Computationally Assessing the Suitability of Whole Genome Geographic Inference Methods

Alex Fassone, 22937

Department of Physics, University of Bath, Bath BA2 7AY, United Kingdom

E-mail: af683@bath.ac.uk

Abstract. Whole Genome Sequencing (WGS) technology presents a new and potentially extremely valuable tool for understanding the dynamics of the spread of disease. Inference of disease dynamics using WGS has predominantly relied on phylogenetic trees which can only be performed retrospectively. This led to the creation of two rapid statistical methods that can instantly infer the origin of a patient's disease. However, a lack of WGS databases meant determining the extent to which these inference methods may be useful required the development of computational models. The models proposed investigate the properties of diseases necessary for successful inference, and the impact of population homogenization and genome acquired noise on the inference method's performance. It is found that samples originating from communities of a disease with a strong genetic identity are easier to classify. The resolution of the inference methods is therefore dependent on the level of homogeneity between individual communities; this implies that inter-community control policies not only diminish the spread of the disease but also help to maintain inference resolution. The models suggest the development of the genetic strength of a community will be greater for diseases that have lower infection rates than mutation rates. However, communities with low infection rates are also more susceptible to homogenization from externally introduced patients. The implementation of a basic control policy highlighted the difficulty in preventing genetic homogenization. Hence, any attempted application of such a policy should be focused on the geographic locations where inference resolution is most desirable.

1. Introduction

There are many existential risks currently facing the human race: climate change, the presence of nuclear weapons, the development of super-intelligent AI, to name a few. One risk, which we seem to be deeply unprepared for, is the risk of a global pandemic. Pandemics have proved to be some of the most deadly events in human history. From 1918 to 1920, Spanish Influenza infected a third of the world's population resulting in the deaths of between 50 and 100 million people [1]. Globalisation has caused the modern world to become increasingly interconnected, creating an environment where an unknown pathogen (a bacterium or virus that causes a disease) could easily spread across the globe within days. Therefore, gaining a better understanding of the epidemiology (the incidence, distribution and possible control) of diseases is of paramount importance. In 2001 Francis Collins, the director of the National Human Genome Research Institute, described the genome as a book with three uses: "It's a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease." Since 2007 the cost of Whole Genome Sequencing (WGS) has dropped from almost $10 million per human genome to just over $1,000. This is largely due to the development of nanopore technology, pioneered by Oxford Nanopore Technologies. Their invention of a mobile phone sized USB device offers more accurately sequenced data, faster and cheaper than its counterparts. This creates the possibility of a future where, aspiring to Dr Collins' transformative textbook of medicine, hospitals and healthcare centres will be able to routinely sequence the genomes of patients and their pathogens. The sequencing of pathogens offers an extremely valuable opportunity to better understand the dynamics of the spread of disease. This can subsequently be harnessed to implement more effective control policies.

Currently, pathogen WGS data can be used to create phylogenetic trees via the analysis of Single Nucleotide Polymorphisms (SNPs). An SNP is the substitution of a single nucleotide that occurs at a specific position within a genome (i.e. it is the location of a mutation). Investigating the phylogeny of a pathogen during an infection and over an epidemic relies on the fact that pathogens naturally accumulate random mutations in their genomes due to error-prone genome replication. Phylogenetic trees are branching diagrams that employ statistical methods, using the differences between genomes, to infer the evolutionary relationships amongst a biological population. They are primarily used to infer transmission events via one of two methods. Transmission events, the passing of a pathogen from one host to another, are the building blocks of the dynamics of any infectious disease. The first method combines epidemiological data, such as the time of symptom onset, geographical location and social ties, with phylogenetic analysis to statistically infer transmission events [2][3]. Uncertainties arise from four key processes: the time between infections, the time between infection and sampling, the in-host pathogen dynamics and mutation [4]. The second method uses the interpretative experience of a trained expert, the availability of whom is not always guaranteed. In addition to the individual issues of each of these methods, relying on phylogenetic trees has two fundamental flaws. Firstly, the statistical methods used to create phylogenetic trees are extremely computationally expensive. It is estimated that the computational cost of the neighbour joining and the maximum likelihood methods, two examples of such methods, increase by 2.5 and 5 times respectively with the addition of just 5 genomes [5]. Secondly, due to the computational cost, these methods are performed retrospectively. Meaning, as tools to inform the implementation of policy aimed at reducing the spread of an epidemic as it occurs, these methods become redundant.

1.1. Methicillin-Resistant Staphylococcus Aureus (MRSA)

Extensive phylogenetic analysis has been performed using the data of MRSA [6]. Staphylococcus Aureus (SA) is a bacterium that is commonly found in the upper respiratory tract and on the skin. It is estimated that 20% - 30% of the human population are persistent carriers and another 60% are intermittent carriers [7]. Although SA usually coexists with its host, it can occasionally become an opportunistic pathogen. It is a common cause for skin infections and bacteraemia (the presence of bacteria within blood). Initially, most outbreaks of SA were found to be treatable with drugs from the penicillin family such as methicillin, cloxacillin and flucloxacillin. However, the widespread clinical use of these drugs quickly caused the emergence of penicillin resistant strains which are now referred to as MRSA [8]. Infections caused by these strains continue to pose a significant problem within healthcare systems across the world. The difficulty in treating systematic infections and their propensity to infect debilitated patients has led to MRSA becoming one of the top five most common causes of hospital acquired infections. The tendency for MRSA to exist exclusively within healthcare systems and not within the external community simplifies the modelling process. This is one of a number of favourable factors that led to MRSA becoming the chosen pathogen for this investigation.

