Deep characterization of *Escherichia coli*
in a cohort of
mothers and their infants

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1328

ISSN 1501-7710

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika publishing.
The thesis is produced by Akademika publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

ACKNOWLEGMENTS

I would like to thank my supervisors: Nils Christian Stenseth, Kjersti Rønningen, Knut Rudi and Pål Trosvik. Without their guidance and support this thesis would not have been possible. I would also like to thank my co-authors for providing the samples and much needed technical assistance. Further, I would like to thank my colleagues at Nofima, FHI and at UiO. Throughout this scientific journey I relied on both their input and direction. The scientific environment at Nofima and CEES has made this challenging adventure an enjoyable experience. I would also like to thank my family and friends for helping me keep the important things in perspective.

Finally, and in particular, I would like to extend a special thanks to Pål Trosvik. Much of this thesis owes particular credit to him. Not only was the genesis of the project made possible by the techniques that he initially developed. Ongoing scientific collaboration, guidance, technical and analytical support by him made this series of studies both possible and an intense learning experience.

ABBREVIATIONS

Cfu	colony forming units
DAEC	diffusely adherent *E. coli*
E. coli	*Escherichia coli*
EAEC	enteroaggregative *E. coli*
EHEC	enterohemorrhagic *E. coli*
EIEC	enteroinvasive *E. coli*
ENC	effective number of codons
EPEC	enteropathogenic *E. coli*
ETEC	enterotoxigenic *E. coli*
ExPEC	extraintestinal pathogenic *E. coli*
IBD	inflammatory bowel disease
LT	heat-labile toxin
MLST	multilocus sequence typing
PCR	polymerase chain reaction
qPCR	quantitative PCR
ST	heat-stable toxin
sHULK	high μ_{max} and low K_s

LIST OF PAPERS

Paper I

Diversity, transmission and persistence of Escherichia coli in a cohort of mothers and their infants

Eric J. de Muinck, Torbjørn Øien, Ola Storrø, Roar Johnsen, Nils Christian Stenseth, Kjersti S. Rønningen, Knut Rudi

Environmental Microbiology Reports, (June 2011) 3: 352-359

Paper II

Context-dependence in a bacterial community

Eric J. de Muinck, Pål Trosvik, Daniel Sachse, Jan vander Roost, Kjersti S. Rønningen, Knut Rudi, Nils Chr. Stenseth

In revision for the ISME Journal

Paper III

Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche

Eric J. de Muinck, Karin Lagesen, Jan Egil Afset, Xavier Didelot, Kjersti S. Rønningen, Knut Rudi, Nils Chr. Stenseth, Pål Trosvik

Submitted to Genome Biology

TABLE OF CONTENTS

INTRODUCTION

Summary

The bacterial species *Escherichia coli* is still not fully characterized despite being one of the most thoroughly studied organisms. This thesis presents a deep characterization of *E. coli* strains in a cohort of infants and their mothers and extends current understanding of the ecology of this ubiquitous organism. The samples for this study were obtained from Trondheim, Norway, and allows for a much needed geographic perspective that makes it possible to link phylogenetic, ecological, and molecular data with a defined location. The initial nested case-control study was designed to examine the impact of bacterial community colonization on the development of atopic disease in a cohort of infants (Storrø et al., 2010; Storrø et al., 2011). Quantitative polymerase chain reaction (qPCR) was used to identify and quantify the microbial fecal composition of several classes of bacteria in the infants over time and this was matched with cytokine profile development. From this work it was found that early *E. coli* colonization in this cohort was linked to protection from atopy and that the mother was a likely source of the infant colonization (Rudi et al., 2012). These findings, as well as the plethora of tools available for the study of *E. coli* led us to focus on this species and characterize its colonization patterns within this cohort of infants.

This thesis develops methodology and then characterizes population structure and dynamics of *E. coli* colonization within the larger study framework. We first developed a simple and novel technique that allowed us to uncover limits on the diversity of colonizing strains and found evidence of transmission from the mothers to the infants (Paper

I). We placed these colonizing *E. coli* strains into a phylogenetic context and further placed these strains into overall *E. coli* diversity. This allowed us to understand and compare strains colonizing infants in a defined geographic area with the wider population structure of this species. Additional investigation found differences in growth characteristics of the *E. coli* strains that were either early or late colonizers of the infant gut (Paper II). *In vitro* competition studies revealed potential mechanisms that modulate strain competitive dynamics. Finally, through genome sequencing, we compared several phenotypic characteristics using differential gene content in order to determine enrichment profiles that may explain these traits (Paper III). Enrichment comparisons included: phylogenetic, pathogenic *vs.* commensal, growth rate, and early or late colonization. The signatures we found can be used for further investigations into genotype-phenotype connections within *E. coli* strain ecology. Overall, we developed much needed insight into modern colonization patterns in a geographically defined cohort.

Brief introduction to E. coli

E. coli was first isolated from the feces of a newborn in 1885 by Theodor Escherich and is a gram negative, facultative anaerobic bacillus that is able to use glucose as a sole carbon source for growth (Escherich, 1989). It belongs to the Domain *bacteria*, Phylum *Proteobacteria*, Class *γ-Proteobacteria*, Order *Enterobacteriaceae*, and Genus *Escherichia* of which there are seven members. The primary habitat is believed to be the animal gastrointestinal system, however, this should and has been extended to include extra-intestinal environments (Luo et al., 2011). The ease of culture and ubiquity of *E. coli* has lead to its usefulness as a molecular biology workhorse.

Extensive basic research using *E. coli* as a model organism has also afforded rare insight into the molecular mechanisms of its characteristics. Today, a Pubmed search for '*Escherichia coli*' returns 294,153 results. Much of this is basic research but a large part is devoted to the understanding of the pathogenic role of *E. coli* and the characteristics that drive this normally commensal organism towards pathogenesis (Kaper et al., 2004). Limited work has also looked into a probiotic role and the ability of some strains to directly protect the host from pathogens or modulate immune responses to help the host to maintain health (Fuller, 1989).

E. coli is a diverse bacterial species and encompasses a large number of strains. The general diversity of *E. coli* lies mostly in commensal strains of the gut (Tenaillon et al., 2010). This diversity is typically divided into clades or groups A, B1, B2, D, and sometime including E and F (Jaureguy et al., 2008). Humans are thought to be mostly colonized by *E. coli* of the B2 and A groups while B1 derives from domesticated animals, although this depends some on the geography (Tenaillon et al., 2010).

E. coli Genomics

Much genomic information of *E. coli* has already been collected and Genbank has cataloged 60 chromosomal genomes and 346 scaffolds or contigs (as of this writing) with most of the sequencing effort directed toward pathogenic strains. The use of these full genomes has become important for understanding the role of differential gene content in determining a realized ecological niche (Luo et al., 2011). Previous comparative analysis of the genomes of 61 isolates (Lukjancenko et al., 2010) has further developed a new view of the *E. coli* community

structure that highlights diversity: at the genome level, on average, *E. coli* is only 20% core and 80% non-core. The limitations of MLST make consistent fine scale architecture of phylogeny difficult (Sahl et al., 2012) while whole genome sequencing has offered accurate phylogenetic placement of the *E. coli* strains.

The comparison of 61 genome-sequenced strains showed that the total genome sizes range from ~4.5 Mb to almost 6 Mb containing ~4,200-~6,000 genes (Lukjancenko et al., 2010). The biological functions for many of these genes are still unknown. For example, from the genomes of the canonical K-12 MG1655 and derivative W3110 *E. coli* strains, 2,403 or 54% of the genes have known functions based on experimental data; 1,425 (32%) are genes that are only computationally predicted and the remaining 616 (14%) 'genes' are categorized as unknown (Riley et al., 2006). This data only represents a small fraction of the number of genes within the species. Due to its enormous genetic diversity, as little as 20% of the genome is common to all strains (Lukjancenko et al., 2010), and the core genome of *E. coli* as a species is estimated to be between 900 and ~2,000 genes whereas the pan-genome of all *E. coli* strains is estimated to be 18,000 genes and growing as more strains are sequenced.

E. coli in the human gastrointestinal system

Commensal

Advances in technology have changed our understanding of the microbial community inhabiting the gastrointestinal system (Zoetendal et al., 2004). Due to its ease of culture from the fecal samples, *E. coli* has long been thought to have a solid place in this large and complex community. However, 16S rRNA gene sequencing has since shown

that only about 10-50% of the gut microbial species are cultivable using current knowledge. Infants are born sterile and, under normal circumstances, *E. coli* is one of the early colonizing members (Bettelheim and Lennox-King, 1976). The neonatal gut is rich in oxygen and promotes the establishment of aerobic and facultative anaerobic organisms (Adlerberth, 2008). These deplete available oxygen and promote the growth of obligate anaerobic organisms which then come to dominate. Broad colonization patterns of the infant over time show *Proteobacteria* expand to maximal relative abundance at about four months (Koenig et al., 2011). In the mature community, *E. coli* is outnumbered by anaerobic bacteria by a factor of 100 to 10,000 and constitutes only a small fraction (0.1%) of the relative abundance. Nevertheless, *E. coli* is the predominant facultative anaerobe in the gastrointestinal tract and adults carry about 10^8 cfu per gram of feces (Tenaillon et al., 2010).

The ecological role of *E. coli* in the gut microbial community is less clear. We know little about the effects of *E. coli* presence on the community structure as a whole. Trosvik et al. (Trosvik et al., 2010a) addressed this issue in a simplified model gut system that included representatives of four of the main gut bacterial groups: *Bacteriodes thetaiotaomicron, Bifidobacterium longum, Clostridium perfringens,* and *Escherichia coli*. By following the experimental community over time, an interaction map of these groups was established showing that *E. coli* abundance was negatively affected by the abundances of *C. perfringens* and *B. thetaiotaomicron* (Figure 1).

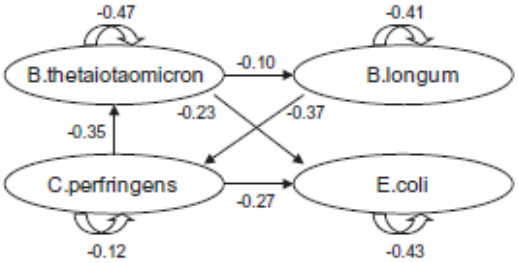


Figure 1. Interaction map of four model bacteria species in a chemostat system. Reproduced from Trosvik et al. (Trosvik et al., 2010a)

This seemed to follow some of the key patterns often observed in the actual gut environment as both of these classes of bacteria are extremely dominant in the mature gut. Dysbiosis of this normal relative abundance of these groups has been linked with diseases such as diabetes, inflammatory bowl diseases (IBD), allergy, and obesity (Frank et al., 2007; Giongo et al., 2011; Larsen et al., 2010; Turnbaugh et al., 2006). Specifically, decreases in Firmicutes (represented by Clostridium in the previous model community) and Bacteriodetes with a corresponding increase in Proteobacteria (represented by *E. coli* in the model) are associated with IBD. Others found increases in specific *E. coli* strains are associated with IBD (Baumgart et al., 2007). Since this work, there has been increased effort to understand the mechanism of the interaction between the host and bacteria and to characterize the strains associated with IBD (Chassaing et al., 2011; Kotlowski et al., 2007; Rolhion and Darfeuille-Michaud, 2007; Sepehri et al., 2009).

Pathogenic E. coli

In addition to being a well-known commensal, *E. coli* is also a pathogenic species that may cause fatal diseases. One of the first reported cases of *E. coli* causing disease is an outbreak of diarrhea among infants in 1935 (Merritt and Paige, 1935). The diarrhea-inducing *E. coli* pathogens are divided into six classes depending on strain characteristics that result in distinct features in pathogenesis: enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC), enterohemorrhagic *E. coli* (EHEC), and diffusely adherent *E. coli* (DAEC) (Kaper et al., 2004; Todar, 2012). Extraintestinal pathogenic *E.coli* (ExPEC) infect a variety of other tissues outside the intestinal system but are related to both commensal and pathogens of the gastrointestinal tact as many of the factors that allow a strain to become virulent are also important as general fitness factors for gastrointestinal colonization (Pitout, 2012).

Enterotoxigenic *E. coli* (ETEC) is an important cause of diarrhea in infants and travelers in underdeveloped countries of poor sanitation. The bacteria colonize the GI tract by means of fimbrial adhesion molecules, e.g. CFA I and CFA II, and are noninvasive, but cause pathology by producing one or more plasmid-encoded enterotoxins. Enterotoxins produced by ETEC include the LT (heat-labile) toxin and/or the ST (heat-stable) toxin. The LT enterotoxin has an enzymatic activity that is identical to that of the cholera toxin and binds to the same intestinal receptors that are recognized by the cholera toxin. ST causes an increase in cyclic GMP in host cell cytoplasm which in turn leads to secretion of fluid and electrolytes resulting in diarrhea.

Enteroinvasive *E. coli* (EIEC) closely resemble *Shigella* in their pathogenic mechanisms and clinical presentation they produce. The clinical symptoms resemble *Shigella* dysentery and include a dysentery-like diarrhea with fever. Like *Shigella*, EIEC are invasive organisms and they penetrate and multiply within epithelial cells of the colon causing widespread cell destruction. Unlike *Shigella*, they do not produce the shiga toxin, or the LT or ST toxins.

Enteropathogenic *E. coli* (EPEC) induce a watery diarrhea similar to ETEC, but they do not possess the same colonization factors and do not produce ST or LT toxins. Rather, EPEC strains adhere to the intestinal mucosa through a complicated process and produces dramatic effects in the ultra-structure of the cells resulting in rearrangements of actin in the vicinity of adherent bacteria. The diarrhea and other symptoms of EPEC infections are probably caused by inflammatory responses of host cells to bacterial invasion and interference with normal cellular signal transduction, rather than by production of toxins.

Enteroaggregative *E. coli* (EAEC) attach to tissue culture cells in an aggregative manner. The significance of EAEC strains in human disease is controversial but it has been associated with persistent diarrhea in young children. They resemble ETEC strains in that the bacteria adhere to the intestinal mucosa and cause non-bloody diarrhea without invading or causing inflammation.

Enterohemorrhagic *E. coli* (EHEC) is represented by a single strain (serotype O157:H7), which causes a diarrheal syndrome distinct from

EIEC (and *Shigella*) in that there is copious bloody discharge and no fever. Pediatric diarrhea caused by this strain can be fatal due to acute kidney failure (hemolytic uremic syndrome [HUS]). The bacteria do not invade mucosal cells as readily as *Shigella*, but EHEC strains produce a toxin that is virtually identical to the Shiga toxin. The toxin plays a role in the intense inflammatory response produced by EHEC strains and may explain the ability of EHEC strains to cause HUS.

Diffusely adherent *E. coli* (DAEC) induces a characteristic, diffuse pattern of adherence to HEp-2 cell monolayers, a human epithelial cell line derived from larynx carcinoma. DAEC express fimbrial adhesin molecules that bind receptors on the intestinal epithelial cells which in turn induce inflammatory responses and cytopathic effects. DAEC has been implicated in diarrhea in children over 1 years of age.

Extraintestinal pathogenic *E.coli* (ExPEC) can cause pathologies in a several tissues in addition to infections of the intestine (Russo and Johnson, 2000). Some common targets of infection include the urinary tract, meninges and intra-abdominal areas and these infections are often accompanied by bacteremia. Although these strains have been previously claimed to be completely distinct from commensal strains, they have subsequently been shown to have overlapping properties with gastrointestinally colonizing *E.coli* (Diard et al., 2010; Le Gall et al., 2007). Examples of this overlap can be observed in cases such as when a urinary pathogenic *E.coli* strain lives commensally in the gut (Foxman, 2010), or the presence of virulence factors such as Type 1 fimbrea (Nielubowicz and Mobley, 2010). Type 1 fimbrea are a well characterized virulence factor almost always found in human uropathogenic *E. coli* strains but also often found in intestinally derived

strains as well. On the genome level, ExPECs may exhibit a clearer distinction from commensal isolates than seen in comparisons of the broad category of pathogenic *E.coli* with commensal strains (Brzuszkiewicz et al., 2006; Chen et al., 2006; Moriel et al., 2010). However, this genetic distinction is related with the specific pathology and host history of the strain (Rasko et al., 2008; Touchon et al., 2009)

Importantly, although it has been well established that many types of *E. coli* can cause disease, the dividing line between pathogenic *E. coli* and commensal *E. coli,* as was seen in our genome analysis and will be discussed further, is blurry. Strains containing well-characterized virulence determinants such as the LEE pathogenicity island can be isolated from healthy individuals. In the clinical setting, the cause of an "infection" is attributed to a pathogenic strain because that is what was found in the patient's stool sample. However, gastroenteritis can be caused by a myriad of agents (Alter et al., 2011; Wilhelmi et al., 2003).

Probiotic E. coli

Even less well understood are the possibilities of *E. coli* as a probiotic. The role of a probiotic is threefold. A probiotic can protect the host from infection by a pathogen by direct competition for a particular niche. A probiotic can interact with the host to make the host less susceptible to infection or some other diseases. A probiotic can also protect one host, e.g. humans, from a pathogen by displacing that pathogen in a non-susceptible host, *e.g.* ruminants that carry the human pathogens as commensals. *E. coli* as a probiotic has been

involved in all three roles either alone or in some combination of these effects.

The most studied probiotic *E. coli* is marketed under the name Mutaflor®. This strain, also called Nissle 1917 or DSM 6601, was isolated by Professor Alfred Nissle from a German soldier who remained healthy during an outbreak of Shigellosis during the First World War (Nissle , 1918). Since then, this strain has been used for addressing a variety of conditions and has been the subject of many studies (Pubmed search yields 146 hits). The long list of treatable conditions includes protection from other pathogens (Altenhoefer et al., 2004), maintaining remission or treatment of ulcerative colitis, irritable bowel syndrome, and constipation(Kruis et al., 1997; Kruis et al., 2004; Kruis et al., 2012). The ability of this strain to perform all of these functions seems to be a combination of itself being an effective colonizer that outcompetes other bacteria, the antimicrobial peptides it produces (Patzer et al., 2003) and the wide variety of stimulating interactions with the host immune system such as cytokine production, increased secretion of IgA, mucin and human β-defensin-2 induced by this strain (Jacobi and Malfertheiner, 2011).

E. coli has also been linked to long term protection from allergy. In addition to the IMPACT study associated with this work (Rudi et al., 2012) that found early colonization by *E. coli* in general protect later development of allergy, other studies have linked early colonization by a probiotic *E. coli* with allergy protection (Frank et al., 2007; Kim et al., 2005; Kocourkova et al., 2007; Lodinova-Zadnikova et al., 2003; Lodinova-Zadnikova et al., 2010; Penders et al., 2007; Weise et al., 2011). The exact mechanism of this protection is not well understood

but the hygiene hypothesis suggesting that a lack of early childhood exposure to microorganisms increases susceptibility to allergies due to suppressed development of the immune system (Strachan, 1989), is an actively pursued hypothesis.

The third source of protection that probiotic bacteria can afford is depletion of a pathogenic bacterial species in a separate host population. Cattle feedlots contain endemic populations of O157:H7 *E. coli* that are highly pathogenic to humans but benign to the cattle. Efforts have been made to outcompete these pathogenic populations with other species of *E. coli* in cattle (Schamberger et al., 2004; Zhao et al., 1998).

Genomic comparative analysis of four different probiotic E. coli strains have found that that not surprisingly, they are more related to a non-pathogenic commensal strain (K12) than to a pathogenic EHEC strain (Willenbrock et al., 2007). Importantly, no virulence genes were detected in the probiotic isolates apart from one hemolysin gene that in itself was not sufficient to characterize an isolate pathogenic. Each probiotic strain also contained ~100 unique genes not found in the control genomes (K12 and an O157:H7 EHEC strain) and a few of them were predicted to have general metabolic functions. A closer analysis will be needed to assess whether some of these genes may provide improved fitness for colonization for these probiotic strains.

Ecological theory and the gut microbiota

Our knowledge about the types of ecological interactions that occur between species in the gut is very limited. We also know little about the relative contribution of the bacterial-bacterial interaction and host-

bacterial interactions (Costello et al., 2012). The ecological development of the gut microbiota and the control of its ultimate structure can be divided into three parts: initial colonization, host-bacterial interaction, and bacterial-bacterial interactions.

Initial colonization

As human, we provide habitat for a set of distinct 'microbiomes' within an individual. Four proposed scenarios for colonization of these sites have been proposed:

a) environmental selection: habitats with initially similar conditions select for similar assemblages. This could account for different bacterial assemblages in different body sites such as between the skin (aerobic) and intestine (anaerobic) as each habitat will select for organisms with distinct abilities.

b) historical contingency: habitat does not control colonization, timing and order of colonization determine community structure. The same body site in different individuals provides similar habitat and would support similar communities so that difference in the communities is only determined by timing of exposure to different colonizers.

c) random sampling: random draws from the species pool determine the final community. In contrast to Costello et al.,(Costello et al., 2012) who find this may explain differences between monozygotic twin colonization, I would assert that this is less important except as embedded within the other scenarios or for extremely transient colonization.

d*) dispersal limitation: local communities determine the population that the novel host has access to for colonization*: In contrast to historical microbial theory that assumes "everything is everywhere and the environment selects" (O'Malley, 2007), this theory states that the available species pool for colonization is limited by local availability.

Host-bacterial interactions

Non-immune control

The primary way that the gastrointestinal tract controls the bacterial community structure is by its physical parameters. A colonizing microbe has to survive the pH and digestive enzymes of the stomach and then the anaerobic environment maintained by the intestine. In addition, the intestine is a dynamic place with layers of viscous mucus and intestinal epithelial cells that are constantly turning over and being sloughed off. Finally, host diet affects bacterial community structure and host secreted nutrients can promote the growth of certain species (Garrido et al., 2012).

Immune interactions

The complete immune system requires stimulation by a colonizing microbiota for proper development (Hooper et al., 2012). In germ-free mice, gut-specific lymphoid structures, secretory IgA and CD8αβ intraepithelial lymphocytes all fail to develop normally. One of the main goals of immune control over the microbial community is the containment of bacteria and related immune responses to the intestine, preventing their spread to systemic sites. Important tools include defensins, antimicrobial peptides secreted by the host, IgA secretion that binds to bacteria and prevents their crossing of the intestinal epithelium, and cytokines secreted by T cells and innate

lymphoid cells. Secreted anti-microbial proteins can not only keep the microbiota within confined locations, certain members such as the human α-defensin-5, can even shape the overall community composition. The intestinal immune system actively sample luminal bacteria content and produce protective secretory IgAs against commensals. In contrast, pathogenic bacteria penetrate to the systemic secondary lymphoid tissues and elicit a systemic immune response characterized by IgG production (Hooper et al., 2012).

Interestingly, the commensals shape the host immune responses which in turn may affect susceptibility for infectious or autoimmune diseases. The most striking example is the dependence of one important pro-inflammatory T cell response known as the T_H17 response, on the presence of one particular commensal bacteria in the murine gut, the segmented filamentous bacteria (Gaboriau-Routhiau et al., 2009; Ivanov et al., 2009). Other examples include expansion of immune-modulating systemic T_{reg} response by certain Clostridial strains (Ivanov et al., 2009), and induction of IL-10 by polysaccharide A of *Bacteroides fragilis* (Round et al., 2011).

Bacterial-bacterial interactions

In addition to the colonization and host forces determining the bacterial composition in the gut, bacterial-bacterial interactions among the trillions of bacteria and hundreds to thousands of species must contribute in shaping species content and relative abundance (Faust and Raes, 2012). Indeed, Trosvik and colleagues (Trosvik et al., 2010b) showed that main group bacterial interactions during infant colonization could be predictive of later community structure.

Interactions between microorganisms can result in win-win, win-loss, or in rare cases a neutral relationship. Detecting these types of relationships within the complex ecosystem of the gut requires two complementary approaches. One requires studying the *in situ* relationships operating within the natural system such as performed by Trosvik (Trosvik et al., 2010b). The other is to apply well-controlled *in vitro* experiments that can parse out the specific relationships that are possible between members of the microbiota in simplified systems (Paper II).

Most studies are designed to provide snapshots of the bacterial community profile among cohorts (Ley et al., 2006; Yatsunenko et al., 2012). While these descriptive studies are creating a benchmark for understanding general diversity, they do not take into account the temporal variation inside individuals. We lack robust models of bacterial interactions in the human gut and an understanding of the consequences of these interactions on the bacterial community structure as a whole. This is mostly due to a deficit of data sets that are appropriate for rigorous statistical treatment. Thus far, the largest time series of gut colonization is daily sampling of two adult individuals, one for 15 months and the other for 6 months (Caporaso et al., 2011). Infant gut dynamics are notably more diverse than adult dynamics (Palmer et al., 2007). Despite this variation, as our group found, these dynamics can significantly influence the final stable community structure (Trosvik et al., 2010b). Another study has analyzed 60 samples collected over 2.5 years from a single infant (Koenig et al., 2011). They presented strong evidence that colonization patterns followed distinct stages and that it was seemingly

"nonrandom" and determined by ecological interactions. However, the nature of these interactions was left undescribed.

In addition to work describing ecological profiles of microbial communities, there is a rich and long history of applying *in vitro* experiments using microorganisms to understand the basic principles that may generate these various structures (Gause, 1934). These experiments allow the investigator to operate with relatively short time scales, large populations and the opportunity for replication (Buckling et al., 2009). In our work, we have performed experiments using gut bacterial strains that have been isolated from human infants in order to generate models of interaction that can be used to understand more complex community structures.

In vitro competition results that we observed in paper II invoke a relationship that we defined as sHULK, (high μ_{max} and low K_s) that would allow two strains to be competitively superior in different phases of the batch culture. The acronym derives from the Monod equation $\mu = \mu_{max}s/(K_s+s)$(monod, 1942) where the outcome of competition is determined by the maximum growth rate (μ_{max}) and the Monod constant (K_s), where K_s is the nutrient concentration at which a species has a growth rate of $\mu_{max}/2$ (Figure 2). In the special case where one species has a higher μ_{max} while the other has a lower K_s, there exists a nutrient concentration (s) at which the growth rates are equal. Below that concentration, the species with the lower K_s will win, while higher concentrations favor the species with the higher μ_{max}. The exact value of s permitting coexistence is for practical reasons very difficult to achieve in a chemostat. In theory, and in contrast to a chemostat system, this relationship in conjunction with the fluctuating nutrient

levels inherent in a batch culture can be used to promote coexistence of strains on a single nutrient (Rainey et al., 2000; Stewart and Levin, 1973).

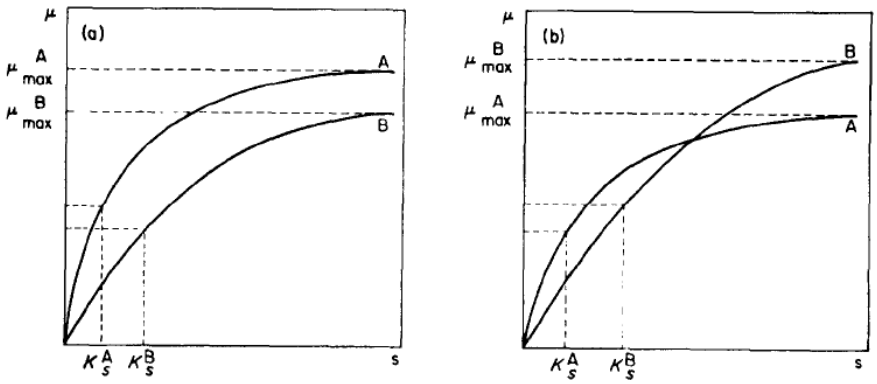


Figure 2. *μ-s* relationship of two organisms A and B. (a) $K_{s(A)} < K_{s(B)}$ and $\mu_{max(A)} > \mu_{max(B)}$, (b) $K_{s(A)} < K_{s(B)}$ and $\mu_{max(A)} < \mu_{max(B)}$. From (Veldkamp, 1970)

These dynamics are a special case for pairs of species in which one (sometimes called a "gleaner") has a higher growth rate at low nutrient concentrations and another (sometimes called an "exploiter" or "opportunist") has a higher growth rate at high nutrient concentrations (Fredrickson and Stephanopoulos, 1981; Gottschal, 1993). In fact, some experimental evidence does support this type of proposed interaction (Grover, 1997; Levin, 1972). However, much controversy still remains concerning the conditions that lead to coexistence (Abrams, 2004).

PAPERS

Paper 1

In the first part of the thesis we utilized a novel, growth-independent, direct typing approach to describe *E. coli* mother-to-child transmission and persistence within infants in a well-defined cohort. This technique has since been used to create strain resolved analysis of *Bacteriodes fragilis* in the same cohort (Bjerke et al., 2011). We then performed seven gene multilocus sequence typing (MLST) of 28 *E. coli* study isolates, three probiotic strains, eight Norwegian pathogenic isolates plus the 72 strains of the reference ECOR strain collection (Ochman and Selander, 1984), which added a phylogenetic framework to the direct sequencing data. We found that a type B2 subpopulation of the maternal *E. coli* strains was the main group transmitted to the infants and that the proportion of children carrying their mother's strain decreased as the children aged. Using species richness estimates we also found a limited number of strains within the cohort compared with the total *E. coli* diversity, constraints on infant colonization, and that infant strain diversity levels increased towards maternal diversity levels over time. These results support the idea of 'dispersal limitation' having a strong effect on potential colonization patterns. This is supported by other work that found differences in initial *E. coli* colonization rates between vaginal and caesarean section delivered infants (Nowrouzian et al., 2003).

The direct typing approach that we developed used a single gene to differentiate strains of *E. coli*. This gene was amplified directly from the stool samples and polymorphisms between the strains in this particular gene were used for quantification of relative abundance. However, this method does have limitations. First, strain differentiation

is dependent on polymorphisms in this single gene. In addition, the relative abundance measurement using sequencing reactions of this gene has an effective detections threshold of 10%. This means that we can determine fine scale relative abundance differences but not the log-fold differences that serial dilution techniques are better suited for measuring. Even in the world of high-throughput genomics, this technique still has inherent value that can be applied in a clinical setting. *E. coli* are of low relative abundance in the human gut microbiome and it will require high metagenomic coverage and extensive bioinformatic analysis in order to achieve strain resolution (Morowitz et al., 2011).

Paper 2

Very little is known about how biotic interactions influence community dynamics and the ecological processes that generate the establishment and maintenance of a normal microbial community in the human gastrointestinal tract. In this study we investigated isolates derived from one particular infant. These isolates were obtained from an infant at day ten, four months and two years of age. We used these strains and a simplified model microbiota to investigate intra-species competition and demonstrate that the colonization process on the strain level can be context dependent, in the sense that the outcome of intra-specific competition may be determined by the composition of the background community. Further experimentation outlined a possible mechanism by which strain preponderance may be modulated through biotic interactions with distantly related species.

The intra-specific competitions generated two seemingly contradictory sets of observations that can be explained by two different models. We

observed two different competitive outcomes between the three strains studied. One strain (EDM106) had a shorter generation time at low nutrient levels than the other two strains (EDM116, EDM530) while at high nutrient levels EDM106 had a longer generation time than the other two. When EDM106 and EDM530 were inoculated together, EDM530 outcompeted EDM106 under high nutrient conditions whereas EDM106 outcompeted EDM530 under low nutrient conditions. However, when EDM106 and EDM116 were inoculated together in a high nutrient concentration regime they coexisted.

Our results from the EDM106 and EDM530 strain competitions are consistent with the sHULK model for competition between organisms where one competitor has adapted to low resource and high stress environments whereas the other is optimized for rapid reproduction when resources are abundant. The inherent limitations of our model system make it all the more surprising that we found some of the same competitive outcomes that were observed in the gut environment. Additionally, genomic profiles of *E. coli* strains representing these differing ecological strategies provided clues for deciphering the genetic underpinnings of niche adaptation within a single species. Several pathways were identified in the genomes that could have influenced competitive outcomes and suggest further investigations to solidify a bridge between gene content and competitive outcomes in the natural environment. It is difficult to ascribe these results to particular genes or pathways because of the number of unique genes that are annotated as hypothetical proteins in each of the strains. However, we do provide genomic profiles of an ecological gleaner in comparison with two exploiter phenotypes.

