An investigation into the link between antibiotic resistance and antibiotic consumption among Recombinogenic bacteria.

Layman’s Summary:

Bacterial pathogens are leading causes of mortality across the world, particularly in developing countries. While treatment with antibiotics in many cases is still effective, the growing spread of resistance amongst pathogenic species, especially the development of multi-drug resistant (MDR) lineages, represents a major global health concern. While a link between increased antibiotic consumption over time and the growth of resistant lineages has been observed for pathogenic bacteria that have low rates of Horizontal gene transfer (HGT), the link between consumption and resistance among species with high levels of HGT is much more ambiguous. This work will develop a new methodology that will allow for the analysis of the past population dynamics of highly recombinogenic bacterial species such as *Streptococcus pneumoniae* (the pneumococcus). Using this methodology, we will then be able to assess under what conditions of antibiotic consumption resistant bacteria can emerge and spread within a population. Specifically, we will look at data from Germany where we already have an established set of samples. The conclusions of this work will enable us to directly inform policy surrounding antibiotic consumption, with the aim of reducing the likelihood of resistance emerging and ensuring that bacterial infections are still treatable in the foreseeable future.

Background:

Infections from highly recombinogenic pathogenic bacterial species, such as the pneumococcus and *Neisseria meningitidis*, still account for a large number of deaths globally each year, with an estimated 1.5 million people dying from pneumococcal pneumonia in 2015 and a further 73,000 from meningitis caused by *N. meningitidis* in the same year (1). Furthermore, the global dissemination of MDR genotypes of these important human pathogens represents a major public health concern (2).

In the past the recombinogenic nature of the pneumococcus has allowed it to rapidly adapt to the introduction of a vaccine (3). For instance, MDR genotypes whose serotypes were included in the original heptavalent PCV7 vaccine were found to remerge within the population via serotype switching (2). This occurs via recombination around the *cps* locus, which encodes the polysaccharide coat of the bacteria that determines the serotype (3).

While the recombinogenic nature of the pneumococcus in this case has allowed for MDR genotypes to be maintained in a population, it is surprising to note that in general, cases of MDR invasive pneumococcal disease (IPD) are caused by only a few closely related clones (4). Given the recombinogenic nature of the pneumococcus and its high carriage rate, and thus high levels of exposure to antibiotics, it would perhaps be reasonable to expect that MDR genotypes would evolve independently multiple times. Interestingly this does not appear to be the case.

Furthermore, the level of antibiotic resistance amongst this highly recombinogenic pathogen has remained stable in Europe over the last 15 years, at around 15% (5). This is another puzzling aspect of the evolution of MDR within the recombinogenic pneumococcus, which suggests a nuanced relationship between consumption and resistance of antibiotics. Further understanding of this relationship is of great importance in mitigating the development of further resistance.

Past studies have sought to understand this link by looking at the correlation between a country’s use of antibiotics and the level of resistance within that country, such as Goosens *et al* 2005 (6). This type of analysis however, only allows you to infer a correlation between antibiotic use and resistance, it doesn’t allow you to infer the nature of the relationship between consumption and the development of resistance over time. Recent studies have addressed this via using phylodynamic methods that incorporate genomic data from dated isolates, allowing for the reconstruction of the past population dynamics.

Phylodynamic methods are powerful tools in epidemiology, allowing for inference of the spatial spread of pathogens and estimation of the time of origin of pathogenic lineages from genetic data (7). Volz and Didelot 2018, used these tools to depict that the growth of a beta lactam resistant clade of *Staphylococcus aureus* in the US in the late 1990s and early 2000s was significantly correlated with increased beta lactam consumption, and the subsequent decline in growth rate of the clade was also associated with a decline in beta lactam consumption.

The phylodynamic technique of estimating past effective population size (Ne) through time using dated genetic data is currently only feasible in species which recombine very infrequently, such as *S.aurues (8*). This is as recombination can confound phylogenetic reconstructions, upon which estimates of Ne are based (9). In previous work we have been able to test the association of macrolide consumption and the growth of a macrolide resistant clade of pneumococcus in Germany. This was only possible due to the insertion of a mobile genetic element (MGE) within the competence machinery of these isolates, which prevented further HGT among these isolates, allowing for phylogenetic reconstruction and inference of the Ne of this clade through time. From this the data we saw no significant association between macrolide consumption and the growth of this resistant lineage. However, the data we used were only for macrolide consumption, we looked at no other class of antibiotics, and were only broad population wide averages of consumption from 1992 to 2010. The further work proposed here would look to acquire greater resolution data for antibiotic consumption, with a breakdown by age class, to assess the nature of the relationship between the consumption of antibiotics and the development of resistance among bacteria.

