UPPSALA
UNIVERSITET

UPTEC X 22013
Examensarbete 30 hp
Juni 2022

Genomic comparison of Shiga toxin-producing E. coli O157:H7 from ruminants and humans

Linnéa Good

Civilingenjörsprogrammet i molekylär bioteknik

Genomic comparison of Shiga toxin-producing E. coli O157:H7 from ruminants and humans

Linnéa Good

Abstract

Shiga toxin-producing *E. coli* (STEC) are zoonotic pathogens that frequently colonise ruminants without them showing any symptoms. In humans, STEC cause diarrhoeal disease and occasionally leads to the life-threatening disease haemolytic-uraemic syndrome (HUS). In this study, the aim is to identify any genomic differences between Swedish STEC O157:H7 isolates that have caused HUS and isolates that did not, as well as between isolates taken from animals and isolates taken from humans. I constructed a pan-genome analysis pipeline and performed statistical analyses to find genes that differed between these groups. I also constructed a phylogenetic analysis pipeline to visualise any clustering of isolates based on different categories. The results from the phylogenetic analysis showed that the isolates tended to not form clear clusters based on their category. When comparing isolates from animals to isolates from humans, an elastic net regression analyses yielded a list of 23 genes that differed between them, while a statistical analysis using Scoary found 1854 genes. The genes found by the regression analysis consists largely of genes associated with metabolism, with other notable genes being transposases as well as two genes from the *prp* operon. Gene ontology analysis of the genes from Scoary showed that no particular molecular functions or biological processes stand out when compared to the background frequency of gene ontology terms. When comparing isolates that caused HUS against isolates that did not, no genes were found to be statistically significant. In order to find more conclusive results about the genomic differences between STEC in animals and humans, as well as between STEC that leads to HUS and STEC that does not, further studies are needed.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Robert Söderlund Ämnesgranskare: Lionel Guy

Examinator: Siv Andersson

På jakt efter gener – En genomisk jämförelse av EHEC

Populärvetenskaplig sammanfattning

Linnéa Good

*

Escherichia coli (*E. coli*) är en av de vanligaste bakterierna vi människor har i vår tarmflora, som i vanliga fall är helt ofarlig. Men det finns också patogena *E. coli* som kan orsaka svår sjukdom hos människor. En av dessa är enterohemorragisk *E. coli* (EHEC), som karaktäriseras av att de producerar Shiga toxin. Vanliga symptom vid infektion av EHEC är magont, diarré som oftast blir blodig efter några dagar, och ibland även feber och illamående. Det finns också risk att man utvecklar följdsjukdomen hemolytiskt uremiskt syndrom (HUS), som främst drabbar små barn kan vara dödligt eller leda till livslång sjukdom. EHEC är en zoonos, vilket innebär att det smittar människor från djur, och koloniserar ofta tarmarna av idisslare utan att de visar några symptom. EHEC kan spridas från dessa idisslare till oss människor på flera olika sätt: genom mat från kontaminerade djur, genom miljön på grund av kontakt med kontaminerad mark eller vatten, eller genom direkt kontakt med infekterade djur. Det kan också spridas genom nära kontakt mellan människor. De flesta EHEC fall är från små, isolerade utbrott eller enstaka fall, men ibland sker det även större utbrott som kan spåras till en specifik källa, ofta nationellt distribuerade livsmedel som har kontaminerats.

Målet med denna studie är att identifiera gener som är potentiellt associerade med ökad risk för infektion hos människor och utveckling av HUS. Jag har genomfört en genomisk analys för att hitta gener som skiljer sig mellan EHEC som tagits från djur och människor, samt gener som skiljer sig mellan EHEC som tagits från patienter som utvecklat HUS och som inte utvecklat HUS. Jag har också visualiserat hur EHEC grupperar sig baserat på de variablerna region, om de leder till HUS eller inte, eller om de togs från djur eller människa.

Resultaten från den fylogenetiska analysen visade inga tydliga indelningar baserat på de olika variablerna, endast svaga grupperingar i vissa fall. Från den genomiska analysen fick jag ut en lista på gener som skiljer sig mellan EHEC från djur och människor, vilket gav ledtrådar till vilken typ av gener som kan påverka ökad risk för infektion hos människor. När jag körde den genomiska analysen för att hitta gener som skiljer sig mellan EHEC som tagits från patienter som fick HUS och som inte fick HUS fick jag inte ut några gener alls. Det tyder på att det inte finns några stora genetiska skillnader mellan EHEC som orsakat HUS och EHEC som inte orsakat HUS. Dessa resultat kan dock till stor del ha påverkats av datasetet som användes, och för att kunna dra konkreta slutsatser skulle ytterligare studier behöva genomföras.

Table of contents

Abbreviations

A/E	Attaching and effacing
EHEC	Enterohemorrhagic E. coli
GO	Gene ontology
HUS	Haemolytic-uraemic syndrome
LEE	Locus of enterocyte effacement
STEC	Shiga toxin-producing E. coli
Stx	Shiga toxin

1 Background

1.1 *Escherichia coli*

Escherichia coli (*E. coli*) are gram-negative facultative anaerobic bacteria commonly found in the gastrointestinal tract of warm-blooded animals, including humans. *E. coli* is the most abundant facultative anaerobe of the human intestinal microflora and coexist with their human host with mutual benefit. Commensal *E. coli* rarely cause disease except when the normal gastrointestinal barriers are breached (Kaper *et al.* 2004). E. coli can be divided into different serotypes, characterised by differences in their O (lipopolysaccharide) antigen and their H (flagellar) antigen.

1.1.1 Pathogenic *E. coli*

Although *E. coli* usually are a harmless part of the gut microbiome, there are some strains that have developed certain virulence which can cause severe disease in humans even within the gastrointestinal tract. These strains are grouped into different pathotypes based on their virulence factors and the symptoms they commonly cause in their human hosts: enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), and diffusely adherent *E. coli* (DAEC). These intestinal pathotypes all cause diarrhoeal disease. In addition, there are also extraintestinal pathogenic *E. coli*: uropathogenic *E. coli* (UPEC) which causes UTIs, and meningitis-associated *E. coli* (MNEC) which causes meningitis and sepsis (Kaper *et al.* 2004).

1.1.2 Shiga toxin-producing *E. coli*

EHEC stands out from the rest of the intestinal *E. coli* pathotypes due to the risk of infected people developing the life-threatening disease haemolytic-uraemic syndrome (HUS) (see section 1.2.2). EHEC are zoonotic pathogens and are characterised by their ability to produce Shiga toxin (Stx). The term EHEC refers only to the subset of Stx-producing strains that are pathogenic and cause haemorrhagic colitis in humans. The broader group of *E. coli* that produce Stx that are not necessarily pathogenic are commonly referred to as Shiga toxin-producing *E. coli*, or STEC. However, since it can be hard to know if a Stx-producing strain is pathogenic or not, the terms EHEC and STEC are often used interchangeably. STEC is also sometimes known as verotoxin-producing *E. coli*, or VTEC, as verotoxin is another term for Shiga toxin. In this report, the terms STEC and Shiga toxin (Stx) will be used.

1.1.3 O157:H7

Previous studies have shown that in Sweden as well as many other parts of the world, the most common serotype of STEC causing infection and HUS in humans is O157:H7 (Karmali *et al.* 2003, Anonymous 2014). The serotype O157:H7 can also be further divided into different clades based on their genetic relatedness, with varying virulence levels. O157:H7 infections from clade 8 have been found to be significantly more likely to lead to severe disease and HUS than other clades (Manning *et al.* 2008, Iyoda *et al.* 2014).

Evolutionary analysis of O157:H7 has shown that it evolved from the *E. coli* serotype O55:H7, which is non-toxigenic and is associated with infantile diarrhoea (Wick *et al.* 2005). Typical O157:H7 strains emerged from O55:H7 through four evolutionary steps: (i) acquisition of Stx2, (ii) switch of O antigen from 55 to 157 and acquisition of the virulence plasmid pO157, (iii) acquisition of Stx1, (iv) loss of ability to ferment sorbitol and loss of β-glucuronidase expression. The inability to ferment sorbitol is used to detect O157:H7 by growing it on sorbitol-MacConkey agar (March & Ratnam 1986). The importance of Stx2, pO157 and Stx1 will be discussed further in section 1.2.3.

1.2 Human O157:H7 infection

1.2.1 Symptoms

The typical symptoms of human STEC O157:H7 infection are abdominal pain and diarrhoea which develops into bloody diarrhoea after a few days in about 80% of cases (Mead & Griffin 1998, Tarr *et al.* 2005). In some cases, fever and vomiting has also been reported. The incubation period is around 3 days, and if no further complications occur the symptoms typically resolve spontaneously after about 7 days on average (Tarr *et al.* 2005). O157:H7 infections can also be asymptomatic. In a 1996 outbreak in Scotland, 12% of people who tested positive for O157 were asymptomatic, and in a 2009 English outbreak 16% who tested positive were asymptomatic (Pennington 2010). However, an accurate estimation of the rate of asymptomatic cases is hard to determine due to the lack of population-wide surveys.

