Genome-wide association studies of *Shigella spp.* and Enteroinvasive *Escherichia coli* isolates demonstrate an absence of genetic markers for prediction of disease severity

Amber C. A. Hendriks
National Institute for Public Health and the Environment
Frans A.G. Reubsaet
National Institute for Public Health and the Environment
A.M.D. (Mirjam) Kooistra-Smid
Certe and University of Groningen, University Medical Center Groningen
John W. A. Rossen
University of Groningen, University Medical Center Groningen
Bas E. Dutilh
Theoretical Biology and Bioinformatics, Science for life, Utrecht University and Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre
Aldert L. Zomer
Faculty of Veterinary Medicine, Utrecht University
Maaike J. C. van den Beld (maaike.van.den.beld@rivm.nl)
National Institute for Public Health and the Environment https://orcid.org/0000-0001-8720-8434

Research article

Keywords: GWAS, Shigellosis, Shigella, EIEC, Escherichia coli, E. coli, disease severity, symptoms, disease control guidelines

Posted Date: February 4th, 2020

DOI: https://doi.org/10.21203/rs.2.12350/v3

Version of Record: A version of this preprint was published at BMC Genomics on February 10th, 2020.
See the published version at https://doi.org/10.1186/s12864-020-6555-7.

Abstract

Background: We investigated the association of symptoms and disease severity of shigellosis patients with genetic determinants of infecting *Shigella* and entero-invasive *Escherichia coli* (EIEC), because determinants that predict disease outcome per individual patient could be used to prioritize control measures. For this purpose, genome wide association studies (GWAS) were performed using presence or absence of single genes, combinations of genes, and k-mers. All genetic variants were derived from draft genome sequences of isolates from a multicenter cross-sectional study conducted in the Netherlands during 2016 and 2017. Clinical data of patients consisting of binary/dichotomous representation of symptoms and their calculated severity scores were also available from this study. To verify the suitability of the methods used, the genetic differences between the genera *Shigella* and *Escherichia* were used as control.

Results: The isolates obtained were representative of the population structure encountered in other Western European countries. No association was found between single genes or combinations of genes and separate symptoms or disease severity scores. Our benchmark characteristic, genus, resulted in eight associated genes and >3,000,000 k-mers, indicating adequate performance of the algorithms used.

Conclusions: To conclude, using several microbial GWAS methods, genetic variants in *Shigella spp.* and EIEC that can predict specific symptoms or a more severe course of disease were not identified, suggesting that disease severity of shigellosis is dependent on other factors than the genetic variation of the infecting bacteria. Specific genes or gene fragments of isolates from patients are unsuitable to predict outcomes and cannot be used for development, prioritization and optimization of guidelines for control measures of shigellosis or infections with EIEC.

Background

Shigellosis is caused by the gram-negative bacterium *Shigella* and can lead to dysentery (1). The genus *Shigella* is divided in four species; *Shigella dysenteriae*, *Shigella flexneri*, *Shigella boydii*, and *Shigella sonnei*. All *Shigella spp.* are genetically closely related to *Escherichia coli* to the extent that they should be classified as one species (2, 3). However, it is a taxonomical decision based on historical and clinical arguments that has maintained the current classification (4). Entero-invasive *E. coli* (EIEC) is a pathotype of *E. coli*, which also can cause dysentery (5, 6). Because of the similarity in pathogenetic features of EIEC and *Shigella spp*, differentiation using diagnostic laboratory tests is difficult (7).

As in many other countries, shigellosis is a notifiable disease in the Netherlands. This means that in each case health authorities are notified, and consequently, control measures are activated (8-11). These control measures consist of source tracing for every shigellosis case, which places a burden on our public health system. Case definitions for shigelloses in the Dutch guidelines require confirmation with culture techniques (8). The sensitivity of the culturing of S*higella spp.* and EIEC is low (12). Additionally, most

laboratories perform a molecular prescreening based on the *ipaH* gene, which is present in both *Shigella spp* and EIEC. From approximately half of fecal samples positive in the molecular prescreening an isolate cannot be obtained in culture (12, 13). Shigellosis cases that are diagnosed purely by molecular procedures are not notifiable.

In contrast to cultured *Shigella spp.*, infections with EIEC are not notifiable in the Netherlands. Because of the high genetic similarities, identical disease outcomes and the low sensitivity of culturing, the two infective agents are often not detected in culture at all or are misidentified. Consequently, accurate application of the guidelines is challenging (14). Genes of pathogens that are predictive for disease outcomes can help in the prioritization of infectious disease control measures. Moreover, the presence of genes is more easily detected by using molecular procedures as opposed to the current used culture techniques required for notification.

A few studies have investigated the association of virulence genes with disease severity for shigellosis, using Pearson's correlation and regression analyses (15, 16). In one of these studies, the virulence gene *sepA* was associated with abdominal pain and the combination of *sepA*, *sigA* and *iaI* genes with bloody stools (16). Another study found that detection of the *sen* (shET-2) gene was associated with diarrhea and the *virA* gene was associated with fever (15). Both studies had a limited sample number, did not correct for multiple testing, and in one study the presence of virulence genes was established using direct detection in fecal samples. This approach is problematic, because different *Enterobacteriaceae* present in fecal samples may carry these genes, for example, on average, 2-3 *E. coli* strains are detected in the feces of a single person (17). Therefore, assessment of single isolates would be more appropriate. Furthermore, the association with only a limited number of targeted virulence genes was conducted in these previous studies, while genomic approaches would analyze all harbored genes, gene variants, or other genetic content.

The purpose of our study is to investigate whether there is an association between symptoms and disease severity of the patients and genetic determinants of infecting *Shigella* and EIEC isolates in the Netherlands. To address this, microbial genome-wide association methods (GWAS) were applied. We hypothesize that genetic variants associated with symptoms or severity of disease allow development of specific molecular diagnostics that could predict the disease outcome per individual patient and prioritize the employment of control measures for infections with *Shigella spp* and EIEC.

Results

Data preparation and exploration

To assess whether other pathogens present in the fecal samples caused the symptoms and severity of patients, presence of symptoms and severity scores of patients with coinfection were compared to those of patients without coinfection. In 15.5% of the patients, a coinfection was detected. The symptom blood in stool, known as a typical symptom of shigellosis (18), was significantly less present in patients with a coinfection (chi-square, $p = 0.019$), while the presence of other symptoms was not statistically different

(chi-square, p > 0.05). The lower fraction of patients with coinfection that experienced blood in stool was also reflected in the de Wit severity score, in which blood in stool is a criterion with double weighing, as it was significantly lower for patients with coinfection (T-test, p = 0.017). The Modified Vesikari Score (MVS), in which blood in stool is not a considered factor, showed no significant difference between patients with and patients without coinfection (T-test, p = 0.076).

The assemblies of 277 isolates were used to construct a gene presence/absence table and k-mers of variable length. This resulted in a gene presence/absence table consisting of 2,890 core genes (i.e. present in all 277 isolates) and 9,869 genes in total. K-mer counting yielded 28,551,795 genetic variants.

