# Introduction

Bacterial typing is critical for identification of outbreaks, tracking transmission and understanding the epidemiology of bacterial pathogens (REF). Knowledge which is key to effective healthcare intervention and public health policy. To address this, multiple typing schemes have been devised, all using different methods. Some are capable of typing bacteria across species and genus, while others are species or genus specific.

Hierarchical, core genome multi locus sequence typing (cgMLST) is able to type at both the genus and species levels [1]; covering high intra-species genetic diversity whilst being hierarchical and therefore able to be highly granular. SnapperDB type genes through core genome single nucleotide polymorphisms (SNPs), in a hierarchical manner, generating a SNP address in the process [2]. Add further description.

With the introduction of bacterial pathogen surveillance, particularly with the addition of whole genome sequencing, the number of isolates being typed has rapidly increased. It has been observed that even at the most granular level, typing schemes are sometimes grouping isolates together into large groups (REF). In this study we examine the effectiveness of three currently used typing schemes for the typing of *S. sonnei* during on-going surveillance. More description.

*Shigella sonnei* is transmitted person-to-person and is relatively less genetically diverse than other bacterial pathogens, such as *E. coli*, while still having enough genetic diversity to create a SNP-based maximum likelihood phylogeny with strong support (REF). Besides being a good model organism based on within population genetic diversity, *S. sonnei* is also interesting as strains circulating within the community of men who have sex with men (MSM) have a distinct epidemiology (REF). Rapid transmission within this community produces quick clonal expansion of strains within this community which can be identified on the phylogeny as highly related lineages, predominantly isolated from adult men (REF). Highly clonal lineages are particularly difficult to type with enough granularity to identify or track outbreaks. The variation of genetic diversity between MSM and non-MSM associated *S. sonnei* isolates enables us to examine the specific effects rapid transmission and/or high clonality of strains on the typing scheme outputs.

We have chosen to compare three commonly used typing schemes: core genome MLST, SNP address and sonneiTyping. All are used by The UKHSA for typing *S. sonnei* during surveillance. SonneiTyping is a *S. sonnei* specific typing scheme, based on genetic variation with a panel of genes, identified as core to *S. sonnei* [3]. All schemes are hierarchical, providing multiple levels of granularity.

# Methods

## Sample collection and sequencing

N samples, collected and sequenced by UKHSA between, 2016 and 2020, were included in this study (this accounts for all *S. sonnei* isolates sequenced by UKHSA during this time. Isolates were collected and sequenced according to standard UKHSA procedures.

## Typing methods

Three different typing schemes were compared, 1) the *S. sonnei* genotyping scheme (ref), 2) SNP address (ref), and 3) hierarchical, core genome MLST (ref). Every included isolate was typed using all three programs according to standard UKHSA procedure.

For the genotyping scheme, not all levels are designated for every isolate by the program, every missing level was, therefore, manual designated 0. Note that all higher/less granular levels are designated by the software, thus 1 or more ‘0’s were added on to the end of the program-generated genotype for every isolate without a genotype given at every granular level.

Each typing level generates a typing name for each isolate. Those with the same name are determined to be genetically similar by the typing scheme and are therefore referred to as ‘grouped together’ throughout, while typing names are referred to as ‘groups’.

SNP address levels were defined as just the number corresponding to that level, with the exception of SNP address which is the full address. Genotyping scheme group names include all prior levels. This is because the SNP address group names are unique at each level in the ANP address, while the genotyping scheme is only unique when the prior levels are included in the name.

## Year-based sub-setting

Year-based data subsets were, predominantly, cumulative, in that isolates from each year were added to the isolates from each prior year. This is a more accurate model of on-going surveillance than sub-setting isolates by individual years. In cases where individual year subsets are used or shown, this is clearly indicated.

## Group size determination

We used to definitions of group size to assess the how typing scheme group sizes change through time. The first method was used to model the real-life scenario of using typing schemes to type isolates collected as part of on-going surveillance. For this purpose, year-based data subsets were created before the group size was determined. In this way, group size is determined separately for each year-based subset, thus the maximum group size increases, as the sample size increases, every year.