The use of antibiotics to treat STEC O157:H7 infections is not recommended as it has been shown to increase the risk of developing HUS in children, theorised to be due to the increased release of Stx in dead and dying bacterial cells (Wong *et al.* 2000, Goldwater & Bettelheim 2012). Patients with STEC-associated bloody diarrhoea are recommended to be hospitalised, given intravenous expansion, and monitored for any development of HUS (Karch *et al.* 2005).

1.2.2 Haemolytic-uraemic syndrome (HUS)

Human STEC infection occasionally progresses to haemolytic-uraemic syndrome (HUS). HUS is characterised by causing haemolytic anaemia (which causes the destruction of red blood cells), thrombocytopenia (low count of blood platelets) and acute kidney injury, which can be fatal or cause life-long disability. The mortality rate of HUS is 1-4%, and about 30% of patients develop long-term sequelae (Spinale *et al.* 2013). The percentage of people who are infected with O157:H7 that develop HUS varies, but it commonly ranges from 5-10% although in some outbreaks the percentage has been as high as about 20% (Mead & Griffin 1998, Keithlin *et al.* 2014). Young children are particularly vulnerable for developing HUS, and the risk of a child below the age of 10 infected with O157:H7 developing HUS is about

15% (Tarr *et al.* 2005). Elderly people also have a larger risk of developing HUS compared to other adults (Gould *et al.* 2009).

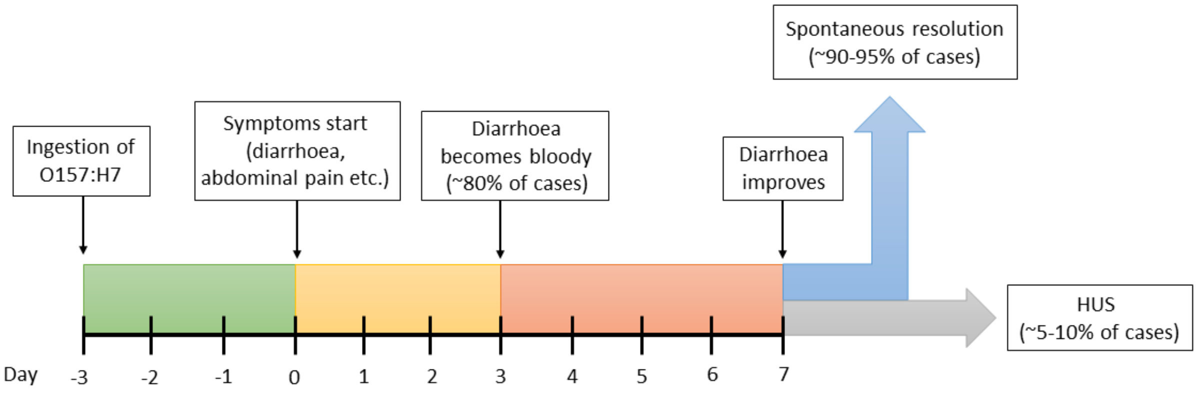

Figure 1. An overview of an average O157:H7 infection from ingestion to resolution. Adapted from Tarr *et al.* (2005).

1.2.3 Virulence factors

A key virulence factor of STEC is Shiga toxin (Stx), which causes damage to the colon leading to bloody diarrhoea and haemorrhagic colitis. Stx is also systemically absorbed by the host and travels via the bloodstream to the kidney where it causes endothelial cell damage, which can lead to the development HUS (Kaper *et al.* 2004). STEC frequently carry more than one Stx gene, and different serotypes of STEC and different clades of O157:H7 carry distinct variants and combinations of Stx genes. Stx are divided into two types: Stx1 and Stx2, which are then further divided into the subtypes Stx1a, Stx1c, and Stx1d as well as Stx2a to Stx2g (Scheutz *et al.* 2012). STEC can carry only Stx1, only Stx2, or a combination of both types. Stx2a and Stx2c in particular have been shown to be significantly associated with O157:H7 clade 8 and an increased risk of severe disease and HUS (Eklund *et al.* 2002, Iyoda *et al.* 2014).

STEC will adhere to the colon, inducing attaching and effacing (A/E) lesions. These lesions cause the destruction of brush border microvilli and cytoskeletal derangement, limiting the absorbing ability of the colon, leading to diarrhoea (Frankel *et al.* 1998). The proteins involved in A/E are located on the pathogenicity island called locus of enterocyte effacement (LEE). LEE is another key virulence factor for STEC since strains that do not contain LEE pathogenicity islands are commonly not pathogenic, although some LEE-negative STEC strains use different mechanisms to interact with the host and induce disease (Newton *et al.* 2009). The genes encoded by LEE are organised into three major regions with known functions: the genes *eae* (encoding the adhesin intimin) and *tir* (encoding an intimin receptor), genes encoding the type III secretion system, and genes secreted by the type III secretion system (such as *EspA*, *EspB*, *EspD* and *EspF* which are involved with host signal transduction pathways) (Frankel *et al.* 1998, Perna *et al.* 1998). Intimin has been shown to be involved with the intimate adherence of the pathogen to host epithelial cells, making it necessary for inducing A/E lesions (Jerse *et al.* 1990, Dean-Nystrom *et al.* 1998).

In addition to Stx and LEE, STEC O157:H7 also contains a virulence plasmid called pO157. The pO157 contains 35 genes that are presumably involved in pathogenesis of O157:H7, but out of these only 19 have been characterised (Burland *et al.* 1998, Lim *et al.* 2010). The characterised genes that the pO157 encodes has been shown to influence adherence to eukaryotic cells, colonisation of cattle, and acid resistance (Lim *et al.* 2010). However, the significance of pO157 as a virulence factor is still not fully understood.

1.3 STEC in animals

Many animals including ruminants are frequently colonised by STEC without showing any signs of illness. Only neonatal calves infected by O157:H7, unlike older calves or adult cattle, have been shown to be susceptible to A/E lesions which cause diarrhoea (Dean-Nystrom *et al.* 1997). Cattle are also not affected by the Shiga toxins as they lack the vascular receptors for it (Pruimboom-Brees *et al.* 2000).

1.4 Transmission of STEC

Due to the low infectious dose required for transmission, estimated at less than 100 bacteria (Tilden *et al.* 1996), STEC can be transmitted from animals to humans in a variety of ways. Food items are commonly associated with STEC infections as they tend to be the source of infection in larger outbreaks, usually products such as undercooked meat, unpasteurised dairy products, or vegetables that have been watered from a contaminated source. Other important sources of contamination that are particularly common in sporadic cases are through direct contact with animals, or environmentally, for example through contact with contaminated water or soil (Kaper *et al.* 2004). In addition, person-to-person transmission has been frequently recorded, commonly via people in the same household or in children's day care facilities (Pennington 2010).

1.5 Outbreaks

The first reported case of human STEC infection occurred in 1975 in the United States, and the first major outbreaks occurred in 1982, also in the United States (Wells *et al.* 1983). Since then, human STEC infections have been reported on all continents except Antarctica (Chase-Topping *et al.* 2008). Most of the cases of human STEC infection are sporadic or small isolated outbreaks, but occasionally there are larger outbreaks that can be traced to a common source. Two of the largest outbreaks in Sweden were due to the O157:H7 serotype (Folkhälsomyndigheten 2022). One was in 2005 where 135 people in southern Sweden were infected and 11 of them developed HUS. The source was traced back to lettuce that had been watered with contaminated water. The second was an outbreak in 2018 where 116 people in different parts of Sweden were infected with STEC and 14 developed HUS, which was suspected to be linked to some food item that had been nationally distributed. Swedish national authorities perform surveillance activities, outbreak investigation and design of interventions to understand and reduce the prevalence of high-virulence STEC among animals in order to reduce the incidence of severe disease among humans.

1.6 Aims

In this project, the identity and function of genes in STEC potentially associated with an increased risk of human infection and HUS will be investigated by comparing Swedish STEC O157:H7 isolates taken from animals and humans and identifying genomic differences. The results will inform future efforts to combat STEC in animals for the benefit of human health.

2 Material and methods

In this section, the data that was used is presented along with the methods for the analyses that were performed in this project. All scripts used can be found at the following GitHub page: https://github.com/LinneaGG/exjobb.

2.1 Overview of pipelines

In this project, two different pipelines were created in order to perform two different analyses: A phylogenetic analysis to correlate phylogeny visualised as trees with host preference, disease severity and geographic origin, as well as a pan-genome analysis pipeline to generate a list of genes that significantly differ between animal and human isolates, as well as isolates that did and did not cause HUS (see figure 2). The phylogenetic analysis pipeline consists of the steps subsampling, trimming, core SNP analysis and generating the phylogenies. The pan-genome analysis pipeline consists of the steps subsampling, trimming, assembly, annotation, pan-genome analysis, statistical analyses and finally a GO and pathway analysis. The subsampling and trimming steps are part of the pre-processing of the data and is identical in both pipelines. For the pan-genome analysis pipeline, only isolates that belong to clade 8 were analysed in order to reduce noise from differences in accessory genome content between clades in the results.