A phylogenetic tree was created based on the core genome SNPs, and the distribution of the severity scores, coinfection and the effects of underlying diseases were visualized (Figure 1). The core SNP analysis resulted in some species-specific clusters. However, clusters that contain multiple species were also present (Figure 1). In addition, severity scores, effects of underlying diseases and coinfection were randomly distributed over the isolates in the tree (Figure 1). For the GWAS analysis, only isolates sequenced during this study and displayed in Figure 1 were used. However, for contextualization of the position of the isolates in this study compared to the global population structure of *Shigella spp.* and EIEC, an additional tree was inferred including genomes from each of the main lineages and phylogenetic groups (Additional File 1). It showed that the population structure of our EIEC isolates was mainly concentrated in three clusters containing ST270, ST6 and ST99 based on isolates from the United Kingdom (UK) (19). The UK ST270 cluster corresponded with cluster 8, the large EIEC cluster from Pettengill *et al.*(3). In our analysis, EIEC isolates belonging to cluster 4, EIEC small or cluster 7, the EIEC/EHEC/EAEC cluster were not included (3). For *S. flexneri*, a few isolates related to travel to Asia belonged to PG6 and PG2 (Figure 1 and Additional File 1). However, the majority of isolates were PG3, consisting solely of isolates with serotype 2a or Y, and PG1, consisting of isolates of serotypes 1a, 1b, 1c, Yv and 4av. For *S. sonnei*, almost all isolates were of lineage III, only a few isolates within lineage II were detected (Figure 1 and Additional File 1). The presence of large clusters of EIEC isolates, the presence and distribution of serotypes over the PGs for *S. flexneri* and the predominance of *S. sonnei* lineage III were described before, and are representative of population structures found in other western European countries (19-22).

GWAS using gene presence/absence of single genes

None of the tested symptoms and severity scales resulted in significantly associated genes with a sensitivity and specificity above 85%. However, eight significantly associated genes were found with sensitivity above 92% and a specificity of 87% for the characteristic “genus”, that was used as a benchmark to evaluate algorithm performance. The gene with the highest association, produces a hypothetical protein and had a Benjamini Hochberg corrected p-value of 7.01E-27 and a sensitivity and specificity of 99% and 87%, respectively.

Additionally, the p-values of all characteristics were compared to random permutation datasets by plotting the log transformed expected and observed p-values against each other (Figure 2). The gene

associations with the tested severity scales (Figure 2A and 2B) and symptoms (Figure 2C) displayed similar plots as the random permutation datasets, indicating a performance as random cases. This did not apply to the benchmark characteristic “genus”, that plot showed a clear difference between expected and observed p-values, which was supported by the low Benjamini Hochberg corrected p-values (Figure 2D).

It followed from the sensitivity analysis based on the benchmark characteristic “genus” that genes present in 0.7% of total isolates within the smallest group (*Escherichia*, n=30), corresponding to two isolates of the total number of isolates, resulted in significant p-values. This indicated that a gene presence in a minimum of two isolates from the smallest group was enough to detect significance, if these genes were not present in the other larger group (Additional File 2).

GWAS using gene presence/absence of multiple genes

The generated random forest model, created using isolates from the training set resulted in an out-of-bag (OOB) estimate of error rates when testing the isolates from the test set. A random error rate of 66.7% for the severity scores and 50% for the symptoms and genus was expected, as respectively three and two classes were predicted. OOB error rates in the created random forest models using 5000 trees for the prediction of symptoms and severity scales of patients were as expected for random datasets when applied to the test set. Error rates ranged from 40.8% to 53.1% for all symptoms and 65.1% to 70.1% for the two severity scales (Table 1). The construction of additional trees did not lead to better predicting models.

In contrast, the OOB error rate of the model that predicted the benchmark characteristic genus was 15.9%, much lower than the random expected error rate of 50% (Table 1). The created model for genus prediction was further explored by examining the location of the misclassified isolates in the phylogenetic tree (Figure 1). Comparing them with the traditional laboratory results that were obtained during the IBESS-study showed that six out of ten discrepant isolates were so-called hybrid isolates and also had an uncertain assignment using the traditional laboratory tests (Table 2).

GWAS using k-mers

Associating k-mers with different characteristics using Pyseer did not lead to any significant k-mers for abdominal pain, abdominal cramps, blood in stool, fever, headache, mucus in stool, nausea, vomiting, and the severity score of MVS (Table 1). In contrast, 156 k-mers were associated with diarrhea, however, all k-mers had an invalid chi squared test and likelihood-ratio test (LRT) p-values higher than 0.313. The de Wit severity score resulted in 17 associated k- mers, whereof 15 k-mers with an LRT p-value lower than 0.05. An assembly of these 15 k-mers resulted in a single consensus sequence of 100 bp, based on overlapping k-mers. A BLASTn search of the consensus sequence against the database of the National Center for Biotechnology Information (NCBI, Bethesda, USA) revealed that the significant k-mers are located between two genes (Additional Figure 3), including a type II toxin-antitoxin gene (AYE47152.1) and a gene coding for DUF1391 (AYE48123.1), a protein of unknown function. A potential promoter

region in the k-mer was found with a -10 box (CATTATTTT) at position 58 and a -35 box (TTGACG) at position 36 of the sequence (Additional Figure 3).

To validate the potential of the k-mer to predict the severity score of de Wit scale, the k-mer was queried by BLAST against a database with all isolate assemblies from our study. For every sample, the bit-score of the best scoring hit was plotted against the corresponding severity score (Figure 3A). Roughly, three groups resulted, one with a bit-score of >175 corresponding with a full-length match with the k-mer, one with a bit-score of 50-175 corresponding to a partial match and <50 corresponding to no match. Subsequently, the Kruskal-Wallis test was performed to investigate the difference in the de Wit severity score between the groups (Figure 3B). No statistically significant difference between the groups was found, with a p-value of 0.6.

To check the suitability of the Pyseer method for the association of k-mers with characteristics in our data-set, the benchmark characteristic “genus” was used and resulted in 3,036,507 potential associated k-mers.

Discussion

The purpose of our study was to investigate associations between genetic determinants of infecting *Shigella spp.* and EIEC isolates and the symptoms and disease severity of the patients. If such associating genetic determinants were found, diagnostics could be developed that predict the severity of the resulting disease. Additionally, it could guide prioritization and optimization of infectious disease control measures regarding shigellosis. In the Netherlands, the severity predicting capabilities of genes of other pathogens have been used previously in prioritization of control measures. In 2016, case definitions for Shiga producing *E. coli* (STEC), another pathotype of *E. coli*, were extended from culture confirmation alone to the detection of STEC by Polymerase Chain Reaction (PCR) targeting the stx_1 and stx_2 genes and particular virulence genes. These combination of genes within STEC bacteria are known to have associations with a higher risk for severe disease and clinical complications (24).

However, for *Shigella spp.* and EIEC in the present study, the association of the presence or absence of single genes resulted in no statistically significant association between genes with specific symptoms or severity scores with high sensitivity and specificity. Second, the association of multiple genes resulted again in no statistically significant association with specific symptoms and severity scores of patients, indicating that no complex genetic interactions that may explain disease severity could be found. Third, the association of k-mers resulted in a consensus sequence consisting of multiple aligned k-mers that was associated with a high severity score of de Wit. The sequence of 100 bp, containing multiple associated k-mers, was located between two genes with a putative promoter region with an optimal inter-base distance of 16 bases but an unclear TATAAT box. When blasting the consensus k-mer against all assemblies, three difference bit scores were observed, suggesting there are three different genetic variants of this locus. Performing a Kruskal-Wallis test on these three different bit score groups, showed that the k-mer was not valid ($p = 0.6$), and presumably was a false positive.