The stabile co-existence and increased carrying capacity relative to EDM106 when EDM106 and EDM116 are co-cultured suggests that some form of cross-feeding is occurring. This invokes a second model that is based on the observed increase in carrying capacity of the co-culture. If between-strain competition is weaker than within-strain competition there is a theoretical equilibrium point for coexistence. This state can be brought about by cooperative interactions like cross-feeding, resulting in increased productivity in the co-culture. Additionally, the fact that these strains were found to coexist in the infant gut, as well as in co-culture suggests that they could occupy overlapping niches in their natural environment, and that the observed interaction is ecologically relevant.

We also present evidence of context dependent competition in bacteria in Paper II, and we propose mechanisms that can promote this phenomenon. The previously described competition outcomes between EDM106 and EDM530 could be altered by the presence of Clostridia but not by the presence of Bacteriodes. We can easily envision two scenarios in an actual gut where one of the two *E. coli* strains outcompetes the other, depending who is dominating the background community (Clostridia or Bacteroides).

When all of our competition results were taken together, we were able to replicate some of the outcomes of strain competition observed in the actual infant gut. This does not mean that these were the actual factors responsible for the outcomes in that complex environment. The large number of species and need for further information preclude definitive conclusions. Our findings however, do extend the role of

ecological theory in understanding microbial systems and the conceptual toolbox for describing microbial community dynamics.

Paper 3

Together with Jan Egil Afset at NTNU, we sequenced twelve commensal and four pathogenic strains from the previously described cohort. We compared differential gene content in order to determine enrichment profiles that may explain phenotypic traits. We found signatures that relate to phylogeny, early *vs.* late colonization, pathogenicity, and growth rate characteristics that show comparable enrichments in biological processes but use different genetic elements. Embedded in the sequencing were two pairs of strains isolated from the same infant and that were clonally related. Genome sequencing revealed gene content and codon use changes that could be attributed to adaption to the host or other microbes. Methodological challenges included alleviating potential bias in our gene content comparisons between genomes due to fragmented assembly of the genomes by 454 sequencing.

There are three important categories of findings in this paper. The first is the gene content enrichment profiles for the individual phenotypic categories of: phylogeny, early *vs.* late colonization, pathogenicity, and growth rate. Additionally, we found genomic profiles that are associated with evolution towards a late colonizer using two lines of evidence. "Re-sequencing" of a strain that had been in an infant for four months showed that three genes that were present in the ancestral strain and that all belonged to the early enrichment genes, were lost from the evolved version. These genes included a tellurite resistance protein that has been linked to resisting host defense

(Morowitz et al., 2011; Taylor, 1999). Secondly, we observed an increased anaerobic generation time of the isolates of the same strain from the same infant four month later. In addition, the evolved EDM123c had an elevated genome-wide Effective number of codons (ENC) (and thus also ΔENC) relative to the parent strain. This indicates a selection pressure for synonymous mutations toward reduced codon usage bias from the parent to the evolved strain. Reduced codon bias and growth rate have previously been associated with late gut colonization (Vieira-Silva and Rocha, 2010), suggesting that isolate EDM123c has in fact evolved toward a late colonizer profile.

Lastly, and perhaps most importantly, we found that strains use different genetic elements to attain enrichments for similar biological processes. There were several instances where clear gene content enrichment profiles could be linked to specific phenotypes. When the lists of genes in these enrichment profiles were categorized into biological processes, strong similarities between the enrichments arose. This suggests that there could be strong selection towards a defined niche for *E. coli* in the human gut. Nevertheless, many genetic pathways are available to achieve this and fine scale specialization can still direct the evolution of strains. In contrast to previous studies of *E. coli* eco-genomics (Didelot et al., 2012; Lukjancenko et al., 2010; Rasko et al., 2008; Touchon et al., 2009), our isolates come from a population that is narrowly localized both temporally and geographically. This could result in reduced genetic diversity in our samples due to shared ancestry and increased exchange of genes through horizontal transfer (HGT) between strains. We were not particularly interested in HGT but we did see a substantially higher percentage of shared gene content (52.4%) than what has previously

been reported, as well as a smaller pan-genome, indicating that homogenizing forces are increasingly affecting genomic diversity on a local scale. The more homogenous genomic background, as seen in this work, could make it easier to tease out gene content signatures that are ecologically relevant.

Future perspectives

This thesis generates several lines of inquiry that can be useful for uncovering different aspects of *E. coli's* natural history and ultimately be useful in a clinical and basic science perspective. We still do not understand the emergence of pathogenic *E. coli* from its commensal origins even though this is a much studied organism that accounts for billions in health care costs each year (Russo and Johnson, 2003). A larger understanding of commensal *E. coli* strains would shed light on the relationships between gene flow, genetic background, host susceptibility, population structure and how these relate to disease.

Direct sequence typing and clinical diagnostics

The direct sequencing methods developed in the first paper would facilitate rapid evaluation of infectious samples. Modern clinical practice still requires culture to identify bacteria in an infection. The well-known differences in cultivability of different species and strains within a species can bias the colony distribution growing on the plate. Further, it is likely that an individual is colonized by more than one strain of *E. coli* as we and others have reported (Nowrouzian et al., 2003). This cultivability bias, in conjunction with multiple colonizing strains, makes it difficult to ascertain whether the strains that have been isolated accurately represent the true colonization pattern of the

patient. Direct typing using single-gene-polymorphisms offers a simple procedure for validating these culture results.

The gene that we chose for direct sequence typing was malate dehydrogenase *mdh* for reasons outlined in the paper. Since publication of the paper, comparative genomics has found many other candidate genes that could be more informative with regard to increased diversity or phylogenetic accuracy (Sahl et al., 2012). As more genomes are sequenced, continued analysis of these will most likely find even more informative genes.

Context dependence in microbial ecology

Context-dependent competition most likely represents a general phenomenon where community composition at high taxonomic levels determines the outcomes of strain level colonization processes by remodelling the environment to become more permissive to some strains than others. In a system, such as the gut, where a high degree of exploitation competition takes place, the ability of keystone taxa to remodel the biotic environment may have profound effects on community structure. There are few, if any, concrete examples of context-dependent competition on a single trophic level as presented in Paper II. However, this phenomenon can have potentially dramatic effects on which bacteria will successfully establish and persist in the gastrointestinal system, and the principle should be equally applicable to other microbial ecosystems.

Understanding the population ecology of gut bacteria and competition effects across phyla is important because of the growing use of antibiotics and probiotics without consideration of possible cascading

effects (Costello et al., 2012). Results from the second paper could lead to better understanding of important phenotypic characteristics that will lead to more effective probiotic choices, an increasingly important avenue of investigation with the rise of antibiotic resistance. In addition, we believe that extending competitions presented in the paper would lead to complex dynamics that could be modeled and further extend the theoretical understanding of bacterial competition.

Gene-content profiles and genotype-phenotype mapping

Even for the most well characterized genome (*E.coli* K-12) only half of the genes have a function defined by experimental evidence. The remaining genes have purely hypothetical functions or are completely undefined. The approach that we present in the third paper links gene sets with phenotypes which can serve as a starting platform for extending the known functions of genes and assigning functions to previously uncharacterized genes. A larger scale and more systematic approach that applies differential gene content profiling to a series of phenotypic responses to environmental conditions would relate the functional role of genes to ecological variables.

Correlating IMPACT E. coli gene-content profiles to allergy protection

Results from the third paper suggest using the same IMPACT sample collection for investigating colonizing *E. coli* gene content profiles for protection from allergy. Early colonization by *E. coli* was linked to protection from allergy but specific strains were not identified using the direct-sequencing method. By choosing appropriate samples for a case-control study design, we could that take into account *E. coli* gene repertoire to determine profiles and important genetic signatures for protection.

The era of 'population genomics'

We and others have discovered (Lukjancenko et al., 2010) that phylogeny may not be an optimal predictor of important phenotypoic properties such as clinical manifestation. In paper III of this thesis pathogens were separated on two different deep branches whether we used the core or pan-genome for tree construction. However, in the pan-genome tree some of the pathogens formed a tight cluster along with a commensal strain. This suggests that even though virulence can emerge from very different genomic backgrounds, there may still be gene content signatures that are predictive of virulence potential. Commensal isolates that carry the genomic signature of a pathogen could have increased potential for causing disease given the right set of circumstances. Adequete populations scale genomic monitoring of populations of commensal bacteria could provide a predictive framework for implenting preventive strategies.

REFERENCES

Abrams,P.A. (2004) When does periodic variation in resource growth allow robust coexistence of competing consumer species? *Ecology* **85:** 372-382.

Adlerberth,I. (2008) Factors influencing the establishment of the intestinal microbiota in infancy. *Nestle Nutr Workshop Ser Pediatr Program* **62:** 13-29.

Altenhoefer,A., Oswald,S., Sonnenborn,U., Enders,C., Schulze,J., Hacker,J., and Oelschlaeger,T.A. (2004) The probiotic Escherichia coli strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens. *FEMS Immunol Med Microbiol* **40:** 223-229.

Alter,S.J., Vidwan,N.K., Sobande,P.O., Omoloja,A., and Bennett,J.S. (2011) Common childhood bacterial infections. *Curr Probl Pediatr Adolesc Health Care* **41:** 256-283.

Baumgart,M., Dogan,B., Rishniw,M., Weitzman,G., Bosworth,B., Yantiss,R. et al. (2007) Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum *ISME J* **1:** 403-418.

Bettelheim,K.A. and Lennox-King,S.M. (1976) The acquisition of Escherichia coli by new-born babies. *Infection* **4:** 174-179.

Bjerke,G.A., Wilson,R., Storrø,O., Oyen,T., Johnsen,R., and Rudi,K. (2011) Mother-to-child transmission of and multiple-strain colonization by Bacteroides fragilis in a cohort of mothers and their children. *Appl Environ Microbiol* **77:** 8318-8324.

Brzuszkiewicz,E., Bruggemann,H., Liesegang,H., Emmerth,M., Olschlager,T., Nagy,G. et al. (2006) How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains. Proc Natl Acad Sci U S A 103: 12879-12884.

Buckling,A., Craig,M.R., Brockhurst,M.A., and Colegrave,N. (2009) The Beagle in a bottle. *Nature* **457:** 824-829.

Caporaso,J.G., Lauber,C.L., Costello,E.K., Berg-Lyons,D., Gonzalez,A., Stombaugh,J. et al. (2011) Moving pictures of the human microbiome. *Genome Biol* **12:** R50.

Chassaing,B., Rolhion,N., de,V.A., Salim,S.Y., Prorok-Hamon,M., Neut,C. et al. (2011) Crohn disease-associated adherent-invasive E. coli bacteria target mouse and human Peyer's patches via long polar fimbriae. *J Clin Invest* **121:** 966-975.

Chen,S.L., Hung,C.S., Xu,J., Reigstad,C.S., Magrini,V., Sabo,A. et al. (2006) Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A 103: 5977-5982.

Costello,E.K., Stagaman,K., Dethlefsen,L., Bohannan,B.J., and Relman,D.A. (2012) The application of ecological theory toward an understanding of the human microbiome. *Science* **336:** 1255-1262.

Diard,M., Garry,L., Selva,M., Mosser,T., Denamur,E., and Matic,I. (2010) Pathogenicity-associated islands in extraintestinal pathogenic Escherichia coli are fitness elements involved in intestinal colonization. J Bacteriol 192: 4885-4893.

Didelot,X., Meric,G., Falush,D., and Darling,A.E. (2012) Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* **13:** 256.

Escherich,T. (1989) Classics in Infectious-Diseases - the Intestinal Bacteria of the Neonate and Breast-Fed Infant (Reprinted from Fortschritte der Med, Vol 3, 1885) *Rev Inf Dis* **11:** 352-356.

Faust,K. and Raes,J. (2012) Microbial interactions: from networks to models. *Nat Rev Microbiol* **10:** 538-550.

Foxman,B. (2010) The epidemiology of urinary tract infection. Nat Rev Urol 7: 653-660.

Frank,D.N., Amand,A.L.S., Feldman,R.A., Boedeker,E.C., Harpaz,N., and Pace,N.R. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* **104:** 13780-13785.

Fredrickson,A.G. and Stephanopoulos,G. (1981) Microbial Competition. *Science* **213:** 972-979.

Fuller,R. (1989) Probiotics in man and animals. *J Appl Bacteriol* **66:** 365-378.

Gaboriau-Routhiau,V., Rakotobe,S., Lecuyer,E., Mulder,I., Lan,A., Bridonneau,C. et al. (2009) The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. *Immunity* **31:** 677-689.

Garrido,D., Barile,D., and Mills,D.A. (2012) A molecular basis for bifidobacterial enrichment in the infant gastrointestinal tract. *Adv Nutr* **3:** 415S-421S.

Gause,G.F. (1934) *The Struggle for Existence.* Baltimore, MD: Williams and Wilkins.

Giongo,A., Gano,K.A., Crabb,D.B., Mukherjee,N., Novelo,L.L., Casella,G. et al. (2011) Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J* **5:** 82-91.

Gottschal,J.C. (1993) Growth kinetics and competition-some contemporary comments. *Antonie Van Leeuwenhoek* **63:** 299-313.

Grover,J.P. (1997) *Resource competition.* London: Chapman & Hall.

Hooper,L.V., Littman,D.R., and Macpherson,A.J. (2012) Interactions between the microbiota and the immune system. *Science* **336:** 1268-1273.

Ivanov,I.I., Atarashi,K., Manel,N., Brodie,E.L., Shima,T., Karaoz,U. et al. (2009) Induction of intestinal Th17 cells by segmented filamentous bacteria. *Cell* **139:** 485-498.

Jacobi,C.A. and Malfertheiner,P. (2011) Escherichia coli Nissle 1917 (Mutaflor): new insights into an old probiotic bacterium. *Dig Dis* **29:** 600-607.

Jaureguy,F., Landraud,L., Passet,V., Diancourt,L., Frapy,E., Guigon,G. et al. (2008) Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains. *BMC Genomics* **9:** 560.

Kaper,J.B., Nataro,J.P., and Mobley,H.L. (2004) Pathogenic Escherichia coli. *Nat Rev Microbiol* **2:** 123-140.

Kim,H., Kwack,K., Kim,D.Y., and Ji,G.E. (2005) Oral probiotic bacterial administration suppressed allergic responses in an ovalbumin-induced allergy mouse model. *FEMS Immunol Med Microbiol* **45:** 259-267.

Kocourkova,I., Ladnikova,R., Zizka,J., and Rosova,V. (2007) Effect of oral application of a probiotic E. coli strain on the intestinal microflora of children of allergic mothers during the first year of life. *Folia Microbiol (Praha)* **52:** 189-193.

Koenig,J.E., Spor,A., Scalfone,N., Fricker,A.D., Stombaugh,J., Knight,R. et al. (2011) Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108 Suppl 1:** 4578-4585.

Kotlowski,R., Bernstein,C.N., Sepehri,S., and Krause,D.O. (2007) High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut* **56:** 669-675.

Kruis,W., Chrubasik,S., Boehm,S., Stange,C., and Schulze,J. (2012) A double-blind placebo-controlled trial to study therapeutic effects of probiotic Escherichia coli Nissle 1917 in subgroups of patients with irritable bowel syndrome. *Int J Colorectal Dis* **27:** 467-474.

Kruis,W., Fric,P., Pokrotnieks,J., Lukas,M., Fixa,B., Kascak,M. et al. (2004) Maintaining remission of ulcerative colitis with the probiotic Escherichia coli Nissle 1917 is as effective as with standard mesalazine. *Gut* **53:** 1617-1623.

Kruis,W., Schutz,E., Fric,P., Fixa,B., Judmaier,G., and Stolte,M. (1997) Double-blind comparison of an oral Escherichia coli preparation and mesalazine in maintaining remission of ulcerative colitis. *Aliment Pharmacol Ther* **11:** 853-858.

Larsen,N., Vogensen,F.K., van den Berg,F.W.J., Nielsen,D.S., Andreasen,A.S., Pedersen,B.K. et al. (2010) Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *Plos One* **5**: e9085.

Le,G.T., Clermont,O., Gouriou,S., Picard,B., Nassif,X., Denamur,E., and Tenaillon,O. (2007) Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. Mol Biol Evol 24: 2373-2384

Levin,B.R. (1972) Coexistence of two asexual strains on a single resource. *Science* **175:** 1272-1274.

Ley,R.E., Peterson,D.A., and Gordon,J.I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124:** 837-848.

Lodinova-Zadnikova,R., Cukrowska,B., and Tlaskalova-Hogenova,H. (2003) Oral administration of probiotic Escherichia coli after birth reduces frequency of allergies and repeated infections later in life (after 10 and 20 years). *Int Arch Allergy Immunol* **131:** 209-211.

Lodinova-Zadnikova,R., Prokesova,L., Kocourkova,I., Hrdy,J., and Zizka,J. (2010) Prevention of allergy in infants of allergic mothers by probiotic Escherichia coli. *Int Arch Allergy Immunol* **153:** 201-206.

Lukjancenko,O., Wassenaar,T.M., and Ussery,D.W. (2010) Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* **60:** 708-720.

Luo,C., Walk,S.T., Gordon,D.M., Feldgarden,M., Tiedje,J.M., and Konstantinidis,K.T. (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A* **108:** 7200-7205.

Merritt,K.K. and Paige,B.H. (1935) Septicemia due to escherichia acidi-lactici (Bacillus acidi-lactici) in a newborn infant - Report of a case with necropsy. *Am J Dis Child* **50:** 693-698.

Monod,J. (1942) Recherches sur la Croissance des Cultures Bacteriennes. *Hermann & Cie , Paris*.

Moriel,D.G., Bertoldi,I., Spagnuolo,A., Marchi,S., Rosini,R., Nesta,B. et al. (2010) Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic Escherichia coli. Proc Natl Acad Sci U S A 107: 9072-9077.

Morowitz,M.J., Denef,V.J., Costello,E.K., Thomas,B.C., Poroyko,V., Relman,D.A., and Banfield,J.F. (2011) Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc Natl Acad Sci U S A* **108:** 1128-1133.

Nielubowicz,G.R. and Mobley,H.L. (2010) Host-pathogen interactions in urinary tract infection. Nat Rev Urol 7: 430-441.

Nissle ,A. (1918) Die antagonistische Behandlung chronischer Darmstoerungen mit Colibakterien [The antagonistic therapy of chronic intestinal disturbances]. *Med Klinik* 29-33.

Nowrouzian,F., Hesselmar,B., Saalman,R., Strannegard,I.L., Aberg,N., Wold,A.E., and Adlerberth,I. (2003) Escherichia coli in infants' intestinal

microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatr Res* **54:** 8-14.

O'Malley,M.A. (2007) The nineteenth century roots of 'everything is everywhere'. *Nat Rev Microbiol* **5:** 647-651.

Ochman,H. and Selander,R.K. (1984) Standard reference strains of Escherichia coli from natural populations. *J Bacteriol* **157:** 690-693.

Palmer,C., Bik,E.M., DiGiulio,D.B., Relman,D.A., and Brown,P.O. (2007) Development of the human infant intestinal microbiota. *PLoS Biol* **5:** e177.

Patzer,S.I., Baquero,M.R., Bravo,D., Moreno,F., and Hantke,K. (2003) The colicin G, H and X determinants encode microcins M and H47, which might utilize the catecholate siderophore receptors FepA, Cir, Fiu and IroN. *Microbiology* **149:** 2557-2570.

Penders,J., Stobberingh,E.E., van den Brandt,P.A., and Thijs,C. (2007a) The role of the intestinal microbiota in the development of atopic disorders. *Allergy* **62:** 1223-1236.

Penders,J., Thijs,C., van den Brandt,P.A., Kummeling,I., Snijders,B., Stelma,F. et al. (2007b) Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut* **56:** 661-667.

Pitout,J.D. (2012) Extraintestinal Pathogenic Escherichia coli: A Combination of Virulence with Antibiotic Resistance. Front Microbiol 3: 9
Rainey,P.B., Buckling,A., Kassen,R., and Travisano,M. (2000) The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol Evol* **15:** 243-247.

Rasko,D.A., Rosovitz,M.J., Myers,G.S., Mongodin,E.F., Fricke,W.F., Gajer,P. et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* **190:** 6881-6893.

Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R. et al. (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* **34:** 1-9.

Rolhion,N. and Darfeuille-Michaud,A. (2007) Adherent-invasive Escherichia coli in inflammatory bowel disease. *Inflamm Bowel Dis* **13:** 1277-1283.

Round,J.L., Lee,S.M., Li,J., Tran,G., Jabri,B., Chatila,T.A., and Mazmanian,S.K. (2011) The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science* **332:** 974-977.

Rudi,K., Storrø,O., Oien,T., and Johnsen,R. (2012) Modelling bacterial transmission in human allergen-specific IgE sensitization. *Lett Appl Microbiol* **54:** 447-454.

Russo,T.A. and Johnson,J.R. (2000) Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC. J Infect Dis 181: 1753-1754.

Russo,T.A. and Johnson,J.R. (2003) Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes Infect* **5:** 449-456.

Sahl,J.W., Matalka,M.N., and Rasko,D.A. (2012) Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. *Appl Environ Microbiol* **78:** 4884-4892.

Schamberger,G.P., Phillips,R.L., Jacobs,J.L., and Diez-Gonzalez,F. (2004) Reduction of Escherichia coli O157:H7 populations in cattle by addition of colicin E7-producing E. coli to feed. *Appl Environ Microbiol* **70:** 6053-6060.

Sepehri,S., Kotlowski,R., Bernstein,C.N., and Krause,D.O. (2009) Phylogenetic analysis of inflammatory bowel disease associated Escherichia coli and the fimH virulence determinant. *Inflamm Bowel Dis* **15:** 1737-1745.

Stewart,F.M. and Levin,B.R. (1973) Partitioning of resources and outcome of interspecific competition - Model and some general considerations. *Am Nat* **107:** 171-198.

Storrø,O., Oien,T., Dotterud,C.K., Jenssen,J.A., and Johnsen,R. (2010) A primary health-care intervention on pre- and postnatal risk factor behavior to prevent childhood allergy. The Prevention of Allergy among Children in Trondheim (PACT) study. *BMC Public Health* **10:** 443.

Storrø,O., Øien,T., Langsrud,Ø., Rudi,K., Dotterud,C., and Johnsen,R. (2011) Temporal variations in early gut microbial colonization are associated with allergen-specific immunoglobulin E but not atopic eczema at 2 years of age. *Clin Exp Allergy* **41:** 1545-1554.

Strachan,D.P. (1989) Hay-Fever, Hygiene, and Household Size. *BMJ* **299:** 1259-1260.

Taylor,D.E. (1999) Bacterial tellurite resistance. *Trends Microbiol* **7:** 111-115.

Tenaillon,O., Skurnik,D., Picard,B., and Denamur,E. (2010) The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* **8:** 207-217.

Todar,K. Todar's online textbook of bacteriology. University of Wisconsin-Madison, Madison. 2012.

Touchon,M., Hoede,C., Tenaillon,O., Barbe,V., Baeriswyl,S., Bidet,P. et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5:** e1000344.

Trosvik,P., Rudi,K., Straetkvern,K.O., Jakobsen,K.S., Naes,T., and Stenseth,N.C. (2010a) Web of ecological interactions in an experimental gut microbiota. *Environ Microbiol* **12:** 2677-2687.

Trosvik,P., Stenseth,N.C., and Rudi,K. (2010b) Convergent temporal dynamics of the human infant gut microbiota. *ISME J* **4:** 151-158.

Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R., and Gordon,J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444:** 1027-1031.

Veldkamp,H. (1970) Enrichment cultures of prokaryotic organisms. In *Methods in Microbiology*. Norris,J. and Ribbones DW (eds). Academic Press, London, 305-361.

Vieira-Silva,S. and Rocha,E.P. (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* **6:** e1000808.

Weise,C., Zhu,Y., Ernst,D., Kuhl,A.A., and Worm,M. (2011) Oral administration of Escherichia coli Nissle 1917 prevents allergen-induced dermatitis in mice. *Exp Dermatol* **20:** 805-809.

Wilhelmi,I., Roman,E., and Sanchez-Fauquier,A. (2003) Viruses causing gastroenteritis. *Clin Microbiol Infect* **9:** 247-262.

Willenbrock,H., Hallin,P.F., Wassenaar,T.M., and Ussery,D.W. (2007) Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. *Genome Biol* **8:** R267.

Yatsunenko,T., Rey,F.E., Manary,M.J., Trehan,I., Dominquez-Bello,M.G., Contreras,M. et al. (2012) Human gut microbiome viewed across age and geography. *Nature* **486:** 222-227.

Zhao,T., Doyle,M.P., Harmon,B.G., Brown,C.A., Mueller,P.O., and Parks,A.H. (1998) Reduction of carriage of enterohemorrhagic Escherichia coli O157:H7 in cattle by inoculation with probiotic bacteria. *J Clin Microbiol* **36:** 641-647.

Zoetendal,E.G., Cheng,B., Koike,S., and Mackie,R.I. (2004) Molecular microbial ecology of the gastrointestinal tract: from phylogeny to function. *Curr Issues Intest Microbiol* **5:** 31-47.

I

II

***Running Title*:**

Context-dependence in a bacterial community

Title:

Context-dependent competition in a model gut bacterial community

Eric J. de Muinck[1,2,3], Pål Trosvik[1] , Daniel Sachse[4,5], Jan vander Roost[1], Kjersti S. Rønningen[6], Knut Rudi[7], Nils Chr. Stenseth[1 *]

Affiliations: [1]Center for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Norway, [2]Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway, [3]NOFIMA The Norwegian Institute of Food, Fisheries and Aquaculture Research, Ås, Norway, [4]Institute of Clinical Medicine, University of Oslo, Norway, [5]Department of Medical Biochemistry, Oslo University Hospital, Oslo, Norway, [6]Department of Paediatric Research, Oslo University Hospital, Rikshospitalet, Oslo, Norway, [7]Department of Chemistry, Biotechnology and Food Science, University of Life Sciences, Ås, Norway

[*] Corresponding Author: Nils Chr. Stenseth, n.c.stenseth@bio.uio.no, Phone: +47 22 85 44 00, Fax:+47 22 85 40 01

Understanding the ecological processes that generate complex community structures may provide insight into the establishment and maintenance of a normal microbial community in the human gastrointestinal tract, yet very little is known about how biotic interactions influence community dynamics in this system. Here, we use natural strains of *Escherichia coli* and a simplified model microbiota to demonstrate that the colonization process on the strain level can be context dependent, in the sense that the outcome of intra-specific competition may be determined by the composition of the background community. These results are consistent with previous models for competition between organisms where one competitor has adapted to low resource and high stress environments whereas the other is optimized for rapid reproduction when resources are abundant. The genomic profiles of *E.coli* strains representing these differing ecological strategies provide clues for deciphering the genetic underpinnings of niche adaptation within a single species. Our findings extend the role of ecological theory in understanding microbial systems and the conceptual toolbox for describing microbial community dynamics. There are few, if any, concrete examples of context-dependent competition on a single trophic level. However, this phenomenon can have potentially dramatic effects on which bacteria will successfully establish and persist in the gastrointestinal system, and the principle should be equally applicable to other microbial ecosystems.

Subject Category: Microbial population and community ecology

Keywords: *Escherichia coli*/Resource competition/Intestinal colonization

INTRODUCTION

Escherichia coli is a ubiquitous, albeit not abundant, member of the gastrointestinal (GI) microbiota (Tenaillon et al., 2010). Still, a healthy human will normally have more than one billion living in the intestine (Savageau, 1983; Touchon et al., 2009). In addition to acute infections, *E.coli* colonization of the gut has been consistently linked to the chronic condition *inflammatory* bowel disease (IBD) (Rhodes, 2007), implicating particular types of strains in its aetiology (Baumgart et al., 2007; Friswell et al., 2010; Kotlowski et al., 2007; Sepehri et al., 2009). It has also been proposed that competition between strains may be important due to the positive effects of the *E.coli* Nissle 1917 strain in IBD (Vejborg et al., 2011). The wide spectrum of relationships between *E.coli* and humans highlights the importance of understanding colonization on the strain level and linking this to the dynamics of higher taxa.

In theory, and in contrast to a chemostat system, the fluctuating nutrient levels inherent in a batch culture can be used to promote coexistence of strains on a single nutrient (Rainey et al., 2000; Stewart and Levin, 1973). These dynamics are a special case for pairs of species in which one (sometimes called a "gleaner") has a higher growth rate at low nutrient concentrations and another (sometimes called an "exploiter" or "opportunist") has a higher growth rate at high nutrient concentrations (Fredrickson and Stephanopoulos, 1981; Gottschal, 1993). This relationship can be described in terms of the Monod equation $\mu = \mu_{max} s/(K_s+s)$ (Monod 1949) where the outcome of competition is determined by the maximum growth rate (μ_{max}) and the Monod constant (K_s), i.e. the nutrient

concentration at which a species has a growth rate of μ_{max} /2. In the special case where one species has a higher μ_{max} while the other has a lower K_s, there exists a nutrient concentration (s) at which the growth rates are equal. Below that concentration, the species with the lower K_s will win, while higher concentrations favor the species with the higher μ_{max}. We will refer to this case of affairs as 'species-pair with high u_{max} and low K_s' (sHULK). The exact value of s that in theory will permit coexistence is for practical reasons very difficult to achieve in a chemostat. In contrast, the parameter space is widened for possible coexistence in a serial batch culture system due to the dynamic nature of the nutrient environment. In fact, some experimental evidence does support this type of proposed interaction (Grover, 1997; Levin, 1972). However, much controversy still remains concerning the conditions that lead to coexistence (Abrams, 2004). Our use of this type of cyclical, or seasonal, system was due to its tractability and because the intestinal system most likely experiences nutrient pulsing (Johnson, 2000).

In the ecological literature, there is abundant evidence that interactions between pairs of species are dependent on the community context through indirect interactions with other species (Brown et al., 2001; Werner and Peacor, 2003). Such biotic interaction have been classified into two main types; density-mediated and trait-mediated indirect interactions (DMII and TMII respectively) (Abrams, 1995). The former describes the case where the density of one of a focal species pair is affected by a third party species, with effects cascading to the second of the pair. The latter pertains to a situation where a third species causes the interaction between a species pair to change due to phenotypic alterations. TMII has been documented in a number of experimental systems involving three trophic

levels (Wissinger and Mcgrady, 1993), as well as exploitative competitor pairs sharing either a common resource (Peacor and Werner, 1997) or predator (Relyea, 2000). Although these studies focus on behavioral traits, the plastic phenotype also includes more subtle traits like changes in physiological state or gene expression.

Experimental work demonstrating DMII has been carried out using host-parasite systems (Kiesecker and Blaustein, 1999) including exploitatively competing pairs of bacteria and their bacteriophages (Harcombe and Bull, 2005). The host-parasite system is analogous to the predator-prey system, consisting of two distinct trophic levels. However, much less is known about context-dependent interactions among guilds of organisms on one trophic level. This is especially true when it comes to the GI microbiota, although probiotic bacteria have been thought to compete for nutritional substrates and thus alter the microbial structure of the gut (Fuller, 1989).