Strains of MRSA that infect more than one person are said to have become epidemic (EMRSA). When epidemic strains are detected in more than one hospital they are numbered, for example, EMRSA-1. MRSA was first detected in the UK in the 1970s. Despite the apparent disappearance of these strains, their descendants have caused significant problems [9]. EMRSA-15 and 16 developed into prevalent strains within the UK healthcare system in the 1990s. The former is known to have first emerged in the midlands in the mid 1980s, spreading extremely quickly throughout the UK [10]. It is still unclear how this strain was able to spread so far in such a short period of time and what factors attenuated its epidemic behaviour. The emerging epidemic became heavily politicised, resulting in the implementation of a number of control policies within the NHS. The mandatory surveillance of MRSA bacteraemia led to the creation of a relatively large WGS dataset of 1022 sequences, sampled between 2001 and 2010; the second favourable feature of choosing MRSA [11].

The final favourable features are due to two aspects of MRSA's population biology; its rates of mutation and recombination. Within the MRSA database the mean number of mutations per genome was 53.7, these could have occurred in any of the, approximately, 3,000,000 sites that make up an MRSA genome. The probability of two mutations occurring at the same site is very low; in the order of 10^{-5}. Genetic recombination or genetic reshuffling is the exchange of genetic material across a genome, between generations. In terms of SNP sites this would mean the information encoded at one site could suddenly swap with information at another. MRSA has a particularly low recombination rate and so this 'reshuffling' of the sites is extremely rare. These two rates create a situation where once a specific site has gained a mutation, it is extremely unlikely that it will lose that mutation. SNP sites can therefore be considered as clocks, implying that two genomes with a mutation at the same site are likely to have a closely related phylogeny. Consequently, MRSA genomes maintain stable patterns of spatio-temporal variation which is primarily why MRSA is used as a model organism to study the spread of drug-resistant bacteria.

1.2. Use of the MRSA database and the development of Geographic Inference Methods

The relationship between the spatio-temporal signal within the genomes and the structure of the patient referral network within the UK has been investigated by combining data from the NHS Hospital Episode Statistics and the database of MRSA genomes [12]. Genomes were grouped into populations by the hospital from which they were sampled, hospitals were grouped by referral clusters (large regions within the UK containing multiple hospitals) and referral clusters were grouped by nationality. Using a version of Wright's F statistic, a measure of the similarity between genetic populations, it was shown that bacterial populations within each hospital were genetically most similar, followed by those within each region and finally those within each country. Hospitals that shared more patients generally showed genetically more similar populations. This indicates that the spread of MRSA is governed by the hierarchical structure of patient movements within healthcare systems. The identification of patient movement as a potential pathway for the transmission of pathogens could have significant implications on the design of effective control policies.

Each hospital containing more genetically similar populations (a genetic clade) is greatly significant as it implies that the genomes themselves contain information relating to their geographical origin. This led to the idea of rapid statistical methods that could infer the geographic origin of a newly sampled genome by comparing it with a central database. The most basic method for determining if two genomes belong to the same geographic community is measuring whether they are separated by fewer SNPs than a specific threshold. However, there is no agreement on what this threshold should be. Different pathogens have different mutation rates which vary at different stages of infection and will respond to selection pressures in different ways. The determination of the threshold level is therefore largely based on the context of the investigation; meaning different studies are not always comparable [13]. The reasons outlined above led to the development of two statistical methods using the genomic data of MRSA [14]; these methods will be referred to as Geographic Inference Methods (GIMs).

Both GIMs employ classifiers used to determine the most probable hospital (community more generally) from which a newly sampled genome originated. This can be used to infer intra- or inter-hospital transmission. Intra-hospital transmission refers to genomes that are classified as having originated from the hospital in which they were sampled. Inter-hospital transmission refers to genomes that are classified as having originated from an exterior hospital. The ability to determine the origin of a patient's pathogen in real-time would allow for the dynamics of a disease to be instantly mapped. If, in a given hospital, the majority of cases were found to be due to intra-hospital transmission, then resources should be prioritised to re-evaluate infection control procedures. For example, identifying the source of infection and limiting the number of health care workers that have access to that region within the hospital (it has been found that staff hands are the primary route of spread of MRSA within hospitals [15]). In contrast, if the majority of cases were imported from other hospitals then screening on admission and/or prior to referral should be prioritized.

The extent to which GIMs may be useful depends on the range of pathogens and environmental conditions to which they can successfully be deployed. It may be that due to their development with the favourable features of MRSA, they are only able to succeed in very specific circumstances. The lack of WGS databases limits the scope to which this can be tested and hence the development of computational models was required. The models proposed in this paper attempt to gain an understanding of the conditions required for the development of the community specific genetic identities necessary for the success of GIMs. They also aim to gain knowledge on the circumstances under which, even with a genetic identity, the GIMs may fail or one approach may be superior to another.

2. Methods

2.1. Explanation of the Geographic Inference Methods

For both GIMs proposed each hospital in the system is labelled with an index $h = 1, \dots, H$ and each SNP site within a genome is labelled with an index $s = 1, \dots, S$. Both methods utilise an $S \times H$ SNP count matrix, $\boldsymbol{C}$, where the element C_{sh} is the number of sampled genomes from hospital h that contain a mutation at site s. All query genomes are assigned an $S \times 1$ binary vector, $\boldsymbol{x}$, where a 1 indicates a site in which a mutation has occurred (total number of mutations in $\boldsymbol{x}$: $\sigma = \sum_{s=1}^{S} x_s$). These vectors are compared with $\boldsymbol{C}$ to produce a y_h score for each hospital which represents the probability that $\boldsymbol{x}$ originated from hospital h.

