Swansea University
Prifysgol Abertawe

SCHOOL OF MEDICINE
YSGOL FEDDYGAETH

PM304 – Biomolecular Research Project 2011

Single Honours in Medical Genetics

Name: Jake Ireland

Title: Genomic and Phenotypic comparison of *Escherichia coli* isolates from host and environmental sources

Abstract

BACKGROUND: Escherichia coli are human intestinal pathogens that are a frequent cause of food poisoning in humans. Studies into the genetic characteristics of E. coli shows they have a special ability to thrive in secondary habitats like soils and waters. There is also evidence to suggest that E. coli have the ability to specify pathogenicity in a few specific hosts. This has been shown in part to stem from the horizontal transfer of fitness and pathogenicity islands.

AIM: Our aim is to apply comparative growing environments to determine phenotype responses of E. coli strains isolated from host and environmental sources, to determine if there are any preferences for certain environmental conditions.

RESULTS: No difference was determined between host and non host but there was a comparative difference between two small groups of strains that showed preference for a particular growing condition. We identified 36 unique genes to a small set of E. coli strains that prefer to grow in endotherm mimicked conditions. In addition we have found 107 unique genes that are present in strains that prefer moderate temperate climates. We used multi genome analysis to show the presence of certain fitness and virulence factors that correlates with preferential growing conditions.

CONCLUSION: From our analysis we have identified phenotypes of strains that show preferential growth based on the precedence or absence of certain virulence and fitness factors that might be relevant for distinguishing between commensal strains with enhanced growing abilities.

Introduction

E*scherichia coli* (*E. coli)* are ubiquitous to endotherms meaning that they can be found within the intestines of all warm blooded animals and reptiles (Souza *et al.*, 1999). *E. coli* are the most prominent aerobic bacteria found in the gut (Savageau, 1983). Due to the natural passage of *E. coli* they can be found in a large number of other environments besides the intestinal tract known as the secondary habitat (figure 1). *E. coli* also have a moderately sized genome of which ~17-18% is made up of mobile elements that can be altered in order to adapt to a new conditions of its environment (Dobrindt *et al.*, 2003).

The study of ishii 2006 showed that *E. coli* in the soils surrounding waters of northern minnesota showed an autochthonous (indigenous, born from the earth) behaviour and preference for soil type as as *E. coli* strain populations were three times greater in an organic soil than in a sandy soil. DNA fingerprinting also showed that *E. coli* can survive extreme winters and display a recurrent reservoir not attributable to wildlife or river waters acting as a point or non-point source. This is also similar to the findings of (Byappanahalli *et al.*, 2006) who observed the same recurrent reservoir of *E. coli* in the soils of Indiana.

E. coli can cycle between habitats and since the environmental factors of these habitats change such as nutrient availability, temperature and moister it is likely that the expression of genes is managed by turning them on and off in response to these factors. In a sense *E. coli* may display multiple cell types that are dependent on its habitat (Savageau, 1983).

E. coli can also be found in water and in particular waters surrounding sewage treatment plants (STP) (Anastasi *et al.*, 2012). Most of the research surrounding *E. coli's* presence in water is to do with it being used as a biological drinking water indicator. *E. coli* has been used as an indicator since 1890 when induction of the thermo-tolerent coliform test (Edberg *et al.*, 2000) following the death of Prince Consort Albert in 1861. This "golden age" of microbiology is when specific microbes were definitively established as causes of specific diseases and water sanitation began. Most waters surrounding sewage treatment plants contain *E. coli* strains that have pathogenic abilities and It is clear from that the procedures of STPs are not efficient enough to remove all pathogens (Anastasi *et al.*, 2012). This is due in part to the finical restrictions on sanitation and the survival of *E. coli* is dependent on many factors including exposure to sunlight (Filip et al 1987), temperature (Kudryavtseva et al1972), availability of sustained food source (Grabow et al 1975). It has been estimated that the average survival bracket of *E. coli* in water is between 4-12 weeks (Edberg *et al.*, 2000). If mixture of sewage with natural waters occurs then this lifespan can be increased due to chlorinated waters mixing with sewage produces chlorinamines that take a lot longer to kill bacteria (Rice et al 1999).

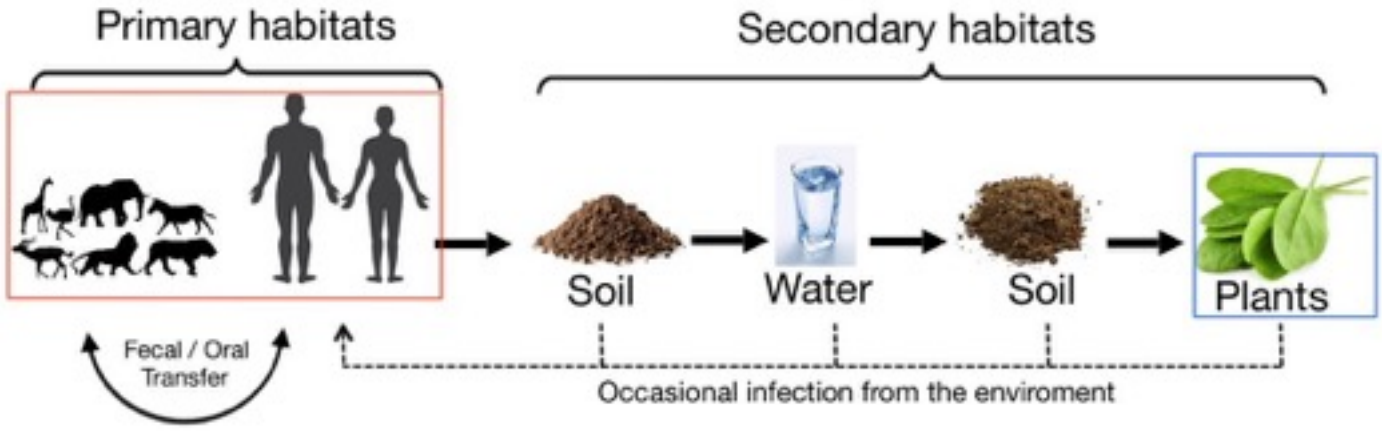

Figure 1: Habitats of *E. coli*.This diagram represents the possible habitat transitions that *E. coli* can go through. It is expected that *E. coli* cell are "born" in an intestine and spend half its life there. They are then excreted onto the earth's surface where they spend the second half of there life.

Previous work

Once *E. coli* is colonised in new borne's a few days after they are borne it is very hard for other strains to compete for colonisation((Sears *et al.*, 1956). This highlights that the intestinal tract is a competitive environment that pathogens have to overcome (Savageau, 1983).

Previous work is skewed largely towards the medical relevance of *E. coli* and pathogens. Pathogenic *E. coli* cause over 160 million cases of dysentery and two million deaths globally per year through both intra-intestinal and extra-intestinal diseases (Wirth *et al.*, 2006), whereas non-pathogenic *E. coli* constitute part of the normal intestinal flora of healthy mammals and birds.

These pathogens draw the attention of governments industry and media. Recent work into the epidemiological factors has found that there are certain factors that can promote a strains fitness and virulence. Genes like alpha-hemolysin, P-fimbrial adhesins, and the Semirough lipopolysaccharide can be considered virulence genes (VG). Genes like microcins, different iron uptake systems, adhesins, and proteases can be considered fitness genes (Grozdanov *et al.*, 2004). The probiotic strain Nissle 1917 is used in the treatment of many intestinal related diseases and its though that the strain having a lack of these virulence genes and presence of a lot of these fitness genes makes it so good at colonising in the gut. *E. coli* pathogens show clustering of these VGs on pathogenicity islands (PIA). The clustering of the fitness genes also occurs on genomic islands (Grozdanov *et al.*, 2004)

Due to this large flux in the genome of *E. coli* they display an ability in evolution unrelated to eukaryotes. In addition the the modifications of the *E. coli* genome through mutations and reordering of genes they have the ability for these islands to go through horizontal gene transfer between different populations (Ochman *et al.*, 2000). Horizontal gene transfer methods can include the clustering of genes onto a plasmid and migration to other bacteria. These plasmids often include tRNA genes which gives them targets for integration into a chromosome (Doolittle, 2002).This was first demonstrated to be involved in a pathogenic ability in the 1980s when an *E. coli* enterobacteria showed the ability to alter is sucrose fermentation pathway by means of a conjugative transposon (Hochhut *et al.*, 1997:94). This gives *E. coli* an extremely dynamic genomes in which substantial amounts of DNA are introduced into and deleted from the chromosome. This also enables *E. coli* to change there characteristics from commensal and pathogenic with the acquisition of virulence gene and functional loss of native genes (Ochman *et al.*, 2000).

