UNIVERSITÀ
DI PAVIA

UNIVERSITY OF PAVIA

Ph.D. course in Genetics, Molecular and Cellular Biology

XXXIV Cycle

Genomics, bioinformatics and epidemiological analyses for the study of microorganisms of public health interest

Ph.D. Thesis

Umberto Postiglione

472836

Scientific tutor: Prof. Davide Sassera

Academic Year: 2018-2021

ABSTRACT IN ENGLISH

For many years, the characterization and epidemiological surveillance of pathogenic bacteria has been based on the use of conventional techniques and methodologies which, although essential for a first form of investigation, do not allow to obtain a detailed and complete picture of the phenotypic and genetic variability of the pathogen itself. The limits of these 'classical' methods are clear in terms of turnaround times and detail of characterization, and thus represent bottlenecks in the timing of the implementation of surveillance measures and in the effectiveness of health treatments.

In this context, the implementation of modern bioinformatics and evolutionary approaches based on genomics and transcriptomics is currently making it possible to considerably expand the knowledge of the various bacterial strains, to fully characterize them and to reconstruct their epidemiological evolutionary history, with significant advantages both at a diagnostic and therapeutic level.

On these basis, in the following pages, I will show practical examples of the application of the latest generation sequencing techniques for the genomic and transcriptomic analysis of the following pathogens of interest: *Mycobacterium tuberculosis*, *Bacillus anthracis*, *Streptococcus agalactiae* and *Staphylococcus aureus*.

For *Mycobacterium tuberculosis* and *Streptococcus agalactiae*, I performed differential gene expression studies (RNA-Seq) to identify the altered molecular processes following the administration of the molecule 11726172 in *Mycobacterium tuberculosis* and the deletion of the *codY* gene in *Streptococcus agalactiae*. The analysis of RNA-Seq data resulted to be very fruitful, as it made it possible to delineate in detail the mechanisms of action of these two therapeutic approaches and to infer the genes involved in these processes.

In *Bacillus anthracis*, genome sequencing with Illumina and Oxford Nanopore technology was fundamental to study the compensatory mechanisms in response to the deletion of the *sap* and *eag* genes, encoding two proteins of the S-layer. The chromosome-level assembly of these genomes made it possible to carry out comparative analyses between wild-type and mutant strains, allowing the identification of variations at the level of gene content and single nucleotide mutations (SNPs) involved in the adaptation process.

Finally, for my main Ph.D. project, I carried out a retrospective study on a nine-year collection of *Staphylococcus aureus* within the San Matteo Hospital in Pavia (Italy). Thanks to the combination of metadata and genomic data, I was able to describe in detail the

phenotypic and genetic variability of the various strains of *Staphylococcus aureus* circulating within this hospital, highlighting the existence of a complex network of sequence types and clonal complexes (CC), of environmental and hospital origin.

My analyses allowed to detect the presence not only of highly abundant lineages, such as CC8 or CC22, but also of rarer STs, some of which with relevant resistance and virulence profiles (e.g ST30). Lastly, phylogenetic analyses based on coreSNPs, allowed to identify clusters of highly-related samples belonging to ST22 and ST8, responsible for persistent episodes of infection throughout the course of the nine analysed years.

In conclusion, this study represents a clear example of how a well-constructed research project can provide a complete depiction of the epidemiological and genetic status of a pathogen in clinical setting and therefore, direct the attention of medical personnel not only towards common variants, but also rarer ones than could escape normal surveillance controls.

TABLE OF CONTENTS

INTRODUCTION

Over the last decades, clinical laboratories and research centers have mostly relied on culture-based methods for the diagnosis, epidemiological surveillance and research on pathogens of clinical importance, which include plate culture for isolation of colonies, differential staining, morphological analysis, biochemical tests and lastly, the typing of isolates for the presence/absence of antigens (toxins or resistance genes) (De Almeida & De Martinis, 2019).

However, this culture-based identification, diagnostics and epidemiology is often not optimal. It can be laborious, time-consuming, unable to estimate portions of the actual microbial composition in the samples under study and only allows to determine a portion of the phenotypic characteristics of a microorganism. For example, this is particularly accentuated for slow-growing or highly pathogenic microorganisms, where the delay for definitive diagnosis can extend to weeks, limiting the correct management of infected patients and slowing down the monitoring of infectious diseases (Buchan & Ledeboer, 2014). To overcome these obstacles, major improvements have been made in recent years to increase sensitivity, specificity and response times for the identification and characterization of microorganisms of interest. In particular, culture-independent methods, such as mass spectrometry-based technologies, nucleic acid-based methods and immunoassays, have been introduced, allowing to swiftly detect emerging pathogens, characterize them and promptly respond to key epidemiological concerns (Bursle & Robson, 2016). However, even these approaches have severe limitations and that is why there is great interest in alternative methods that can complement these traditional procedures.

In this context, the recent development of many new bioinformatics techniques and integrated databases has transformed infectious disease research, facilitating the understanding of host-bacteria interaction and raising the expectation of better microorganisms control. Above all, the advent of high-throughput Next Generation Sequencing (NGS) has marked the beginning of a new era in data generation and analysis, enabling biological issues to be addressed at the genomic scale. The main impact of this technology has been to automate the sequencing of an organism's entire genome/ transcriptome in a single run, with relatively low cost and fast response times, providing comprehensive pathogen characterization.

SHORT-READ AND LONG-READ SEQUENCING

Overall, modern high-throughput platforms can be broadly divided into two categories: “short-reads” and “long-reads” sequencing technologies. Each platform offers its own advantages and disadvantages, in terms of accuracy, efficiency, and cost, depending on what the experiment is aiming to accomplish.

Short-read sequencing has been the standard for many microbial applications for over 10 years now, both in research and public health. Illumina sequencing, in particular, is the most widely used sequencing technology in these fields. Illumina sequencing workflow includes three basic steps: library preparation, sequencing, and data analysis. During library preparation, the genomic DNA is fragmented and adapters are added to both ends of these small segments. Then, the library is loaded onto a flow cell which has milions of adapters attached to the surface, which can match the adapters added at the ends of the DNA fragment in the previous process. Each fragment is then amplified into a cluster of segments through bridge amplification PCR. Finally, sequencing reagents are added, including DNA polymerase, connector primers and 4 dNTP with base-specific fluorescent markers. The fluorescence marker is excited by laser and the signal is recorded by optical equipment and translated into sequencing bases. (Buermans & den Dunnen, 2014).

This procedure allows to produce millions of low-error (0.1%), generally 100–300 bp in length, single-end or paired-end reads. In single-end reading, the sequencer reads a fragment only from one end. Instead, in paired-end reading, the sequencer starts at one end, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment. This last approach is more effective than single-end reading, making it possible to improve the ability to identify the relative positions of various reads in the genome and to solve repetitive regions or structural rearrangements, such as gene insertions, deletions, or inversions. Due to these characteristics, Illumina reads result to be extremely suitable for applications such as SNPs calling or characterization of loci (e.g virulence/resistance genes in the accessory genome) (Didelot et al., 2012).

However Illumina technology, as well as all others short-read platforms, comes with one significant limitation: the inability to sequence long stretches of DNA and, as a consequence, to generate a sufficient overlap between fragments. Because of this, these methodologies are unable to solve highly complex and repetitive areas and to fully reconstruct genomic structures of interest (on chromosomes or mobile genetic elements) that extend beyond the maximum read length generated (George et al., 2017).

In microbiology this is for example an issue with plasmid characterization, which is nigh impossible with such long reads.

On the other hand, long-reads sequencing technologies are able to overcome these limitations by generating reads with a median length of 5–30 kb (Loman et al., 2015), which make them suitable to solve complex repetitive regions and for de-novo assembly (for species with no or low quality reference), which assumes no prior knowledge of the source DNA sequence length, layout or composition.

One of the predominant long-read sequencing technologies available is Oxford Nanopore Technologies (ONT). This platform uses a flow cells which contain an array of nanopores embedded in an electro-resistant membrane. Each nanopore corresponds to its own electrode connected to a channel and sensor chip, which measures the electric current that flows through the nanopore. When molecules such as DNA or RNA move through the nanopores, they cause disruption in the current. The signal disruption is then decoded using base calling algorithms to determine the DNA or RNA sequence in real time (Buermans & den Dunnen, 2014). This technology has the great advantage of being capable of producing extremely long reads (up to 1 milion base pairs) but, at the same time, the sequencing error rate is particularly high, being estimated at 11–15 %.

One possible solution to these issues has been to complement short-reads (e.g Illumina) with long-reads (e.g ONT), using an “hybrid approach”. With this method, long reads provides information regarding the structure of the genome and short reads facilitate assembly at local scales, and can be used to correct errors in long reads (Risse et al., 2015), enabling the generation of high-quality “finished” microbial genomes with low sequencing errors.

Very recently, improvements in long reads technologies have been extremely promising. Error rates are lowering, together with prices, and breakthrough such as hifi sequencing (Wenger et al., 2020) open the realistic possibility for long reads to fully displace Illumina in microbiology in the near future.

NGS APPLICATIONS

Starting from single or polymicrobial specimens, NGS sequencing analyses can be used in microbiology for multiple purposes, such as (1) typing of pathogens (detection of emerging viral or resistant determinants), (2) outbreak reconstruction or (3) differential genes expression analysis (Köser et al., 2012). The first two applications rely on genomics while the third on transcriptomics.

TYPING OF PATHOGENS

NGS can serve as a perfect one-step tool to study and characterize a broad range of pathogens, providing information about relevant genomic features, such as sequence of typing markers, virulence or resistance genes (ESCMID Study Group on Molecular Epidemiological Markers (ESGEM), 2021). Knowledge of the virulence and resistance profile of a pathogen is crucial to estimate the severity of a specific disease, the outcome of the infection and to allow risk assessment during the early onset of the infection.

- **Typing of virulence factors**: In the process of host–pathogen interactions, microbes always employ specific genes to adapt to new niches and cause damage or diseases to the host. In this context, virulence factors (VFs) play a major role. Some are intrinsic to the pathogen but for the major part they are laterally acquired through horizontal gene transfer (HGT) and encoded on chromosome or plasmids. According to their function, VFs can be roughly classified in six categories (i) adherence and colonization factors, (ii) Type I to VI secretion systems, (iii) immune evasion factors, (iv) toxins, (v) siderophores for iron absorption and (vi) invasion genes (Niu et al., 2013). Identification of these molecules using NGS can be performed in different ways , such as by homology search or through comparative genomic analyses (comparing the genomes of non-virulent and virulent strains) (Bakour et al., 2016). Above all, homology-based approaches have proved to be effective in the search for known and conserved VFs and several types of investigations can be carried out. Some examples are:

1. Sequence Similarity Search: it aims at obtaining orthologous sequences corresponding to a given query. The best known algorithm for this task is the Basic Local Alignment Search Tool (BLAST) algorithm.
2. Sequence Motif search: with this approach it is possible to identify patterns of amino acids or nuclotides which are characteristic of a specific biochemical function.
3. Domain Search: functional domains of a protein are generally well conserved in terms of sequence and folding. Thus, the domain information of an unannotated

protein sequence can be used to predict its function (Chaudhuri & Ramachandran, 2014).

These methodologies are implemented in open-source tools, curated databases or stand-alone tools and widely used in numerous diagnostic applications. However, they can only detect by similarity, and they result to be useless for unknown or evolutionary distant virulence genes.

- **Typing of antimicrobial resistances**: Antimicrobial resistance (AMR) is one of the major public health burdens of the twenty-first century, responsible worldwide for increasing mortality rates and economic costs (Tacconelli & Pezzani, 2019). Resistant bacterial strains are difficult to treat because identifying what antibiotic is still effective is time-consuming and in a growing number of cases, physicians have also to deal with multiresistant microbes or, in some, fortunately rare cases, pan-resistant microbes, which are resistant to all known antimicrobial agents.

  Resistance to antibiotic can occur through various mechanisms including: (i) degradation or enzymatic modification of the antimicrobial, (ii) overproduction, protection or modification of the antimicrobial target, (iii) antimicrobial efflux and (iv) change in cell permeability resulting in restricted access to the target site (Wright, 2010). In general, these mechanisms are either the result of chromosomal point mutations or more frequently the result of horizontal gene transfer events (Van Hoek et al., 2011). To determine AMR, clinical laboratories and hospitals measure the so-called minimum inhibitory concentrations (MICs), which is defined as the lowest concentration of an antimicrobial that inhibits the visible growth of a microorganism. MIC values are then compared to clinical breakpoints to establish if a microbe is clinically resistant or susceptible to a certain drug. As for virulence factors, molecular methods are necessary, in that they allow resistance genes or point mutations to be detected using either homology-based methods or doing comparative analyses. Either way, these approaches are not recommended for novel or remote homologous AMR genes/mutations.

  Lastly, a promising field for AMR detection is the application of machine learning algorithms to predict resistance phenotypes and/or AMR determinants (Nguyen et al., 2019). These algorithms may be extremely useful for the detection of unknown AMR features. In fact, it is not uncommon to observe bacterial strains without known genetic components at the basis of phenotypic resistance and, even though these cases are fewer compared to the overall population, these microorganisms represent an important challenge. By using NGS sequencing information and phenotypic data (e.g antimicrobial susceptibility test results), machine learning algorithms could be trained and optimized to predict resistance phenotypes and rank AMR determinants by their importance in discriminating the resistance from the susceptible phenotypes (Sunuwar & Azad, 2021).

Overall, all of these strategies represent a valuable set of assets, but they still suffer from several drawbacks. From a biological point of view, the mere presence of a determinant may not be sufficient to explain a resistant/virulent phenotype because many other genes or regulatory pathways may be involved and influence the outcome. Furthermore, these genes can also be found in non-pathogenic strains (Niu et al., 2013), also highlighting the importance of the environment in the development of pathogenicity. Therefore, a complete replacement of standard laboratory procedures (phenotypic measurements and biochemical tests) with predictive algorithms or homology-based approaches is not recommended, as microbial strains continue to evolve and display new mechanisms of resistance and virulence, which may be overlooked if not represented in databases or in the datasets used to train machine learning models. Phenotypic testing of a representative genomic diversity of strains needs to be maintained to ensure that genotypic results do not diverge from the true phenotype over time.

OUTBREAK RECONSTRUCTION

Most of outbreak management protocols are based on phenotypic and molecular tests (serotyping, molecular typing, susceptibility tests and mass spectrometry) to monitor and limit the spread of dangerous pathogens, such as multi-drug resistant bacteria (MDR).

However these tools, despite being still essential, often fail to correctly reconstruct outbreaks and to detect relevant virulence/resistance determinants. This is partly due to their limited resolution power and to the target-specific nature of outbreak analysis approaches. For instance, in the fight against MDR bacteria, much attention is given to antimicrobial resistances but not enough to virulence genes, which are known predictors of higher mortality rates and longer hospitalization stays (Leopold et al., 2014). To overcome these limitations, novel technologies such NGS can be used to retrieve detailed and full information about the entire genome of a pathogen.

NGS has proved to be important and helpful in disclosing and tracing the dissemination of emerging pathogens, playing an essential role in epidemiological studies and public health investigations. The ability to precisely infer transmission linkage between isolates from different sources (environmental, animal or human) and to reconstruct the phylogenetic relationship between isolates, have pushed many countries (e.g the United States, Denmark, the United Kingdom, Germany, and The Netherlands) to implement NGS as complementary typing method for national surveillance screening (ECDC 2016).

From a bioinformatics point of view, once assemblies are obtained and genomes are fully characterized (taxonomy identification, presence/absence of genes), comparative genomic analyses are performed to to detect relatedness between strains. In a first step, the similarity between different genomes is estimated by different approaches, leading to the generation of a distance matrix. Then, various clustering strategies (e.g. neighbor-joining trees, minimum-spanning trees, hierarchical clustering) can be applied in order to allow the construction of a phylogenetic tree. Thus, all samples falling below a specified distance threshold can be considered to belong to the same cluster. So far, several methods have been developed to assess and cluster closely-related strains, such as cgMLST, K-mer and SNP-calling-based strategies (Uelze et al., 2020).

- **cgMLST**: Multilocus sequence typing (MLST) is a gene-based procedure used for the characterization of bacterial isolates. It uses the sequences of internal species-specific house-keeping genes and, for each gene, all unique sequences are assigned allele numbers and combined into an allelic profile, defining a sequence type (ST). An extension of MLST is the so-called core-genome MLST (cgMLST), an approach with an increased number of core-genes (genes present in all members of a given population subset). The genetic similarity between genomes is calculated using a gene-by-gene approach comprising thousands of genes and associated allele sequences. Each isolate, in a species-specific way, is characterized by its allele profile and the cross-comparison of samples yields the allele distance matrix. This method is efficient for species with many publically available schemes but useless for less characterized microbes. Plus, the absence of an organized allele nomenclature system makes it still unsuitable for wide investigational studies.
- **K-mer**: genomes are splitted into nucleotide blocks of a defined length k (K-mer). Then, the pair-wise comparison of the k-mer content between a set of genomes is used to evaluate their phylogenetic relatedness. One major downside of K-mer-based methods is the proper K-mer size determination, as it is time-consuming and a wrong choice can be associated with low quality results.

- **SNP-calling**: selectively neutral SNPs tend to accumulate at uniform rate across the time as result of spontaneous mutations, and are found along the entire length of genomes, falling in genic and intergenic regions. The usually high number of point mutations can be then used to measure the evolutionary distance between genomes, minimizing assembly errors and the impact of strong selective pressure (Shakya et al., 2020). SNPs are identified by aligning assembled draft genomes to an annotated reference genome and only reference SNPs that are covered by all query genomes are considered, which define a set of core SNPs. All possible combinations of pairwise SNP distances determine the SNP distance matrix which is ultimately used for phylogenetic analysis. This approach possesses the highest discriminatory power of all comparative genomics approaches and it's strongly recommended for those species having a high-quality reference. If not the case, also non-reference-based SNP analysis can be employed (Quainoo et al., 2017).

However, for high-quality results, the proper choice of the reference genome is essential. Ideally, the reference should be closely related to the set of strains under investigation in order to maximize the number of SNPs discovered. In addition, SNP-based methods are also affected by some inherent bias, such as HGT, recombination, and rate heterogeneity, thus requiring a careful curation of the obtained results.

DIFFERENTIAL GENES EXPRESSION ANALYSIS

Bacteria can easily and quickly adapt, through changes in gene expression, in response to multiple stimuli. Such changes are fundamental for surviving and can provide valuable clues about the composition of gene networks, the nature of the underlying genes and also tissue type, genotype and species (Coate & Doyle, 2015). Various technologies have been developed to study the transcriptome, mostly based on hybridization or sequence-based approaches. The power of hybridization methods lies in the capacity to map relatively large regions of the transcriptome, up to tens of thousands of various mRNA transcripts, at very high resolution and low costs. However, these methods present hard constraints, which include: a priori knowledge about genome sequence, high background levels due to cross-hybridization and the necessity to normalize results to compare expression levels across different experiments.

Sanger sequencing technology instead offers the chance to detect only a limited portion of the transcriptome without a priori knowledge about the sequence. Nonetheless, many disadvantages come with this methodology: high costs, impossibility to distinguish transcript isoforms and low power of resolution. In conclusion, the study of the transcriptome is a difficult challenge and still little is known about the absolute composition of the mRNA population within bacterial cells.

Only with the advent of NGS has been possible to provide a new technology capable of mapping and quantifying transcriptomes. This method, termed RNA Sequencing (RNA-Seq) has clear advantages over existing approaches and has emerged as a new powerful method for studying the transcriptome of microbes, creating many opportunities to improve functional genomics experiments, such as differential gene expression analyses (Z. Wang et al., 2010).

An adequate planning of sequencing experiments is crucial in order to obtain error-free data. In general, the first step is extraction and enrichment of RNA from cells. Typically, ribosomal RNA (rRNA) is highly abundant, constituting up to 90% of total RNA in the cell, therefore it is a standard practice to deplete the rRNA from a total RNA sample such that the reads in an RNA-seq experiment derive predominantly from messenger RNA (mRNA, 1-2%). Once RNA is extracted and purified, there is the sequencing phase: RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences (single-end or pair-end sequencing), of typical 150-250 bp. For the sequencing process, the library size is an important factor to be taken into account. With more reads, the sample is sequenced to a deeper level and as a consequence, transcript quantification will be more precise (Mortazavi et al., 2008). However, experimental results suggest that also the over-generation of reads must be avoided as it could lead to detection of transcriptional noise and off-target transcripts. Another important aspect is the choice of the number of replicates. Usually, three replicates for each biological sample are considered to provide good results, limiting technical bias and normalizing the biological variability of the system under study.

Following sequencing, a series of major steps are required for the analysis of RNA-Seq data, including quality control, read alignment (with and without a reference genome) and quantification of transcripts expression.

1. **Quality check**: quality of reads is assessed in terms of sequence quality, GC content, presence of adaptors, overrepresented k-mers and duplicated reads. Sequence quality is generally expressed as “phred score” (Ewing & Green, 1998) which is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing. All the other parameters, such as

duplication levels, k-mer and GC content, are organism-specific values and should be homogeneous across all samples in the same experimental condition. One popular tool for quality check of Illumina reads is FastQC (*FastQC*, 2015). For low-quality reads, tools such as Trimmomatic (Bolger et al., 2014), can be used to eliminate poor-quality bases.