In our study, the genes that were associated with specific symptoms in earlier studies (15, 16), were not confirmed. In another study that was conducted in Brazil among children with shigellosis, *sepA* was associated with abdominal pain, and the combination of *sepA*, *sigA* and *ial* genes with bloody diarrhea (16). However, it is not clear if univariate or multivariate testing for virulence genes was performed. In another study from Brazil, a case-control study was conducted. They found that the *sen* (shET-2) gene was associated with diarrhea in children in general, but not with specific symptoms of shigellosis patients. They associated the *virA* gene with fever in children with shigellosis, however *virA* was also found in 44% of controls (15). In our study, we have used a larger sample size consisting of patients with other demographics in another setting, analyzed all genes harbored instead of a predefined selection, used other methods with higher resolution as it was based on whole genomes, and included correction for multiple testing.

Because all algorithms used in our study generated negative results for association, the “genus” was also tested as a benchmark. The algorithms used performed adequate, as they resulted in relevant genetic variants. Furthermore, a sensitivity analysis indicated that the group distribution of the characteristic “genus” was suitable for significant detection of associated single genes. This characteristic had an adverse unequal group distribution of 10% versus 90%, indicating that the number of isolates and the distribution over the groups was suitable for associating genetic content with all symptoms and severity, except for “diarrhea”, which was the only characteristic with a more unequal group distribution than “genus”. Moreover, other studies found genetic variants significantly associated with their tested traits using the microbial GWAS methods that were used in our study (25-29).

Using Scoary, single genes that had association with the characteristic “genus” were found, with low p-values and high sensitivity and specificity. Further, with Pyseer, over 3,000,000 potentially associated k-mers were found. This is in concordance with another study that demonstrated the suitability of k-mers for identification of *Shigella spp.* and *E. coli* isolates based on whole genome sequences (30). Moreover, using Random Forest, OOB estimate error rate for the benchmark characteristic “genus” was 15.9%. This indicated that the model that predicts the genus of unknown isolates performed better than random, however, it does not accurately predict the genus of some isolates. Notably, six out of ten discrepant isolates also had an uncertain assignment with traditional laboratory tests. If we exclude these isolates, the OOB estimate error rate is 1.9%, indicating that it was not the method used but rather the nature of these isolates and their possession of characteristics of both *Shigella spp.* and *E. coli* that caused the uncertain assignments. The Random Forest method performed almost equally as well as the traditional laboratory tests and could be used for identification of the genus if whole genome data is available, although more isolates should be tested to validate this. Additionally, it would be useful to test the applicability of Random Forest for identification to species and serotype level. Furthermore, in a future study, the results of the traditional laboratory tests specifically can be associated with genetic variants. Consequently, if associated variants could be found, traditional tests could be omitted. This will save costs in workflows that already consist of draft genome sequencing of isolates for other purposes, for instance surveillance.

In addition to the methods using gene presence/absence and k-mers that were used in our study, other types of genetic variants can be used as input for microbial GWAS (31). The k-mer approach used in this study is able to detect different genetic variants such as SNPs, indels, variable promotor regions and gene content simultaneously (32). This indicates that adding purely SNP-based methods to the methods used is redundant as SNPs are already encompassed in the k-mer method performed. Another genetic variant that can be used in GWAS is based on De Bruijn Graphs. However, it is mainly based on the creation of overlaps of k-mers, therefore, it probably would not generate associations with symptoms or disease severity using the data from our study (33).

One of the strengths of our study was the availability of isolates representative

of the population structure encountered in other western European countries, as well as the clinical data of the patients that they were infecting. Second, results of the traditional laboratory tests performed to determine the species of the bacteria were available for all isolates. Finally, another strength of our study is that several potential genetic variants were associated with the trait “genus”, and a sensitivity analysis was performed, both proving the suitability of the algorithms used.

Some considerations with regard to our study should be taken into account. The impact of several factors regarding host-variability is unknown, as the symptoms and severity of disease were characteristics of the patients and not directly of the bacterial isolates. First, the immune status of the patients was not taken into account because data was not available, although the need for correction of the effects of underlying disease was investigated. Second, the clinical characteristics used in our study were self-reported and not objectively measured, therefore subject to the judgment and memory of the patients. To overcome these difficulties of host-variability, an infection model can be used for future investigations into genetic factors of *Shigella* isolates that influence the disease severity of patients. Because *Shigellaspp.* are host-adapted to humans only, recently developed human intestinal enteroids are more appropriate for this purpose than animal models (34). Additionally, inconsistencies between the two scoring methods used were present (Figure 1 and Additional File 4). Because each scale uses different criteria with different weighing for calculation of the score, patients can be classified in different severity classes depending on the severity scale used. Therefore, conclusions based on research into severity of gastro-intestinal infections in general are highly dependent on the chosen severity scale. To rule out this dependency, in this study, both scores were associated with genetic content separately. Another consideration was that genus level was associated as a characteristic, while other GWAS studies have concentrated on bacterial isolates of the same species (35, 36). However, according to multiple research groups (3, 37, 38) *Shigella spp.* and *E. coli* should be considered as one species based on their genetic relatedness, if present, their differences are more phenotypical. Next to this, the number of isolates for *S. boydii* and *S. dysenteriae* in our study were inadequate with two and no isolates, respectively. However, we believe the total number of isolates to be adequate, as studies with similar sample sizes have been performed in the past in which genetic variation in pathogens was identified that had predictive value for the course of disease (29, 39). Finally, the dataset used only contained isolates encountered in the

Netherlands, resulting in a geographical biased set (40, 41). Therefore, to avoid missing serotypes in future studies, the current dataset should be supplemented with isolates from other geographic areas.

Conclusions

Using several microbial GWAS methods, genetic variants in *Shigella spp.* and EIEC that can predict specific symptoms or a higher disease severity were not found. In contrast to adjustment of the guidelines of STEC, genes or gene fragments that indicate higher risks for a more severe course of disease does not appear to exist for shigellosis, whether caused by *Shigella* or EIEC, using the dataset in our study. Therefore, the bacterial specific genes or gene fragments from patient isolates are not suitable to predict outcomes in individual patients or to use in development, prioritization and optimization of guidelines for control measures of shigellosis or EIEC. As GWAS in our study associated genetic fragments with genus, future studies can be performed in which GWAS could support the distinction of *Shigella spp.* from EIEC. Additionally, the prediction of results of traditional laboratory tests using draft genome sequences could be performed using GWAS. The results of these suggested follow-up studies could improve diagnostics and guidelines for control measures of shigellosis.