The presence of identical genes in pathogenic and nonpathogenic variants of one species indicates that some of their encoded factors (such as adhesins, iron uptake systems, or proteases) contribute to general adaptability, fitness, and competitiveness rather than to particular virulence trait .

The initial work into the genetic diversity of *E. coli* was done by roger milkman (Hille and Berger, 2011). This work has since been expanded into the human flora responsible for medical illnesses and has started to expand into the investigation of strains in secondary habitats (Touchon *et al.*, 2009). The investigation into primary and secondary habitats has formed a basis for a clonal paradigm of the genetic structure of bacterial populations (Souza *et al.*, 1999). It is observed by (Souza *et al.*, 1999) that genetic diversity among *E. coli* isolated from mammalians, reptiles and wild bird populations is not randomly distributed and that the *E. coli* found in these secondary and primary habitats have developed a specificity for hosts. It can be concluded from the work of (Langridge *et al.*, 2015) that bacterial species can indeed modify there genomes to a degree as to specify pathogenicity in only one or a few hosts. These “host adaptations” include the acquisition of pathogenicity islands comprised of multiple genes involved in disease.

It was also shown by “” that Pathogens can occur not just through the acquisition of virulence genes but also through the loss of genes. These were nicknamed the blackholes of the genome, and opened up the possibility for a new evolutionary pathway of bacteria (Maurelli *et al.*, 1998).

The need for gene loss comes from a mechanical property of genome size. The bacterial genome size isn't increasing throughout time and shows a deletion bias is a major force shaping the genome (Maurelli *et al.*, 1998). This deletion bias has been speculated to enhance the pathogenic ability of uropathogenic *E. coli* by showing that losses correlate with prolonged persistence of infections (Hacker *et al.*, 2003). A study into the pathogenic chromosome of *E. coli* showed that the species can also retain genes and alter there activation state by internal genomic shuffling in reference to promotor regions. This can also show the lack of gene expression due to phase variation slipping and mispairing the nucleotide sequence with the promotors (Hacker *et al.*, 2003).

Although *E. coli* have shown the ability to adapt to a specific environment it is hypothesised by (Arber, 2000) that there evolution is indirect within populations. This leads to the realisation that in spite of all the new gene combinations, mutations and reordering of the genomes, they are characterised as second order selection. The genes that are associated with evolution are not programatically directing evolution towards a specify goal. Rather, a steady interplay between natural selection and mixed populations of genetic variants gives microbial evolution the darwinian existence of survival of the fittest (Arber, 2000).

Although *E. coli* appear to have an environmental adaptability there is not always evidence to suggest that there numbers and level of virulence are determinant on a specific factor of that environment as shown by the work of (Jones *et al.*, 2014).

The ability of food borne bacteria has to deal with very harsh environments compared to host environments. Therefor they must have developed systems that enable them to endeavour in that environment if they are to survive as pathogens otherwise they may die (Hacker and Carniel, 2001).

A darwinian approach would state that *E. coli's* evolution is driven by the overall need to increase the fitness of the species. This evolutional strategy is seen in pathogenic *E. coli* where the acquisition and loss of genetic material is the result of them needing to survive an environment (Hentschel and Hacker, 2001).

The presence of pathogens in endotherms is a representation of increased fitness as is the ability of *E. coli* to be food born pathogens. The duel function of factors like iron uptake systems or adherence can be found on genomic islands in commensals although this harmless intestinal commensal can be considered uropathogenic and demonstrate fitness in a specific environment (Hacker and Carniel, 2001).

AIM

The research present on *E. coli* is skewed towards the needs of environmental and public health which has left a lack of knowledge about commensal strains. A better understanding of the commensal niche is necessary to understand how a useful commensal can become a harmful pathogen (Tenaillon *et al*., 2010). It is estimated that half of the *E. coli* population lives in there secondary habitats (Savageau, 1983) and warrants the investigation into how there growth may be different to commercial *E. coli.* There is even evidence to show that these habitats can support growth of particular stains that can feed on organic matter and that this is a temperature dependent factor (Power *et al*., 2005; Solo-Gabriele *et al*., 2000).
It is the aim of this study to investigate the response of *E. coli* strains isolated from both hosts and non-hosts to observe the potential difference in growing abilities.

Materials and Methods

Table 1: *E. coli* strains used to compare genomic and phenotypic differences

Strain name	Source	Phylogroup	Group
ECOR-03	Host (Dog)	A	HOST
ECOR-16	Host (Leopard)	A	HOST
ECOR-31	Host (Leopard)	E	HOST
ECOR-32	Host (Giraffe)	B1	HOST
ECOR-46	Host (Ape)	D	HOST
ECOR-47	Host (Sheep)	D	HOST
ECOR-65	Host (Ape)	B2	HOST
ECOR-66	Host (Ape)	B2	HOST
ECOR-68	Host (Giraffe)	B1	HOST
SS_203	Host (other)	B2	HOST
SS_227	Host (bird)	A	HOST
SS_229	Host (bird)	A	HOST
SS_236	Host (Bovine)	D	HOST
SS_245	Host (Bovine)	D	HOST
SS_261	Host (human)	D	HOST
SS_271	Host (human)	B2	HOST
SS_272	Host (human)	E	HOST
SS_293	Host (ruminant)	B1	HOST
SS_297	Host (ruminant)	B1	HOST
SS_304	Host (bird)	D	HOST

Strain name	Source	Phylogroup	Group
GMB02	Nonhost (plants)	A	NONHOST
GMB04	Nonhost (plants)	D	NONHOST
GMB06	Nonhost (plants)	A	NONHOST
GMB101	Nonhost (plants)	B2	NONHOST
GMB104	Nonhost (plants)	A	NONHOST
GMB15	Nonhost (plants)	D	NONHOST
GMB21	Nonhost (plants)	D	NONHOST
GMB34	Nonhost (plants)	A	NONHOST
GMB38	Nonhost (plants)	B1	NONHOST
GMB41	Nonhost (plants)	B1	NONHOST
GMB44	Nonhost (plants)	D	NONHOST
GMB45	Nonhost (plants)	B2	NONHOST
GMB47	Nonhost (plants)	B2	NONHOST
GMB50	Nonhost (plants)	D	NONHOST
GMB52	Nonhost (plants)	D	NONHOST
GMB77	Nonhost (plants)	E	NONHOST
GMB78	Nonhost (plants)	E	NONHOST
GMB83	Nonhost (soil)	B1	NONHOST
GMB89	Nonhost (plants)	B1	NONHOST
GMB93	Nonhost (plants)	B2	NONHOST

Phylogenetic Analysis

All *E. coli* strains used in this study are shown in table 1. During this study *E. coli* master stocks were moved from -80 freezer to -20 freezer and kept in a 50% v/v glycerol stock. Cells were recovered and cultured on to TBX agar plates (Oxoid, UK) and incubated for a period of 24h at 37C. A single bacterial colony was then harvested from the agar plate and resuspended in TSB inoculation media (Oxoid, UK). The cells were grown in inoculation media in a shaking plate incubator (200 RPM) to inhibit biofilm formation for 24h at 37C. After 24h the cells optical density (OD) is measured in a 10% v/v PBS solution (OD at 600nm). Each of the 40 strains from the inoculation media has 2 x 1.5uL inoculated into separate 500uL wells on a 96 well plate (Thermo Sci Nunc, UK). The cells are inoculated at a dilution factor of 1:100. The plate is immediately placed into a variable atmospheric plate reader (OMEGA FluroStar) in an carefully controlled environment. Oxygen concentrations are controlled with nitrogen gas and CO2 levels are controlled with CO2. The OMEGA maintains temperature oxygen and carbon dioxide concentrations continuously. The environment of the OMEGA is set to one of the specified conditions detailed below. The cells growth is measured every 30mins over a 24h period with a 10 second multidirectional oscillation (200 RPM) before every reading is taken. Growth of the cells was observed with the lag phase finishing on average within the first half an hour and exponential ending between an average of 4-6 hours. The cells go through stationary growth for the next 18-20 hours. Aerobic conditions were controlled by gas permeable moister barrier seals (4titude, UK) placed over the 96 well plates.