2. **Read alignment**: two alternatives are possible when a reference sequence is available: either mapping of reads onto the genome or onto the transcriptome. Instead, if the organism does not possess a genome to be used as reference, reads are first assembled into longer contigs ("de-novo assembly") which are later used as the expressed transcriptome to which reads are mapped back again for quantification. In this phase, an important indicator of mapping quality is the percentage of mapped reads. Reads may map uniquely (they can be assigned to only one position in the reference), they may map onto multiple loci or, in the worst scenario, cannot be mapped at all. A high percentage of unmapped or multi-mapped reads is synonymous of low accuracy and of the presence of contaminating DNA (Conesa et al., 2016). One popular tool for this step is Bowtie (Langmead et al., 2009).
3. **Transcript quantification**: By utilizing a gene transfer format (GTF) file containing the genome coordinates of genes it is possible to quantify the number of reads that map to each transcript sequence. However, to compare reads count across samples, it is important to normalize raw counts, as these values may be affected by factors such as transcript length, total number of reads, and sequencing biases. Measures such as RPKM (reads per kilobase of exon model per million reads), FPKM (fragments per kilobase of exon model per million mapped reads) or TPM (transcripts per million) are the most frequently reported RNA-seq gene expression values. Tools such as featureCounts (Liao et al., 2014) are widely used for gene quantification and normalization.
4. **Differential gene expression analysis**: Lastly, the normalised read count data are obtained and statistical analyses are performed to estimate quantitative changes in expression levels between experimental groups. For any given gene, the observed difference in read counts is considered significant if it is greater than what would be expected just due to natural random variation. Differential gene expression tools make use of different algorithms to characterize changes in expression levels. For instance, DESeq2 (Love et al., 2014) uses a negative binomial generalized model

to compute differentially expressed genes (DEGs) and, for each gene, it provides multiple parameters such as a p-value corrected for multiple testing (the Benjamini and Hochberg method) and log2 fold change (log2FC) between treatment and control. The log2FC indicates how much the gene or transcript's expression seems to have changed between the comparison and control groups. If the log2FC is a positive value, the gene is up-regulated in the treated samples; on the opposite, if the log2FC has a negative value, the gene is down-regulated in the treated samples. Thus, the log2FC value expresses how much the gene is differentially expressed. Thus, if the p-value is lower than the threshold (usually set at 0.05), the gene is differentially expressed (down regulated or up regulated). On the contrary, If the p-value is bigger than the threshold, the differential expression is not statistically significant.

RNA-Seq has several advantages compared to other methods that aim to characterize gene expression: the whole transcriptome is studied in an unbiased manner, without requiring any sort of probe (attractive solution for non-model organisms) or complex normalization. This enables the study of all transcripts at nucleotide resolution, allowing the discovery of novel genetic features, detection of different isoforms and permitting the delineation of operons and untranslated regions. Additionally, RNA-Seq allows the quantification of gene expression, counting the amount of reads matching a given coding sequence (CDS) (Croucher & Thomson, 2010).

Besides some technical hurdles, like other high-throughput sequencing technologies, RNA-Seq faces several informatics challenges, including the need of infrastructures to store and process large amounts of data. Furthermore, the mapping step can be tricky when reads match multiple locations in the genome, lowering the resolution power of detection. Lastly, the proper choice of coverage depth is fundamental and strictly related to the experimental biological question: greater coverage requires more sequencing depth, which is essential to detect a rare transcript or variant.

PREFACE TO WORK

In the following sections I have included a collection of research works performed during my period as a doctoral student at the University of Pavia. For each microorganism, (1) *Mycobacterium tuberculosis*, (2) *Bacillus anthracis*, (3) *Streptococcus agalactiae* and (4) *Staphylococcus aureus*, I am going to describe the bioinformatics techniques I used and the main results obtained.

Mycobacterium tuberculosis

INTRODUCTION

Mycobacterium tuberculosis is a gram-positive acid-fast bacterium belonging to the *Mycobacteriaceae* family and it is a causative agent of tuberculosis (TB). It is a top infectious killer, responsible for 1.8 million deaths per year (over 95% of cases and deaths are in developing countries) and for around a quarter of all deaths caused by AMR bacteria, with nearly half a million estimated multidrug-resistant tuberculosis cases annually (WHO, 2021). The bacterium is transmitted by airborne droplets generated by coughing (Churchyard et al., 2017) and, once in the human body, it can be either immediately eliminated, it can survive inside macrophages in a quiescent state, causing latent TB infection (LTBI) or it can cause primary active TB. From a clinical point of view, *M. tuberculosis* affects mostly the lungs and rarely other sites (e.g brain, kidneys or the spine), causing a plethora of symptoms, with seriousness ranging from asymptomatic infections to life-threatening conditions, such as pulmonary TB. The probability of developing TB disease is much higher among individuals who already suffer from conditions that impair the immune system (e.g. HIV) or with undernutrition, alcohol and tobacco smoking disorders (WHO 2021).

This disease is curable and preventable, with about 85% of patients successfully treated with a 6-month drug regimen. This therapy involves the use of a combination of multiple drugs: the first 2 months patients are treated with isoniazid (INH), rifampicin (RIF), ethambutol and pyrazinamide followed by 4 months of isoniazid and rifampicin (WHO, 1991). This regimen has proved to be highly effective for drug-susceptible TB (Frieden et al., 1995; Murray, 1996) but less for resistant strains.

Mechanisms of drug resistance

Drug resistant *M. tuberculosis* strains are a major global health concern, responsible for costlier and longer treatment regimens and poor prognosis.

According to their level of resistance, they can be classified as multidrug-resistant (MDR) or extensively drug-resistant (XDR). MDR *M. tuberculosis* are resistant to both first-line tuberculosis drugs, rifampicin and isoniazid, while XDR *M. tuberculosis* manifest resistance also to fluoroquinolones and second-line drugs, lowering the rate of success to 54% and 28%, respectively (WHO 2021).

Unlike what happens in most other bacteria, drug resistance in *M. tuberculosis* is not generally due to gene acquisition through horizontal transfer (Ducati et al., 2006) but it has evolved mostly through several mutational-based mechanisms which include compensatory evolution, epistasis, clonal interference, cell envelope impermeability, efflux pumps, drug degradation and modification, target mimicry and phenotypic drug tolerance (Al-Saeedi & Al-Hajoj, 2017). Thus, failure in TB treatment may be due to intrinsic (naturally occurring high levels of antibiotic resistance) and extrinsic (newly acquired mutations) antibiotic resistance.

For instance, resistance to RIF is mainly monofactorial and it is due to mutations in the active site of the gene *rpoB*, a DNA-dependent RNA polymerase. Alterations in this site decrease the affinity of RIF for its target (Meftahi et al., 2016). Instead, resistance to INH involves multiple mechanisms. In fact, INH is a pro-drug activated by catalase-peroxidase (KatG), and target the subsequent enzymes namely, enoyl acyl carrier protein (ACP) reductase (InhA) and beta-ketoacyl ACP synthase (KasA). Thus, the mechanisms associated to INH resistance can involve: abrogated prodrug activation due to missense mutations in *katG* (responsible of isoniazid activation) (Heym et al., 1995), drug target alteration of *inhA* or overexpression of *kasA* (Morlock et al., 2003).

Besides the genetic characteristics of the pathogen, the spread and evolution of *M. tuberculosis* resistant strains is also exacerbated by a series of socioeconomic and behavioral factors that have led to more dangerous forms of resistant TB. Failure to identify and appropriately treat resistant-TB patients, the inadequate implementation of control measures, administration of low-quality drugs and patient non-adherence are all factors that have limited the proper management of patients and the effectiveness of surveillance strategies (Tola et al., 2015).

Due to the lack of effective and definitive treatments to treat *M. tuberculosis* drug-resistant strains, new diagnostic tools, more effective drugs and adjunct therapies are urgently needed to improve the treatment outcome. Different approaches (e.g drug repurposing, nano-based drugs, high-throughput screening) have been pursued to find new anti-TB candidates but the rate of discovery is limited by our lack of knowledge of all resistance mechanisms. In this context, the analysis of transcriptional responses to antimycobacterial compounds is useful in order to study the principal mode of action of such compounds and to provide a better understanding of the effect of antibiotics on *M. tuberculosis*.

AIM OF THE WORK

As part of the collaboration with The Molecular Microbiology Laboratory of the University of Pavia, I investigated the mechanism of action of a new compound named 11726172 (4-nitrobenzo[c][1,2,5]thiadiazol-5-yl thiazolidine-3-carbodithioate), discovered by Dr. Vadim Makarov (A. N. Bach Institute of Biochemistry, Russian Academy of Sciences, Moscow, Russia). It was found that this compound was endowed with antitubercular activity (MIC 0.25 μg/ml) so, using a transcriptomic approach (RNA-Seq), I studied the transcriptional response of *M. tuberculosis H37Rv* upon 11726172 exposure.

I investigated variations in gene expression levels in two independent experiments:

1. Differential gene expression analysis between *M. tuberculosis H37Rv* cells treated with 2.5 μg/ml of the drug 11726172 (10-Fold MIC) and untreated *M. tuberculosis H37Rv* cells;
2. Differential gene expression analysis between *M. tuberculosis H37Rv* cells treated with 7.5 μg/ml of the drug 11726172 (30-Fold MIC) and untreated *M. tuberculosis H37Rv* cells.

MATERIALS AND METHODS

Bacterial strains, culture conditions and chemicals

The *M. tuberculosis H37Rv* strains were grown at 37°C in 7H9 broth (Difco) supplemented with 0.2% glycerol, 0.05% Tween 80 or Tyloxapol in final concentration 0.05%, or on 7H11 (Difco) plates supplemented with 0.5% glycerol, both supplemented with 10% oleic acid-albumin-dextrose-catalase (OADC, Middlebrook). The compounds were dissolved in DMSO (Sigma-Aldrich). All the experiments with *M. tuberculosis H37Rv* were performed in Biosafety level-3 laboratory by authorized and trained researchers.

Determination of Minimal Inhibitory Concentration (MIC)

The drug susceptibility of *M. tuberculosis H37Rv* strains was determined using the resazurin microtiter assay (REMA), as previously described (Palomino et al., 2002). Briefly, log-phase bacterial cultures were diluted to a theoretical OD600=0.0005 and grown in a 96-well black plate (Fluoronunc, Thermo Fisher) in the presence of two-fold serial compound dilution. In some cases, sterile solutions of CuSO4, NiSO4, CoCl2 and ZnSO4 in the concentration range from 5 to 50 µM were added to determine the influence of metal cations to MIC values. A growth control containing no compound and a sterile control without inoculum were also included. After 7 days of incubation at 37°C, 10 µg/L of resazurin (0.05% w/v) were added and fluorescence was measured after 24 hours further incubation using a FluoroskanTM Microplate Fluorometer (Thermo Fisher Scientific; excitation=544 nm, emission=590 nm). Bacterial viability was calculated as a percentage of resazurin turnover in the absence of compound.

Preparation of *M. tuberculosis* cultures

M. tuberculosis H37Rv strain was grown in Middlebrook 7H9 at 37°C, starting from an OD600 of 0.06. When the cultures reached an approximately OD600 of 0.4 (exponential phase), samples were treated with 11726172 compound at the following final concentrations: 2.5 µg/ml and 7.5 µg/ml, corresponding respectively to 10-fold and 30-fold MIC of the wild-type strain. As a control, an untreated *H37Rv* strain was used. After drug addition, the *M. tuberculosis* cultures were incubated at 37°C for 4 hours. Then, the cells were collected by centrifugation (3500 rpm for 6 minutes) and stored at -80°C until use. Three biological replicates per condition were set up.

RNA extraction

RNA was extracted as previously described (Uplekar et al., 2013). *M. tuberculosis* cells stored at -80°C were resuspended in 1 ml Trizol (Ambion) in a 2-ml screw-cap tube containing 0.5 ml zirconia beads (BioSpec Products) for disruption of *M. tuberculosis* cell wall by vortexing (twice for 5 minutes with a 2-minute interval on ice). Then, each sample was centrifuged at 4100 rpm for 1 min and the supernatant was transferred to a new tube, where 1 volume of chloroform-isoamylalcohol (24:1) solution was added. The reaction was centrifuged at 12000 rpm for 5 minutes and the supernatant was then transferred in a new tube. RNA was precipitated by adding 1/10 volume of sodium acetate (2M, pH 5.2) and 0.7 volume of isopropanol. After a 30-minutes -80°C incubation, and a 30-minutes 4°C centrifugation at 12000 rpm, the supernatant was eliminated, and RNA was washed with 1 ml of 70% ethanol. The sample was centrifuged for 30 minutes at 12000 rpm at 4°C. The supernatant was eliminated and the sample was air-dried and resuspended in 88 µl DEPC-treated water. DNase treatment was performed using TURBO DNA-free™ Kit (Ambion), following the manufacturer's recommendations. To ensure that all the contaminant DNA was eliminated from the samples, the treatment was repeated twice. Then, the reactions were subsequently cleaned up by phenol-chloroform extraction and ethanol precipitation. The amount and purity of RNA were determined spectrophotometrically and integrity of RNA was assessed on a 1% agarose gel.

RNA sequencing

RNA-seq was performed by IGA Technology Service (https://igatechnology.com/, Udine, Italy) using the Illumina platform. Strand-specific Illumina libraries of total RNA were prepared after a ribosomal RNA depletion step, using the Universal Prokaryotic RNA-Seq, Prokaryotic AnyDeplete® library preparation kit (NuGEN, San Carlos, CA). Sequencing libraries were then generated using the resulting ribosomal transcript-depleted RNA and sequenced in single-end 75 bp mode on NextSeq 500 (Illumina, San Diego, CA).

Transcriptome and bioinformatics analysis

Raw reads were quality checked using FASTQC (*FastQC*, 2015) and processed by Trimmomatic (Bolger et al., 2014) to trim the adaptor sequences and remove low-quality sequences. The remaining clean reads were mapped onto the reference genome *M. tuberculosis H37Rv* using Bowtie2 (Langmead et al., 2009). To quantify the known transcripts, the alignment results were inputted into featureCounts (Liao et al., 2014).

Lastly, the R package DESeq2 (Love et al., 2014) was used to test for differential expression. I defined genes as differentially expressed using the following criteria: log2FoldChange >= |2.5| and FDR < 0.05. For the meta-analysis, I used the R package metaRNASeq (Rau et al., 2014) and finally, I analysed differences in the enrichment of Gene Ontology (GO) categories for the DEGs using the DAVID (FDR threshold < 0.05) functional annotation analysis tool (Huang et al., 2009).

Real-time PCR

1µg of purified total RNA was retro-transcribed using QuantiTect® Reverse Transcription kit (Qiagen). Real-Time PCR experiments were performed using QuantiTect SYBR Green PCR Master Mix (Qiagen) kit and the thermocycler "Rotor Gene 6000" (Corbett Life Science) to amplify and quantify cDNA sequences of interest. Housekeeping gene *sigA* was chosen as a reference gene, encoding σA factor of *M. tuberculosis*.

RESULTS AND DISCUSSION

Transcripts quantification and exploratory analyses

To investigate the global gene expression pattern of *Mycobacterium tuberculosis* in response to treatment with the molecule 11726172, genome-wide expression analysis using RNA-Seq was performed. About 439 million single-end reads were generated from the three biological groups (10-Fold MIC, 30-Fold MIC and untreated cells), each with three biological replicates.

High-quality trimmed reads (min 26.4, max 77.7 milions) were mapped onto the reference genome *Mycobacterium tuberculosis H37Rv,* resulting in a high percentage of mapped reads (min 94.75%, max 98.59%) which were later counted (23.7 ± 9.2 million reads per sample) and summarized at CDS level (Table 1).

Samples	Raw_Reads	After_Trimming	Mapped_%	Counted_Reads
Sample1_Control	33833242	33350594	98,1	17038874
Sample2_Control	77767940	76601722	98,59	40218563
Sample3_Control	49902511	49164007	97,63	28024691
Sample4_10Fold_MIC	47319152	46647939	95,13	27294037
Sample5_10Fold_MIC	78920468	77757204	97,43	33109131
Sample6_10Fold_MIC	34334013	33783582	96,48	18418477
Sample7_30Fold_MIC	52155575	51300106	97,61	20465892
Sample8_30Fold_MIC	26807666	26425464	96,95	10312389
Sample9_30Fold_MIC	38063599	37517224	94,75	18459726

Table 1: The RNA-Seq data for the 9 *M. tuberculosis* samples

Before starting with the differential expression analysis, the overall similarity between samples was assessed, in order to ensure that our replicates clustered together according to the experimental group they belong to. Using a regularized log transform (rlog) of the normalized counts, a principal component analysis (PCA) was performed in order to identify any sample outliers, which may need to be explored further to determine whether they need to be removed prior to DE analysis. In particular, we see (Fig 1) that the untreated (control group) and treated samples (10-Fold MIC and 30-Fold MIC groups) are well separated along the first direction (PC1), which represents 76% of the total variance. Instead, the 10-Fold MIC and 30-Fold MIC samples appear to be best separated along the second direction (PC2) which collects only 10% of the variance. These results confirm the solidity of the project's experimental design, as samples with similar expression levels for genes that significantly contribute to the variation are expected to be drawn close to each other. Conversely, samples with extremely different gene expression patterns should be plotted much further along the principal components.

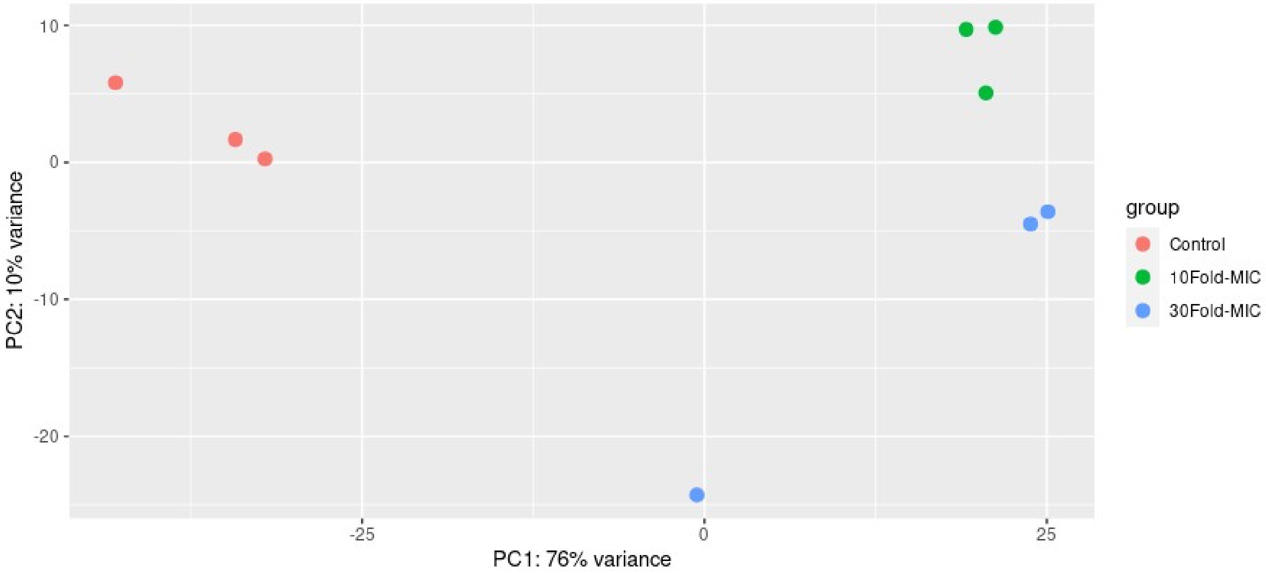


Fig 1: Principal component analysis (PCA) plot of *M. tuberculosis* samples

Differential gene expression analysis

Out of 3906 coding DNA sequences, 167 (4.2%) and 146 (3.7%) genes were found to be differentially expressed in *M. tuberculosis H37Rv* cells treated with 2.5 and 7.5 µg/ml of the drug 11726172, respectively. As it is possible to observe from the Volcano plot below (Fig 2), in the first experimental comparison, 126 genes were found to be up-regulated (2.5 - 7.99 fold) and 41 genes down-regulated (2.5 - 4.07 fold). Instead, in the second comparison, 127 genes were found to be up-regulated (2.5 – 8.02 fold) and 19 genes down-regulated (2.5 - 4.19 fold).

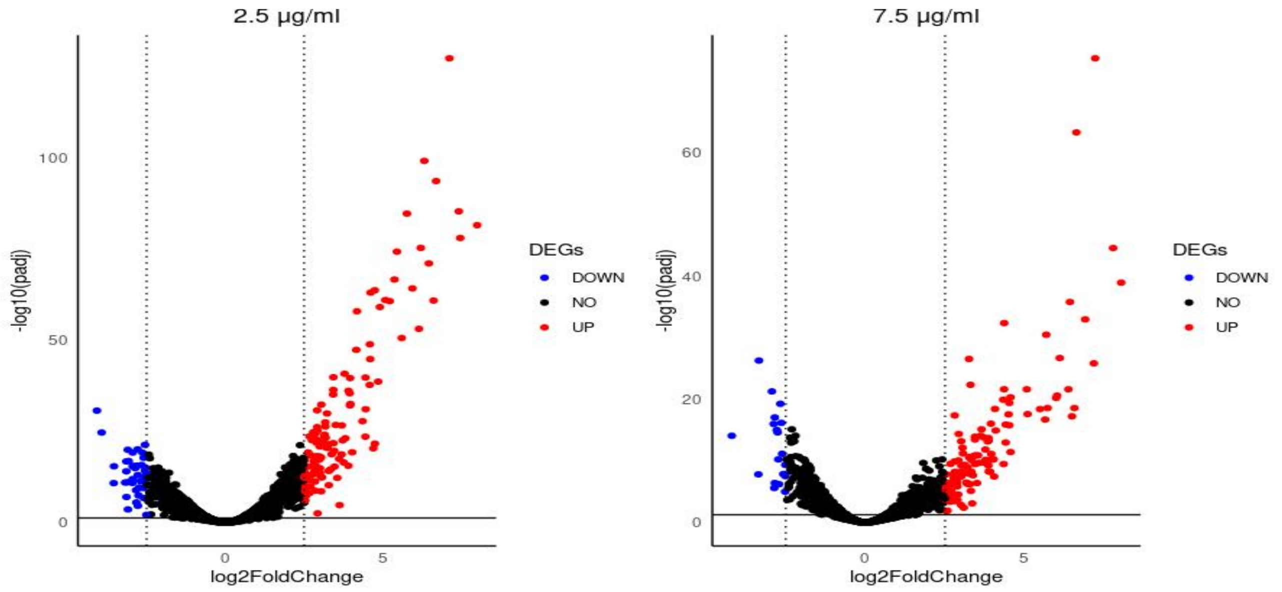


Fig 2: Volcano plots of differentially expressed genes for each condition (2.5 and 7.5 μg/ml)

Meta-analysis

Usually, RNA-seq experiments are performed on very few biological replicates, and therefore analyses to detect DEGs tend to lack detection power. Plus, biological and technical variability among samples (e.g sample preparation, library protocols, batch effects) may contribute to furtherly reduce the significance of the results. Thus, to increase the detection power of my analysis and to exclude possible inconsistencies, I decided to use the R package “metaRNASeq”. This package implements two p-value techniques (inverse normal and Fisher methods) to furtherly confirm the identification of DEGs obtained during the differential expression analysis.