The second method was used to assess changing group size through time without the effect of increasing sample size through time. This was achieved by determining group sizes from the full dataset rather than the year-based subsets.

We also quantified the number of groups with a size below threshold (10 and 50) to determine if the number of small groups changed across time. In this case, the sample set was subset into individual years and the groups size determined from these year subsets.

## Assessing the effect of rapid person-to-person transmission

MSM transmission – splitting the dataset into full, non-MSM and MSM-only to compare how this group with different transmission patterns and highly clonal sub-lineages in the phylogeny effects the utility of typing schemes. Patient sexuality was unknown, an MSM clade was defined by being predominantly from male patients above 16 years old. All isolates within these clades were defined as part of the MSM subset regardless of patient age or gender, as these clades were assumed to be circulating predominantly within the MSM community.

## Statistical analysis

We determined that typing levels had similar levels of granularity when the difference in groups size was within 10% of the total group size for either typing level.

We used Theil’s U to assess

# Results

## Level of granularity

The number of groups at each typing level was used to directly compare the level of granularity between the three typing schemes, and the rate of increase from less granular typing levels to more granular typing levels (Figure 1).

A picture containing screenshot, square, rectangle, line

Description automatically generated

Figure 1. Comparison of the number of groups at each level for each typing scheme.

Add description.

The genotyping scheme did not increase much in granularity between each level, in the end being nearly equivalent to the 100 SNP threshold level (Genotype = N groups, 100 SNP threshold = N groups). While the level of granularity of SNP address and cgMLST did not follow the exact same upwards trend, they two typing schemes were similarly granular with several of the typing levels (which?) being within a X% difference in number of groups for both typing levels being compared. CgMLST did reach higher granularity (X groups) than SNP threshold (X groups), however.

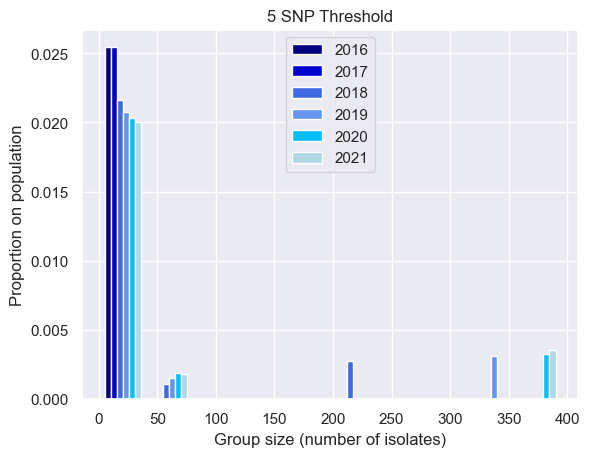
Most typing scheme levels are lacking in enough granularity to be able to identify and track outbreaks from surveillance data. Based on the level of granularity we believe the 10 SNP threshold and below or HC5 level and below to be useful for outbreak identification and tracking, thus this report focusses on these levels. Figures showing the results for other levels can be found in the supplementary data.

## Group size through time

To see if group size trended towards bigger group sizes with time, during on-going surveillance, we created histograms of group size of cumulative year-based subsets. We defined group size with two different methods 1) group size per year-based subset, and 2) groups size of the whole dataset.

When group size was defined for each year-based subset (method 1), the cumulative effect of increasing sample size with time resulted in increasing group size with time, across all typing schemes and levels (Figure 2). This was true even in the non-MSM dataset; however, the very large groups were absent when the MSM isolates were removed (Figure 2).

Add comparison between typing schemes.

 A picture containing text, screenshot, plot, diagram

Description automatically generated

Full dataset

Non-MSM dataset

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, plot, number

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 2. Group size increases with time during surveillance, with very large group size driven by MSM clades.