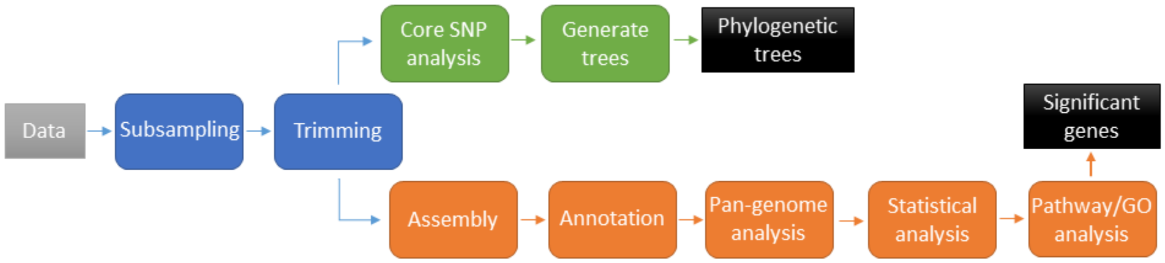

Figure 2. Overview of the two pipelines created: The phylogenetic analysis pipeline (blue and green), and the pan-genome analysis pipeline (blue and orange).

2.2 Data

Two sets of read data from Swedish STEC isolates of the serotype O157:H7 were used in this project. The first was data from isolates recovered from cattle and sheep during nationwide slaughterhouse prevalence studies and targeted regional sampling in areas with high prevalence of O157:H7 clade 8 over the past decade, and the second was data from isolates recovered from human patients across Sweden, received by the Swedish Public Health Agency (Folkhälsomyndigheten). The animal isolate data is paired-end short-read Illumina data while the human isolate data is single-end short-read Ion-Torrent data. There are 157 animal isolates and 227 human isolates, taken from different regions of Sweden. For information on how many isolates were obtained from each region, see appendix A. 73 (46%) of the animal isolates belongs to clade 8. Out of the 227 human isolates, 21 (9%) of them are from patients who developed HUS. 10 of the isolates has no information on whether the patient developed HUS or not. 132 (58%) of the human isolates belongs to clade 8, and 16 (76%) of patients who developed HUS had been infected with a clade 8 strain.

The reference used in the phylogenetic analysis was the complete genome of the E. coli O157:H7 clade 8 TW14359 strain (GenBank accession number CP001368) responsible for a spinach-associated outbreak in the United States in 2006 (Kulasekara *et al.* 2009).

2.3 Pre-processing and quality control

Before any analysis was done, the data was first pre-processed and quality controlled. FastQC ver. 0.11.9 (Andrews 2010) and MultiQC ver. 1.12 (Ewels *et al.* 2016) was used to check the quality of the reads before and after pre-processing. Any data that had more than 100x coverage was randomly subsampled in order to reduce the time and computing resources needed to run the analyses. The data was then trimmed using Trimmomatic ver. 0.39 (Bolger *et al.* 2014) to remove adapters and low-quality bases. The settings used were ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:2:True LEADING:3 TRAILING:20 MINLEN:36. These settings remove Nextera adapters, removes bases from the start and end of reads if below a certain Phred quality score threshold (3 and 20, respectively), and discards any reads that are shorter than 36 base pairs.

2.4 Phylogenetic analysis

A phylogenetic analysis was performed on the trimmed reads by running Snippy ver. 4.6.0 (Seeman 2015), a tool for variant calling and creating a core SNP alignment. This was done using the snippy-multi command, specifying the --minfrac option (the minimum proportion for variant evidence) as 0.9. The minimum site depth required for calling alleles was used as the default value 10. Grapetree ver.1.5.0 (Zhou *et al.* 2018) was then used to create trees in Newick format. The –method option was specified as "NJ", creating FastME 2.0 Neighbour Joining trees (Lefort *et al.* 2015). The trees were then visualised in the Grapetree web application ver. 1.5 (Zhou *et al.* 2018), also using metadata to visualise any potential clustering of different categories. The animal isolates were visualised based on region as well as year, and the human isolates were visualised based on region as well as whether or not they caused HUS. Trees were also made with both the animal and human isolates, visualising the difference between them.

2.5 Assembly

The assembly was done with SPAdes ver. 3.15.3 (Prjibelski *et al.* 2020) using the --careful option, which tries to reduce the number of mismatches and short indels. The assemblies were then filtered by removing any contigs shorter than 500 base pairs as well as contigs with a k-mer coverage of < 1, since these contigs are likely to not be anything valuable and would only create noise in the upcoming analyses. The quality of the assemblies was assessed using Quast ver. 5.2 (Gurevich *et al.* 2013), and any assemblies that seemed to be contaminated were discarded from the analysis. These were determined by looking at the total length of the assembly and discarding any that were unreasonably large compared to the size of an *E. coli* O157:H7 genome.

2.6 Annotation

The annotation was done with Prokka ver. 1.45 (Seemann 2014), using the included Escherichia-specific BLAST database.

2.7 Pan-genome analysis

The pan-genome analysis was performed by running Roary ver. 3.13 (Page *et al.* 2015). Here, the minimum percentage identity was specified as 90 instead of the default 95 to make it less stringent. Roary takes annotated assemblies in GFF3 format and calculates the pan-genome of them. Roary also outputs a file listing which isolates contain each protein, which was then used in further statistical analyses.

2.8 Statistical analysis

A statistical analysis in the form of an elastic net regression analysis was performed in RStudio ver. 1.3 (RStudio Team 2020) on the pan-genome analysis data in order to estimate the relationships between traits and genes. This outputs a coefficient for each gene where a higher absolute value indicates a stronger relationship between the gene and the trait. Elastic net regression is essentially a mix of ridge regression and lasso regression. In ridge regression it is assumed that every variable (or gene in this case) is important to the trait, and this means no coefficient can be equal to zero. With lasso regression however, it is assumed that many of the variables are not correlated with the trait and variables are allowed to have coefficients equal to zero, meaning they have no impact on the trait. The traits HUS/no HUS and human/animal among the isolates were analysed to identify any genes that differ between isolates that caused HUS and isolates that did not cause HUS as well as between animal and human isolates. The Roary output file was first transformed to fit the conditions for the analysis; unnecessary columns were removed so that we only had the isolates and genes in the table, the table was transposed, and the content of the table were changed so that 1 = gene presence and 0 = gene absence. Any genes that were present in every isolate was removed as these are non-informative. Finally all variables were converted to factors. An elastic net regression model was then fit to the data using the caret package. Leave-one-out cross-validation was used to determine the optimal value for the shrinkage parameter λ, testing values from 0 to 10 with intervals of 0.1. The shrinkage parameter λ shrinks the coefficients which then reduces variance. The optimal value for λ was determined to be 0.4. For elastic net regressions, an α value is set to determine how close to ridge regression or lasso regression the analysis should be, where a value of 0 corresponds to ridge regression and a value of 1 corresponds to lasso regression. The value for α was set as 0.4 after manually testing different values to see which best suited the dataset. The genes with non-zero coefficients were then plotted to visualise the results.

Additionally, a statistical analysis was also done using Scoary ver. 1.6.16 (Brynildsrud et al. 2016), a software designed to do a statistical analysis of the association between genes and traits using the output from Roary. Scoary uses a pairwise comparisons algorithm to determine the likelihood of each gene being associated with the trait. The Scoary output contains several statistical measurements including Bonferroni-adjusted p-values, which were used to determine the cut-off of significant genes at p-value < 0.05.

2.9 Gene ontology and pathway analysis

A gene ontology (GO) analysis was performed using Panther (Protein Analysis Through Evolutionary Relationships) ver. 16 (Mi *et al.* 2021), as well as using QuickGO (Binns *et al.* 2009). This was done in order to find out which molecular functions and biological processes are common among the genes found by the statistical analyses. GO terms for the significant genes found by the statistical analyses were determined, and their frequency were compared to the background frequency of GO terms in the complete gene set by a statistical overrepresentation test.

A pathway analysis was performed using KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway database (Kanehisa & Goto 2000). This was done to determine which pathways could be involved in the difference between the traits tested, as well as to see if any of the significant genes found are part of the same pathway.

3 Results

In this section, the results from the phylogenetic analysis, the statistical analysis, and the GO and pathway analysis will be presented.