Here, we present observations of *E.coli* strains isolated from a cohort of infants and a series of competition experiments with strains isolated over a two-year period from a single infant. Previous analysis of the fecal samples derived from this infant cohort discovered that *E.coli* colonization of the gut before the age of 1 year reduced the likelihood of IgE sensitization and that early colonization was likely to have originated from the mother (de Muinck et al., 2011; Rudi et al., 2012; Storro et al., 2011). Our ability to use unmodified strains that were isolated from a single infant for the competition experiments allows us to make some tentative connections to the intestinal

ecosystem. By measuring life history traits (Vasi et al., 1994) and conducting competition experiments using different serial transfer regimes, we demonstrate that colonization at the strain level may be explained by relatively simple models. However, our experiments also reveal a mechanism by which the *E.coli* colonization process may be dramatically changed through a dynamic nutrient environment and interaction with gut community members belonging to a different phylum. Specifically, we show that the relative abundance of members of Firmicutes and Bacteriodetes may have the potential to influence strain level competition in *E.coli*. The balance between Firmicutes and Bacteriodetes, has been linked to a number of diseases including type 1 and type 2 diabetes, IBD, and obesity (Frank et al., 2007; Giongo et al., 2011; Larsen et al., 2010; Turnbaugh et al., 2006). Thus, insights into the colonization dynamics in the infant gut on high taxonomic levels (Trosvik et al., 2010b) and the potential to influence patterns on lower taxa are of potentially great value (Kau et al., 2011). Genomic comparison of the competing strains allows us to generate new hypotheses of specific genetic factors that contribute to the competitive phenotype of these bacteria in the serial transfer regime. Our results provide a specific illustration of the concept of context-dependent competition in bacteria on the same trophic level, where the outcome of competition between closely related strains during colonization can be determined by how the nutrient structure in the environment is modified by the established community.

Materials and Methods:

Strains, life history trait measurement and competitions

Strains were obtained from the IMPACT cohort and all are described as part of that study except EDM530 (de Muinck et al., 2011). EDM530 was isolated after the previous study. The model background microbiota has been described elsewhere (Trosvik et al., 2010a). Growth rates and competition experiments (see supplementary methods for details) were performed in Oxoid anaerobe basal broth, a medium designed to accommodate GI bacteria, unless otherwise stated. Growth rates were measured in triplicate by monitoring OD600 with a Bioscreen (Oy Growth Curves Ab Ltd, Finland). Carrying capacities were measured by freeze drying and weighing (see supplementary methods for details).

Determination of co-existence of EDM106 and EDM116 in one year fecal sample

A 600 bp region of the mdh gene was sequenced directly from the DNA extracted from the fecal sample. From the electropherogram we identified mixed peaks in 7 positions that indicated two co-existing strains with one in lower abundance. Visual decomposition of the mixed electropherogram into two sequences and subsequent alignment with the pure sequences of EDM116 and EDM106 found perfect identity for both subsequences. Multiple linear regression of the mixed spectrum against the strain unit spectra (see below) found relative abundances of 84% and 16% for EDM116 and EDM106 respectively (R-squared=0.975, $p<0.0001$), providing evidence of the co-existence of these two strains in the one year sample.

Sample processing and quantification of relative abundances

DNA extraction, PCR amplification, and sequencing of *mdh* and 16s rRNA gene fragments were performed using previously described methods (Trosvik et al., 2007; de Muinck et al., 2011). The relative abundances of strains and species in the competitions were determined by multivariate decomposition of mixed *mdh* or 16SrRNA gene sequence electropherograms as previously described (Trosvik et al., 2007; de Muinck et al., 2011).

Bacteriocin screening and plasmid detection

All strains were subjected to a bacteriocin screen using the overlay method with each used both as an overlay and stab inoculate. Plasmid extraction from the three strains used in the competitions was carried out using the Promega (Madison, WI, USA) Wizard® *Plus*SV Miniprep system and isolates were visualized on a gel.

Proton nuclear magnetic resonance (HNMR) spectra

Samples were spun down at 12,000 rpm (13,400g) at 4ºC and the supernatant was removed and combined with potassium salts buffer at pH 7.4 and TSP as an internal standard. NMR spectra were obtained from a Bruker Avance 600 NMR spectrometer. 1d NOESY spectra were acquired with presaturation for water suppression. The spectra were referenced to the TSP signal at 0 ppm, baseline corrected and normalized to a constant sum of 1. Prior to PCA, the spectra were mean centered and scaled to unit variance. All statistical computations were carried out using R (R Development Core Team, 2011).

Genome sequencing and assembly

EDM530, EDM106, and EDM116 were sequenced using Roche 454 GS (FLX Titanium) pyrosequencing according to standard protocol. Number of contigs, median depth and N50 were 198, 17.5 and 1209, 585.8, 17.5 and 4007, and 864, 8.2 and 2714 for the strains, respectively. De Novo assembly was performed using Newbler v2.3 with default settings and contigs were annotated with RAST (Aziz et al., 2008). Characterization of shared and unique gene content was done by BLASTing all annotated genes of all strains against one another using 85% sequence identity and an e-value cutoff of $<1e^{-25}$ for assignment of presence (see supplementary methods for details). Gene enrichment analysis and GO assignment were carried out using the Blast2GO software package (Conesa et al., 2005). All DNA sequences will be made available in Genbank and accession numbers will be provided.

RESULTS

Growth rate and colonization

In general, infant GI-bacterial species appear to have shorter generation times when compared with adult GI-bacterial species (Vieira-Silva and Rocha, 2010). This was attributed to faster growing organisms having an advantage during early colonization (Leveque C, 2003). We measured the generation times of 23 different *E.coli* isolates under aerobic and anaerobic growth conditions (Table S1) and found that low generation

times in aerobic culture have a tendency to coincide with early colonization ($p=0.034$, one-tailed Mann-Whitney U test) (Figure S1). That this effect was observed under aerobic but not anaerobic conditions ($p=0.95$) coincides with the aerobic environment thought to exist in the early infant gut (Adlerberth, 2008). Of note, no trade-off was found between anaerobic and aerobic growth rates for the collection of *E.coli* isolates. Instead, a significant positive correlation was found between maximal aerobic growth rate and maximal anaerobic growth rate ($p<0.001$)(Figure S2A). Further analysis revealed a strong relationship between an increase in anaerobic growth rate and difference between anaerobic and aerobic growth rates ($p<0.001$)(Figure S2C). One strain (EDM106) however, did not follow the strict relationship between anaerobic and aerobic growth rates and had a faster growth rate under anaerobic conditions than would have been expected (Figures S2A and S2B). EDM106 was an early colonizer of an infant and became a focus of further investigation.

Competition between strains isolated from an infant

We were curious about the effect of the above described aberrant growth rate characteristics on strain ecology in a competitive context. Also, in order to focus our investigation on strain competitiveness during colonization, we looked more closely into the *E.coli* strain succession in the infant that harboured EDM106. Three different strains with no apparent clonal relationship, as determined from genome analysis *(below),* were isolated from samples collected at three different time points from the infant over a period of two years. These strains were the only three found in the infant over the two

year period as determined from *mdh* fragment amplification and sequencing as described in (de Muinck et al., 2011). A single strain was found at four days of age (EDM106), whereas EDM106 and EDM116 were found to coexist in the sample collected at one year of age (see materials and methods). At two years of age, both of these strains had been supplanted by EDM530.

Despite the abnormally high growth rate of EDM106 under anaerobic conditions, when compared with the entire collection of isolates, EDM106 had a longer anaerobic generation time (56.50±0.30 min.; mean ±s.e. with n=3) relative to EDM116 (50.97±0.32 min.; $p < 0.01$) and EDM530 (49.03±0.24 min.; $p < 0.01$, t-test). Thus, the final replacement event is in agreement with these measurements, assuming this is the main phenotype determining competition outcomes. Under aerobic conditions, EDM106 had a shorter generation time (38.12±0.35 min.) relative to both the EDM116 (40.98±0.49 min.; $p < 0.01$) and EDM530 (41.56±0.72 min.; $p < 0.01$) and thus, this strain fits the profile of an early colonizer with respect to aerobic growth rate (Adlerberth, 2008).

Pair-wise competition experiments found that both EDM116 and EDM106 were individually outcompeted by EDM530 (Figures S3A and S7). If all three strains were competed together, EDM530 dominated the cultures (Figure S3B). Bacteriocins are a commonly ascribed determinant of bacterial intra-specific competition. Due to their potential effect and the fact that bacteriocin genes were discovered in the genomes, a susceptibility screen was performed. No effect on the strains used in the competition

experiments was found. Plasmid transfer may also affect the the competition experiments. The increased coverage depth of several contigs relative to the rest of the contigs of a genome sequence showed plasmid carriage by EDM106 and EDM116 (Table S3). In contrast, EDM530 does not seem to carry plasmids. After five days of co-culture of EDM530 and EDM106, no plasmid was isolated from EDM530, indicating a lack of plasmid transfer.

The co-occurrence of EDM106 and EDM116 in the sample taken at one year of age is difficult to explain given the substantially different generation times of these strains in anaerobic conditions. Pair-wise competitions found that EDM106 and EDM116 were also able to coexist in culture at stable densities for a substantial amount of time under both one and two day transfer regimes (Figure S4). The sHULK relationship described above could explain the observed competition outcomes between these strains in the batch culture regime. Growth rate measurements at different nutrient concentrations showed that these strains do in fact have this kind of relationship (Figure S5). Additionally, the co-culture showed an increased carrying capacity relative to the one expected from combining the single strain carrying capacities ($p = 0.024$, one-sample t-test)(Figure S6A). The generation of isoclines using these carrying capacities and relative abundances of strains in the co-culture supports that a stable equilibrium point can be reached by EDM106 and EDM116 (Figure S6B).

EDM106 was outcompeted by EDM530 in serial batch culture. A series of competition experiments using different transfer times (12hours, 1day, 2days, 3days, and 4days)(Figure S7) found the same outcome under all conditions. There was, however, a positive non-linear relationship between increased time spent in low nutrient competition and number of transfers to out-competition (R^2=0.95). This was further investigated using long term stationary phase cultures in which EDM106 eventually did outcompete EDM530 (Figure S8), supporting the hypothesis that EDM106 is a "gleaner" whereas EDM530 is an "exploiter" or "opportunist".

Competition between E.coli isolates in a model gut microbiota

In order to investigate whether the strain competitions would be influenced by the presence of a background community, we used a simplified model gut microbiota with species representing the four main phyla (Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria) inhabiting the human gut (Trosvik et al., 2010a). *Clostridium perfringens*, *Bacteroides thetaiotaomicron*, and *Bifidobacterium longum* were inoculated into the batch culture system along with either EDM106 and EDM530 or EDM116 and EDM530. Species and *E.coli* strain dynamics were then followed over time. During the initial stages of the competition, *E.coli* strain dynamics remained consistent with previously described observations and continued as such for the competition between EDM116 and EDM530 (Figure S9). However in the competitions between EDM 106 and EDM 530, EDM530 no longer continued on the trajectory towards dominance after day 10 (Figure 1 and Figure S10). Instead, EDM106 came to dominate. This change in *E.coli* strain competition

dynamics seemed to occur in conjunction with *C.perfringens* dominance in the batch culture. As a control experiment we attempted to manipulate the species dynamics of the system by allowing small amounts of air into the culture flasks. This resulted in partial inhibition of the anaerobes, allowing for increased dominance of *E.coli* (Figure S11). Under these conditions we did not observe any modulation of the *E.coli* competitions. The effect of *C.perfringens* on strain competition was further investigated by reviving cryogenic stocks collected on day 10 of a competition between EDM106 and EDM530, and spiking these cultures with high doses of either *C.perfringens* or *B.thetaiotaomicron* (Figures 2A and 2B). The competition trajectories remained unchanged in the competition inoculated with *B.thetaiotaomicron*. However, the competition trajectories reversed in the competition inoculated with *C.perfringens* indicating that *C.perfringens* dominance is the causative factor of reversed trajectories.

Nutritional and metabolomic profiling of competitions

Investigation into potential reasons of altered strain competition outcomes by the addition of *C.perfringens* was begun by growing *C.perfringens* to saturation and then removing cells by filtration to create a spent rich medium that included any modulating factors released into the media. Repeating the *E.coli* strain competitions in 90/10, 50/50, or 10/90 (spent/fresh) mediums did not produce any effects on competition outcomes (Figure S12). However, using rich/minimal-salts medium proportions of 90/10, 50/50, and 10/90 (Table S2) did find a reversal of the trajectories in the low nutrient (10/90) competition (Figure 3). HNMR spectra of the supernatants from these competitions identified differences

linked to the competition outcomes, suggesting a chemical underpinning of the observations (Figure S13). HNMR signal peaks of complex samples can be attributed to many different types of molecules and therefore definitive identification is difficult.

We then performed competitions using high, intermediate, and low concentrations of glucose and peptone (Table S2), with the high nutrient regimes corresponding to the concentrations in the original rich medium to see if the reversal of competitive outcome was a result of strict density effects or related to specific compounds. EDM530 dominated under all three scenarios with glucose as the sole carbon source (Figure S14), indicating that the reversal was not strictly due to density effects. However, a low concentration of peptone resulted in a change to the competition trajectory ($p = 0.019$, logistic mixed effects model, see supplementary methods)(Figure 4) suggesting a pathway by which the strain level competition may be modulated through ecological interaction with bacteria belonging to a completely different phylum.

Genome comparisons

Genomes were sequenced as described in materials and methods. A general comparison of the annotated genomes of the three sequenced strains found a core genome of 3535 genes (72% of the annotated total) that drops by 7.5% to 3271 genes if *E.coli* MG1655 (K12) is included in the comparison (Figures S15 and S16). An enrichment comparison of some candidate pathways identified potential functional categories (discussed below) that could influence competition results (Figure 5 and Supplementary files). The large

differences in gene content strongly indicate that these strains do not have a recent clonal relationship

DISCUSSION

It would most likely be a question of "who gets there first" that determines the colonization pattern in the intestine. In most cases, it would be the mother who has first opportunity of exposing the fetus to the new microbiota. Still, growth rate has been previously linked to intestinal colonization (Vieira-Silva and Rocha, 2010), and we found a tendency of early colonizers to have a faster aerobic growth rate than late colonizers (Figure S1). Early colonizers of the infant gut are often a combination of aerobic and facultative anaerobic bacteria that use up the available oxygen and thus allow for the succession to obligate anaerobes which dominate the gut flora in the mature intestine (Adlerberth, 2008). The positive correlation between the aerobic and anaerobic generation times suggest that some strains are intrinsically more efficient at using nutrients than other strains regardless of whether they are in an aerobic or anaerobic environment (Figure S2A). The relationship discovered between the increase in minimal generation times and the difference between the anaerobic generation times and aerobic generation times (Figure S2C) suggests that metabolic efficiency reaches a plateau as a strain's 'intrinsic' doubling time decreases. This could signal a physical constraint on efficient nutrient utilization that is independent of internal cellular machinery.

Isolate competitions

The maximal growth rates of the three strains used in competitions matched the previously described pattern of colonization with the faster growing strain under aerobic conditions isolated from an earlier sample. Despite different growth rates of two of the isolates (EDM106 and EDM116) stable co-existence was observed in both one and two day transfer regimes. These observations can be explained by two different models. The first model invokes the sHULK relationship which would allow the two strains to be competitively superior in different phases of the batch culture. This would be in accordance with our growth rate measurements (Figure S5), and is a previously reported mechanism for maintenance of coexistence (Rainey et al., 2000). The second model is based on the observed increase in carrying capacity of the co-culture. If between-strain competition is weaker than within-strain competition there is a theoretical equilibrium point for coexistence (Figure S6B). This state can be brought about cooperative interactions like cross-feeding, which is indicated in by the increased productivity in the co-culture (Figure S6A). Acetate cross-feeding has been reported to evolve readily in *E.coli* cultures (Treves et al., 1998), but from our data we cannot determine the exact nature of the interaction between EDM106 and EDM 116. Neither model excludes the other, but further experimentation is required for uncovering the mechanism of coexistence. However, the fact that these strains were found to coexist in the infant gut, as well as in co-culture suggests that they could occupy overlapping niches in their natural environment, and that the observed interaction is ecologically relevant.

In contrast, competitions with EDM106 and EDM530 under several different transfer regimes were consistently dominated by EDM530 (Figure S7). We did find a strong

positive correlation between transfer rate and the number of transfers to outcompetition, and long term stationary cultures were dominated by EDM106 (Figure S8). These results fit a model where the exploiter gains ground during the log period of growth while the gleaner does so during stationary phase. When all the competition results are taken together, we were able to replicate the outcomes of strain competition observed in the actual infant gut but this does not mean that these were the actual factors responsible for the outcomes in that complex environment.

Context dependant competitive effects

Competition between EDM106 and EDM530 in the model gut background found that the competitive trajectories of the strains changed as *C.perfringens* became dominant (Figure 1), but this effect was not observed in the micro-aerophilic cultures (Figure S11). This indicates that the community structure of the surrounding species can modify the interaction between competing strains. While this may seem intuitive, examples of these types of interactions in bacterial systems are, to our knowledge, lacking.

The altered strain interactions in the presence of *C.perfringens* (and not *B.thetaiotaomicron*) suggested that the *C.perfringens* was modifying the environment through either resource consumption or by releasing a factor to which EDM530 was more susceptible (Figure 2A and 2B). The latter hypothesis was rejected after repeating the *E.coli* strain competitions in the presence of filtered supernatant failed to reverse the competition trajectories (Figure S10). HNMR spectroscopy also failed to identify any

specific compound affecting the competitions. However, simply lowering the starting nutrient concentration resulted in the same change in competitive dynamics (Figure 3) as observed when *C.perfringens* was added to the medium. We were able to attribute this effect to peptone availability (Figure 4). These observations are in accordance with growth rate measurements that suggest a sHULK relationship between EDM106 and EDM530. They also suggest a concrete mechanism through which context dependent interactions can occur between organisms on one trophic level.

Dogma dictates that organisms evolve towards maximal metabolic efficiency, but this does not explain how a metabolically effective organism can be at a disadvantage when a resource is abundant. However, a phenotype has been observed in several species of bacteria, including *E.coli,* showing growth inhibition in the presence of high concentrations of certain amino acids or combinations thereof (De et al., 1979). This may provide insight into reasons why high nutrient concentrations could affect the observed strain level competitive dynamics. *C.perfringens* is known to be able to drastically alter environments rich in amino acids because of enhanced proteolytic capabilities and high growth rates (Fonknechten et al., 2010; Shimizu et al., 2002). This may, in turn, change the intra-specific competitive interaction in favour of the low K_s strain rather than the high μ_{max} strain.

It is difficult to classify the context-dependent interaction that we observed as either TMII or DMII. While the density effect of *C.perfringens* on the abundance of the two *E.coli*

strains in co-culture is evident, the effect on the strain interaction does not seem to be density dependent (Figure S12). Rather the effect is mediated through an environmental intermediary. It stands to reason that this would alter the physiological state, and hence the phenotype, of the *E.coli* strains, possibly in a differential manner, but that is not something we can conclude based on our data. Even so it would be incorrect to call this strain competition strict TMII since there is no direct biotic link between the context-dependent effect on *E.coli* and the intervening species.

Genome comparison

Previous investigation into colonization determinants of *E.coli* strains has mostly focused on adhesins and other factors that mediate host interactions. This is understandable considering the importance of pathogenic *E.coli* in human health. However, gut bacteria reside in complex communities in which bacteria-bacteria interactions occur mainly through resource competition, and the term virulence deserves, and has begun to receive, a broader definition that includes metabolic capabilities (Brussow et al., 2004; Kamada et al., 2012). This is especially important in the case *E.coli* because commensal and pathogenic strains share so many genomic features (Rasko et al., 2008).

A general genome comparison of the three strains competed in this study found the number of genes in core genome, 72% of the annotated total, were well within the normal variance seen between *E.coli* genomes (Lukjancenko et al., 2010). Investigation into several gene pathways hypothesized to be important for competition in our model system

found differences between strains in peptide uptake and utilization, sugar uptake and utilization, and quorum sensing pathways (Supplementary files). EDM106 is enriched compared with EDM530 for small molecule and secondary metabolic processes (Figure 5). This strain also encodes a unique putative oligo-peptide ABC transporter that could help explain the increased peptide affinity at low concentrations. Competition outcomes could also be influenced by strain specific differences in quorum sensing (Vendeville et al., 2005). EDM106 has the *luxS* gene for production of autoinducer-2, but lacks a functional suite of *lsr* genes for quorum response. 'Signal blind' mutants had been found to have higher fitness than their wild type parent strain when present as a minority (Diggle et al., 2007; Hibbing et al., 2010). The relative enrichment of stress response genes of EDM106 (Figure 5) may help explain the advantage of EDM106 later in the competition, when toxic metabolic by-product concentrations are high, although further work would be required to make this claim.

EDM530 is enriched for oxidation reduction processes relative to both EDM106 and EDM116 which probably relates to the high anaerobic growth rate relative to the other two strains (Figure 5). EDM530 also encodes several unique systems for transmembrane transport and catabolism of sugars, including the *yihO* gene encoding a glucuronide transporter. This transporter could have impacts on competition and colonization in the infant gut since glucuronide is a major carbon source for *E.coli* in the intestine (Chang et al., 2004; Miranda et al., 2004).

Strengths, limitations and concluding remarks

The inherent limitations of our model system make it all the more surprising that we found some of the same competitive outcomes that were observed in the gut environment. Several pathways were identified in the genomes that could have influenced competitive outcomes and could lead to further investigations to solidify a bridge between gene content and competitive outcomes in the natural environment. It is difficult to ascribe our results to particular genes or pathways, especially due to the number of unique genes that are annotated as hypothetical proteins in each of the strains. However, we do provide genomic profiles of an ecological gleaner in comparison with two exploiter phenotypes. We also present evidence of context dependent competition in bacteria, and we propose mechanisms that can promote this phenomenon. Understanding the population ecology of gut bacteria is of increasing importance with increased use of antibiotics and probiotics for therapeutic ends without knowledge of possible cascading effects (Costello et al., 2012).

Our observations offer a general mechanism by which the fine scale dynamics of microbial communities can be determined by biotic processes. In a system where a high degree of exploitation competition takes place on several taxonomic levels, the ability of keystone taxa to remodel the abiotic environment may have profound effects on community structure. In the present case we can easily envision two scenarios where one of the two *E.coli* strains outcompetes the other, depending who is dominating the background community (*Clostridia* or *Bacteroides*). Context-dependent competition most

likely represents a general phenomenon where community composition at high taxonomic levels determines the outcomes of strain level colonization processes by remodelling the environment to become more permissive to some strains than others. This suggests a mechanism by which temporal changes in a limiting nutrient concentration can promote coexistence by changing the competitive interactions between strains. Context-dependent competition may be especially important in the GI-system since it is subject to a pulse-like nutrient regime where key resources alternate between high and low concentrations.

Acknowledgments

Tim Coulson and Thomas Owens Svennungsen for close reading and very helpful comments. Eva Aas for laboratory assistance. The Norwegian Foundation for Health and Rehabilitation and the Centre for Ecological and Evolutionary Synthesis (CEES) for funding this research and Department of Chemistry, University of Oslo, Norway for use of the NMR facility.

Abrams PA (1995). Implications of Dynamically Variable Traits for Identifying, Classifying, and Measuring Direct and Indirect Effects in Ecological Communities. *Am Nat* **146:** 112-134.

Abrams PA (2004). When does periodic variation in resource growth allow robust coexistence of competing consumer species? *Ecology* **85:** 372-382.

Adlerberth I (2008). Factors influencing the establishment of the intestinal microbiota in infancy. *Nestle Nutrition workshop series Paediatric programme* **62:** 13-29; discussion 29-33.

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

Baumgart M, Dogan B, Rishniw M, Weitzman G, Bosworth B, Yantiss R *et al* (2007). Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *The ISME journal* **1:** 403-418.

Brown JH, Whitham TG, Morgan Ernest SK, Gehring CA (2001). Complex species interactions and the dynamics of ecological systems: long-term experiments. *Science* **293:** 643-650.

Brussow H, Canchaya C, Hardt WD (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and molecular biology reviews : MMBR* **68:** 560-602.

Chang DE, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ *et al* (2004). Carbon nutrition of Escherichia coli in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 7427-7432.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21:** 3674-3676.

Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336:** 1255-1262.

De Felice M, Levinthal M, Iaccarino M, Guardiola J (1979). Growth inhibition as a consequence of antagonism between related amino acids: effect of valine in Escherichia coli K-12. *Microbiological reviews* **43:** 42-58.

de Muinck EJ, Oien T, Storro O, Johnsen R, Stenseth NC, Ronningen KS *et al* (2011). Diversity, transmission and persistence of Escherichia coli in a cohort of mothers and their infants. *Env Microbiol Rep* **3:** 352-359.

Diggle SP, Griffin AS, Campbell GS, West SA (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature* **450:** 411-414.

Fonknechten N, Chaussonnerie S, Tricot S, Lajus A, Andreesen JR, Perchat N *et al* (2010). Clostridium sticklandii, a specialist in amino acid degradation:revisiting its metabolism through its genome sequence. *BMC genomics* **11:** 555.

Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 13780-13785.

Fredrickson AG, Stephanopoulos G (1981). Microbial competition. *Science* **213:** 972-979.

Friswell M, Campbell B, Rhodes J (2010). The role of bacteria in the pathogenesis of inflammatory bowel disease. *Gut and liver* **4:** 295-306.

Fuller R (1989). Probiotics in man and animals. *The Journal of applied bacteriology* **66:** 365-378.

Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G *et al* (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *The ISME journal* **5:** 82-91.

Gottschal JC (1993). Growth kinetics and competition--some contemporary comments. *Antonie van Leeuwenhoek* **63:** 299-313.

Grover JP (1997). *Resource competition*, 1st edn. Chapman & Hall: London ; New York.

Harcombe WR, Bull JJ (2005). Impact of phages on two-species bacterial communities. *Appl Environ Microbiol* **71:** 5254-5259.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature reviews Microbiology* **8:** 15-25.

Johnson LR (2000). *Gastrointestinal physiology*, 6th edn. Mosby: St. Louis.

Kamada N, Kim YG, Sham HP, Vallance BA, Puente JL, Martens EC *et al* (2012). Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. *Science* **336:** 1325-1329.

Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI (2011). Human nutrition, the gut microbiome and the immune system. *Nature* **474:** 327-336.

Kiesecker JM, Blaustein AR (1999). Pathogen reverses competition between larval amphibians. *Ecology* **80:** 2442-2448.

575 Kotlowski R, Bernstein CN, Sepehri S, Krause DO (2007). High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. *Gut* **56:** 669-675.

Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, Pedersen BK *et* 580 *al* (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS one* **5:** e9085.

Lévêque C (2003). *Ecology from ecosystem to biosphere*. Science Publishers: Enfield, NH.

585

Levin BR (1972). Coexistence of two asexual strains on a single resource. *Science* **175:** 1272-1274.

Lukjancenko O, Wassenaar TM, Ussery DW (2010). Comparison of 61 sequenced 590 Escherichia coli genomes. *Microbial ecology* **60:** 708-720.

Miranda RL, Conway T, Leatham MP, Chang DE, Norris WE, Allen JH *et al* (2004). Glycolytic and gluconeogenic growth of Escherichia coli O157:H7 (EDL933) and E.coli K-12 (MG1655) in the mouse intestine. *Infection and immunity* **72:** 1666-1676.

595

Monod J (1949). The Growth of Bacterial Cultures. *Annual review of microbiology* **3:** 371-394.

Peacor SD, Werner EE (1997). Trait-mediated indirect interactions in a simple aquatic
600 food web. *Ecology* **78:** 1146-1156.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

605 Rainey PB, Buckling A, Kassen R, Travisano M (2000). The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol Evol* **15:** 243-247.

Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P *et al* (2008).
610 The pangenome structure of Escherichia coli: Comparative genomic analysis of E-coli commensal and pathogenic isolates. *J Bacteriol* **190:** 6881-6893.

Relyea RA (2000). Trait-mediated indirect effects in larval anurans: Reversing competition with the threat of predation. *Ecology* **81:** 2278-2289.

615

Rhodes JM (2007). The role of Escherichia coli in inflammatory bowel disease. *Gut* **56:** 610-612.

Rudi K, Storro O, Oien T, Johnsen R (2012). Modelling bacterial transmission in human
620 allergen-specific IgE sensitization. *Letters in applied microbiology* **54:** 447-454.

Savageau MA (1983). Escherichia-Coli Habitats, Cell-Types, and Molecular Mechanisms of Gene-Control. *Am Nat* **122:** 732-744.

625 Sepehri S, Kotlowski R, Bernstein CN, Krause DO (2009). Phylogenetic Analysis of Inflammatory Bowel Disease Associated Escherichia coli and the FimH Virulence Determinant. *Inflammatory bowel diseases* **15:** 1737-1745.

Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T *et al* (2002).
630 Complete genome sequence of Clostridium perfringens, an anaerobic flesh-eater. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 996-1001.

Stewart FM, Levin BR (1973). Partitioning of Resources and Outcome of Interspecific Competition - Model and Some General Considerations. *Am Nat* **107:** 171-198.

Storro O, Oien T, Langsrud O, Rudi K, Dotterud C, Johnsen R (2011). Temporal variations in early gut microbial colonization are associated with allergen-specific immunoglobulin E but not atopic eczema at 2 years of age. *Clinical and Experimental Allergy* **41:** 1545-1554.

Tenaillon O, Skurnik D, Picard B, Denamur E (2010). The population genetics of commensal Escherichia coli. *Nature reviews Microbiology* **8:** 207-217.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P *et al* (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genetics* **5:** e1000344.

Treves DS, Manning S, Adams J (1998). Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Molecular biology and evolution* **15:** 789-797.

Trosvik P, Skanseng B, Jakobsen KS, Stenseth NC, Naes T, Rudi K (2007). Multivariate analysis of complex DNA sequence electropherograms for high-throughput quantitative analysis of mixed microbial populations. *Appl Environ Microbiol* **73:** 4975-4983.

Trosvik P, Rudi K, Straetkvern KO, Jakobsen KS, Naes T, Stenseth NC (2010). Web of ecological interactions in an experimental gut microbiota. *Environmental microbiology* **12:** 2677-2687.

Trosvik P, Stenseth NC, Rudi K (2010). Convergent temporal dynamics of the human infant gut microbiota. *The ISME journal* **4:** 151-158.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444:** 1027-1031.

Vasi F, Travisano M, Lenski RE (1994). Long-Term Experimental Evolution in Escherichia-Coli .2. Changes in Life-History Traits during Adaptation to a Seasonal Environment. *Am Nat* **144:** 432-456.

Vejborg RM, Hancock V, Petersen AM, Krogfelt KA, Klemm P (2011). Comparative genomics of Escherichia coli isolated from patients with inflammatory bowel disease. *BMC genomics* **12**.

675

Vendeville A, Winzer K, Heurlier K, Tang CM, Hardie KR (2005). Making 'sense' of metabolism: autoinducer-2, LuxS and pathogenic bacteria. *Nature reviews Microbiology* **3:** 383-396.

680 Vieira-Silva S, Rocha EP (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS genetics* **6:** e1000808.

Werner EE, Peacor SD (2003). A review of trait-mediated indirect interactions in ecological communities. *Ecology* **84:** 1083-1100.

685

Wissinger S, Mcgrady J (1993). Intraguild Predation and Competition between Larval Dragonflies - Direct and Indirect Effects on Shared Prey. *Ecology* **74:** 207-218.

690

Figure 1: Intra-specific competition is modulated by the resident background community. Percent relative abundances are plotted as a function of time for the competitions between *C.perfringens* (C.perf), *B.thetaiotaomicron* (B.tio), *B. longum* and *E.coli* strains EDM106 and EDM530. Experiments were carried out in duplicate. A. Relative *E.coli* strain abundances. B. Relative species abundances. After day ten, competitive strain trajectories change, coinciding with *C.perfringens* dominance.

Figure 2A and 2B: Intra-specific competition is context dependent. One competition was inoculated with *B.thetaiotaomicron* (B.tio) (A and B) while the other was inoculated with *C. perfringens* (C and D) at day 12 of the experiment. Experiments were carried out in duplicate. Top panels (A and C) show *E.coli* strain relative abundances. Bottom panels (B and D) show species relative abundances. Competition trajectories of strains EDM106 and EDM530 reversed in the competition inoculated with *C.perfringens* relative to when the strains were alone or inoculated with *B.thetaiotaomicron.*

Figure 3: Strain competition outcomes are contingent on resource availability. After two days equilibration period in rich medium, aliquots were transferred into either 90/10 (A), 50/50 (B), 10/90 (C) rich/minimal salts medium (see Table S2 and supplementary methods for details) and the competitions were continued under these conditions. Experiments were carried out in duplicate. Strain competition trajectories shift in the low nutrient competition (C).

