth rate it influences if any, and whether it exerts its effects independently from the *lac* operon, or in interaction. To address these questions, we compared growth rate data of our *lac* operon constructs with those of the corresponding original isolates. All *lac* operon constructs were maintained in the same genomic background; for each of the original isolates, genomic background was different.

Naively, since genomic background is expected to have an influence on overall growth rate, one would expect μ_{max}, maximal growth rate on abundant lactose, to be determined by both the *lac* operon sequence and genomic background. K_s, on the other hand, which represents the sensitivity to low concentrations of lactose, we would expect to depend only on the *lac* operon sequence.

Comparing growth rate parameters of *lac* operons of natural *E. coli* isolates on plasmids and in their original backgrounds, we found that for all but one of the strains, maximal growth rate on unlimited lactose (μ_{max}) was higher in the original background than in the K12 background

(not counting one infant strain which did not grow on minimal medium). The most simple and parsimonious explanation for this advantage of the original backgrounds is that lab strain K12 has a reduced fitness on minimal medium compared to most natural strains of *E. coli* and Escherichia clade I, III and IV. Such a reduced fitness on K12 on minimal medium might be due to adaptation to a high nutrient environment, which might have occurred over its almost century-long lab propagation. An alternative possibility is that this strain happened to grow relatively poorly on minimal medium to begin with.

Alternatively, one might hypothesize that maximal growth rate is higher for the original isolates because they have the *lac* operon on the chromosome, instead of on a plasmid. It is conceivable that due to e.g. copy number effects, strains with the *lac* operon on a plasmid would differ systematically in growth rate from strains with the *lac* operon on the chromosome. However, if this were the reason, the same difference should exist between the wildtype K12 strain and the K12 Δ *lac* strain with the *lac* operon on a plasmid. This is not the case.

For all constructs, μ_{max} estimates overlap with those of K12, suggesting that all compared *lac* operons are equivalent with regard to the maximal growth rate they can confer. If maximal growth rate on lactose medium would depend on interactions between the *lac* operon with the genomic background, such that operons in their non-native background are at a disadvantage compared to operons in their original background, we would expect μ_{max} estimates for the constructs to be more variable (since not all constructs would be lacking the same things with regard to their original backgrounds). In addition, if this were the only reason for the disadvantage of constructs with regard to the original isolates, this disadvantage should not exist for construct K12.

While our data do not give reason to believe that this is the case, since the μ_{max} estimate for K12 is not higher than for the other constructs, we cannot exclude that interactions with the genomic background occur in principle. Although less parsimonious, it is conceivable that in addition to the K12 background conferring a low maximal growth rate, *lac* operons are dependent on their own genomic background to reach their maximal μ_{max}. To assess whether this is the case, I would need to measure growth rates of my constructs each in multiple genomic backgrounds other than K12 (each carrying a similar small deletion of the *lac* operon on the chromosome). In addition, to find out whether the pattern of maximal growth conferred by the original backgrounds is specific to lactose medium, growth rate on minimal medium with a different carbon source would be informative.

K_s estimates overlapped between constructs and the corresponding original isolates. There is thus no reason to assume that genomic background has any influence on the lactose concen-

tration at which growth is half maximal. This corresponds to our initial expectations; however, as rightly noted by (Tryon, 2001), 'absence of positive evidence for statistical difference does not constitute presence of positive evidence for statistical equivalence'. In other words, while our not finding a statistically significant different between K_s of constructs and original isolates matches our expectations, we cannot conclude from this that there is no effect of genomic background on K_s. Indeed, in spite of attempts to optimize growth curve fitting, the variability of my data is quite high, and there might be a relevant effect hidden in the spread of our parameter estimates (although we have no biological reason to expect this). Finally, it can be noted that the spread of both the μ_{max} and the K_s parameter are larger for the original isolates than for the *lac* operon constructs. I think the reason for this is that these strains tended to clump more in the wells, leading to less reliable growth curve fitting.

3.3.4 Conclusions

We showed that natural phenotypic variation in the *lac* promoter region, upstream of the canonical -10 RNA polymerase binding site, predicts natural variation in gene expression as measured in a LacZ activity assay. However, this variation in LacZ activity does not predict growth rate on different concentrations of lactose. This suggested that neither natural variation in *lac* operon expression, nor molecular activity of the LacZ protein, is an important determinant of variation in growth rate. Instead, variation in the molecular sequence of the lactose permease might underlie natural variation in growth rate on lactose; alternatively, growth rate variation might not have a strong common genetic component.

Lac operons of human isolates were not systematically different in phenotype from *lac* operons of environmental isolates of *E. coli* and divergent *Escherichia* clades. We thus have found no support for the hypothesis that these isolates represent strains adapted to different lifestyles. It may, however, be the case that the *lac* operons of the isolates I compared differ systematically from each other on dimensions we did not investigate, such as growth on a different substrate such as galactosyl-glycerol, combinations of lactose and other sugars, or brief pulses of available lactose. If there really is no systematic difference between phenotypes of human and environmental isolates of *E. coli* and divergent *Escherichia* clades, this could have several reasons, such as regular cycling of *E. coli* and *Escherichia* clade I-V strains through the mammal gut environment, the presence of chemically similar compounds to lactose in temperate soil and water environments, slow growth in temperate soil and water environments, or population bottlenecks upon gut colonization.

The maximal growth rate on unlimited lactose across natural variants of the *lac* operon was similar to that of K12 when these *lac* operons were in a K12 background, but almost universally higher than K12 in their original genomic backgrounds. This suggests that genomic background affects maximal growth rate on minimal medium and/or on lactose, and that the K12 background is compromised relative to the genomic background of natural *E. coli* and *Escherichia* isolates on minimal medium with lactose as sole carbon source.

While my data do not give reason to expect an interaction between the *lac* operon and its genomic background w.r.t. maximal growth rate on lactose, unfortunately my experimental design does not permit me to exclude that this is the case. I found that the lactose concentration at which growth is half maximal is similar across constructs, as well as between *lac* operons on plasmids and the strains they were isolated from. There thus is no reason to assume that genomic background has any influence on this parameter, although we cannot formally exclude this.

3.4 Methods

3.4.1 Strains

The same strains were used as described in Chapter 2. See the Methods section of Chapter 2 and **Table 2.1** for an overview of these strains, their origin and location of isolation.

3.4.2 Plasmid construction

All enzymes were obtained from New England Biolabs (Ipswich, MA) except where stated otherwise. Plasmids were constructed from a pZs* backbone (Lutz and Bujard, 1997) carrying a kanamycin resistance gene and the Venus fluorescence gene. The backbone was isolated from overnight liquid bacterial culture using an Invitrogen PureLink HiPure Midiprep plasmid purification kit. The Venus gene was excised from the plasmid backbone by restriction with HindIII and XhoI restriction enzymes, and replaced by the PCR amplified *lac* operons of the wild isolates. PCRs were carried out using Phusion polymerase (ThermoFisher Scientific, Waltham, MA). PCR primers were designed with an overhang containing HindIII and XhoI restriction sites and supplied by Sigma Aldrich (St. Louis, MO); after amplification, the PCR products were restricted with these enzymes and ligated together with the restricted plasmid backbone. PCR primers annealed in the genes flanking the *lac* operon, which are cynX and mphR in strain K12 MG1655. In several of the more divergent wild isolates, the *lac* operon is flanked by

other genes; for these strains, alternative primer sets were used (see supplement for primer sequences).