All DEGs obtained from the two independent conditions were confirmed to be differentially expressed according to the inverse-normal and Fisher methods techniques. However, when considering the log2FoldChange, 4 DEGs (*Rv1575*, *Rv3541c*, *Rv1886c* and *Rv2490a*), two for each condition, resulted to change expression in opposite direction, so we decided to not include them in the following analyses. Thus, at the end of this analysis, we ended up with a final list of 165 DEGs from the first experiment, and 144 DEGs from the second experiment, for a total number of 92 DEGs (88 up-regulated and 4 down-regulated) which showed congruent expression variation in both experiments.

Up and down regulated DEGs

Among the up-regulated DEGs in common between the two conditions, I found several known genes with possible links to development of antibiotic resistance. *CadI* (encoding a putative metal transporter, mean log2FoldChange: 7.8) and *furA* (encoding a ferric uptake regulation protein, mean log2FoldChange:5.3) are two genes involved in metals-metabolism, which is an important pathway in regulating drug tolerance (Sepehri et al., 2017). *CysK2* (mean log2FoldChange:3.2) is involved in the biosynthesis of mycothiol, a signaling molecule triggering responses upon exposure to reactive oxygen species. In particular, it has been seen that addition of small thiols along with isoniazid and rifampicin prevents the emergence of drug-tolerant but also drug-resistant cells leading to sterilization of the cultures in vitro (Vilchèze et al., 2017). I also found the *cyp135A1* (mean log2FoldChange: 6.9), a gene coding for a cytochrome P450. Cytochromes P450 are a group of heme-thiolate monooxygenases, which oxidize a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. Thus, this specific c*yp135A1* could be involved in the detoxification process correlated to 11726172. Another induced gene in this category is *trxC* (mean log2FoldChange: 5.3), coding for a thioredoxin. Thioredoxins participate in various redox reactions through the reversible oxidation of its active center dithiol, to a disulfide, and catalyzes dithiol-disulfide exchange reactions. It forms together with thioredoxin reductase and NADPH a redox active system which donates electrons to a wide variety of different metabolic process.

Instead, among the four down-regulated genes in common, the most repressed gene was r*v2274A* (mean log2FoldChange: -4), coding for the antitoxin MazE8. *M. tuberculosis* has multiple toxin-antitoxin systems that are involved in regulating adaptive responses to stresses associated with the host environment and drug treatment. We speculate that, the repression of the antitoxin MazE8 could help the compound in its toxic effect against *M. tuberculosis* cells.

Taken together, these data agree with the analysis above in indicating that 11726172 could have a pleiotropic effect on *M. tuberculosis*, triggering general stress responses, perturbing metal homeostasis and cytoplasmic redox potential.

Gene functional annotation

Functional annotation analysis performed independently on the DEGs found in the first (n=165) and in the second condition (n=144) highlighted the presence of a core of enriched GO terms (FDR < 0.05), which include (i) response to copper ion (GO:0046688), (ii) cysteine biosynthetic process (GO0019344), (iii) response to cadmium ion (GO:0046686), (iv) protein disulfide oxidoreductase activity (GO:0015035) and (v) cysteine biosynthetic process from serine (GO:0006535). The only differences between the two experimental conditions are represented by the following GO terms: glycerol ether metabolic process (GO:0006662) and cell redox homeostasis (GO:0045454) which are found to be enriched (FDR: 0.05) in the first experiment, and defense response to virus (GO:0051607) which is enriched (FDR < 0.05) in the second experiment.

Taken together, these data suggest that 11726172 could have a pleiotropic effect on *M. tuberculosis bacteria*, triggering general stress responses. In particular, it seems to perturbs metal homeostasis and cytoplasmic redox potential.

Gene expression validation using Real-Time PCR

To validate the expression profile obtained by RNA-Seq analysis, we performed an experimental validation by qPCR. To this aim, in both conditions of treatment with 11726172, three differentially expressed genes (two induced genes, *cyp135A1*, *trxC*, and one repressed, *mazE8*) were selected amongst the most up- or down-regulated genes and their level of expression was analysed.

To perform qPCR experiments, RNA was extracted from *M. tuberculosis H37Rv* cultures treated with 11726172 10- or 30-fold MIC. Not treated cultures were included in the experiment as control. Expression profiles of the selected genes obtained by qPCR were consistent with the patterns of expression revealed by the RNA-Seq. The results were considered as technical validation of the differential expression gene analysis.

CONCLUSIONS AND FUTURE DIRECTIONS

In this study, I explored the biological mechanisms responsible for the antitubercular activity of the molecule 11726172. Upon treatment with varying concentrations of this drug (2.5 µg/ml and 7.5 µg/ml), I observed many genes being differentially expressed. In particular, from the comparison between the 2 conditions, I found 92 DEGs in common, which were subjected to a functional annotation analysis, revealing several enriched pathways, most of them being involved in metals-metabolism and redox homeostasis. Thus, we speculate that alterations in these processes might represent major ways through which this compound carries out its activity.

Based on these findings, at the Molecular Microbiology Laboratory (University of Pavia), experiments to determine the 11726172 MIC in presence of metal ions (Cu2+, Zn2+, Co2+, Ni2+) were performed to evaluate possible variation in MIC. No shifts in MIC were detected in presence of Zn2+, Co2+, Ni2+ ions, while we observed a shift in 11726172 MIC in the presence of increasing concentrations of Cu2+ (from 5 mM to 50 mM) ions. The lowest 11726172 MIC value (0.125 mg/ml) was observed in presence of the highest concentration of Cu2+. It seems that Copper ions gradually reduce MIC values for 11726172. To validate that the activity of 11726172 is affected by the presence of Cu2+ ions, the analysis of metal content in *M. tuberculosis H37Rv* cell lysates after treatment with metals or with both 11726172 and metal ions was performed. Unfortunately, the measurement of concentrations of metals in *M. tuberculosis H37Rv* cell extracts could not demonstrate a direct effect of the 11726172 compound on metals accumulation. Thus, we could not define the 11726172 mechanism of action, in fact this compound does not exert its activity by perturbing metal homeostasis, as hypothesized, even if we cannot exclude an indirect effect on copper-dependent enzymes.

Variation in gene expression is an invaluable tool for a better understanding of the drug effect on *M. tuberculosis* cells, and it could be integrated by metabolomics approaches that measure molecules produced during metabolism, reflecting the biochemical activity of bacteria. So, analysing the metabolome profiling after treatment with antimicrobial compounds completes the knowledge on bacterial physiological response to antibiotic and drug mechanism (Zampieri et al., 2018). Moreover, mycobacterial metabolism changes during the course of the infection, in particular when the bacilli enter the dormant state, in which metabolic activity is reduced, leading to antibiotic resistance (Gengenbacher & Kaufmann, 2012). Thus, the next challenge will be the study of the metabolome profile changes induced by 11726172 exposure, for a better understanding of its mechanism of action, in collaboration with Dr. Katarina Mikusová (Department of Biochemistry, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia).

Bacillus anthracis

INTRODUCTION

A common outermost cell envelope component in prokaryotes is the Surface layer (S-layer). The S-layer is almost universal in Archaea, while it is sparsely present in species belonging to bacterial lineages (Sára & Sleytr, 2000). Structurally speaking, S-layers are 2D paracrystalline (glyco-) protein monolayers formed by S-layer proteins (SLPs) that, once released at the cell surface, self-assemble into a 2D lattice with defined symmetry and non-covalent anchoring to the cell envelope (Sára & Sleytr, 2000) (Mignot et al., 2002). In Archaea, where the cell envelope in most cases consists of a cytoplasmic membrane surrounded by an S-layer, S-layers function as cell shape determinants and supporting structures (Mignot et al., 2002) (Albers & Meyer, 2011). In Bacteria on the other hand, where cell envelopes have evolved as complex multi-layered structures, the S-layer has been reported to carry out a broad variety of functions with marked differences across species (Sára & Sleytr, 2000). Bacterial S-layers have been proposed to play a role as selective barriers for molecular trafficking of nutrients and metabolites, to confer protection from predators and phages, to be involved in cell adhesion as well as to act as virulence factors (Sa & Sleytr, 2000) (Mignot et al., 2002) (Gerbino et al., 2015). Expression of SLPs can account for up to 30% of total protein synthesis (Mignot et al., 2002) (Awram & Smit, 1998) and together with secretion makes the whole process highly expensive from an energetic point of view. Its conservation in several lineages, despite such a high energetic load, clearly indicates the physiological importance of the S-layer and it suggests at the same time the existence of multi-level control mechanisms regulating S-layer synthesis. In this context, an interesting model organism for the study of the S-layer is *Bacillus anthracis*.

Function of the *Bacillus anthracis* S-Layer

Bacillus anthracis is a gram-positive endospore-forming bacterium belonging to the *Bacillus cereus sensu lato* group, renown as the etiological agent of anthrax (Cerquetti et al., 2000) (Chami et al., 1997) (Chateau et al., 2018) and a CDC Category A bioterrorism agent. Exposure to anthrax spores can occur by three routes: cutaneous (through the skin), gastrointestinal (by ingestion), and pulmonary (inhalation). Depending on the exposure, mortality rates can greatly vary and are approximately 20% for cutaneous anthrax, 25 - 75% for gastrointestinal anthrax and 80% or higher for inhalation anthrax.

From the genomic standpoint, *B. anthracis* has a single circular chromosome and two circular, extrachromosomal, double-stranded DNA plasmids, pX01 and pX02, which are responsible for the onset of the disease. Plasmid pX01 harbours three genes (*pag*, *cya* and *lef*) encoding for the anthrax proteins "lethal toxin" and the "edema toxin" while plasmid p0X2 carries an operon named "*capBCADE*" which synthesizes a poly-γ-D-glutamic acid (polyglutamate) capsule, required to evade the host immune system (Moayeri et al., 2015).

In this pathogen, the S-layer has a dual composition which reflects the trend of its life cycle. It is composed of two S-layer proteins (SLPs): the surface array protein (Sap), encoded by the gene *sap*, which is the main S-layer component during the bacterium's exponential phase of growth, and the extractable antigen 1 (EA1), encoded by the gene *eag*, which is the main molecule forming the stationary phase S-layer. Interestingly, the S-layer switch must occur also during systemic infection as both proteins have been found to be immunogenic during human anthrax infection (Chitlaru et al., 2007). Yet it is unclear why *B. anthracis* performs this energetically expensive S-layer remodeling during its life cycle and infection, emphasizing the need to understand the functions of these two S-layers, the environmental and host triggers that induce it as well as the fine regulation that leads to this differential expression to gain a better understanding of S-layer biology. Different studies in which *B. anthracis* strains were either treated with Sap inhibiting nanobodies (Fioravanti et al., 2019) or depleted of Sap or EA1 genes (Kern et al., 2012), revealed that the disruption/absence of the S-layer attenuates bacterium growth and the pathology of anthrax in vivo, highlighting the importance of this structure as a potential therapeutic candidate. *B. anthracis* strains lacking one of the two SLPs, as well as a double S-layer knockout (lacking both proteins) were described viable in vitro but with cell envelope defects, alterations in growth kinetic and reduced capability to face cell envelope stresses, therefore indicating some sort of adaptation.

On the other hand, the Nbs-induced phenotype was more striking than the genetic knockout as Nbs-treated cells would first wrinkle and then collapse, while in the genetic knock out no collapsed cells were found. Such evidence suggests that cells undergoing an acute loss of S-layer cannot adapt by switching to an EA1 S-layer to rescue such defects.

However, a detailed understanding of how the disruption of the S-layer may alter the growth and virulence of *B. anthracis* is still missing and therefore, worthy of being investigated.

AIM OF THE WORK

In the light of these new evidence and methodologies, for the first time we had the chance to perform a comprehensive and detailed study on how the S-layer disruption and/or absence may alter the fitness of this pathogen and reduce its virulence.

To understand the function of the two *B. anthracis* SLPs and how *B.anthracis* is able to counteract the loss of S-layers, I joined the Structural and Molecular Microbiology Laboratory, at the Vrije Universiteit Brussel, headed by Prof. Remaut (VUB-VIB, Belgium) working in the S-layer team lead by Prof. Antonella Fioravanti. There, I performed genomic comparative analyses between wild type (wt) *B. anthracis 7702* (pX01+, pX02-) and *9131* (pX01-, pX02-) and their derivative S-layer single or double knockouts (ko) (7*702 Δsap-eag* and *9131 Δeag*), respectively. In particular, I explored the downstream effects resulting from the genetic deletion of EA1 or/and Sap gene by searching for variations in terms of gene content and for SNPs falling in coding sequences that might have a direct effect on the activity of the corresponding protein.

MATERIALS AND METHODS

DNA extraction

Avirulent *B. anthracis 7702* (pXO1+, pXO2–) and *9131* (pXO1-, pXO2–) are derivative of the Sterne strain, discovered in the 1930s, which has naturally lost its pXO2 plasmid, and consequently its ability to produce a capsule. DNA from *B. anthracis* Sterne *7702* and *9131* and their strain variants, *7702 Δsap-eag* and *9131 Δeag,* was extracted using the QIAGEN Genomic-tips 100/G. The amount and purity of DNA were determined spectrophotometrically and integrity of DNA was assessed on a 1% agarose gel.

DNA sequencing

Whole gDNA obtained for each strain (wt *7702*, wt *9131*, *7702 Δsap-eag* and *9131 Δeag*) was sequenced using both Illumina MiSeq and Oxford Nanopore GridION platforms. MiSeq sequencing was performed using the kit MiSeq v2 300 cycles (10 pM + 1.65% PhiX v3), paired end (151-10-10-151). In parallel, Nanopore sequencing was performed using a single FLO-MIN106 flow cell on a GridION instrument (Oxford Nanopore Technologies). The sequencing library was prepared using the Oxford Nanopore’s ligation sequencing kit (SQK-RBK004). Base calling of gridION raw signals was done using Guppy (v.3.3.0; ONT).

Genome assembly and bioinformatic analyses

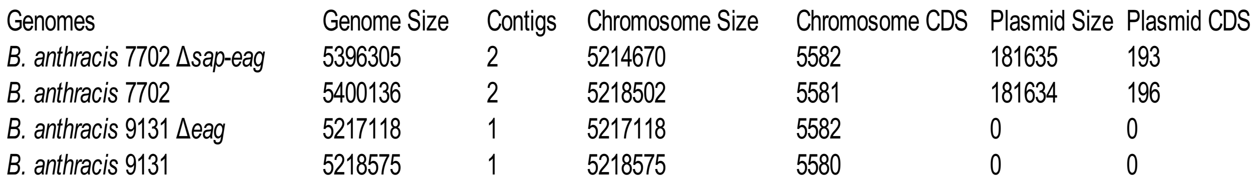
First, Illumina and Nanopore reads were quality checked with the tools FASTQC (*FastQC*, 2015) and Filtlong (https://github.com/rrwick/Filtlong), respectively, and then assembled together using the tool Unicycler (Wick et al., 2017). Once assembled, I utilized the tool Prokka (Seemann, 2014) for the prediction of open reading frames and their relative functions. This predicted proteins were then given as input to Orthofinder (Emms & Kelly, 2018) to infer orthologous gene groups (orthogroups) and their multiple-sequence alignments.

Lastly, for each group of comparison (*B. anthracis* 7702 and *B. anthracis* 7702 Δ*sap-eag*, *B. anthracis* 9131 and *B. anthracis 9131* Δ*eag*), I used the tool MAUVE (Darling et al., 2010) to align ko strains against their wt references. Individual alignments were merged using a Python script to obtain a multi-alignment file, allowing the extraction of coreSNPs (defined as variations of a single nucleotide flanked on each side by two nucleotides conserved in all the genomes analysed).

RESULTS AND DISCUSSION

Genomic response to S-layer deletion

Wild type *B. anthracis 7702* (pX01+, pX02-) and *9131* (pX01-, pX02-) and their derivative S-layer single or double knockouts (*9131* Δ*eag* and *7702* Δ*sap-eag*) genomes resulted in assemblies of either 1 or 2 contigs, corresponding exactly to the number of chromosomes and plasmids possessed. Chromosome size of all 4 assemblies was extremely similar, with a standard deviation of only 1825 base-pairs. As expected, plasmid pX01 was found only in wt *B. anthracis* 7702 and ko *B. anthracis* 7702 Δ*sap-eag*, with only 1 base-pair difference between the two assemblies. Considering the number of open reading frames annotated, as we can see from the table below (Table 1), Prokka identified an almost identical amount of chromosomal (5582, 5581, 5582 and 5580) and plasmidic proteins (193 and 196) among the genomes.



Genomes	Genome Size	Contigs	Chromosome Size	Chromosome CDS	Plasmid Size	Plasmid CDS
B. anthracis 7702 Δ*sap-eag*	5396305	2	5214670	5582	181635	193
B. anthracis 7702	5400136	2	5218502	5581	181634	196
B. anthracis 9131 Δ*eag*	5217118	1	5217118	5582	0	0
B. anthracis 9131	5218575	1	5218575	5580	0	0

Table 1: *B. anthracis* genomes statistics

From these first exploratory analyses, I noticed an extreme similarity among the 4 genomes, both in terms of genome/plasmid size and number of proteins. This is not surprising as *B. anthracis* is highly monomorphic. It shows little genetic variation as it primarily exists in the environment as a highly stable, dormant spore in the soil (Kolstø et al., 2009).

After protein annotation, the knockout strains were deeply characterized and compared to their relative wild types to detect changes (at gene and/or nucleotide level, see below) that might be part of the adaptation process of the knock-out.

Gene content analysis

To compare the genomic content between the wildtypes and the mutants, I partitioned all the proteins coded for by our genomes into ortholog groups, by using the tool “Orthofinder". Orthofinder identified 5156 unique orthogroups: 5150 constituting the core-genome (present in all 4 genomes) and 6 constituting the accessory-genome (variable among the samples). In the comparison between *B. anthracis 7702* and *B. anthracis 7702 Δsap-eag* and*,* between *B. anthracis 9131* and *B. anthracis 9131Δeag,* the only difference is represented by the presence of the gene b*cla*, a major component of the exosporium. This gene encodes a protein containing an internal collagen-like region (CLR) of GXX repeats which includes a large proportion of GPT triplets. This region is extremely polymorphic and can be characterized by a variable number of GXX repeats, which leads to variation in the exosporium filament length (Sylvestre et al., 2003). In our analysis, *B. anthracis 9131* and *B. anthracis 7702 Δsap-eag* possess a longer *bcla* gene than the one found in *B. anthracis 7702* and *B. anthracis 9131 Δeag.* Explaining such a difference can be challenging and various reasons can be cited. Firstly, repetitive stretches of DNA can introduce errors during the sequencing and as a consequence, during the assembly phase (misassemblies) (Tørresen et al., 2019). Secondly, *B. anthracis 7702* and *B. anthracis 9131* may have accumulated distinctive genomic differences across their cell replication cycles which may have led to such a difference, Lastly, even though it is not known yet, differences in *bcla* gene length might be a sort of response to the disruption of the S-layers which could help these mutated cells to adapt to these new conditions.

SNPs analysis

The goal of this analysis was to detect the presence of mutations falling in interesting regions of our genomes. Indeed, SNPs may change a codon in a synonymous or non-synonymous way, or they may occur in noncoding regions, potentially altering the transcription rate of the downstream gene and therefore its abundance. By comparing the wt with the ko strains, I evaluated whether *sap* and/or *eag* loss led to changes at nucleotide/amino acid level.

For each group of comparison, the ko is aligned against its relative reference and coreSNPs are extracted and further explored with particular focus on non-synonymous ones. In *B. anthracis 7702 Δsap-eag*, I found 7 synonymous and 7 non-synonymous SNPs, these last ones falling into 4 proteins coding for a polyamine antiporter (n=1), HAMP domain containing histidine kinase (n=1), UDP-N-acetylglucosamine 2-epimerase (n=1) and a transposase (n=4). Instead, in *B. anthracis 9131 Δeag*, I found 1 synonymous and 13 non-synonymous SNPs, these last ones falling into 4 proteins coding for an aspartokinase (n=7), dephospho-CoA kinase (n=1), transposase (n=4) and (n=1).

Unfortunately, at the moment, none of these genes is fully annotated, and little is known about their function and therefore, we cannot yet hypothesize their involvement in the adaptation process of the ko strains. However, it is interesting to notice that two proteins, the UDP-N-acetylglucosamine 2 epimerase and the transposase, are found to be mutated in both the experimental comparisons. In particular, the UDP-N-acetylglucosamine 2 epimerase is an enzyme involved in the synthesis of the capsule precursor UDP-ManNAc, a component of the pyruvylated secondary cell wall polysaccharide (SCWP), which is linked by the S-layers proteins Sap and EA1 and S-layer-associated (BSL) proteins via the S-layer homology (SLH) domain. Thus, a proper interaction between the SCWP and the S-layers is fundamental for subsequent cell division events in vegetativ*e B. anthracis* cells (Y. T. Wang et al., 2014)

Hence, because of this known interaction, mutations of the UDP-N-acetylglucosamine 2 epimerase gene might be one of the forms of adaptation adopted by *B. anthracis 7702* Δ*sap-eag* and *B. anthracis 9131* Δ*eag* strains.