Methods

Bacterial isolates and clinical data

The data used in our study was collected during the Invasive Bacteria *E. coli-Shigella* Study (IBESS). IBESS was a cross-sectional study in the Netherlands, of which one of the aims was to fill the gap of knowledge about the incidence, clinical implications and impact on public health of infections caused by EIEC. During this study, in 2016 and 2017, EIEC and *Shigella* isolates were collected, together with epidemiological patient data (van den Beld *et al.*, manuscript submitted). Isolates were identified using an identification scheme, using traditional laboratory tests as previously described (42). In short, it consists of a Polymerase Chain Reaction (PCR) of the *ipaH* gene, followed by thorough phenotyping and classical *E. coli* and *Shigella* O-antigen serotyping by agglutination. The draft genome sequences of a set of 277 bacterial isolates, of which patient data was available, were used as genetic input data. The set comprises *S. sonnei* (n=163), *S. boydii* (n=1), *S. flexneri* (n=77), EIEC (n=30), provisional *Shigella* (n=5), which are *Shigella* isolates with an undescribed serotype, and one isolate of which the distinction between *S. flexneri* and EIEC was unclear, using the traditional laboratory tests.

The clinical characteristics that were used in this GWAS study were symptoms and disease severity of patients infected with *Shigella spp.* or EIEC isolates included in IBESS. For all patients, a list of symptoms including abdominal pain, abdominal cramps, blood in stool, diarrhea, fever, headache, mucus in stool, nausea, and vomiting was available and was used as binary input (Additional File 4). Additionally, disease severity was calculated using two severity scales, both are modifications of the Vesikari scale, a widely used method in clinical studies (43). These modifications, the MVS (44) and the modified score of de Wit *et al.* (45), were both developed and validated for outpatient settings in high-resource areas. With

these severity scores, lower scores indicate a milder course of disease (44, 45). The calculated scores were stratified into scales representing mild, moderate and severe disease according to their own categorization and used as dichotomous input in the GWAS methods that were assessing the presence/absence of genes in this study (Additional File 4). For the de Wit severity score, a score of 0-3 is considered mild, 4-6 moderate and ≥7 severe (45). The designers of the MVS score consider a score of 0-8 as mild, 9-10 as moderate and ≥11 as severe disease (44). Alternatively, for the GWAS method that assessed k-mers, both severity scores were used as a continuous input.

Laboratories that participated in IBESS, reported other bacteria, viruses and parasites detected in the fecal samples by molecular, culture and microscopy methods as well (Additional File 4). Other pathogens were detected in 15.5% of the patients from whom the *Shigella spp.* and EIEC isolates were isolated that were used in this GWAS study. The statistical differences in symptoms and severity scores of patients with and without coinfections were assessed, in order to establish if these pathogens have impact on the symptoms experienced. If necessary, data about effects of underlying diseases of the patients were used as a correction. Additionally, the genus of the bacteria was used as directly derived characteristic to use as a control to verify the suitability of the methods used. The patient data used in the GWAS studies are depicted in Additional File 4.

Genome sequencing and data preparation

DNA isolation and short-read Illumina sequencing was performed as described earlier (42). For preparation of the genomes, an in-house assembly pipeline available at GitHub (https://github.com/Papos92/assembly_pipeline) was used. It consists of raw data quality assessment using FastQC v. 0.11.8 (46) and MultiQC v. 1.7 (47), read trimming using ERNE v. 2.1.1 (48), contamination filtering using CLARK v. 1.2.5.1 (49), contigs and scaffold assembly using SPAdes v. 3.10.0 (50), and assembly quality assessment using QUASTv. 4.4 (51). Contigs smaller than 200 bp or with a coverage <10 were filtered out. CheckM v. 1.0.11 (52) (taxonomy_wf: genus 'Shigella') was used for quality assessment, genome completeness and contamination checks of the assemblies. Isolates with completeness above 99% and a contamination below 2% were included for further analyses. Sequences of isolates are available from the Sequence Read Archive (SRA) with study number PRJEB32617 (https://www.ncbi.nlm.nih.gov/sra/), accession numbers are indicated in detail in Additional File 3.

Prokka v. 1.1 (53) was used without cleanup for annotation of the genomes. Gene presence/absence for all genomes was determined using Roary v. 3.12.0 (54), using a BLAST identity cutoff of 80% and with paralog splitting disabled. Phylogenetic trees based on core genome SNPs were constructed with Parsnp v.1.2 (55). To include contextualization of the position of the isolates sequenced in this study relative to the main lineages of EIEC and *S. sonnei* and the phylogenetic groups (PG) of *S. flexneri* randomly selected genomes from each lineage or phylogenetic group were included in the phylogenetic tree (3, 19, 22, 56, 57). Details of these representatives and their accession numbers are depicted in Additional File 5. Data was visualized using iTol v. 4.3 (58).

GWAS using gene presence/absence of single genes

Scoary v. 1.6.16 (26) was used to associate gene presence and absence with the symptoms and severity of patients and the genus of the isolates, using a p-value cut-off of 0.5. Output was generated as a list of associated genes per characteristic with their best pairwise comparison p-values, sensitivity, and specificity. For each characteristic, as benchmark, a 1000 random datasets were created by shuffling the original traits randomly for a thousand times using a custom script (59). For each symptom and severity scale, 1000 genes with the lowest 'best pairwise p-value' were used, this p-value takes population structure into account. The observed p-values of the traits were log transformed and plotted against the log transformed expected p-values of the permutation benchmark using a custom script (59). For the characteristic 'genus', Benjamini-Hochberg's method for multiple comparisons correction is used instead of pairwise p-values as the latter cannot be used to find genetic differences between the species and genera. Additionally, a sensitivity analysis including corrections for multiple testing and the population structure was performed. To assess the minimal number of isolates with gene presence that is needed to detect a significant association, the corrected p-values from the output for the association of genes with the characteristic "genus" were log transformed and plotted against the percentage of isolates in which the corresponding genes were present (Additional File 2).

GWAS using gene presence/absence of multiple genes

Random Forest classification was executed using R v. 3.4.4 (60) and the randomForest package v. 4.6-14 (61). The gene presence/absence table derived from Roary and the symptoms and severity of patients and the genus of the isolates were used as input. The dataset was divided over a test set and a training set. Potential class size differences were corrected by using two-thirds of the smallest class as the sample size to create models based on gene presence/absence of multiple genes in the training set, using 5000, 8000 and 10,000 trees respectively. The performance of these models was validated by predicting the outcome of each trait using the genomes of the isolates in the test set.

GWAS using k-mers

To generate the k-mers that were associated with the characteristics, first, a population structure estimation was made using mash v. 2.0 (62). Second, k-mer counting was performed using fsm-lite v. 2.0.3, and the optimal number of dimensions to use as co-factors in the analysis was determined (32, 63). Subsequently, to estimate the effect of the k-mers on the severity scores and patient symptoms, Pyseer v. 1.1.2 was used with the following settings: a maximum of six dimensions, a filter p-value of 1E-8, a minimum allele frequency of 0.02 and a maximum allele frequency of 0.98.

The resulting k-mers were aligned using ClustalW v. 2.1, which resulted in one consensus sequence (64). To identify the position of the k-mers in the genome, the resulting consensus sequence was aligned using the nucleotide Basic Local Alignment Search Tool (BLASTn) v. 2.8.1 with default settings (65). To investigate whether the k-mer contained a promotor, BPROM was used (66).