Very large groups, present in the full dataset (left), are missing in the non-MSM dataset. Add more granular levels too

A picture containing text, screenshot, diagram, plot

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

MSM dataset

Figure 3. MSM-only dataset group size through time, by SNP address level (left) and cgMLST level (right).

Group size increases with time, with some groups becoming very large. For SNP address, the majority of the dataset belongs to large groups (> 200 isolates) by 2018. At the more granular level (HC2), the cgMLST typing groups the MSM-only datatset into predominantly small groups (< 50 isolates) across all years. Add more granular levels too

When examining the MSM-clades in isolation, the number of very large groups (size?) increases rapidly with time (Figure 3). In the case of HC5, and SNP thresholds 5 and 10, the majority of the MSM-clade isolates belonged to these very large groups (Figure 3). HC2 seemed to retain a higher proportion of small group sizes with time than the aforementioned SNP address and cgMLST levels (Figure 3). Add the more granular levels too.

When groups size is determined from the dataset as a whole (method 2), the effect of increasing sample size is removed. Despite this, In the full dataset we still observe a trend of decreasing numbers of small groups (<50 isolates) and increasing large groups, with time (Figure 4). This trend is not as obvious when the MSM isolates are removed, however there is slight increase in the number of medium size groups (50-150 isolates) (Figure 4). Add typing scheme comparison.

When examining the MSM-clades in isolation, using method 2, we see a very different trend to both the full and non-MSM datasets. The first three years, group size is stable. In most cases there is then a decrease in the number of large groups and an increase in the small and medium groups. – look at this in the supplementary data too.

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generated

Non-MSM dataset

Full dataset

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, plot, number

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated A picture containing text, screenshot, plot, number

Description automatically generated

Figure 4. Group size increases with time during surveillance, even when accounting for the effect of increasing sample size with time.

Very large groups, present in the full dataset (left), are missing in the non-MSM dataset (right) , showing that very large groups size remains exclusively MSM clade related. Add more granular levels too

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

MSM dataset

A picture containing text, screenshot, plot, diagram

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 5. MSM-only dataset group size remains constant or decreases through time, by SNP address level (left) and cgMLST level (right).

Majority of MSM clades are grouped into large groups (SNP address and HC5 cgMLST level), however, the HC2 cgMLST level (top right) groups the MSM clades predominantly into small group sizes.

## Association between variables

Using Theil’s U to determine the ….. We can see that in the full dataset, there are high levels of reciprocal informativity between the cgMLST and SNP address typing schemes. They are not 100% informative of each other.

Genotyping scheme levels are highly informative of levels which are less granular, while the same is not true of cgMLST or SNP address (except the full SNP address, which is 100% informative of all SNP threshold levels).

A close-up of a grid

Description automatically generated with low confidence

Figure 6. Full dataset Theil’s U comparison of all variables

SNP address is highly associated with year (symmetrical), SNP address is highly associated with both other typing schemes, at all levels. There is strong, symmetrical association between SNP address and cgMLST for levels with similar granularity.