Figure 4: Peptones can act as a mediator of context-dependent competition. Two independent strain competition experiments were performed in duplicate in flasks containing minimal salts medium and different concentrations of peptone (Table S2). All experiments were carried out in duplicate. For the experiments shown in (A-C), The revived stocks were allowed two days equilibration period in Oxoid anaerobe basal broth, aliquots were then transferred into 90/10 (A), 50/50 (B), or 10/90 (C) peptone/minimal salts medium (A-C). Peptone concentration had a pronounced effect on the competition trajectories ($p = 0.019$, logistic mixed effects model). For experiments shown in (D-F), a separate stock culture with a low relative starting abundance of strain EDM530 was inoculated into flasks into 200/10 (D), 50/50 (E), or 20/80 (F) peptone/minimal salts medium. Peptone concentration had a pronounced effect on the competition trajectories.

Figure 5: Comparisons of gene enrichment between strains. Strains were compared for GO enrichment using the SEED categorization derived from ontology level 3 biological process assignments using Blast2GO. Columns represent the six possible pairwise comparisons of EDM530, EDM106, and EDM116. (*)'s indicate the strain that is enriched for genes assigned to the biological process categories.

Figure 1

735

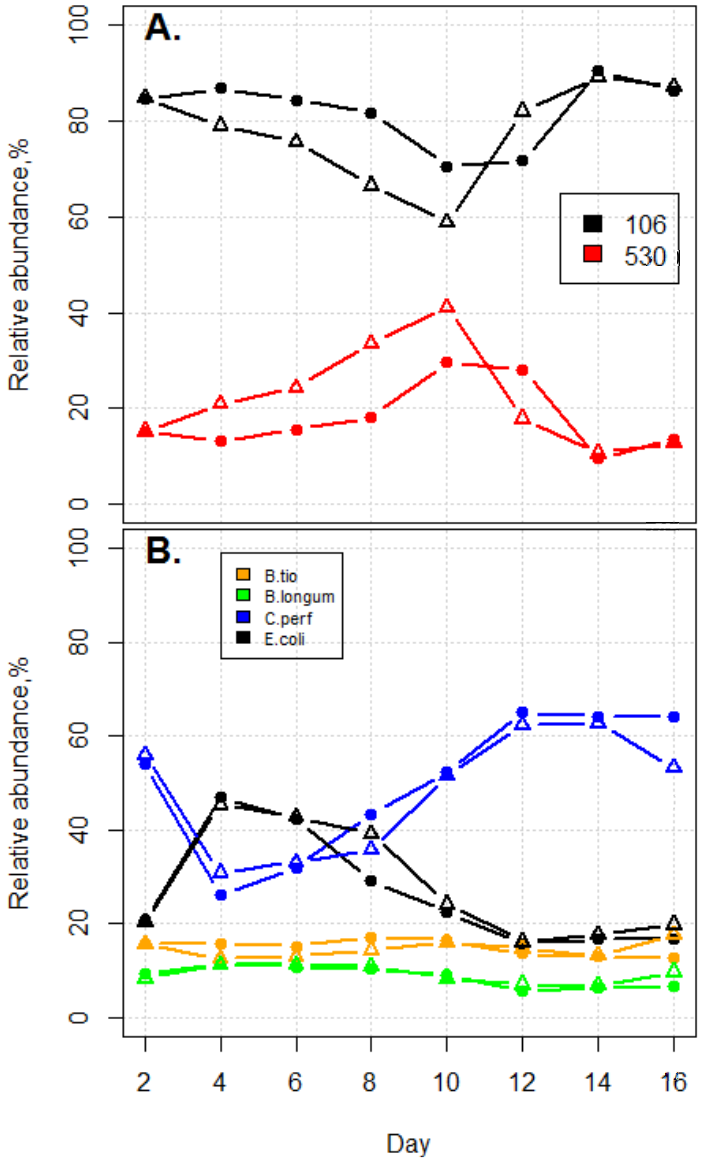
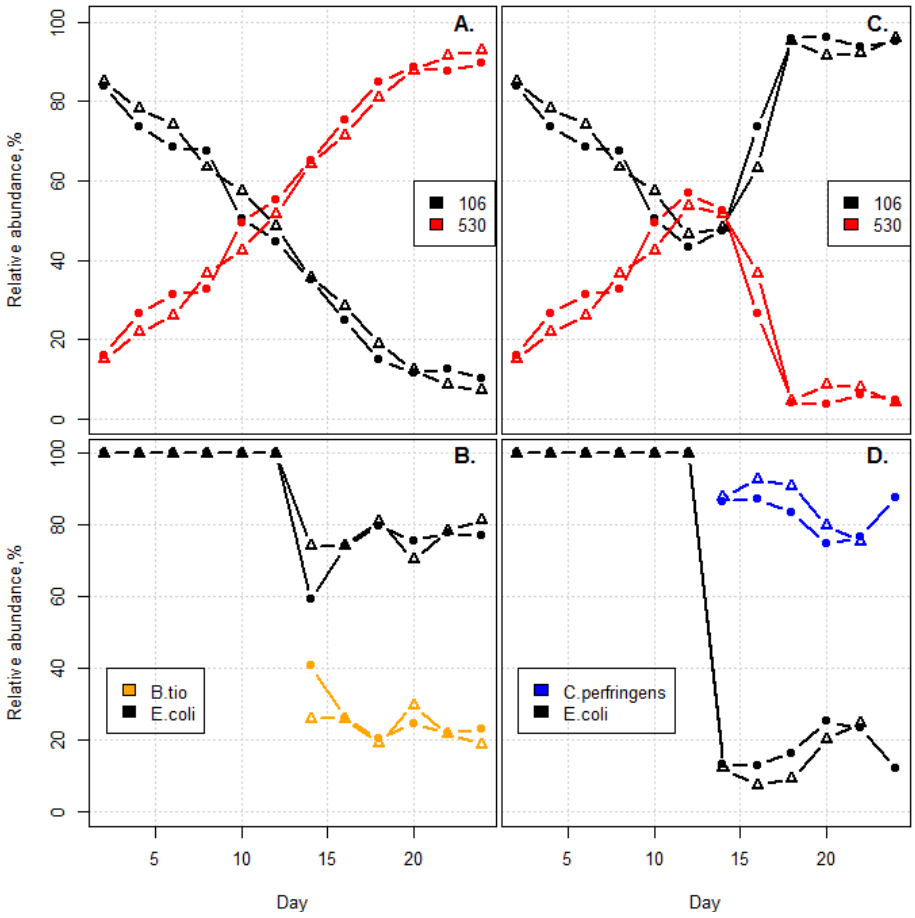

A.
Relative abundance,%
0
20
40
60
80
100
106
530
B.
B.tio
B.longum
C.perf
E.coli
2
4
6
8
10
12
14
16
Day

Figure 2

760



765

Figure 3

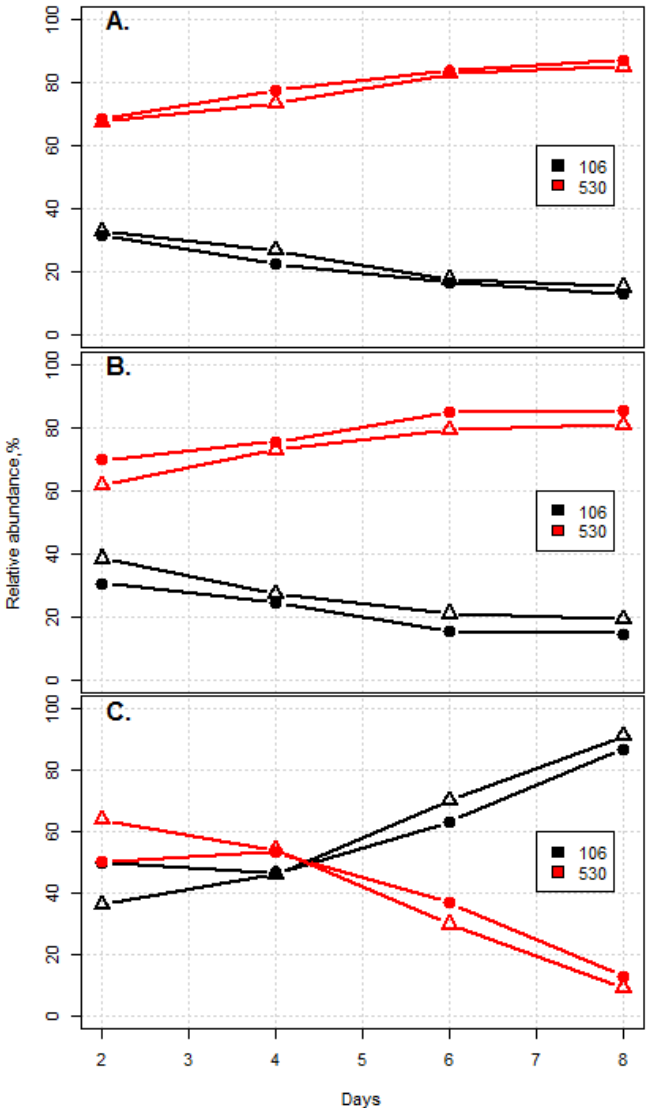


Figure 4

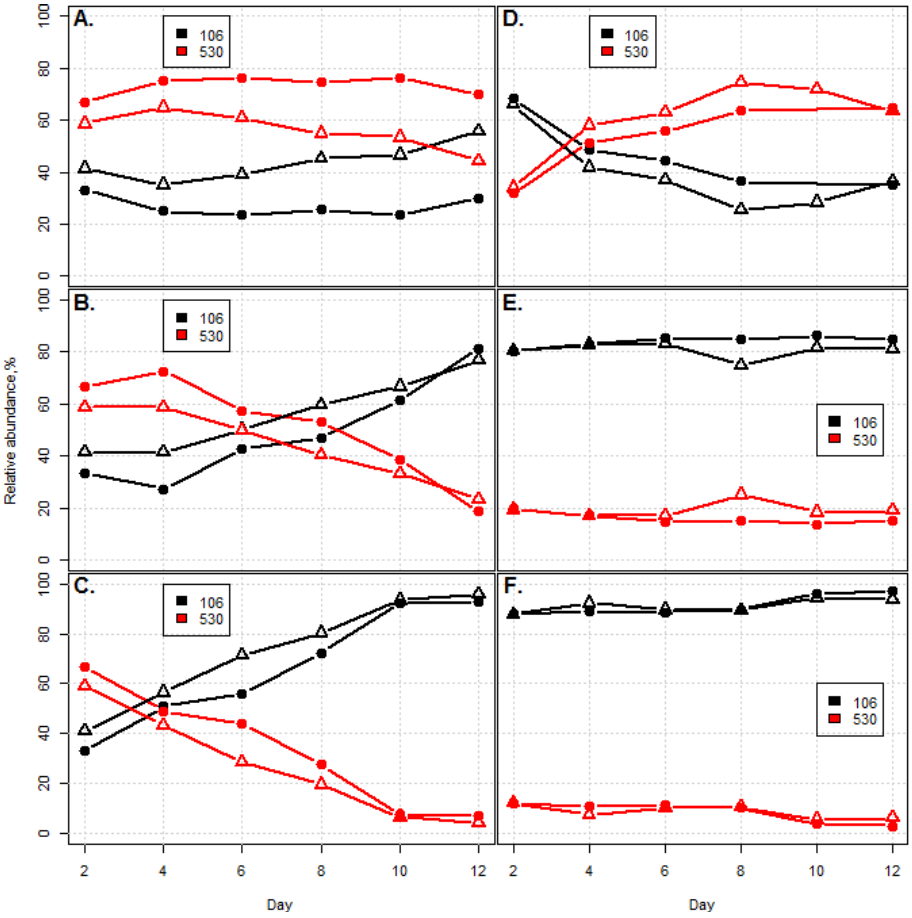
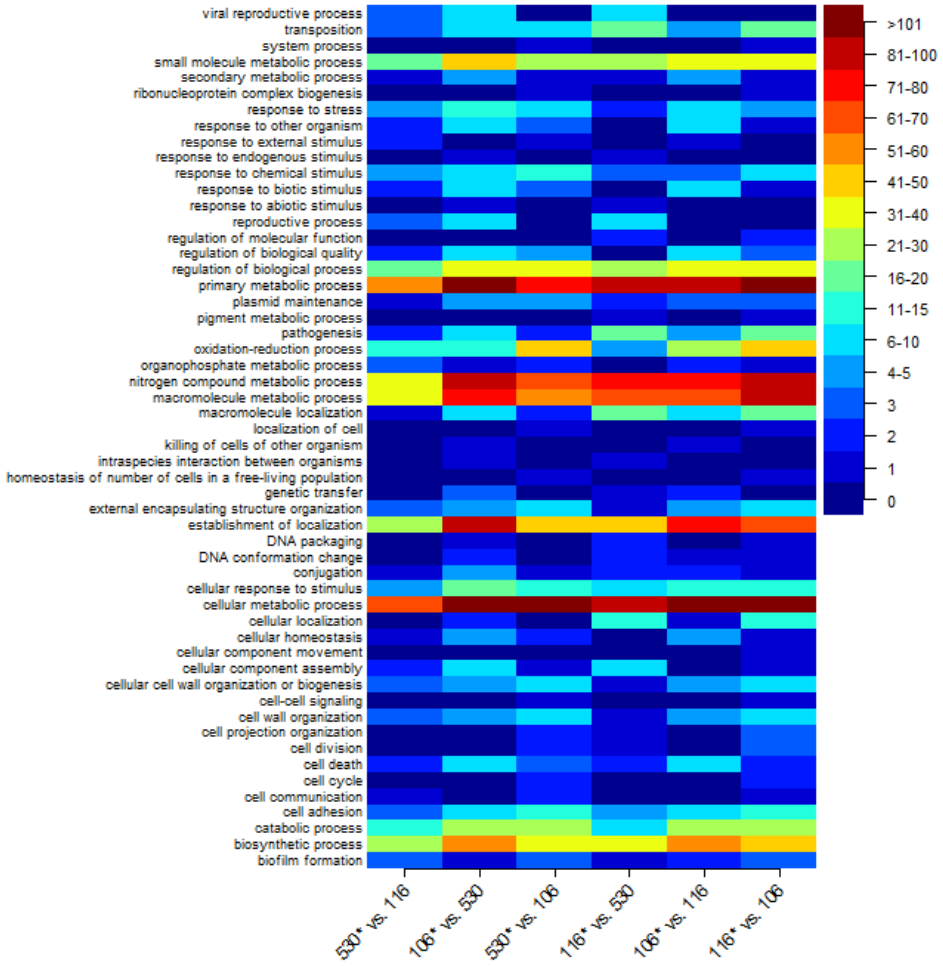

A.
B.
C.
D.
E.
F.
106
530
Relative abundance,%
Day
0
20
40
60
80
100
2
4
6
8
10
12

Figure 5

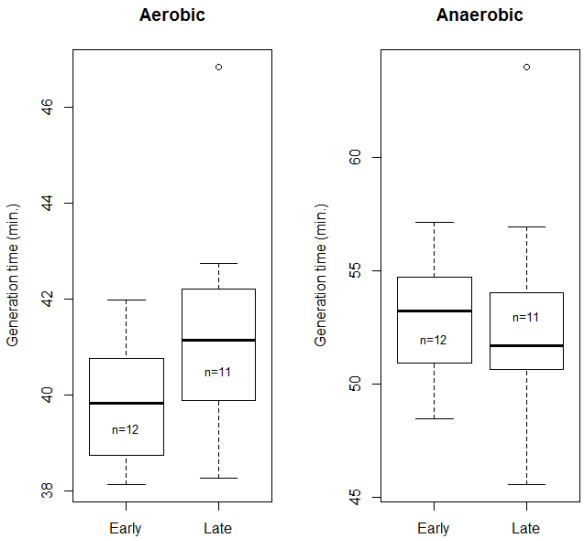

viral reproductive process
transposition
system process
small molecule metabolic process
secondary metabolic process
ribonucleoprotein complex biogenesis
response to stress
response to other organism
response to external stimulus
response to endogenous stimulus
response to chemical stimulus
response to biotic stimulus
response to abiotic stimulus
reproductive process
regulation of molecular function
regulation of biological quality
regulation of biological process
primary metabolic process
plasmid maintenance
pigment metabolic process
pathogenesis
oxidation-reduction process
organophosphate metabolic process
nitrogen compound metabolic process
macromolecule metabolic process
macromolecule localization
localization of cell
killing of cells of other organism
intraspecies interaction between organisms
homeostasis of number of cells in a free-living population
genetic transfer
external encapsulating structure organization
establishment of localization
DNA packaging
DNA conformation change
conjugation
cellular response to stimulus
cellular metabolic process
cellular localization
cellular homeostasis
cellular component movement
cellular component assembly
cellular cell wall organization or biogenesis
cell-cell signaling
cell wall organization
cell projection organization
cell division
cell death
cell cycle
cell communication
cell adhesion
catabolic process
biosynthetic process
biofilm formation
530* vs. 116
106* vs. 530
530* vs. 106
116* vs. 530
106* vs. 116
116* vs. 106
>101
81-100
71-80
61-70
51-60
41-50
31-40
21-30
16-20
11-15
6-10
4-5
3
2
1
0

795

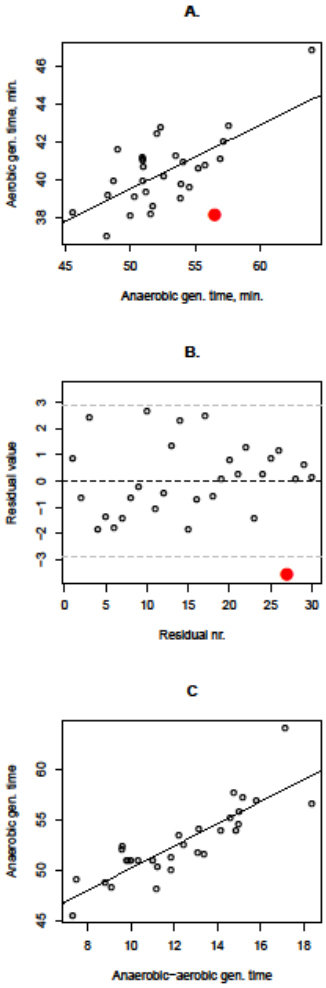
Supplementary figures (1-16),

Supplementary Tables (1-3),

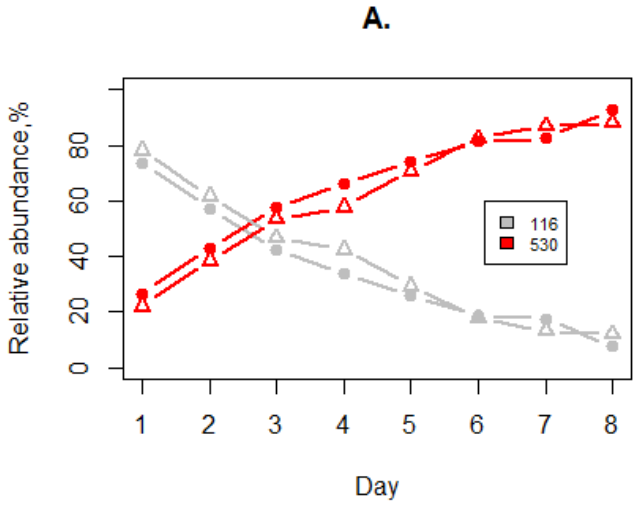
& Supplementary Methods

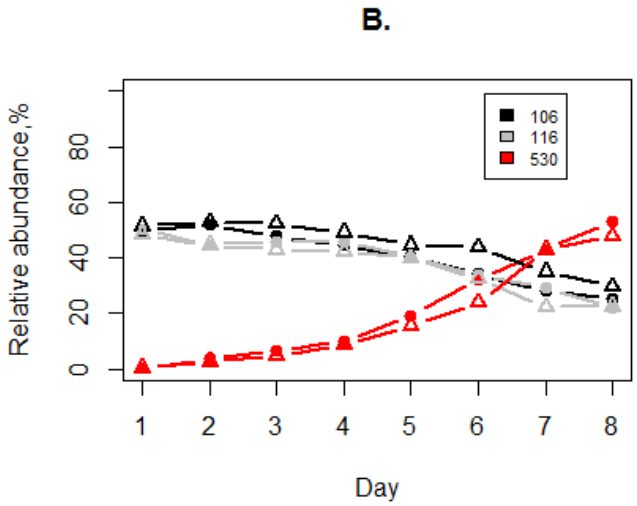


Supplementary Figure 1: **Early colonizers have a tendency towards faster growth rates than later colonizers under aerobic conditions.** 23 different *E. coli* strains were categorized as either early or late colonizers (see TABLE S1 for categorization) for comparison of growth rates in aerobic and anaerobic conditions. There was a tendency for early colonizers to have a shorter generation time than late colonizers in the aerobic environment ($p = 0.03$, one-tailed Mann-Whitney U test).

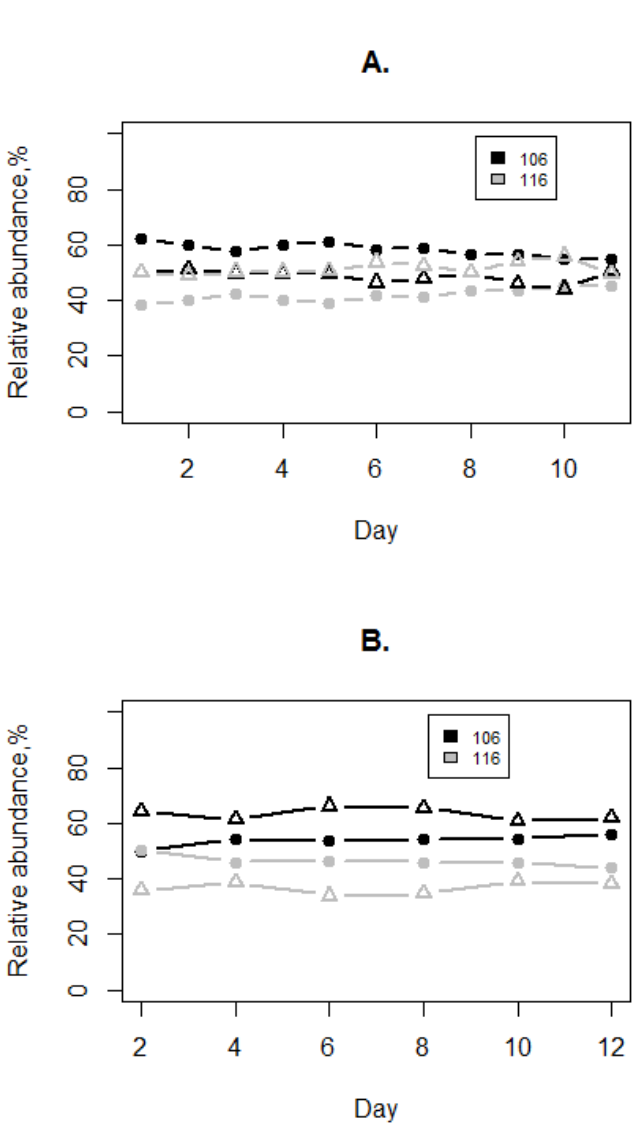


Supplementary Figure 2: Positive correlation between aerobic and aerobic growth rates. (A) A strong positive correlation is seen between aerobic and anaerobic growth rates. However, strain EDM106 (red) does not seem to follow the trend of the other strains. **(B)** This strain (red) deviates from the trend by more than two standard deviations and is due to a faster aerobic growth rate than should be for its anaerobic rate. **(C)** Even with this deviation, strain EDM106 and all the others show a strong relationship between anaerobic generation time and the difference between anaerobic and aerobic generation time. This indicates that fast growing strains are optimized for both conditions.

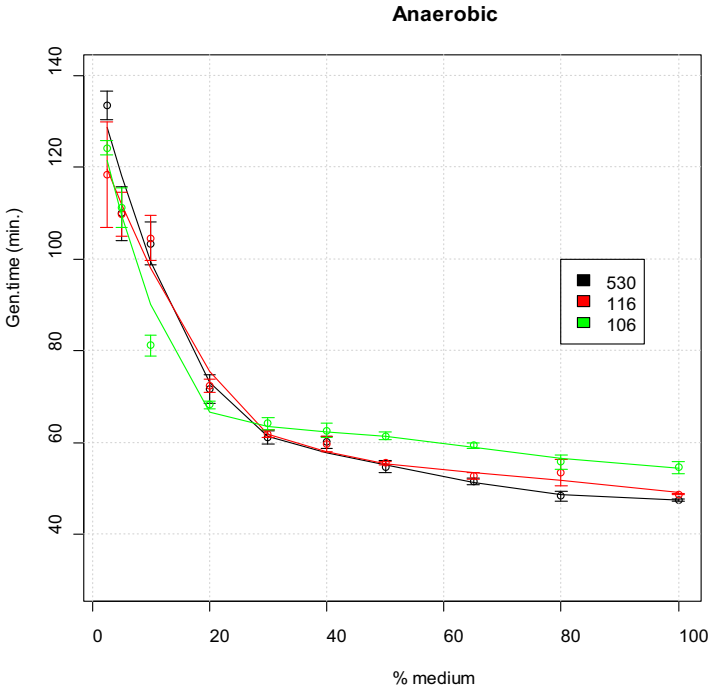




Supplementary Figure 3: Strain EDM530 dominates in competition with strain EDM116 as well as with the combination of strains EDM106 and EDM116. (**A**) Strain EDM116 vs. EDM530. (**B**) Strain EDM106 vs. EDM116 vs. 530. Batch culture competitions were performed in Oxoid anaerobe basal broth. See supplementary methods for details. The two independent replicates are represented by solid circles and open triangles.



Supplementary Figure 4: Strains EDM106 and EDM116 coexist in one day and two day culture regimes. Batch culture competitions of strains EDM106 and EDM116 were performed in Oxoid anaerobe basal broth (rich medium). (**A**) One day tranfer regime. (**B**) Two day transfer regime. The two independent replicates are represented by solid circles and open triangles.



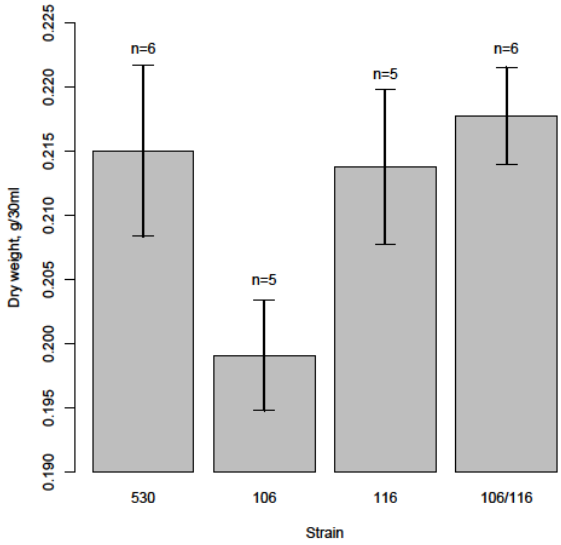
875

Supplementary Figure 5: Strain EDM106 has shorter generation times under low nutrient conditions than strain EDM116 and strain EDM530. Doubling times were measured for each of the three strains with different concentrations of Oxoid anaerobic basal broth medium (%medium) and minimal salts solution under anaerobic conditions.

880

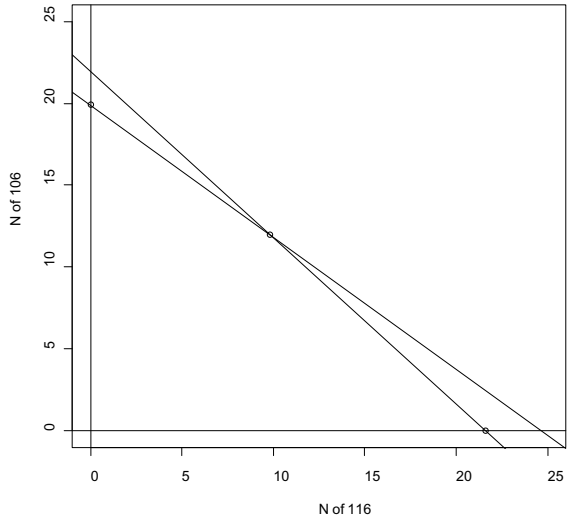
885

A

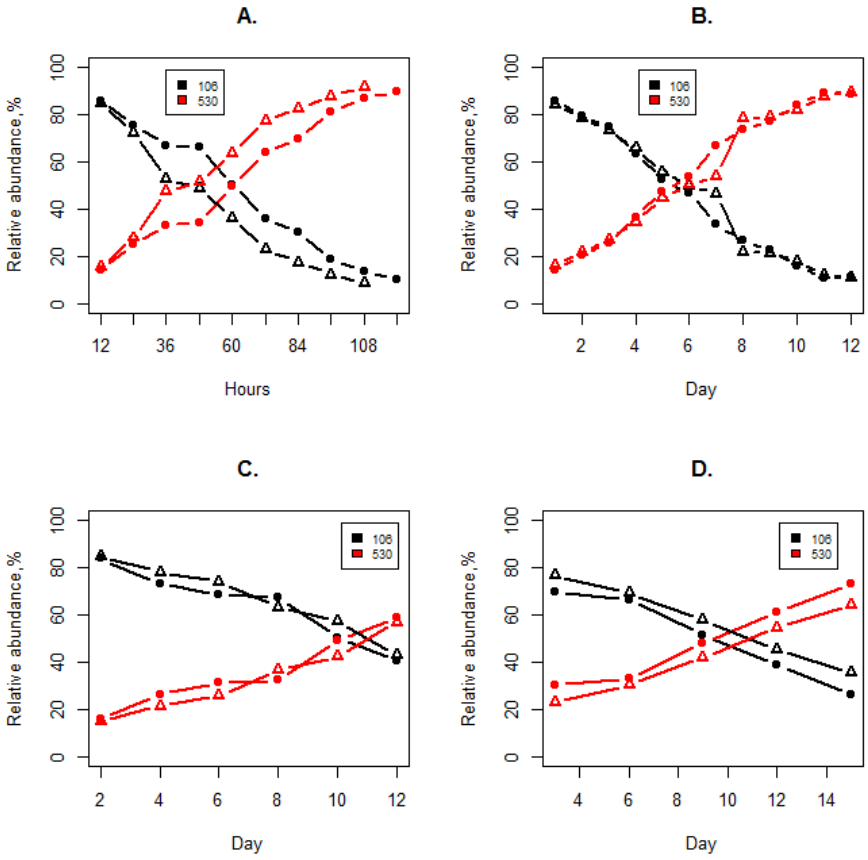

Dry weight, g/30ml
0.190
0.195
0.200
0.205
0.210
0.215
0.220
0.225
n=6
n=5
n=5
n=6
530
106
116
106/116
Strain

890

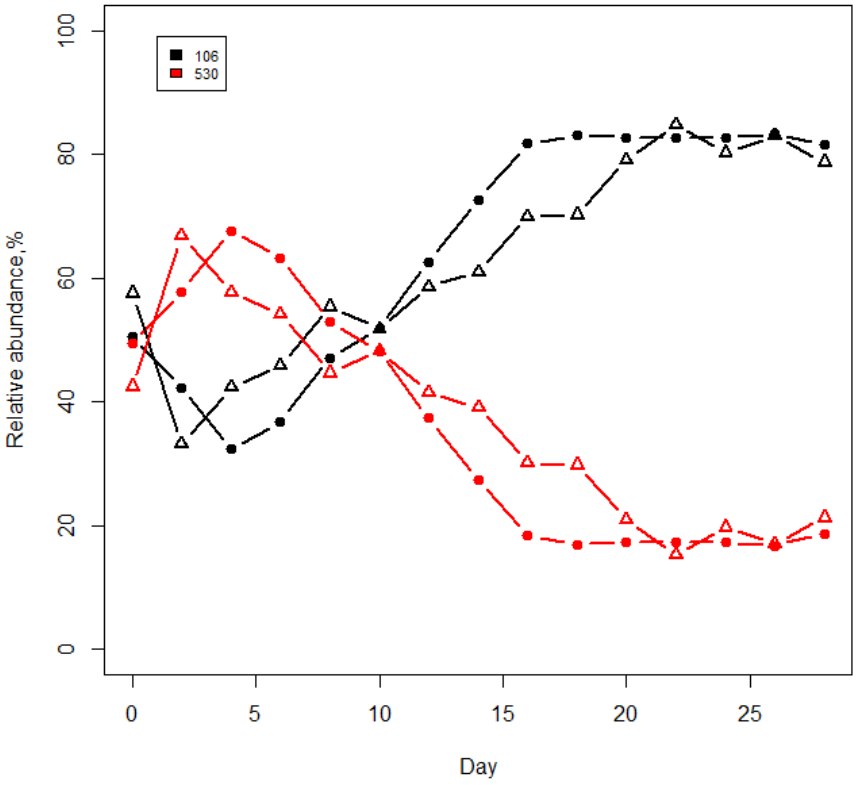
B


N of 106
0
5
10
15
20
25
N of 116

Supplementary Figure 6: Increased carrying capacity in co-cultures of strains EDM106 and EDM116 facilitate co-existence (**A**) Carrying capacity of competitor strains in dry weight (grams) per 30ml medium (± s.e). The co-culture has a higher carrying capacity ($p = 0.024$, one sample t-test) than expected from combining the mean carrying capacities of the individual strains in the approximate proportions provided by the competition experiments (55% strain EDM106 and 45% strain EDM116 during co-culture). (**B**) Isoclines of carrying capacities were calculated using the approximate proportions provided by competition experiments. The intersecting isoclines show a stability point for co-existence. Carrying capacities used in the model were 19.9grams for strain 106, 21.6grams for strain 116, and 21.8 grams for the co-culture of strains EDM106 and EDM116.