CONCLUSIONS AND FUTURE DIRECTIONS

Previous studies found that acute *B. anthracis* Sap S-layer disruption by nanobodies results in severe cell surface defects, growth attenuation and clearance of infection. S-layer knockouts also present peculiarity in cell growth, mechanics and stress responses if compared to their parental strain, but how the S-layer is integrated in the cell physiology and how cells react to its disruption is still unknown.

Our study stands out as the first step towards the identification of all those genetic features and metabolic pathways involved in the adaptation process following the loss of the S-layers in *Bacillus anthracis*.

From this investigation, I did not find any difference in gene content but several mutations in proteins that might allow the survival of ko strains. However, Vogler, Amy J et al. (2002) (Vogler et al., 2002) estimated the in-vitro mutation rate of *B. anthracis* to be 5.2×10^{-10} mutations/bp/generation. This implies that limited expansions of colonies in vitro would hardly result in the accumulation of mutations. Plus, due to the lengthy spore phase of its life cycle, *B. anthracis* has evolved very slowly and has a very narrow pan-genome, with an estimated core/pan-genome ratio of 0.99 (M. et al., 2016). This leads to one possible scenario where S-layer single or double knockouts might be ‘rescued’ not by mutations altering protein activities, but through changes at transcriptional level. Thus, as a next step, we argue here that deep transcriptome comparisons (RNA-Seq) may allow to determine the differences in ko lines with more precision than genome sequencing only.

Streptococcus agalactiae

INTRODUCTION

Streptococcus agalactiae (also known as group *B* streptococcus or GBS) is a gram-positive harmless commensal bacterium, normal constituent of the human microbiota and colonizer of the gastrointestinal and genitourinary tract of up to 30% of healthy human adults. Yet, GBS has highly invasive potential, being able to cause infections in people with compromised immune systems, in the elderly and in newborns (Raabe & Shane, 2019). In the latter, GBS is a leading cause of morbidity and mortality, responsible for a plethora of clinical syndromes, including sepsis, pneumonia, bacteremia and meningitis.

Depending on the age of the infant at the time of disease manifestation, GBS disease in newborns can be classified as early-onset disease (EOD) or late-onset disease (LOD). In EOD, the disease occurs within the first seven days of life while LOD can present in infants up to several months in age (7–90 days) (Rajagopal, 2009). Usually, a neonate acquires GBS vertically (vertical transmission), during passage through the birth canal or shortly thereafter and it becomes a normal constituent of the child's microbiome. This process is generally safe, as 99% of newborns do not develop invasive conditions. However, when GBS does not adapt or it is acquired later, from environmental sources (horizontal transmission), it can cause severe invasive disease and tissue damage, causing mortality in 10% of cases and severe long-term consequences in 30% of survivors (Landwehr-Kenzel & Henneke, 2014). GBS is also an emerging pathogen of adult humans. People with diabetes, cancer or gastrointestinal infections are particularly at risk, with clinical manifestations which include skin, soft tissue and urinary tract infections, bacteremia, pneumonia, arthritis and endocarditis (Furfaro et al., 2018).

In addition, the severity of GBS syndrome is furtherly influenced by the genetic background of the infecting strains. GBS are currently divided into ten serotypes (Ia, Ib, II, III, IV, V, VI, VII, VIII, and IX) which harbour a variable endowment of virulence and resistance factors.

In general, *Streptococcus agalactiae* is susceptible to most beta-lactam antibiotics, including ampicillin, first-, second-, and third-generation cephalosporins, and carbapenems. These drugs are the mainstay of treatment for GBS invasive disease, for both infants and adults but, although administering antibiotics during childbirth to women with GBS has helped reducing the incidence of vertically acquired infections, this type of strategy did not help reduce the incidence of horizontally transmitted infections after delivery (Raabe & Shane, 2019).

The role of codY in survival

During infection, the ability of GBS to invade different host niches (e.g., vaginal epithelium in pregnant mothers, blood, brain and lungs in the newborn) reflects its capacity to adapt to various environmental conditions. This versatility is allowed by the activity of several transcriptional regulators which, in response to environmental signals, control the expression of proteins involved in nutrient acquisition, adhesion, and immune evasion. In particular, the capacity to rapidly adapt to fluctuations in nutrient source and availability, contribute to the pathogenic potential of *Streptococcus agalactiae*. Such adaptation is finely tuned in response to multiple internal and external signals and, in this context, the global regulator *codY* plays a fundamental role.

Little is known about the complete role of *codY* in *Streptococcus agalactiae* but in model organisms such as *Bacillus anthracis*, *Clostridium difficile*, *Enterococcus faecalis, Staphylococcus aureus* or *Listeria monocytogenes,* this protein has been thoroughly investigated.

The gene *codY* encodes for a GTP-sensing transcriptional pleiotropic repressor, highly conserved across many Gram-positive bacteria. This transcriptional factor acts as a nutritional sensing molecular system, controlling the expression of more than a hundred genes, that are typically repressed during rapid growth and induced when cells experience nutrient deprivation (Sonenshein, 2005).

During the post-exponential growth phase, when nutrients such as GTP and BCAAs become limiting, *codY* is repressed, allowing the expression of a myriad of genes involved in adaptation to starvation, throughout mechanisms which regulate, for instance, carbon metabolism, iron uptake and biosynthesis of branched-chain amino acids (Feng et al., 2016). This gene is also an important modulator of virulence. In response to nutrients deprivation, *codY* protein regulates cellular motility, the expression of virulence factors and biofilm formation (Stenz et al., 2011). For instance, It has been seen that *Staphylococcus aureus codY*-null strains overexpress several virulence genes (e.g proteases, leukocidins and hemolysins) and are characterized by higher hemolytic activities, produce more polysaccharide intercellular adhesin, and form more robust biofilms (Rivera et al., 2012). On the contrary, in *Bacillus anthracis* the disruption of c*odY* led to attenuated virulence of a wild-type *B. anthracis* strain in a mouse model of infection (Château et al., 2011).

AIM OF THE WORK

Nowadays, as no effective vaccine is available yet, new diagnostic and therapeutic tools are urgently required to prevent and treat GBS associated diseases. In this context, due to its involvement in the regulation of global metabolism and virulence, *codY* represents a perfect candidate for further studies and future therapeutic applications for *Streptococcus agalactiae* infections.

Thus, as part of the collaboration with The Genetics and Microbiology Laboratory of the University of Pavia, I investigated the role of c*odY* in *Streptococcus agalactiae* CC17 BM110 cells (serotype III), into which, a marker-free, in frame deletion of *codY* was introduced. Thus, a differential expression analysis (RNA-Seq) was performed to study the transcriptional response of *Streptococcus agalactiae BM110* (Accession: NZ_LT714196.1) upon deletion of the gene codY (*ΔcodY*).

MATERIALS AND METHODS

Bacterial culture

All GBS strains (wild-type and Δ*codY*) are derivatives of BM110 and were grown in THY medium (Todd Hewitt's medium supplemented with 5 g/liter of yeast extract) or in chemically defined medium (CDM) at 37°C, 5% CO 2 , steady state. *E. coli* strains were cultured in Luria Bertani (LB) broth at 37°C. Antibiotics were used at the appropriate concentrations. For E. coli: kanamycin 50 µg/ml; erythromycin 150 µg/ml. For GBS: kanamycin 1 mg/ml; erythromycin 10 µg/ml.

Strains construction

To prepare the pG1-°*codY* vector used to create BM110 Δ*codY,* the BM110 genomic regions located upstream and downstream of *codY* were amplified by using pG1_codYUpF + BM_codYFusR and BM_codYFusF + pG1_BM_codYDwR primers, respectively. The temperature-sensitive pG1 plasmid was amplified with pG1R and pG1F oligonucleotides. All three fragments were fused by Gibson assembly using the NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs) and the reaction product was electroporated into *E. coli* XL1 blue. The obtained pG1-°*codY* plasmid was verified by sequencing and then electroporated in GBS. Transformants were selected at 30°C on TH plates supplemented with erythromycin. Plasmid integration and excision were performed as previously described (Biswas et al., 1993). The resulting in-frame deletion of *codY* on genomic DNA was verified by sequencing using primers COH1_1525FUp and COH1_1527RDw.

RNA extraction

GBS total RNA from cells collected at mid-exponential phase of growth was extracted using the Quick-RNA Fungal/Bacterial Miniprep Kit (Zymo Research) following manufacturer's instructions. Reverse transcription was performed in a single step using the iTaq Universal SYBR Green One-Step Kit (Bio-Rad). The reactions were performed in 20 µl volumes using 4 ng of DNAse I treated RNA, according to the manufacturer's protocol. RNA was retro transcribed at 50°C for 10 min and the reaction mixture was then incubated at 95°C for 1 min followed by 35 amplification cycles with 95°C for 10 sec and 55°C for 30 sec.

Transcriptome and bioinformatics analysis

Raw reads were quality checked using FASTQC (*FastQC*, 2015) and processed by Trimmomatic (Bolger et al., 2014) to trim the adaptor sequences and remove low-quality sequences. The remaining clean reads were mapped onto the reference genome *Streptococcus agalactiae BM110* (Accession: NZ_LT714196.1) using Bowtie2 (Langmead et al., 2009). To quantify the known transcripts, the alignment results were input into featureCounts (Liao et al., 2014). Lastly, the R package DESeq2 (Love et al., 2014) was used to test for differential expression. We defined genes as differentially expressed using the following criteria: log2FoldChange >= | 2 | and FDR < 0.05.

Finally, prediction of orthologous groups was performed using COGnitor (Tatusov et al., 2000).

RESULTS AND DISCUSSION

Transcripts quantification and exploratory analyses

To investigate the global gene expression pattern of *Streptococcus agalactiae* in response to the deletion of the gene *codY*, I performed genome-wide expression analysis using RNA-Seq. About 190 million pair-end reads were generated from the 2 biological groups, each with 4 biological replicates.

High-quality trimmed reads (min 22.1, max 25.5 milions) were mapped onto the reference genome *Streptococcus agalactiae BM110* (Accession: NZ_LT714196.1), resulting in a high percentage of mapped reads (min 93.1%, max 98.1%) which were later counted (around 21.2 milions reads per sample) and summarized at CDS level. (Table 1).

Samples	Raw_Reads	Mapped_%	Counted_Reads	Counted_Reads_%
421753_Control	23379794	96.16%	21149524	90.5
421754_ΔcodY	22161327	93.10%	19981370	90.2
421755_Control	23884015	94.42%	21276275	89.1
421756_ΔcodY	24541865	96.92%	21818789	88.9
421757_Control	23214327	96.25%	20473321	88.2
421758_ΔcodY	24346373	97.53%	21629412	88.8
421759_Control	25525882	96.80%	23048390	90.3
421760_ΔcodY	22983742	98.10%	20813160	90.6

Table 1: The RNA-Seq data for the 8 samples

Before starting with the differential expression analysis, I assessed the overall similarity between samples, in order to ensure that our replicates clustered together according to the experimental group they belong to. Using a regularized log transform (rlog) of the normalized counts, I performed a hierarchical clustering method in order to identify any sample outliers, which may need to be explored further to determine whether they need to be removed prior to DE analysis.

The hierarchical clustering heatmap shown below (Fig 1) displays the correlation of gene expression for all pairwise combinations of samples in the dataset and, as we can see, samples more similar to each other cluster together as a block for each experimental group. These results confirm the solidity of the project's experimental design, as the 4 replicates of the “control group” cluster together as well as the 4 replicates missing the *codY* gene.

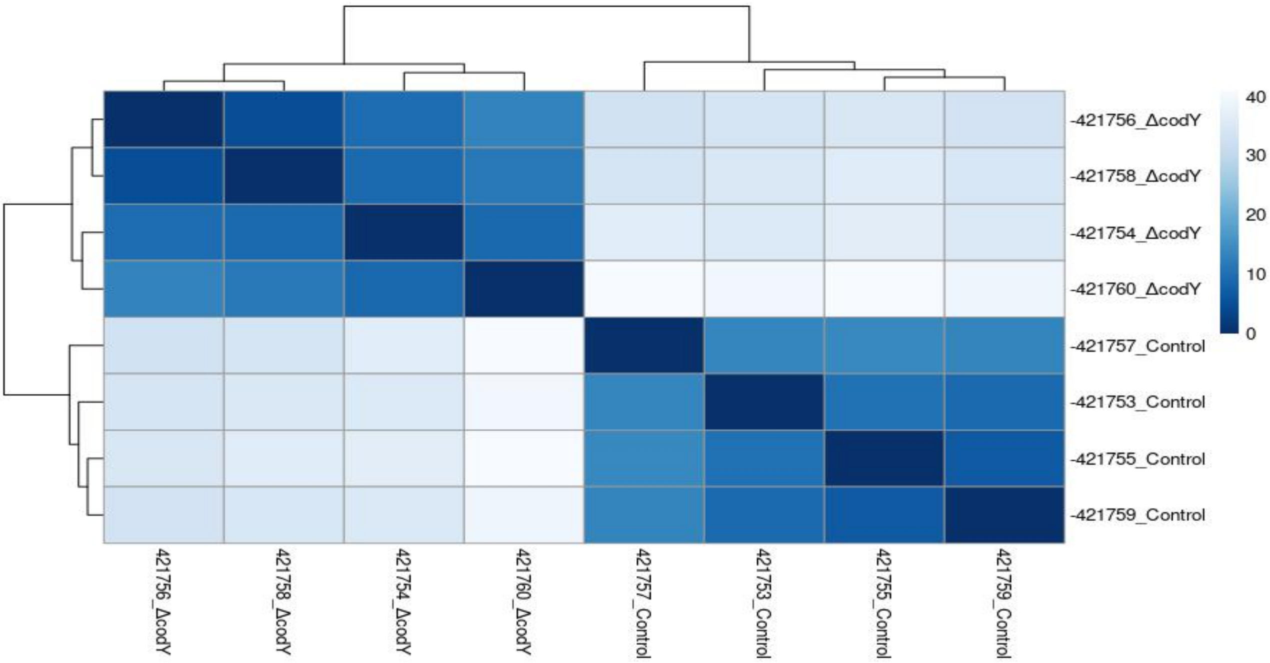


Fig 1: Hierarchical clustering plot of samples in control and mutated (Δ*codY)* bacteria

Differential gene expression analysis

Out of 2128 coding DNA sequences, 99 (4.6%) genes were found to be differentially expressed in *Streptococcus agalactiae BM110* Δ*codY*. Of these, 95 genes are annotated as CDS and 4 as pseudo-genes.

As it is possible to observe from the Volcano plot below (Fig 2), 94 genes were found to be up-regulated (2 – 8.4 fold) and 5 genes down-regulated (2.1 – 5.8 fold). One piece of evidence confirming the correctness of the analysis is represented by the proper expression of *codY* gene (internal control). In fact, it appears to be expressed about 6 times more in the controls than in mutant cells.

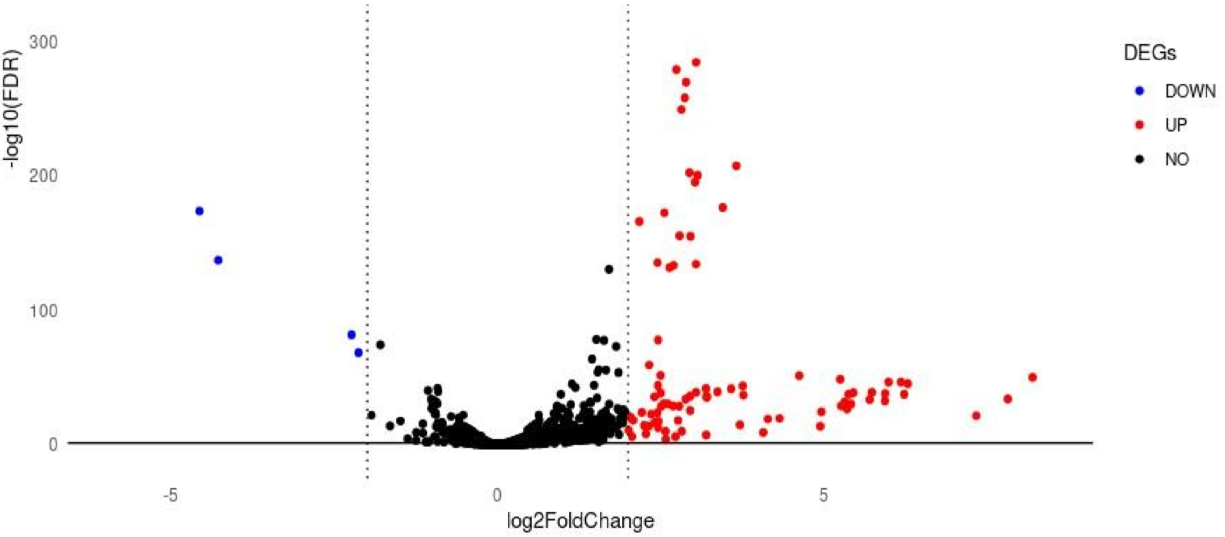


Fig 2: Volcano plot of DEGs between control and mutated *S. agalactiae* samples

Up and down regulated DEGs

Among the five down-regulated genes (besides the positive control *codY),* I report the gene *cfb* (log2FoldChange: -4.5), encoding for the cAMP factor pore-forming toxin and *ahpC* (log2FoldChange: -2.1) and *ahpF* (log2FoldChange: -2.2)*,* which are two alkyl-hydroperoxide reductase subunits.

Instead, among the up-regulated gene, I report the presence of genes encoding for branched chain amino acids transporters (*braB*, *brnQ*, *livK-G* operon), the peptide permease *oppA1-F*, cell-wall anchored adhesins and serine peptidases carrying the LPxTG motif, as well as proteins involved in DNA replication, recombination and repair. Interestingly, also genes of the *cas* operon and the virulence factor *srr2* were among the most up-regulated in the Δ*codY* mutant.

Gene functional annotation

Based on the previously obtained list of differentially expressed genes, the next step is represented by the enrichment analysis, which aims to identify particularly enriched biological pathways.

Unfortunately, a problem encountered at this stage was the lack of annotation for many of our genes. In fact, *Streptococcus agalactiae BM110* (Accession: NZ_LT714196.1) is not the reference strain for the *Streptococcus agalactiae* species and consequently, due to the many absent data, it was not possible to perform the classic enrichment analysis.

To overcome this problem, it was therefore decided to conduct an orthology analysis in order to identify the most represented Cluster of Orthologous Groups (COGs) among our differentially expressed genes. Considering the down-regulated DEGs, we found 4 COGs, among which "defense mechanism" was the most populated. Instead, among the up-regulated DEGs we identified 17 COGs, of which "amino-acid transport and metabolism", "cell wall/membrane/envelope biogenesis", "mobilome: prophages and transposons" and "replication, recombination and repair" being the most abundant (Fig 3).

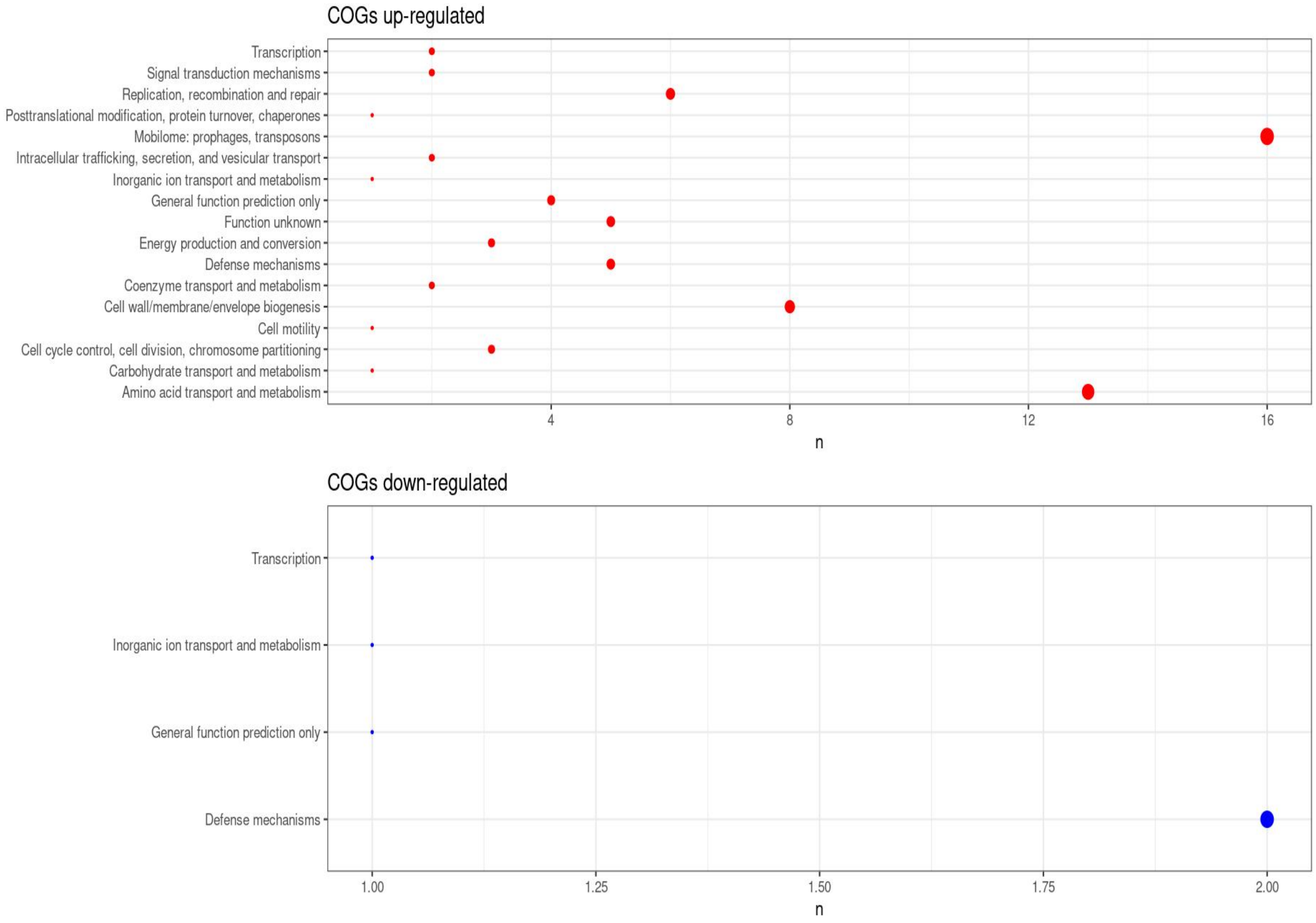


Fig 3: COGs enrichment analysis of DEGs after deletion of *codY*

CONCLUSIONS AND FUTURE DIRECTIONS

The evidence collected so far confirms that c*odY* is a global regulator of gene expression. RNA-Seq analysis found 99 DEGs, of which 94 being up-regulated and 5 down-regulated in the Δ*codY* mutant in comparison to wild-type bacteria, supporting a role for *codY* mainly as a repressor of gene expression.