PCR products were gel purified before restriction using a Zymoclean Gel DNA Recovery kit and column purified after restriction using a Zymo Clean&Concentrator kit (Zymo Research, Irvine, CA). Restrictions were carried out for 3 hours at 37°C. The plasmid backbone was dephosphorylated after restriction by adding shrimp alkaline phosphatase (rSAP) to the mixture and incubating for one hour at 37°C. Enzymes were deactivated for 20 minutes at 80°C. Subsequently, the restricted backbone was gel purified as described above. Ligations were carried out for 1 hour at room temperature or overnight at 16°C, using T4 DNA ligase, using a 1:6 backbone:insert volume ratio when possible, except when DNA concentrations were low; in the latter case, added volumes were maximised. Ligation mixtures were column purified as described above and eluted in 20 ml nuclease-free water. 5 μl of column purified ligation mixture was electroporated at 1800 mV, path length 2 mm, into 70 μl electrocompetent cells of cloning strain DH5α and incubated for around 1.5 hours in 700 μl SOC medium. This entire mixture was then spread on MacConkey agar plates with lactose (powder from VWR, Radnor, PA) containing 25 μg/ml kanamycin, and incubated for one to two days at 37°C. Constructs containing a functional *lac* operon were recognisable as dark red colonies. From these red colonies, overnight cultures were inoculated in LB medium with 25 μg/ml kanamycin. The next morning, plasmids were prepped from 2 ml of each overnight culture, using a Zyppy plasmid isolation kit (Zymo Research, Irvine, CA), eluting with 20 μl nuclease-free water. Of each of the the thus purified plasmids, 2 μl was electroporated into strain HG105, which is K12 MG1655 with a small deletion of the *lac* operon, kindly shared by Rob Philips (Garcia and Phillips, 2011).

3.4.3 Frozen stock plates

Growth rate measurements and β-galactosidase activity assays for the different *lac* operon constructs were started from overnight precultures, which had been inoculated by pinning from a 96 well 'master' plate containing 4 replicate frozen stocks of each of the 21 constructs. To prepare these frozen stock plates, the original frozen stocks of the construct-bearing HG105 strains were streaked out on MacConkey agar plates with kanamycin. From each of these streaks , 4 separate colonies were selected to inoculate 4 separate overnight cultures in 5 ml LB with 25 μg/ml kanamycin. The following morning, cultures were put on ice and of each culture, 140 μl was then added to one well of the aforementioned 96 well plates, which were kept on ice; 4 replicate frozen stock plates were made. The order of the constructs on the plate was randomised using R, with the constraint that each construct should occur at least once in

one of the outermost wells; five constructs occurred twice. The wells were then filled up with 60 μl of 50% glycerol solution. Plates were frozen at -80 °C.

3.4.4 Growth rate measurements

Growth rate measurements were started by pinning from precultures grown in a 96 well plate. Precultures for growth rate measurements were inoculated by pinning from a frozen stock plate, which had been thawed on ice for 75 minutes at 4°C. Precultures were grown for 24 hours in M9 medium with 1mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.3% glycerol, and for strains bearing constructs, 25 μg/ml kanamycin, at 37 °C, on a 96 well-plate shaker, shaking at 900 rpm.
Growth rate measurements were carried out in M9 medium with 1mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.001% TritonX, and lactose at the specified concentrations, in a Biotek plate reader at 37 °C, under fast double orbital shaking, with optical density being measured at 600 nm every 10 minutes for 20 hours, or until optical density had stopped increasing.
Growth rates were calculated as the slope of a linear function fit to $\ln(OD - reference)$ for a defined time window, starting from the first time point from which OD values remained consistently at a value of 0.0015 or more above reference. The reference OD value was calculated as the average of the first 10 OD values measured during that experiment in the respective well. The length of the time window was chosen such that the number of time points included was maximized, while only time points in the exponential phase were included; a fixed time window was used per replicate run. To correct for that fact that due to the logarithmic transformation, measurement errors have a larger impact at lower OD values , measurements within the time window were weighted according to the formula $(\frac{OD-reference}{0.0005})^2$[4], where 0.0005 is the limit of the resolution of the plate reader.

3.4.5 Lactose metabolism assays

The activity of beta-galactosidase was quantified by an ONPG assay modified from (Dodd *et al.*, 2001), following a protocol kindly shared by Adam Palmer. In preparation for this assay, overnight cultures were started by pinning from the frozen stock plate described above, after it had been thawed on ice for 75 minutes at 4°C. Overnight precultures were carried out on a plate shaker at 37 °Cin LB with 25 μg/ml kanamycin and 0.05, 0.1, 0.2 or 0.5 mM of the inducer IPTG, with any one preculture plate containing two different IPTG concentrations across different wells. The following morning, 2 μl from each well of the overnight preculture plate was

[4]thanks to Bor Kavcic for pointing out this problem, as well as its solution.

transferred to a 96 well plate containing fresh medium and shaken at 37 °Cfor 90 minutes. Subsequently, this dilution step was repeated and the cultures were once more grown for 90 minutes, after which the lid was replaced and the plate was transferred to the plate reader, and growth was continued until optical density values as measured at 600 nm were mostly over 0.2, and effects of condensation on OD readings had worn off. At this point, the plate was placed on top of a metal cooling block in an ice bucket for 5 minutes, after which 10 μl per well was transferred to a plate containing prewarmed lysis buffer, and shaken for 30 minutes on a plate shaker at 37 °C. Lysis buffer consisted of 100 mM Tris-HCl (pH 8), 1 mM $MgSO_4$, 10 mM KCl, 126.7 mg/l polymyxin B, and 12.67 ml/l β-mercaptoethanol. After 30 minutes of lysis, 40 μl of 4 mg/ml ONPG solution, containing Tris-HCl, $MgSO_4$, and KCl at equal molarities to the lysis buffer, was added to each well, and the plate was transferred to a plate reader where absorbance at 414 nm was measured every minute for 2 hours. To the thus obtained sequence of increasing values, a linear function was fit. β-galactosidase activity was then calculated from the slope *s* of this function, volume *v* of culture added, and final OD of the respective well after pregrowth, using the formula $\frac{s*200000}{OD*v}$.

Since the variance of these measurements scaled with the mean, data were transformed by taking the natural logarithm before the analysis, which removed this heteroscedasticity and rendered them normally distributed. The log transformed data were analyzed with a linear mixed effect model, starting with IPTG concentration, strain origin (human or environmental) and predicted gene expression as fixed factors, and construct identity and run (whether the experiment was part of the first or second set of experiments) as random factors, using the lme4 and lmerTest packages in R. Models were compared using ANOVA; in addition, for each model the AIC score was calculated.

4 Conclusions

In this work, we tried first of all to get a view on the distribution and fitness of the *lac* operon across 20 natural isolates of *E. coli* and divergent clades of Escherichia. With this, we aimed to get insight into the ecology of *E. coli* and Escherichia in the wild. We found that the *lac* operon has been conserved across natural isolates of *E. coli* and divergent clades of Escherichia, with respect to its organization as well as its function. With the exception of one *lac* operon that lost its function due to a frameshift mutation, and two *lac* operons conferring lower growth rates on lactose and lower LacZ activity, all *lac* operons we studied appeared to be of equivalent fitness as far as our methods could distinguish. Correspondingly, dN/dS ratios were the same for human and environmental branches in the phylogeny for all genes except the *lacZ* gene. For the *lacZ* gene, environmental branches even had a lower dN/dS ratio than human branches; however, our phenotypic data did not show a corresponding difference. *Lac* operons of human isolates did not differ systematically from those of environmental isolates with respect to growth rate conferred on lactose medium or LacZ activity.