3.1 Phylogenetic analysis

I have visualised the relationships of the isolates by creating neighbour joining phylogenetic trees from core SNP alignments and grouping the isolates by different categories. The isolates taken from animals are grouped by region (figure 3), and the isolates taken from humans are grouped by region and whether infection led to HUS or not (figure 4 and 5). Both the animal and human isolates are shown in the same tree, grouped by if the isolate came from an animal or a human (figure 6). Looking at the phylogenetic trees, we can see that the isolates mostly do not tend to form clear clusters based on any grouping. In the phylogeny of isolates from animals in figure 3, the only apparent cluster that can be seen is that most of the isolates from Kalmar are clustered together alongside the rest of the clade 8 isolates. We can see that a majority of the isolates from Kalmar, Skåne, Blekinge and Kronoberg belong to clade 8 and make up most of the isolates in the clade 8 group, as seen in figure 3. For the human isolates in figure 4 we can see that similarly to the animal isolates, a majority of isolates from Kalmar, Skåne, and Kronoberg are clustered and belong to clade 8, as well as a majority of isolates from Halland, Uppsala, Jönköping, and Värmland. We can also see that most isolates from Östergötland are closely clustered together. In figure 5, besides the most of the HUS cases being caused by clade 8 isolates, there are also four cases of HUS from three different regions that cluster together that do not belong to clade 8. Finally, when comparing isolates from animals against isolates from humans in figure 6 isolates are only loosely grouped together based on their source, with some clusters forming consisting largely of isolates from one of the sources. There is also a larger diversity among human isolates compared to animal isolates.

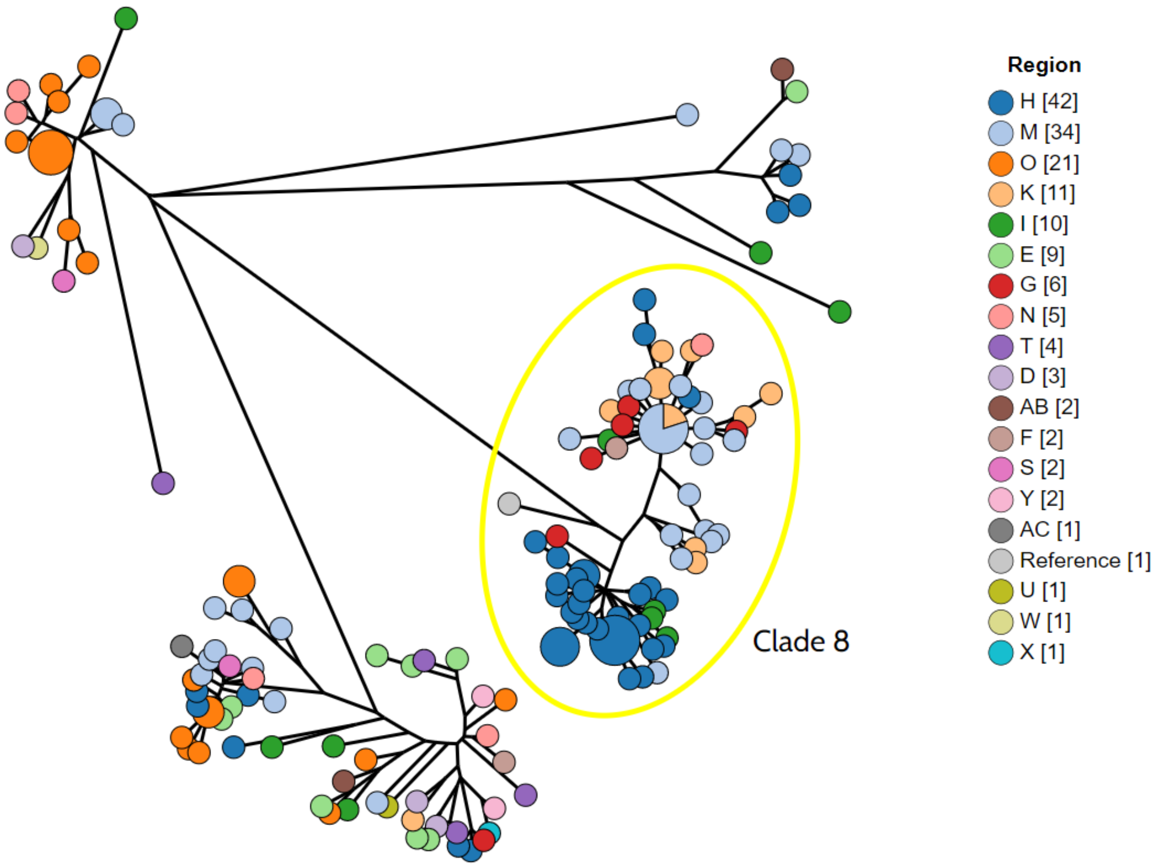

Figure 3. Phylogeny of all animal O157:H7 isolates, coloured by region. H = Kalmar, M = Skåne, O = Västra Götaland, K = Blekinge, I = Gotland, E = Östergötland, G = Kronoberg, N = Halland, T = Örebro, D = Sörmland, AB = Stockholm, F = Jönköping, S = Värmland, Y = Västernorrland, AC = Västerbotten, U = Västmanland, W = Dalarna, X = Gävleborg. The nodes are scaled based on the number of isolates. For nodes containing isolates from multiple regions, a pie chart gives the relative number of nodes per region.

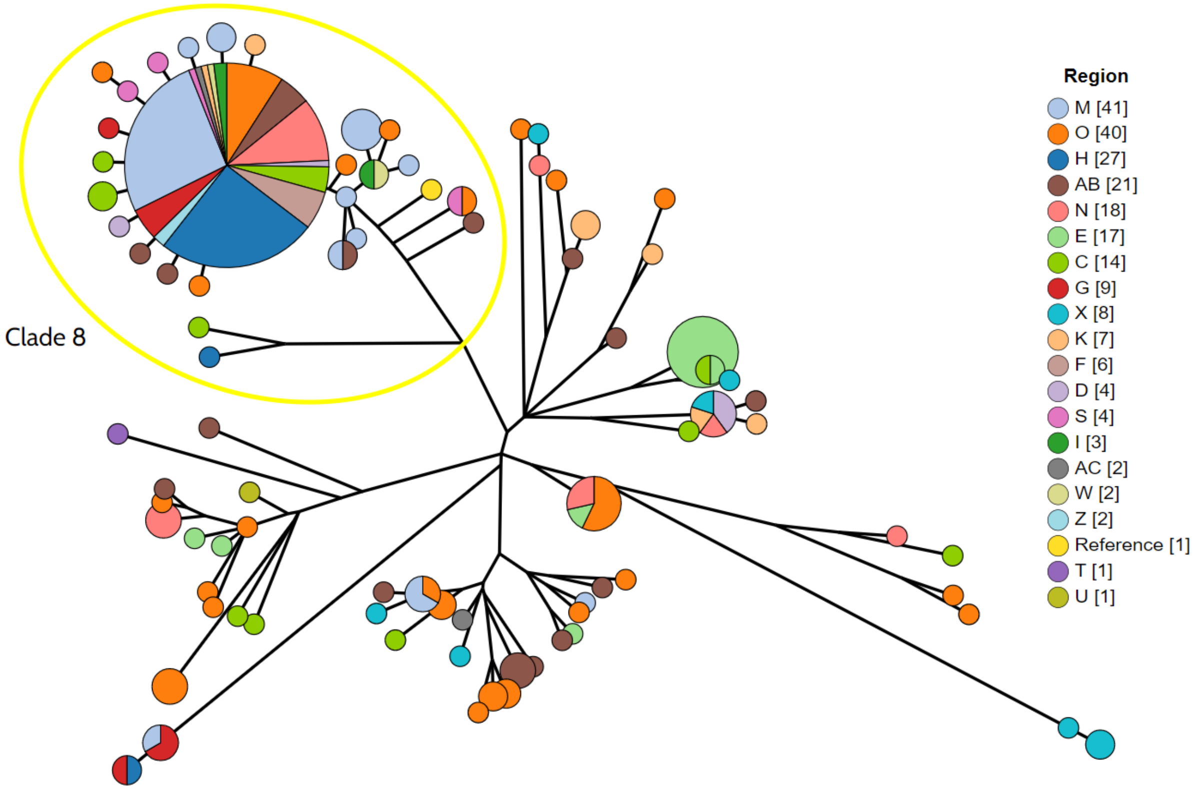

Figure 4. Phylogeny of all human O157:H7 isolates, coloured by region. M = Skåne, H = Kalmar, O = Västra Götaland, N = Halland, E = Östergötland, AB = Stockholm, C = Uppsala, F = Jönköping, G = Kronoberg, S = Värmland, I = Gotland, D = Sörmland, K = Blekinge, W = Dalarna, Z = Jämtland Härjedalen, AC = Västerbotten, T = Örebro, U = Västmanland. The nodes are scaled based on the number of isolates. For nodes containing isolates from multiple regions, a pie chart gives the relative number of nodes per region.