The first method uses a simple heuristic approach that gives a weighting to each mutation site per hospital that is proportional to the inverse of the overall frequency of SNPs at that site. This was achieved by using $\boldsymbol{C}$ to create a new $S \times H$ matrix, $\boldsymbol{z}$, using equation 1.

$$z_{sh} = \frac{C_{sh}}{\sum_{h=1}^{H} C_{sh}} \qquad\qquad y_h = \frac{1}{\sigma} \sum_{s=1}^{S} x_s z_{sh} \tag{1, 2}$$

The element z_{sh} represents the fraction of the total number of SNPs that occur at site s in hospital h. This creates a scoring system whereby a SNP seen in many hospitals will produce a low z-score for that site in each of the hospitals in which it occurs. A SNP seen in few hospitals will produce a high z-score for that site in each of the hospitals in which it occurs. For example, if all the detected mutations at site S occur within hospital H, it is evidence that S is geographically linked with H, therefore $z_{SH} = 1$ and $z_{Sh \neq H} = 0$. The total weight of the z-scores for each SNP site is 1, so the probability of $\boldsymbol{x}$ belonging to each of the hospitals is calculated using equation 2.

The second method employs a naïve Bayes classifier. The hospital probabilities, y_h, are posterior probabilities $y_h = \frac{L_h P_h}{\Omega}$ where L_h is the likelihood function (the likelihood that $\boldsymbol{x}$ contains the mutations at the sites it does, given it comes from hospital h), P_h is any prior expectation that $\boldsymbol{x}$ comes from h and Ω is the normalisation constant ($\Omega = \sum_{h=1}^{H} L_h P_h$). Due to the binary nature of each site a Binomial likelihood function was chosen with a Beta prior using hyper-parameters α and β [16]. These were used to create another $S \times H$ matrix, $\boldsymbol{p}$, which estimates the probability that a sample from hospital h has a mutation at site s using equation 3, where n_h is the number of samples taken from hospital h.

$$p_{sh} = \frac{C_{sh} + \alpha}{n_h + \alpha + \beta} \qquad\qquad L_h = \prod_{s=1}^{S} p_{sh}^{x_s} (1 - p_{sh})^{1 - x_s} \tag{3, 4}$$

The probability that a sample from hospital h does not have a mutation at site s is therefore $1 - p_{sh}$. These probabilities are then used to construct the likelihood function; equation 4. When applied to the MRSA database, P_h consisted of two prior distributions; one to account for the uneven sampling of hospitals and the other to account for the known geographical clustering of patient referrals. It is noted that neither of these prior distributions made a significant difference to the classification and hence they have not been included.

These two methods approach the problem in very different ways. The heuristic method compares every hospital in the system giving preferential weighting to sites that occur in individual hospitals. The naïve Bayesian classifier looks at each hospital as a whole. It calculates the probability that a query genome originates from that hospital given it has mutations on the sites that it does, giving an equal weighting to all sites. An extension to the heuristic method was also implemented. Instead of looking across all SNP sites to generate the $\boldsymbol{z}$-matrix, it identifies sites that are common amongst the populations of each hospital (i.e. looking at the sites that define each hospitals genetic clade) reasoning that these sites contain the desired geographical information. This was called the informed heuristic method.

2.2. Principles of the computational models

The computational models developed use the kinetic Monte Carlo (KMC) method to run a set of stochastic simulations. KMC models are used to simulate the time evolution of a system of processes, often occurring in nature, using the rates of these processes as inputs for the model. KMC modelling involves the repeated application of three steps; (i) generating simulated data, (ii) performing some statistical procedure, and (iii) recording the results. Studies performed in this way are limited by their finite nature and are therefore subject to sampling variability. This between-simulation variability is called Monte Carlo Error (MCE) and is defined as the standard deviation of the Monte Carlo estimator (the value of interest, for example, the percentage of correct predictions made by a GIM) [17]. All errors quoted are calculated using this method. The models build on the basics of compartmental modelling, considering each hospital as a compartment containing a population of genomes. Each simulation progresses through discrete time via a Markov chain. The rates that define each of the model's processes govern the movement of the system from the current timestep to the next. Over 20,000 SNP sites were identified within the MRSA database, therefore our models assign 20,000 possible mutation sites to each genome. The rates incorporated into our models are: (i) transfer rates between hospitals – the probability of an individual genome being transferred to another hospital; assuming that the strain encoded by the genome transferred from hospital x to hospital y still exists within hospital x after transfer, (ii) mutation rate – the probability that an individual genome gains a mutation in any of its 20,000 possible sites, (iii) infection rate – the probability that a patient infects another patient within the same hospital; the infecting genome is duplicated and replaces the infected genome, (iv) introduction rate – the probability of an external strain being introduced into a hospital (similar to the transfer rate but applied to a single hospital system), (v) misread rate – the probability that a given mutation site within a genome is incorrectly read; for example a 0 being read as a 1, and finally (vi) screening rate – the proportion of patients entering a hospital that are 'tested' for the presence of an external MRSA strain and subsequently placed into quarantine.

The first sub-section within the results and discussion section concerns itself with how the performance of the GIMs is impacted by various processes, under the assumptions that each hospital harbours a homogenous population of its own genetically unique clade. These assumptions are challenged in the second sub-section. These models assume that the population of MRSA genomes within each hospital remains constant. The justification for this originated in early versions of the models that used a deterministic framework, using ODEs adapted from birth-death processes to model the population dynamics of the hospitals. Arguing that hospitals can only accommodate a fixed, maximum number of patients and therefore MRSA strains, a logistic equation was applied which reduced the transfer rate into a hospital as it approached full capacity. This highlighted two separate periods of transience. The first of these transient periods was the fill-up period, the time it took for the populations within each hospital to reach a steady state. The second was the mixing period. Each hospital was initially given a population of its own identifiable genomes. The mixing period was the time taken for the system to reach a homogeneous population of genomes in every hospital. The mixing period was significantly longer than the fill-up period and hence a constant population was assumed when the system was altered into a stochastic framework.