Each strain is grown in total of six times as each 96well plate will accommodate 2 technical replicates. This plate is then repeated 3 times totalling 6 individual growths of the same strain. These six replicates were repeated four times in four different environmental conditions:

- At 37C with 5% oxygen
- At 37C with atmospheric oxygen
- At 25C with 5% oxygen
- At 25C with atmospheric oxygen

These four conditions were chosen to mimics the natural environment in which the 40 strains were isolated. 37C 5% oxygen mimics the environment found on the intestinal tract (primary habitat), 37C with atmospheric oxygen mimics the environment on the surface of skin. 25C 5% oxygen mimics the environment found in water or sediments and 25C atmospheric oxygen mimics the environment found on plant leafs.

Spectrometer Analysis

The spectrophotometer was initially used to form a calibration curve by pairing optical density (OD) at 600nm against the number of colonies formed (CFU) when 1ml of the liquid solution was smeared on an LB agar plates. The trend line generated from the calibration curve was used to predict the concentrations of the samples OD readings every week thereafter. This was used to obtain the initial concentration of the bacterial inoculation media before phenotype analysis in the Omega was carried out. It was necessary to make sure that all phenotype results were carried out with accuracy and consistency. Concentration readings estimated from the calibration curve were tested with a one way anova test to check consistency between the 40 strains and within each of the 6 repeat inoculations of a individual strain.

Spectrophotometric readings were taken each week for two reasons 1) To check for contamination 2) To check bacteria have grown well enough for phenotype analysis. Contamination was check during every stage of incubation with the addition of blanks but it was observed that contamination of one or a few strain could occurred that did not affect the blanks. The spectrophotometric analysis was used to eliminate any obvious abnormal growths in the liquid media that were either too high or too low.

Media Mixtures and Master Stocks

- TSB: The TSB media (Oxoid, UK) was used as an inoculation media to grow the bacteria in liquid culture so that they could be transferred to the 96well plates and so that Spectrophotometer quality control could be carried out to check for contamination. TSB media came in powder form and was made into a liquid media in house. The powder was measured out on a analytical Laboratory 4 decimal place balance and mixed with purified water from a type 3 reverse osmosis (RO) water purification system

(Merek Millipore, UK). The media was then autoclaved at 121C for 15mins. The media is then kept in sealed containers until needed where it is used under sterile conditions with a bunsen burner.

- TBX: TBX media was also prepared in house as above with the same balance, water and autoclaving process. The media was then poured into petri dishes under a positive pressure clean hood and stored in a 4C fridge.
- MASTER STOCKS: The master stocks of the *E. coli* strains were made by streaking the strains onto pre made LB agar petri dishes (Oxoid, UK). These strains were grown for 24h in an incubator at 37C. The grown colonies were then collected form the plate and placed into 50% V/V glycerol stocks and placed in to a -20 freezer.

Bioinformatics

40 *E. coli* strains were randomly selected from the BIGSdb multi species database at Swansea University. 20 host isolates from a variety of geographical locations were selected. 35% of these host strains came from zoo animals (7/20), 25% from livestock (5/20), 0.05% from dogs (1/20), 0.05% from sheep (1/20), 15% from wild birds (3/20) and 15% from humans (3/20). 20 non-host strains were also selected. 90% of these strains were isolated from agricultural plants in the UK (18/20) 0.5% from agricultural plants in Italy (1/20) and 0.5% was a soil sample from the UK (1/20).

The genomes for these 40 strains were then collected from the BIGSdb database and annotated using the RAST annotation service (Aziz *et al*., 2008). These annotated strains were then used to compile a super-genome using the method detailed by (Méric *et al*., 2014). The BIGSdb genome comparator BLAST function was then used to determine a presence and absence matrix for genes in each strain in reference to the super-genome. The BIGSdb web server was then used to create a concatenated alignment of the genes with MUSCLE (Edgar, 2004). The cutoffs for what constitutes a homolog was >70% similarity in nucleotide sequence and >50% sequence length similarity. Missing nucleotides and or genes were replaced with empty spaces. The algorithmic program RapidNJ was then used to determine the neighbour joining relationships between the genomes. This program uses a Tajima's D statistical testing method (Korneliussen *et al*., 2013) to construct the relatedness between strains based on whole genome analysis and gene presence absence. This data was then input into the Mega 6 program to establish a neighbour joining tree and cladogram as seen in figures 4 and 5. The Seed Viewer Web-server (http://www.theseed.org/wiki/Main_Page) was used to compare functional genome proportions and to identify the unique genes.

Comparison of Phenotype Data

The comparison between growth curves for all 40 strains was performed on the data analysis software Graph Pad Prism. Graphing of the averages between the 20 host and 20 non-host strains produced a graph that could be used to determine growth difference and growth similarities. Graphing of the average growths of all 40 strains in all conditions revealed the differences in growth and was used to establish groups of strains that showed preference for a particular condition. This preference was arbitrarily determined by strains meeting the following rule; Present in the top third percentile in one condition and in the bottom third percentile of another condition.

Spectrograph Analysis

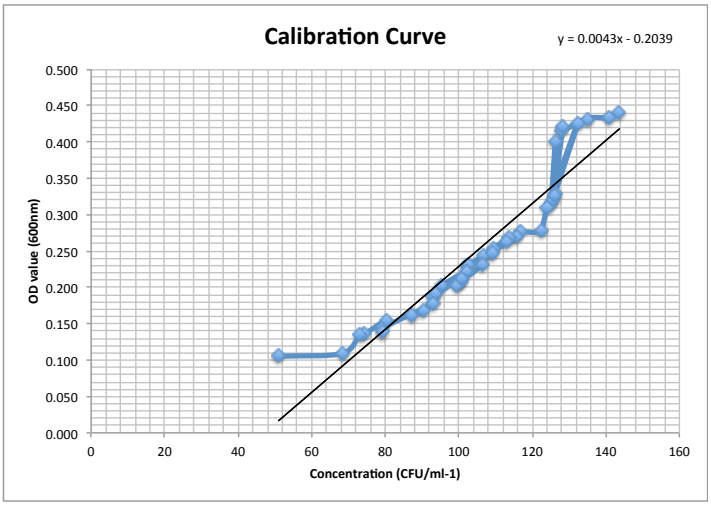

Figure 2: Calibration curve. Absorbance values were taken for each strain in triplicate after a single colony had grown in the liquid growth media for 24h. It was observed that there is a correlation between optical density and the number of colonies that formed on an LB agar plate. The correlation between absorbance and concentration gave a trend-line with an equation that can be used to determine the concentration of other growth media without having to grow and count colonies. Concentrations were labeled as colony forming units (CFUl) per ml of liquid culture.
The results of the spectrophotometer and CFU count developed a good consistent growth curve of absorbance Vs CFU. This graph was then used to generate the concentrations (CFU ml-1) for each strain each week which were then used to test the consistency of the liquid growth media.