Functional annotation based on COGs allowed the identification of the most represented pathways under *codY* control which include defense mechanisms, replication, recombination and repair, cell-wall biosynthesis and amino acid transport and metabolism. In particular, amino acid transport and metabolism is, as expected, particularly enriched, as *codY* is a major regulator of transcription in response to starvation and nutrients deprivation. As a consequence, genes such as *livK,* encoding for a BCAA transporter*,* and *braB*, encoding for a BCAA permease, result to be highly up-regulated.

To further explore the role of *codY* in controlling GBS virulence, the Genetics and Microbiology Laboratory of the University of Pavia in collaboration with the University of Messina (Prof.ssa Beninati) performed various in-vitro and in-vivo experiments. Deletion of *codY* did not affect cell growth but strongly impaired the ability of *S. agalactiae* to adhere to epithelial cells monolayers, reducing up to 50% (p-value <0.05) adherence to epithelial cell lines. In newborn mice infected with the BM110*codY* mutant, higher survival rates were detected when compared to mice infected with wild-type BM110 cells. Moreover, the number of bacteria detected in blood, brain, and liver revealed that the *codY* deletion mutant has a reduced ability to persist in blood, and to colonize the organs of infected mice. Finally, they also observed that deletion of *codY* determined an increased ability to form thicker and more compact biofilm compared to the wild-type strain.

All together these last findings highlight the importance of *codY in* virulence. *CodY* is involved in the regulation of the adhesion properties of *S. agalactiae*, therefore suggesting a possible role for this regulator in the control of the initial steps of host colonization. Plus, these data support the evidence that *codY* is required for GBS virulence, influencing the ability to form biofilms and to persist in the host, once again remarking the importance of this protein as a possible future target for therapeutic applications.

Staphylococcus aureus

INTRODUCTION

Staphylococcus aureus is a gram-positive bacterium, distributed worldwide, and is a common member of the microbiota of the body, also often found in the upper respiratory tract and on the skin. Although it usually acts as commensal, sometimes it can also become an opportunistic pathogen, thus infecting hosts with a weakened immune system or those that have undergone surgery procedures (Tong et al., 2015).

S. aureus can spread through skin-to-skin contact with colonized individuals and can cause infections both in the community and in the medical settings. Its entrance in the human body occurs through a breach in the skin or mucosa, affecting local or distant organs, generating life-threatening invasive infections such as bacteremia, pneumonia, endocarditis and osteomyelitis (Tong et al., 2015).

A key factor in the success of *S. aureus* both as a colonizer and as a pathogen is its ability to easily acquire new DNA by horizontal gene transfer (HGT) (e.g. transposons, bacteriophages, insertion sequences, pathogenicity islands, cassette..) and to spread clonally, although chromosomal mutations can also be important (Haaber et al., 2017). Using these mechanisms, over the last 70 years, strains of this bacterium have been able to become resistant to a wide spectrum of chemotherapeutic molecules, including methicillin and derivatives, widely used to combat *Staphylococcus* infections.

The first emergence of antibiotic resistant *S. aureus* broke out in the mid-1940s and was caused by strains resistant to penicillin: thanks to the production of a plasmid-encoded penicillinase BlaZ that hydrolyzes the beta-lactam ring of penicillin, essential for its antimicrobial activity. A second wave of resistance rose in the early 1960s; it was marked by the onset of strains resistant to methicillin, the so-called Methicillin-Resistant *Staphylococcus aureus* (MRSA). Methicillin resistance is due to the acquisition of a new gene, *mecA*, that codes for a novel penicillin-binding protein (PBP). This gene is contained in a mobile genetic element named Staphylococcal chromosomal cassette mec (SCCmec), that is chromosomally integrated and which makes the strain resistant to all beta-lactam antibiotics, including penicillins, cephalosporins, and carbapenems (Monaco et al., 2017) .

In response to this issue of great sanitary importance, a number of newly developed antibiotics that display good anti-MRSA activity have been released on the market. Vancomycin, linezolid and daptomycin, nowadays, represent the antibiotics of choice for particularly aggressive MRSA infections. However, although these antibiotics are useful and necessary their improper use has inevitably led to the origin of new resistance patterns. In the 1990s, for example, *S. aureus* strains showing increased resistance to vancomycin were discovered, known as vancomycin intermediate-resistant *S. aureus* (VISA) (MIC = 4-8 μg/mL). Later, in 2002, in U.S, strains completely resistant to vancomycin, known as vancomycin-resistant *S. aureus* (VRSA) (MIC ≥ 16 μg/mL) were also reported (Mcguinness et al., 2017). Since then, the total number of human VRSA and VISA infections has been increasing, but luckily, or thanks to a higher awareness, is still very limited. For example, no VRSA have ever been reported in Italy.

Until recently, MRSA strains were primarily limited to hospitals and other institutional settings. However, since the 1980s, more and more reports of community-acquired MRSA (CA-MRSA) infections have appeared, marking the beginning of a new epidemic era.

Since the very beginning, CA-MRSA have appeared to be different from those encoding methicillin resistance and acquired in hospitals (HA-MRSA). CA-MRSA infections tend to occur in individuals in the community, who are generally healthy (children and young adults) with unknown risk factors, generally causing minor skin infections. Additionally, CA and HA- MRSA strains are genetically and phenotypically distinct. CA typically bear SCCmec type IV or V while HA type I-III. Plus, CA strains result susceptible to a wider range of anti-staphylococcal antibiotics and they often produce the panton valentine leukocidin (PVL), a staphylococcal virulence factor that destroys the white blood cells (Vysakh & Jeya, 2013).

Even though considerable variations between geographical areas exist, in the clinical setting most *Staphylococcus aureus* infections are caused by strains belonging to a relatively restricted number of lineages. Indeed, in the hospital setting, where there is the implementation of strategies to control the spread and transmission of *S.aureus* strains, clones are relatively stable and mainly diversify by the accumulation of single nucleotide substitutions in the absence of frequent interstrain recombination (Feil et al., 2003).

On these bases, a deeper understanding of the genetic variability and evolution of *S. aureus* would be pivotal, because it would allow to support local and global health authorities in preventing the extensive dissemination of dangerous clones and to provide an accurate management of patients.

In this perspective, the real-time integration of whole genome sequencing (WGS) technologies in clinical practice could potentially provide a single platform for extracting all the information required to resolve outbreaks and identify emerging resistant lineages.

Furtherly, these technologies coupled with epidemiological surveillance methods would allow precise tracking and monitoring of strains, and assessment of their clinical significance, and could provide early warning of emerging pathogenic clones.

In order to obtain this goal of ‘real-time genomic surveillance’, we need to start with properly designed retrospective studies that can provide a temporal and spatial background describing the epidemiological landscape in different regions and hospitals.

AIM OF THE WORK

At the San Matteo Hospital (Pavia, Italy), there is a strictly implemented surveillance programme aimed to manage the onset and the diffusion of *S. aureus* infections. All patients positive for *S. aureus* undergo antimicrobial susceptibility test in order to detect possible drug resistances and to assure susceptibility to drugs of choice for their particular infections. Therefore, all MIC values and related EUCAST interpretations are archived in a local hospital database for future references and investigations.

Leveraging this database as a starting point for my analyses, the aim of my Ph.D. project has been to provide an advanced characterization of all *S. aureus* infections reported at the San Matteo Hospital between June 2011 and April 2019, through the use of genomics, bioinformatics, and the statistical analysis of metadata information.

The work here presented is divided into two major sections:

1. Analysis of antibiotic resistance trends and selection of samples to be subjected to whole genome sequencing;
2. Genomic and epidemiological characterization of *S.aureus* genomes.

MATERIALS AND METHODS

Determination of antimicrobial susceptibility by minimal inhibitory concentration (MIC)

Staphylococcus aureus isolates from patients admitted to San Matteo Hospital (Pavia, Italy) are routinely investigated by BD Phoenix for the detection of antimicrobial resistances and determination of minimum inhibitory concentration (MIC) values. These results, combined with the EUCAST expert rules, provide an accurate categorization of the organism as Susceptible, Intermediate or Resistant to an antibiotic. For our study, I retrieved all *Staphylococcus aureus* MIC values collected between June 30th 2011 and December 31st 2019, encompassing 15 antibiotic compounds of clinical relevance: Benzylpenicillin, Teicoplanin, Clindamycin, Daptomycin, Co-Trimoxazole, Oxacillin, Erythromycin, Linezolid, Rifampicin, Vancomycin, Gentamicin, Fosfomycin, Cefoxitin, Ciprofloxacin, Tetracycline.

Unless otherwise stated, only the first isolate per patient was used for all the analyses.

Data polishing and cluster detection

All antibiograms with missing MICs, MICs classified as “Intermediate” or ambiguous MIC values were removed from the dataset. Then, the dataset was transformed into binary by converting "Resistant" and "Susceptible" values to "1" and "0", respectively. The detection of clusters of isolates with similar resistance binary profile was conducted in R.

Dataset dimensionality was first reduced with a technique called “Uniform Manifold Approximation and Projection” (UMAP) (McInnes et al., 2018), using the R package "uwot". Afterwards, on a such lower dimensional space, I applied a clustering algorithm called “Density-Based Spatial Clustering of Applications with Noise” (DBSCAN) (Ester et al., 1996), using the R package "dbscan".

DNA extraction and sequencing

DNA was extracted from 226 blood cultured isolates using Qiagen DNeasy Blood and Tissue kit (Qiagen) and quality and quantity was assessed with gel electrophoresis (TAE, 1.5% agarose).

Whole-genome sequencing was performed using paired-end (2×150 bp) sequencing on a NovaSeq instrument (Illumina). Reads were quality checked with “FastQC” (*FastQC*, 2015) and subsequently assembled into contigs by “SPAdes” (version 3.14.1) (Bankevich et al., 2012). Quality of the assemblies was assessed with the “assembly-stats” free software and statistical analysis summarized with the software R.

In silico Typing

Multilocus sequence typing (MLST) was performed by scanning the strains’ genome assemblies against the *S. aureus* MLST database (PubMLST), using an in-house python script.

Typing of the SCCmec element was done with the “SCCmecFinder” web-based tool (Ito et al., 2014) with the “min_coverage” 60 and “percentage identity” 90 options.

Statistical analyses

R software (v3.2.4) was used for all statistical analyses. Comparison of prevalences was performed using the chisq.test function, while the proportion test was assessed using the function prop.test. The significance level was set to 0.05.

Linear regression, with the function lm, was used for the analyses of temporal trends. The slope (beta value) of the linear regression lines is used as an indicator of the time course.

Unless otherwise stated, the R package ‘ggplot2’ was used for visualizations.

CDC classification

According to the Centers for Disease Control and Prevention (CDCs), an isolate is considered CA if: a positive wound culture was taken within 48 hours of hospital admission, and the patient did not have surgery, did not live in a long-term care facility, or undergo hemodialysis/peritoneal dialysis during the past year, and the patient did not undergo catheterization or insertion of indwelling percutaneous devices during present hospital admission. Instead, it is considered HA if a subsequently positive wound culture was taken after 48 h of hospital admission.

Detection of Selected Virulence Factors

The presence of virulence factors (n=32) was addressed with the tool “VirulenceFinder2.0” (Joensen et al., 2014), with the min_coverage 80 and percentage identity 90 options.

Phylogenetic analyses

First, we performed a global phylogenetic analysis including only our *S.aureus* samples. High-quality SNPs were called, and maximum likelihood phylogeny inferred using RaxML (Stamatakis, 2014).

Later, a SNP-based phylogeny was also generated for the major clonal complexes as follows. The novel genome sequences were added to a selected dataset of highly related *S. aureus* genomes extracted from the PATRIC database (Wattam et al., 2017). In detail, each genome was compared to all PATRIC genomes using Mash (Ondov et al., 2016) and the 50 best hits, representing the most similar genomes, were selected. All the obtained best hits lists were merged to obtain the final genomic dataset for each CC. CoreSNPs were extracted from the resulting dataset following a published method (Gaiarsa et al., 2015). Briefly, the Mauve software (Darling et al., 2010) was used to align each of the novel genomes and the similar PATRIC genomes to a well-characterized complete genome reference. Individual alignments were merged using a Python script to obtain a multi-alignment file, allowing to extract coreSNPs (defined as variations of a single nucleotide flanked on each side by two nucleotides conserved in all the genomes analysed). The coreSNPs alignment was used to perform a phylogenetic analysis using the software RaxML. The tree topology reliability was tested using 100 bootstrap replicates.

RESULTS AND DISCUSSION

At the San Matteo Hospital in Pavia all the results of antibiograms are routinely stored as MIC values in a local hospital database and their classification as “Resistant”, “Intermediate” or “Susceptible” is regulated by the most updated EUCAST guidelines. Between 2011 and 2019, at the Hospital, 5233 patients (inpatients: 3107, outpatients: 2126) resulted positive for *Staphylococcus aureus*, resulting in 7523 isolates, which were all tested against a panel of 15 antibiotics. The isolates were sampled in various hospital wards and obtained from different body sites including skin (31.2%), blood (16%), urine (4.6%), respiratory tract (13%), soft-tissues (30%), orthopedic-tissue (2.2%) and others (3%).

Overall, during the period 2011-2019 there was a rising numbers of hospital admissions, from 112 patients in 2011 to 791 in 2019, with an estimated number of 58 new more cases per year (beta: 58.42). This growth affected both inpatients and outpatients.

Antibiotic resistance

Out of 5233 samples, 1398 (26.7%) were found to be resistant to methicillin, with higher rates of resistance among isolates from soft-tissues (30%), skin (24%), blood (17.7%) and respiratory tract (17.4%). However, these values drastically change when comparing inpatients and outpatients. *S. aureus* strains infecting inpatients were more resistant to antimicrobials and clearly more invasive (bacteremia: 23.2%, respiratory-tract infections: 22.4%) while SA isolates from outpatients result mostly sampled from skin (38.8%) and soft-tissues (35.5%).

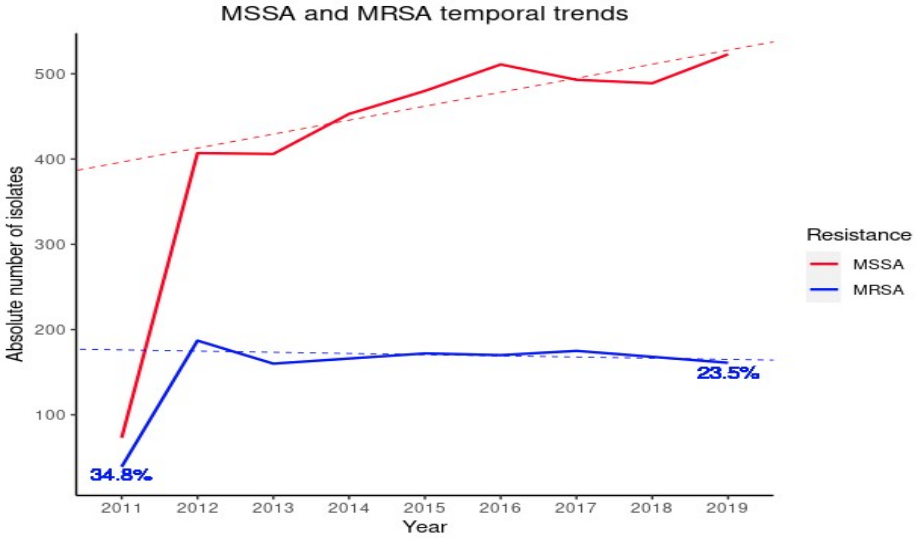


Fig 1: Absolute number of MR and MS isolates in the Hospital between 2011-2019

The percentage of MR *S. aureus* isolates varied significantly during the studied period, decreasing from 34.8% in 2011 to 23.5% in 2019. However, when dissecting this trend, we noticed that the absolute number of MRSA did not decline but rather remained almost stable, with a mean value of 170 (sd=8.6) cases per year. On the other hand, during the same period, the number of MSSA shifted from 73 to 523 (Fig 1), with an increase of 16.4 new cases per year (95% CI: 10.2-22.6, RSE: 20.54).

Comparing the complete antibiograms of MRSA and MSSA isolates, we observed a higher number of multidrug-resistant isolates in the former: resistance to penicillin (100%), ciprofloxacin (82.5%), erythromycin (57.8%), clindamycin (56.6%) and gentamicin (32.8%) were the most common. A lower number of MRSA isolates showed resistance to tetracycline (14%), trimethoprim with sulfamethoxazole (8.8%), fosfomycin (14.2%) and rifampicin (5.4%). MSSA samples showed generally much lower frequencies of resistance, with percentages spanning between 0 and 18.6% depending on the antibiotic, with the only exception for penicillin (78.9%). While this last value is much higher than any other resistance in MSSA, it is still significantly lower than the corresponding value in MRSA (100%). For teicoplanin, daptomycin and linezolid the resistance rate was below 1% in the entire dataset. No isolate resistant to vancomycin was detected.

By using linear regression models where time is used as a predictor, we detected opposing trends (Fig 2): in MRSA strains, resistance to clindamycin (64.1 to 53.4%, beta:-2.3), erythromycin (64.1 to 56%, beta:-1.9), fosfomycin (20.5 to 11.9%, beta:-1.12), ciprofloxacin (84.6 to 72.7%, beta:-2) and rifampicin (10.2 to 2.5%, beta:-0.81) declined markedly (p-value < 0.05). By contrast, resistance rates of MSSA strains to clindamycin (9.6 to 18.2%, beta:1.2) and erythromycin (15 to 22%, beta:1) increased significantly (p-value < 0.05), while penicillin resistance rate decreased over time (82.2 to 73.4%, beta:-0.92). For the remaining antibiotics, no changes in resistance levels were considered significant by these models.

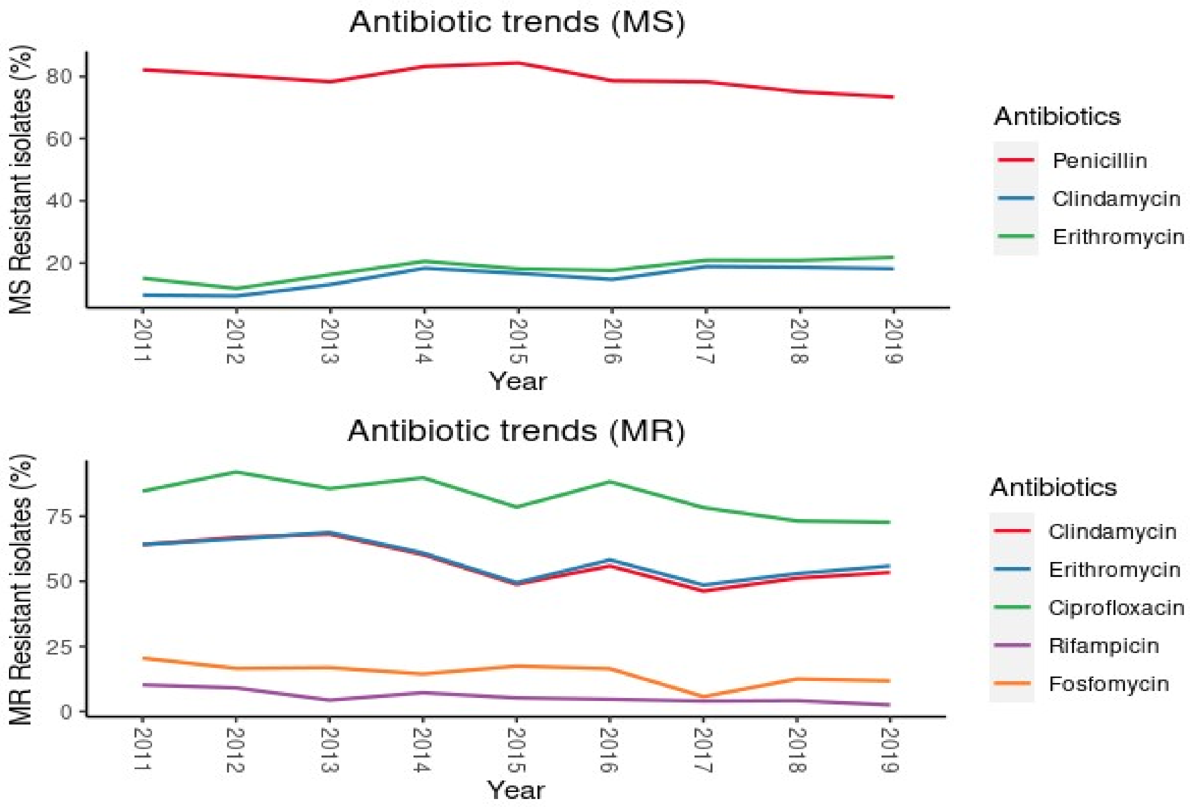


Fig 2: Time trend of percentages of resistance to different antibiotics throughout the studied period in MSSA and MRSA

Community and hospital acquired infections

During the nine years of study, 3107 patients positive for SA were hospitalized at the San Matteo Hospital. Based on the CDC epidemiological criteria mentioned in Materials and Methods, 54.1% of these samples were classified as Community Acquired (CA n=1681, mean age=57) and 45.9% as Hospital Acquired (HA n=1426, mean age=61). During the studied period, the incidence of recorded CA infection significantly (p-value < 0.05) increased by 16.6%, from 41.1% to 57.7% (with a corresponding drop in HA percentages), with the community becoming the primary source of infection.