The second sub-section investigates the conditions necessary for a hospital to develop its own unique genetic clade and the strength of said clade. The hospitals here are no longer initiated with a fixed population size, instead they begin with a single blank genome (0's in all 20,000 sites) which is allowed to mutate and infect. This was designed to imitate the dynamics of the EMRSA-15 outbreak that initially spread quickly and then slowed down, allowing for the colonisation of unique genetic clades within each hospital. The infection process allows for the population of MRSA infected patients to grow within the hospital. The simulations are stopped once the hospital fills up to its maximum population, subsequently the genetic diversity of this population is inspected.

3. Results and Discussion

3.1. Testing the robustness of the GIMs, assuming hospital specific clades

3.1.1. Homogenization via patient transfer The relationship between the geographic signal within the genomes and the structure of the UK healthcare system depends on the movement of patients between hospitals and regions. Therefore, this investigation attempts to gauge the correlation between the performance of the GIMs and the rate of patient referral. When applied to the MRSA database, the heuristic and Bayesian methods predicted approximately equal proportions of intra- and inter-hospital transmission. This was noted as a surprising discovery; due to the strong spatial structuring of the genomes it might have been expected that intra-hospital transmission is much more frequent. Of the inter-hospital transmission events identified, 39% were identified as long-range events between hospitals in different referral clusters; again a surprisingly high number. The first model was created to test the idea that an over-estimation of long-range events could be explained by the loss of resolution of short-range events between hospitals that exchange a large quantity of patients.

A three hospital system was created. To isolate the impact of population mixing, the homogeneous starting populations within each hospital were distinct from one another. This was implemented though a genetic fingerprint which consists of specific sites within each genome gaining mutations at $t = 0$. The fingerprint length is the number of unique mutations given to each hospital's population. For this experiment the fingerprint length was set to 10 (for example, each genome initiated in hospital 1 gained mutations in the first 10 sites). A genetic haplotype is a collection of specific mutations that are likely to be inherited together, hence each hospital can be considered to initially harbour a population entirely of its own haplotype. Hospitals 1 and 2 were in close proximity, exchanging a large quantity of patients, while hospital 3 was placed significantly further away with hospital 2 in the middle, see figure 1. Each simulation was run for 100 timesteps; the mutation rate was set to 0.5 such that each genome would be expected to gain 50 mutations on average, approximately the same number that was identified within the MRSA database. The total time was split into two sections, the first 80 time steps allowed the system to exchange genomes between each of the hospitals and for genomes to gain mutations. During the final 20 time steps the processes of mutation and transfer continued however, during this period a fixed number of samples were taken, constructing the central database. The splitting of time in this way was designed to mimic the formation of the MRSA database. The disease spread quickly in the 1980's however samples were taken two decades later; between 2001 and 2010. Figure 2a shows how the transfer rates impacted the haplotype populations within each hospital. Query genomes were used to assess the performance of each of the GIMs by comparing GIM predictions with the known query origin. 90 query genomes were taken at the end of each simulation. These 90 queries were produced such that an equal number of each haplotype population was taken. Due to the uneven proportions of haplotypes within each hospital, queries produced via random sampling would have caused an under representation in the assessment of certain populations. Unless otherwise specified, all investigations using a three hospital model used the same mutation rate, transfer rates, the splitting of time and the production of queries discussed above.

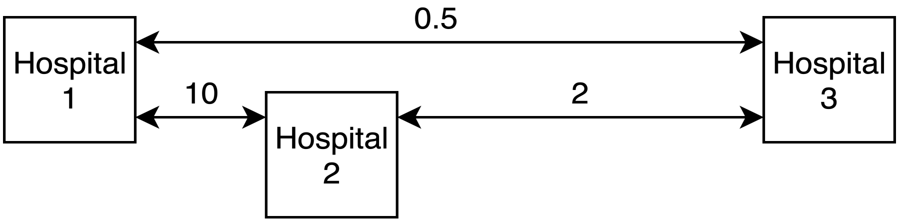
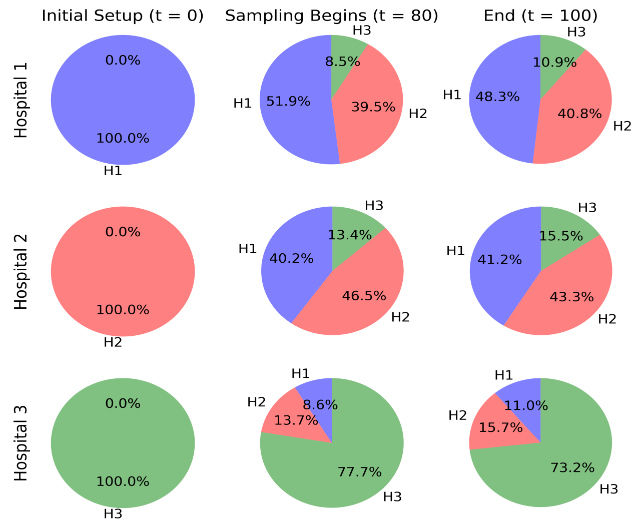


Figure 1: Schematic diagram showing the transfer flows between the three hospitals. The numbers above each arrow indicates the expected number of genome transfers per timestep, each hospital has a constant population of 360 genomes.