A one way ANOVA was employed to test the consistency between the 40 different strains and between the individual strains growth each week. This developed an F value of 0.8713 which fell below the F critical value of 1.4273. The P value of 0.695 was also above the set alpha value of 0.05. Both of these show that the null hypothesis “there is equal concentrations between the final growth volumes” is true which confirms that there is consistency in the bacterial growth method. This method also allowed the removal of obvious abnormal growth results and eliminated 30% (5/17) of all growth cultures from the analysis.

Core and Pan genomes

Having gathered the 40 genomes for the *E. coli* we first decided to identify the core and pan genomes, i.e. genes present among all the genomes and the orthologs present in each genome. In the comparison of the 40 genomes it was found that the average genome size is 4688 genes. In addition the average core genome size was 3350 and the average reference genome size was 1422. Comparison of the genomes revealed that any one strain contains only 47% of the reference pan genome suggesting that no single strain is fully representative of the reference pan genome. The comparison between the core and reference genomes showed a very close resemblance between host and non host and summarised the understanding that if any genetic difference was to be found to account for a change in growth phenotype, it would be a small genetic variation. However, it is interesting to see that the core and Pan genome sizes do not reach a clear plateau, even when 20 genomes are sampled. This shows the diversity of the genomes as the more genomes are added the the lower the core genome size becomes and the higher the pan genome size becomes (Rasko *et al.*, 2008).

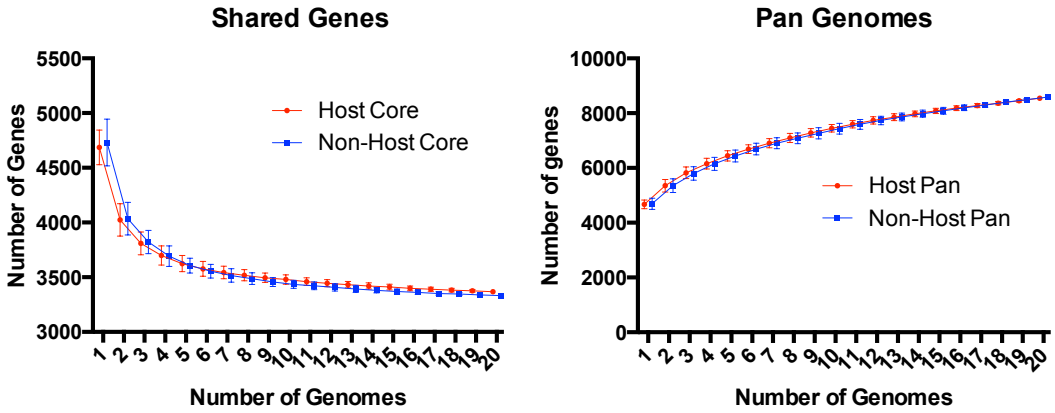

Figure 3:
Accumulation curve. The number of shared genes represents the difference in gene volume between the host and non-host strains. The Pan genomes show how the number of genes increases as you include more strains indicating the reference genome is highly variable. Comparisons are made based on a matrices of gene presence/absence derived from the reference pan genome. The method of genome sampling was randomised and carried out 100 times to obtain the average number of genes for each sample. At 20 sequenced genomes, the average core-genome had 3350 genes (33.7% of the pan-genome).

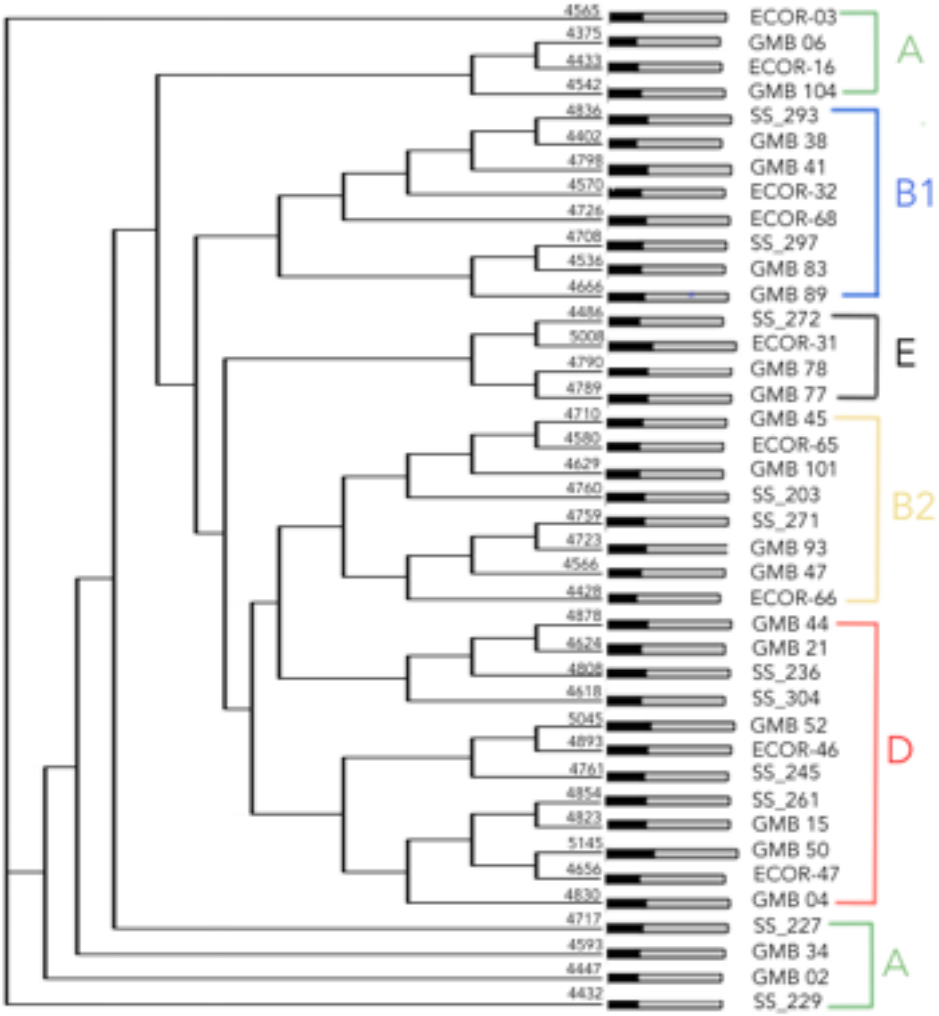

Figure 4: Cladogram. The cladogram shows the relatedness of all 40 strains and grouping of the strains into there respective phylogenetic linages (represented by the vertical coloured lines). Each of the 40 strain exhibited a variable reference genome size as indicated by the black regions in the percentage bars. The core genomes of the strains represent an average of 69% of the genome as can be seen by the grey regions of the percentage bars. The total genome size for each strain is label at the ends of the trees branches showing that genome size is fairly conserved.