By additionally partitioning the HA and CA samples into MR and MS subgroups, we observed significant changes of CA-MSSA and HA-MRSA in inpatients. The percentage of CA-MSSA increased from 29.4 to 44.1%, while HA-MRSA declined from 27.9 to 14.8%.

For the other 2 categories, HA-MSSA (30.9 to 27.5%) and CA-MRSA (11.8 to 13.6%), no meaningful variations are noticed (Fig 3).

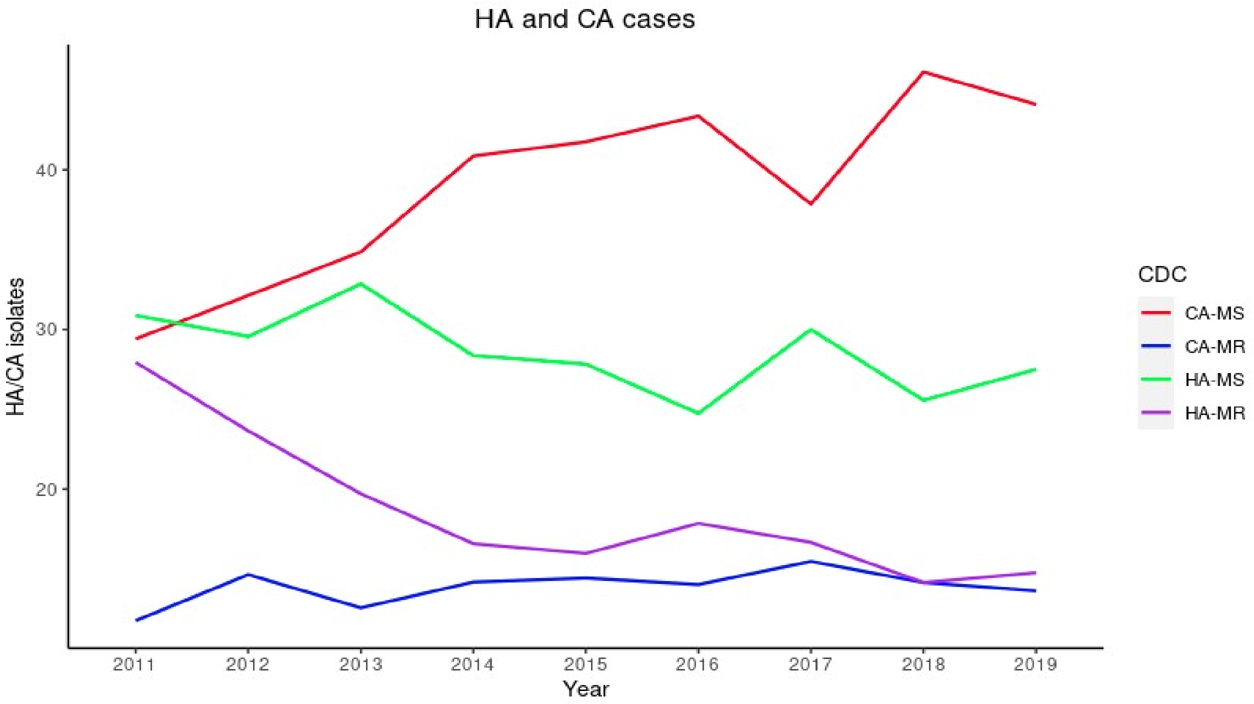


Fig 3: Temporal trends of CA-MS, CA-MR, HA-MS and HA-MR isolates in the San Matteo Hospital between 2011-2019

CA and HA infections also significantly vary by body sites. CA isolates were more associated to patients with skin (23.7%) or soft-tissue infections (33%) while HA isolates were more associated to blood (27.4%) and respiratory tract (28.1%) infections. Comparison of resistance patterns of CA and HA samples showed HA strains to be more resistant than CA strains to most antibiotics, with larger differences for methicillin (38.3% and 26%, respectively) and ciprofloxacin (40.1% and 27.1%, respectively). When considering MS and MR subgroups, negligible differences are observed in the comparison of the CA-MS and the HA-MS, with differences in resistance levels reaching at most 2.4%. On the other hand, when comparing CA-MR and HA-MR this difference goes up to 8.4%.

Overall, our results document an increase in the proportion of CA infections, especially CA-MSSA, starting as a minority in 2011 and becoming the majority in 2019, overcoming the number of HA. Phenotypically, these CA strains show relevant differences with respect to HA : besides the site of infection, where HA strains seem to be involved in more invasive infections (pneumonia and bacteremia), differences in the susceptibility of CA and HA isolates are also detected, with HA strains being more resistant to most of the antibiotics.

Mortality rate

The mortality rate (last isolate per patient), calculated within 28 days from hospitalization, shows that the considered infections resulted in at least 435 casualties (14%, mean age=71.8). However, between 2011 and 2019, deaths associated to *S. aureus* infections declined, from 15.5% to 10%. The site of infection as expected influences the outcome, with higher mortality rates among respiratory tract (30%) and bloodstream (45%) infections.

Considering susceptibility to methicillin, which is a well-known predictor of mortality, the attributable mortality rate for MSSA and MRSA infections was 10.7% (n=225) and 20.7% (n=210), respectively. This pattern does not surprise as on average MRSA strains show an enhanced resistance to multiple compounds and and are also generally associated to more severe infections. Furthermore, longer hospitalisation stays, which are associated with negative outcomes, are reported among deceased inpatients with MRSA infections, with an average length of 28 days, compared to only 20 days for MSSA infected inpatients. Going more into details, the percentage of both MSSA and MRSA caused deaths decreased from 11% in 2011 to 8% in 2019 and from 23% in 2011 to 14% in 2019, respectively.

Considering CDC categories, the percentage of both deaths caused by CA and HA infections decreased through the years, from 13.6% in 2011 to 9.3% in 2019 for CA and from 16.6% in 2011 to 10.8% in 2019 for HA.

All this data together suggests that at the San Matteo Hospital, mortality rate reduced over time, with a similar decreasing trend among MR and MS patients, regardless of the source of infection (community or hospital).

Clustering

As explained in the “Materials and Methods” section above, we devised a strategy to exploit the available resistance patterns to cluster all isolates. This approach then allowed us to select a certain number of isolates for genome sequencing with the goal of *maximizing* the detected genomic diversity. As hospitals often are subjected to outbreaks and are populated by well-defined clones at defined time periods, the risk of random sampling is to sequence the more common strains multiple times while rare strains are consequently overlooked. We reasoned that antibiotic resistance profiles can be considered as a raw proxy of genomic relatedness, meaning that it is much more probable that two isolates sharing multiple resistances belong to the same genomic clone than two isolates with completely different resistance profiles. Therefore, we decided to use an algorithm able to cluster “similar” resistance profiles and then we sequenced by sampling from all the resulting groups.

Clustering was obtained through a two-steps procedure. UMAP was first used as pre-processing step to reduce the dimensionality of the dataset. After that, we applied the DBSCAN algorithm to exploit the modular structure of the data and partition the observations into clusters.

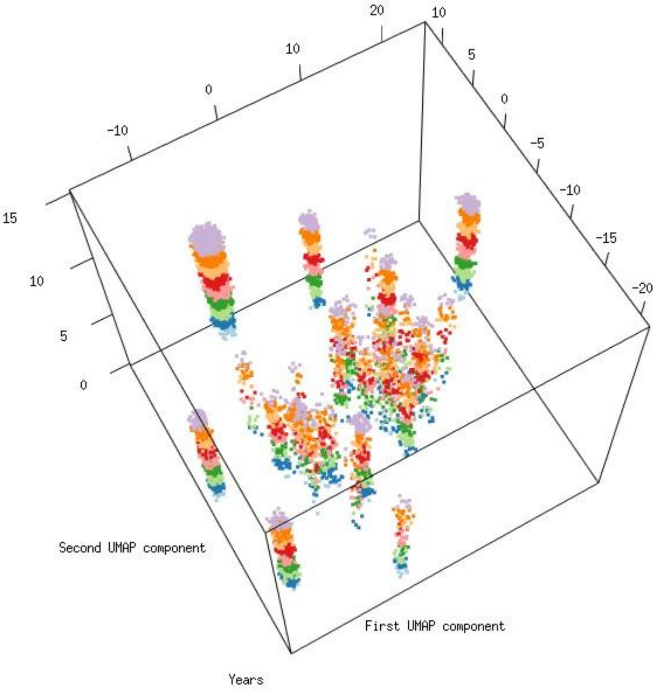


Fig 4: Temporal trends of the 24 clusters defined based on antibiotic profiles

In practice, this procedure summarized the 243 unique resistance profiles of the 7523 isolates into 24 clusters (Fig 4), each grouping isolates with similar patterns of antibiotic resistance. UMAP identified 16 methicillin susceptible (MS) clusters and 8 methicillin resistant (MR) clusters.

Selection of isolates for genome sequencing

Leveraging the power of the clustering analysis, we carried out a meticulous manual selection for genome sequencing, knowing that we had access only to bacteremia isolates, as these are routinely stored. The goal of the selection was to include at least one isolate from each UMAP cluster for each year of the study and to also include those profiles with higher antibiotic resistance levels. This procedure led us selecting 226 samples, representing 79 distinct profiles, 44% of which were MR, so respecting the proportions seen in the initial dataset in terms of isolates and profiles.

As we sequenced only these strains, we had no way to measure how much we increased the capability to capture genomic variability compared to a random selection of isolates. Therefore, to infer this, we compared the antibiotic profiles obtained by our procedure with 100 random sampling runs from the original list of isolates. The random sampling selected an average of 49 unique profiles (standard deviation=4), which is significantly less than the 79 unique profiles sampled using the UMAP-based strategy. Additionally, as a consequence, the samples were also much more biased towards very abundant strains. On the contrary, our strategy (dimensionality reduction of the antibiotic resistance profiles, followed by clustering and sampling for genomics) enabled us to observe very rare resistance profiles (with frequencies below 1%), allowing a better description of the genomic variability present in the Hospital.

Multi-Locus Sequence Typing

The 226 genomes were sequenced and assembled, (average N50: 245178.5, average genome length: 2,817,127, average N50n: 7, average number of contigs: 72). MLST typing identified 49 different Sequence Types (STs), including 16 novel STs (6438 – 6452, 6461), submitted to pubmlst. Among them, the most abundant are ST1 (n=16), ST5 (n=20), ST8 (n=29), ST22 (n=54), ST45 (n=10), ST72 (n=10) and ST228 (n=10), together accounting for 66% of the dataset. Out of 226 samples, 192 were assigned to 8 different Clonal Complexes (CCs): CC1 (n=20), CC15 (n=6), CC22 (n=59), CC30 (n=10), CC45 (n=15), CC5 (n=35), CC8 (n=40) and CC97 (n=7). For the remaining 34 it was not possible to assign any CC.

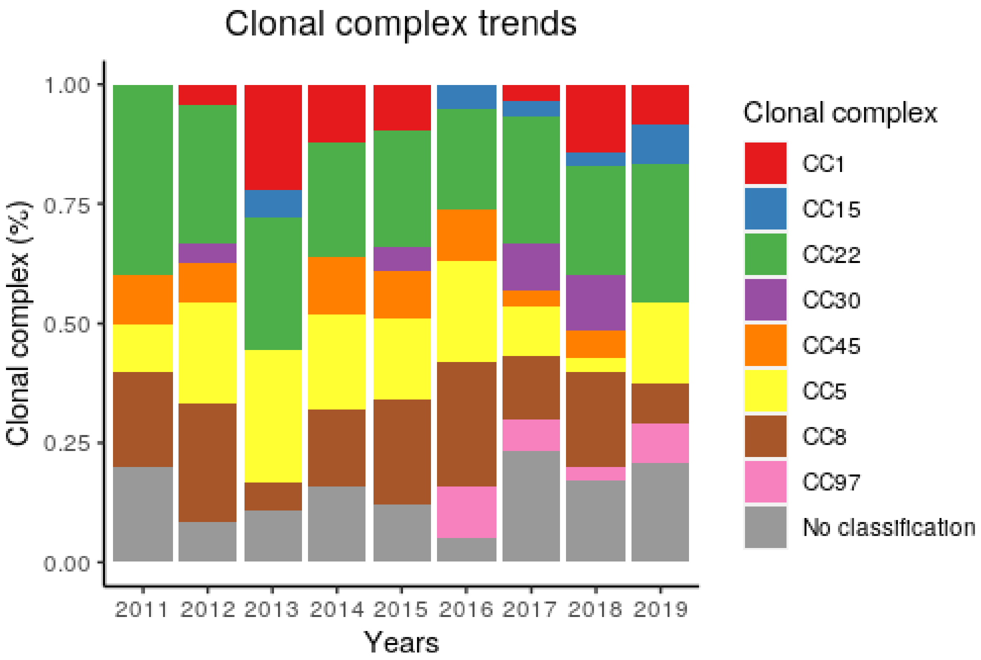


Fig 5: Temporal trends of CCs in the San Matteo Hospital between 2011-2019

Over the years, the clonal structure composition analysis of our bacteremia isolates report CC5, CC8 and CC22, being consistently present in the Hospital, while the others CCs and STs are subjected to variable fluctuations (Fig 5).

Staphylococcal Cassette Chromosome mec (SCCmec)

From a genomic perspective, SCCmec classification identified 4 different SCCmec types, for a total of 97 MR samples. Among these, classical nosocomial SCCmec types I (n=11) and II (n=6) represent a minority and are all included in CC5, (except for 1 sample included into CC22). Instead, community-associated cassette IV (n=74) encompasses 76% of all MRSA samples and is found in different CCs, especially CC8 (60%) and CC22 (64%). Lastly, SCCmec V, another community-associated cassette, is detected in only 6 genomes with no meaningful distribution among CCs. Temporal statistical analyses show a significant downtrend (beta: -5.8) of SCCmec type I prevalence over the years: starting from 2013, when this cassette is found in 45% of genomes, all the way down to 2019, where it is completely absent. On the other hand, SCCmec type IV, for the nine analysed years, showed an uprising frequency: starting from 2011, where this cassette is found in almost 60% of the genomes, up to 81.1% in 2019. These findings highlight that, in the Hospital, strains with SCCmec-IV have gradually overtaken the typical HA-MRSA strains endowed with SCCmec type I and II, confirming the trend observed in many recent studies

(Valsesia et al., 2010) reporting the infiltration of CA-SCCmec type IV in hospital settings, replacing the traditional HA-MRSA strains. Interestingly, we came across some discrepancies between the phenotypic and genomic testing of methicillin resistance (discrepancy rate: 6%): (I) SCCmecFinder detected 2 sample without any type of cassette that was phenotypically classified as MRSA; (ii) SCCmecFinder detected 4 samples presenting a SCCmec cassette that was phenotypically classified as MSSA.

Community and hospital acquired infections

CDC classification, antimicrobial resistance patterns, the distribution of the SCCmec cassette, the expression of PVL gene and CC affiliation are all features we used to discriminate community from hospital acquired infections, thus investigating the relations between isolate origin and these characteristics.

According to CDC epidemiological criteria, 35% (n=78) of our bacteremic samples were classified as CA and 65% (n=144) as HA. For both CDC categories, we did not find statistically significant linear trends over the 9 years but rather waves of infections.

Among MRSA samples, SCCmec cassettes did not correlate with the type of infection, even though we noticed a certain preponderance of samples carrying cassette I or II classified as hospital-acquired (with the exception of 3 samples classified as CA).

Instead, SCCmecIV is homogeneously distributed (33% in both groups) between the hospital and community environments, representing the primary source of methicillin resistance in the two categories.

HA-MRSA strains showed greater resistance than CA-MRSA strains to most of antibiotics (see table), with higher differences for clindamycin (43.7% and 32%), methicillin (45% and 36%) and fosfomycin (16% and 9%, respectively). On the other hand, CA-MRSA strains had higher rates of resistance to tetracycline (25.6% in CA and 13% in HA). Lastly, PVL gene was detected in only 8 samples: 3 were CA and 4 HA.

Taken together, these differences in the antimicrobial resistance patterns, SCCmec distribution and PVL gene presence, were not stastically significant (p-value > 0.05) and didn’t correlate with neither CA nor HA infections. Therefore, it seems that also inside the San Matteo Hospital, as already observed in many other clinical institutions, the epidemiological and molecular features used to classify staphylococcal infections are no longer valid and concordant. Most of the nosocomial strains harbour a community genomic background which confirms how the two lineages may have mixed up over the last years, contributing to the onset of new clones with new genetic traits.

Outcome

The isolates selected for our study were collected from 201 patients whose median age was 66 years (range 0-92 years), with 6% being younger than 16 years. The total length of stay (LOS) in hospital per patient ranged from less than a full day up to 191 days, with a median of 24 days.

The mortality rate was calculated within 28 days from the last hospitalization, with 29.16% (n=53) of deceased patients at the time of data collection; 22.4% (n=24) were infected by MSSA (n=29) and 25.5% by MRSA.

None of the CCs was found to be significantly associated with an increased risk of death, even though we found CC22, CC8 and CC30 to have a rate of mortality above the mean, standing at 41%, 38% and 33% , respectively.

In our study, CC30 is a relatively small group of 10 MSSA samples, of which 63% bears the gene *tst*, a virulence factor responsible for the toxic shock syndrome condition, an acute disease characterized by high fever, skin rash followed by skin peeling, hypotension, vomiting, diarrhea and potentially leading to multisystem organ failure and death (Schaefers et al., 2012). Thus, the presence of the gene *tst* may be a plausible explanation for such a high mortality rate.

In CC8 and CC22, all samples are either susceptible to methicillin (n=13) or mecIV carrier (n=18). Among the latter, 89% (n=16) of them were previously classified as HA.

From an antibiotic viewpoint, CC8 and CC22 did not show relevant differences, with the only exceptions represented by gentamicin, trimethoprim, fosfomycin and tetracycline that in CC8 (57.5%, 30%, 32.5% and 12.5%) had a resistance rate higher than in CC22 (30%, 10%, 13.5% and 1.6%).

Greater significant dissimilarities (p-value < 0.05) were instead noticed at virulence level: presence/absence of serine proteases (*splA*, *splB*, *splE*), leukotoxin (*lukED*), enterotoxins (*sed*, *sej*, *ser*, *seg*, *sel*, *sem*, *sen*, *seo*, *seu*) and immune-modulating genes (*sak*, *chp*, *scn*) deeply correlated with the clonal complex (Fig 6).

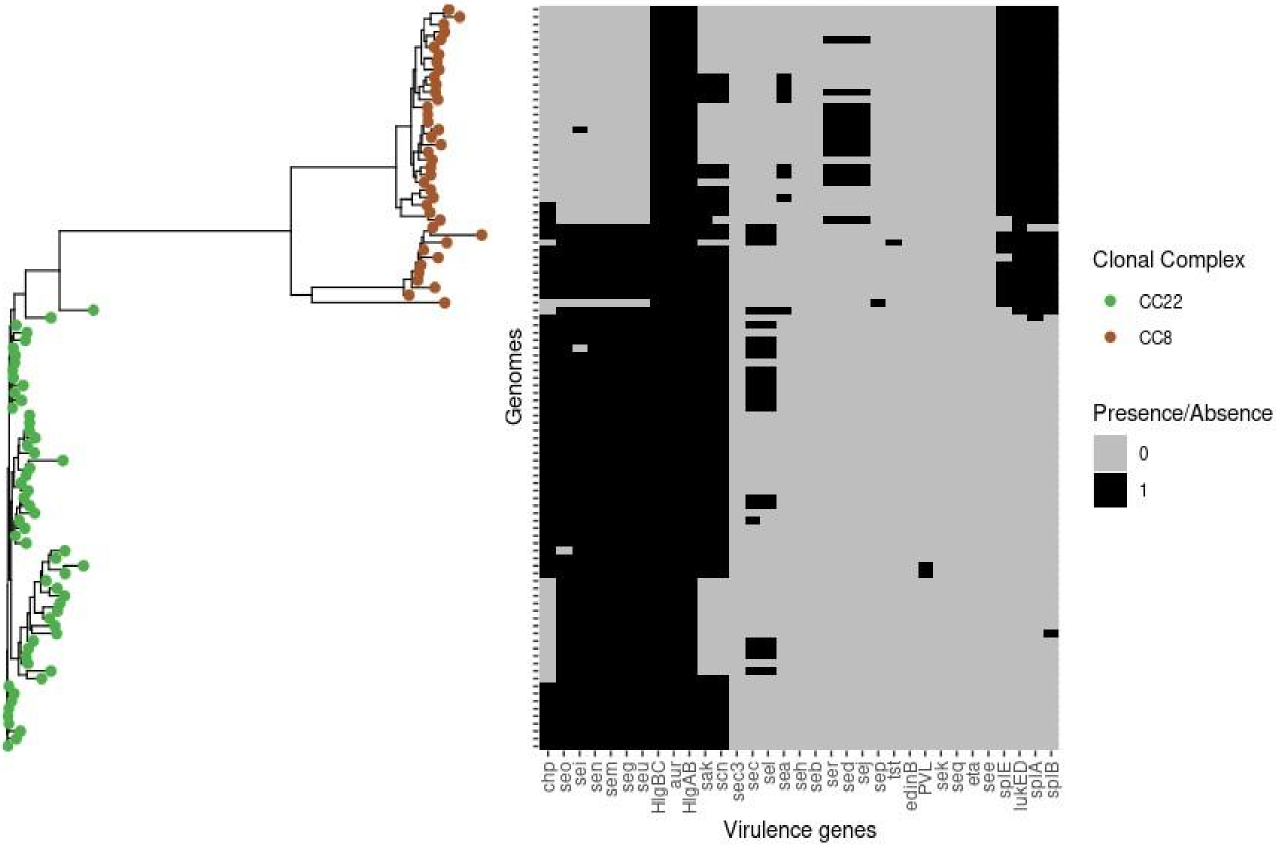


Fig 6: Virulence factors distribution among CC22 and CC8 samples

Phylogenetic reconstruction

A total of 143,221 coreSNP sites within the core genome were identified by mapping the entire dataset against a single reference genome, HO 5096 0412 (ST22).

