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

Layman's Summary:

Bacterial infectious agents, 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 rapid analysis of the past population dynamics and spatial spread of bacterial species with high rates of HGT, 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* (the meningococcus) , 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 the meningococcus 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]. However, this type of analysis only allows you to infer a correlation between antibiotic use and resistance, but does not 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 (N_e) through time using dated genetic data is currently only feasible in species which recombine very infrequently, such as *S.aureus* [8]. This is as recombination can confound phylogenetic reconstructions, upon which estimates of N_e 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 N_e of this clade through time. From this 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, not the many other classes 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, as well as their spatial spread and gene content flux through time. 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, as well as work allowing for spatial spread of pathogens to be inferred from phylogenies such as Lemey *et al* 2009. Finally we will also incorporate techniques allowing for the detection of how these resistance genes move through bacterial populations, building on work done by Didelot *et al* 2009.

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

<u>Scientific methodology:</u>

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 dynamics 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. 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 software such as BWA [10]. Our methodology to detect recombination events within these sequences will be developed from Gubbins [11], 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 [12]. 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 [13]. The TreeTime methodology for inferring a molecular clock dated phylogeny aims to strike a balance between faster simple heuristic methods implemented in LSD [14] and the more sophisticated, but computationally intensive, Bayesian framework employed by BEAST [15]. 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 N_e via modelling the growth rate of N_e as a simple stochastic autoregressive process, rather than modelling N_e directly as a Brownian motion process [16]. In our previous work we found this to better estimate the N_e closer to the present than the equivalent skyline estimator used in BEAST [15]. The Skygrowth framework also allows for incorporating time varying covariate data when reconstructing the N_e 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 the 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.

The next step in our methodology would be to incorporate spatial reconstruction of spread from our clonal frame phylogeny. We would follow on from work by Lemey *et al* 2009, implementing a discrete state phylogeographic model to allow for the spatial spread of bacterial pathogens to be inferred from their phylogeny, sample location and phylogenetic rate matrix [17] . This is preferable to quicker parsimony based methods for phylogeographic reconstruction as these are not suitable for organisms with rapid rates of evolution, which bacteria often have. Unlike Lemey *et al* 2009, which uses more sophisticated Bayesian analyses of Phylogeography, we would use a ML framework for increased computational speed.

The final part of our methodology would be to infer the genomic flux of these bacteria through time, this would follow on from work by Didelot *et al* 2009. This method will use the initial input alignment to detect a section of the bacterial pangenome which moves around the population, and reconstructs how these sections move around on our clonal phylogeny [18]. This is currently implemented in a Bayesian framework; the researcher would look to adapt this methodology to an ML framework to enable faster computation time.

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, while my Co-I will also create a new database of WGSs of the meningococcus from Germany for further testing of our tool. Furthermore 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 N_e 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 and geographic spread. 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. Given the increasing use of WGS data by public health bodies, our methodology will allow for these bodies, who have limited expertise in utilizing all the separate complex Bayesian analyses our approach collates, to produce rapid and informative results ideal for weekly reports about the spread of resistance among bacteria for instance .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.

Contribution of the Co-I:

As co-investigator in this important study, my role will be as coordinator, data miner and data manager to firstly, extract all available pneumococcal isolate strain data from surveillance records across Germany, held at The German National Reference Centre for Streptococci, and secondly, to coordinate sequencing of meningococcus isolates. Regarding the pneumococcal surveillance data, the principle investigator's previous work has utilized sequences from 1992-2006, however, pneumococcal surveillance in Germany has been conducted since 1992 to the present day. Therefore, my work will centre around extending the scope of the current dataset to include sequences up to at least 2010 inclusive, as well as to source the corresponding years' antibiotic consumption data from the Institute of Medical Statistics, Frankfurt. My current role at the Institute of Medical Statistics in Frankfurt and fluency in German mean I am well positioned to find and read these data. I have worked in statistics and bacterial research for many years in Germany which will help liaison with government departments, research groups and hospitals to gain access to any additional data sources that may not be freely accessible. Such sources may include the Federal Ministry of Health (Bundesministerium für Gesundheit) and the European Centre for Disease Prevention and Control (ECDC). The ECDC collects population-level data on antibiotic consumption but may be able to help locate sources of data at lower-administrative levels that feed into these country-level statistics.