If lactose, as is commonly assumed, is a substrate typical for the mammal gut, but very rare in other environments, the question is what explains the presence of these conserved *lac* operons in environmental isolates. While tests for lactose metabolism are part of some phenotypic tests for *E. coli*, likely causing an isolation bias for strains with functional *lac* operons, this does not explain how these *lac* operons would have retained their function and fitness over extended periods of time in the absence of any other form of selection.

Three possible reasons can be envisaged for the presence of conserved *lac* operons across environmental isolates. First of all, one might imagine that these *lac* operons were acquired by horizontal transfer from gut-adapted *E. coli*. In combination with the isolation bias mentioned above, this could explain why almost only strains with *lac* operons very similar to gut *E. coli* are isolated from soil and water environments. Secondly, environmental *E. coli* and *Escherichia* strains might all still regularly cycle through the mammal gut. If growth rates in the environment are sufficiently low, a substantial part of the selection on growth of these strains might take

place in the gut, and generation time in temperate soil and water might be long enough to render loss of fitness by mutation accumulation very slow. Finally, the *lac* operon might have a different function in the environment, such as the metabolism of other lactose-like compounds.

The first hypothesis of horizontal transfer of functional *lac* operons from gut *E. coli* is disproved by our finding that the phylogeny of *lac* operons of divergent environmental clades of *Escherichia* is the same as the whole-genome phylogeny of these strains. We did find evidence for a homologous recombination event of part of the *lac* operon between *Escherichia* clades. This event involved most of the *lacY* gene, which was the gene with the strongest signature of purifying selection, which might be most relevant for fitness of the operon. One might speculate that this recombinant was indeed selected for because it restored *lac* operon fitness. While our results thus suggest that there has been a limited amount of homologous recombination, the data suggest too few horizontal gene transfer events to explain the overall presence of functional *lac* operons in environmental *E. coli* and *Escherichia* strains and clades.

Thus, our results suggest that either these strains and clades cycle regularly through the human gut, or are regularly exposed to a lactose-like compound in the environment. Unfortunately, our data do not enable us to distinguish between these two hypotheses; future work could try to disentangle these by testing for growth rate on galactosyl-glycerol, a *lac* operon substrate which occurs in the gut of plant-eating animals, and might thus be more widespread than lactose. If *lac* operons of environmental clades would systematically differ in growth on this substrate, either in maximal growth rate or in substrate concentration of half maximal growth, this would indicate that these clades cycle less through the mammal gut, and that the selection pressure maintaining their *lac* operons may well be caused by this substrate.

The second main question I tried to address in this work is where in the *lac* operon lies the genetic variation underlying variation in phenotype, and potentially variation in fitness.
We found variation in the *lac* promoter region to predict LacZ activity, using a thermodynamic model of (Tugrul, 2016) which is based on the RNA polymerase binding probability matrix inferred by (Kinney *et al.*, 2010) and sums the probability of polymerase binding across the sites in the promoter region. While all *lac* operons were identical in the canonical -10 and -35 binding locations, the model predicted variation in LacZ activity scores of *lac* constructs based on sequence variation upstream of these sites.

However, this variation in LacZ activity, a combined measure of gene expression and molecular activity, did not predict variation in growth rate on lactose medium. Thus, we have no evidence that the variation in gene expression or molecular activity of LacZ across the natural

variants of the *lac* operon we studied is relevant for fitness; although it might be important for aspects of fitness we did not measure, such as carbon source switching. Variation in growth rate across constructs might be most strongly determined by variation in LacY, the protein which was found in previous work to have the highest control coefficient and steepest fitness surface w.r.t. growth on lactose, and which we correspondingly found to be most conserved across isolates. Alternatively, the lac of correlation between LacZ activity and growth rate might be due to the *lac* operons under study being mostly equivalent with respect to growth rate; in that case, the growth rate variation on lactose we measured would not have a strong genetic component.

If the variation in growth rate we found would have correlated with environment of isolation, this would indicate that it is a result of differences in selection pressure, which would imply a genetic basis. As it is, this variation may or may not be a result of unknown differences in selective history between the strains we studied.

Future work could try to correlate variation in the LacY protein sequence with variation in growth rate or growth rate parameters. In addition, one might try to assess LacY activity directly by measuring the remaining lactose concentration in the medium. If either of these measures would predict variation in growth rate, this would indicate that variation in LacY underlies natural variation in growth rate on lactose.

Finally, I tried to get a view on the effect of genomic background on phenotypic variation in growth on lactose. Fitting a Monod model with Hill coefficient to growth rate on lactose of constructs and the corresponding isolates with their original genomic backgrounds, I found that genomic background affects maximal growth rate at unlimited lactose, but found no evidence that it affects the substrate concentration at which growth is half maximal. This is in line with expectation, since doubling time is affected by many more factors than metabolism alone. However, my setup did not enable me to infer whether the *lac* operon interacts with the genomic background in shaping the phenotype of growth on lactose.

It is surprising that after more than half a decade of research, the *lac* operon still poses so many mysteries. While I have asked more questions than I have answered, I hope to have contributed a wider perspective on the place of the *lac* operon, and more broadly, of *E. coli* and *Escherichia*, in nature.

Bibliography

Jeff Abramson, Irina Smirnova, Vladimir Kasho, H. Ronald Kaback, and So Iwata, "Structure and Mechanism of the Lactose Permease of Escherichia coli," *Journal of Biological Chemistry*, 301:610–615, 2003.

John F Andrews, "A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates," *Biotechnology and Bioengineering*, 10(6):707–723, 1968.

K J Andrews and E C C Lin, "Thiogalactoside transacetylase of the lactose operon as an enzyme for detoxification.," *Journal of bacteriology*, 128(1):510–513, oct 1976.

Ronald M Atlas, *Handbook of Microbiological Media*, CRC Press, 4th edition, 2010.

J Barrick, D Yu, S Yoon, H Jeong, T Oh, D Schneider, R Lenski, and J Kim, "Genome evolution and adaptation in a long-term experiment with Escherichia coli," *Nature*, 461(7268):1243–1247, 2009.

Frederic Bertels, Olin K Silander, Mikhail Pachkov, Paul B Rainey, and Erik van Nimwegen, "Automated reconstruction of whole genome phylogenies from short sequence reads.," *Molecular biology and evolution*, 31(5):1077–1088, 2014.

Zachary D Blount, "The unexhausted potential of ¡i¿E. coli¡/i¿," *eLife*, 4:1–12, 2015.

Winfried Boos, "Synthesis of (2R)-glycerol-o-β-d-galactopyranoside by β-galactosidase," *Methods in enzymology*, 89:59–64, 1982.

B Y Roy J Britten and Eric H Davidson, "Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty," 46(2), 1971.

Roy J Britten and Eric H Davidson, "Gene Regulation for Higher Cells: A Theory," *Science*, 165(3891):349–357, 1969.