Analysis of core SNPs resolved the population into 8 clonal complexes (CC1, CC15, CC22, CC30, CC45, CC5, CC8, CC97) and rare STs, as previously defined by multilocus sequence typing (Fig 7).

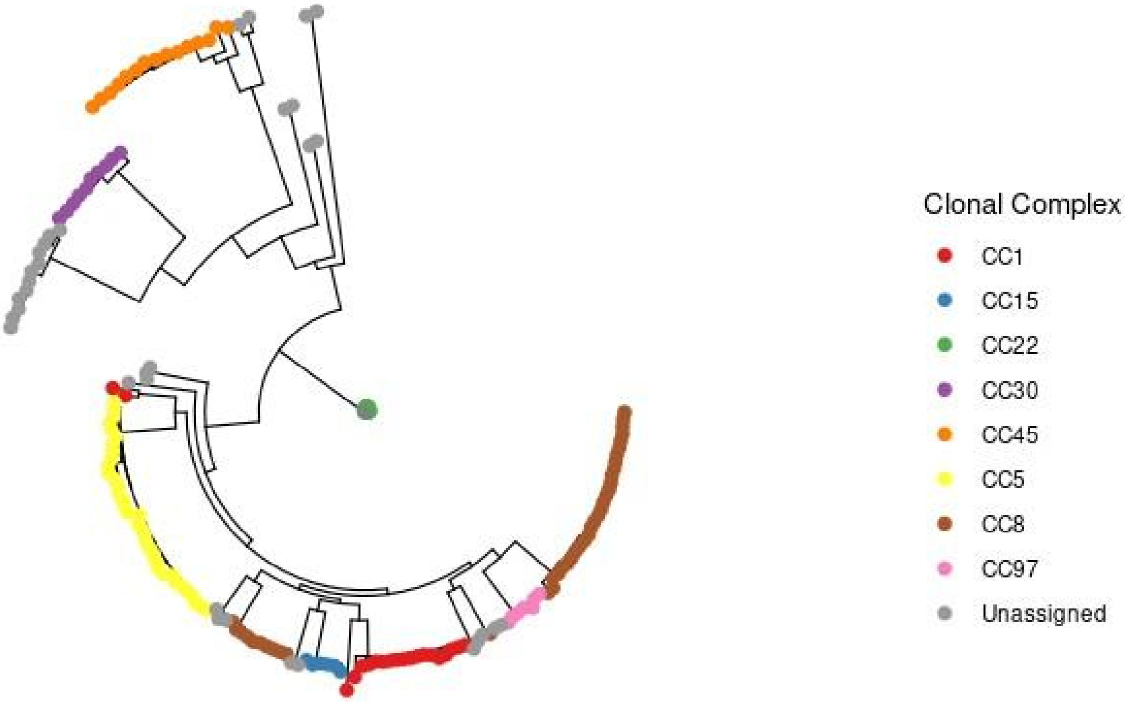


Fig 7: Phylogenetic analysis of all sequenced 226 *S. aureus* samples

To gain more insights about the phylogenetic relationships among our sample, for CC22 and CC8, which are the largest clonal complexes in our study, we performed further specific analyses.

CC22: In our study, we detected 5 different STs within CC22, for a total of 59 isolates, 91.5% of which being ST22. The alignment of all these samples against a single reference genome (ST22, GenBank: CP022291.1) resulted in the identification of 6434 unique SNPs which were used to measure intra-CC diversity.

As we can see from the dendrogram below (Fig 8), we identify a large clade of 39 highly related samples (max SNPs: 499), 37 of which being ST22 and, 2 being of ST1327 and ST3863.

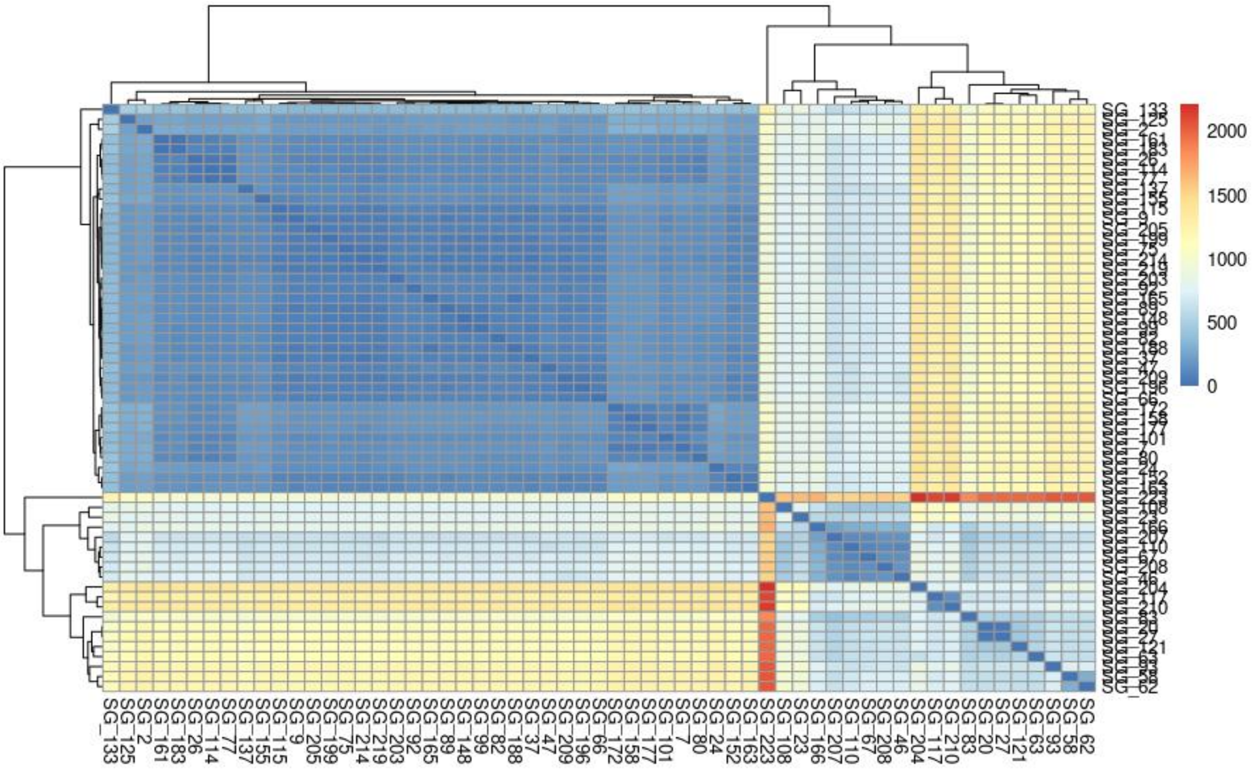


Fig 8: Dendrogram of the SNP distances for all samples of CC22

As opposed to the other samples, this clade embraces more virulent and resistant samples

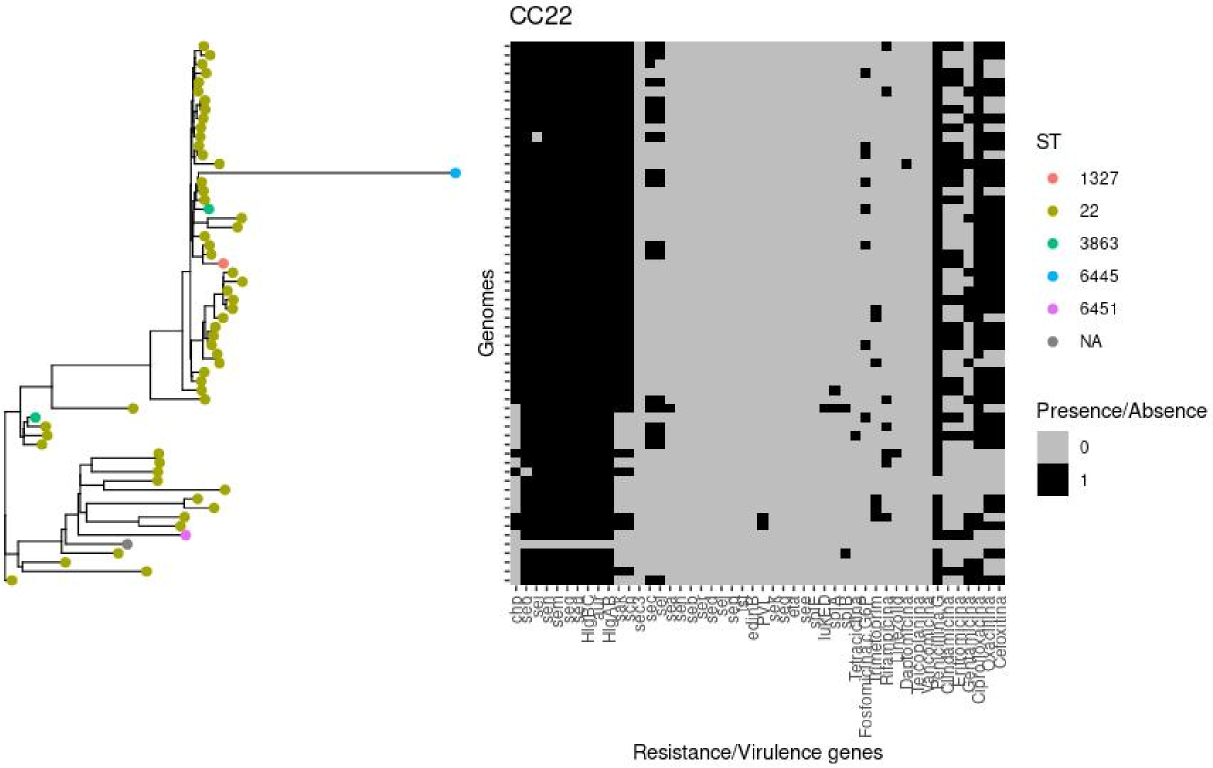


Fig 9: Phylogenetic reconstruction of CC22

Indeed, as we can see from the heatmap above (Fig 9), in addition to all the conserved virulence factors, this clade is characterized also by the presence of the virulence genes *sak*, *chp* and *scn*. These 3 genes are immune modulators clustered on the conserved 3′ end of β-hemolysin (*hlb*)-converting bacteriophages (β*C-φs*) (Van Wamel et al., 2006). Instead, from an antibiotic resistance point of view, this clade of samples displays higher levels of resistance for most antibiotics, with the exception of trimethoprim, rifampicin and gentamicin.

To gain more insights about the nature of this clade, we also performed a global phylogenetic analysis including 194 CC22 genomes downloaded from Patric. From the tree (Fig 10) we can notice that most of the isolates belonging to the local clade of 39 samples falls into two sub-clades, of 26 (mean SNPs: 153) and 11 (mean SNPs: 92) samples, respectively.

As reported from our databases, the first monophyletic sub-clade populated the hospital for all of the 9 years of the study, with a mixture of MSSA (n=7) and SCCmecIV (n=19), classified as either CA (n=10) or HA (n=16). The second monophyletic sub-clade populated the hospital starting from 2013 until 2019, missing in 2016. Even in this case, there is a mixture of MSSA (n=3) and SCCmecIV (n=8), classified as either CA (n=2) or HA (n=9). For both clades, we do not observe clusterization in terms of methicillin resistance nor based on CDC categorization.

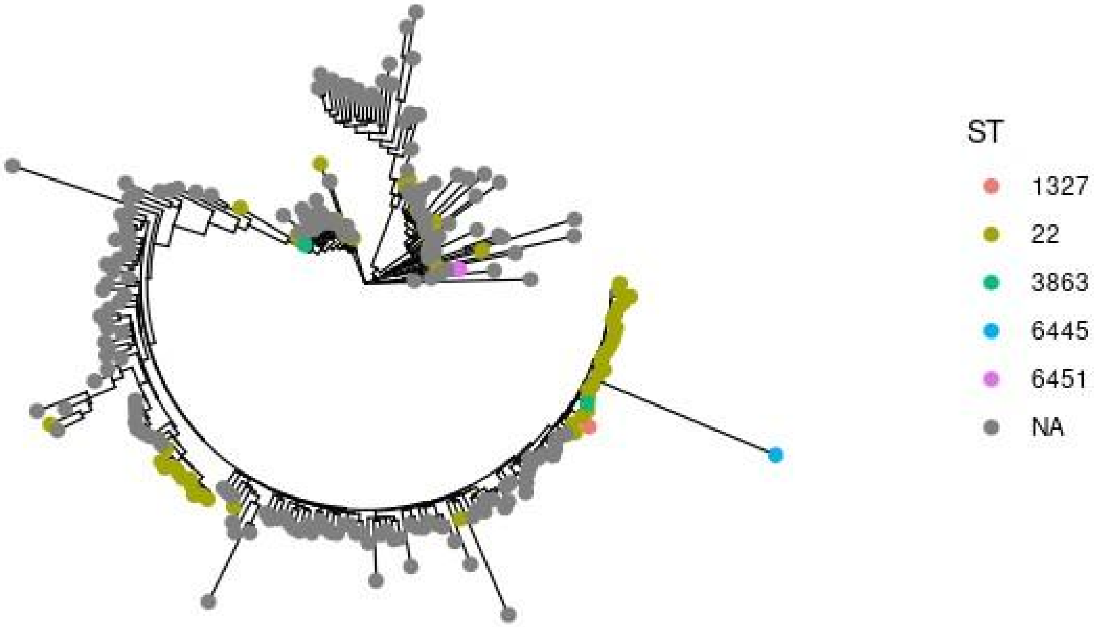


Fig 10: Phylogenetic analysis of CC22 including CC22 samples from Patric

In particular, it is important to notice that ST22-SCCmec IV samples in our study belong to the epidemic clone EMRSA-15, isolated both in hospital and community settings, initially found in England but then endemic in many countries in the world, such as Italy, like in our case (O’Neill et al., 2001).

CC8: In our study, CC8 is represented by 40 isolates encompassing 3 different STs, 72.5% of which being ST8 and 25% being ST72. The alignment of all these samples against a single reference genome (ST8, GenBank: JYAB01000001.1) resulted in the identification of 15437 unique SNPs, which were used to measure intra-CC diversity.
ST8 forms a single cluster of 29 genomes (max SNPs: 933), which persistently populated the hospital over the period 2011-2019, 80% of which being MRSA-SCCmecIV (n=23) and 20% MSSA (n=6). Besides a few exceptions (n=8), most isolates correspond to spa type t008. ST72 encompasses 10 genomes (max SNPs: 647) and besides 1 sample SCCmec-IV, the rest is classified as MSSA (Fig 11).

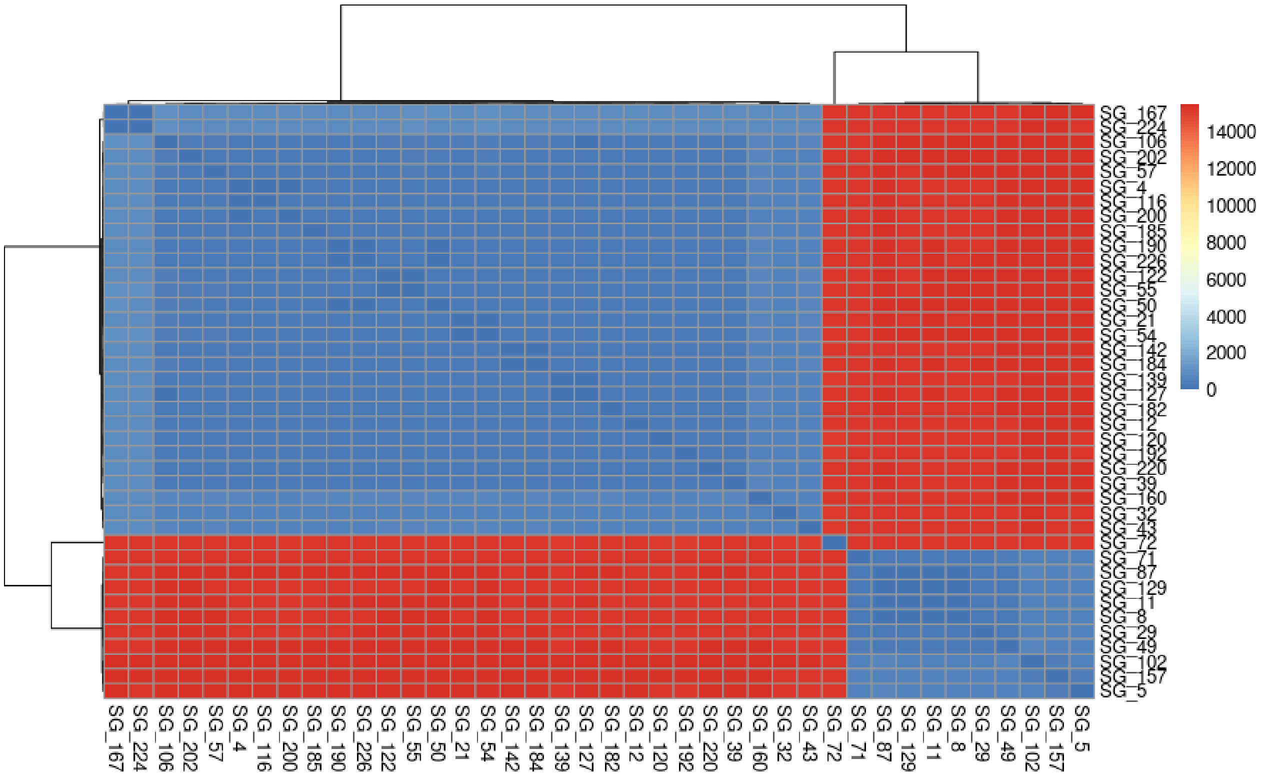


Fig 11: Dendrogram of the SNP distances for all samples of CC8

As we can observe from the heatmap below (Fig 12), from the comparison between ST8 and ST72, we notice that the cluster of genomes of the first sequence type is more resistant to antibiotics but less virulente. Indeed, ST72 is characterized by the further presence of *chp* and, also by the presence of *seg*, *sei*, *sem*, *sen*, *seo*, *seu* genes, which form a cluster of enterotoxins localized on the pathogenicity island υSaβ .

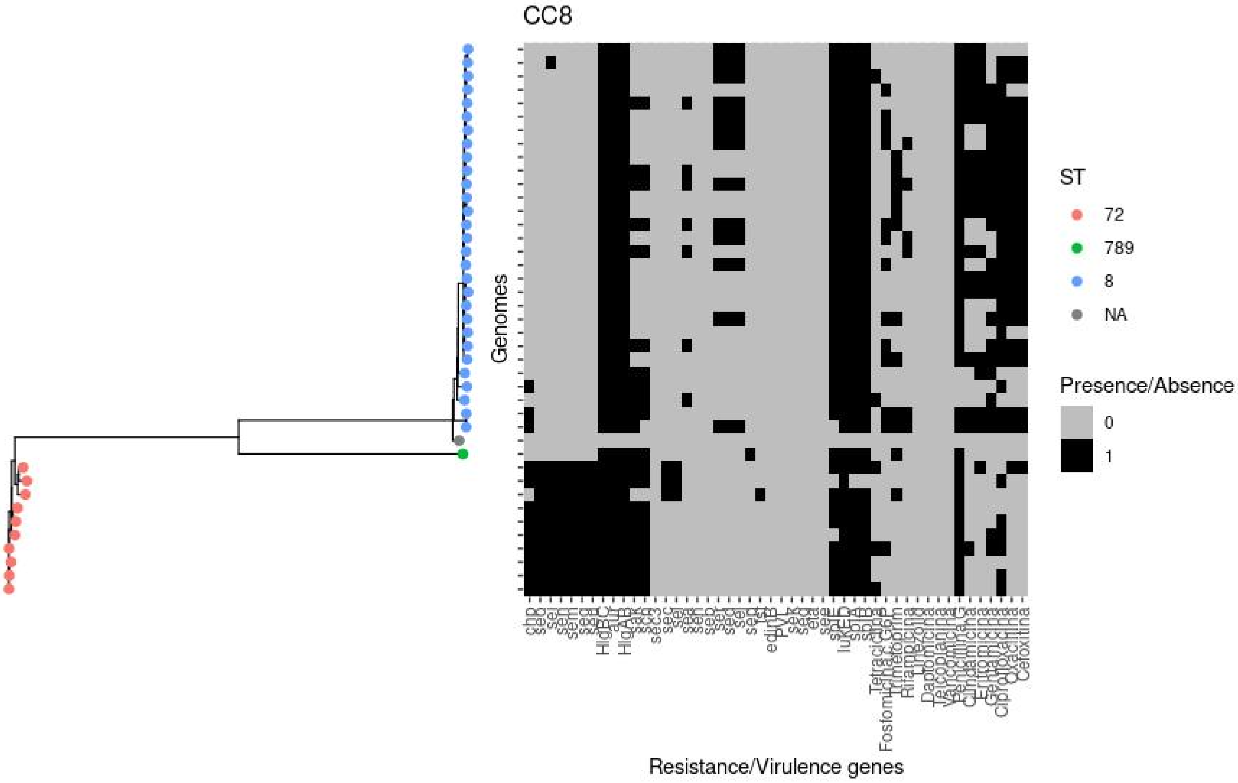


Fig 12: Phylogenetic reconstruction of CC22

To gain more insights about the nature of this clade, as for CC22, we performed a global phylogenetic analysis including 233 CC8 genomes downloaded from Patric (Fig 13).