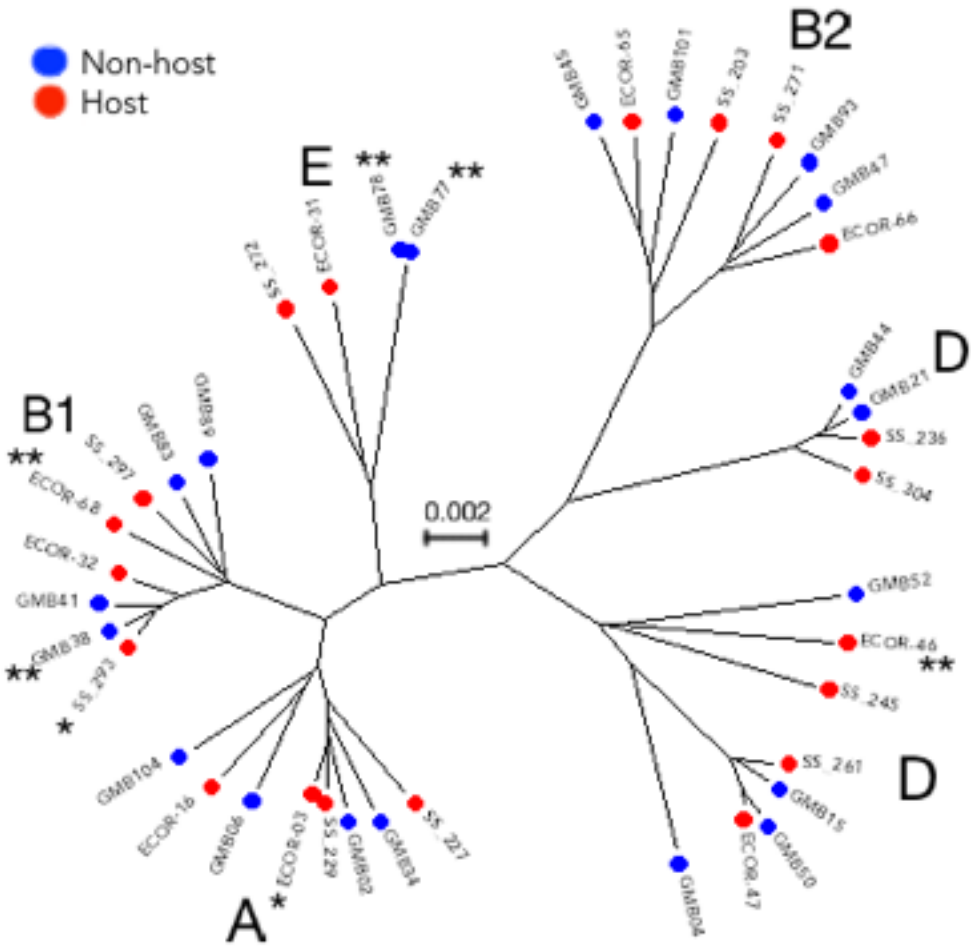

Figure 5: Neighbour Joining Tree. Maximum likelihood tree of 20 genomes belonging to *E. coli* host sources shown in red and 20 non-host sources shown in blue. The scale bar indicates the estimated number of substitutions per gene site. The tree is separated into 5 groups (A, B1, B2, D and E) that represent the main linages that the species is separated into. The genes of the 40 strains were aligned on a gene-by-gene basis using MUSCLE (Edgar, 2004) and then linked into a contiguous sequence. This data was then used to construct neighbour joining tree in Mega 6.

* Strains that showed a growth preference at 37C 5% Oxygen.

** Strains that showed a growth preference to 25C Atmospheric oxygen.

Phenotype Analysis Between Host and Non-host

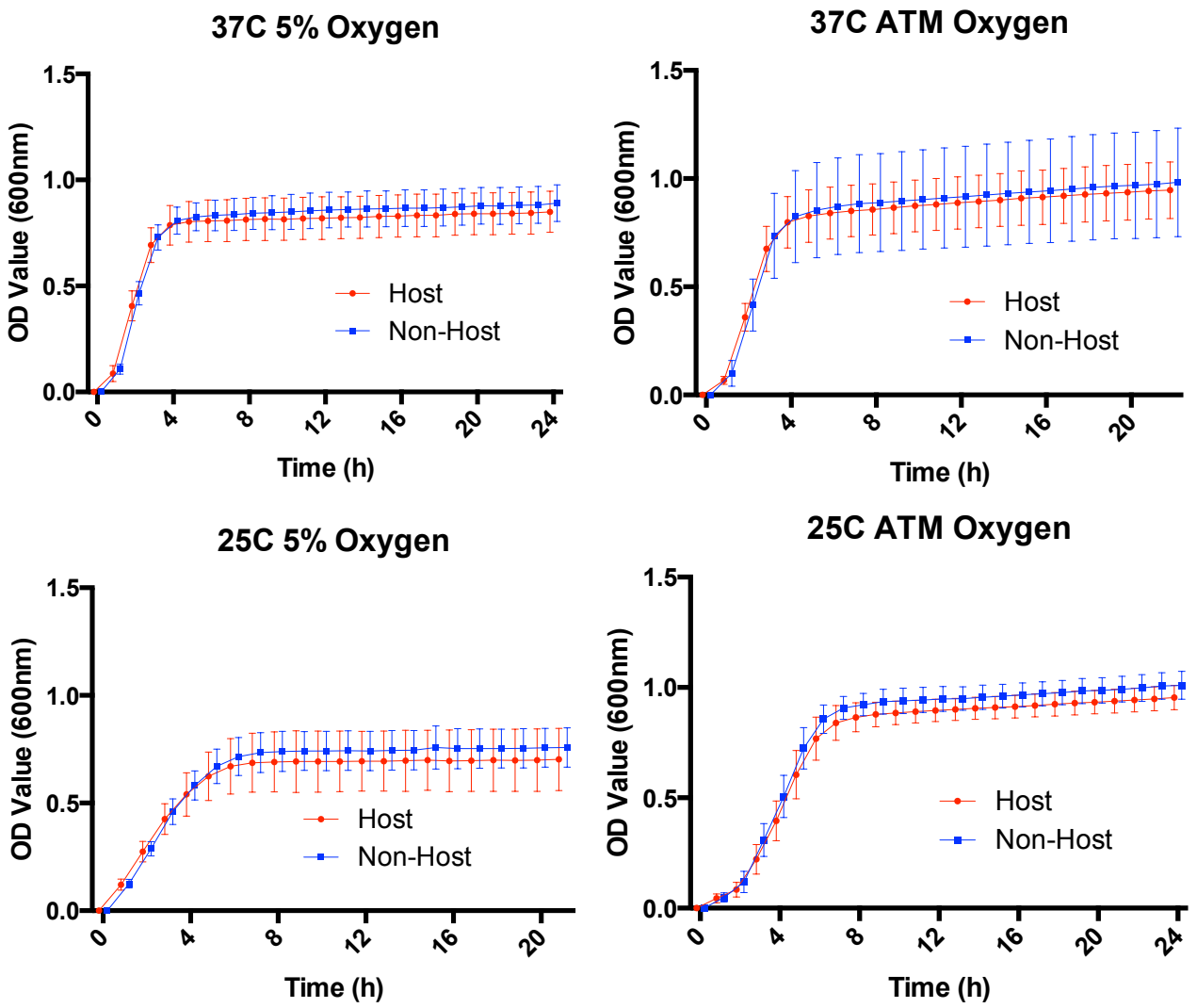

Figure 6: Comparison of growth curves. 40 strains of *E. coli* (20 host and 20 non-host) were grown in a variable atmospheric microplate reader. Cells were grown in four conditions with temperature and oxygen concentration variations. Measurements were taken every 30 mins for 24 hours. Each graph shows the concurrent lag, growth and stationary phase of the growth cycle.

There are three variables involved in the analysis of the growth curves, 1) The growth time (GT) also known as the log phase 2) the Lag phase duration (LPD) 3) the maximum population density (MPD). The phenotype analysis between the four conditions showed that there is a general depression in MPD under anaerobic conditions, absorbance is an average 0.17 units lower. The MPD is an average of 0.06 CFUs higher at 37C compared 25C, the lag phase is and average of 50mins shorter at 37C compared to 25C and the GT is an average of 2h 43mins shorter at 37C compared to 25C.

Phenotype Analysis Between All Strains

Due to the poor comparative results between host and non-host groups it was decided to narrow the criteria of all strains to see if we could find strains with a preference for a particular condition. The new criteria were set to look at the differences between conditions that mimic a human intestinal environment and mimics of an exposed plant leaf environment (37C 5% oxygen vs 25C Atmospheric Oxygen). The collection of growth curves for all 40 strains were graphed and split into thirds denoting high, medium or low levels of growth in that condition.