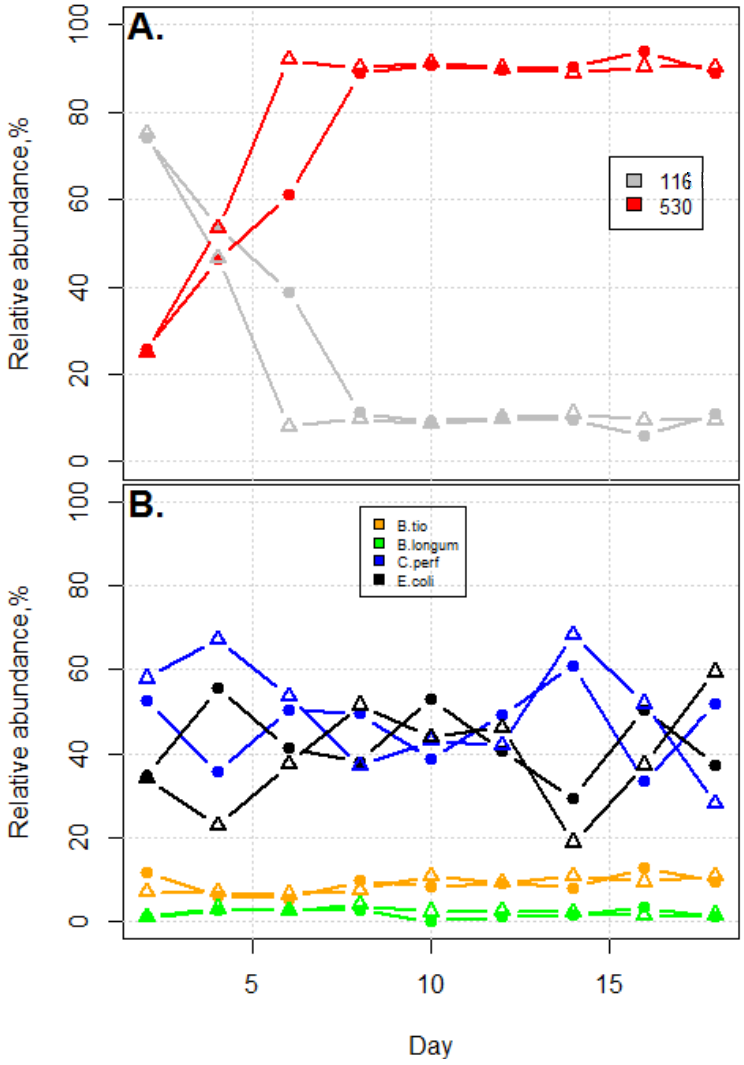


Supplementary Figure 7: Strain EDM106 was dominated by strain EDM530 under four different transfer regimes. Cultures were transferred every 12 hours (**A**), 24 hours (**B**), 2 days (**C**) or 3 days (**D**). Batch culture competitions were performed in Oxoid anaerobe basal broth. The two independent replicates are represented by solid circles and open triangles.

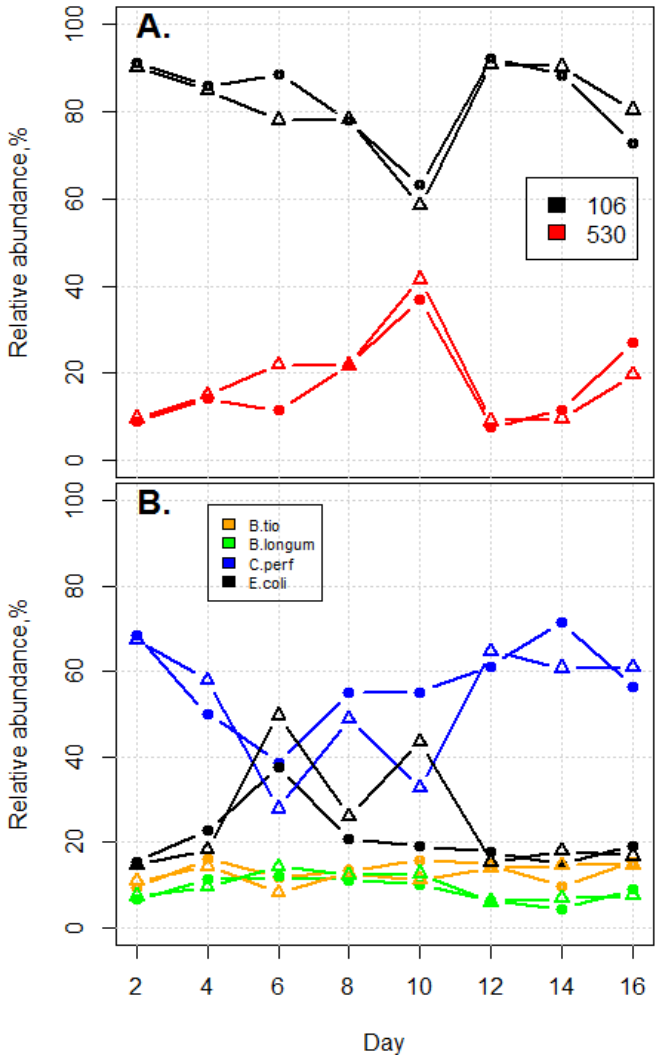


Supplementary Figure 8: Long term stationary phase cultures of strains EDM 106 and EDM530. Cultures were sampled for relative abundance measurement every two days but no fresh media was added to the cultures. The two independent replicates are represented by solid circles and open triangles.

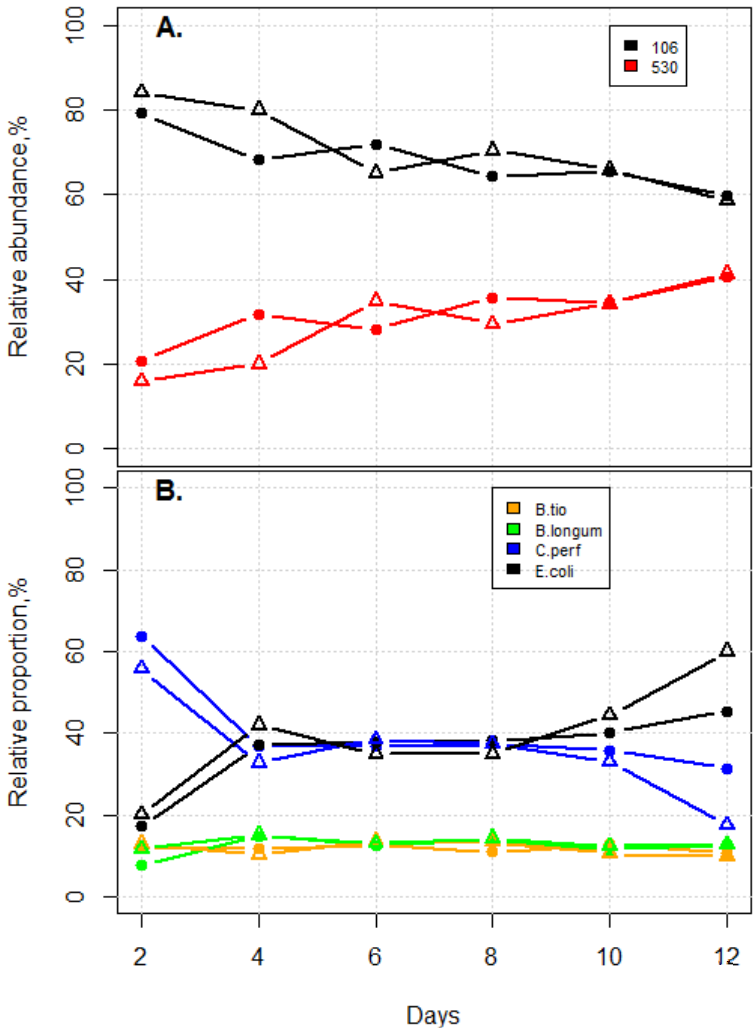
930



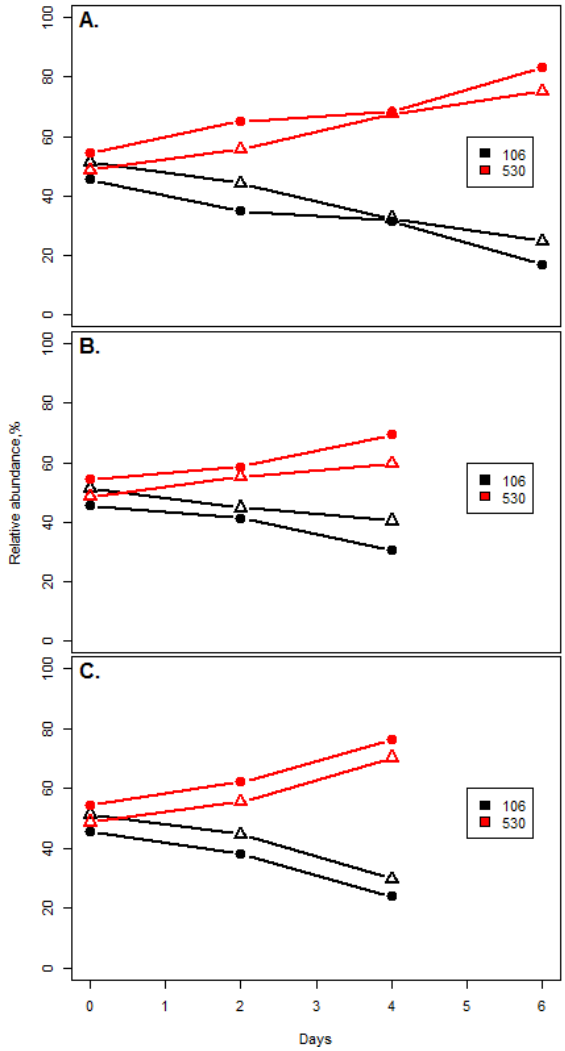
Supplementary Figure 9: The presence of a background flora does not change the outcome of competition between strains EDM116 and EDM530. Competition between *C. perfringens* (C.perf), *B.thetaiotaomicron* (B.tio) , *B.longum* and *E.coli* strains EDM116 and EDM530. Each experiment was performed in duplicate. **(A)** *E. coli* strain competition trajectories. **(B)** Relative species abundances at corresponding time-points. Batch culture competitions were performed in Oxoid anaerobe basal broth. The two independent replicates are represented by solid circles and open triangles.



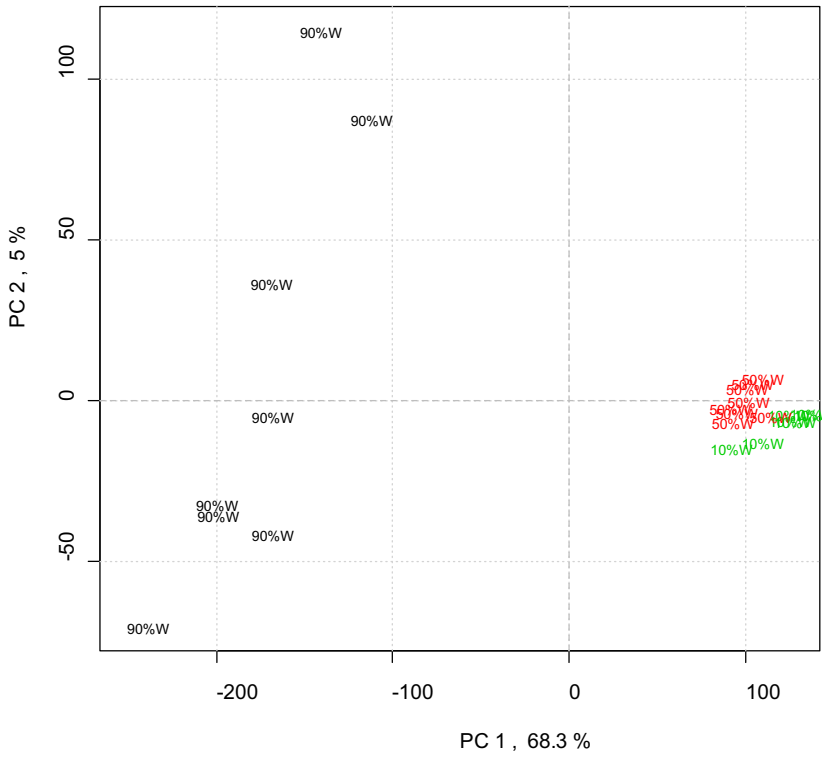
Supplementary Figure 10: Replicate of experiment presented in Figure 1: Intra-specific competition is modulated by the resident background community. Percent relative abundances are plotted as a function of time for the competitions between *C.perfringens* (C.perf), *B.thetaiotaomicron* (B.tio), *B.longum* and *E.coli* strains EDM106 and EDM530. Experiments were carried out in duplicate. **(A)** Relative E.coli strain abundances. **(B)** Relative species abundances. After day ten, competitive strain trajectories change, coinciding with C.perfringens dominance. The two independent replicates are represented by solid circles and open triangles.



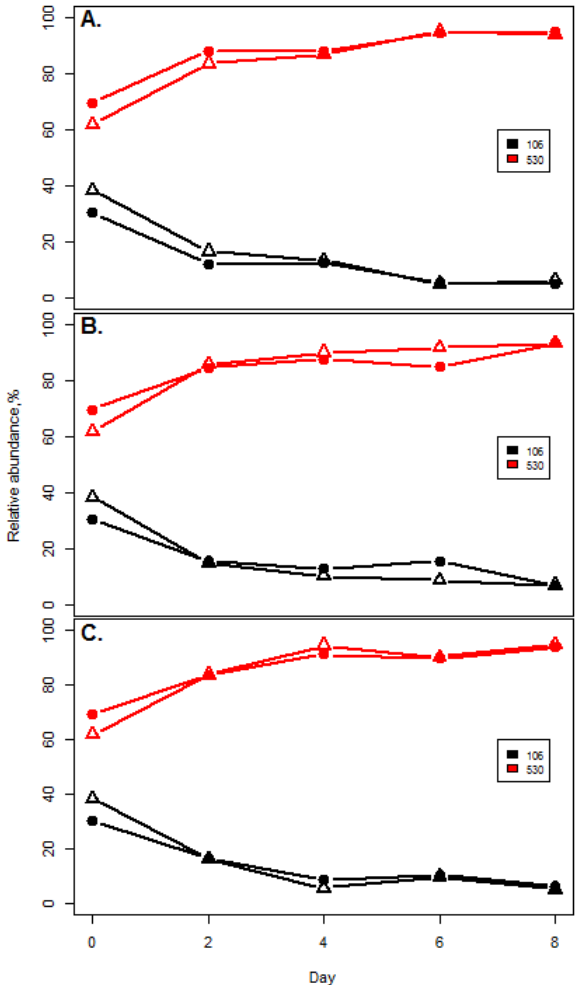
Supplementary Figure 11: Micro-aerophilic conditions increase *E.coli* predominance in the model microbiota, resulting in unmodulated *E.coli* strain competition trajectories. Percent relative abundances are plotted as a function of time for the competitions between *C.perfringens* (C.perf), *B.thetaiotaomicron* (B.tio), *B.longum* and *E.coli* strains EDM106 and EDM530. Experiments were carried out in duplicate. **(A)** Relative E.coli strain abundances. **(B)** Relative species abundances. After day ten, competitive strain trajectories change, coinciding with C.perfringens dominance. The two independent replicates are represented by solid circles and open triangles.



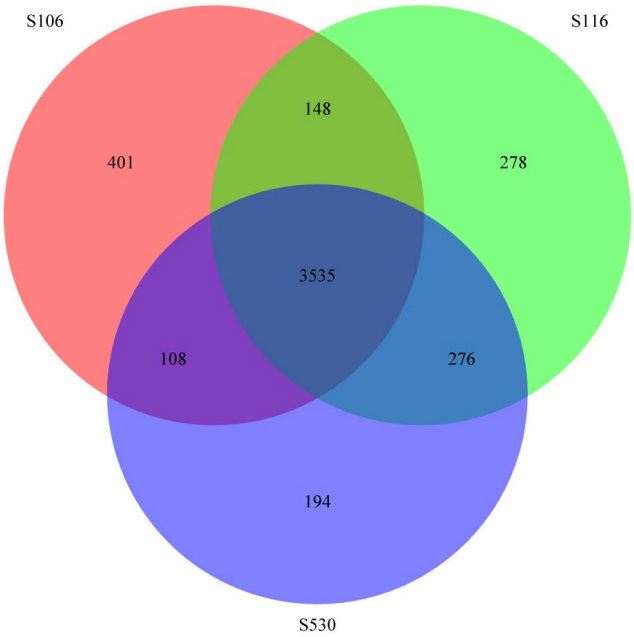
Supplementary Figure 12: Spent medium from a *C. perfringens* culture did not affect competition between strains EDM106 and EDM530. Investigation into potential factors released into the media was performed by growing *C. perfringens* to saturation in Oxoid anaerobic basal broth and then removing cells by filtration to create a spent rich medium. (**A**) 90/10, (**B**) 50/50, or (**C**) 10/90 (spent/fresh) medium proportions were then used for the *E.coli* strain competitions. The two independent replicates are represented by solid circles and open triangles.



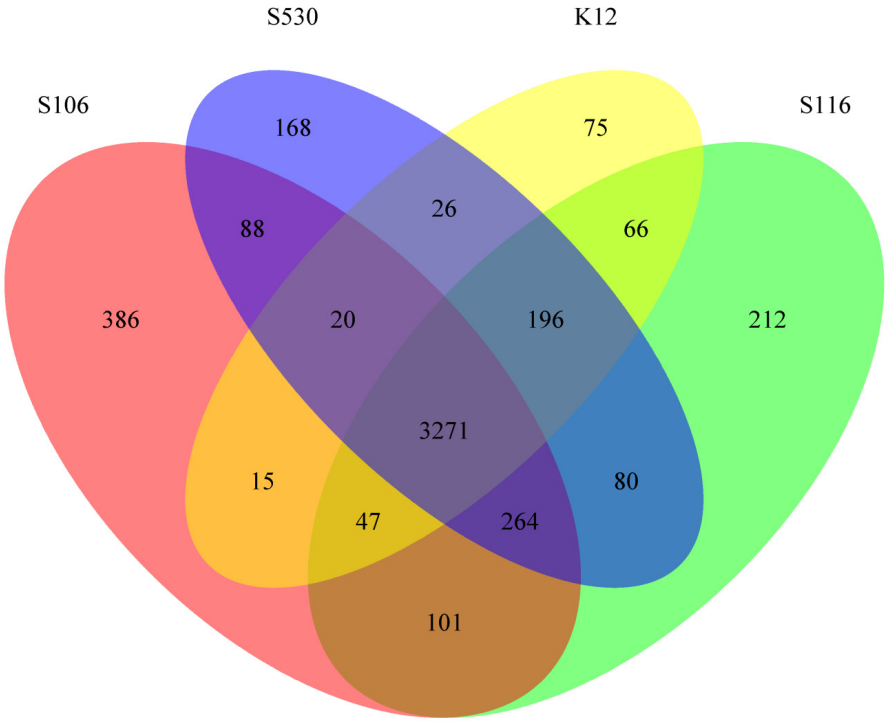
Supplementary Figure 13: Competitions in low nutrient media produce qualitatively different media than more concentrated media. PCA of normalized HNMR spectra of competition time points in 90/10, 50/50, 10/90 minimal salts (W)/ Oxoid anaerobic basal broth (rich medium). Clustering differentiates the spectra of the 10/90 and 50/50 (W)/rich medium competition supernatants compared with 90/10 medium.



Supplementary Figure 14: Strain EDM106 and EDM530 competition trajectories remain unchanged at different levels of glucose availability. Minimal salts medium with different amounts of glucose as the sole carbon source was used for the competitions (Table S2). Revived frozen stocks from day 10 of strain EDM106 and strain EDM530 competition were used to start the competition. After two days equilibration in Oxoid anaerobic basal broth (rich medium), aliquots were transferred into either (**A**) 90/10, (**B**) 50/50 or (**C**) 10/90 glucose/minimal salts medium, relative to the amount of glucose in the rich medium (table S2). The two independent replicates are represented by solid circles and open triangles.



Supplementary Figure 15: Comparison of gene content between strains. The three strains had totals of 4,192 genes (S106), 4,237 genes (S116) and 4113 genes (S530). The relative percents of unique genes found S106 with the largest (9.7%), 6.6% for S116, and S530 had the least (4.7%). Core genes represented 72% of the annotated total. See Supplementary files for listings of non-core gene annotations.



Supplementary Figure 16: Comparison of gene content between strains. The three strains had totals of 4,192 genes (S106), 4,237 genes (S116) and 4113 genes (S530). The relative percents of unique genes found S106 with the largest (9.7%), 6.6% for S116, and S530 had the least (4.7%). Core genes represented 72% of the annotated total.

Isolate number	Age at sampling	Aero	s.e	Anaero.	s.e.	Colonization Catagory
EDM24	7 days	39.91	0.1369	48.71	0.6465	Early*
EDM123	4 months	40.75	0.5328	55.75	1.116	Late
EDM530	2 years	41.56	0.7187	49.03	0.2424	Late
EDM124	8-10 days	38.12	1.082	49.98	1.79	Early
EDM70	10 days	39	0.3685	53.86	0.1729	Early*
EDM71	10 days	39.57	0.4597	54.54	0.4518	Early*
EDM49	4 days	40.62	0.3767	55.21	0.2947	Early
EDM1.2	10 days	40.13	0.4572	52.56	0.3634	Early
EDM111	1 year	46.85	0.1906	63.99	0.5959	Late
EDM130	10 days	39.72	0.4042	53.88	0.3736	Early
EDM3	1 year	39.1	0.2759	50.32	0.3407	Late
EDM51	7 days	41.13	0.1506	50.92	0.8549	Early
EDM10	1 year	42.45	0.6447	52.04	0.7133	Late
EDM6	11 days	38.18	0.2411	51.57	0.6825	Early
EDM21	2 years	41.11	0.2318	56.91	0.1064	Late*
EDM5	1 year	42.74	0.5692	52.34	0.4431	Late
EDM101	11 days	41.97	0.6982	57.15	0.1964	Early
EDM69	2 years	42.83	0.1181	57.58	1.319	Late*
EDM11	1 year	38.26	0.5393	45.57	0.4071	Late
EDM131	1 year	38.61	0.1038	51.71	1.342	Late
EDM16	7 days	39.17	0.1958	48.26	0.128	Early*
EDM110	1 year	40.66	0.5612	51	0.4244	Late
EDM116	1year	40.98	0.4935	50.97	0.3206	Late*
EDM106	4 days	38.12	0.3495	56.5	0.3043	Early
EDM103	4 days	40.91	0.1273	54.05	0.4431	Early
EDM108	1 year	41.28	0.4512	53.51	0.8802	Late*
EDM118	4 days	39.93	0.3581	50.94	0.6398	Early

Supplementary Table 1: Table of isolates examined in this study with growth rates and standard errors. Colonization categorization is also listed in the final column. Strains isolated before the first two weeks of are categorized as early colonizer strains and strains isolated after two weeks of age are categorized as late colonizers. The isolates marked with an asterisk in the Category column have a sister isolate from the same sample. Means of the two independent growth rate measurements were used for testing.

Media Preparations

	Oxiod anaerobic basal broth (rich media)				Glucose			Peptone (A-C)			Peptone (D-F)		
	100 %	90 %	50 %	10 %	90 %	50 %	10 %	90 %	50 %	10 %	200 %	90 %	20 %
Typical Formula*	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**	**gm/litre**
Peptone	16	14.4	8	1.6				14.4	8	1.6	32	14.4	3.2
Yeast extract	7	6.3	3.5	0.7									
Sodium chloride	5	4.5	2.5	0.5									
Starch	1	0.9	0.5	0.1									
Dextrose	1	0.9	0.5	0.1	0.9	0.5	0.1						
Sodium pyruvate	1	0.9	0.5	0.1									
Arginine	1	0.9	0.5	0.1									
Sodium succinate	0.5	0.45	0.25	0.05									
L-cysteine HCl	0.5	0.45	0.25	0.05									
Sodium bicarbonate	0.4	0.36	0.2	0.04									
Ferric pyrophosphate	0.5	0.45	0.25	0.05									
Haemin	0.005	0.0045	0.0025	0.0005									
Vitamin K	0.0005	0.00045	0.00025	0.00005									
Sodium thioglycollate	0.5	0.45	0.25	0.05									
Dithiothreitol	1	0.9	0.5	0.1									
7 g/L KH2PO4					7	7	7	7	7	7	7	7	7
2 g/L K2HPO4					2	2	2	2	2	2	2	2	2
1 g/L (NH4)2SO4					1	1	1	1	1	1	1	1	1
50 mg/L MgSO4					0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Supplementary Table 2: Table of media preparations used for the competitions. Formula show the complete recipe Oxoid anaerobic basal broth and the columns show the recipe for 200%, 90%, 50%, 20% and 10% concentrations. The competitions with glucose or peptone as sole carbon sources were used proportions relative to the complete Oxoid anaerobic basal broth. Peptone broths A-C used Bacto™ Peptone from BD Biosciences. Peptone broths D-F used Fluka peptone from Sigma-Aldrich.

	Contig nr.	Contig Length	depth	Best BLAST score
Strain 106	542	1527	433.9	E.coli plasmid pKL1, pO26-S1, pMG828-1, pSERB2 RepA and DfrA genes
	584	7968	162.4	E.coli SMS-3-5 plasmid pSMS35_8, pKY1 (from S.sonnei) cea, immunity protein (imm), kil genes, pColE1-EC39
	585	4078	137.3	E.coli SE11 pSE11-6 DNA, SMS-3-5 pSMS35_4, pIGWZ12, pMG828-2, O111:H- str. 11128 pO111_4
	563	1672	61.1	16s rRNA gene
	380	3245	58.9	23S rRNA gene
Strain 116	428	738	212.8	InsAB' transposase, IS1 protein InsA, IS1 protein InsB, plasmid associated
	863	4233	84.8	Associated with plasmids in Klebsiella, Salmonella, E.coli and Shigella, poor query coverage
	862	5512	80.7	E.coli pECO29, Salmonella enteritidis serovar Enteritidis pC
	444	2410	73.1	E.coli UM146 plasmid pUM146, E.coli strain CFT073 pathogenicity island II
	412	2454	59.1	protein
	795	1472	58.3	16S rRNA gene
	464	3238	52.5	23S rRNA gene, 5S rRNA gene
Strain 530	75	1286	145.1	transposase IS116/IS110/IS902 family protein
	93	3103	126.2	23S rRNA gene
	163	1683	117	16S rRNA gene
	42	767	68.9	IS1 insertion element
	81	1010	67.9	transposase IS3/IS911 family protein
	74	472	65.2	rshA, possible capsular biosynthesis

Supplementary Table 3: Listing of high depth contigs from the genome sequencing.

Supplementary Methods:

Batch culture competitions

All competition experiments were carried out in 100ml flasks containing 90ml of medium and a water trap to allow gas to escape but kept the flasks anaerobic after the oxygen in the head space was consumed. The only exception to this protocol was the experiment shown in Supplementary Figure 11, where we attempted to give *E.coli* more of a competitive advantage relative to the background flora. In this experiment the water traps were emptied in order to allow small amounts of oxygen to enter the flasks. Flasks were incubated at 125rpm and 37ºC. Oxoid anaerobic basal broth was either prepared fresh or boiled in a water bath for at least ten minutes to de-gas the media. A typical experiment (2 replicates) used 90 ml of media per replicate for each time point. Mostly, 500ml batches of media were prepared but sometimes as much as 1 liter was made. This required many separate batches of media to be used throughout most of the individual experiments and also across the entire study. At the beginning of each of the competition experiments, initial inocula were diluted into each of the replicate flasks. Thereafter, 100µl of the cultures were transferred to flasks containing fresh medium. This procedure established two independent but parallel series of competition time points. The competitions of strain EDM106 and strain EDM530 presented in figures 2A, 2B, 3, 4, S9, and S12, are continued from revived frozen stocks from day 10 of the two day strain EDM106 and strain EDM530 competition presented in figure S6. Aliquots were revived and continued for two days in complete Oxoid anaerobic basal broth before individual competition treatments were performed. Long term stationary phase cultures used the same conditions as described above, except that no transfers to fresh medium was carried out and medium was not replenished after inoculation. The peptone media experiments were conducted using different concentrations of peptone and the individual experiments were performed using two different brands of peptone media. Experiments shown in Figure 4A-C were conducted using Bacto™ Peptone manufactured by BD Biosciences. Experiments shown in Figure 4D-F were conducted using Fluka peptone from Sigma-Aldrich. Both types of peptone are enzymatic digests of animal protein.

Measurement of medium carrying capacity

Strains were grown to saturation (16hour culture) in Oxoid anaerobic basal broth. 90ml of culture was separated into three 50ml conical vials with perforated tops of 30ml each and freeze dried. After freeze drying, all tubes were allowed to equilibrate to atmospheric moisture levels for 1week before weighing. The t.test that is presented is of strain EDM106 against the EDM106/EDM116 mixed culture.

Mixed effects modeling

The mixed effects model was fitted using the nlme package for R. Logit-transformed proportions were modeled as a function of the interaction of time and treatment (peptone concentration) assuming non-random intercepts and fitting a first-order continuous autoregressive process in the errors. The test was done for unequal slopes of strain proportions between the low (1.6 g/L) and high (16 g/L) peptone competitions.

Characterization of shared and unique gene content

Due to the large number of contigs, determination of gene presence included additional processing steps *(manuscript in preparation).* Briefly, in order to recover genes split into separate contigs or genes that did not receive an annotation from RAST due to sequencing errors, we carried out an additional blast search using annotations of sixteen *E.coli* genomes. If any annotated gene was found in a subset of the genome collection, this was re-BLASTed against the raw assemblies of the strains not included in that subset. Re-assignment of the recovered sequence to the annotation required 90% identity and an e-value of $<1e^{-25}$. Genes with redundant functional RAST assignments were collapsed into one single assignment prior to further analysis.