Sean B Carroll, "Endless forms: the evolution of gene regulation and morphological diversity.," *Cell*, 101(6):577–580, 2000.

G Cornelis, “Sequence relationships between plasmids carrying genes for lactose utilization.,” *Journal of general microbiology*, 124(1):91–7, may 1981.

Alejandro Cravioto, Rosa E Reyes, Francisca Trujillo, Felipe Uribe, Armando Navarro, Jose M de la Roca, Juan M Hernandez, Gabriel Perez, and Virginia Vasquez, “Risk of diarrhea during the first year of life associated with initial and subsequent colonization by specific enteropathogens,” *American Journal of Epidemiology*, 131(5), 1990.

Rolf Daniel, “The metagenomics of soil.,” *Nature reviews. Microbiology*, 3(6):470–478, 2005.

A. M. Dean, “A molecular investigation of genotype by environment interactions,” *Genetics*, 139(1):19–33, jan 1995.

Antony M Dean, “Selection and Neutrality in Lactose Operons of Escherichia coli,” *Genetics*, 1989.

Erez Dekel and Uri Alon, “Optimality and evolutionary tuning of the expression level of a protein,” *Nature*, 436(7050):588–592, 2005.

I B Dodd, a J Perkins, D Tsemitsidis, and J B Egan, “Octamerization of λ CI repressor is nedded for effective repressions of P_{RM} and efficient switching from lysogeny,” *Genes Dev.*, 15:3013–3022, 2001.

D E Dykhuizen, a M Dean, and D L Hartl, “Metabolic flux and fitness.,” *Genetics*, 115(1):25–31, jan 1987.

Daniel Dykhuizen and Daniel Hartl, “Transport by the lactose permease of Escherichia coli as the basis of lactose killing . Transport by the Lactose Pernease of Escherichia coli as the Basis of Lactose Killing,” *Journal of bacteriology*, 135(3):876, 1978.

Matt Eames and Tanja Kortemme, “Cost-benefit tradeoffs in engineered lac operons.,” *Science*, 336(6083):911–5, may 2012.

Richard Egel, “The lac-operon for lactose degradation, or rather for the utilization of galactosylglycerols from galactolipids?,” *Journal of Theoretical Biology*, 79(1):117–119, 1979.

Richard Egel, “The ’lac’ operon: an irrelevant paradox?,” *Trends in Genetics*, 4(2):31, 1988.

J.J. Farmer, G.R.Fanning, B.R.Davis, C.M.O’Hara, C.Riddle, W.Hickman-Brenner, M.A.Asbury, V.A.Lowery, and D. Brenner, “Escherichia fergusonii and Enterobacter taylorae.species of Enterobacteriaceae isolated from clinical specimens,” *Journal of clinical microbiology*, 21(21):77–81, 1985.

C. F. Favier, E. E. Vaughan, W. M. De Vos, and A. D L Akkermans, "Molecular monitoring of succession of bacterial communities in human neonates," *Applied and Environmental Microbiology*, 68(1):219–226, 2002.

Hernan G Garcia and Rob Phillips, "Quantitative dissection of the simple repression input-output function.," *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12173–12178, 2011.

Olivier Gascuel, "BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data," *Molecular biology and evolution*, 14(7):685–695, 1997.

W Gilbert and B Müller-Hill, "The lac operator is DNA.," *Proceedings of the National Academy of Sciences of the United States of America*, 58(6):2415–2421, 1967.

N Goldman and Z Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences.," *Molecular biology and evolution*, 11(5):725–736, 1994.

Nick Goldman, "Statistical Tests of Models of DNA Substitution," *Journal of Molecular Evolution*, 36:182–198, 1993.

C. P L Grady, Barth F. Smets, and Daniel S. Barbeau, "Variability in kinetic parameter estimates: A review of possible causes and a proposed terminology," *Water Research*, 30(3):742–748, 1996.

Nicolas Guex and Manuel C. Peitsch, "SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling," *Electrophoresis*, 18(15):2714–2723, jan 1997.

Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.," *Systematic biology*, 59(3):307–21, may 2010.

Nicole Guiso and Agnes Ullmann, "Expression and Regulation of Lactose Genes Carried by Plasmids," *Journal of bacteriology*, 127(2):691–697, 1976.

Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano, "Dating of the Human-Ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, 22:160–174, 1985.

Hopi E. Hoekstra and Jerry A. Coyne, "The locus of evolution: Evo devo and the genetics of adaptation," *Evolution*, 61(5):995–1016, 2007.

John P. Huelsenbeck and Bruce Rannala, "Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context," *Science*, 276(5310):227–232, apr 1997.

Danielle J Ingle, Olivier Clermont, David Skurnik, Erick Denamur, Seth T Walk, and David M Gordon, "Biofilm formation by and thermal niche and virulence characteristics of Escherichia spp.," *Applied and environmental microbiology*, 77(8):2695–700, apr 2011.

Satoshi Ishii, Winfried B Ksoll, Randall E Hicks, and Michael J Sadowsky, "Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds," *Applied and environmental microbiology*, 72(1):612–621, 2006.

Satoshi Ishii and Michael J. Sadowsky, "Escherichia coli in the Environment: Implications for Water Quality and Human Health," *Microbes and Environments*, 23(2):101–108, 2008.

François Jacob and Jacques Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, 3(3):318–356, 1961.

François Jacob and Jacques Monod, "On the Regulation of Gene Activity," *Cold Spring Harb Symp Quant Biol*, 26:193–211, 1961.

R H Jacobson, X J Zhang, R F DuBose, and B W Matthews, "Three-dimensional structure of beta-galactosidase from E. coli.," *Nature*, 369(6483):761–766, 1994.

Alan Jobe, John E. Sadler, and Suzanne Bourgeois, "lac Repressor-operator interaction," *Journal of Molecular Biology*, 69:397–408, 1972.

Douglas H Juers, Brian W Matthews, and Reuben E Huber, "LacZ β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance," *Protein Science*, 21(12):1792–1807, 2012.

James B Kaper, "Pathogenic Escherichia coli.," *International journal of medical microbiology : IJMM*, 295(6-7):355–356, 2005.

Mary-Claire King and A C Wilson, "Evolution at two levels in humans and chimpanzees," *Science*, 188(4184):107–116, 1975.

Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox, "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):9158–9163, 2010.

Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost, "Automated phylogenetic detection of recombination using a genetic algorithm.," *Molecular biology and evolution*, 23(10):1891–901, oct 2006.

J G Lawrence and H Ochman, "Molecular archaeology of the Escherichia coli genome.," *Proceedings of the National Academy of Sciences of the United States of America*, 95(16):9413–7, aug 1998.

Jeffrey Lawrence, "Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes," *Current Opinion in Genetics and Development*, 9(6):642–648, 1999.

Jeffrey G. Lawrence and John R. Roth, "Selfish operons: Horizontal transfer may drive the evolution of gene clusters," *Genetics*, 143(4):1843–1860, 1996.

Si Quang Le and Olivier Gascuel, "An improved general amino acid replacement matrix.," *Molecular biology and evolution*, 25(7):1307–20, jul 2008.

B. R. Levin, "Periodic selection, infectious gene exchange and the genetic structure of E. coli populations.," *Genetics*, 99(1):1–23, 1981.