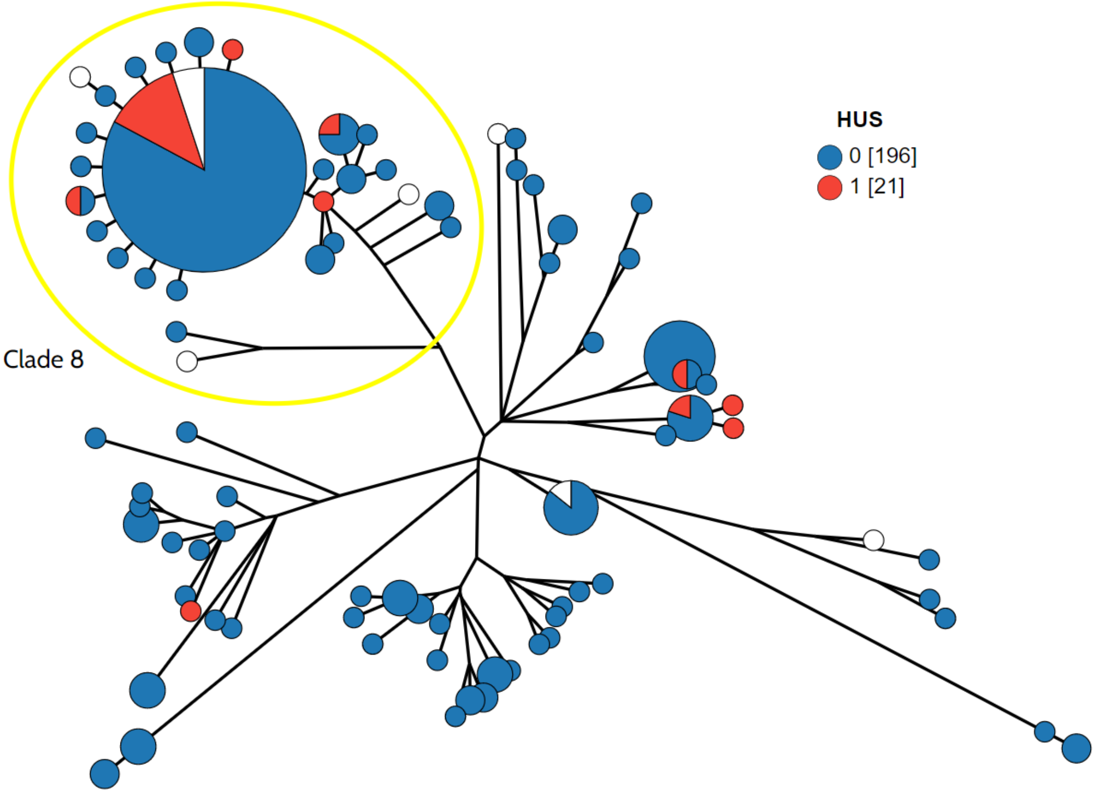

Figure 5. Phylogeny of all human isolates coloured based on whether infection led to HUS or not. 1 indicates HUS, 0 indicates no HUS. White nodes indicate no data. The nodes are scaled based on the number of isolates. For nodes containing both isolates that led to HUS and isolates that did not, a pie chart gives the relative number of nodes per trait.

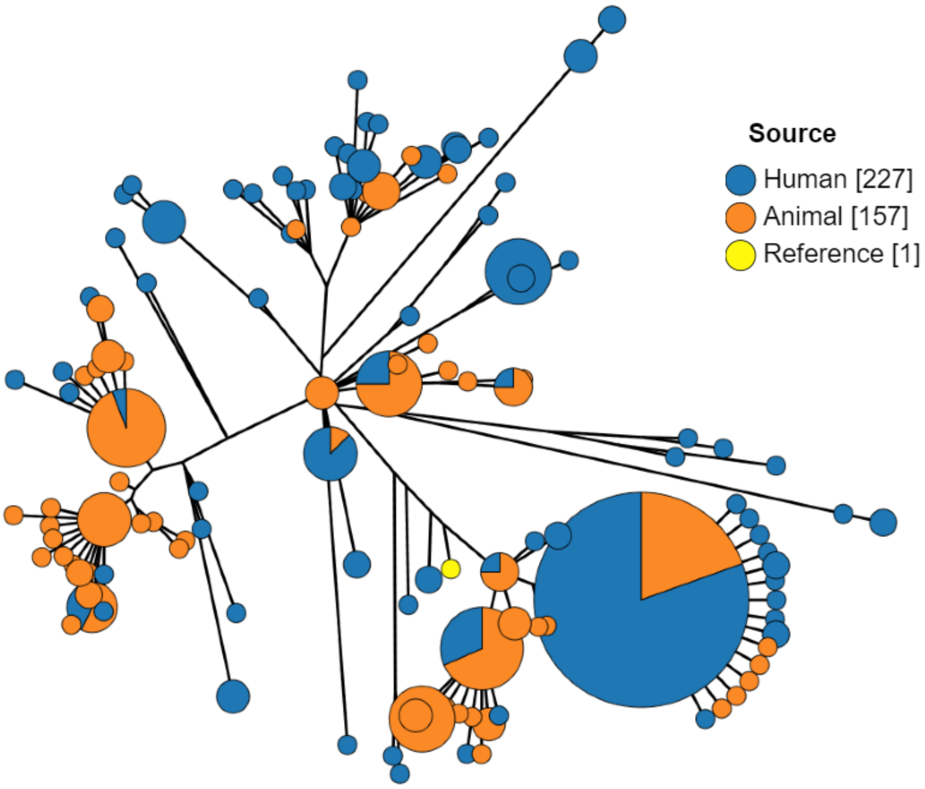

Figure 6. Phylogeny of all isolates coloured by source. The nodes are scaled based on the number of isolates. For nodes containing isolates from both sources, a pie chart gives the relative number of nodes per source.

3.2 Statistical analysis

Elastic net regression analyses were performed in order to identify genes that significantly differ between clade 8 isolates that did or did not cause HUS, as well as animal and human clade 8 isolates. This analysis yields a list of genes with either a positive or a negative coefficient indicating if the gene is positively or negatively correlated with the trait tested, with a larger absolute value indicating a stronger correlation. Analyses were also run using the statistical analysis software Scoary, which uses a pairwise comparisons algorithm to determine the likelihood of each gene being associated with the trait.

When running the statistical analyses comparing the human isolates that led to HUS to the ones that did not, no statistically significant genes were found in the elastic net regression analysis nor in the Scoary analysis.

However, the elastic net regression analysis comparing isolates from animals to isolates from humans yielded 40 genes with non-zero coefficients, meaning 40 genes were found as significant in the analysis (see figure 7). Out of these, 17 were only annotated as “hypothetical protein”, meaning an open-reading frame was found but there was no hit in the databases searched by Prokka. Of the annotated genes, 9 of them had positive coefficients, meaning those genes are more associated to the isolates taken from humans. 14 of the annotated genes had a negative coefficient, meaning they are more associated with the isolates from animals. One thing to note is that two different genes both annotated as *prpE* shows up both among the human-associated genes and the animal-associated genes.