In addition, to validate the association of the resulted consensus k-mer with the characteristics it was aligned against a BLAST database of all assembled genomes from this study, created using BLASTn v.

2.2.31+ (67). Best scoring hits including bit-score were collected for all isolates, plotted against the severity score and a Kruskal-Wallis test was performed using GraphPad prism v. 7.04 (GraphPad Software, La Jolla California USA).

Declarations

Ethics approval and consent to participate

Data from the IBESS-study was used, that is registered as observational study under number 23481 in the Dutch Trial Register. The medical ethics review board (METC) in Utrecht, the Netherlands, stated that this study was not subject to the requirements of “medical research with human subjects” laws (protocol number 15-414/C). All patients gave their informed consent for participation, in case of minor, parents or guardians participated. Data handling complied with the Dutch Personal Data Protection Act and with the EU General Data Protection Regulation.

Consent for publication

Not applicable.

Availability of data and material

Sequences of isolates were submitted to the European Nucleotide Archive (ENA, EMBL-EBI, Cambridge, United Kingdom) with study number PRJEB32617 (https://www.ebi.ac.uk/ena).

The in-house assembly pipeline is available on Github: https://github.com/Papos92/assembly_pipeline. All scripts and commands used for data preparation and GWAS are available on zonodo (https://doi.org/10.5281/zenodo.3626738). All other data generated or analyzed during this study are included in this published article and its supplementary files.

Competing interests

The authors declare that they have no competing interests.

Funding

The collection of patient data during the IBESS-study was (partially) supported by the research fund of the Dutch National Institute for Public Health and environment (RIVM) for local Public Health Services. BD was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004.

Authors’ contributions

FR, MK, JR, and MB conceptualized the project. AH, BD, AZ and MB designed the experiments. AH and AZ performed experiments and analyzed the data. AH, BD, AZ and MB interpreted results. AZ and MB

supervised the project. AH and MB wrote the manuscript. All authors read, reviewed and approved the final manuscript.

Acknowledgements

The IBESS group provided isolates and patient data, and consists of:

- J. C. van den Beld, Infectious Disease Research, Diagnostics and laboratory Surveillance, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
- Warmelink, Public Health Service GGD Groningen, Groningen, the Netherlands
- M. D. Kooistra-Smid, Department of Medical Microbiology, Certe, Groningen, the Netherlands and Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
- W. Friedrich, Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
- A. G. Reubsaet, Infectious Disease Research, Diagnostics and laboratory Surveillance, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
- W. Notermans, Infectious Disease Research, Diagnostics and laboratory Surveillance, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
- W. F. Petrignani, Public Health Service GGD Amsterdam, Amsterdam, the Netherlands and National Coordination Centre for Communicable Disease Control, Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands
- H. F. M. Waegemaekers, Public Health Service GGD Gelderland-Midden, Arnhem, the Netherlands and National Coordination Centre for Communicable Disease Control, Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands
- W. A. Rossen, Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
- P. van Dam, Amsterdam Health Service, Amsterdam, the Netherlands
- Svraka-Latifovic, CBSL, Tergooi, Hilversum, the Netherlands
- J. Verweij, Elisabeth-TweeSteden Hospital, Laboratory for Medical Microbiology and Immunology, Tilburg, the Netherlands
- E. S. Bruijnesteijn van Coppenraet, Isala, Laboratory for Medical Microbiology and Infectious diseases, Zwolle, the Netherlands
- Waar, Izore, Centre for Infectious Diseases Friesland, Leeuwarden, the Netherlands
- Hermans, Jeroen Bosch Ziekenhuis, Laboratorium Medische Microbiologie, ‘s-Hertogenbosch, the Netherlands

- L. J. Hess, LabMicTA, Laboratory for Medical Microbiology and Public Health, Hengelo, the Netherlands
- J. M. van Mook, Microvida location Amphia, Breda, the Netherlands
- C. Bergmans, Microvida location Bravis, Roosendaal, the Netherlands
- R. Jansen, OLVG, Medical Microbiological Laboratory, Amsterdam, the Netherlands
- H. B. van de Bovenkamp, PAMM Laboratory for Medical Microbiology, Veldhoven, the Netherlands
- Demeulemeester, SHL-group, Etten-Leur, the Netherlands
- Reinders, St. Antonius Ziekenhuis, Medical Microbiology and Immunology, Nieuwegein, the Netherlands
- F. M. Linssen, Zuyderland Medical Centre, Medical Microbiology, Heerlen, the Netherlands
- And all adjacent Public Health Services

Abbreviations

BLAST Basic local alignment search tool

EIEC entero-invasive *Escherichia coli*

GWAS genome wide association studies

IBESS Invasive Bacteria *E. coli-Shigella* Study

LRT likehood-ratio test

MVS Modified Vesikari Score

NCBI National Center of Biotechnology Information

OOB out-of-bag

PCR Polymerase Chain Reaction

PG Phylogenetic Group

RF Random Forest

SNP single nucleotide polymorphism

ST sequence type

STEC Shiga toxin- producing *Escherichia coli*

References

1. Hale TL. Genetic basis of virulence in *Shigella species*. Microbiol Rev. 1991;55(2):206-24.
2. Lan R, Reeves PR. *Escherichia coli* in disguise: molecular origins of *Shigella*. Microbes Infect. 2002;4(11):1125-32.
3. Pettengill EA, Pettengill JB, Binet R. Phylogenetic analyses of *Shigella* and enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers: whole-genome comparative analysis does not support distinct genera designation. Front Microbiol. 2015;6:1573.
4. Strockbine NA, Maurelli, A.T. Genus XXXV. *Shigella*. Bergey's manual of systemic bacteriology. 2. second ed. New York, USA: Springer science and business Media, Inc.; 2005. p. 811-23.
5. Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. J Infect Dis. 1987;155(3):377-89.
6. DuPont HL, Formal SB, Hornick RB, Snyder MJ, Libonati JP, Sheahan DG, et al. Pathogenesis of *Escherichia coli* diarrhea. N Engl J Med. 1971;285(1):1-9.
7. van den Beld MJ, Reubsaet FA. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. Eur J Clin Microbiol Infect Dis. 2012;31(6):899-904.
8. RIVM. LCI Richtlijn shigellose 2017 [Available from: https://lci.rivm.nl/richtlijnen/shigellose.
9. EU. Comission Implementing Decision (EU) 2018/945 of 22 June 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions 2018 [updated 6 July 2018.
10. CDC. Shigellosis (*Shigella spp.*) 2017 Case Definition 2017 [Available from: https://wwwn.cdc.gov/nndss/conditions/shigellosis/case-definition/2017/.
11. CDNA. Shigellosis Surveillance Case Definition 2018 [Available from: http://www.health.gov.au/internet/main/publishing.nsf/Content/cda-surveil-nndss-casedefs-cd_shigel.htm.
12. Van Lint P, De Witte E, Ursi JP, Van Herendael B, Van Schaeren J. A screening algorithm for diagnosing bacterial gastroenteritis by real-time PCR in combination with guided culture. Diagn Microbiol Infect Dis. 2016;85(2):255-9.
13. Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, et al. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. Lancet. 2016;388(10051):1291-301.
14. Lede IO K-DM, van den Kerkhof JHTC, Notermans DW. Gebrek aan uniformiteit bij meldingen van Shigatoxineproducerende *Escherichia coli* en *Shigella* aan en door GGDen. Infect Bull. 2012;23:116-8.
15. Bona M, Medeiros PH, Santos AK, Freitas T, Prata M, Veras H, et al. Virulence-related genes are associated with clinical and nutritional outcomes of *Shigella*/Enteroinvasive *Escherichia coli* pathotype infection in children from Brazilian semiarid region: A community case-control study. Int J Med Microbiol. 2019;309(2):151-8.
16. Medeiros P, Lima AAM, Guedes MM, Havt A, Bona MD, Rey LC, et al. Molecular characterization of virulence and antimicrobial resistance profile of *Shigella species* isolated from children with

moderate to severe diarrhea in northeastern Brazil. Diagn Microbiol Infect Dis. 2018;90(3):198-205.