Mitchell Lewis, Geoffrey Chang, Nancy C. Horton, Michele A. Kercher, Helen C. Pace, Maria A. Schumacher, Richard G. Brennan, and Ponzy Lu, "Crystal structure of the lactose operon repressor and its complexes with DNA and inducer.," *Science*, 271(5253):1247–1254, 1996.

Nathalie Lombard, Emmanuel Prestat, Jan Dirk van Elsas, and Pascal Simonet, "Soil-specific limitations for access and analysis of soil microbial communities by metagenomics," *FEMS Microbiology Ecology*, 78(1):31–49, 2011.

Chengwei Luo, Seth T Walk, David M Gordon, Michael Feldgarden, James M Tiedje, and Konstantinos T Konstantinidis, "Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species.," *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7200–5, apr 2011.

Rolf Lutz and Hermann Bujard, "Independent and tight regulation of transcriptional units in Escherichia coli via the LacR / O , the TetR / O and AraC / I 1 -I 2 regulatory elements," *Nucleic Acids Research*, 25(6):1203–1210, 1997.

Mahapatra A, S Mahapatra, and A Mahapatra, "Escherichia fergusonii: An emerging pathogen in South Orissa.," *Indian Journal of Medical Microbiology*, 23:204–204, 2005.

Anja Marbach and Katja Bettenbrock, "lac operon induction in Escherichia coli: Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA.," *Journal of biotechnology*, 157(1):82–8, jan 2012.

L J Mata and J J Urrutia, "Intestinal colonization of breast-fed children in a rural area of low socioeconomic level," *Ann NY Acad Sci*, 25:1380–1390, 1971.

Benno Müller-Hill, *The lac operon: a short history of a genetic paradigm*, Walter de Gruyter, 1996.

Frederick C Neidhardt and Roy Curtiss, *Escherichia coli and Salmonella : cellular and molecular biology*, ASM Press, 1996.

Rasmus Nielsen and Ziheng Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, 148(3):929–936, 1998.

H Ochman, J G Lawrence, and E a Groisman, "Lateral gene transfer and the nature of bacterial innovation.," *Nature*, 405(6784):299–304, 2000.

S Oehler, E R Eismann, H Krämer, and B Müller-Hill, "The three operators of the lac operon cooperate in repression.," *The EMBO journal*, 9(4):973–979, 1990.

Marina V Omelchenko, Kira S Makarova, Yuri I Wolf, Igor B Rogozin, and Eugene V Koonin, "Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ.," *Genome biology*, 4(9):R55, 2003.

Tadasuke Ooka, Kazuko Seto, Kimiko Kawano, Hideki Kobayashi, Yoshiki Etoh, Sachiko Ichihara, Akiko Kaneko, Junko Isobe, Keiji Yamaguchi, Kazumi Horikawa, Marjorie Bardiau, Jacques G Mainil, Lothar Beutin, Yoshitoshi Ogura, and Tetsuya Hayashi, "Clinical Significance of Escherichia albertii," *Emerging Infectious Diseases*, 18(3):488–492, 2012.

Ertugrul M Ozbudak, Mukund Thattai, Han N Lim, Boris I Shraiman, and Alexander Van Oudenaarden, "Multistability in the lactose utilization network of Escherichia coli.," *Nature*, 427(6976):737–40, feb 2004.

Csaba Pal and Laurence D. Hurst, "Evidence against the selfish operon theory," *Trends in Genetics*, 20(6):232–234, 2004.

Chana Palmer, Elisabeth M. Bik, Daniel B. DiGiulio, David A. Relman, and Patrick O. Brown, "Development of the human infant intestinal microbiota," *PLoS Biology*, 5(7):1556–1573, 2007.

Arthur B. Pardee, François Jacob, and Jacques Monod, "The genetic control and cytoplasmic expression of Inducibility in the synthesis of β-galactosidase by E. coli," *Journal of Molecular Biology*, 1(2):165–178, 1959.

Morgan N Price and Adam P Arkin, "Operon Formation is Driven by Co-Regulation and Not by Horizontal Gene Transfer," *Genome Research*, 15:809–819, 2005.

Selwyn Quan, J Christian J Ray, Zakari Kwota, Trang Duong, Gábor Balázsi, Tim F Cooper, and Russell D Monds, "Adaptive evolution of the lactose utilization network in experimentally evolved populations of Escherichia coli.," *PLoS genetics*, 8(1):e1002444, jan 2012.

M Razo-Mejia, J Q Boedicker, D Jones, A DeLuna, J B Kinney, and R Phillips, "Comparison of the theoretical and real-world evolutionary potential of a genetic circuit.," *Physical biology*, 11(2):026005, 2014.

J. A. Robinson and J. M. Tiedje, "Nonlinear estimation of monod growth kinetic parameters from a single substrate depletion curve," *Applied and Environmental Microbiology*, 45(5):1453–1458, 1983.

Joseph A Robinson, "Determining microbial kinetic parameters using nonlinear regression analysis," *Advances in Microbial Ecology*, 8:61–114, 1985.

Eduardo P C Rocha, John Maynard Smith, Laurence D. Hurst, Matthew T G Holden, Jessica E. Cooper, Noel H. Smith, and Edward J. Feil, "Comparisons of dN/dS are time dependent for closely related bacterial genomes," *Journal of Theoretical Biology*, 239(2):226–235, 2006.

Steven L Roderick, "The lac operon galactoside acetyltransferase," *Comptes Rendus Biologies*, 328:568–575, 2005.

Michael A Savageau, "Escherichia coli habitats, cell types, and molecular mechanisms of gene control," *The American naturalist*, 122(6):732–744, 1983.

Howard A. Shuman and Thomas J. Silhavy, "Microbial genetics: The art and design of genetic screens: Escherichia coli," *Nature Reviews Genetics*, 4(6):419–431, 2003.

David L Stern, "Perspective: Evolutionary Developmental Biology and the Problem of Variation," *Evolution*, 54(544):1079–1091, 2000.

David L. Stern and Virginie Orgogozo, "The loci of evolution: How predictable is genetic evolution?," *Evolution*, 62(9):2155–2177, 2008.

Daniel M Stoebel, "Lack of evidence for horizontal transfer of the lac operon into Escherichia coli.," *Molecular biology and evolution*, 22(3):683–90, mar 2005.

Daniel M Stoebel, Antony M Dean, and Daniel E Dykhuizen, "The cost of expression of Escherichia coli lac operon proteins is in the process, not in the products.," *Genetics*, 178(3):1653–60, mar 2008.

Michael Travisano and Richard E. Lenski, "Long-Term Experimental Evolution in Escherichia coli. IV. Targets of Selection and the Specificity of Adaptation," *Genetics*, 143(May):15–26, 1996.

W W Tryon, "Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests.," *Psychological methods*, 6(4):371–386, 2001.

Murat Tugrul, *Evolution of transcriptional regulatory sequences*, PhD thesis, 2016.

Parag A. Vaishampayan, Jennifer V. Kuehl, Jeffrey L. Froula, Jenna L. Morgan, Howard Ochman, and M. Pilar Francino, "Comparative metagenomics and population dynamics of the gut microbiota in mother and infant," *Genome Biology and Evolution*, 2(1):53–66, 2010.