(a)

Heuristic - Overall Performance: 62% ± 10%

	H1	H2	H3
Hospital 1	56% ± 25%	48% ± 24%	75% ± 14%
Hospital 2	53% ± 26%	54% ± 24%	78% ± 14%
Hospital 3	52% ± 24%	44% ± 25%	94% ± 9%

Bayesian - Overall Performance: 63% ± 10%

	H1	H2	H3
Hospital 1	60% ± 23%	51% ± 24%	75% ± 15%
Hospital 2	54% ± 24%	55% ± 25%	78% ± 13%
Hospital 3	53% ± 21%	47% ± 23%	94% ± 8%

Informed Heuristic - Overall Performance: 75% ± 23%

	H1	H2	H3
Hospital 1	78% ± 41%	47% ± 49%	100% ± 0%
Hospital 2	78% ± 41%	47% ± 49%	100% ± 0%
Hospital 3	78% ± 41%	47% ± 50%	100% ± 0%

(b)

Figure 2: (a) The average proportions of haplotypes (H1 correspsonds to haplotype 1 etc.) within each hospital at three significant points in time; (i) initialisation, (ii) the beginning of sampling and (iii) the end. (b) The performance (percentage of successfully identified haplotype origins) of the GIMs. Each row represents the hospital from which a query is sampled; each column, the haplotype number (a proxy for the hosptial in which it originated).

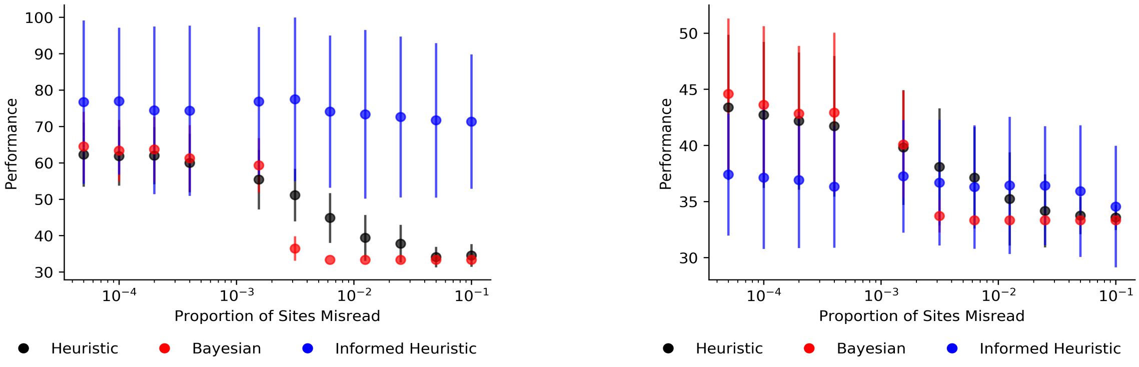
Figure 2b shows the results from 111 simulations, approximately 10,000 total queries. Each row within the tables represents the hospital from which the haplotype was sampled, each column represents the haplotype number (a proxy for the hospital from which the genome originated). Each cell displays the percentage of queries whose origin was correctly identified. The two methods employing a heuristic framework follow a clear trend: haplotypes originating from more mixed hospitals are harder to identify. The Bayesian method performed approximately equally for identifying both haplotype 1's and 2's. Interestingly it performed worse than the informed heuristic method at identifying haplotype 1's but better at identifying haplotype 2's. These differences can be explained by the different way each approach tackles the problem. The heuristic framework makes a direct comparison between every hospital in the system (equation 1). Consequently, the determination of the origin of each haplotype relies on the difference in proportions of said haplotype within each hospital. Figure 2a reveals the average ending proportion of haplotype 2 within hospitals 1 and 2 are very similar (40.8% and 43.3% respectively). Therefore, within a single simulation there is high probability of hospital 1 containing a higher proportion of haplotype 2 than hospital 2. In this situation heuristic based methods would determine hospital 1 as the most likely origin of haplotype 2's. The probability of hospital 2 containing a higher proportion of haplotype 1 than hospital 1 is significantly lower. Therefore, this explains the difference between the heuristic based methods' ability to identify haplotype 1's and 2's. In contrast, the Bayesian method looks at each hospital independently (equation 3). Therefore, it finds hospitals with similar haplotype proportions hard to distinguish between. Hospitals 1 and 2 both contain significant, and almost equal, proportions of haplotypes 1 and 2 and a small proportion of haplotype 3. This explains the Bayesian method's approximately equal performance identifying haplotypes 1 and 2. The improved performance of the informed heuristic method is due to its lack of contribution from the noise of the random mutations (unless they occur within the sites identified, although with a fingerprint length of 10 this is unlikely to impact its performance). Understanding the different regimes in which each GIM fails is crucial to determine the most impactful method for different situations.

Every GIM confidently identified the origin of haplotype 3's regardless of where they were found. There was a slight difference in the ability of the heuristic and Bayesian methods to correctly predict the origin of haplotype 3's depending on whether they were sampled from hospitals 1 and 2 or hospital 3. This was due to the sampling methods used to generate the central database and the query genomes.

There was a greater probability that haplotype 3 queries taken from hospital 3 also existed within the database making them significantly easier to identify. The impact of this could have been reduced by increasing the population sizes of each hospital. However, this would have significantly increased the computational cost of each simulation. Figure 2a shows that hospital 3, on average, maintains a haplotype 3 population of approximately 75%, i.e. it holds a very strong identity within the system. This allows a simple but significant conclusion to be drawn: haplotypes that originate from hospitals that maintain a strong genetic identity can be more consistently identified. This conclusion also highlights that the ability for GIMs to distinguish between two hospitals significantly reduces as the hospitals exchange more patients. This provides evidence that the proportion of detected inter- and long distance inter-transmission using the MRSA data set was in fact an over-representation caused by an inability for the GIMs to resolve inter-transmission among closely related hospitals. Therefore, policies designed to limit the spread of disease between communities will have the added benefit of maintaining the resolution of GIMs.

3.1.2. Contribution of genome acquired noise The previous section investigated how the homogenization of pathogen populations influenced the performance of the GIMs. The informed heuristic method was able to remove the impact of noise, introduced through the gain in mutations, while the other two methods could not. Here, the inclusion of additional noise, implemented via a misread rate, is investigated. Current WGS technologies still incur a significant proportion of mistakes. It is estimated that current nanopore sequencers can produce a misread proportion of up to 10% [18]. The misread proportion is the fraction of sites within a genome that are expected to contain a mistake. For this analysis a mistake would mean that a mutated site, signalled with a 1, would be read as a 0 and vice versa. The three hospital model was repeated with an increasing misread proportion, varied between 0% and 10%, doubling the expected number of misreads per genome for each data point. Figure 3 shows the performance of the GIMs with initial fingerprint lengths of 10 and 1. This reduction in fingerprint length did not impact the shape of the graphs for any of the methods. It did however cause a significant reduction in the overall performance of all three GIMs. The relationship between fingerprint length and performance is discussed later. The performance of the informed heuristic method remained relatively stable as the proportion of misreads increased. This follows the theory that by selecting only the most informative sites, the impact of noise is removed. This has only been shown for a maximum misread rate of 10%, however developments in WGS technology should contribute to decrease the misread rate from the current 10%.

The heuristic and the Bayesian methods displayed a drop in performance as the misread proportion exceeded 10^{-3}. A misread rate of 10% within a 20,000 site genome would be expected to produce 2000 misreads, the majority of which would represent 0's being read as 1's, far more noise



(a)

(b)

Figure 3: The impact of the proportion of misreads per genome (varied from 0% to 10%) on the overall performance of the GIMs with an initial fingerprint length of (a) 10 and (b) 1.

than the 50 mutations each genome is expected to gain. Both methods look at every mutation site, therefore all additional noise is incorporated into their predictions. The performance of the heuristic method can be seen to gradually drop down to 33% (the same percentage as would be expected by a method of randomly guessing). The Bayesian method exhibits a much steeper drop in performance. Due to its equal treatment of every site, every misread will equally contribute to its failure. The heuristic method was designed to weight sites according to the inverse of their overall frequency (equation 1); 'reasoning that any classifier giving equal weight to all sites might be prone to having the most useful information swamped by a mass of less differentiating information' [14]. Therefore, it can be concluded that heuristic style methods are more resistant to genome acquired noise.

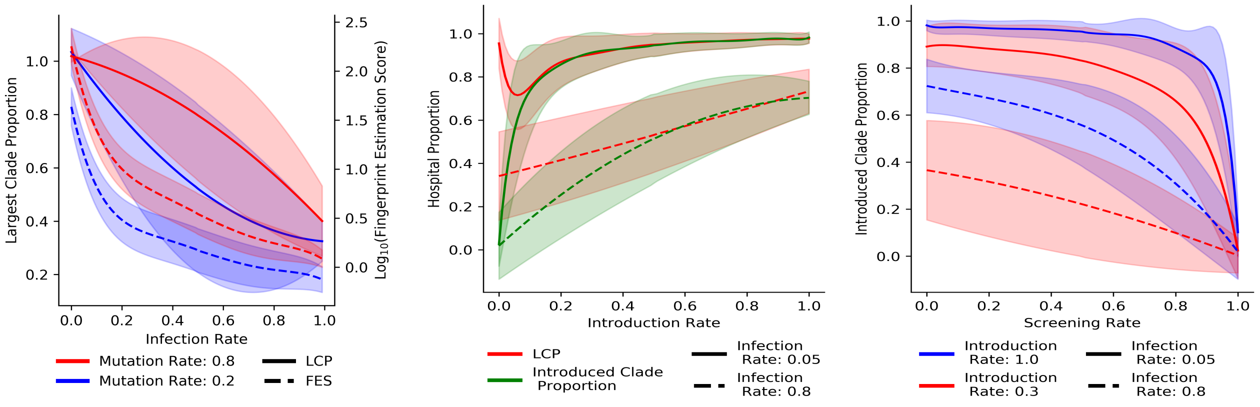
Another mechanism for increasing the proportion of noise within each genome is reducing the number of sites on which mutations can occur. Bacterial pathogens have a wide range of genome lengths; anywhere between $10^5 - 10^7$ sites. Therefore, measuring the number of sites and the mutation rate of a specific pathogen and the misread rate of the technology can provide an estimation for the expected percentage of noise within each genome. This will help to inform decisions regarding the implementation of GIMs, based on their expected success, and if implemented, the chosen style of method.

3.1.3. Fingerprint length The performance of the GIMs was tested against several other factors, most notably the length of the fingerprint; this can be considered as a measure of the depth of the genetic clade within each hospital. Unsurprisingly, it was found that increasing the fingerprint length, varied from 1-10, increased the performance of all three methods. The informed heuristic outperformed the other two methods for all fingerprint lengths except 1 and 2 (see the first two data points in Figure 3b compared with Figure 3a). Smaller fingerprints create a greater opportunity for the misidentification of the important sites. This also explains the higher variance nature of the informed heuristic method (see the error bars in figure 3 and quoted errors in figure 2b). The identification of informative sites leads to strong performances however the misidentification of sites leads to significantly weaker performances. Therefore, although promising, the adoption of this method should be treated with caution. A development in the criteria used to identify sites of importance could significantly improve its consistency.