17. Gordon DM, O'Brien CL, Pavli P. *Escherichia coli* diversity in the lower intestinal tract of humans. Environ Microbiol Rep. 2015;7(4):642-8.
18. Lampel KA, Formal SB, Maurelli AT. A Brief History of Shigella. EcoSal Plus. 2018;8(1).
19. Cowley LA, Oresegun DR, Chattaway MA, Dallman TJ, Jenkins C. Phylogenetic comparison of enteroinvasive *Escherichia coli* isolated from cases of diarrhoeal disease in England, 2005-2016. J Med Microbiol. 2018;67:884-8.
20. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H, Day M, et al. Genomic epidemiology of *Shigella* in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance. Sci Rep. 2018;8(1):7389.
21. Connor TR, Barker CR, Baker KS, Weill FX, Talukder KA, Smith AM, et al. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. Elife. 2015;4:e07335.
22. Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, et al. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012;44(9):1056-9.
23. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53(8):2410-26.
24. RIVM. LCI richtlijn Shigatoxineproducerende *E.coli* (STEC)-infectie 2016 [Available from: https://lci.rivm.nl/richtlijnen/shigatoxineproducerende-ecoli-stec-infectie.
25. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci U S A. 2013;110(29):11923-7.
26. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17(1):238.
27. Bazinet AL. Pan-genome and phylogeny of *Bacillus cereus* sensu lato. BMC Evol Biol. 2017;17(1):176.
28. Wegener A, Broens EM, Zomer A, Spaninks M, Wagenaar JA, Duim B. Comparative genomics of phenotypic antimicrobial resistances in methicillin-resistant *Staphylococcus pseudintermedius* of canine origin. Vet Microbiol. 2018;225:125-31.
29. Cremers AJH, Mobegi FM, van der Gaast-de Jongh C, van Weert M, van Opzeeland FJ, Vehkala M, et al. The Contribution of Genetic Variation of *Streptococcus pneumoniae* to the Clinical Manifestation of Invasive Pneumococcal Disease. Clin Infect Dis. 2019;68(1):61-9.
30. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli* and *Shigella Species* from Whole-Genome Sequences. J Clin Microbiol. 2017;55(2):616-23.

31. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25:17-24.
32. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7:12797.
33. Jaillard M, Lima L, Tournoud M, Mahe P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet. 2018;14(11):e1007758.
34. Koestler BJ, Ward CM, Fisher CR, Rajan A, Maresso AW, Payne SM. Human Intestinal Enteroids as a Model System of Shigella Pathogenesis. Infect Immun. 2019;87(4).
35. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2013;45(10):1183-9.
36. Alam MT, Petit RA, 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. Genome Biol Evol. 2014;6(5):1174-85.
37. Brenner DJ, Fanning GR, Steigerwalt AG, Orskov I, Orskov F. Polynucleotide sequence relatedness among three groups of pathogenic *Escherichia coli* strains. Infect Immun. 1972;6(3):308-15.
38. Brenner DJ, Steigerwalt AG, Wathen HG, Gross RJ, Rowe B. Confirmation of aerogenic strains of *Shigella boydii* 13 and further study of *Shigella* serotypes by DNA relatedness. J Clin Microbiol. 1982;16(3):432-6.
39. Tunjungputri RN, Mobegi FM, Cremers AJ, van der Gaast-de Jongh CE, Ferwerda G, Meis JF, et al. Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease. MBio. 2017;8(1).
40. Khatun F, Faruque AS, Koeck JL, Olliaro P, Millet P, Paris N, et al. Changing species distribution and antimicrobial susceptibility pattern of *Shigella* over a 29-year period (1980-2008). Epidemiol Infect. 2011;139(3):446-52.
41. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, et al. *Shigella* isolates from the global enteric multicenter study inform vaccine development. Clin Infect Dis. 2014;59(7):933-41.
42. van den Beld MJC, de Boer RF, Reubsaet FAG, Rossen JWA, Zhou K, Kuiling S, et al. Evaluation of a culture dependent algorithm and a molecular algorithm for identification of *Shigella spp.*, *Escherichia coli,* and enteroinvasive *E. coli* (EIEC). J Clin Microbiol. 2018;56:e00510-18.
43. Ruuska T, Vesikari T. Rotavirus disease in Finnish children: use of numerical scores for clinical severity of diarrhoeal episodes. Scandinavian journal of infectious diseases. 1990;22(3):259-67.
44. Freedman SB, Eltorky M, Gorelick M, Pediatric Emergency Research Canada Gastroenteritis Study G. Evaluation of a gastroenteritis severity score for use in outpatient settings. Pediatrics. 2010;125(6):e1278-85.

45. de Wit MA, Kortbeek LM, Koopmans MP, de Jager CJ, Wannet WJ, Bartelds AI, et al. A comparison of gastroenteritis in a general practice-based study and a community-based study. Epidemiol Infect. 2001;127(3):389-97.
46. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.
47. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017.
48. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.
49. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32(24):3823-5.
50. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
51. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
52. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043-55.
53. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
54. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
55. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):524.
56. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H, Day M, et al. Horizontal antimicrobial resistance transfer drives epidemics of multiple *Shigella species*. Nat Commun. 2018;9(1):1462.
57. Baker KS, Campos J, Pichel M, Della Gaspera A, Duarte-Martinez F, Campos-Chacon E, et al. Whole genome sequencing of *Shigella sonnei* through PulseNet Latin America and Caribbean: advancing global surveillance of foodborne illnesses. Clin Microbiol Infect. 2017;23(11):845-53.
58. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44(W1):W242-5.
59. Hendriks ACA, Reubsaet FAG, Kooistra-Smid AMDM, Rossen JWA, Dutilh BE, Zomer AL, et al. Genome-wide association studies of Shigella spp. and Enteroinvasive Escherichia coli isolates demonstrate an absence of genetic markers for prediction of disease severity. https://doi.org/10.5281/zenodo.36267382020.
60. R_core_team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2018 [Available from: https://www.R-project.org/.
61. Liaw A, Wiener MJR. Classification and Regression by RandomForest. R News. 2002;2(3):18-22.

62. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.
63. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018;34(24):4310-2.
64. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8.
65. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1-2):203-14.
66. Solovyev W, Salamov A. Automatic annotation of microbial genomes and metagenomic sequences. In: Li RW, editor. Metagenomics and its applications in agriculture, biomedicine and environmental studies: Nova Science Pub Inc; 2011. p. 61-78.
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

Tables

Table 1 Results of Random Forest classification and k-mer association

Characteristic	Random Forest	K-mer association with Pyseer	
	OOB error rate	No. of k-mers	Lowest LRT p-value
MVS severity scale	70.1%	0	NA
De Wit severity scale	65.1%	17	0.015
Abdominal cramps	52.7%	0	NA
Abdominal pain	40.8%	0	NA
Blood in stool	41.2%	0	NA
Diarrhea	51.6%	156	0.313
Fever	47.7%	0	NA
Headache	46.6%	0	NA
Mucus in stool	43.3%	0	NA
Nausea	53.1%	0	NA
Vomiting	51.6%	0	NA
Genus	15.9%	3,036,507	1.94E-153

Table 2 Comparison of misclassified isolates with Random Forest to traditional laboratory testing

Isolate	Phenotype[a]	Random Forest (RF)[a]	Votes[b]	Location in SNP tree	Serotype *Shigella*/ *E. coli* (agglutination)	Properties against RF classification
IBESS811	E	S	0.99	Within *S. sonnei*	*S. sonnei* phase 1/ O-negative	Motility
IBESS97	E	S	0.80	Within *S. flexneri*	*S. flexneri*, inconclusive/ O135	Inconclusive Shigella serotype
IBESS1163	E	S	0.76	Within *S. flexneri*	*S. flexneri*, inconclusive/ O135	Inconclusive Shigella serotype
IBESS911	E	S	0.68	Within *S. flexneri*	*S. flexneri*, inconclusive/ O135	Inconclusive Shigella serotype
IBESS996	S	E	0.53	Within EIEC / *S. flexneri*	*S. flexneri* 3a/ O135	None, hybrid isolate[d]
IBESS988	S	E	0.56	Within EIEC / *S. flexneri*	*S. flexneri* 3b/ O135	None, hybrid isolate[d]
IBESS419	S	E	0.57	Within *S. flexneri*	Provisional/O-negative	None, hybrid isolate, provisional *Shigella*[d]
IBESS232	S	E	0.60	Within *S. flexneri*	Provisional/O-negative	None, hybrid isolate, provisional *Shigella*[d]
IBESS470	S	E	0.82	Within EIEC	Provisional/O-negative	None, hybrid isolate, provisional *Shigella*[d]
IBESS810	S	E	0.89	Within EIEC	Auto agglutinable[c]	None, hybrid isolate, provisional *Shigella*[d]

RF = Random Forest. [a]E = Escherchia, S = Shigella. [b]fraction of votes for classification in Random Forest. [c]In-silico serotype, using E. coli serotypeFinder 2.0 of the Center for Genomic Epidemiology (23): provisional/O-negative. [d] Hybrid isolates = isolates that possess characteristics of both Shigella spp. and E. coli

Figures

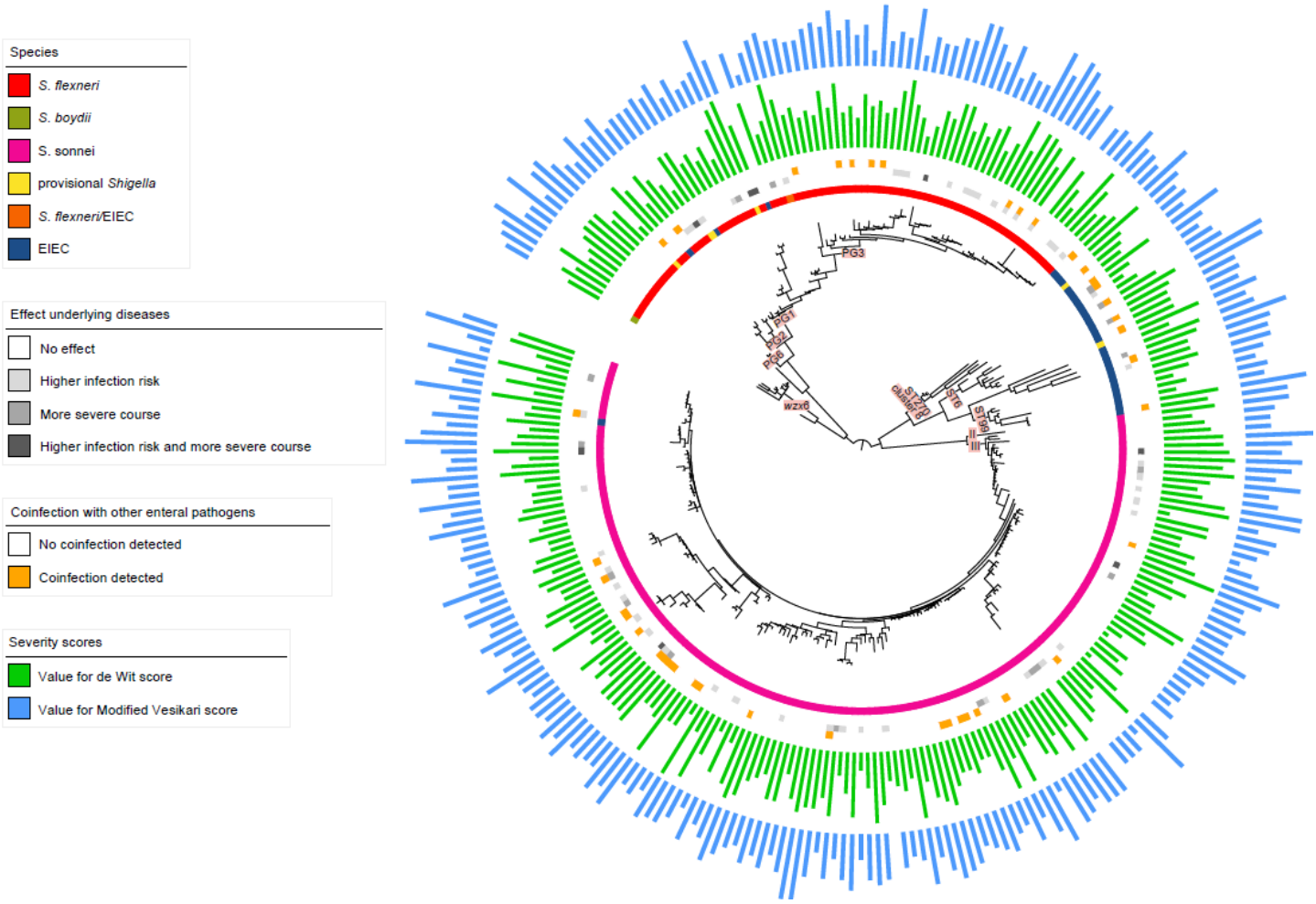

Figure 1

Phylogenetic tree based on core genome SNPs with species indication, underlying diseases and severity scores Within the salmon squares are the main lineages or phylogroups depicted. wzx6 = S. flexneri serotype 6. PGx = phylogenetic group of S. flexneri. STxxx = Warwick sequence type of EIEC. II and III = S. sonnei lineage II and III.