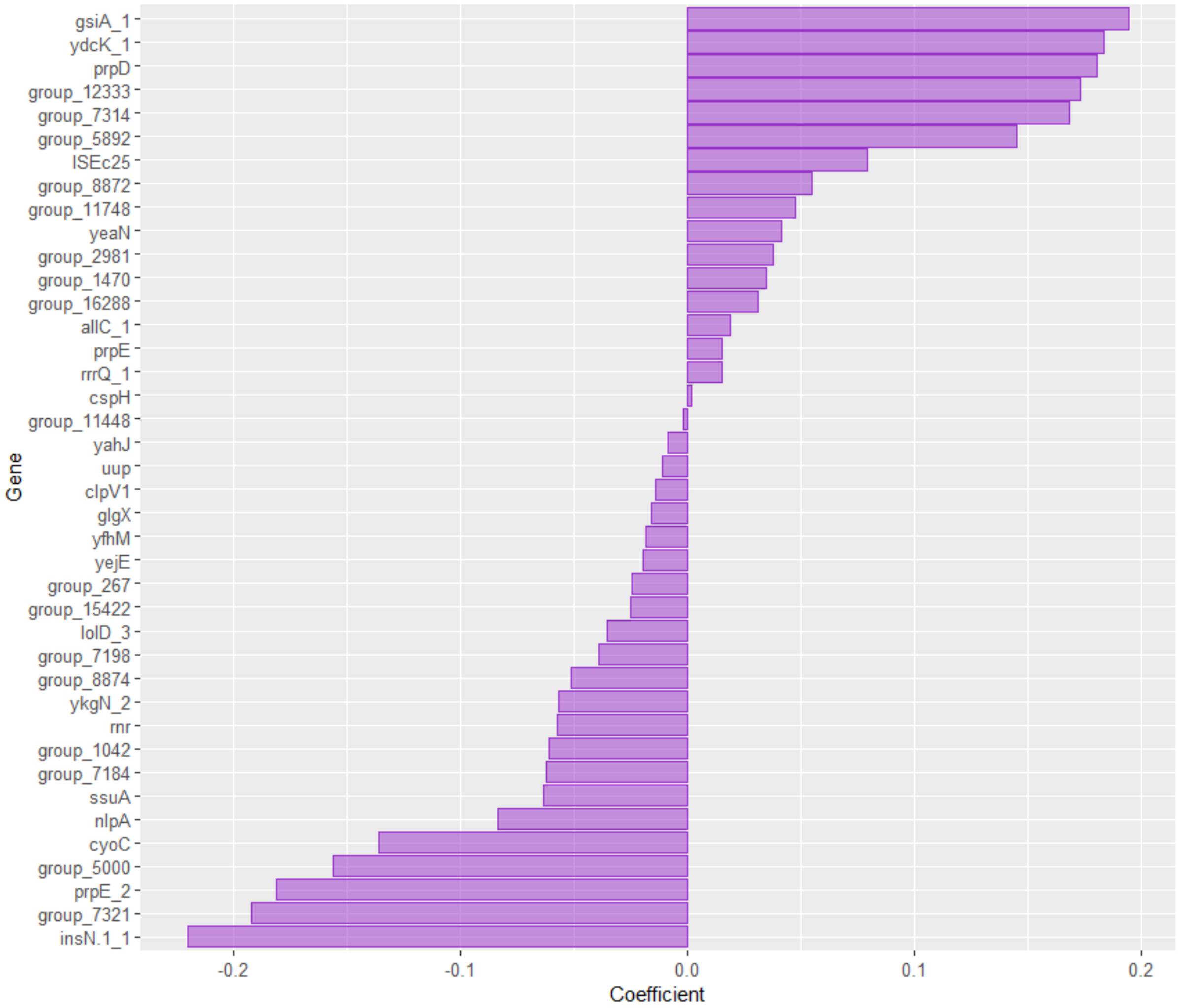

Figure 7. Graph visualising the non-zero coefficients yielded from the elastic-net regression analysis comparing isolates taken from humans against isolates taken from animals. A positive coefficient indicates association with isolates from humans, and a negative coefficient indicates association with isolates from animals. Genes named “group_...” were annotated as hypothetical proteins.

When running Scoary comparing human and animal clade 8 isolates, 1854 significant genes were yielded. Out of these, 558 are non-hypothetical, unique genes associated with animal isolates. 237 of them are non-hypothetical, unique genes associated with human isolates. Among the genes associated with human isolates we can find some of the genes previously mentioned in section 1.2.3 such as *eae*, *tir*, *espF*, as well as *stxB* (Shiga toxin subunit B). Every significant gene that was found by the elastic net regression analysis was also found by the Scoary analysis.

3.3 GO/Pathway analysis

GO terms for molecular function and biological process were annotated for the genes found by the elastic net regression analysis using QuickGO, and KEGG was used to find the pathways. GO terms and pathways that were found for the human-associated genes can be seen in table 1 and GO terms and pathways for animal-associated genes can be seen in table 2, although for some of the genes no GO terms or pathways were found.

Table 1. Genes associated with human isolates found by the elastic net regression analysis. Molecular function and biological process GO terms found by QuickGO and pathways found by KEGG are noted for each gene.

Gene	GO – Molecular Function	GO – Biological Process	Pathway
gsiA	ATP binding; ATPase-coupled transmembrane transporter, Hydrolase activity	Peptide transport	ABC transporters
ydcK	Acyltransferase activity	-	-
prpD	Hydro-lyase activity; Iron-sulfur cluster binding	Tricarboxylic acid cycle; Propionate metabolic process, methylcitrate cycle	Propanoate metabolism
ISEc25	Transposase activity; DNA binding	Transposition, DNA-mediated	-
yeaN	Transmembrane transporter activity	Response to antibiotic	-
allC	Hydrolase activity; Protein homodimerization activity; Transition metal ion binding	Allantoin assimilation pathway; Purine nucleobase metabolic process	Purine metabolism
prpE	ATP binding; Propionate-CoA ligase activity	Propionate catabolic process, 2-methylcitrate cycle	Propanoate metabolism
rrrQ	Lysozyme activity	Macromolecule catabolic process; Cytolysis; Defense response to bacterium	-
cspH	Nucleic acid binding	Regulation of gene expression	-

Table 2. Genes associated with animal isolates found by the elastic net regression analysis. Molecular function and biological process GO terms found by QuickGO and pathways found by KEGG are noted for each gene.

Gene	GO – Molecular function	GO – Biological Process	Pathway
insN1	DNA binding; Transposase activity	Transposition, DNA-mediated	-
prpE	ATP binding; Propionate-CoA ligase activity	Propionate catabolic process, 2-methylcitrate cycle	Propanoate metabolism
cyoC	Oxidoreduction-driven active transmembrane transporter activity; Proton transmembrane transporter activity; Electron transfer activity	Aerobic electron transport chain	Oxidative phosphorylation
nlpA	-	-	ABC transporters
ssuA	Alkanesulfonate transmembrane transporter activity; ATPase-coupled transmembrane transporter activity	Alkanesulfonate transport; Cellular response to sulfur starvation; Sulfur compound metabolic process	Sulfur metabolism; ABC transporters
rnr	Ribonuclease activity; RNA binding	mRNA catabolic process; ncRNA processing; Response to cold	RNA degradation
ykgN	DNA binding; Transposase activity	Transposition, DNA-mediated	-
lolD	ATP binding; Lipoprotein releasing activity; Transmembrane transporter activity	Lipoprotein transport; Transmembrane transport	ABC transporters
yejE	-	Microcin transport; Oligopeptide transmembrane transport; Peptide transport	ABC transporters
yfhM	Endopeptidase inhibitor activity	Negative regulation of endopeptidase activity	-
glgX	Amylo-alpha-1,6-glucosidase activity; Glycogen debranching enzyme activity	Cellular response to DNA damage stimulus; Glycogen catabolic process	Starch and sucrose metabolism
clpV1	ATP binding; ATP hydrolysis activity; Peptidase activity	-	-
uup	ATP binding; ATP hydrolysis activity; DNA binding; Ribosome binding	Postreplication repair; Regulation of transposon integration; Response to radiation	-
yahJ	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines	-	-

For the significant genes found by Scoary, statistical overrepresentation tests were performed using Panther. Tests were performed for genes associated with human isolates and genes associated with animal isolates, looking at both molecular functions and biological processes, however there were no significant GO terms that were found to be over- or underrepresented in any of the tests. This means that the frequency of different GO terms among both human-associated isolates and animal-associated isolates is essentially the same as the background frequency among all genes found in all of the isolates (see figure 8). For the human-associated genes found by Scoary there were 112 GO terms for molecular functions found, and 140 GO terms for biological processes found. For the animal-associated genes there were 257 GO terms for molecular functions found, and 363 GO terms for biological processes found. For the set of all genes that were used in the statistical analyses there were 1592 GO terms for molecular functions found, and 2285 GO terms for biological processes found.