As well as gathering more detailed consumption data to elucidate the link between consumption and resistance, we will also develop a new methodology allowing us to estimate the past population dynamics of recombinogenic isolates. This will build on previous work by Croucher et al 2015 and Didelot and Wilson 2015, that enables the clonal frame of a collection of whole genome sequence (WGS) data to be estimated using spatial scanning statistics, implemented within a maximum likelihood (ML) framework for computational efficiency. This will then be combined with techniques allowing us to formally test the association between antibiotic consumption and the growth rate of these resistant clades, this building on the work of Volz and Didelot 2018.

The aim of this work is to understand the link between antibiotic consumption and resistance developing among recombinogenic pathogens such as the pneumococcus. In the process we will develop a tool allowing for the evaluation of the past population dynamics of recombinogenic bacteria that can be used by other research groups to look at a range of other problems posed by recombinogenic bacteria.

Scientific methodology:

To meet our objectives with the proposal, a three phase work schedule will be adopted: (i) Development of methodology and software for reconstructing the past population dyanmics of recombinogenic bacteria. (ii) Prepare and modify methodology for distribution via github. (iii) Application of methodology to pneumococcal WGS data and antibiotic consumption data collected from Germany.

**Phase 1:**

The Post-doc student will set out to build upon the work of Croucher et al 2015 and Didelot and Wilson 2015 for the estimation of recombination events within collections of WGS data from closely related samples. Then they will further the work of Sagulenko et al 2017 in forming an accurate ML estimate of a time calibrated phylogeny for the isolates. This will then feed into work by Volz and Didelot 2018, enabling for the estimation of the association between Ne of a clade of isolates and the consumption of antibiotics.

The starting input for the analysis will be an alignment of the dated WGS samples of closely related recombinogenic bacteria. This could be produced via the mauve software (10). Our methodology to detect recombination events within these sequences will be developed from Gubbins, Croucher et al 2015, produced by our collaborators within Imperial College. This detection therefore would work via initially producing a phylogeny from the polymorphic sites in the alignment, which would then be used to produce a joint ancestral sequence reconstruction of the base substitutions in the alignment (11). While marginal ancestral reconstruction is less computationally taxing, joint estimates used in this methodology are more likely to match the global optima. From this joint ancestral state reconstruction, we can place the number and positions of substitutions on the branches of the phylogeny. These can then be analyzed using a sliding window scan statistic tool, to detect higher than expected base substitution density which indicates possible recombination. This technique could produce false positives, as selection differentially acting on areas of the genome would also produce a higher than expected substitution density. However, as Croucher et al 2015 note, there have been relatively small differences in point mutation rates noted across bacterial genomes.

More sophisticated Bayesian techniques for estimating possible recombination within bacterial alignments have been implemented in ClonalFrame from Didelot and Falush 2007. These are more sensitive to recombination events, however are only feasible computationally on multi-locus sequence typing (MLST) data and small sets of WGS data. Therefore, for broader scale analysis our ML based inference of recombination is more applicable.

Our recombination analysis would then output an alignment of the input WGS data without the identified regions of recombination. An ML tree of this alignment will then be formed and subsequently parsed to the opensource ClockTree module of functions from the TreeTime python based framework for phylodynamic analysis (12). The TreeTime methodology for inferring a molecular clock dated phylogeny aims to strike a balance between faster simple heuristic methods implemented in LSD (13) and the more sophisticated, but computationally intensive, Bayesian framework employed by BEAST (14). As such it would allow for the analysis of larger WGS data that we intend to input for many sequences over an acceptable time frame, whilst still giving us a good estimate of the true time scaled tree. The ClockTree framework assumes the input tree topology is the correct one, but optimizes the branch lengths based on the dating information. This is currently designed mainly for shorter viral phylogenies, and thus may need to be further optimized for the longer bacterial sequences we intend to use this framework for.