Seth T Walk, Elizabeth W Alm, David M Gordon, Jeffrey L Ram, Gary a Toranzos, James M Tiedje, and Thomas S Whittam, "Cryptic lineages of the genus Escherichia.," *Applied and environmental microbiology*, 75(20):6534–44, oct 2009.

X. Xiong, N. De La Cruz, and W. S. Reznikoff, "Downstream deletion analysis of the lac promoter," *Journal of Bacteriology*, 173(15):4570–4577, 1991.

Ziheng Yang, "PAML 4: Phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

Ziheng Yang and Rasmus Nielsen, "Synonymous and nonsynonymous rate variation in nuclear genes of mammals," *Journal of Molecular Evolution*, 46(4):409–418, 1998.

Ziheng Yang, Rasmus Nielsen, Nick Goldman, and Anne-Mette Krabbe Pedersen, "Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites," *Molecular biology and evolution*, 19(1):49–57, 2000.

Ziheng Yang, Wendy S W Wong, and Rasmus Nielsen, "Bayes empirical Bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, 22(4):1107–1118, 2005.

A Appendix 1

Table A.1: **Occurrances of the *lac* operon in other Enterobacteriae, as well as published genomes of pathogenic E. coli.** This is not intended to be an exhaustive list, but rather an illustration of the variety of forms in which the lac operon can be found across Enterobacteriae. No attempt was made to systematically cover all available genomes of Enterobacteriae. Rather, the *lac* operon sequence was used to guide targeted BLAST queries in different Enterobacteric genomes, attempting to span a variety of different species. While strains without a lac operon have been included occasionally for illustrative purposes, the ratio of their representation in this table should not be taken to reflect their relative abundance in nature. In addition, it should be kept in mind that available genome data reflect sequencing and publishing biases at least as much as the underlying natural diversity.

Species	Subspecies	Strain	Accession	Lac genes	Adjacent?	Similarity to *E. coli* K12	Remarks
E. coli	O157:H7	EDL933	AE005174.2	all	y	98%	
E. coli	O104:H4		AFOB02000005.1	all	y	99%	
E. coli	O104:H21	CFSAN002236	AUQC01000006.1	I,Z,Y	y	99%	
E. coli	O121:H19	2011C-3609	JASV01000004.1	-			
E. coli	O145:H28	RM13516	CP006262.1	all	y	98%	
E. coli		NC101	AEFA01000004.1	all	y	97%	
E. coli		53638	AAKB02000001.1	I,Z	y	99%	Flanked by insertion sequence, transposase orfA
E. coli		APEC O1	CP000468.1	all	y	97%	
E. coli	O83:H1		CP001855.1	all	y	97%	
Shigella sp.		PAMC 28760	CP014768.1	all	y	99%	
Shigella flexneri		2a str. 2457T	AE014073.1	-			
Escherichia fergusonii		ATCC 35469	CU928158.2	Z		76%	
Escherichia albertii		TW08933;	AEJU00000000.1;				
		TW15818	AEJY00000000.1				
Salmonella enterica	**arizonae**	serovar 62:z4,z23;	CP000880.1;	all	y	74%	
		serovar 62:z36 str. RKS2983	CP006693.1				
		11-01854;	CP011292.1;				
		11-01853;	CP011289.1;				
		11-01855	CP011288.1				
Salmonella enterica	**enterica**	serovar Typhimurium str. DT2;	HG326213.1;	-			
		serovar Montevideo str. USDA-ARS-USMARC-1903	CP007222.1				
		N268-08;					
		NCTC 12419;					
		serovar 48:z41 str. RKS3044					
Citrobacter sp.		FDAARGOS_156	CP014030.1	I,Z	y	76%	
Citrobacter amalonaticus		FDAARGOS_122	CP014015.1	all	y	76%	
Citrobacter freundii		P10159	CP012554.1	I,Z,Y	y	80%	
Citrobacter freundii		CAV1321;	CP011612.1;	-all	--	77%	
		CAV1741	CP011657.1				
Serratia sp.		S4	NZ_KB861120.1	Z,Y	n	Z 65%; Y 62%	
Serratia odorifera		DSM 4582	NZ_GG753567.1	Z,Y	n	Z: 52% Y: 65%	
Serratia marcescens		CAV1492	CP011642.1	Z		65%	
Klebsiella sp.		T17-2	KU505145.1	all	y	76%	
Yersinia pestis		CO92	AL590842.1	Z, (Y)	n	63%	Y pseudogene, disrupted by insertion sequence, lacking 19 aa at N-terminal end
Yersinia pestis		KIM10+	AE009952.1	Z, Y	n	Z 62%, Y 67%	Y flanked by insertion element
Pantoea sp. PSNIH2			CP009866.1	I,Z,Y	y	97%	Flanked by IS1 transposase disrupting lacA - missing 473 bp
Pantoea ananatis		AJ13355, LMG 20103, R100, LMG 5342	AP012032.2, CP001875.2, CP014207.1, HE617160.1	Z,Y	y	Z 74%, Y 80%	
Pantoea vagans		B1	CP002206.1	Z		68%	
Cronobacter sakazakii		ATCC BAA-894	CP000783.1	Z,Y	y	66%	
Cronobacter sakazakii		SP291	CP004091.1	I,Z,Y	y	97%	Flanked by IS1 transposase disrupting lacA - missing 473 bp
Cronobacter malonaticus		CMCC45402	CP006731.1	I,Z,Y	y	97%	Flanked by IS1 transposase disrupting lacA - missing last 473 bp.
Raoultella ornithinolytica		Yangling I2	CP013338.1	I,Z,Y	y	97%	Mobile elements disrupting lacA, missing last 473 bp.
Raoultella ornithinolytica		S12	CP010557.1	I,Z,Y	y	73%	

B Appendix 2

Table B.1: **Variability of the *lac* operon across the isolates used in this study.** Numbers and percentages apply to the *lac* operon region as counted from the start codon of *lacI* until the stop codon of *lacA*. 'Promoter region' denotes the complete intergenic region between *lacI* and *lacZ*, which includes the RNA polymerase binding site, the CRP binding site, operator O1 and part of O3.

	length (nt)	variable sites (nt)	pairwise identity (nt)	length (aa)	variable sites (aa)
LacI	1080	227 (21%)	93.8%	360	42 (11.7%)
LacZ	3069	669 (21.8%)	94.3%	1023	176 (17.2%)
LacY	1248	171 (13.6%)	96.9%	416	21 (5%)
LacA	612	114 (18.6%)	95.1%	202	33 (16.3%)
promoter region	122	19 (15.6%)	95.6%		
operator O1	21	0			
operator O2	21	1 (4.8%)			
operator O3	21	3 (14.3%)			
CRP binding site	22	1 (4.5%)			
total lac operon	6262	1241 (19.8%)	95%		

Table B.2: **Percentages of pairwise nucleotide differences for the entire *lac* operon sequence**