III

Comparisons of infant *Escherichia coli* isolates link

genomic profiles with adaptation to the ecological niche

Eric J. de Muinck[1,2,3], Karin Lagesen[1] , Jan Egil Afset[4,5], Xavier Didelot[6], Kjersti S. Rønningen[7], Knut Rudi[8], Nils Chr. Stenseth[1], Pål Trosvik[1,*]

* Corresponding author: Pål Trosvik pal.trosvik@bio.uio.no

1 Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Po Box 1066, 0316 Oslo Norway

2 Division of Epidemiology, Norwegian Institute of Public Health, PO Box 4404, 0456 Oslo, Norway

3 NOFIMA - The Norwegian Institute of Food, Fisheries and Aquaculture Research, PO Box 210, 1430 Ås, Norway

4 Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, PO Box 8905, 7491 Trondheim, Norway

5 Department of Medical Microbiology , St Olavs Hospital, PO Box 3250, 7006 Trondheim, Norway

6 Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK

7 Department of Pediatric Research, Oslo University Hospital, Rikshospitalet, PO Box 4905, 0424 Oslo, Norway

8 Department of Chemistry, Biotechnology and Food Science, University of Life Sciences, PO Box 5003, 1432 Ås, Norway

E-mail addresses:

Eric J. de Muinck (ericdemuinck@gmail.com)[1,2,3]

Karin Lagesen (karin.lagesen@bio.uio.no)[1]

Jan Egil Afset (jan.afset@ntnu.no)[4,5]

Xavier Didelot (x.didelot@imperial.ac.uk)[6]

Kjersti S. Rønningen (kjersti.skjold.ronningen@gmail.com)[7]

Knut Rudi (knut.rudi@umb.no)[8]

Nils Chr. Stenseth (n.c.stenseth@bio.uio.no)[1]

Pål Trosvik (pal.trosvik@bio.uio.no)[1]

Abstract

Background: Despite being one of the most intensely studied model organisms, many questions still remain about the evolutionary biology and ecology of *Escherichia coli*. An important step toward achieving a more complete understanding of *E.coli* biology entails elucidating relationships between gene content and adaptation to the ecological niche.

Results: Here, we present genome comparisons of 16 *E.coli* strains that represent commensals and pathogens isolated from infants during a specific time period in Trondheim, Norway. Using differential gene content, we characterized enrichment profiles of the collection of strains relating to phylogeny, early vs. late colonization, pathogenicity and growth rate. We found clear gene content distinction relating the various grouping criteria. We also found that categories of strains use different genetic elements for similar biological processes. The sequenced genomes included two pairs of strains where each pair was isolated from the same infant at different time points. One pair, in which the strains were isolated four months apart, showed maintenance of an early colonizer genome profile but also gene content and codon usage changes toward the late colonizer profile.

Conclusions: Our results indicate a general pattern where alternative genetic pathways lead toward a consistent ecological role for *E.coli* as a species. Within this framework however, we saw selection shaping the coding repertoire of *E.coli* strains toward distinct ecotypes with different phenotypic properties.

Keywords

Escherichia coli, comparative genomics, infant gut, commensal, pathogen, generation time, codon usage bias

Background

Awareness of the importance of the gut microbial colonization for human health is growing as numerous links with a multitude of diseases are being discovered [1]. Recent advances in sequencing technology have generated massive amounts of data but much remains to be understood about the processes important for maintaining a healthy community structure. *E.coli*, as well as being a much studied model organism, is an important and ubiquitous member of the human gut microbial community. Although *E.coli* constitutes only a small fraction of the total gastrointestinal microbiota, it has a wide spectrum of potential interactions with the human host, ranging from probiotic to commensal and on to pathogenic [2].

As one of the most intensely studied organisms, much genomic information on this species has already been collected. Genbank has cataloged 60 complete chromosomal genomes and 346 draft genomes (at the time of writing). However, most of the sequencing effort has been directed toward pathogenic *E.coli* strains. Previous comparative analysis of the genome sequences of 61 isolates has helped develop the new view of the *E.coli* genetic landscape which highlights diversity at the genome level [3]. A typical *E.coli* strain carries between 4,000 and 5,500 genes. On average, an *E.coli* strain will share about 40% of these with all other members of the species, while the remainder forms part of the pan-genome [4,5]. Following these approaches, differential gene content between strains is thought to subdivide *E.coli* into ecological classes that may be more biologically informative than traditional phylogenetic categorization based for example on Multi-Locus Sequence Typing (MLST). Use of full genomes and subsequent gene-content profiling has thus become important for understanding the role of genome contents for defining a realized ecological niche [6].

This work is the continuation of a deep characterization of *E.coli* strains isolated from a cohort of infants and their mothers in Trondheim, Norway. The original study design was a nested case-control format created to examine the impact of whole gut microbial colonization on the development of atopic disease [7,8]. In this original characterization of the cohort, qPCR was used to identify and quantify the microbial fecal composition of several classes of bacteria and these data were matched with cytokine profile development. From this, it was observed that early *E.coli* colonization was linked to protection from atopy and the mother was found to be a likely source of the infant colonization [9]. We have previously characterized the *E.coli* colonization pattern of a sub-cohort of this larger study, 85 infants and their mothers, and found limits on the diversity of strains and further evidence of transmission from the mothers to the infants [10]. Deeper characterization placed these strains into a phylogenetic context of the larger *E.coli* diversity.

Here, we built upon these earlier observations using whole genome sequencing. We compare the genomic content of strains with different phylogenetic, pathogenic vs. commensal, growth rate and early vs. late colonization characteristics in order to determine enrichment profiles that may explain these ecological traits. The signatures that were observed can be used for further investigations into genotype-phenotype mapping within the context of ecological adaptation and for investigating the role of the many hypothetical proteins that we found differentiating the groups. The collection of strains that were used for this analysis offer insight into a temporally and geographically coherent population of gut colonizing *E.coli*, with additional context afforded by our previous characterizations of these strains [10]. Methodological challenges that were addressed included developing a strategy for compensating for incomplete assembly, small sequencing errors, and potential loss of genetic information derived from genomes sequenced by 454 single-end shotgun sequencing. Dealing with incompletely assembled draft genomes, as we have done, may become less problematic for single isolate analysis as assembly algorithms and sequencing technologies progress. However, costs may

hinder coverage for large collections of isolates and also for complex samples such as the soil or mammalian gut, which at minimum contains several hundred genomes [11,12].

Results

Methodological challenges

454 sequencing of the 16 genomes yielded coverage levels, the median number of times that a specific genomic position is included in a sequencing read, ranging from 7.55 to 20.1 (Table 1). Good coverage is crucial to aid assembly of the reads into as few contigs as possible. This was indicated in our data by the significant decrease of contig numbers as the median coverage depth increased (R^2=0.69, p<0.0001, Figure S1A in Additional file 1). However, there was a plateau at about 13x coverage, above which that trend subsided. We also found a strong positive correlation between the number of RAST annotated genes per base in a genome and the number of contigs in the assembly (R^2=0.52, p=0.0017) (Figure S1B in Additional file 1). Furthermore, there was an even stronger trend for mean annotated gene length to decline as assemblies became more fragmented (R^2=0.94, p<0.000001, Figure S1C in Additional file 1). This indicated that partial gene sequences were more often retrieved from the more fragmented assemblies. The main cause of this phenomenon was genes being split onto two different contigs due to reduced coverage at contig edges in low read depth assemblies (Figure S1D in Additional file 1), or small sequencing errors, usually in homopolymer tracts (a known shortcoming of pyrosequencing), producing spurious frame-shifts in the coding sequence. Both of these causes can result in coding sequences being un-annotated by RAST. To circumvent this issue, we applied an additional gene recovery step which resulted in a positive relationship between the number of BLAST hits retrieved and the number of contigs in an assembly (R^2=0.57, p=0.0008, Figure S1E in Additional file 1), with a total of 8,322 genes being recovered from all the strains combined. Following this curation of the genome annotations, we re-examined the bias in the relationship between the number of annotated genes per base and the

number of contigs and found that the pre-treatment bias had been completely removed (R^2=0.0001, p=0.97, Figure S1F in Additional file 1). This information, and the continued strong correlation (R^2=0.92) between the number of gene families and genome size (Figure S1G in Additional file 1) suggest that the updated annotations were correct. One outlier (JEA297p) showed a different gene density than the other strains, and will be discussed below.

Phylogenetic and gene content comparisons

Comparative analysis of the genomes content revealed that 52.4% of the genes are shared by our 16 genomes (Figure 1). However, inclusion of strain MG1655 (K12) reduced the proportion of shared genes to 50.2%. This is higher than results reported by other studies [3,5] and could be attributed to the localized sampling of our *E.coli* population. The pan-genome of our 16 strains includes 6,152 gene families, which increases only to 6,181 when K12 is included (Figure 2). The structure of the dendrograms generated from the genome collection were comparable whether we used homology in the core genome or gene content differences to determine their relationships, with the deepest subdivision being between the clades denoted 1 and 2 (Figures 3A and 3B). *E.coli* that colonize humans are generally grouped into the four phylogroups; A,B1,B2 and D [2]. Here clade 2 contained the strains previously categorized into the B2 phylogroup whereas members of clade 1 belonged to phylogroups A, B1 and D. The next subdivision in the ClonalFrame tree (Figure 3A) was to separate JEA297p from the rest of clade 1, whereas the gene-content tree (Figure 3B) further divided the phylogeny into four subclades. One of these four subclades contains three of the four pathogens (JEA124p, JEA179p and JEA297p) despite the complete phylogenetic unrelatedness of these strains (Figure 3A). This clade also included a commensal strain (EDM116c) which was not related to the pathogens according to the ClonalFrame core phylogeny, but contained a pathogenicity island and some of the genetic profile of a pathogen (Figure 6).

Gene content enrichment analysis using the split between clades 1 and 2 to define the groupings (criteria I; Table 2) found 305 genes (Additional files files 2 and 3) differentiating the clades with several of the gene sets falling into the same biological process categories (Figure 4, Figure 10). A relatively even distribution of genes (~150) were associated with each clade, and the level of enrichment of the clades was significantly higher than expected by chance ($p<0.0001$; permutation test; Figure S2 in Additional file 1). Both clades are enriched for different cell adhesion proteins while clade 1 is differentially enriched for several additional iron acquisition proteins including an additional hemoglobin receptor, hemin transport protein, and yersiniobactin siderophore system. Clade 2 is differentially enriched for small molecule usage including an alternative pathway for obtaining nitrogen from cyanate, aromatic compound decomposition and resistance to potential toxins such as arsenic.

Pathogen and commensal comparisons

Identifying the genetic elements that differentiate commensal and pathogenic strains is extremely important. Multiple correspondence analysis (see methods) of the sequenced genomes highlighted differences between two groups of strains separated on the first axes, matched the previously described cladistic structure (Figure 5), and showed a clustering of some of the pathogenic strains using an overall similarity profile despite the categorical phylogenetic differences between the strains. We began by exploring the most strict criteria of group differentiation; where a gene would have to be in 100% of the pathogenic strains and not found in the commensal group and vice versa. Surprisingly, this approach only identified a few chaperone genes. Relaxing the criteria (criteria II, Table 2) still yielded only 33 genes enriched in the commensal group but 164 in the pathogenic enrichment (Figure 6, Additional files 4 and 5). The probabilities of the commensal and pathogenic enrichments to happen by chance were equal to $p=0.18$ and $p=0.02$ respectively (permutation test;

Figure S3 in Additional file 1). Most of the commensal group gene enrichments were either related to fatty acid metabolism or sugar utilization pathways.

A noticeable contributor to the pathogen enrichment was the pathogenicity island carrying the type III secretion system (T3SS) and several effecter molecules associated with it. BLASTing the large contig sequences generated from the Newbler assemblies against a complete enterocyte effacement pathogenicity island (LEE) (35,624bp) [13] revealed significant identity for many of the strains (Table S1 in Additional file 1). All four pathogenic strains, two commensal isolates from the EPEC study, and one commensal isolate from the IMPACT study contained the pathogenicity island [10,14]. Categories that were enriched in the pathogenic grouping relative to the commensal grouping included both nitrogen and primary metabolic processing (Figure 10). Not surprisingly, since the pathogenic strains were initially selected based on the presence of the intimin *eae* gene, and therefore belonged to pathotype enteropathogenic *E.coli* (Table 1), intimin enrichment was also observed in the pathogenic grouping.

Growth rate comparisons

Plotting relationships between the anaerobic generation times and the ratio of anaerobic to aerobic generation times of isolates showed a strong positive correlation (Figure 7). Highlighting strains from which we have genome sequences showed three clusters that we then defined as fast, medium and slow growers. Evaluation of growth rate clustering, using a relatively complicated enrichment test due to the differences in numbers of strains in each of the groups (criteria III, Table 2), saw strong gene content distinction between the fast group (group of two) and slow group (group of four) with p-values of the gene content profile amounts equal to 0.09 and 0.04 respectively (Figure S4 in Additional file 1). The strains with

the medium growth rate (group of four) did not have a significant (p=0.56) number of distinguishing genes . Enrichment profiles of the fast, medium and slow groups showed overrepresentation of 227, 47 and 324 gene families, respectively (Figure 8, Additional files 6-8). Relative GO category enrichment in the slow group included primary metabolic process, nitrogen metabolism and macromolecular processes plus several genes important for iron uptake and utilization (Figure 10). The fast growing group was uniquely enriched for several GO categories including response to chemical stimuli and cell wall organization. The medium growth rate group seemed split between the slow and fast growers but the majority of genes enriched in this group were phage related.

Early vs. late colonizer comparisons

In this collection of strains we categorized a strain as an early colonizer if the strain was isolated from an infant within the first two weeks of life. These strains have a higher likelihood of coming from the mother than isolates from later age categories. Late colonizer strains were isolated from infants aged four months to two years. Comparison between the early colonizer and late colonizer strains found 416 genes that were differentially enriched between the two groups (criteria IV, Table 2, Figure 9, Additional files 9 and 10). One of the late colonizer strains was an early colonizer strain that had remained in the infant for four months (see below), this strain was considered an early colonizer and the enrichment criteria were modified to minimize bias. The 6 early colonizers and the 6 late colonizers had gene content profiles with p-values equal to 0.02 and 0.05 respectively (Figure S5 in Additional file 1). Early colonizers were distinguished by 238 genes including capsular genes, fimbrial genes, yersiniabactin and other iron uptake systems as was seen in the cladistic enrichment. Additionally, we found enrichment for type four pili, required for localized adherence and auto-aggregation phenotypes [15]. Late colonizer strains were distinguished by 178 genes

including GO category biological enrichment for oxidation reduction processes and response to chemical stimuli (Figure 10).

Codon usage bias and generation times

Codon usage bias in highly expressed genes has been found to be a strong predictor of maximal growth rate in prokaryotes [16]. In order to investigate this relationship in our data we looked at correlations between our effective number of codons (ENC) estimates and growth rates under aerobic and anaerobic conditions. Mean genome wide ENC for the 10 EDM strains was 49.044±0.182, while mean ENC for ribosomal protein genes was 35.790±0.052. Mean generation times were 40.3±1.2 min. and 52.7±3.0 min. under aerobic and anaerobic conditions, respectively. We first looked at the relationship between whole genome ENC and growth rate. We found a positive correlation (two-sided Spearman correlation ρ=0.71, p-value=0.03) with anaerobic growth rate (Figure S6A in Additional file 1), but no significant relationship with aerobic growth rate (ρ=0.05, p-value=0.89). This result indicated that faster growing isolates tend to have more pronounced overall codon bias. As expected the within species variation in ENC for ribosomal protein genes was minimal and no relationship could be found between this index and growth rates. Also due to a lack of variation in ribosomal protein ENC the relationships between ΔENC and growth rates were essentially the same as the genome wide correlations (ρ=0.72, p-value=0.02 and ρ=0.26, p-value=0.47 for anaerobic and aerobic generation times, respectively) (Figure S6B in Additional file 1).

Strain evolution in the infant gut

Two pairs of strains from two infants (child 1891 and 1360, Table 1) were isolated at two different time points and had matching MLST profiles. The strains were isolated at four and eleven days of age

(EDM49c and EDM101c) and at ten days and four months (EDM1c and EDM123c) of age respectively. The isolates from the same child had almost identical genome contents (Figure 2) and were subjected to closer scrutiny in order to shed light upon the selective pressures in a novel infant gut environment. The earlier isolate in each pair was thus defined as "the parent strain" and the later as "the evolved strain".

Both genome content and codon usage indicated strain evolution in the infant gut from an early colonizer to late colonizer phenotype. From the EDM49c and EDM101c, only three genes, possibly phage related, were found in the parent strain but not in the evolved strain, and no genes were unique to the evolved strain relative to the parent strain. From EDM1c and EDM123c, 16 genes were found in the parent strain that were not in the evolved strain and 13 genes were found in the evolved strain not in the parent strain (Additional files 11 and 12). Interestingly, three of the genes unique to the parent strain were also called in the early colonization enrichment list whereas none of the genes unique to the parent strain were found in the list of genes from the late colonization enrichment. The three genes that were matched to the genes in the early enrichment list were GO categorized as a type-f conjugative transfer system pilin chaperone, hypothetical protein c4302 [uropathogenic, *E.coli* CFT073, NC_004431.1] and a tellurite resistance protein with transposon elements encoded nearby. Other genes unique to the evolved strain relative to the parent strain included a mercury resistance operon that has evidence of being carried on the transposon Tn21. Genome wide ENC and ΔENC comparison of EDM1c and EDM123c found reduced codon bias in the evolved strain (Figure S6A and B in Additional file 1).

Discussion

Genome analysis methods

The methodological challenges we addressed in order to generate the genotype-phenotype profiles presented in this work require some discussion. The 454 pyrosequencing single-end shotgun data presented difficulties that would, in several cases, not have been ameliorated by increasing the sequencing coverage (Figure S1D in Additional file 1). This is partly due to the intrinsic variability of *E.coli* genomic content, which made it impossible to rely on reference-based assembly and necessitated the use of *de novo* assembly methods, but also because of the relatively error prone nature of the technology. Alternative sequencing technologies or laborious and costly paired-end/mate-pair DNA sample preparation would have been required to reduce the number of contigs. However, the single-end shotgun approach offers a number of advantages due to its simplicity and lower cost compared with paired-end library preparation [17]. Furthermore, even though improvements in sequencing technologies will help genome assembly of bacterial isolates due to increased read length, sequencing of complex mixtures of bacteria such as gut or soil communities will continue to face some of the same challenges that we have addressed. The additional post-annotation search step employed in this study appears to have alleviated some of the biases introduced by an imperfect assembly (Figure S1 in Additional file 1).

Pathogens vs. commensals

The factors that distinguish a pathogenic from a commensal *E.coli* remain contentious. Previous studies have failed to come up with pathovar specific genomic cores for strains classified as enteropathogenic or enterotoxigenic *E.coli* (EPEC and ETEC, respectively) [18,19], but there have been studies reporting specific gene content profiles in extraintestinal pathogenic *E.coli* (ExPEC) [20,21]. However, recent work indicates that many of these genes are primarily associated with gut colonization and that virulence is an incidental by-product of commensalism [22,23]. In our case, using strict 100% presence/absence as an enrichment criterion failed to detect genes that separated commensals and pathogens (all four pathogenic strains were EPEC).

Relaxing the criteria resulted in a significant set of 164 genes that were preferentially found in the pathogenic group, but there was substantial gene overlap with commensal strains. The 33 genes enriched in the commensal group may represent a small part of the wide variety of genes necessary to be a successful colonizer. The weak commensal signature, compared with the pathogenic one, suggests that the term commensal may not be a meaningful descriptor in a phenotypic or evolutionary context as our analyses identified 'pathogen-like' commensals (e.g. commensal isolate EDM16c is closer to the pathogenic isolates when it comes to functional genetic profile than it is to the other commensals (Figures 3 & 6)) which may suggest a virulence potential of certain commensal strains. This is especially highlighted by the large pathogenicity island carrying the TTSS which was shared by all the pathogenic strains and a subset of the commensals (Table S1). Recent work has shown that this system is important for bacterial competition in the gut in addition to its role in host interactions [24]. If virulence is indeed an accidental by-product of adaptation to the gut environment it would explain why it is hard to find a non-clinical distinction between pathogenic and commensal strains, as virulence may rather be a matter of context and opportunism [25]. Thus, genomic signatures may nevertheless identify strains that have greater capacity to make the transition from commensalism to virulence, and could thus aid in designing preventive strategies.

Minimal generation time

Growth rate is a phenotype with quintessentially complex genetic underpinnings, and can hardly be ascribed to specific genes or alleles. Insight into the mechanisms underlying growth rate differences is highly desirable as it is related to other phenotypes of fundamental importance, such as virulence [26]. Minimal generation time in a study comparing 214 bacterial and archaeal species was found to correlate with genomic features such rRNA and tRNA copy number and codon usage bias [16].

However, minimal generation times were found to vary considerably within the *E.coli* isolates in our collection, even though these particular features were similar among our isolates.

We could not find any significant correlation between generation time and rRNA and tRNA copy number (results not shown), and codon usage bias was also found to be a poor predictor of aerobic generation time. Surprisingly, it correlated strongly with anaerobic generation time.

In contrast to the study by Vieira-Silva, we found a positive correlation between generation time and codon usage bias in highly expressed genes (ΔENC). This result is not necessarily in conflict with previous findings, as it may be explained by the fact that we were looking at strain level rather than species level relationships. Specifically, in contrast to the previous work covering many diverse species, the ribosomal protein genes were extremely conserved and the spread of ENC values for this set of sequences was less than a third of what was observed for genome wide ENC. Whole genome bias dominated our analysis and gave rise to the interpretation that a narrower general codon usage profile is associated with shorter anaerobic generation times. It is noteworthy that this relationship did not hold for aerobic growth. At face value it may seem paradoxical that codon usage specialization should be more important under anaerobic conditions when translation efficiency is presumably less of a limiting factor than under intrinsically faster aerobic growth. One explanation for this could be that gut adapted *E.coli* are primarily selected for anaerobic growth properties as the gut community matures and that aerobic growth leaves comparatively little systemic imprint on their genomes. Even though we found a significant correlation between aerobic and anaerobic generation time (R^2=0.41, p<0.001), we found and even stronger correlation between anaerobic generation time and anaerobic to aerobic generation time ratio (R^2=0.51, p<0.0001), suggesting that slow anaerobic growth entails disproportionally fast aerobic growth, and that the genomic bases for these two modes of growth might, at least in part, be uncoupled. This interpretation is supported by the fact that codon usage bias correlated with anaerobic but not aerobic growth rates. It would be interesting

to compare these results with environmentally adapted *E.coli* isolates [6,27] and discern if the genomic imprint of aerobic growth might be more visible.

Gene content analysis represents an entirely different approach to investigating the genomic basis of differential generation time, and one that would not be appropriate for inter-species comparisons. The fact that our *E.coli* isolates are closely related, as witnessed by the shared genomic core, yet display relatively high variation in generation time begs the question of whether there are signatures of coding potential that relate to this phenotypic diversity. To our knowledge, the results presented here are the first attempt at correlating growth rate phenotype with differential gene content. Even though the small sample sizes warrant some caution in interpreting the results, the gene profiles of the fast and slow growing groups are quite unlikely to have arisen by chance. It is also noteworthy, albeit perhaps not surprising, that the intermediate group failed to produce a significant enrichment profile and that differences are only visible when comparing the extremes.

Relative enrichment in the slow group (324 genes) compared to the fast growing group (227 genes) found that many of the same GO categories were enriched but the slow growing group had a greater enrichment in several metabolic processes, including nitrogen, macromolecular, and several genes important for iron uptake and utilization (Figure 10). In contrast, the fast growers had a larger relative enrichment for genes involved in response to chemical stimuli and cell wall organization. Perhaps, this represents an ability to quickly adapt to changes in the environment. The fact that we observed relatively clear gene content signatures in both the fast and slow groups may reflect an evolutionary trade-off between short minimal generation time and scavenging potential. Copiotrophic, fast growing bacteria tend to have low affinity transporters typically representing an adaptation towards “feast” conditions, resulting in reduced competitiveness during nutrient starvation [28]. Slow growers, on the other hand, tend to have high affinity transporters, making

them competitive in low nutrient environments, while at the same time making them susceptible to saturation or toxic effects when resources are plentiful [29]. This interpretation is further supported by the enhanced presence of scavenging-associated genes in our slow growing isolates.

Early and late colonization

The infant gut environment is temporally dynamic in terms of reduction potential, nutrient availability, immune function and the structure of the resident microbial community [30,31]. The infant gut microbiome has been found to undergo a smooth increase in phylogenetic diversity over the first few years, while broad scale taxonomic patterns are characterized by abrupt events, eventually conforming to a mature profile [32]. The same study found concomitant changes in metagenomic content indicating that the community as a whole is responding to a changing environment. Selection pressures faced by members of the gut microbiota may therefore differ widely between the earlier and later stages of infancy. This pressure is reflected in the reduced relative abundance of *E.coli* in the mature microbiota relative to the infant gut community [31,33] and suggests that strains present at different stages of development could differ widely in their characteristics. Dramatic changes in the gut microbiota of pregnant women have also been shown from the first to third trimester, resulting in increased abundances of Proteobacteria and Actinobacteria and reduced taxonomic richness [34]; a community state more reminiscent of the infant gut structure. The mother may somehow prime the gut microbiota with a qualitatively different environment in preparation for transfer to the infant.

Both early and late colonizers had significant differential gene content profiles (178 and 238 gene families respectively). We found that early colonizers were enriched for type IV secretion system and fimbrial genes that are important for attachment and interaction with the host. This group also had

an increased presence of colicin resistance genes, which may reflect the importance of competition with bacteria of the same or closely related species is in the low diversity conditions of the early gut environment. Furthermore, we found an increased number of genes involved in biosynthetic processes in the early colonizer group. This could also be an adaptation to low diversity conditions where production of secondary metabolites and secreted growth factors is potentially limited. The late colonizers were enriched for resistance to toxins such as arsenate and cyanate. This could indicate the importance of these pathways for survival in the complex ecosystem of the mature gut.

Evolution towards a late colonizer genomic profile

There is ample evidence that, given some selective regime, microbial evolution in the laboratory can be exceedingly rapid [35]. A few studies have documented the evolution of pathogenic bacteria in infected individuals [36,37] but reports of real-time evolution in natural environments remain scarce, and to our knowledge there are no such studies focusing on bacteria of the human gut. Isolate EDM123c was categorized as a late colonizer due to the fact that it was isolated from an infant at four months of age. EDM123c is by all probability clonally descended from EDM1c which had colonized that same infant already at 10 days after birth. Since this strain had spent nearly four months in the infant gut during an environmental transition period, we hypothesized that selection would push it toward a late colonizer genomic profile. There are two lines of evidence to suggest that this is the case. First, three of the genes that were present in the ancestral strain but lost from the evolved version matched genes in the early enrichment list. This list included a tellurite resistance protein which has been linked to resisting host defense [38,39]. Further experimentation is necessary to fully characterize the effect of these particular genes on early colonizing ability and possible reasons for negative selection in a more mature microbiota. Secondly, we observed an increased anaerobic generation time from isolate EDM1c (52.6±0.4 min.) to EDM123c (55.8±1.1 min.). Interestingly, EDM123c also had and elevated genome wide ENC (and thus also ΔENC) (Figure S6 in

Additional file 1) relative to the parent strain. This indicates that from the parent to the evolved strain there has been selection for synonymous mutations pushing the strain toward reduced codon usage bias. Reduced codon bias and growth rate have previously been associated with late gut colonization [16], indicating that isolate EDM123c has in fact evolved toward a late colonizer profile.

Given the close relatedness between EDM1c and EDM123C, as witnessed by both sequence similarity and gene content (Figure 3), there can be little doubt that these isolates are clonally related, and genomic differences are probably due to evolution taking place in the gut. Indeed the other pair of parental (EDM49c) and evolved (EDM101c) strains displayed practically no divergence in gene content or codon usage bias, probably due to the fact that they were isolated only 7 days apart. We cannot discount the possibility that clonally related strains were introduced, outcompeted and then re-introduced at a later time. In this case at least part of any evolution taken place would have done so in a different environment. In the case of EDM123c, however, we feel that this is an unlikely scenario since adaptation took the direction predicted if the isolate had evolved in a maturing infant gut.

Cross category enrichment comparisons

Even though the different enrichment comparisons were fruitful for understanding functional categories, using this information across the different comparisons gave a better and more nuanced view. The main clade comparisons are very informative as they link a strain's evolutionary history to a measure of functional differentiation which can help define its ecological niche. For example, all early colonizers except EDM16c (which had an atypical gene content profile for an early colonizer) belong to clade2. The late colonizers all belong to clade1 except EDM123c, which is the evolved EDM1c and thus an atypical late colonizer. Thus there appears to be a phylogenetic split defining these ecological

categories, and this split is reinforced by disparate gene content. Also, three of four pathogens group to clade1 which is comprised solely of B2 isolates, a group known to be pathogen rich [2]. Furthermore, and in contrast with the core genome phylogeny, the pan-genome phylogeny places the commensal strain EDM116c within the same subclade as these three pathogens (Figure 3). One could speculate that although EDM116c is an ostensibly asymptomatic isolate, its genetic makeup is such that given the right circumstances it may cause symptoms similar to known EPEC strains. The pathogenic isolate JEA242p, on the other hand, is placed within the otherwise exclusive commensal clade 2, demonstrating that virulence can emerge from quite different genomic backgrounds.

The two isolates classified as fast growing in this sample set of genome sequenced strains were both late colonizers (thus belonging to clade1) while early colonizers in this set tended to be slow growing (within clade2), but with disproportionally short aerobic relative to anaerobic generation times (Figure 7). This trend is not consistent with a previous study [16], but the disagreement is most likely attributable to sampling bias. Nevertheless some interesting associations emerged when making cross-grouping comparisons. Comparing the similarities in the gene content enrichments between all groupings found that the combined clade1-late-fast and clade2-early-slow designations shared the most (57 and 49 respectively; Additional files 13 and 14) (Figure S7 in Additional file 1). Unique phosphotransferase system (PTS) were enriched in each cross category grouping which are thought to enhance sugar utilization in general and possible bacterial uptake of sugars from breast milk [40]. A similar general differential gene content profile was seen between the same combined groups in glycosyl transferases and glycosyl hydrolase genes which are important for obtaining nutrients from the host and correct “assembly of a microbiota” [41]. The combined clade2-early-slow group further encoded arylsulfate sulfotransferase, which has been claimed to play a role in the detoxification of phenolic compounds [42]. On the other hand, a gamma aminobutyrate utilization gene was enriched in the combined clade1-late-fast group. This polyamine utilization gene has roles in proliferation

under stressful conditions and utilization of alternative sources of carbon and nitrogen, which could be an adaptation to the difficult conditions of a mature gut microbiota [43,44]. Lastly, the clade1-late-fast group showed enrichment for the hydrogenase-4 operon, which is important in anaerobic growth [45]. These cross-category comparisons provide a tentative link between the evolutionary history and functional phenotypes of our isolates where the two main branches of the core and pan-genome phylogenies may represent adaptive paths leading toward distinctive ecological properties.

Conclusions

This study addresses the role of gene repertoire in bacterial niche ecology, including the genomic bases of phenotypes that are not directly linked with pathogenicity. This aspect of *E.coli* ecology has not been thoroughly explored, but may shed light on the evolutionary history of the species [6]. The relatively small sample size and need for further molecular work precludes definitive conclusions regarding relationships between the array of genetic pathways and specific phenotypes. However, our results indicate a general pattern where alternative genetic pathways lead toward a consistent ecological role for *E.coli* as a species (Figure 10). Within this framework however, we saw selection shaping the coding repertoire of *E.coli* strains toward distinct ecotypes with different phenotypic properties. Additionally, the profiles we present should lead to further investigation and may lend insight into the biological roles of genes whose previously assigned biological function is incomplete and also for the large number of hypothetical proteins that were outlined using this method.