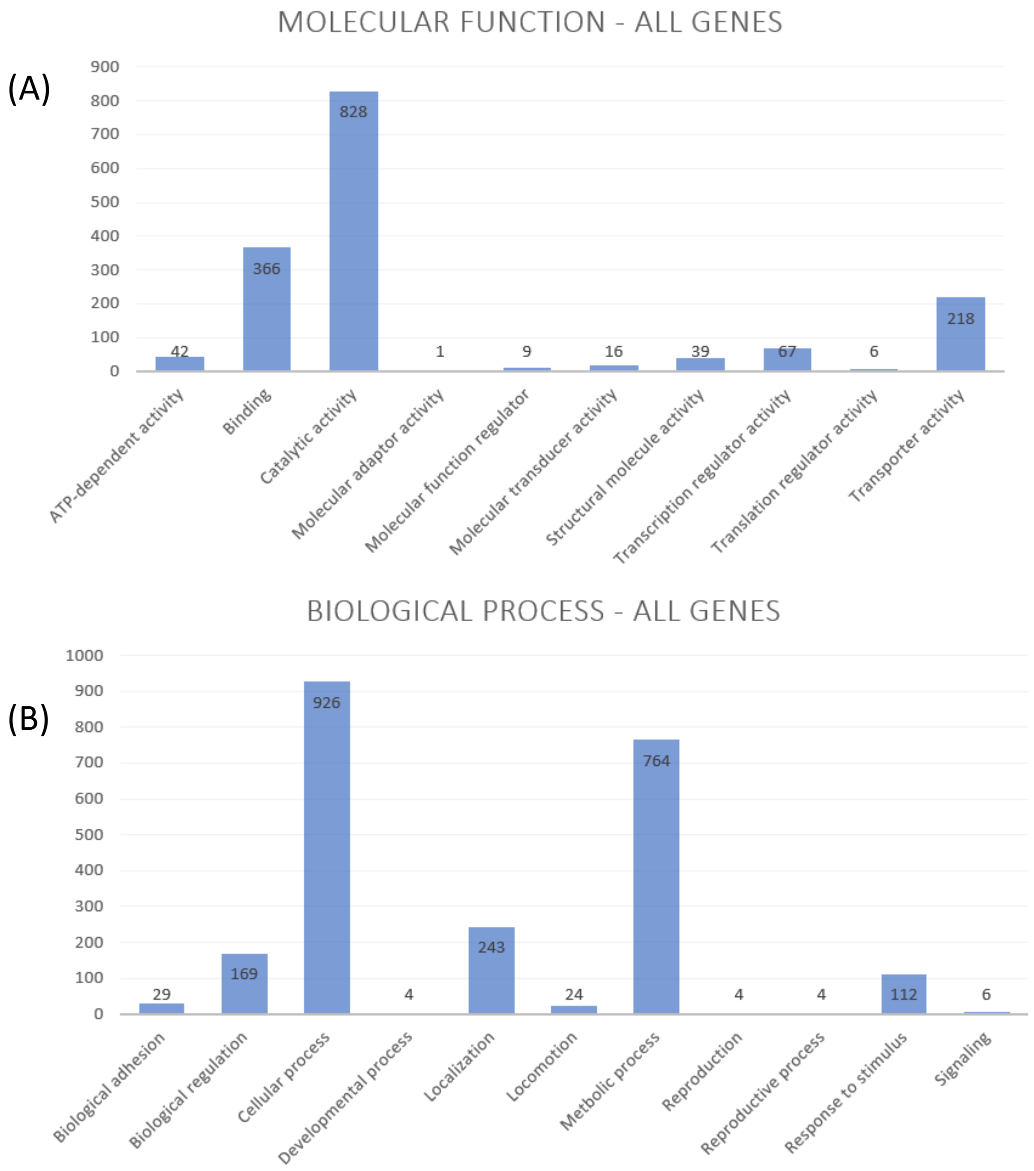

Figure 8. The frequency of different GO terms in the set of all genes. (A) The frequency of molecular function GO terms in the set of all genes. (B) The frequency of biological process GO terms in the set of all genes.

4 Discussion

When grouping isolates by region in the phylogenetic analysis, the isolates do not tend to form perfect clusters, and we can only see vague clusters forming in some cases. This indicates that strains are not necessarily region-specific, and that very similar strains can be found even in different regions of the country. In the phylogeny showing which isolates caused HUS, some of the HUS cases from different regions clustered together while some parts of the tree have no HUS cases. This could be due to the small number of isolates that caused HUS in the dataset, but it could be interesting to study this further with a larger dataset and see how the HUS cases cluster. When grouping the human isolates along with the animal isolates, a larger diversity can be observed among the human isolates. This could be because despite the fact that all patients the isolates were taken from were infected in Sweden, we do not know the source of their infection. This means that they could have been infected by for example an imported food item, and that isolate will therefore be different from the rest of the Swedish isolates.

The statistical analyses comparing isolates from humans that developed HUS to isolates from humans that did not develop HUS yielded no genes that significantly differed between the two groups, and this could have happened due to several reasons. The set of isolates tested could have been too small as there were only 16 isolates that had caused HUS among the clade 8 isolates that were used in the pan-genome analysis pipeline. Such a small number of isolates makes it more difficult to find any genes of statistical significance. There is also the possibility that no specific genes in the STEC are affecting the rate causing HUS in this case, as we already know that certain groups of people are much more vulnerable to developing HUS. However, we also know that different outbreaks have had large variations in the frequency of people that develop HUS, indicating that different strains may in fact have some genetic difference to affect this. To study this further perhaps it would be better to use a larger set of isolates, as well as to specifically study isolates from outbreaks with a low rate of HUS against isolates from outbreaks with a high rate of HUS, although this might be hard to do in practice.

The genes found in the elastic net regression analysis comparing isolates taken from humans to isolates taken from animals largely consist of genes associated with metabolism such as transporters. There are also some transposases found both among human-associated genes and animal-associated genes, including the most significantly associated gene of the animal-associated genes, *insN1*. Also, two genes in the same operon were found, *prpD* and *prpE*. However, two different gene copies of *prpE* showed up in each set of significant genes, one in the human-associated genes and one in the animal-associated genes. This could mean that there are gene copies or paralogs of the *prpE* gene that can be found in STEC O157:H7 and that animal isolates tend to have one of them more often and human isolates tend to have the other more often. To further examine the importance of this list of genes yielded, additional studies should be performed with different datasets to see if the results will be replicated.

The number of genes found in the Scoary analysis is too large to examine each individually, however I did find some genes more associated with human isolates that are known to be relevant to virulence in humans like *eae, tir*, *espP* and *stxB* since these virulence factors have no effect on animals like cattle as they are not susceptible to A/E lesions or Shiga toxin.

The statistical overrepresentation tests found no significant results, indicating that the frequency of GO terms in the set of all genes are essentially the same as the frequencies among the human-associated genes and the animal-associated genes. This means that no specific molecular function or biological process was more or less represented among the genes found by the statistical analyses. The number of genes found by Scoary is high enough where this result should not be due to there simply being too few genes to find anything statistically significant. This could mean that any genetic differences between the human and animal isolates are too subtle to be picked up on in such an analysis and that we cannot attribute any potential differences to any specific function or process. Although one thing that affected these results is that the statistical overrepresentation test was done using Panther, which I noticed was not as good at finding GO terms for genes as QuickGO was. When using Panther to look up GO terms for the genes found by the elastic net regression analysis, GO terms were only found for less than half of the genes. Compared to this, when looking up the genes on QuickGO, GO terms were found for nearly all genes for both molecular function and biological process. However, since QuickGO does not have a feature for doing statistical overrepresentation tests and you can only search for one gene ID at a time, Panther had to be used for this analysis.

The Scoary analysis found a lot more significant genes than the elastic net regression analysis. This is partly because the number of significant genes in the elastic net regression analysis is to a certain degree due to the value of α. The value of α was chosen simply based on what seemed to fit the dataset best. A higher α means fewer genes get a non-zero coefficient, which means fewer genes are found that are correlated to the tested trait. A lower α means more genes get a non-zero coefficient, but more of the genes will have very low coefficients and therefore be less significant. So the value of α was chosen to yield as many genes as possible without a large number of them having very small, close to zero coefficients, in order to make sure that the genes found can actually plausibly be correlated to the trait tested. Although the difference in numbers is still quite large, even despite using the Bonferroni-adjusted p-value as the cut-off in Scoary, which is known to be stringent and good at avoiding false positives (Diz *et al.* 2011). But regression analysis and the pairwise comparison analysis that Scoary uses are very different analysis methods, and it seems in this case the analysis performed by Scoary was a lot more sensitive and managed to pick up on more subtle differences between the groups of isolates. In addition, it is important to keep in mind that these statistical analyses are not perfect, and there are some things we cannot find out just from these analyses. For example, we have no way of finding out how genes are intercorrelated. So if a gene shows up in our statistical analysis it may not actually be significant for the trait, but it will still show up in our list of genes simply because it is correlated with another gene that actually is significant to the trait. Another limitation is from the pan-genome analysis itself which limits the analysis to the presence or absence of genes, and differences due to specific alleles within the groups are not detected.

In conclusion, the phylogenetic clustering (or lack thereof) of Swedish *E. coli* O157:H7 isolates has been visualised based on the categories region and year as well as if the isolates caused HUS and whether they were taken from humans or animals. A list of genes that significantly differ between isolates taken from animals and isolates taken from humans has been produced, although the importance of the genes would need to be confirmed by additional studies on different datasets to see if the results are similar. No significantly differing genes were found between isolates that caused HUS and isolates that did not, and further studies with different datasets would be needed in order to identify any potential genetic factors that influence the risk of developing HUS. In future studies, the developed pipelines will be applied to other serotypes of STEC and may also be used as groundwork for genomic comparisons of other pathogens.

5 Acknowledgements

I would like to thank my supervisor Robert Söderlund for all of his support and help, and for giving me the opportunity to work on such an interesting degree project. I want to thank my subject reviewer Lionel Guy for looking over my work and always providing helpful feedback. I also want to thank Lena-Mari Tamminen for helping me create the statistical analysis and making sense of its results.

References

Andrews. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Anonymous. 2014. Infektion med EHEC/VTEC - Ett nationellt strategidokument.

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. 2009. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25: 3045–3046.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology 17: 238.

Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR. 1998. The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. Nucleic Acids Research 26: 4196–4204.

Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. 2008. Super-Shedding and the Link Between Human Infection and Livestock Carriage of Escherichia Coli O157. Nature reviews Microbiology 6: 904–912.

Dean-Nystrom EA, Bosworth BT, Moon HW. 1997. Pathogenesis of O157:H7 Escherichia Coli Infection in Neonatal Calves. In: Paul PS, Francis DH, Benfield DA (ed.). Mechanisms in the Pathogenesis of Enteric Diseases, pp. 47–51. Springer US, Boston, MA.

Dean-Nystrom EA, Bosworth BT, Moon HW, O'Brien AD. 1998. Escherichia coli O157:H7 Requires Intimin for Enteropathogenicity in Calves. Infection and Immunity 66: 4560–4563.

Diz AP, Carvajal-Rodríguez A, Skibinski DOF. 2011. Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work. Molecular & Cellular Proteomics : MCP 10: M110.004374.