	K12	SC1_A5	SC1_F10	SC1_G8	SC1_G10	SC1_H3	M1	M2	M3	M4	M5	M6	M7	TW10509	TW15838	TW11588	TW14182	TW15844	TW09276	TW09231	TW09308
K12		98.8	98.3	97.8	98.4	98.6	98.7	97.4	99.1	97.6	99.7	98.8	98.9	94.6	95	92.7	92.8	92.7	91.7	91.5	90.3
SC1_A5	98.8		98.6	98	98.6	98.5	98.1	97.5	98.3	97.7	98.7	98.7	98.4	94.7	95.2	92.8	92.9	92.8	91.9	91.6	90.3
SC1_F10	98.3	98.6		98	98	98.1	97.6	97.3	97.8	97.7	98.2	97.9	97.9	94.6	95.2	92.6	92.6	92.6	91.8	91.5	90.4
SC1_G8	97.8	98	98		97.8	97.9	97.4	97.4	97.3	98.8	97.8	97.8	97.3	94.8	95.2	92.9	92.9	92.9	91.9	91.8	90.5
SC1_G10	98.4	98.6	98	97.8		98.3	98	97.8	98.3	97.9	98.4	98.2	97.9	94.8	95.1	92.9	92.9	92.8	92.1	91.8	90.5
SC1_H3	98.6	98.5	98.1	97.9	98.3		98.3	97.7	98.3	98	98.7	98.9	97.9	94.8	95.4	92.8	92.8	92.8	92	91.7	90.3
M1	98.7	98.1	97.6	97.4	98	98.3		97.3	98	97.4	98.7	98.3	98.6	94.7	95.1	92.8	92.9	92.9	91.8	91.6	90.2
M2	97.4	97.5	97.3	97.4	97.8	97.7	97.3		97.2	97.5	97.5	97.5	96.9	94.8	95.3	92.7	92.8	92.7	92.1	91.8	90.6
M3	99.1	98.3	97.8	97.3	98.3	98.3	98	97.2		97.5	98.9	98.5	98.3	94.6	95.1	92.5	92.6	92.5	91.7	91.5	90.1
M4	97.6	97.7	97.7	98.8	97.9	98	97.4	97.5	97.5		97.7	97.6	97.2	94.6	95.2	92.6	92.6	92.6	91.8	91.6	90.4
M5	99.7	98.7	98.2	97.8	98.4	98.7	98.7	97.5	98.9	97.7		98.9	98.6	94.7	95.1	92.7	92.8	92.7	91.7	91.5	90.2
M6	98.8	98.7	97.9	97.8	98.2	98.9	98.3	97.5	98.5	97.6	98.9		98.2	94.7	95.1	92.8	92.9	92.8	91.9	91.6	90.2
M7	98.9	98.4	97.9	97.3	97.9	97.9	98.6	96.9	98.3	97.2	98.6	98.2		94.5	95	92.6	92.7	92.7	91.6	91.4	90.3
TW10509	94.6	94.7	94.6	94.8	94.8	94.8	94.7	94.8	94.6	94.6	94.7	94.7	94.5		98.2	91.8	91.9	91.7	91.4	91.3	89.4
TW15838	95	95.2	95.2	95.2	95.1	95.4	95.1	95.3	95.1	95.2	95.1	95.1	95	98.2		92	92	91.9	91.3	91.2	89.7
TW14182	92.7	92.8	92.6	92.9	92.9	92.8	92.8	92.7	92.5	92.6	92.7	92.8	92.6	91.8	92		99.4	99.4	94.8	95	90.6
TW11588	92.8	92.9	92.6	92.9	92.9	92.8	92.9	92.8	92.6	92.6	92.8	92.9	92.7	91.9	92	99.4		99.3	94.6	94.7	90.4
TW15844	92.7	92.8	92.6	92.9	92.8	92.8	92.9	92.7	92.5	92.6	92.7	92.8	92.7	91.7	91.9	99.4	99.3		94.7	94.9	90.7
TW09231	91.7	91.9	91.8	91.9	92.1	92	91.8	92.1	91.7	91.8	91.7	91.9	91.6	91.4	91.3	94.8	94.6	94.7		98.8	90.1
TW09276	91.5	91.6	91.5	91.8	91.8	91.7	91.6	91.8	91.5	91.6	91.5	91.6	91.4	91.3	91.2	95	94.7	94.9	98.8		90.1
TW09308	90.3	90.3	90.4	90.5	90.5	90.3	90.2	90.6	90.1	90.4	90.2	90.2	90.3	89.4	89.7	90.6	90.4	90.7	90.1	90.1	

Table B.3: **# of pairwise amino acid differences in LacI**

	K12	SC1_A5	SC1_F10	SC1_G8	SC1_G10	SC1_H3	M1	M2	M3	M4	M5	M6	M7	TW10509	TW15838	TW11588	TW14182	TW15844	TW09276	TW09231	TW09308
K12		2	3	3	1	0	1	3	0	5	0	0	2	10	11	20	20	19	19	19	28
SC1_A5	2		3	5	1	2	1	3	2	7	2	2	0	10	11	21	21	20	20	20	29
SC1_F10	3	3		2	2	3	2	2	3	4	3	3	3	7	8	18	18	17	17	17	26
SC1_G8	3	5	2		4	3	4	4	3	2	3	3	5	7	8	19	19	18	18	18	27
SC1_G10	1	1	2	4		1	0	2	1	6	1	1	1	9	10	20	20	19	19	19	28
SC1_H3	0	2	3	3	1		1	3	0	5	0	0	2	10	11	20	20	19	19	19	28
M1	1	1	2	4	0	1		2	1	6	1	1	1	9	10	20	20	19	19	19	28
M2	3	3	2	4	2	3	2		3	6	3	3	3	9	10	20	20	19	19	19	27
M3	0	2	3	3	1	0	1	3		5	0	0	2	10	11	20	20	19	19	19	28
M4	5	7	4	2	6	5	6	6	5		5	5	7	9	9	21	21	20	20	20	29
M5	0	2	3	3	1	0	1	3	0	5		0	2	10	11	20	20	19	19	19	28
M6	0	2	3	3	1	0	1	3	0	5	0		2	10	11	20	20	19	19	19	28
M7	2	0	3	5	1	2	1	3	2	7	2	2		10	11	21	21	20	20	20	29
TW10509	10	10	7	7	9	10	9	9	10	9	10	10	10		1	20	20	21	21	21	28
TW15838	11	11	8	8	10	11	10	10	11	9	11	11	11	1		21	21	22	22	22	29
TW11588	20	21	18	19	20	20	20	20	20	21	20	20	21	20	21		0	1	3	3	25
TW14182	20	21	18	19	20	20	20	20	20	21	20	20	21	20	21	0		1	3	3	25
TW15844	19	20	17	18	19	19	19	19	19	20	19	19	20	21	22	1	1		2	2	24
TW09276	19	20	17	18	19	19	19	19	19	20	19	19	20	21	22	3	3	2		2	24
TW09231	19	20	17	18	19	19	19	19	19	20	19	19	20	21	22	3	3	2	2		24
TW09308	28	29	26	27	28	28	28	27	28	29	28	28	29	28	29	25	25	24	24	24	