In contrast to previous studies of *E.coli* eco-genomics [3,5,18,46] our isolates come from a population that is narrowly localized both temporally and geographically. This could entail reduced genetic diversity due to shared ancestry and increased exchange of genes through horizontal transfer (HGT) between strains. Although the present study was not in particular concerned with HGT we did see a

substantially higher percentage of shared gene content (52.4%) than what has previously been reported, as well as a smaller pan-genome, indicating that homogenizing forces are increasingly effecting genomic diversity on a local scale. Nevertheless there were several instances where relatively clear gene enrichment profiles could be linked to specific phenotypes and ecological characters. Due to the disparate nature of *E.coli* genomes identification of such gene suites might be impeded if similar phenotypes can arise through different mechanisms and evolutionary histories, as is the case with clinical phenotypes of many pathogenic *E.coli* [5]. A more homogenous genomic background, as seen in this work, could make it easier to tease out gene content signatures that are ecologically relevant.

Material and Methods

Strains and culture conditions

The bacterial strains used in this study have been previously described in [10,47] and (de Muinck et al., manuscript submitted) (Table 1). Six strains were selected for genome sequencing from [47] because they were *eae*-positive and represented the previously reported diversity of phylogenetic groups. Two of these strains were from healthy children while four were isolated from children with diarrhea and these isolates were further classified as enteropathogenic *E.coli* (EPEC). A further ten strains were selected from [10], all of which were isolated from healthy children. All strains were grown to saturation in LB media and DNA extraction was performed using the DNeasy kit from Qiagen.

Genome sequencing and annotation

DNA was single end shotgun sequenced using Roche 454 GS (FLX Titanium) pyrosequencing. Sequences have been deposited in the EMBL-EBI Sequence Read Archive. Accession numbers are listed in Table 1. *De novo* assembly was performed using Roche's program Newbler v2.3 (performed at the freely available Bioportal computing service, http://www.bioportal.uio.no). Annotation was done using RAST version 4.0 [48]. The RAST annotated genes of each of the genomes were BLASTed [49] against all the other annotated genomes using criteria of 85% identity and an e-value of less than 1x10^-25 to signify a gene match. Due to the large number of contigs, determination of gene presence included additional processing steps to recover genes split into separate contigs or genes that were not included in the annotation. Briefly, we used the complete set of annotated genes from all of the genomes as a reference pool. If a gene in the reference pool was not found in all of the analyzed genomes, the longest copy of the gene was re-BLASTed against the Newbler assemblies of each of the genomes in which the gene was initially not found. This gene was then added to the annotation of a genome if a partial hit was found that was at least 90% identical and an e-value of less than 1x10^-25. Genes were grouped as a family if they matched with the BLAST criteria just mentioned, or if they received identical functional annotations from RAST.

Core genome phylogeny and pan genome tree

A multiple alignment of the *de novo* genome assemblies was performed using progressiveMauve version 2.3.0 [50]. The regions shared by all genomes were then extracted and used to generate a phylogenetic tree using ClonalFrame version 1.2 [51]. In addition to this phylogeny based on the core genome, we constructed a tree based on the pan genome as follows: a gene content matrix consisting of 1s and 0s was constructed where the columns correspond to the different strains and the rows to different gene families. An entry of 1 means presence of a gene family in a given strain, whereas a 0 means absence. This matrix was used for calculation of Manhattan distances between

strains, which were then used for hierarchical clustering in order to construct the pan genome tree. These computations were done using R [52].

Gene family enrichment analysis

Enrichment for gene families was found using the gene content matrix described above, combined with previous knowledge of the isolates. Isolates were grouped according to phenotypic or phylogenetic criteria and then gene families overrepresented in one group relative to others were counted in the matrix. Group sizes and cutoff values used to define overrepresentation are shown in Table 2. Results were plotted as heat maps in R. To assess the statistical significance of these results, we designed a permutation test in which we used the same group sizes as above but assigned group membership randomly according to a combinatorial scheme. This procedure produces the numerical distribution of gene family enrichments for all possible combinations of group members given some fixed set of group sizes and enrichment criteria, with which our results could be compared. This procedure provides an indication of whether our results could arise from random associations, although the limited strain sample means that subtle associations may go undetected. P-values for our focal enrichments were derived from the computed distributions as the empirical probability of observing an enrichment of equivalent or higher rank. Genes enriched in each of the groups and cross category comparisons are listed in Additional files 2-10 ,13 and 14.

Multiple correspondence analysis

Multiple correspondence analysis was carried out as described by Nenadic and Greenacre using singular value decomposition of the scaled gene content indicator matrix [53].

Enrichment for biological processes and re-annotation of enriched genes

The lists of genes generated by the gene family enrichment analysis and found to be overrepresented within each of the categories were used to generate the biological process scores using Blast2GO (www.Blast2GO.com) [54]. This software annotates coding sequences and assigns them to gene ontology (GO) categories. Blast2GO gene annotations of enriched and unique gene sets can be found in Additional files 11, 12 and 15-23.

Codon usage bias analysis

Genome wide codon usage tables were computed from the annotated coding sequences for each strain. Codon usage for highly expressed genes was computed from the 54 ribosomal protein gene sequences extracted from the annotation of each EDM strain. Effective number of codons (ENC) was computed according to the method of Wright [55]. This provides a metric for the evenness of codon usage with smaller values indicating a bias toward more specialized codon usage while higher values signify more uniform usage. The index of bias in highly expressed genes, ΔENC, was computed as the scaled difference between genome wide ENC and highly expressed gene (ribosomal protein gene) ENC [16]. We did not apply correction for differential G+C content in our ENC calculations as this did not vary significantly across genomes.

Abbreviations

BLAST: basic local alignment search tool; ENC: effective number of codons; EPEC: enteropathogenic *E.coli*; GO: gene ontology; HGT: horizontal gene transfer; MLST: Multi-Locus Sequence Typing; PTS: phosphotransferase system; RAST : rapid annotations using subsystems technology

Competing interests

The authors declare that no competing interests exist.

Authors' contributions

Conceived and designed the experiments: ED and PT. Performed the experiments: ED and PT. Analyzed the data: ED, PT, KL, and XD. Contributed reagents/materials/analysis tools: JEA, NCS and XD. Wrote the paper: ED and PT (edits by KL, JEA, XD, KSR, KR, and NCS). All authors have read and approved the manuscript for publication.

Acknowledgements

The Norwegian Foundation for Health and Rehabilitation, The Norwegian Diabetes Association, Centre for Ecological and Evolutionary Synthesis (CEES), The Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU) (Project 46023900) for funding this research. We acknowledge Liselotte Buarø for technical assistance and Lex Nederbragt for useful discussions. The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities.

Additional files

Additional file 1: Supporting information - Word document containing Supplementary table 1 and Supplementary figures 1-7.

Additional files 2-10: Txt files containing the RAST annotations of the individual enrichment analyses.

Additional files 11 and 12: Txt files containing blast2GO re-annotations unique genes in the parental strain (EDM1c) and evolved strain (EDM123c).

Additional files 13 and 14: Txt files containing RAST annotations of the cross-category enrichments analyses.

Additional files 15-23: Txt files containing blast2GO re-annotations of the individual enrichment analyses.

Reference List

1. Cho I, Blaser MJ: **The human microbiome: at the interface of health and disease.** *Nat Rev Genet* 2012, **13:**260-270.

2. Tenaillon O, Skurnik D, Picard B, Denamur E: **The population genetics of commensal Escherichia coli.** *Nat Rev Microbiol* 2010, **8:**207-217.

3. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced Escherichia coli genomes.** *Microb Ecol* 2010, **60:**708-720.

4. Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW: **Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray.** *Genome Biol* 2007, **8:**R267.

5. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O et al.: **Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.

6. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT: **Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species.** *Proc Natl Acad Sci U S A* 2011, **108:**7200-7205.

7. Storro O, Oien T, Dotterud CK, Jenssen JA, Johnsen R: **A primary health-care intervention on pre- and postnatal risk factor behavior to prevent childhood allergy. The Prevention of Allergy among Children in Trondheim (PACT) study.** *BMC Public Health* 2010, **10:**443.

8. Storro O, Oien T, Langsrud O, Rudi K, Dotterud C, Johnsen R: **Temporal variations in early gut microbial colonization are associated with allergen-specific immunoglobulin E but not atopic eczema at 2 years of age.** *Clin Exp Allergy* 2011, **41:**1545-1554.

9. Rudi K, Storro O, Oien T, Johnsen R: **Modelling bacterial transmission in human allergen-specific IgE sensitization.** *Lett Appl Microbiol* 2012, **54:**447-454.

10. de Muinck EJ, Oien T, Storro O, Johnsen R, Stenseth NC, Ronningen KS, Rudi K: **Diversity, transmission and persistence of Escherichia coli in a cohort of mothers and their infants.** *Environmental Microbiology Reports* 2011, **3:**352-359.

11. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R: **Diversity, stability and resilience of the human gut microbiota.** *Nature* 2012, **489:**220-230.

12. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM: **Accessing the soil metagenome for studies of microbial diversity.** *Appl Environ Microbiol* 2011, **77:**1315-1324.

13. Elliott SJ, Wainwright LA, McDaniel TK, Jarvis KG, Deng YK, Lai LC, McNamara BP, Donnenberg MS, Kaper JB: **The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic Escherichia coli E2348/69.** *Mol Microbiol* 1998, **28:**1-4.

14. Oien T, Storro O, Johnsen R: **Intestinal microbiota and its effect on the immune system--a nested case-cohort study on prevention of atopy among small children in Trondheim: the IMPACT study.** *Contemp Clin Trials* 2006, **27:**389-395.

15. Ramer SW, Schoolnik GK, Wu CY, Hwang J, Schmidt SA, Bieber D: **The type IV pilus assembly complex: biogenic interactions among the bundle-forming pilus proteins of enteropathogenic Escherichia coli.** *J Bacteriol* 2002, **184:**3457-3465.

16. Vieira-Silva S, Rocha EP: **The systemic imprint of growth and its uses in ecological (meta)genomics.** *PLoS Genet* 2010, **6:**e1000808.

17. Wetzel J, Kingsford C, Pop M: **Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies.** *BMC Bioinformatics* 2011, **12:**95.

18. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R et al.: **The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates.** *J Bacteriol* 2008, **190:**6881-6893.

19. Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham AF, Petty NK, Mahon V, Brinkley C, Hobman JL et al.: **A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic Escherichia coli strain H10407.** *J Bacteriol* 2010, **192:**5822-5831.

20. Brzuszkiewicz E, Bruggemann H, Liesegang H, Emmerth M, Olschlager T, Nagy G, Albermann K, Wagner C, Buchrieser C, Emody L et al.: **How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains.** *Proc Natl Acad Sci U S A* 2006, **103:**12879-12884.

21. Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P et al.: **Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach.** *Proc Natl Acad Sci U S A* 2006, **103:**5977-5982.

22. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I: **Pathogenicity-associated islands in extraintestinal pathogenic Escherichia coli are fitness elements involved in intestinal colonization.** *J Bacteriol* 2010, **192:**4885-4893.

23. Le GT, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O: **Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains.** *Mol Biol Evol* 2007, **24:**2373-2384.

24. Kamada N, Kim YG, Sham HP, Vallance BA, Puente JL, Martens EC, Nunez G: **Regulated virulence controls the ability of a pathogen to compete with the gut microbiota.** *Science* 2012, **336:**1325-1329.

25. Alteri CJ, Mobley HL: **Escherichia coli physiology and metabolism dictates adaptation to diverse host microenvironments.** *Curr Opin Microbiol* 2012, **15:**3-9.

26. Read AF: **The evolution of virulence.** *Trends Microbiol* 1994, **2:**73-76.

27. Meric G, Kemsley EK, Falush D, Saggers EJ, Lucchini S: **Phylogenetic distribution of traits associated with plant colonization in Escherichia coli.** *Environ Microbiol* 2012.

28. Koch AL: **The adaptive responses of Escherichia coli to a feast and famine existence.** *Adv Microb Physiol* 1971, **6:**147-217.

29. Button DK: **Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the michaelis constant.** *Appl Environ Microbiol* 1991, **57:**2033-2038.

30. Adlerberth I: **Factors influencing the establishment of the intestinal microbiota in infancy.** *Nestle Nutr Workshop Ser Pediatr Program* 2008, **62:**13-29.

31. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO: **Development of the human infant intestinal microbiota.** *PLoS Biol* 2007, **5:**e177.

32. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE: **Succession of microbial consortia in the developing infant gut microbiome.** *Proc Natl Acad Sci U S A* 2011, **108 Suppl 1:**4578-4585.

33. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.

34. Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Kling BH, Gonzalez A, Werner JJ, Angenent LT, Knight R et al.: **Host Remodeling of the Gut Microbiome and Metabolic Changes during Pregnancy.** *Cell* 2012, **150:**470-480.

35. Buckling A, Craig MR, Brockhurst MA, Colegrave N: **The Beagle in a bottle.** *Nature* 2009, **457:**824-829.

36. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M et al.: **Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection.** *Nat Genet* 2011, **43:**482-486.

37. Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, Dougan G, Gordon MA: **High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal Salmonella typhimurium disease.** *Clin Infect Dis* 2012, **54:**955-963.

38. Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, Banfield JF: **Strain-resolved community genomic analysis of gut microbial colonization in a premature infant.** *Proc Natl Acad Sci U S A* 2011, **108:**1128-1133.

39. Taylor DE: **Bacterial tellurite resistance.** *Trends Microbiol* 1999, **7:**111-115.

40. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP et al.: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** *DNA Res* 2007, **14:**169-181.

41. Martens EC, Chiang HC, Gordon JI: **Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont.** *Cell Host Microbe* 2008, **4:**447-457.

42. Kim DH, Konishi L, Kobashi K: **Purification, characterization and reaction mechanism of novel arylsulfotransferase obtained from an anaerobic bacterium of human intestine.** *Biochim Biophys Acta* 1986, **872:**33-41.

43. Kurihara S, Oda S, Kumagai H, Suzuki H: **Gamma-glutamyl-gamma-aminobutyrate hydrolase in the putrescine utilization pathway of Escherichia coli K-12.** *FEMS Microbiol Lett* 2006, **256:**318-323.

44. Kurihara S, Oda S, Tsuboi Y, Kim HG, Oshida M, Kumagai H, Suzuki H: **gamma-Glutamylputrescine synthetase in the putrescine utilization pathway of Escherichia coli K-12.** *J Biol Chem* 2008, **283:**19981-19990.

45. Skibinski DA, Golby P, Chang YS, Sargent F, Hoffman R, Harper R, Guest JR, Attwood MM, Berks BC, Andrews SC: **Regulation of the hydrogenase-4 operon of Escherichia coli by the sigma(54)-dependent transcriptional activators FhlA and HyfR.** *J Bacteriol* 2002, **184:**6642-6653.

46. Didelot X, Meric G, Falush D, Darling AE: **Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli.** *BMC Genomics* 2012, **13:**256.

47. Afset JE, Anderssen E, Bruant G, Harel J, Wieler L, Bergh K: **Phylogenetic backgrounds and virulence profiles of atypical enteropathogenic Escherichia coli strains from a case-control**

study using multilocus sequence typing and DNA microarray analysis. *J Clin Microbiol* 2008, **46:**2280-2290.

48. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M et al.: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9:**75.

49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

50. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *Plos One* 2010, **5:**e11147.

51. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175:**1251-1266.

52. R Foundation for Statistical Computing. **R:A language and environment for statistical computing.** Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R- project.org.

53. Nenadic O, Greenacre M: **Correspondence analysis in R, with two- and three-dimensional graphics: The ca package.** *Journal of Statistical Software* 2007, **20**.

54. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.

55. Wright F: **The 'effective number of codons' used in a gene.** *Gene* 1990, **87:**23-29.

Table 1. List of strains used in this study with corresponding genome information.

ID (alt. ID)	child number	Clinical condition	Age at Sampling	Phylogenetic group	Nr. contigs	Median coverage	Colonization Category	Accession Number
EDM1c	1360	Healthy	10days	B2	712	7,55	early	EMBL:ERS155053
EDM3c	1360	Healthy	1year	B1	306	12,4	late	EMBL:ERS155049
EDM16c	1870	Healthy	7days	B1	204	13,75	early	EMBL:ERS155051
EDM70c	1997	Healthy	10days	B2	562	8,5	early	EMBL:ERS155055
EDM49c	1891	Healthy	4days	B2	163	14,5	early	EMBL:ERS155056
EDM101c	1891	Healthy	11days	B2	169	16	early	EMBL:ERS155057
EDM106c	123	Healthy	4days	B2	585	8	early	EMBL:ERS155058
EDM116c	123	Healthy	1year	A	864	8,2	late	EMBL:ERS155052
EDM123c	1360	Healthy	4months	B2	669	8,5	late	EMBL:ERS155054
EDM530c	123	Healthy	2years	NA	198	17,5	late	EMBL:ERS155050
JEA117c (Trh9)*	117c	Healthy	1year	B2	284	10,1	late	EMBL:ERS178156
JEA242p (Trh52)*	242p	Diarrhoea	3years	B2	140	13,2	NC	EMBL:ERS178157
JEA297p (Trh58)*	297p	Diarrhoea	2years	D	521	11,1	NC	EMBL:ERS178158
JEA179p (Trh39)*	179p	Diarrhoea	4years	B1	848	7,8	NC	EMBL:ERS178159
JEA160c (Trh12)*	160c	Healthy	2years	A	188	20,1	late	EMBL:ERS178160
JEA124p (Trh29)*	124p	Diarrhoea	2years	A	800	8,9	NC	EMBL:ERS178161

c = commensal isolate, isolated from healthy child

p = pathogenic isolate, isolated from child with diarrhoea

* Isolation based on positive PCR analysis for intimin gene *eae*.

NC=not categorized

The genome assembly statistics are results of Newbler *de novo* assembly. The colonization categorizations are the ones used for gene enrichment comparison between early and late colonizers. Alternative strain IDs are from [47].

Table 2. Criteria used for gene enrichment analyses.

Sorting criteria	Focal group (nr. strains)	Gene presence in focal group	Gene absence in non-focal group
Criteria I, cladistic comparison	Clade1 (8)	≥7	≥7
	Clade2 (8)	≥7	≥7
Criteria II, pathogen/ commensal comparison	Pathogen (4)	≥3	≥9
	Commensal (12)	≥9	≥3
Criteria III, growth rate comparison	Fast (2)	2	≥6
	Medium (4)	≥3	≥4
	Slow (4)	≥3	≥4
Criteria IV, colonization time comparison	Early (6)	≥ 5	≥4
	Late (6)	≥4	≥5

Criteria I: *Criteria used for discriminating cladistic gene content enrichments. Each of the two clades contained 8 strains and enrichment required a gene to be present in at least 7 strains of one clade (focal group) while being absent from at least 7 strains in the other clade (non-focal group).*

Criteria II: *Criteria used for discriminating pathogen vs. commensal gene content enrichments. Since the two groups are of unequal size, a pathogen enriched gene had to present in at least 3 of 4 pathogenic strains and absent from at least 9 of 12 commensal strains. A commensal enriched gene had to be present in at least 9 of 12 commensal strains and absent from at least 3 of 4 pathogenic strains.*

Criteria III: *Criteria used for discriminating growth rate related gene content enrichments. The three growth rate categories (slow, medium and fast) contained 2,4, and 4 strains respectively. For a gene to be considered enriched in the fast category, a gene had to be preset in both fast strains and absent from at least 6 of 8 of the combined slow and medium strains. For a gene to be considered enriched in the medium category, a gene had to be preset in at least 3 of 4 medium strains and absent from at least 4 of 6 of the combined slow and fast strains. For a gene to be considered enriched in the slow category, a gene had to be present in at least 3 of 4 slow strains and absent from at least 4 of 6 of the combined medium and fast strains.*

Criteria IV: *Criteria used for discriminating early vs. late colonizer gene content enrichments. The two groups contain 6 strains each. Since one of the strains in the early group was also isolated in the late group (EDM123c). An asymmetrical enrichment profile was designed which required an early enriched gene to be present in at least 5 of 6 early strains and absent in at least 4 of 6 late strains. A gene enriched in the late colonizer group had to be present in at least 4 of 6 late strains and absent from at least 5 of 6 early strains.*

Figure legends:

Figure 1: Overview of relative and cumulative proportions of genes as the number of included genomes increases. Bright and dark bars show relative and cumulative proportions, respectively. All duplicated gene annotations were removed for this analysis. Less than 20% of annotated genes are unique to one strain and while 52.4% are common to all.

Figure 2: Heat map of total gene content comparisons. Gene presence is shown in blue and gene absence in yellow. The number of genes is depicted on the x-axis. Strains are listed in the order following hierarchical clustering created using a Manhattan distance matrix based on the gene presence/absence gene content matrix.

Figure 3: Comparison of genome trees generated by core and pan-genomes. A. The core genome phylogeny was created using ClonalFrame. **B**. Pan-genome tree generation was created using a Manhattan distance matrix based on the gene presence/absence gene content matrix. The scale below the pan-genome tree indicates Manhattan distances. Both methods separated the strains into two main clades (1 and 2).

Figure 4: **Gene content enrichment comparing main clades 1 and 2.** Enrichment analysis was carried out using criteria I (Table 2). The distribution of possible strain group permutations is presented in Figure S2.

Figure 5: Multiple correspondence analysis of the gene content matrix. The plot shows principal coordinates along the two main components. Each point on the graph represents a gene and the color of the point relates the number of genomes in which it is present. The positions of the genome labels represent the relative distances of the genomes along the respective components.

Figure 6: **Gene content profiles of pathogenic and commensal strains**. Enrichment analysis was carried out following criteria II (Table 2). 164 genes (p= 0.02) were found to be enriched in the pathogenic group while only 33 genes (p=0.18) were enriched in the commensal group. The complete distributions of possible gene enrichments are presented in Figure S3.

Figure 7: Ratio of anaerobic/aerobic generation times related to anaerobic generation times of IMPACT isolates (de Muinck et al. submitted). Circled strains are the ones for which we present genome sequences in this study. Blue circled strains are categorized as fast growers, green have a medium growth rate, and red circled strains are slow growing strains. R^2=0.51, p<0.0001.

Figure 8: Gene content profiles of slow, medium and fast growing strains. Enrichment analysis was carried out using criteria III (Table 2). The fast category had 227 (p=0.09) genes enriched. The medium growth rate category had only 46 (p=0.56) genes enriched while the slow category had 324 (p=0.04) genes. Distributions of possible enrichment profiles are shown in Figure S4.

Figure 9: **Gene content enrichment profiles of early and late colonizer strains.** Enrichment analysis was carried out using criteria IV (Table 2). Both early and late colonizers show significant enrichments (p=0.02 and p=0.05, respectively). The complete distributions of possible enrichment profiles are shown in Figure S5 in Additional file 1. EDM1c is an early colonizer that is clonally related to the late colonizer EDM123c. EDM123c maintains the early colonizer genomic profile but has lost genes found in the early colonizer profile.

Figure 10: General comparison of the enrichment profiles of the strain categories. Each column is created from the gene enrichment list for each grouping (Additional files 15-23). Each list of gene sequences was evaluated for ontology level 3 biological process categorization using Blast2GO for SEED assignments. The coloring scheme corresponds to enrichment scores assigned by Blast2GO. Grouping categories are shown on the x-axis, and the different comparisons are separated by white lines. Enrichment comparisons were performed between clade1 and clade2; pathogen and commensal; slow, medium and fast growth rates; early and late colonization. The color key indicates the enrichments scores for the biological processes.

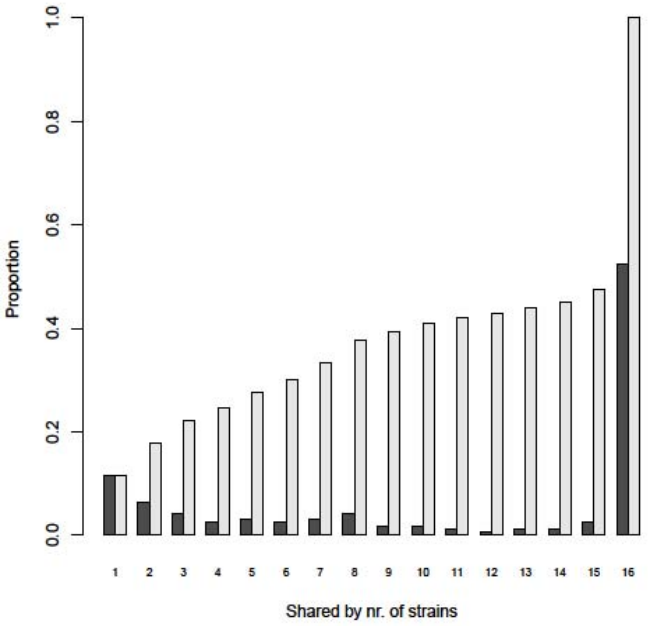

Proportion
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Shared by nr. of strains

Figure 1

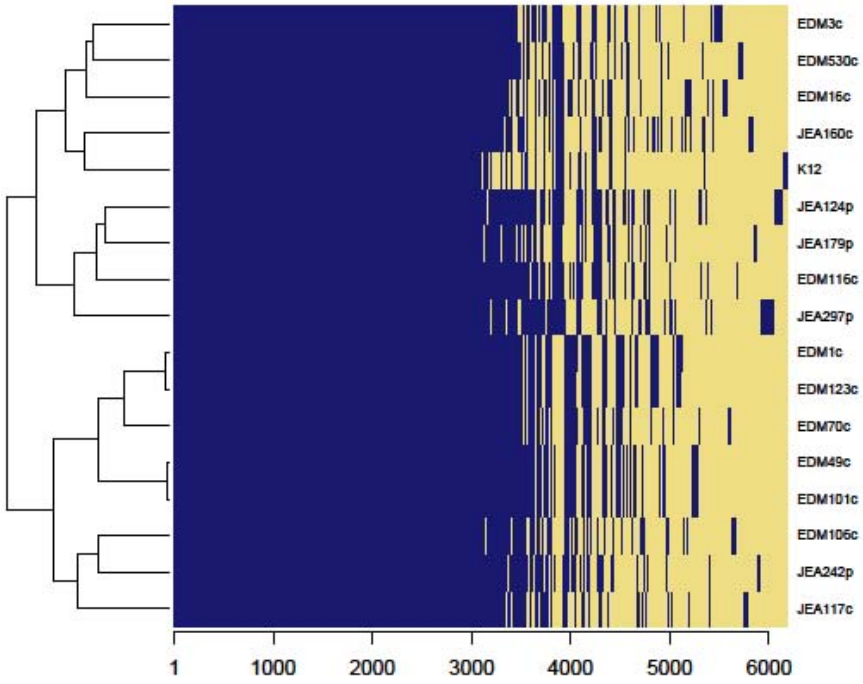

EDM3c
EDM530c
EDM16c
JEA160c
K12
JEA124p
JEA179p
EDM116c
JEA297p
EDM1c
EDM123c
EDM70c
EDM49c
EDM101c
EDM106c
JEA242p
JEA117c
1
1000
2000
3000
4000
5000
6000

Figure 2

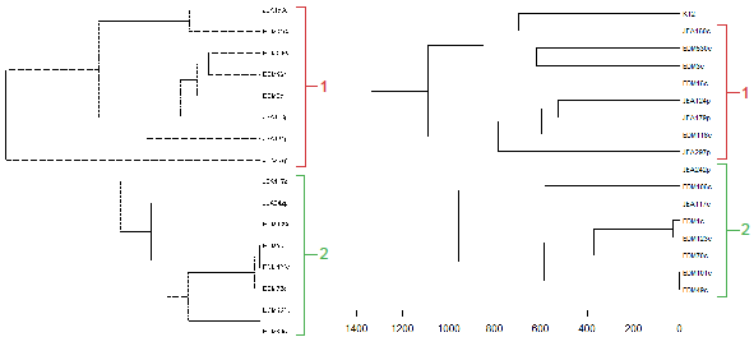


Figure 3

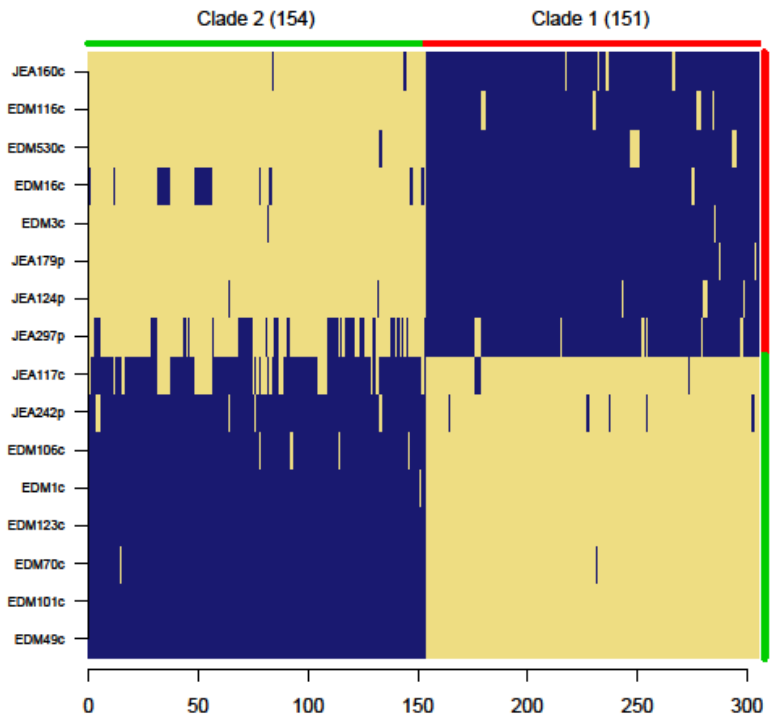

Clade 2 (154)
Clade 1 (151)
JEA160c
EDM116c
EDM530c
EDM16c
EDM3c
JEA179p
JEA124p
JEA297p
JEA117c
JEA242p
EDM106c
EDM1c
EDM123c
EDM70c
EDM101c
EDM49c
0
50
100
150
200
250
300

Figure 4

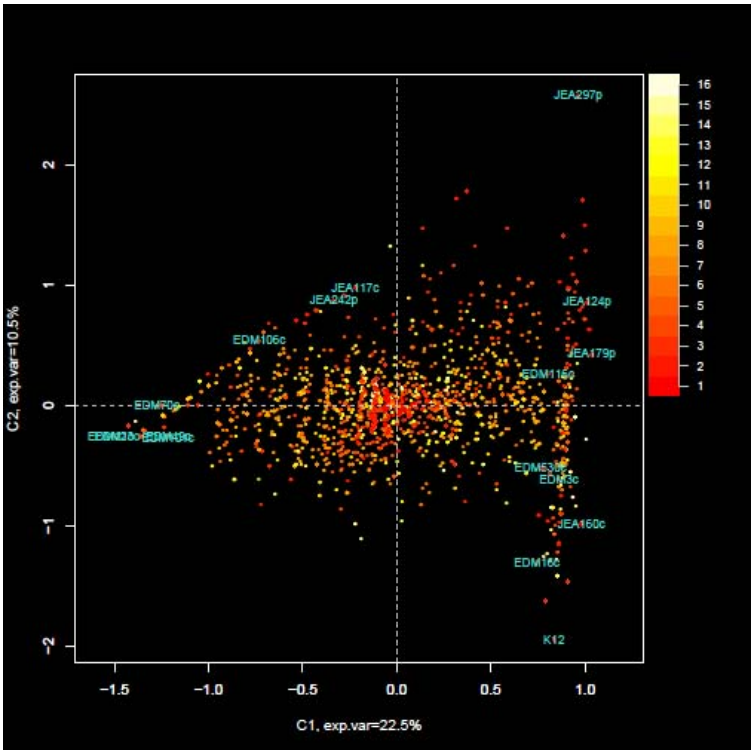

JEA297p
JEA117c
JEA242p
JEA124p
EDM106c
JEA179p
EDM116c
EDM160c
K12
C1, exp.var=22.5%
C2, exp.var=10.5%
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