It was found that 2/40 strains were present in the high growth percentile at 37C 5% oxygen and also present in the low growth percentile at 25C Atmospheric Oxygen. There were also 5/40 strains present in the high growth percentile at 25C Atmospheric Oxygen and also present in the low growth percentile at 37C 5% oxygen. The averages of the strains growth in each condition are shown graphically in figure 7. The graphs

are a representation of there growths in each of the two conditions (37C 5% oxygen vs 25C Atmospheric Oxygen).

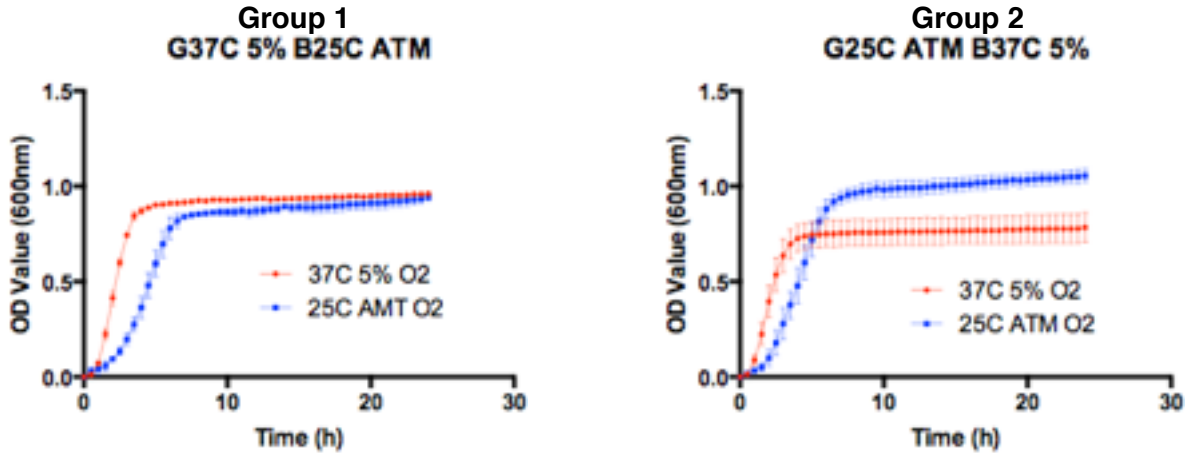

Figure 7: Comparison of growth curves. 40 strains of *E. coli* were grown in a variable atmospheric microplate reader and there growth rates used to identify two groups of cell. The two groups are comprised of strains that show preference for a particular growing condition dependent on a set temperature and oxygen level combination. Measurements were taken every 30 mins for 24 hours. Each graph shows the concurrent lag, growth and stationary phase of the growth cycles.

Both graphs show that the LPD at 25C is longer than at 37C but that GT and MPD are dependent on there groups. In the graph on the left the GT is less and MPD is higher at 37C 5% O2. In the graph on the right the GT is longer but MPD is higher at 25C ATM O2.

Variation in Functional Genes

(25/40) 62% of strains show grater than 92% similarity between there genomes and from the literature it can be concluded that these strains will most likely share the same pathogenic and virulence activity (Ishii *et al.*, 2006). The highest similarity observed between two genomes was 99.8% where they can almost be considered the same strain. The average gene similarity between the two new groups was 87% showing a 13% difference between the genomes which includes some unique genes shown in the figure 8.
By analysing the difference in functional gene categories and identifying unique genes it is possible to draw some conclusions as to why there are differences in the growth rates of these two groups. Analysing the difference between the two functional groups alone shows there is significant difference between the two groups (Chi^2 = 64.01, d.f. = 25; p= 0.000028). The proportion of genes present in the groups 1 and 2 where different with group 2 showing higher proportions of iron uptake systems, Cell wall and capsule and membrane transporters.

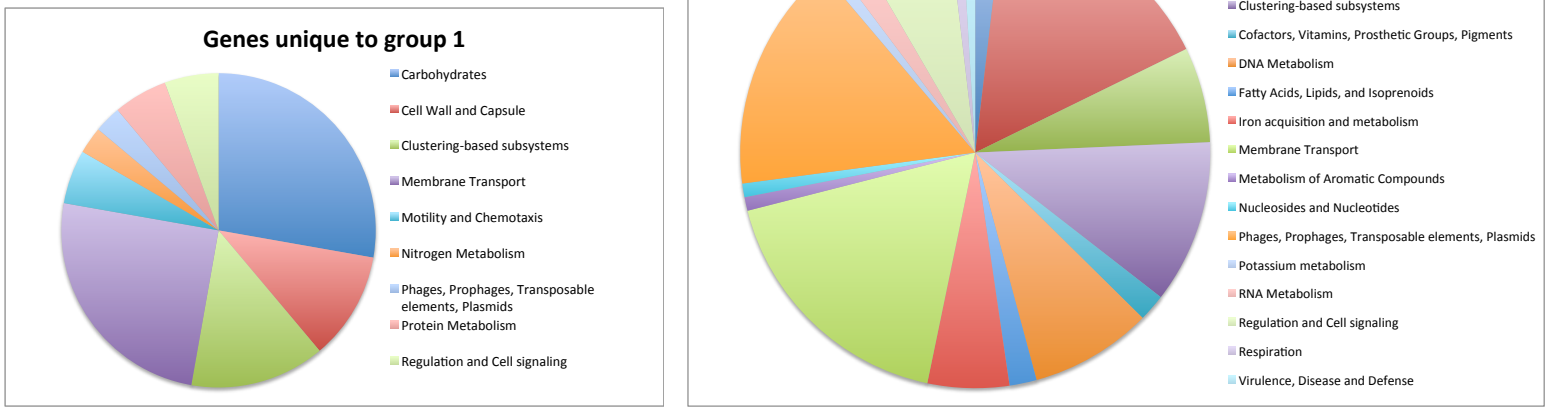

Figure 8: Comparison of unique genes. There are 36 genes unique to group1 and 107 genes unique to group 2. These pie charts represent the relative percentage each type of gene within the unique gene set.

It can be seen in group two there is a presence of unique genes associated with virulence and and defence that are not seen in group 1. This finding is consistent with the hypothesis that strains adapted to less favourable secondary habitats are more closely related to pathogens and display more virulence genes than commensal bacteria.

Table 2: Functional gene comparison. Using the RAST seed web server a functional comparison of genes present within the genomes of groups 1 and 2 shows that there is a significant difference in functional gene proportion (Chi^2 = 64.01, d.f. = 25; p= 0.000028). RAST was able to find gene functions for 60% of the genome of group 1 and 57% of group 2.

Functional Categories	G37C 5% O2 B25C ATM O2	G25C ATM O2 B37C 5% O2
Carbohydrates	813 (18.71%)	851 (18.24%)
Amino Acids and Derivatives	409 (9.41%)	432 (9.26%)
Membrane Transport	325 (7.48%)	361 (7.74%)
Cell Wall and Capsule	296 (6.81%)	312 (6.69%)
Cofactors, Vitamins, Prosthetic Groups, Pigments	288 (6.63%)	304 (6.52%)
Protein Metabolism	273 (6.28%)	267 (5.72%)
RNA Metabolism	248 (5.71%)	252 (5.7%)
Stress Response	192 (4.42%)	197 (4.22%)
Respiration	189 (4.35%)	190 (4.07%)
Regulation and Cell signaling	169 (3.89%)	182 (3.9%)
Nucleosides and Nucleotides	152 (3.5%)	154 (3.3%)
Fatty Acids, Lipids, and Isoprenoids	136 (3.13%)	147 (3.15%)
DNA Metabolism	135 (3.11%)	227 (4.87%)
Motility and Chemotaxis	126 (2.9%)	83 (1.78%)
Virulence, Disease and Defense	117 (2.69%)	115 (2.47%)
Nitrogen Metabolism	79 (1.82%)	77 (1.65%)
Phages, Prophages, Transposable elements, Plasmids	72 (1.66%)	154 (3.30%)
Miscellaneous	64 (1.47%)	64 (1.37%)
Sulfur Metabolism	57 (1.31%)	57 (1.22%)
Phosphorus Metabolism	54 (1.24%)	53 (1.14%)
Cell Division and Cell Cycle	41 (0.94%)	44 (0.94%)
Metabolism of Aromatic Compounds	30 (0.69%)	38 (0.81%)
Potassium metabolism	28 (0.64%)	29 (0.62%)
Secondary Metabolism	26 (0.60%)	27 (0.58%)
Iron acquisition and metabolism	22 (0.51%)	43 (0.92%)
Dormancy and Sporulation	5 (0.12%)	5 (0.11%)

Gompertz comparison

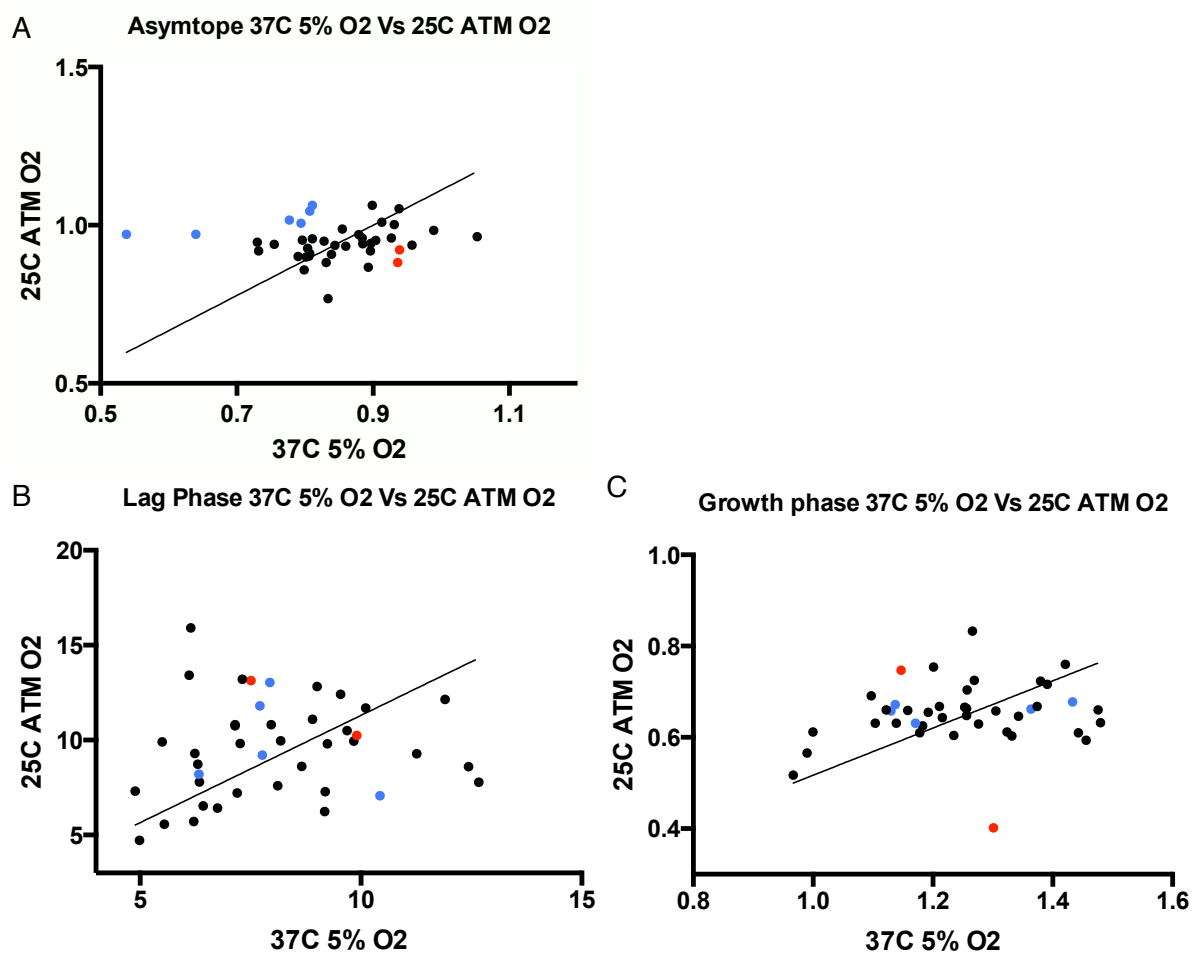

Figure 9: Comparison of Gompertz parameters. Graphs A, B and C show the comparison of asymptotes, Lag phase and Growth phase respectively predicted by the Gompertz model. These values were predicted using a nonlinear regression analysis in SPSS using the Gompertz model formula to generate predictions of the three parameters of the growth curves. The Trend-line through the centre of each plot is used as a guide to show the point at which a strain will behave equal well in each condition.

I used nonlinear regression in SPSS to estimate the parameters of a Gompertz model for each strain (Zwietering *et al.*, 1990). Focusing on the two new groups I chose to compare strains grown in the 37C 5% oxygen condition and pair it against the 25C Atmospheric oxygen condition. This data can be seen in figure 9. From the comparisons of MPD (or the asymptote as it is seen on the graph) we can see that there is some difference between the strains found in group 1 compared to group 2. From looking at the axis of these graphs it is still maintained that growth rate is higher at 37C compared to 25C but there is little difference in terms of lag phase duration. There was no groups association to be draw from the comparisons of Lag phase as the groups are intersperse throughout the plot.

Discussion

The process of understanding the genetic difference between pathogens and there relation to commensals is becoming more active and is narrowing the definitions of these two groups. The acquisition of only a few genes is necessary to convert a harmless comensial into a deadly pathogen.

The fact that there are a small number of strains that showed specificity for certain conditions is validated by similar findings in other studies. Drawing attention to the findings of (Ishii *et al.*, 2006) who showed cells have a genome alterable to preferred growing in soils and Anastasi *et al.*, 2012; Edberg *et al.*, 2000) showing the

validity in water further influences the findings that there are 5 specific strains that preferred growing in conditions that mimicked this environment.
The finding that two strains in group 1 showed preference for endotherm intestinal mimicked conditions and the fact that they showed loss of genes compared to groups 2 highlights the hypothesis that genome losses or black holes can contribute to host specification.

The fact that only a small number of the strains isolated from host or non-host sources showed these condition preferences is indicative that populations of strains dont grow towards a specific goal and that the genes responsible for the specificity are not largely popular throughout the genomes of these strains. Further analysis on strains isolated from the same animals or animals in close proximity, could be fed specific strains or put under conditions that are though to promote host specific strains. This sort of study could help highlight how these gene losses and acquisitions may spread throughout the population once they have shown a competitive ability in a few animals.

It is clear from the work of (Anastasi *et al*., 2012) that strains surrounding sewage treatment plants have a competitive ability to 1) survive the treatment process of STPs 2) to colonise and grow in the surrounding waters and soils. This is clearly due to specific gene losses and acquisitions associated with pathogenicity. It is interesting to see that the presence of unique genes in the group 2 strains are associated with virulence as the most common strains found in non-host environments have enhanced virulence ability due to the presence and loss of these particular genes.

During the study the cells were removed from the freezer and placed on dry ice. Despite maintaining a cold environment theres assumed to fluctuations in the temperatures of the *E. coli*. It has been shown that temperature fluctuations experienced by chilled foods during defrosting cycles of retail display cases significantly affects the bacterial growth of *E. coli* (Jones *et al*., 2004). This hypothesis may have merit to the observed fluctuations in bacterial growth in this study as the *E. coli* will go through possible temperature changes during preparation of the strains. Summarised in figure 10.

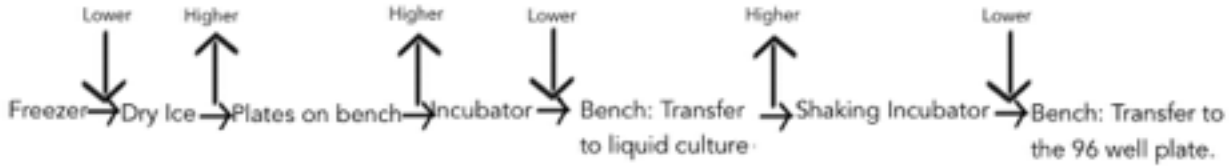

Figure 10: Temperature changes. There a 6 temperature changes predicted through the path of freezer to OMEGA. These temperature changes are small but incubation steps would increase temperature by about 17C from the room temperature.

Even though there was consistency between the strains growth and phylogenetic results it is still from a small sample set. Therefor although there is correlation with other studies they are not validated or proving of virulence/fitness gene hypothesise. It is certainly worth remembering this (and not just for this study) when talking about the validity of results. The study (Ishii *et al*., 2006) showed that *E. coli* incubated in soils increased growth when the temperature was raised from 15C to 37C. This shows that soilborn *E. coli* can persists in an environment for extended periods of time but are ready to grow when favourable conditions occurs (Jones *et al*., 2004).

Temperature and oxygen concentration were used as my two variable factors in the analysis, but its clear from the literature that there are many factors that help shape the genomes of the host and environmental strains. More studies should be carried out into the nature of commensal to pathogenic pathways. Future studies should focus on *E. coli* in a variety of host and a much larger selection of environmental sources. Due to the fact that there are a lot of hypothetical proteins uncovered during annotation investigations into the molecular understanding of *E. coli* could be used to guide future studies towards commensal niches.

Acknowledgements

Sheppard Laboratory: Sam Sheppard, Guillaume Méric, Ben Pascoe, Susan Murray, Leonardos Mageiros, Tom Humphrey, Daniel Falush, Tom Wilkinson, Angharad Davies, LLinos Harris, Jane Mikhail.

- A special thank you to Professor Michael Gravenor for all his help with Gompertz modelling and statistical comparisons.

- A special Thank you to Jessica Rees for her support and warmth throughout the study.

Reference

Anastasi, E.M., Matthews, B., Stratton, H.M., and Katouli, M. (2012) Pathogenic Escherichia coli Found in Sewage Treatment Plants and Environmental Waters. *Appl Environ Microbiol* **78**: 5536–5541.

Arber, W. (2000) Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol Rev* **24**: 1–7.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., *et al*. (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75.

Byappanahalli, M.N., Whitman, R.L., Shively, D.A., Sadowsky, M.J., and Ishii, S. (2006) Population structure, persistence, and seasonality of autochthonous Escherichia coli in temperate, coastal forest soil from a Great Lakes watershed. *Environ Microbiol* **8**: 504–513.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., *et al*. (2003) Analysis of Genome Plasticity in Pathogenic and Commensal Escherichia coli Isolates by Use of DNA Arrays. *J Bacteriol* **185**: 1831–1840.

Doolittle, R.F. (2002) Biodiversity: Microbial genomes multiply. *Nature* **416**: 697–700.

Edberg, Sc., Rice, E.W., Karlin, R.J., and Allen, M.J. (2000) Escherichia coli: the best biological drinking water indicator for public health protection. *J Appl Microbiol* **88**: 106S–116S.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Grozdanov, L., Raasch, C., Schulze, J., Sonnenborn, U., Gottschalk, G., Hacker, J., and Dobrindt, U. (2004) Analysis of the Genome Structure of the Nonpathogenic Probiotic Escherichia coli Strain Nissle 1917. *J Bacteriol* **186**: 5432–5441.

Hacker, J., and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. *EMBO Rep* **2**: 376–381.

Hacker, J., Hentschel, U., and Dobrindt, U. (2003) Prokaryotic chromosomes and disease. *Science* **301**: 790–793.

Hentschel, U., and Hacker, J. (2001) Pathogenicity islands: the tip of the iceberg. *Microbes Infect* **3**: 545–548.

Hille, B., and Berger, E. (2011) Roger Dawson Milkman (1930-2011) Geneticist. *Genetics* **188**: 489–490.

Hochhut, B., Jahreis, K., Lengeler, J.W., and Schmid, K. (1997) CTnscr94, a conjugative transposon found in enterobacteria. *J Bacteriol* **179**: 2097–2102.

Ishii, S., Ksoll, W.B., Hicks, R.E., and Sadowsky, M.J. (2006) Presence and Growth of Naturalized Escherichia coli in Temperate Soils from Lake Superior Watersheds. *Appl Environ Microbiol* **72**: 612–621.

Jones, L.A., Worobo, R.W., and Smart, C.D. (2014) Plant-Pathogenic Oomycetes, Escherichia coli Strains, and Salmonella spp. Frequently Found in Surface Water Used for Irrigation of Fruit and Vegetable Crops in New York State. *Appl Environ Microbiol* **80**: 4814–4820.

Jones, T., Gill, C.O., and McMullen, L.M. (2004) The behaviour of log phase Escherichia coli at temperatures that fluctuate about the minimum for growth. *Lett Appl Microbiol* **39**: 296–300.

Korneliussen, T.S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* **14**: 289.

Langridge, G.C., Fookes, M., Connor, T.R., Feltwell, T., Feasey, N., Parsons, B.N., *et al*. (2015) Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci* **112**: 863–868.

Maurelli, A.T., Fernández, R.E., Bloch, C.A., Rode, C.K., and Fasano, A. (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. *Proc Natl Acad Sci* **95**: 3943–3948.

Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M.C.J., Jolley, K.A., and Sheppard, S.K. (2014) A Reference Pan-Genome Approach to Comparative Bacterial Genomics: Identification of Novel Epidemiological Markers in Pathogenic Campylobacter. *PLoS ONE* **9**: e92798.

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

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., and Slade, M.B. (2005) Phenotypic and genotypic characterization of encapsulated Escherichia coli isolated from blooms in two Australian lakes. *Environ Microbiol* **7**: 631–640.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., 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.

Savageau, M.A. (1983) Escherichia coli habitats, cell types, and molecular mechanisms of gene control. *Am Nat* 732–744.

Sears, H.J., Janes, H., Saloum, R., Brownlee, I., and Lamoreaux, L.F. (1956) Persistence of individual strains of Escherichia coli in man and dog under varying conditions. *J Bacteriol* **71**: 370.

Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R., and Palmer, C.J. (2000) Sources of Escherichia coli in a coastal subtropical environment. *Appl Environ Microbiol* **66**: 230–237.

Souza, V., Rocha, M., Valera, A., and Eguiarte, L.E. (1999) Genetic Structure of Natural Populations ofEscherichia coli in Wild Hosts on Different Continents. *Appl Environ Microbiol* **65**: 3373–3385.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* **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 Genet* **5**: e1000344.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., *et al*. (2006) Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol Microbiol* **60**: 1136–1151.

Zwietering, M.H., Jongenburger, I., Rombouts, F.M., and Van't Riet, K. (1990) Modeling of the bacterial growth curve. *Appl Environ Microbiol* **56**: 1875–1881.

Swansea University
Prifysgol Abertawe

SCHOOL OF MEDICINE

Biomolecular Research Project

DECLARATION OF ORIGINALITY

Project Title:

Genomic and Phenotypic comparison of *Escherichia coli* isolates from host and environmental sources .....
.....

Declaration: I confirm that I understand the term plagiarism and the University's regulations regarding the consequences of plagiarism. I have generated this report myself, and the work described is my own, except where otherwise acknowledged. It has not been copied from any other person's work (published or unpublished) and has not previously been submitted for assessment.

SIGNATURE OF STUDENT:.....

STUDENT'S NAME (PRINTED): JAKE IRELAND.....

DATE OF SUBMISSION: ……27- MAR -2015……