With regard to meningococcus, I will coordinate with the Federal Ministry of Health to sequence available isolates since this has not yet been done. I have a track record of working with the Federal Ministry of Health to analyse bacterial surveillance data and thus am well placed to meet with relevant officials, draw up necessary permissions and contracts, and recommend sequencing methodologies and laboratories to utilise.

Once all data are collected, I will conduct data cleaning and merging from different sources before splitting each respective bacterial dataset into epidemiologically important age groups and by the type of drug and class of antibiotic prescribed, as well as into geographical units within Germany. The size of geographic units to be used will depend on the spread of sequences and consumption data across the country. At the end of the project I will help the principle investigator to disseminate key findings within Germany to key research and policy stakeholders as deemed appropriate.

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				■	
Sequencing			■		
Writing up and engagement					■

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
Culturing and sequencing isolates	85	400	34,000
Overheads	6,034.60	24	144,830.30
			290,238.30

Justification of resources:

Outlay	Justification
Post-doc	Given the complex nature of the work building and combining different bioinformatic tools, as well as developing novel statistical techniques and developing a software platform for public distribution, I believe we require 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 researcher 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
Culturing and sequencing isolates	Creating an extra dataset of WGS for the meningococcus will allow us to further test the link between antibiotic consumption and resistance developing among important recombinogenic bacterial pathogens. Sequencing will take place in the German public health body where isolates are stored.

Impact statement:

The potential impact from the proposed scheme of work is multifaceted and affects numerous areas of society. On the one hand, as a direct result of the work, we will distribute a novel software package allowing researchers the ability, for the first time, to apply powerful phylodynamic techniques to investigate epidemiological questions about an important group of human pathogens. This has the potential to greatly improve our understanding of the selection pressures that pathogens, such as the pneumococcus and the meningococcus, are exposed to through our own clinical interventions. Given our previous work has highlighted how resistance genes seen in these pathogens are spread across bacterial taxa, further work understanding how these resistance genes evolve has potential to mitigate the development of resistance in a whole range of bacterial pathogens.

Furthermore, given the increasing use of WGS data by public health bodies for surveillance, this fast streamlined tool will allow for greater analysis of these datasets by non-experts unfamiliar with the gamut of Bayesian tools which are incorporated our methodology. Our decision to decision to use an ML framework for inference means these results will be produced rapidly allowing for weekly updates in reports on the spread of resistance and potential hotspots of resistance development.

As well as producing a powerful tool to be disseminated publicly, this work will also tease apart the link between antibiotic consumption and antibiotic resistance developing within the pneumococcus and meningococcus. This work will directly impact policy surrounding the consumption of antibiotics. We aim to liaise directly with the German department of Health, the Bundesgesundheitsministerium, about our findings, whilst also communicating effectively with the press to increase public awareness about antibiotic consumption and antibiotic resistance.

References

1. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388**, 1459-1544 (2016).
2. Croucher, N. J. *et al*. Rapid Pneumococcal Evolution in Response to Clinical Interventions. **331**, 430-434 (2011).
3. Watkins, E. R. *et al*. Vaccination Drives Changes in Metabolic and Virulence Profiles of Streptococcus pneumoniae. *PLoS pathogens* **11**, e1005034 (2015).
4. Klugman, K. P. The successful clone: the vector of dissemination of resistance in Streptococcus pneumoniae. *J. Antimicrob. Chemother.* **50 Suppl S2**, 1-5 (2002).
5. Sonja Lehtinen *et al*. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 1075-1080 (2017).
6. Goossens, H., Ferech, M., Vander Stichele, R. & Elseviers, M. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* **365**, 579-587 (2005).
7. Grenfell, B. T. *et al*. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327-332 (2004).
8. Edward J. Feil *et al*. How Clonal Is Staphylococcus aureus? *Journal of Bacteriology* **185**, 3307-3316 (2003).
9. Faria, N. R. *et al*. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56-61 (2014).
10. Li, H & Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform *Bioinformatics* **25** 1754-1760 (2009).
11. Croucher, N. J. *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43** (2015).
12. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251-1266 (2007).
13. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4** (2018).
14. To, T., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst Biol* **65**, 82-97 (2016).
15. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969-1973 (2012).
16. Volz, E. M. & Didelot, X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Syst Biol* (2018).
17. Lemey, P. *et al* Bayesian phylogeography finds its roots *PLoS computational biology* **5** (2009)
18. Didelot, X *et al* Inferring genomic flux in bacteria *Genome Research* **19** (2009).
Didelot, X & Wilson, D J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes *PLos computational biology* **11** (2015).