The output of this step of the methodology will produce a time dated phylogeny, with which we can employ further phylodynamic techniques to reconstruct the past population dynamics of our bacterial samples. This inference will be done using the Skygrowth framework developed by Volz and Didelot 2018. Skygrowth differs to other skyline estimators of Ne via modelling the growth rate of Ne as a simple stochastic autoregressive process, rather than modelling Ne directly as a Brownian motion process (15). In our previous work we found this to better estimate the Ne closer to the present than the equivalent skyline estimator used in BEAST (14). The Skygrowth framework also allows for incorporating time varying covariate data when reconstructing the Ne to test for association between the growth rate and this covariate. This allowed us, in previous work, to estimate the link between macrolide consumption and the growth of macrolide resistant clade in Germany. This framework would therefore be incorporated to allow for further testing of the link between the growth of a resistant lineage and the change in antibiotic consumption, or indeed any other time varying covariate that may affect population sizes.

This would represent the final step in our methodology, which would be combined by the post-doc into one pipeline for further analysis.

**Phase 2:**

Once developed into a single pipeline of tools, we would have to test the methodology and prepare it for distribution. This would involve simulation runs, building in informative error messages and preparing walk-throughs and a message board to enable use by third parties. This would be distributed from GitHub, which is an established environment for open source software distribution.

**Phase 3:**

Once the methodology is developed we will test it with analysis of the link between antibiotic consumption and the spread of antibiotic resistance among German isolates of pneumococcus. We already have a large WGS dataset for IPD isolates from Germany which we will use for an input to our tool. My Co-I will extract a set of data with greater resolution of antibiotic consumption, as described in the Co-I section below. Using this we will test if consumption among certain age groups is associated with an increase in Ne of resistant bacteria and work out the context of the spread of resistance, to detect if any class of antibiotics is associated with increased, or decreased, spread of resistance. The output of this will allow us to make policy recommendations for antibiotic consumption, and given the generalizability of our methodology this can be further tailored to individual country data to provide more relevant recommendations.

Timeliness and novelty:

To our knowledge this is the first attempt to explicitly infer the clonal frame of recombinogenic bacteria and then use this to evaluate their past population dynamics. The output of a molecular clock scaled phylogenetic tree will enable powerful phylodynamic tools to be employed for the first time among these recombinogenic bacteria, enabling inference of a range of epidemiological process, from spatial spread to population dynamics. Our methodology in particular allows for the assessment of the past population dynamics of these important pathogens, from this we have a particular focus on interrogating the link between antibiotic consumption and antibiotic resistance. We aim to provide evidence based policy recommendations to mitigate the further spread of resistance. Given the alarmingly rapid global spread of antibiotic resistance, and the immense public health costs resistance imposes, we believe this work to be both necessary and timely.

Programme of work:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Activity | Time (months) | | | | |
| 0-2 months | 2-14 months | 14-20 months | 20-22 months | 22-24 months |
| Phase 1: Literature review |  |  |  |  |  |
| Phase 1: Pipeline assembly |  |  |  |  |  |
| Phase 2: Distribution |  |  |  |  |  |
| Phase 3: Analysis |  |  |  |  |  |
| Writing up |  |  |  |  |  |

Costing:

|  |  |  |  |
| --- | --- | --- | --- |
| Outlay | Price per unit (£) | No. units | Total Cost (£) |
| Post-doc research associate | 4,642 | 24 | 111,408 |
| Technical software developer | 10,000 | 3 | 30,000 |
| Dataset costs | 10,000 | 1 | 10,000 |
| Overheads | 6,034.60 | 24 | 144,830.3 |
|  | | | **266,238** |

Justification of resources:

|  |  |
| --- | --- |
| Outlay | Justification |
| Post-doc | Given the complex nature of the work building and combing different bioinformatic tools, as well as developing novel statistical techniques and developing a software platform for public distribution, I believe this work requires the greater skillset of a post-doc student, over a post-grad or doctoral student. The candidate would preferably hold a PhD in Statistics or Bioinformatics with relevant coding experience. |
| Technical software developer | The distribution of a software package is a complex undertaking, and the post doc student themselves may not have much relevant experience. Further technical assistance for a period of three months during the distribution phase of the work will allow for a faster development time and a more efficient and professional tool that is likely to be used by a wider range of researchers |
| Dataset costs | While the majority of the German public is covered by a public health insurance scheme, around 15%hold private insurance. Access to these record anonymized could potentially incur costs. This is therefore a contingency fund in case records do need to be purchased from these systems. |

Impact statement: