Integration of microbiological, epidemiological and next generation sequencing technologies data for the managing of nosocomial infections

Matteo Brilli, Francesco Comandatore, Aurora Piazza, Gian Vincenzo Zuccotti, Claudio Bandi

SkyNet UNIMI Laboratory - Piattaforma di Epidemiologia Genomica e Microbiologia Sperimentale, Pediatric Research Center Romeo ed Enrica Invernizzi, "L. Sacco" Department of Biomedical and Clinical Sciences, University of Milan,

Summary

At its core, the work of clinical microbiologists consists in the retrieving of a few bytes of information (species identification; metabolic capacities; staining and antigenic properties; antibiotic resistance profiles, etc.) from pathogenic agents. The development of next generation sequencing technologies (NGS), and the possibility to determine the entire genome for bacterial pathogens, fungi and protozoans will likely introduce a breakthrough in the amount of information generated by clinical microbiology laboratories: from bytes to Megabytes of information, for a single isolate. In parallel, the development of novel informatics tools, designed for the management and analysis of the so-called Big Data, offers the possibility to search for patterns in databases collecting genomic and microbiological information on the pathogens, as well as epidemiological data and information on the clinical parameters of the patients. Nosocomial infections and antibiotic resistance will likely represent major challenges for clinical microbiologists, in the next decades. In this paper, we describe how bacterial genomics based on NGS, integrated with novel informatic tools, could contribute to the control of hospital infections and multi-drug resistant pathogens.

Correspondence: Claudio Bandi, Centro Ricerca Pediatrica Romeo ed Enrica Invernizzi, Università degli Studi di Milano, Via Celoria 26, 20133 Milano, Italy.
Tel.: + 39.02.5031.4710/9824.
E-mail: claudio.bandi@unimi.it

Key words: next generation sequencing technologies, nosocomial infections, multi-drug resistant pathogens.

Contributions: the authors contributed equally.

Conflict of interest: the authors declare no potential conflict of interest.

Received for publication: 19 February 2018.
Accepted for publication: 19 February 2018.

Microbiologia Medica 2017; 32:7359
doi:10.4081/mm.2017.7359

Introduction

The last years have witnessed a large increase in the number of people getting infected by antibiotic resistant bacteria, while staying in hospitals. This is related to both a much more complicated clinical picture of patients, that are often characterized by comorbidities and immunosuppression, and by the growing diffusion of multi drug resistant bacteria (8). Where a bacterial infection is suspected to be the cause of a potentially life-threatening pathology, antibiotic treatment is normally not delayed until microbiologists provide a name for the pathogen and an antibiotic resistance profile. Therefore, the normal approach is to start a wide-spectrum and highly empirical antibiotic treatment (that can then be changed once the microbiological results are available). The efficacy of this empirical approach depends on the knowledge of the local epidemiology, *i.e.* a knowledge on the most widespread pathogens, their antibiotic resistance patterns, the risk factors of the patients to be treated, *etc*. Very often, pathogens have multiple resistances, which limits the available therapeutic options and increase the probability of failure of the empirical treatment.

The growing complexity of the clinical picture of patients, together with the resistance characteristics of the pathogen and the big size of modern hospitals, makes the development of real-time surveillance systems a fundamental issue. Existing active surveillance systems include the screening of patients entering high risk hospital wards (*e.g.* intensive care and surgery) through rectal swab and plating on selective media to evaluate the presence of *important* bacterial pathogens (*e.g.* carbapenem resistant *Enterobacteriaceae*). In addition to pathogen identification during the screenings, and to the determination of their antibiotic resistance profiles, it would be very useful to acquire as more genetic information as possible from the isolated strains, to uncover hidden resistance factors, virulence genes, and other genetic determinants. In addition, a deep genetic characterization would allow to reconstruct the relationships among the isolates, thus to determine the possible clonality of an endemic pathogen presence, or of an outbreak. At the moment, the elected techniques for genetic characterization of bacterial isolates in the clinical context are PFGE (4) and multi-locus sequence typing, which however requires long time and specialized personnel, making a large scale use almost impossible.

Next Generation Sequencing (NGS) technologies have the potential of making feasible the generation of full bacterial genome sequences in the context of clinical microbiology and are therefore ideal candidates for complementing or replacing existing tools for antibiotic resistance and pathogen monitoring in the hospital. Genome analysis can provide most of the information required for the characterization of nosocomial infections and for

the reconstruction of epidemic outbreaks, potentially providing an important contribution to their prevention and resolution. A full genome sequencing of a bacterium, that in the past required long times and economic efforts, can today be completed in about 24h, and several bacterial genomes can be obtained at once through multiplexing, in variable number, depending on the sequencing machine used. When the sequencer is not present *in situ*, sequencing services can be accessed at reasonable prices, but the delay to obtain the results is increased, making clinical use difficult.

NGS advent has surely and radically modified our approach towards many biological problems, and we are convinced that time is ready for heavy deployment in clinical microbiology. Integration with additional information that are normally recorded for patients, such as the contact network of patients, could even boost its usefulness. In this paper, we show how the recent discoveries about epidemic spreading in complex networks together with massive *real time* sequencing efforts in hospital settings might allow to understand promptly if and how an infection outbreak is taking place, together with detailed characterizations of the involved strains, that are not always possible to obtain using standard microbiology practices.

We are indeed convinced that the time is ready for developing surveillance systems where the genome sequences obtained after routine monitoring sampling from the hospital environment and the patients are integrated with information about risk factors of the patients themselves and their movements within the hospital, or among different hospitals. Such a system, integrated with the already existing monitoring systems, would allow to depict the pattern of antibiotic resistance and pathogen presence and their *migration routes* into the hospital, to quickly reconstruct nosocomial epidemic outbreaks, to predict the risk factors for infections and outbreaks, and thus to prevent them.

Contact network of patients

An important object to predict the risk of an epidemic outbreak in hospitals is the so-called contact network of patients which is obtained by recording the room or the rooms where each patient spend (part of) his stay. Preliminary analysis (data not shown) suggest that this network have a scale-free topology. This kind of networks is characterized by a linear relationship among the number of connections a node has (called the degree) and the frequency of nodes with that same degree, in a double logarithmic plot. This is a major difference with respect to network of the Erdos-Renyi type, where the degree follows a normal distribution. This feature has very important consequences on the behavior of the network with respect to perturbations, and its global navigability. In practice, scale-free networks are very robust to random node removal (because the vast majority of the nodes in this kind of networks is lowly connected and have therefore small influence on the topology of the entire network) but at the same time they are very fragile to targeted node removal (for instance if removed nodes belong to the hub or bottleneck classes, Figures 1 and 2).

This is because hubs maintain the network connected and therefore their removal, even in small number, can cause fragmentation of the network into many isolate components. Let's consider the world wide web; removing peripheral nodes like the webpage of a few Italian high schools, has no impact on the global structure, while shutting down Google might cause global problems. In the case of the infectious disease epidemiology, in a scale-free contact network, it has been shown that the R0 parameter becomes so little that it is zero in practice (1,2,5,6). Thus, in this type of network, even a pathogen with a very low transmission capacity can potentially spread and persist into the entire population. It is thus clear that the management of hospital infections might be improved by a rational re-distribution of the patients in a way that minimizes the transfer of pathogens.

Monitoring patient positivization and its integration in the contact network

Monitoring patients for the presence of common pathogens is a relatively common practice in hospitals today. For example, swabs can be taken once a week, from body districts informative for a certain pathogen, and plated on selective media. Positive patients are counted and the information collected in a dedicated, often hospital-specific, database. Once the number of positive patients overcomes some arbitrary and hospital specific threshold, an outbreak status is announced. In this situation, the record of who and when carried the pathogen could easily be integrated in the contact network to take informed decisions about how to manage the infection outbreak in a case-specific way. Let's consider the case of a hub patient found infected by a pathogen; all patients that were recently in contact with this patient are at high risk for the

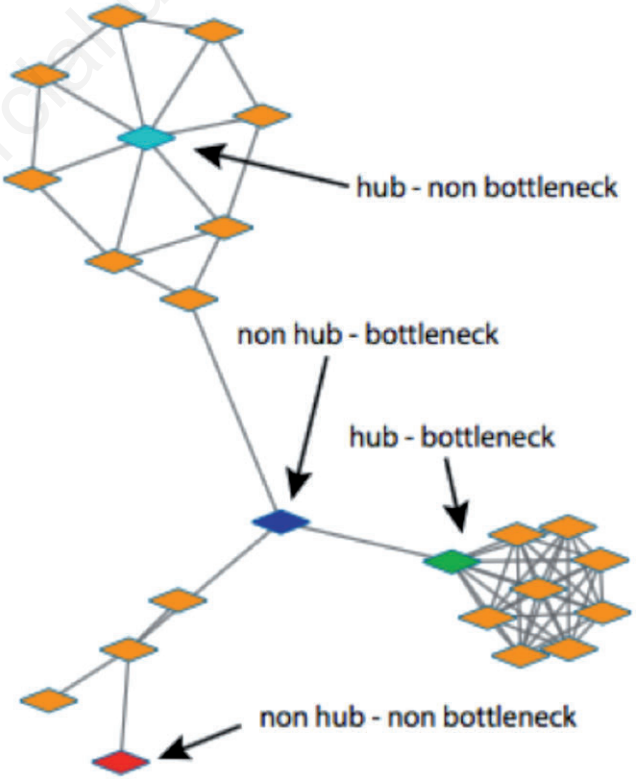

Figure 1. A hypothetical network with node categories defined on the basis of centrality analysis. A hub is a node with many edges; hubs have an important role in keeping the network connected and their removal rapidly causes fragmentation of the network in isolated components. A bottleneck node is a node that is traversed by many shortest-paths. A shortest path is defined as the path crossing the least nodes in going from one node to another. The two characteristics are defined on different properties and can therefore coexist.

same infection. On the converse, the pathogen infecting a peripheral patient should have a slower and reduced propagation. The analysis of the historical record of positives might moreover enable the identification of problems in managing patients, and highlight more or less virtuous wards and associated personnel.

The study of past outbreaks, by integrating the information about which patients were positive to a certain pathogen and the contact network, might moreover allow the identification of recurring patterns and therefore provide an objective risk assessment for future cases. Additional information can also be integrated in this strategy; data concerning antibiotic resistances or clonality could allow a more precise identification of critical situations. It should be noted that the information about the clonality or even phylogenetic relatedness is generally not available; even though several methodologies are available for phylogenetic reconstruction and for the determination of clonality, WGS is for sure the most powerful tool for these purposes.

Clinical genomics

How the information contained in the genome sequence of a pathogen could improve managing infection outbreaks? A thor-

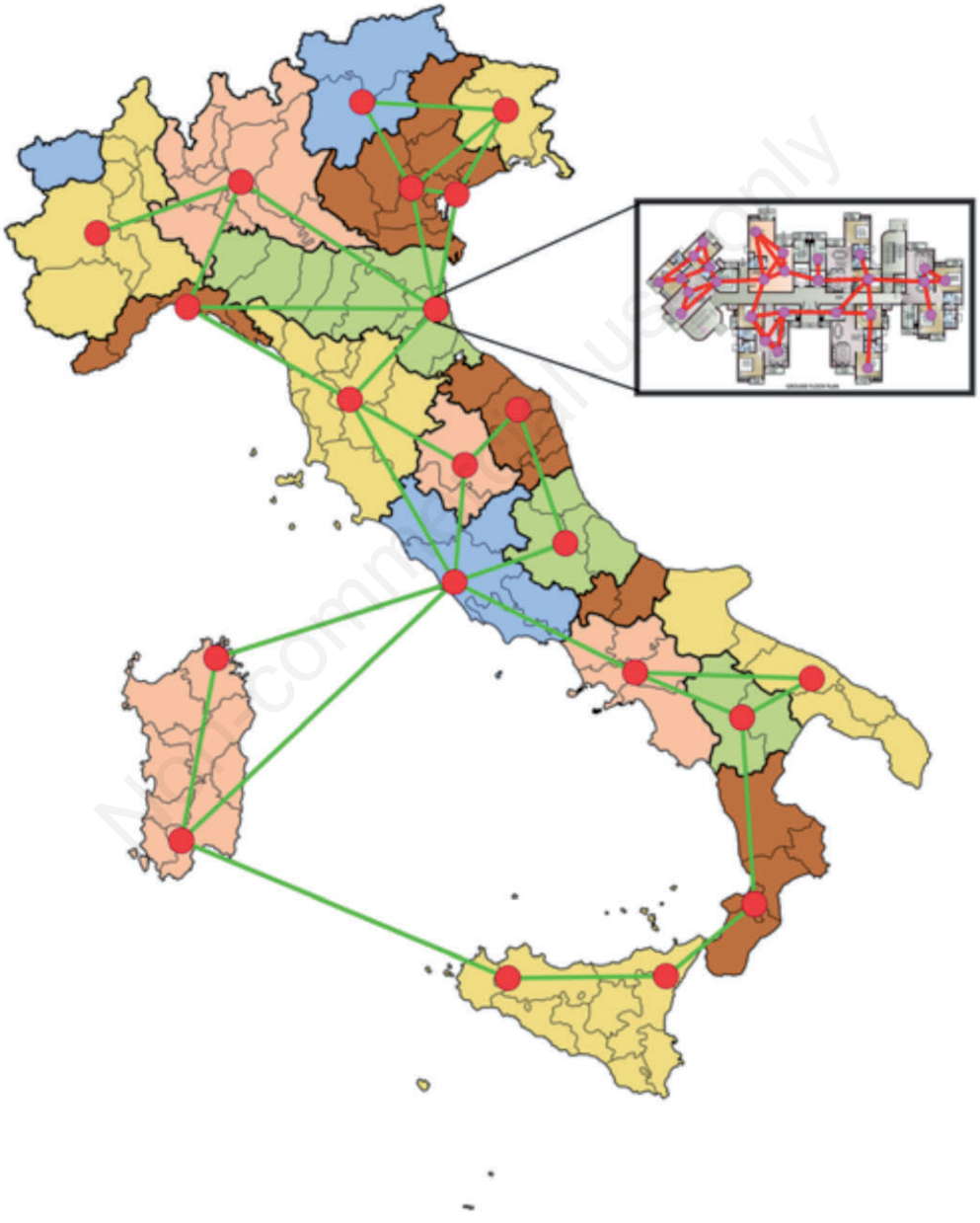

Figure 2. Living in a network of networks. Red nodes represent hypothetical hospitals scattered throughout the country. They are connected on the basis of the flux of patients going from one hospital to another, corresponding to the most likely routes of pathogen movement. Every hospital hosts a network of contact among patients and among patients and personnel. It would be feasible to generate a system in which the patient contact networks are regularly (and automatically) updated, and integrated with genomic information form the isolated pathogens.

ough genomic analysis can integrate/confirm the information provided by standard clinical microbiology techniques, but much more could be extracted from whole pathogen genomes, and translated into applications. The availability of the genome sequence of pathogens sampled in a hospital can provide a precise characterization of the phylogenetic relationships existing among them, as the PFGE, but with greater scalability. Knowledge of phylogenetic relationships is very important to understand how critical is the situation; for instance, strains with very similar resistance profiles (therefore appearing the same or almost the same according to standard microbiological methods) can belong to different clonal groups; knowledge of this situation is of paramount importance because it might indicate an incipient outbreak or the random appearance of strains from different sources in different patients. Moreover, integrating the information about strain isolation time and the contact network might enable an effective prediction of an incoming outbreak. The complete genome sequence additionally enables the comparison with all the genomes from the same species that are deposited in databases, potentially allowing to understand how pathogens move on a global scale and to trace the source of peculiar pathogens.

Antibiograms are obtained using selective media containing antibiotics of the most common families, which is indeed fundamental to treat the infected patients. The availability of the genome sequence would moreover allow a finer analysis of the complement of resistance genes of each strain and would enable the development or refinement of predictive models, able to tell if a certain genome contains the determinants for resistance to almost all known antibiotics for which the corresponding resistance genes are known. Methods using phylogenetic profiles of resistance genes to predict the actual resistances have been developed recently. One of these has shown that the genomic predictions are comparable to standard experimental techniques: in *S. aureus* the sensitivity and specificity of the predictive model turned out to be 99.1% and 99.6%, respectively, while the same approach in *Mycobacterium tubercolosis* resulted in much lower performances, likely for the lower knowledge we have about this pathogen (3). Several other methods have been tested in the last years and a recent work has shown their strengths and weaknesses (7). When resistance genes for a certain species are not very well-known, a possible approach is to use the entire complement of genes in the genome as predictors; in this case, as the number of predictors is much larger than the number of samples (usually in the hundreds, or thousands), it is mandatory to use statistical techniques able to select the important predictors and forget about the others.

A metagenomic approach

The cost of a single bacterial genome sequence is today around 100 Euros. This can be considered very cheap, but a real-time bacterial genomic monitoring might not be feasible in most hospitals, especially with public financing, considering that a medium size hospital might have to sequence 10-100 genomes per week to get a meaningful picture of the circulating pathogens. A possible compromise would be to adopt a shotgun metagenomic approach, where the circulating antibiotic resistance genes are searched by analyzing clinical samples or pooled isolates. Clearly, in this case it would not be possible to associate the sequences to precise isolates, reducing its utility with respect to obtaining the genome sequence of single isolates. Nonetheless, the pattern of abundance of resistance and virulence genes might provide useful information about the *infection status* of a hospital, that can then be used for the managing of the outbreaks, and provide a guide for the *empirical* treatment of the first cases.

Conclusions

To the knowledge of the authors, a hospital-based bacterial monitoring systems has not yet been established in Italy, that integrate genomic information together with the patients' contact networks, and with the other information recorded by the laboratory. However, as *routine* bacterial genome sequencing has come to age, we feel the time is ready for novel and more effective surveillance systems where NGS technologies are used, for getting a clearer picture of what is going on with bacterial pathogens within the hospitals, with a level of detail never approached before. Digitalization of hospital information of any kind, and the standardization of available digital platforms, is mandatory in this context, as it allows the development of software to perform the tasks that we presented in the above sections.

We are moreover convinced that similar platforms will be even more important in the future, when they could be integrated at a regional, national or international scale. The transfer of patients among different hospitals, possibly located in different regions or countries, might indeed be responsible for long-range transfer of pathogens, possibly in a cyclic way. The comprehension of the movement of patients and pathogen strains within and among hospitals would allow the identification of the possible sources of contamination, in a way somehow similar to what happens in food preparation chains. In this case, samples taken at different steps of the food chain are used to identify the source of the contamination; it is time to work for developing a similar system in the hospitals.

References

1. Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. J Theor Biol 2005;235:275-88.
2. Boguñá M, Pastor-Satorra, R, Vespignani A. Absence of epidemic threshold in scale-free networks with degree correlations. Phys Rev Lett 2003;90:1-4.
3. Bradley P, Gordon NC, Walker TM, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun 2015;6:10063.
4. Centers for Disease Control and Prevention. PulseNet. Available from: https://www.cdc.gov/pulsenet/pathogens/pfge.html.
5. Pastor-Satorras R, Vespignani A. Epidemic dynamics in finite size scale-free networks. Phys Rev 2002;65:1-4.
6. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett 2001;86:3200-3.
7. Schleusener V, Köser CU, Beckert P, et al. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Sci Rep 2017;7:46327.
8. WHO. Antimicrobial resistance. Bull World Health Organ 2014;61:383-94.