3.2. Development of genetic clades

The previous analysis assumed that each hospital had developed its own unique genetic clade. This section investigates the conditions under which a community of pathogens is likely to develop its own clade, the size and depth of said clade and the clade's resistance to external introduction with and without a control policy. The models here start with a population of a single unmutated genome which is allowed to mutate and infect until the maximum MRSA population size is reached. The infection mechanism creates a causal link between genomes, allowing the hospital to form its own genetic identity. Once the hospitals reach their maximum population (the fill up point) the genetic diversity is measured using two separate metrics. The first is the proportion of the hospital occupied by its largest clade; Largest Clade Proportion (LCP). Genomes are defined as being part of the same clade if their root site, the site of their first mutation, is the same. Therefore, the maximum value for this metric is 1, indicating that every genome within the hospital has the same root. The LCP is significant as it gives an indication of the probability that samples taken from a hospital will be part of the same clade. The previous section showed that for the GIMs to consistently predict a query's origin, samples taken from each community need to harbour a unique genetic identity. The second is a measure of the number of sites which are shared amongst the population; the depth of the largest clade. This metric is an estimation of the hospital's fingerprint length and is therefore called the Fingerprint Estimation Score (FES). A homogeneous population in which all genomes share exactly ten sites, as seen in the previous section, would score a FES of 10 and a LCP of 1.

Figure 4a shows the relationship between these metrics and the infection rate for two mutation rates; one low (0.2) and one high (0.8). For both mutation rates when the infection rate is low, the LCP score remains close to 1 and the FES score is in the order of $10^{1.5}$. This indicates that not only



(a) (b) (c)

Figure 4: (a) The impact of the infection rate on the largest clade proportion and fingerprint estimation score for mutation rates of 0.2 and 0.8. (b) The Impact of the introduction rate on the largest clade proportion and the proportion of the hospital occupied by the introduced clade for infection rates of 0.05 and 0.8. (c) The impact of the screening rate on the proportion of the hospital occupied by the introduced clade for introduction rates of 1.0 and 0.3 and infection rates of 0.05 and 0.8. All data has been smoothed using a Savitzky-Golay filter.

does the biggest clade occupy the entire community but the depth of the clade is approximately 30 sites long (three times bigger than the hospital communities initialised in sub-section 1). Communities that have evolved in this way therefore contain a strong genetic identity. For both mutation rates, as the infection rate increases both the LCP and the FES decrease; an infection rate of 0.8 causes the LCP to drop to around 0.5 and the FES to drop to approximately 1. It can be concluded that highly infectious pathogens are more likely to produce communities with a weak genetic signal, particularly if their mutation rate is low. It is estimated that MRSA genomes have a mutation rate of 3.75 per year [19] and an infection rate of between 0.5 and 51 per year if unisolated [20]. The model therefore suggests that the infection rates within the hospitals sampled, used to create the MRSA database, were at the lower end of these estimations. If the rates of mutation and infection of any pathogen are known, versions of figure 4a could be used as an initial test to determine the likelihood of GIM success.

This analysis is based on an isolated community. However, the development of genetic clades in the absence of external influence is extremely unlikely. Figure 4b shows the LCP using a mutation rate of 0.2 and infection rates of 0.05 and 0.8. The community now has the additional influx of an external clade. The proportion of the hospital taken up by this external clade at the fill up point is also plotted; when the two lines are equal it implies that the largest clade within the hospital is made from the external haplotype. This is designed to simulate the development of a genetic clade within a community in the presence of another community with which it exchanges patients at a fixed rate. For the infection rate of 0.05 it can be seen that the introduced clade becomes the dominant clade at an introduction rate of just 0.1. For the higher infection rate of 0.8 it takes a significantly higher introduction rate to begin to dominate, approximately 0.5. The infection rate and the introduction rate are both processes that increase the number of strains within the system. Therefore, if the community is to develop its own clade it requires a pathogen that is able to infect, spreading its own clade, faster than the accumulation of an external clade. Within this model it is important to understand two assumptions that provide the introduced clade with an advantage over the community developed clade. Firstly, genomes from the external clade have a probability of being added from the beginning of each simulation. Therefore, this assumes that the external community has already developed a clade at $t = 0$. The population under investigation only contains a single blank haplotype at $t = 0$; having to develop its own clade. Secondly, every genome introduced is a member of the same clade. This therefore assumes that the external community has a LCP ≈ 1. Despite these assumptions, our model still provides an understanding of the mechanisms that contribute to a community of pathogens being resistant to external influence. Understanding the processes that contribute to this resistance provides useful information regarding the expected resolution of the GIMs. If a population of a pathogen lacks resistance to external clades, the resulting genetic homogenization will cause samples to lose their inherent geographical signal. The model suggests that pathogens with low infection rates and communities that regularly trade patients are most susceptible. This supports the claim that the frequent exchange of patients between communities produces more genetically similar populations [12]. A model where two hospitals co-develop, both starting with blank haplotypes, could be used to further understand the relationship between co-development and genetic similarity.

Reducing population homogenization increases the resolution of GIMs. Therefore, policies designed to limit the spread of disease have the added benefit of maintaining inference resolution. The final model was designed to test the impact of one of such policies. A screening rate was implemented representing the proportion of externally introduced patients that were 'tested' and subsequently placed into quarantine. The infection rate of patients in quarantine was reduced to 0, thereby assuming the quarantine was completely effective at limiting the spread of disease. Figure 4c highlights the impact of this policy by plotting the proportion of the externally introduced clade at the fill up point against the screening rate of incoming patients. Communities of pathogens with low infection rates and high transfer rates have been identified as the most at risk of genetic homogenization. Figure 4c shows four of such examples (identified using the same infection rates of 0.05 and 0.8). With an infection rate of 0.05, halving the size of the external clade required a screening rate of over 0.9, even for the lower introduction rate of 0.3. With an infection rate of 0.8, halving the size of the introduced clade still required a screening rate of approximately 0.8. Implementing such a policy is guaranteed to have a strict budget. The model suggests that obtaining impactful results requires a high screening rate, therefore this budget should not be spread too thinly among lots of communities. It should instead be targeted to communities that have been identified as having the highest risk of losing their geographic signals or having the pathogen's most dangerous strain. The screening rate is assumed to have been applied to all incoming patients who each have an equal probability of harbouring the pathogen. Another way to improve the impact of this screening policy would be to form criteria used to identify patients who are more likely to harbour populations of the pathogen. For example, it has been found that patients who have recently had surgery, patients with intra-vascular devices and HIV infections are all more likely to develop MRSA infections [21].

4. Conclusions

It has been shown that the success of GIMs using the WGS data of pathogens depends on the strength of the genetic identity harboured within individual communities of the pathogen. This study suggests that strong genetic communities are more likely to develop with pathogens that have an infection rate lower than their mutation rate. However, communities of pathogens with lower infection rates are also more susceptible to the homogenization of their populations via the exchange of patients. Understanding the process of population homogenization is key to understanding the expected resolution of GIMs. This would be furthered by the development of additional models that attempt to quantify the relationship between community co-development and genetic similarity. The negative impact of genetic homogenization on GIM performance implies the implementation of effective inter-community control policies will not only reduce the spread of a pathogen but will also aide in maintaining GIM resolution. Via the application of a screening rate, a quarantine based policy highlighted the difficulty in preventing genetic homogenization. Therefore, the allocation of resources for such an approach should be targeted to communities for which resolution is most desired.

An additional comprehension of the noise introduced via mutations, the length of the pathogens genome and the misread proportion of the sequencing technology dictates the likelihood of success of different approaches. Heuristic based methods are more resistant to genome acquired noise than Bayesian approaches however they cannot distinguish between communities with equal populations of the same haplotype. Bayesian approaches may be able to make this distinction if each community harbours a more unique population within the overall system.

This analysis is based on the 1990's outbreak of MRSA in the UK, which initially spread rapidly but quickly attenuated its epidemic behaviour. Investigations using models where pathogens spread via different mechanisms are required for a more complete understanding of the suitability of GIMs. However, the models developed highlight the areas on which these investigations should be focused.

Acknowledgements

I would like to thank my project partner, Finlay Duff, for helping to formulate all utilised models and particularly for his work investigating the development of community specific genetic clades. I would also like to thank our supervisor Dr Richard James for his expert guidance throughout the project.

References

[1] Taubenberger JK, Morens DM. 1918 Influenza: the mother of all pandemics. Emerging Infectious Diseases. 2006 Jan;12(1):15–22.

[2] Ypma RJF, Bataille AMA, Stegeman A, et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B. 2011 June;279.

[3] Ypma RJF, van Ballegooijen WM, Wallinga J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. Genetics. 2013 September;195(3):1055–1062.

[4] Klinkenberg D, Backer JA, Didelot X, et al. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol. 2017 May;13(5).

[5] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994 May;11(3):459–468.

[6] Lakhundi S, Zhang K. Methicillin-Resistant Staphylococcus aureus: Molecular Characterization, Evolution, and Epidemiology. Clinical Microbiology Reviews. 2018 Sep;31(4).

[7] Kluytmans J, van Belkum A, Verbrugh H. Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks. Clin Microbiol Rev. 1997 Jul;10(3):505–520.

[8] Shanson DC. Antibiotic-resistant Staphylococcus aureus. J Hosp Infect. 1981 Mar;2(1):11–36.

[9] van Belkum A, Duckworth G, (rapid response). 40 years of methicillin resistant Staphylococcus aureus. BMJ. 2001 Sep;323(7314):644–645.

[10] Holden MTG, Hsu LY, Kurt K, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013 Apr;23(4):653–664.

[11] Johnson AP, Davies J, Guy R, et al. Mandatory surveillance of methicillin-resistant Staphylococcus aureus (MRSA) bacteraemia in England: the first 10 years. J Antimicrob Chemother. 2012 Apr;67(4):802–809.

[12] Donker T, Reuter S, Sciberras J, et al. Population genetic structuring of methicillin-resistant Staphylococcus aureus clone EMRSA-15 within UK reflects patient referral patterns. Microb Genom. 2017 Jul;3(7).

[13] Stimson J, Gardy J, Mathema B, et al. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Mol Biol Evol. 2019 Mar;36(3):587–603.

[14] Aspbury M, Sciberras J, Corander J, et al. Rapid statistical methods for inferring intra- and inter-hospital transmission of nosocomial pathogens from whole genome sequence data. Preprint. 2018 Oct;.

[15] Ayliffe GAJ, Buckles A, Casewell MW, et al. Revised guidelines for the control of methicillin-resistant Staphylococcus aureus infection in hospitals. British Society for Antimicrobial Chemotherapy, Hospital Infection Society and the Infection Control Nurses Association. J Hosp Infect. 1998 Aug;39(4):253–290.

[16] Gelman A, Carlin JB, Stern HS, et al. Bayesian Data Analysis. vol. 1. Chapman and Hall; 2020.

[17] Koehler E, Brown E, Haneuse SJPA. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. Am Stat. 2009 May;63(2):155–162.

[18] Bowden R, Davies RW, Heger A, et al. Sequencing of human genomes with nanopore technology. Nat Commun. 2019 Apr;10(1):1869.

[19] Alam MT, Read TD, Petit RA, et al. Transmission and Microevolution of USA300 MRSA in U.S. Households: Evidence from Whole-Genome Sequencing. mBio. 2015 Mar;6(2).

[20] Tübbicke A, Hübner C, A K, et al. Transmission rates, screening methods and costs of MRSA–a systematic literature review related to the prevalence in Germany. Eur J Clin Microbiol Infect Dis. 2012 Oct;31(10):2497–2511.

[21] Kluytmans J, van Belkum A, Verbrugh H, et al. Nasal Carriage of Staphylococcus aureus: Epidemiology, Underlying Mechanisms, and Associated Risks. CMR. 1997 Jul;10(3):505–520.