Figure 5

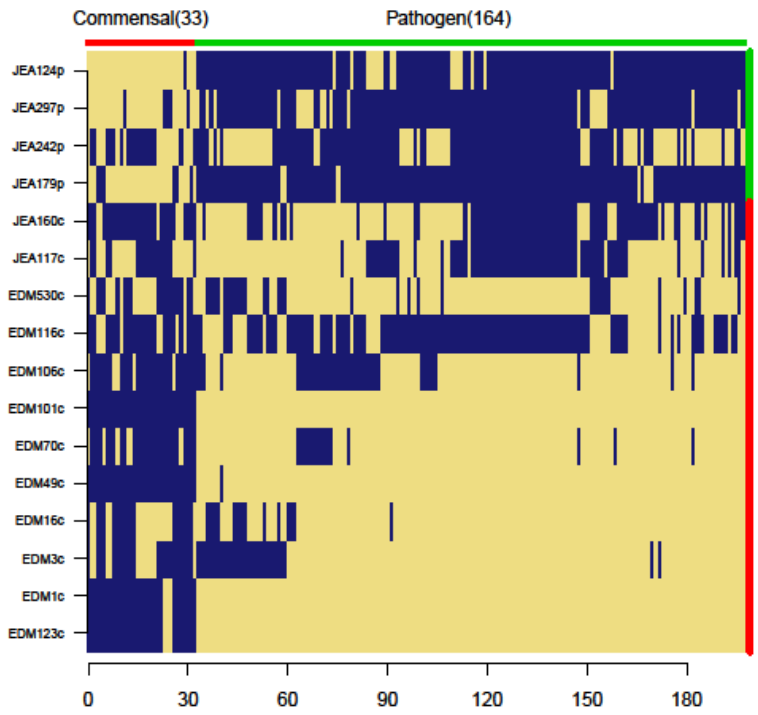

Commensal(33)
Pathogen(164)
JEA124p
JEA297p
JEA242p
JEA179p
JEA160c
JEA117c
EDM530c
EDM116c
EDM106c
EDM101c
EDM70c
EDM49c
EDM16c
EDM3c
EDM1c
EDM123c
0
30
60
90
120
150
180

Figure 6

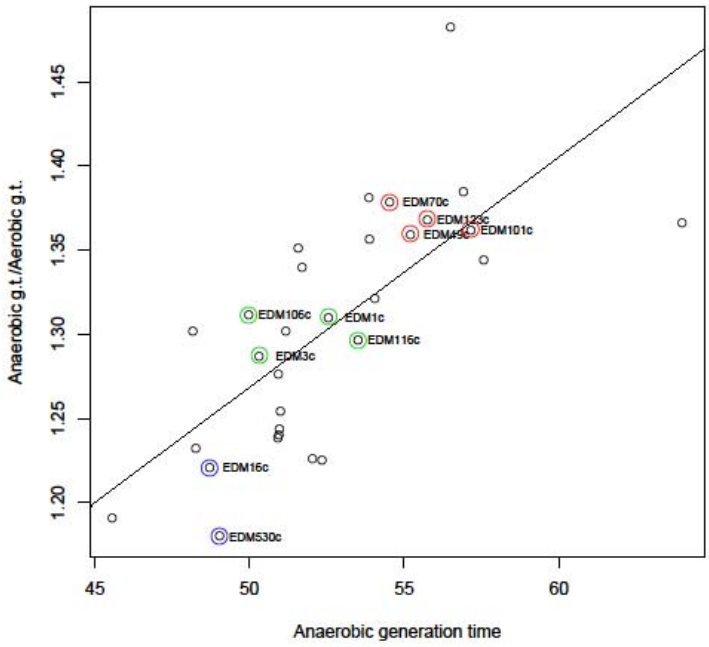

Anaerobic g.t./Aerobic g.t.
1.20
1.25
1.30
1.35
1.40
1.45
45
50
55
60
Anaerobic generation time
EDM70c
EDM123c
EDM49c
EDM101c
EDM106c
EDM1c
EDM116c
EDM3c
EDM16c
EDM530c

Figure 7

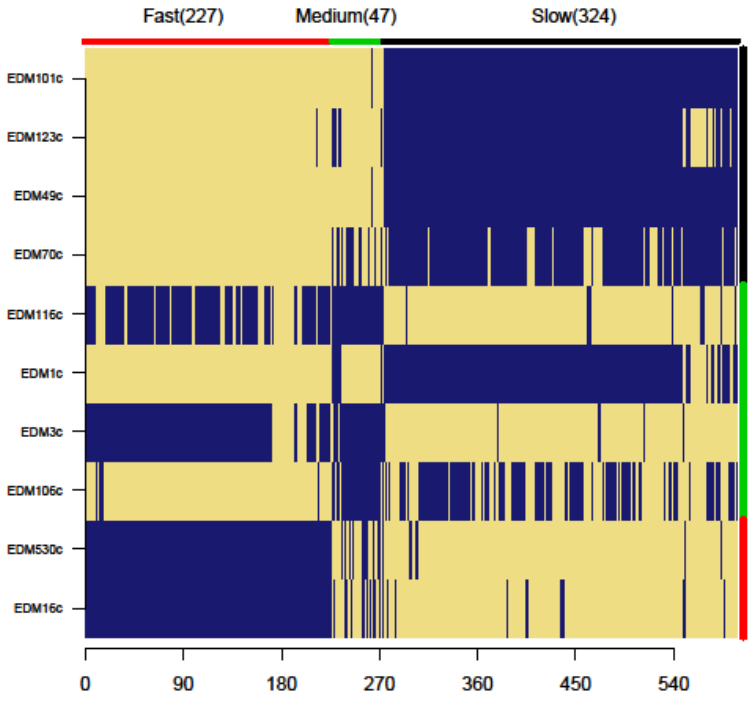

Fast(227)
Medium(47)
Slow(324)
EDM101c
EDM123c
EDM49c
EDM70c
EDM116c
EDM1c
EDM3c
EDM106c
EDM530c
EDM16c
0
90
180
270
360
450
540

Figure 8

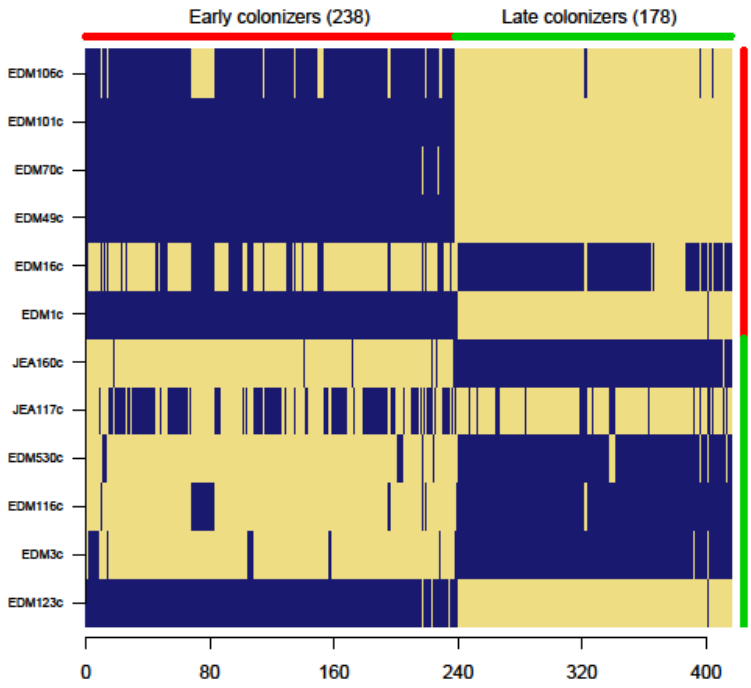

Early colonizers (238)
Late colonizers (178)
EDM106c
EDM101c
EDM70c
EDM49c
EDM16c
EDM1c
JEA160c
JEA117c
EDM530c
EDM116c
EDM3c
EDM123c
0
80
160
240
320
400

Figure 9

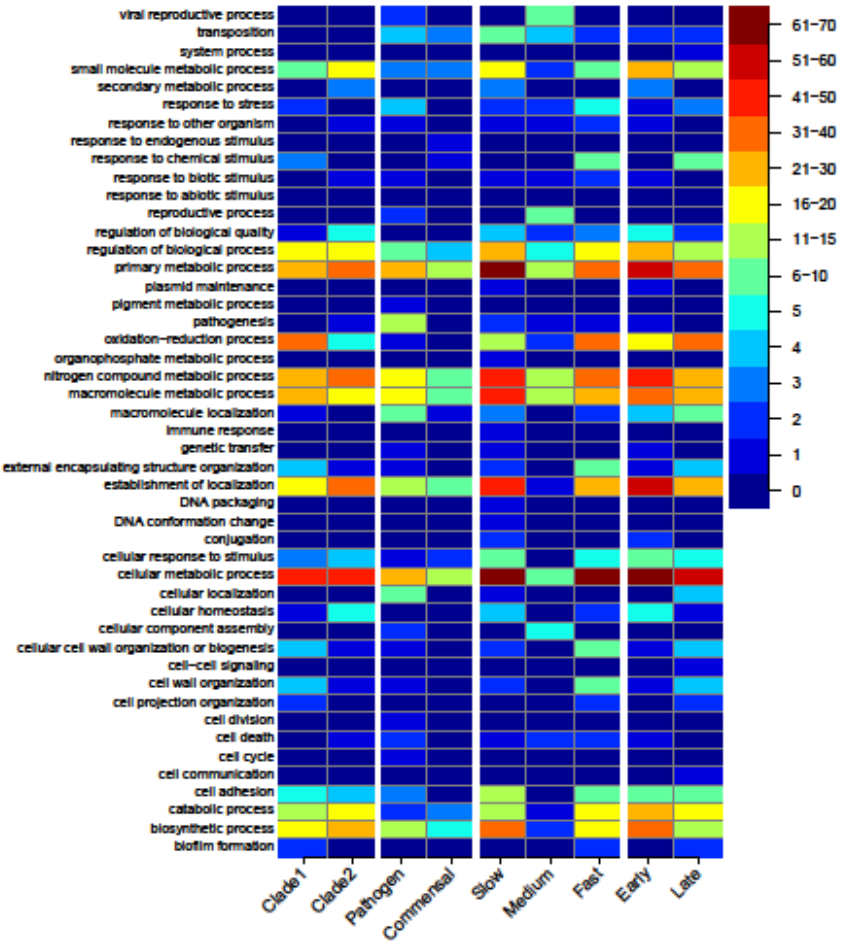

viral reproductive process
transposition
system process
small molecule metabolic process
secondary metabolic process
response to stress
response to other organism
response to endogenous stimulus
response to chemical stimulus
response to biotic stimulus
response to abiotic stimulus
reproductive process
regulation of biological quality
regulation of biological process
primary metabolic process
plasmid maintenance
pigment metabolic process
pathogenesis
oxidation−reduction process
organophosphate metabolic process
nitrogen compound metabolic process
macromolecule metabolic process
macromolecule localization
immune response
genetic transfer
external encapsulating structure organization
establishment of localization
DNA packaging
DNA conformation change
conjugation
cellular response to stimulus
cellular metabolic process
cellular localization
cellular homeostasis
cellular component assembly
cellular cell wall organization or biogenesis
cell−cell signaling
cell wall organization
cell projection organization
cell division
cell death
cell cycle
cell communication
cell adhesion
catabolic process
biosynthetic process
biofilm formation
Clade1
Clade2
Pathogen
Commensal
Slow
Medium
Fast
Early
Late
61−70
51−60
41−50
31−40
21−30
16−20
11−15
6−10
5
4
3
2
1
0

Figure 10

Supporting information: Supplementary table 1. Supplementary figures 1-7.

Comparisons of infant *Escherichia coli* isolates link genomic profiles with adaptation to the ecological niche

Eric J. de Muinck (ericdemuinck@gmail.com)[1,2,3]
Karin Lagesen (karin.lagesen@bio.uio.no)[1]
Jan Egil Afset (jan.afset@ntnu.no)[4,5]
Xavier Didelot (x.didelot@imperial.ac.uk)[6]
Kjersti S. Rønningen (kjersti.skjold.ronningen@gmail.com)[7]
Knut Rudi (knut.rudi@umb.no)[8]
Nils Chr. Stenseth (n.c.stenseth@bio.uio.no)[1]
Pål Trosvik, corresponding author, (pal.trosvik@bio.uio.no)[1]

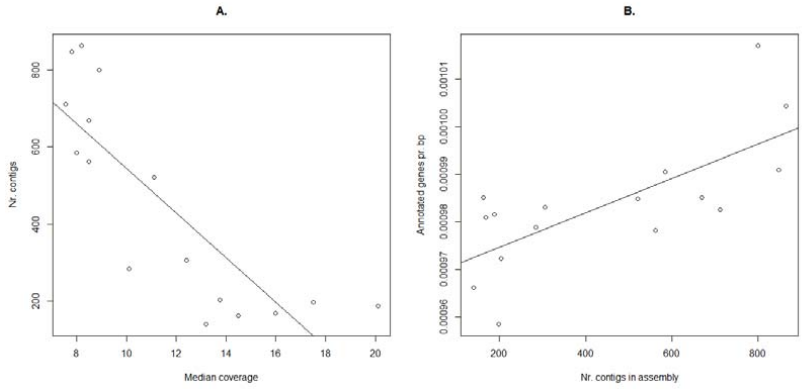
1 Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, Po Box 1066, 0316 Oslo Norway
2 Division of Epidemiology, Norwegian Institute of Public Health, PO Box 4404, 0456 Oslo, Norway
3 NOFIMA - The Norwegian Institute of Food, Fisheries and Aquaculture Research, PO Box 210, 1430 Ås, Norway
4 Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, PO Box 8905, 7491 Trondheim, Norway
5 Department of Medical Microbiology , St Olavs Hospital, PO Box 3250, 7006 Trondheim, Norway
6 Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK
7 Department of Pediatric Research, Oslo University Hospital, Rikshospitalet, PO Box 4905, 0424 Oslo, Norway
8 Department of Chemistry, Biotechnology and Food Science, University of Life Sciences, PO Box 5003, 1432 Ås, Norway

Table S1: Evaluation of genome sequenced strains for the presence of the enterocyte effacement pathogenicity island. The complete 35,624 base pair LEE pathogenicity island of *E.coli* strain E2348/69 (AF022236.1) was BLASTed against each of the genome assemblies. Strains not included in the table showed no significant identity. The lengths of the sequences found in the genome of each strain with greater than 90% identity were summed and the percent of the summed length was compared with the complete 35,624 base pair length of the pathogenicity island.

Strain ID	percent of total alignment length >90% identity
EDM116c	98%
JEA117c	96%
JEA160c	94%
JEA179p	93%
JEA242p	99%
JEA297p	74%,88%*
JEA124p	49%,88%*

*80%identity threshold value

*the same comparison was repeated for some strains using a reduced identity threshold of 80%.



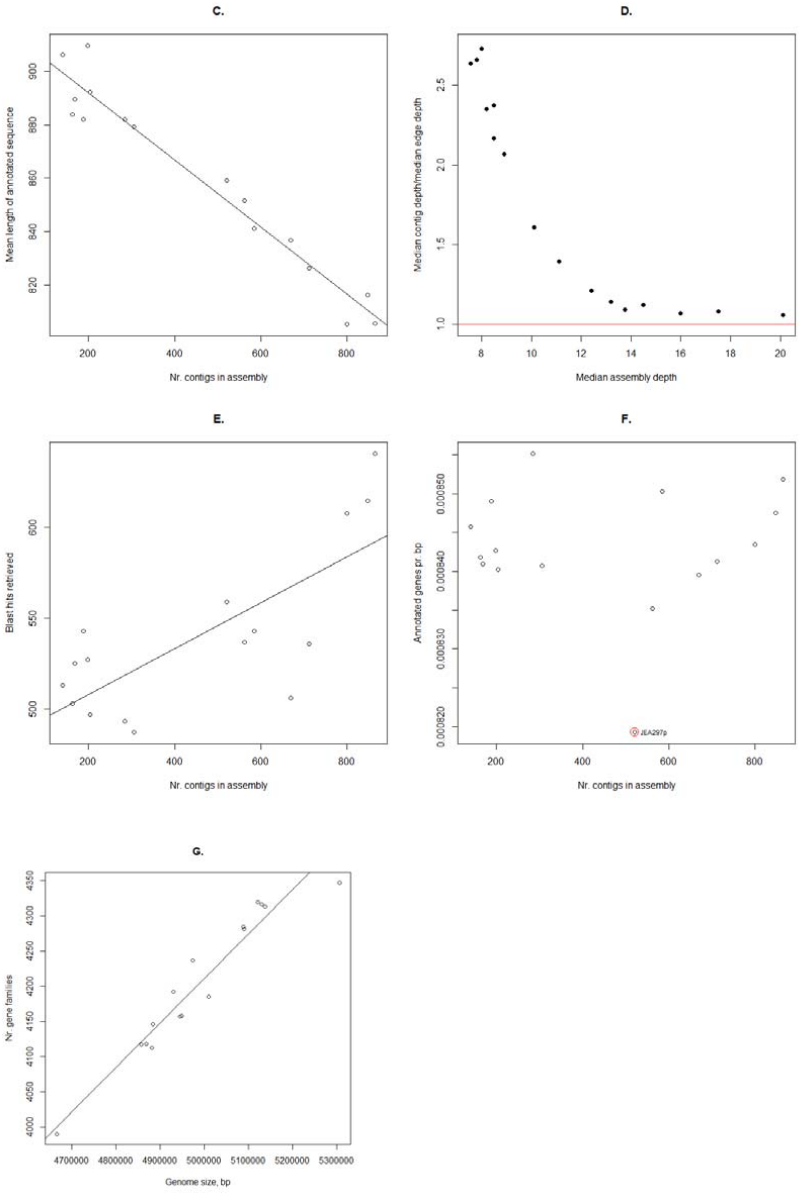


Figure S1: **A**. Decrease in the number of contigs in an assembly as the depth of coverage increases (R^2=0.69, p<0.0001). **B.** Relationship between the number of contigs in an assembly and the number of annotated sequences (R^2=0.52, p=0.0017). **C.** Decrease in the average annotated sequence length relative to the number of contigs in an assembly (R^2=0.94,

$p<0.0001$). **D**. Relationship between the median assembly read depth and the ratio of the median depth of the contigs to the median depth of the contig edges. Edge contig read depths were estimated from the outermost 1% of the total length on either side of all contigs of at least 1000 base pairs within each assembly. **E.** Relationship between the number of contigs in an assembly and the number of partial genes retrieved from re-BLASTing annotated sequences against the complete genome assemblies ($R^2=0.57$, $p=0.0008$). **F**. Relationship between coverage depth and number of genes after the additional processing steps ($R^2<0.0001$, $p=0.97$). **G.** Correlation between the number of gene families and genome size ($R^2=0.92$, $p<0.0001$).

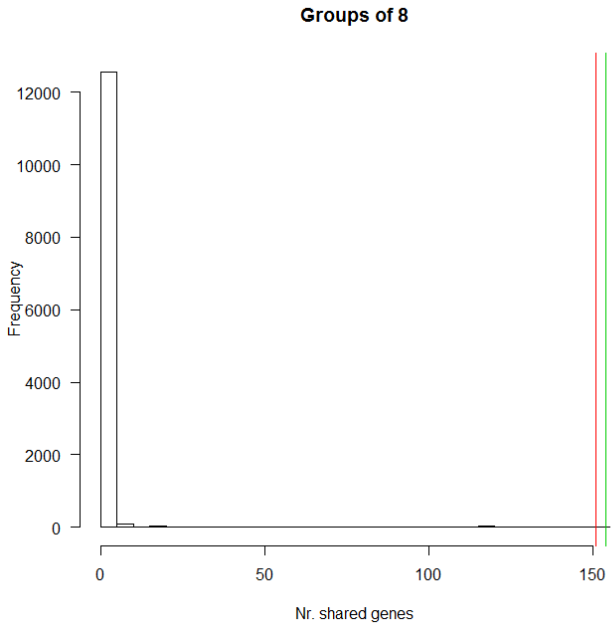


Figure S2: Distribution of possible gene content enrichment profiles using the cladistic enrichment criteria (criteria I, Table 2). 'Groups of 8' describes the number of strains in each category and the numbers of shared genes are shown on the bottom axis. The red and green lines show the number of shared genes in clade1 (151) and clade2 respectively (154). The cladistic grouping had the most significant ($p<0.0001$) distribution of the tested categories.

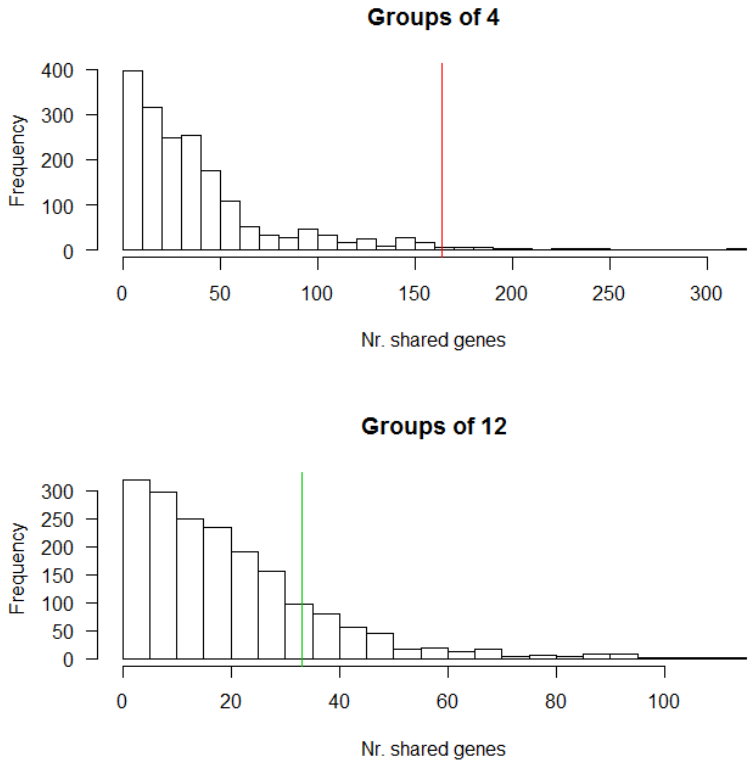


Figure S3: Distribution of possible gene content enrichment profiles using permutations of the groupings described by criteria II (Table2). The number of genes in an enrichment profile is shown on the y-axis. 'Groups of 4' corresponds to the sorting criteria used for pathogens (red line). 'Groups of 12' corresponds to the sorting criteria for the commensals (green line).

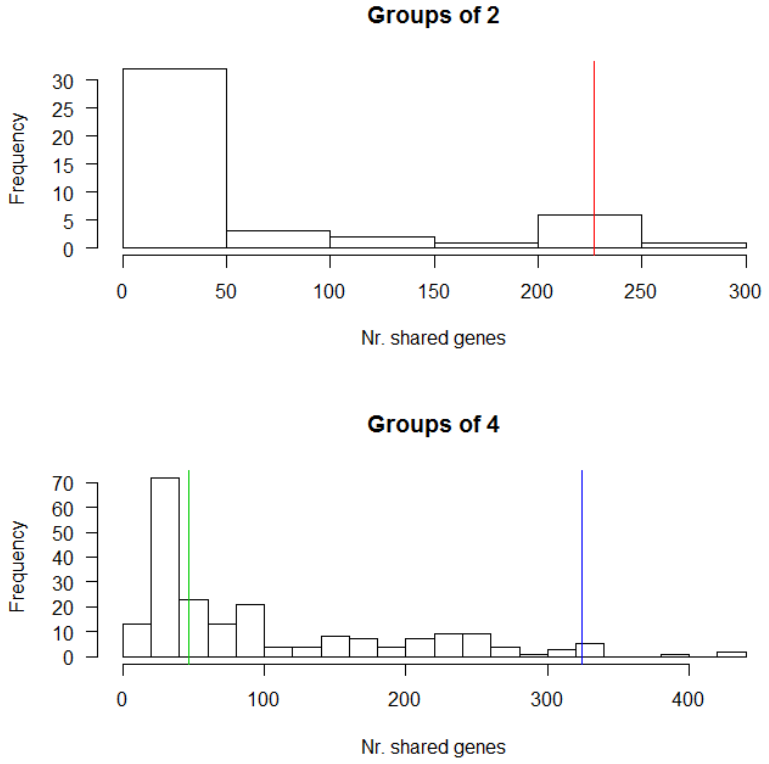


Figure S4: Distribution of possible gene content enrichment profiles of the growth rate groupings described by criteria III (Table2). 'Groups of 2' corresponds to the sorting criteria used for fast growers (red line). 'Groups of 4' corresponds to the sorting criteria used for the medium (green line) and slow growers (blue line).

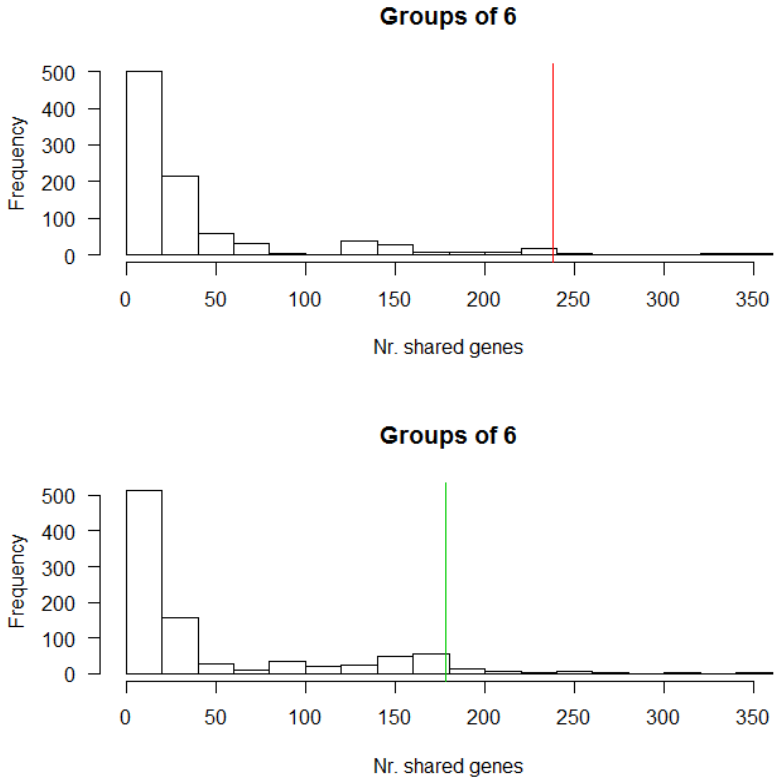


Figure S5: Distribution of possible gene content enrichment profiles of the early and late colonizer groupings described by criteria IV (Table2). The top panel 'Groups of 6' corresponds to the sorting criteria used for early colonizers (red line). The bottom panel 'Groups of 6' corresponds to the sorting criteria used for the late colonizers (green line). The use of two panels is due to the different distributions produced by the asymmetric sorting criteria used for the two gene content enrichment categories.

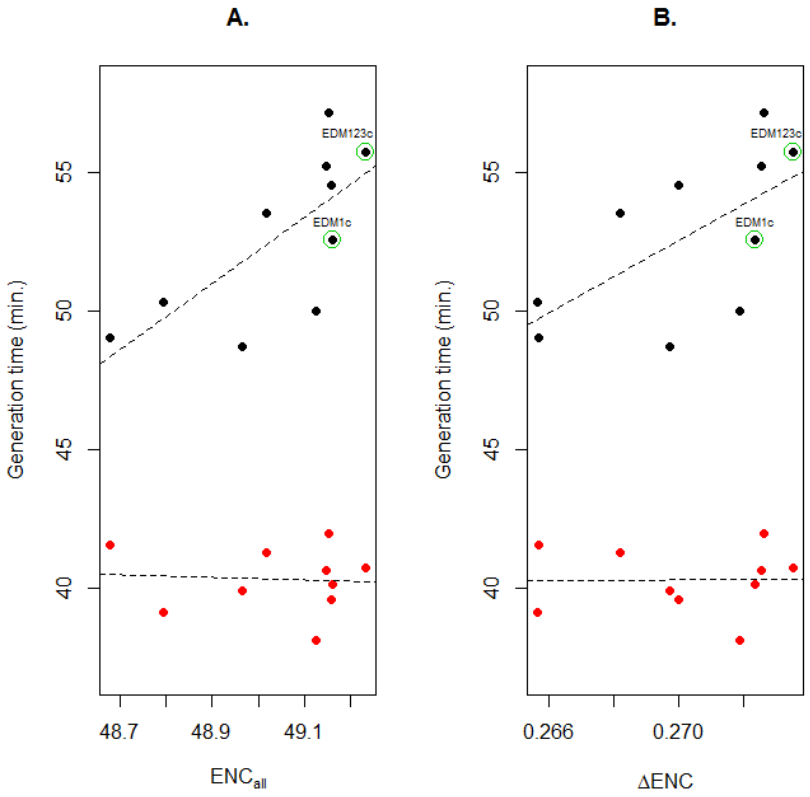


Figure S6: Plots of codon usage bias vs. aerobic (red dots) and anaerobic (black dots) generation times. **A.** Genome wide codon usage bias (ENC_{all}). **B.** Codon usage bias in highly expressed genes (ΔENC), represented by 54 ribosomal protein genes. Dashed lines are linear regression fits. The parent (EDM1c) and evolved (EDM123c) isolates separated by 4 months are marked with green rings.

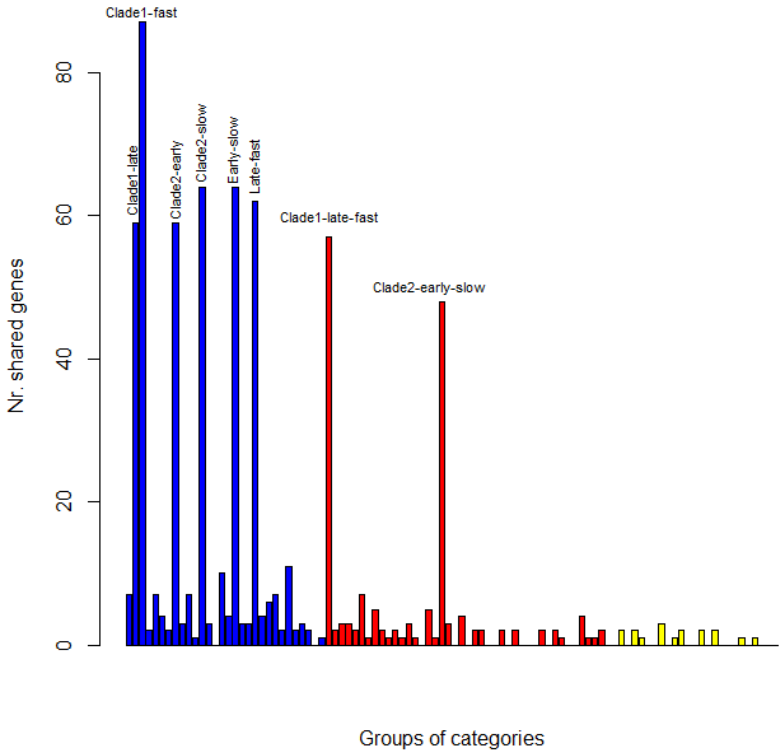


Figure S7: Number of genes enriched across grouping categories. All possible comparisons between main clade provenance, time of colonization, growth rate and pathogenicity are represented on the x-axis with selected outcomes labelled above the bars. The y-axis shows the number of Blast2GO annotated genes that are common to two or more categories. Blue bars are pairwise comparisons. Red bars are three-way comparisons. Yellow bars are four-way comparisons.