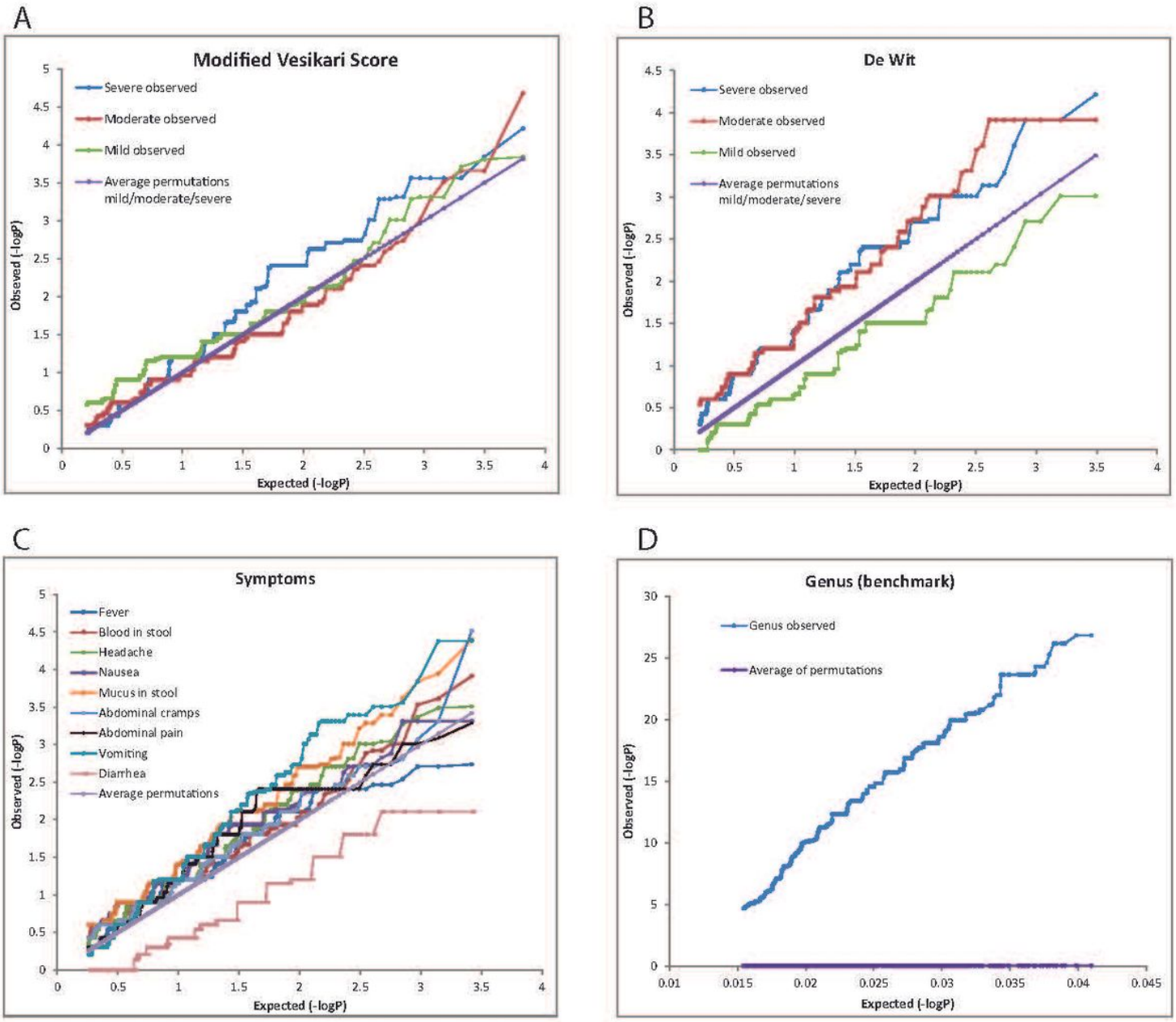

Figure 2

Results of Scoary: the expected versus the observed log transformed p-values Lilac lines indicate the outcomes of the permutation dataset. A. Best comparison test for association of gene presence/absence with de Wit severity score. B. Best comparison test for association of gene presence/absence with Modified Vesikari score. C. Best comparison test for association of gene presence/absence with symptoms. D. Benjamini Hochberg's test for association of gene presence/absence with genus.

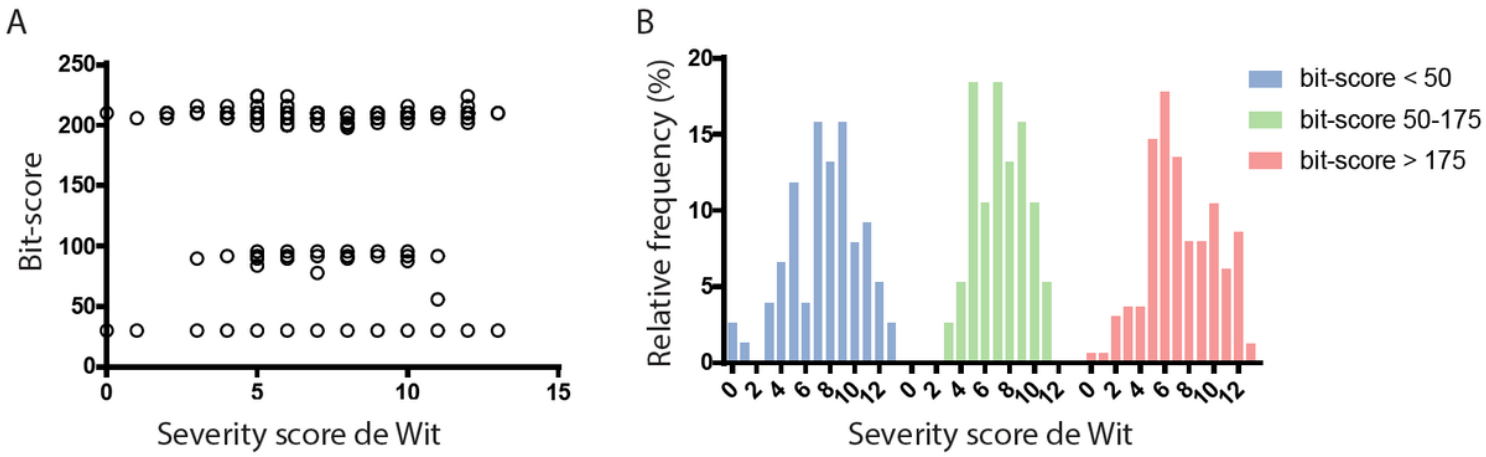

Figure 3

Blast result of k-mers resulting consensus on used isolates A. Blast results versus severity score. B. Histogram of the relative frequency of the severity scores in the dataset versus the severity score of de Wit, displayed for three bit-score categories.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile4REVISION.xlsx
- AdditionalFile6.pdf
- AdditionalFile2V2cropped.pdf
- AdditionalFile1croppedV3REVISION.pdf
- AdditionalFile5V2REVISION.xlsx
- AdditionalFile3V2croppedREVISION.pdf