Table B.4: **# of pairwise amino acid differences in LacZ**

	K12	SC1_A5	SC1_F10	SC1_G8	SC1_G10	SC1_H3	M1	M2	M3	M4	M5	M6	M7	TW10509	TW15838	TW11588	TW14182	TW15844	TW09276	TW09231	TW09308
K12		15	13	18	20	12	19	25	16	23	1	14	18	47	43	52	55	52	59	57	72
SC1_A5	15		18	19	23	13	24	28	27	29	16	13	25	52	49	53	55	53	63	57	74
SC1_F10	13	18		20	27	13	24	30	27	27	14	22	26	48	44	56	57	54	61	58	73
SC1_G8	18	19	20		26	20	29	26	30	18	19	22	28	50	46	53	54	51	60	58	74
SC1_G10	20	23	27	26		24	27	26	16	26	22	23	26	53	49	53	55	53	61	58	74
SC1_H3	12	13	13	20	24		27	29	24	25	13	12	28	51	46	57	58	55	63	58	73
M1	19	24	24	29	27	27		30	27	32	20	25	17	51	48	52	55	51	60	58	69
M2	25	28	30	26	26	29	30		30	29	26	30	29	49	45	54	56	53	59	60	68
M3	16	27	27	30	16	24	27	30		25	17	24	24	46	44	51	54	51	56	53	73
M4	23	29	27	18	26	25	32	29	25		24	28	28	49	45	53	54	51	58	55	68
M5	1	16	14	19	22	13	20	26	17	24		15	19	48	44	53	56	53	60	58	73
M6	14	13	22	22	23	12	25	30	24	28	15		24	52	50	53	56	53	63	58	75
M7	18	25	26	28	26	28	17	29	24	28	19	24		45	42	48	51	47	54	52	66
TW10509	47	52	48	50	53	51	51	49	46	49	48	52	45		22	58	61	58	63	61	77
TW15838	43	49	44	46	49	46	48	45	44	45	44	50	42	22		61	61	58	63	61	73
TW11588	52	53	56	53	53	57	52	54	51	53	53	53	48	58	61		11	8	40	47	77
TW14182	55	55	57	54	55	58	55	56	54	54	56	56	51	61	61	11		5	39	45	75
TW15844	52	53	54	51	53	55	51	53	51	51	53	53	47	58	58	8	5		34	41	72
TW09276	59	63	61	60	61	63	60	59	56	58	60	63	54	63	63	40	39	34		23	75
TW09231	57	57	58	58	58	58	58	60	53	55	58	58	52	61	61	47	45	41	23		73
TW09308	72	74	73	74	74	73	69	68	73	68	73	75	66	77	73	77	75	72	75	73	

Table B.5: **# of pairwise amino acid differences in LacY**

	K12	SC1_A5	SC1_F10	SC1_G8	SC1_G10	SC1_H3	M1	M2	M3	M4	M5	M6	M7	TW10509	TW15838	TW11588	TW14182	TW15844	TW09276	TW09231	TW09308
K12		1	1	0	0	0	1	0	0	3	2	0	0	3	2	3	3	3	3	3	8
SC1_A5	1		2	1	1	1	2	1	1	4	3	1	1	4	3	4	4	4	4	4	9
SC1_F10	1	2		1	1	1	2	1	1	4	3	1	1	4	3	4	4	4	4	4	9
SC1_G8	0	1	1		0	0	1	0	0	3	2	0	0	3	2	3	3	3	3	3	8
SC1_G10	0	1	1	0		0	1	0	0	3	2	0	0	3	2	3	3	3	3	3	8
SC1_H3	0	1	1	0	0		1	0	0	3	2	0	0	3	2	3	3	3	3	3	8
M1	1	2	2	1	1	1		1	1	4	3	1	1	4	3	4	4	4	4	4	9
M2	0	1	1	0	0	0	1		0	3	2	0	0	3	2	3	3	3	3	3	8
M3	0	1	1	0	0	0	1	0		3	2	0	0	3	2	3	3	3	3	3	8
M4	3	4	4	3	3	3	4	3	3		5	3	3	6	5	6	6	6	6	6	11
M5	2	3	3	2	2	2	3	2	2	5		2	2	5	4	5	5	5	5	5	10
M6	0	1	1	0	0	0	1	0	0	3	2		0	3	2	3	3	3	3	3	8
M7	0	1	1	0	0	0	1	0	0	3	2	0		3	2	3	3	3	3	3	8
TW10509	3	4	4	3	3	3	4	3	3	6	5	3	3		1	6	6	6	6	6	11
TW15838	2	3	3	2	2	2	3	2	2	5	4	2	2	1		5	5	5	5	5	10
TW11588	3	4	4	3	3	3	4	3	3	6	5	3	3	6	5		0	0	2	2	7
TW14182	3	4	4	3	3	3	4	3	3	6	5	3	3	6	5	0		0	2	2	7
TW15844	3	4	4	3	3	3	4	3	3	6	5	3	3	6	5	0	0		2	2	7
TW09276	3	4	4	3	3	3	4	3	3	6	5	3	3	6	5	2	2	2		0	9
TW09231	3	4	4	3	3	3	4	3	3	6	5	3	3	6	5	2	2	2	0		9
TW09308	8	9	9	8	8	8	9	8	8	11	10	8	8	11	10	7	7	7	9	9	

Table B.6: **# of pairwise amino acid differences in LacA**

	K12	SC1_A5	SC1_F10	SC1_G8	SC1_G10	SC1_H3	M1	M2	M3	M4	M5	M6	M7	TW10509	TW15838	TW11588	TW14182	TW15844	TW09276	TW09231	TW09308
K12		0	0	5	2	6	7	6	0	5	1	5	0	10	9	9	9	9	15	14	13
SC1_A5	0		0	5	2	6	7	6	0	5	1	5	0	10	9	9	9	9	15	14	13
SC1_F10	0	0		5	2	6	7	6	0	5	1	5	0	10	9	9	9	9	15	14	13
SC1_G8	5	5	5		5	5	6	7	5	0	4	8	5	11	10	10	10	10	15	14	12
SC1_G10	2	2	2	5		4	5	4	2	5	1	3	2	8	7	7	7	7	13	12	11
SC1_H3	6	6	6	5	4		1	4	6	5	5	5	6	8	7	5	5	5	11	10	11
M1	7	7	7	6	5	1		5	7	6	6	6	7	9	8	6	6	6	12	11	12
M2	6	6	6	7	4	4	5		6	7	5	5	6	6	5	7	7	7	13	12	13
M3	0	0	0	5	2	6	7	6		5	1	5	0	10	9	9	9	9	15	14	13
M4	5	5	5	0	5	5	6	7	5		4	8	5	11	10	10	10	10	15	14	12
M5	1	1	1	4	1	5	6	5	1	4		4	1	9	8	8	8	8	14	13	12
M6	5	5	5	8	3	5	6	5	5	8	4		5	8	5	6	6	8	14	13	14
M7	0	0	0	5	2	6	7	6	0	5	1	5		10	9	9	9	9	15	14	13
TW10509	10	10	10	11	8	8	9	6	10	11	9	8	10		7	9	9	9	13	12	17
TW15838	9	9	9	10	7	7	8	5	9	10	8	5	9	7		8	8	10	15	14	15
TW11588	9	9	9	10	7	5	6	7	9	10	8	6	9	9	8		0	2	12	11	14
TW14182	9	9	9	10	7	5	6	7	9	10	8	6	9	9	8	0		2	12	11	14
TW15844	9	9	9	10	7	5	6	7	9	10	8	8	9	9	10	2	2		12	11	14
TW09276	15	15	15	15	13	11	12	13	15	15	14	14	15	13	15	12	12	12		1	16
TW09231	14	14	14	14	12	10	11	12	14	14	13	13	14	12	14	11	11	11	1		15
TW09308	13	13	13	12	11	11	12	13	13	12	12	14	13	17	15	14	14	14	16	15	