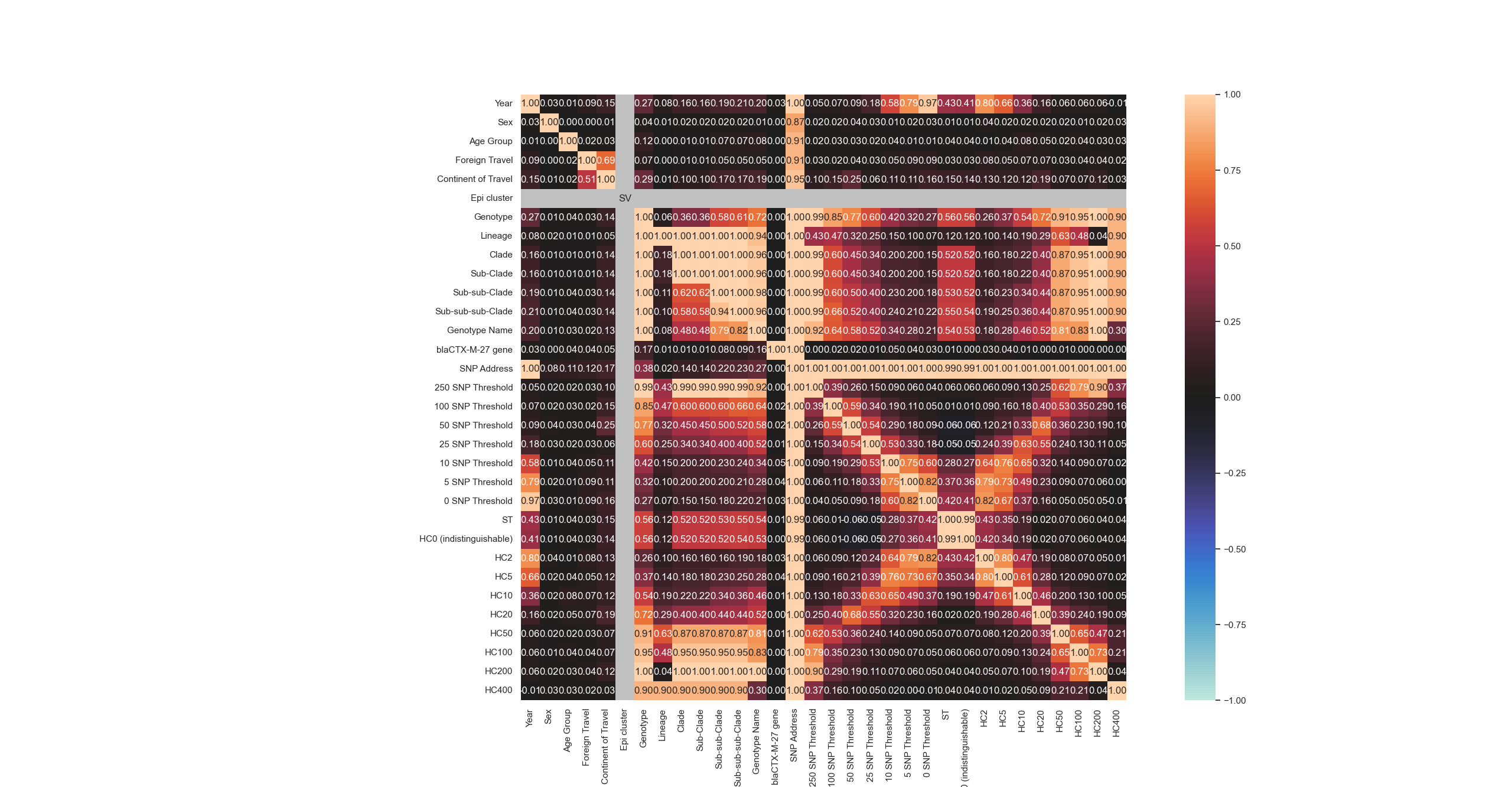


Figure 7. Non-MSM dataset Theil’s U comparison of all variables

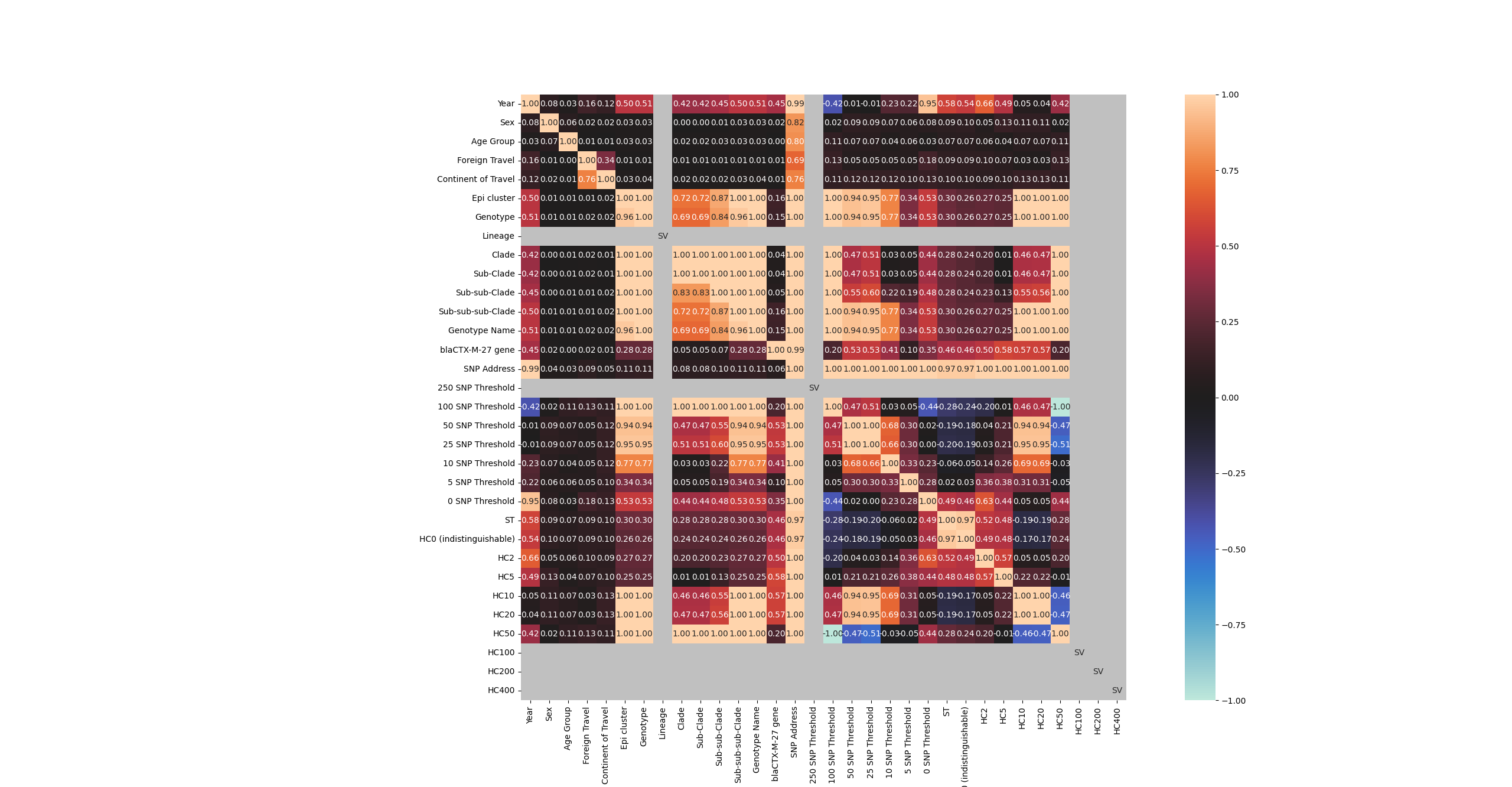


Figure 8. MSM dataset Theil’s U comparison of all variables

# Discussion

Full dataset Theil’s U cgMLST and SNP address. They are not 100% informative of each other, however, showing that there is some difference in the grouping of isolates between the two typing schemes. It is therefore possible that the utility of the two typing schemes is not equal. It is not clear if one is more useful than the other in all scenarios, or even for this dataset.

# References

1. Zhou, Z., J. Charlesworth, and M. Achtman, *HierCC: a multi-level clustering scheme for population assignments based on core genome MLST.* Bioinformatics, 2021. **37**(20): p. 3645-3646.

2. Dallman, T., et al., *SnapperDB: a database solution for routine sequencing analysis of bacterial isolates.* Bioinformatics, 2018. **34**(17): p. 3028-3029.

3. Hawkey, J., et al., *Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei.* Nature Communications, 2021. **12**(1): p. 2684.