Eklund M, Leino K, Siitonen A. 2002. Clinical Escherichia coli Strains Carrying stx Genes: stx Variants and stx-Positive Virulence Profiles. Journal of Clinical Microbiology 40: 4585–4593.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

Folkhälsomyndigheten. 2022. Sjukdomsinformation om enterohemorragisk E. coli-infektion (EHEC) — Folkhälsomyndigheten. WWW document 2022:

https://www.folkhalsomyndigheten.se/smittskydd-beredskap/smittsamma-sjukdomar/enterohemorragisk-e-coli-infektion-ehec/. Accessed 16 March 2022.

Frankel G, Phillips AD, Rosenshine I, Dougan G, Kaper JB, Knutton S. 1998. Enteropathogenic and enterohaemorrhagic Escherichia coli : more subversive elements. Molecular Microbiology 30: 911–921.

Goldwater PN, Bettelheim KA. 2012. Treatment of enterohemorrhagic Escherichia coli (EHEC) infection and hemolytic uremic syndrome (HUS). BMC Medicine 10: 12.

Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, Griffin PM, the Emerging Infections Program FoodNet Working Group. 2009. Hemolytic Uremic Syndrome and Death in Persons with Escherichia coli O157:H7 Infection, Foodborne Diseases Active Surveillance Network Sites, 2000–2006. Clinical Infectious Diseases 49: 1480–1485.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England) 29: 1072–1075.

Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, Ichihara S, Migita Y, Ogata K, Honda M, Kubota T, Kawano K, Matsumoto K, Kudaka J, Asai N, Yabata J, Tominaga K, Terajima J, Morita-Ishihara T, Izumiya H, Ogura Y, Saitoh T, Iguchi A, Kobayashi H, Hara-Kudo Y, Ohnishi M, Arai R, Kawase M, Asano Y, Asoshima N, Chiba K, Furukawa I, Kuroki T, Hamada M, Harada S, Hatakeyama T, Hirochi T, Sakamoto Y, Hiroi M, Takashi K, Horikawa K, Iwabuchi K, Kameyama M, Kasahara H, Kawanishi S, Kikuchi K, Ueno H, Kitahashi T, Kojima Y, Konishi N, Obata H, Kai A, Kono T, Kurazono T, Matsumoto M, Matsumoto Y, Nagai Y, Naitoh H, Nakajima H, Nakamura H, Nakane K, Nishi K, Saitoh E, Satoh H, Takamura M, Shiraki Y, Tanabe J, Tanaka K, Tokoi Y, Yatsuyanagi J. 2014. Phylogenetic Clades 6 and 8 of Enterohemorrhagic Escherichia coli O157:H7 With Particular stx Subtypes are More Frequently Found in Isolates From Hemolytic Uremic Syndrome Patients Than From Asymptomatic Carriers. Open Forum Infectious Diseases 1: ofu061.

Jerse AE, Yu J, Tall BD, Kaper JB. 1990. A genetic locus of enteropathogenic Escherichia coli necessary for the production of attaching and effacing lesions on tissue culture cells. Proceedings of the National Academy of Sciences of the United States of America 87: 7839–7843.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27–30.

Kaper JB, Nataro JP, Mobley HLT. 2004. Pathogenic Escherichia coli. Nature Reviews Microbiology 2: 123–140.

Karch H, Tarr PI, Bielaszewska M. 2005. Enterohaemorrhagic Escherichia coli in human medicine. International Journal of Medical Microbiology 295: 405–418.

Karmali MA, Mascarenhas M, Shen S, Ziebell K, Johnson S, Reid-Smith R, Isaac-Renton J, Clark C, Rahn K, Kaper JB. 2003. Association of Genomic O Island 122 of Escherichia coli EDL 933 with Verocytotoxin-Producing Escherichia coli Seropathotypes That Are Linked to Epidemic and/or Serious Disease. Journal of Clinical Microbiology, doi 10.1128/JCM.41.11.4930-4940.2003.

Keithlin J, Sargeant J, Thomas MK, Fazil A. 2014. Chronic Sequelae of E. coli O157: Systematic Review and Meta-analysis of the Proportion of E. coli O157 Cases That Develop Chronic Sequelae. Foodborne Pathogens and Disease 11: 79–95.

Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, Rohmer L, Ritchie JM, Radey M, McKevitt M, Freeman TL, Hayden H, Haugen E, Gillett W, Fong C, Chang J, Beskhlebnaya V, Waldor MK, Samadpour M, Whittam TS, Kaul R, Brittnacher M, Miller SI. 2009. Analysis of the Genome of the Escherichia coli O157:H7 2006 Spinach-Associated Outbreak Isolate Indicates Candidate Genes That May Enhance Virulence. Infection and Immunity 77: 3713–3721.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution 32: 2798–2800.

Lim JY, Yoon JW, Hovde CJ. 2010. A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157. Journal of microbiology and biotechnology 20: 5–14.

Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, Whittam TS. 2008. Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences of the United States of America 105: 4868–4873.

March SB, Ratnam S. 1986. Sorbitol-MacConkey medium for detection of Escherichia coli O157:H7 associated with hemorrhagic colitis. Journal of Clinical Microbiology 23: 869–872.

Mead PS, Griffin PM. 1998. Escherichia coli O157:H7. The Lancet 352: 1207–1212.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD. 2021. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research 49: D394–D403.

Newton HJ, Sloan J, Bulach DM, Seemann T, Allison CC, Tauschek M, Robins-Browne RM, Paton JC, Whittam TS, Paton AW, Hartland EL. 2009. Shiga Toxin–producing Escherichia coli Strains Negative for Locus of Enterocyte Effacement. Emerging Infectious Diseases 15: 372.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31: 3691–3693.

Pennington H. 2010. Escherichia coli O157. The Lancet 376: 1428–1435.

Perna NT, Mayhew GF, Pósfai G, Elliott S, Donnenberg MS, Kaper JB, Blattner FR. 1998. Molecular Evolution of a Pathogenicity Island from Enterohemorrhagic Escherichia coli O157:H7. Infection and Immunity 66: 3810–3817.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics 70: e102.

Pruimboom-Brees IM, Morgan TW, Ackermann MR, Nystrom ED, Samuel JE, Cornick NA, Moon HW. 2000. Cattle Lack Vascular Receptors for Escherichia coli O157:H7 Shiga Toxins. Proceedings of the National Academy of Sciences of the United States of America 97: 10325–10329.

RStudio Team. 2020. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Scheutz F, Teel LD, Beutin L, Piérard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O’Brien AD. 2012. Multicenter Evaluation of a Sequence-Based Protocol for Subtyping Shiga Toxins and Standardizing Stx Nomenclature. Journal of Clinical Microbiology 50: 2951–2963.

Seeman. 2015. Snippy: Fast bacterial variant calling from NGS reads. https://github.com/tseemann/snippy

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069.

Spinale JM, Ruebner RL, Copelovitch L, Kaplan BS. 2013. Long-term outcomes of Shiga toxin hemolytic uremic syndrome. Pediatric Nephrology 28: 2097–2105.

Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. The Lancet 365: 1073–1086.

Tilden J, Young W, McNamara AM, Custer C, Boesel B, Lambert-Fair MA, Majkowski J, Vugia D, Werner SB, Hollingsworth J, Morris JG. 1996. A new route of transmission for Escherichia coli: infection from dry fermented salami. American Journal of Public Health 86: 1142–1145.

Wells JG, Davis BR, Wachsmuth IK, Riley LW, Remis RS, Sokolow R, Morris GK. 1983. Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare Escherichia coli serotype. Journal of Clinical Microbiology 18: 512–520.

Wick LM, Qi W, Lacher DW, Whittam TS. 2005. Evolution of Genomic Content in the Stepwise Emergence of Escherichia coli O157:H7. Journal of Bacteriology 187: 1783–1791.

Wong CS, Jelacic S, Habeeb RL, Watkins SL, Tarr PI. 2000. THE RISK OF THE HEMOLYTIC–UREMIC SYNDROME AFTER ANTIBIOTIC TREATMENT OF ESCHERICHIA COLI O157:H7 INFECTIONS. The New England journal of medicine 342: 1930–1936.

Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. 2018. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Research 28: 1395–1404.

Appendix A – Overview of data

Table A1. Overview of human isolates.

Region	HUS	Clade 8	Total
M	5	37	**41**
O	1	14	**40**
H	4	26	**27**
AB	2	9	**21**
N	0	10	**18**
E	0	0	**17**
C	2	8	**14**
G	1	6	**9**
X	0	0	**8**
K	3	2	**7**
F	1	6	**6**
D	0	2	**4**
S	0	4	**4**
I	1	3	**3**
Z	1	2	**2**
W	0	2	**2**
AC	0	1	**2**
T	0	0	**1**
U	0	0	**1**
Total	**21**	**132**	**227**

Table A2. Overview of animal isolates.

Region	Clade 8	Total
H	33	**42**
M	17	**34**
O	0	**21**
K	10	**11**
I	4	**10**
E	0	**9**
G	5	**6**
N	1	**5**
T	0	**4**
D	0	**3**
Y	0	**2**
F	1	**2**
AB	0	**2**
S	0	**2**
X	0	**1**
W	0	**1**
U	0	**1**
AC	0	**1**
Total	**73**	**157**