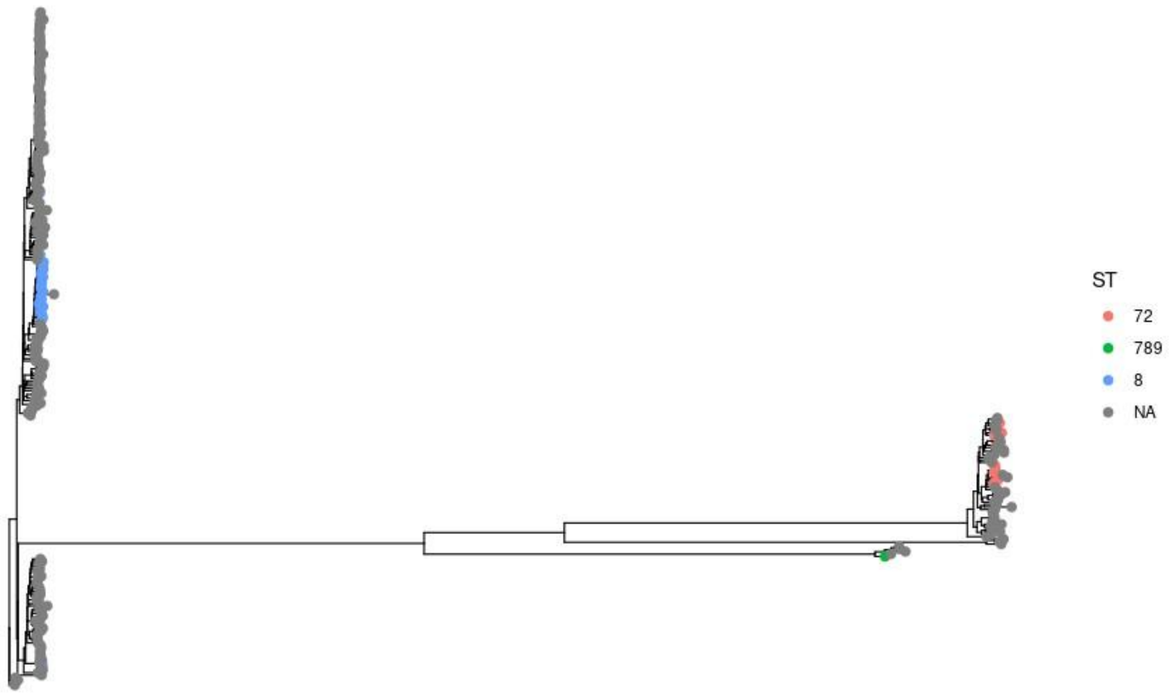


Fig 13: Phylogenetic analysis of CC8 including CC8 samples from Patric

From the tree we can notice that the ST8 forms a big clade (n=24) of MSSA and MRSA samples (mean SNPs: 215), interrupted by the presence of 1 single MRSA genome (1280.16757), collected in Florence (Italy) in 2018 (Ravenni et al., 2018). Within this clade we recognize the presence of the epidemic clone USA300 (ST8-SCCmec IV-t008), originally found in the US, which is a leading cause of community and hospital-acquired infections also in Italy (Read et al., 2018).

Instead, our ST72 samples form a clade of 7 isolates (mean SNPs: 156), all belonging to spa type t148, interrupted by the presence of two non Italian strains, sampled in 2018 in South and North America (1280.15859 and 1280.18629).

For both STs, we do not observe any sort of clusterization, neither in terms of methicillin resistance nor CDC categorization.

CONCLUSIONS AND FUTURE DIRECTIONS

Staphylococcus aureus keeps being an important issue in the clinical setting due to its capacity to adapt to different environments and to swiftly become resistant to multiple antibiotics. Thus, a better understanding of its evolution and an early recognition of high-risk clones is now fundamental for an efficient management of patients.

In this context, the goal of this Ph.D. project has been to provide an accurate characterization of circulating *S. aureus* strains in the San Matteo Hospital (Pavia, Italy), between 2011 and 2019. For this project, I focused on a collection of 226 *Staphylococcus aureus* samples isolated from blood, carefully chosen through a clustering strategy based on antibiograms.

With this approach, compared to random sampling, we were able to increase the frequency of rare sequence types or clonal complexes while demoting those that are far more abundant, allowing to provide a much deeper picture of the total genomic variability of all *S. aureus* strains. Overall, as antibiograms are usually obtained for all isolates, this strategy may be applied widely in hospitals.

Once assembled and quality checked, genomes were exhaustively characterized in terms of MLST, SCCmec cassette, Spa type and for the presence/absence of virulence/resistance genes. Besides these analyses, we also investigated likely correlations between these genomic features and additional information about the hospitalization and outcome of the patients that is available as metadata associated to the samples. At the end, we also evaluated the evolutionary distance among samples (SNPs resolution) with the aim of predicting the presence and dissemination, over the course of the years, of emerging and worrying strains.

The analyses revealed a hospital population with varied genomic background, composed of 8 different clonal complexes, for a total number of 49 different sequence types. Of these clonal complexes, only CC5, CC8 and CC22, are consistently present in the Hospital between 2011-2019, while the presence of other CCs and STs is subjected to fluctuations.

Within this framework, CDC classification of infections outlines a mixed scenario of community- and hospital-acquired strains, coexisting within the hospital. Interestingly, we did not observe any sort of meaningful phenotypic/genomic determinants able to discriminate community from hospital-acquired strains. For instance, among the MRSA strains, the classification of the SCCmec cassette showed that 82% of samples carries SCCmec cassettes of type IV or V, which were typically found in CA-MRSA strains, while just a few isolates is characterized by cassettes of type I and II, historically linked to hospital-acquired infections.

Finally, phylogenetic analyses of the largest groups of our study, which are CC22 and CC8, highlighted the presence of highly related clades of samples, mainly affiliated to ST22 and ST8, which persist inside the hospital for the entire period of study. Within these sequence types, we report the presence of two endemic clones, EMRSA-15 and USA300, which are responsible, also in our study, for community and hospital-acquired infections.

CONCLUSIONS

All the different lines of work presented in this thesis underline the high potential of bioinformatics in general, and more specifically of the analysis of next generation sequencing data, in the field of microbial genomics and genomic epidemiology.

In the projects on *Bacillus anthracis* and *Staphylococcus aureus,* genome sequencing provided a single platform for extracting all the information required to fully characterize strains of interest of these pathogens, perform comparative analyses and resolve local outbreaks.

In *Bacillus anthracis*, the comparison between wild-type (*7702* and *9131*) and mutated strains (*7702 Δsap-eag* and 9*131 Δeag*) made it possible to identify not only variations in gene content but also to accurately detect SNPs located in coding and non-coding regions, outlining the entire set of changes involved in response to the deletion of the S-layers and providing the candidate genes for future studies.

For *Staphylococcus aureus*, the genome sequencing of 226 blood isolates led to a detailed description of the lineages circulating inside the San Matteo Hospital between 2011 and 2019. We were able to fully type (MLST, Spa type and SCCmec) and characterize *S. aureus* strains for the presence of virulence and resistance determinants and also perform phylogenetic analyses in order to identify clusters of highly-related samples responsible for short-time or persistent outbreaks inside the hospital, thus offering a valuable source of information for proper patient management and an effective implementation of surveillance strategies.

The two RNA-Seq projects presented, on *Mycobacterium tuberculosis* and *Streptococcus agalactiae*, allowed to identify the differentially expressed genes and the metabolic pathways altered in response to events, such as administration of a new drug or deletion of a gene, respectively. Therefore, thanks to this explorative information, we were able to understand the molecular processes underlying new therapeutic approaches and provide our collaborators with new evidence on which to focus subsequent studies.

However, when performing comparative experiments, such as done for *M. tuberculosis*, *S. agalactiae* and *B. anthracis*, it is important to consider the influence of the environmental conditions for a correct interpretation of the results. The adaption of microorganisms to changing enviromental niches is the result of the conjuction of genetics and ecology. In bacteria, for instance, it is common to observe phenotypes, which are fixed in a certain population but which are counterselected in other enviroments. These differences are the reflection of different genetic backgrounds which may confer, in response to selective pressures, advantages or disadvantages in the adaptation process (Lenski, 2017).

As a natural consequence, these considerations also apply when performing experimental comparisons. Sometimes, experimental patterns that are identified in bacteria are actually rare or never observed in nature. A classic example is what we see for *Staphylococcus aureus*, where numerous experimental mutations have been found in the *Agr* operon but then, actually, never confirmed in nature (Shopsin et al., 2010).

To conclude, NGS provide a wealth of data that can be exploited with the right bioinformatic tools. The potential is enormous and with the improvement of these technologies and the lowering of costs, it is clear that the next step will be to routinely use these applications in the field of diagnostic microbiology. The aim must now be to further optimize analytical tools which will allow routine genome (and transcriptome) characterization, and to incorporate bioinformatics in diagnostics. This will lead to the ultimate goal of reducing the time span between pathogen isolation and specific treatment and epidemiological characterization, improving the outcome for the patients and limiting the possibility of spread of the pathogens.

REFERENCES

- Al-Saeedi, M., & Al-Hajoj, S. (2017). Diversity and evolution of drug resistance mechanisms in *Mycobacterium tuberculosis*. *Infection and Drug Resistance*, *10*, 333–342. https://doi.org/10.2147/IDR.S144446

- Albers, S. V., & Meyer, B. H. (2011). The archaeal cell envelope. In *Nature Reviews Microbiology* (Vol. 9, Issue 6). https://doi.org/10.1038/nrmicro2576

- Awram, P., & Smit, J. (1998). The *Caulobacter crescentus* paracrystalline S-layer protein is secreted by an ABC transporter (Type I) secretion apparatus. *Journal of Bacteriology*, *180*(12), 3062–3069. https://doi.org/10.1128/jb.180.12.3062-3069.1998

- Bakour, S., Sankar, S. A., Rathored, J., Biagini, P., Raoult, D., & Fournier, P. E. (2016). Identification of virulence factors and antibiotic resistance markers using bacterial genomics. *Future microbiology*, 11(3), 455–466. https://doi.org/10.2217/fmb.15.14

- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

- Biswas, I., Gruss, A., Ehrlich, S. D., & Maguin, E. (1993). High-efficiency gene inactivation and replacement system for gram-positive bacteria. *Journal of Bacteriology*, *175*(11), 3628–3635. https://doi.org/10.1128/jb.175.11.3628-3635.1993

- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

- Buchan, B. W., & Ledeboer, N. A. (2014). Emerging technologies for the clinical microbiology laboratory. *Clinical Microbiology Reviews*, *27*(4), 783–822. https://doi.org/10.1128/CMR.00003-14

- Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

- Bursle, E., & Robson, J. (2016). Non-culture methods for detecting infection. *Australian prescriber,* 39(5), 171–175. https://doi.org/10.18773/austprescr.2016.059

- Cerquetti, M., Molinari, A., Sebastianelli, A., Diociaiuti, M., Petruzzelli, R., Capo, C., & Mastrantonio, P. (2000). Characterization of surface layer proteins from different *Clostridium difficile* clinical isolates. *Microbial Pathogenesis*, *28*(6), 363–372. https://doi.org/10.1006/mpat.2000.0356

- Chami, M., Bayan, N., Peyret, J. L., Gulik-Krzywicki, T., Leblon, G., & Shechter, E. (1997). The S-layer protein of *Corynebacterium glutamicum* is anchored to the cell wall by its C-terminal hydrophobic domain. *Molecular Microbiology*, *23*(3), 483–492. https://doi.org/10.1046/j.1365-2958.1997.d01-1868.x

- Chateau, A., Lunderberg, J. M., Oh, S. Y., Abshire, T., Friedlander, A., Quinn, C. P., Missiakas, D. M., & Schneewind, O. (2018). Galactosylation of the secondary cell wall polysaccharide of *Bacillus anthracis* and its contribution to anthrax pathogenesis. *Journal of Bacteriology*, *200*(5), 1–15. https://doi.org/10.1128/JB.00562-17

- Château, A., Schaik, W., Six, A., Aucher, W., & Fouet, A. (2011). *CodY* regulation is required for full virulence and heme iron acquisition in *Bacillus anthracis* . *The FASEB Journal*, *25*(12), 4445–4456. https://doi.org/10.1096/fj.11-188912

- Chaudhuri, R., & Ramachandran, S. (2014). Prediction of virulence factors using bioinformatics approaches. *Methods in molecular biology*, 1184, 389–400. https://doi.org/10.1007/978-1-4939-1115-8

- Chitlaru, T., Gat, O., Grosfeld, H., Inbar, I., Gozlan, Y., & Shafferman, A. (2007). Identification of in vivo-expressed immunogenic proteins by serological proteome analysis of the *Bacillus anthracis* secretome. *Infection and Immunity*, *75*(6), 2841–2852. https://doi.org/10.1128/IAI.02029-06

- Churchyard, G., Kim, P., Shah, N. S., Rustomjee, R., Gandhi, N., Mathema, B., Dowdy, D., Kasmar, A., & Cardenas, V. (2017). What We Know about Tuberculosis Transmission: An Overview. *Journal of Infectious Diseases*, *216*(Suppl 6), S629–S635. https://doi.org/10.1093/infdis/jix362

- Coate, J. E., & Doyle, J. J. (2015). Variation in transcriptome size: are we getting the message? *Chromosoma*, *124*(1), 27–43. https://doi.org/10.1007/s00412-014-0496-3

- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 1–19. https://doi.org/10.1186/s13059-016-0881-8

- Croucher, N. J., & Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*, *13*(5), 619–624. https://doi.org/10.1016/j.mib.2010.09.009

- Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, *5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

- De Almeida, O. G. G., & De Martinis, E. C. P. (2019). Relating next-generation sequencing and bioinformatics concepts to routine microbiological testing. *Electronic Journal of General Medicine*, *16*(3). https://doi.org/10.29333/ejgm/108690

- Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, *13*(9), 601–612. https://doi.org/10.1038/nrg3226

- Ducati, R. G., Ruffino-Netto, A., Basso, L. A., & Santos, D. S. (2006). The resumption of consumption - A review on tuberculosis. *Memorias Do Instituto Oswaldo Cruz*, *101*(7), 697–714. https://doi.org/10.1590/S0074-02762006000700001

- Emms, D. M., & Kelly, S. (2018). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *BioRxiv*, 1–14. https://doi.org/10.1101/466201

- Aanensen, D. M., Feil, E. J., Holden, M. T., Dordel, J., Yeats, C. A., Fedosejev, A., Goater, R., Castillo-Ramírez, S., Corander, J., Colijn, C., Chlebowicz, M. A., Schouls, L., Heck, M., Pluister, G., Ruimy, R., Kahlmeter, G., Åhman, J., Matuschek, E., Friedrich, A. W., Parkhill, J., … European SRL Working Group (2016). Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe. *mBio*, 7(3), e00444-16. https://doi.org/10.1128/mBio.00444-16

- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

- European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2016. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). Stockholm: ECDC; 2017.

- Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3), 186–194. https://doi.org/10.1101/gr.8.3.186

- *FastQC*. (2015). https://qubeshub.org/resources/fastqc

- Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., Peacock, S. J., Smith, J. M., Murphy, M., Spratt, B. G., Moore, C. E., & Day, N. P. J. (2003). How Clonal Is *Staphylococcus aureus*? *Journal of Bacteriology*, *185*(11), 3307–3316. https://doi.org/10.1128/JB.185.11.3307-3316.2003

- Feng, L., Zhu, J., Chang, H., Gao, X., Gao, C., Wei, X., Yuan, F., & Bei, W. (2016). The CodY regulator is essential for virulence in *Streptococcus suis* serotype 2. *Scientific Reports*, *6*(June 2015), 1–15. https://doi.org/10.1038/srep21241

- Fioravanti, A., Van Hauwermeiren, F., Van der Verren, S. E., Jonckheere, W., Goncalves, A., Pardon, E., Steyaert, J., De Greve, H., Lamkanfi, M., & Remaut, H. (2019). Structure of S-layer protein Sap reveals a mechanism for therapeutic intervention in anthrax. In *Nature Microbiology* (Vol. 4, Issue 11). https://doi.org/10.1038/s41564-019-0499-1

- Frieden, T. R., Fujiwara, P. I., Washko, R. M., & Hamburg, M. A. (1995). Tuberculosis in New York City — Turning the Tide. *New England Journal of Medicine*, *333*(4), 229–233. https://doi.org/10.1056/nejm199507273330406

- Furfaro, L. L., Chang, B. J., & Payne, M. S. (2018). Perinatal *Streptococcus agalactiae* epidemiology and surveillance targets. *Clinical Microbiology Reviews*, *31*(4), 1–18. https://doi.org/10.1128/CMR.00049-18

- Gaiarsa, S., Comandatore, F., Gaibani, P., Corbella, M., Valle, C. D., Epis, S., Scaltriti, E., Carretto, E., Farina, C., Labonia, M., Landini, M. P., Pongolini, S., Sambri, V., Bandi, C., Marone, P., & Sasserac, D. (2015). Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. *Antimicrobial Agents and Chemotherapy*, *59*(1), 389–396. https://doi.org/10.1128/AAC.04224-14

- Gengenbacher, M., & Kaufmann, S. H. E. (2012). *Mycobacterium tuberculosis*: Success through dormancy. *FEMS Microbiology Reviews*, *36*(3), 514–532. https://doi.org/10.1111/j.1574-6976.2012.00331.x

- George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., Mathers, A., Norris, R., Navickaite, I., Eaton, C., Iqbal, Z., Crook, D. W., & Phan, H. T. T. (2017). Resolving plasmid structures in enterobacteriaceae using the MinION nanopore sequencer: Assessment of MinION and MinION/illumina hybrid data assembly approaches. *Microbial Genomics*, *3*(8), 1–8. https://doi.org/10.1099/mgen.0.000118

- Gerbino, E., Carasi, P., Mobili, P., Serradell, M. A., & Gómez-Zavaglia, A. (2015). Role of S-layer proteins in bacteria. In *World Journal of Microbiology and Biotechnology* (Vol. 31, Issue 12). https://doi.org/10.1007/s11274-015-1952-9

- Haaber, J., Penadés, J. R., & Ingmer, H. (2017). Transfer of Antibiotic Resistance in *Staphylococcus aureus*. In *Trends in Microbiology* (Vol. 25, Issue 11, pp. 893–905). Elsevier Ltd. https://doi.org/10.1016/j.tim.2017.05.011

- Heym, B., Alzari, P. M., Honore, N., & Cole, S. T. (1995). Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Molecular Microbiology*, *15*(2), 235–245. https://doi.org/10.1111/j.1365-2958.1995.tb02238.x

- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

- Ito, T., Kuwahara-Arai, K., Katayama, Y., Uehara, Y., Han, X., Kondo, Y., & Hiramatsu, K. (2014). *Staphylococcal Cassette Chromosome mec (SCCmec)* analysis of MRSA. *Methods in molecular biology* (Clifton, N.J.), 1085, 131–148. https://doi.org/10.1007/978-1-62703-664-1_8

- Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *Journal of Clinical Microbiology*, *52*(5), 1501–1510. https://doi.org/10.1128/JCM.03617-13

- Kern, V. J., Kern, J. W., Theriot, J. A., Schneewind, O., & Missiakas, D. (2012). Surface-Layer (S-Layer) Proteins Sap and EA1 Govern the Binding of the S-Layer-Associated Protein BslO at the Cell Septa of *Bacillus anthracis*. *Journal of Bacteriology*, *194*(15). https://doi.org/10.1128/JB.00402-12

- Kolstø, A. B., Tourasse, N. J., & Økstad, O. A. (2009). What sets *Bacillus anthracis* apart from other Bacillus species? *Annual Review of Microbiology*, *63*, 451–476. https://doi.org/10.1146/annurev.micro.091208.073255

- Köser, C. U., Ellington, M. J., Cartwright, E. J. P., Gillespie, S. H., Brown, N. M., Farrington, M., Holden, M. T. G., Dougan, G., Bentley, S. D., Parkhill, J., & Peacock, S. J. (2012). Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. *PLoS Pathogens*, *8*(8). https://doi.org/10.1371/journal.ppat.1002824

- Landwehr-Kenzel, S., & Henneke, P. (2014). Interaction of *Streptococcus agalactiae* and cellular innate immunity in colonization and disease. *Frontiers in Immunology*, *5*(OCT), 1–11. https://doi.org/10.3389/fimmu.2014.00519

- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25

- Lenski R. E. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME journal*, 11(10), 2181–2194. https://doi.org/10.1038/ismej.2017.69

- Leopold, S. R., Goering, R. V., Witten, A., Harmsen, D., & Mellmann, A. (2014). Bacterial whole-genome sequencing revisited: Portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. *Journal of Clinical Microbiology*, *52*(7), 2365–2370. https://doi.org/10.1128/JCM.00262-14

- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
- Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, *12*(8), 733–735. https://doi.org/10.1038/nmeth.3444

- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

- M., M., F., T. Lo, A., A. D., Y., S. N., M., D., & M., N. (2016). Pan-genome analysis of Senegalese and Gambian strains of *Bacillus anthracis*. *African Journal of Biotechnology*, *15*(45), 2538–2546. https://doi.org/10.5897/ajb2016.14902

- Mcguinness, W. A., Malachowa, N., & Deleo, F. R. (2017). Vancomycin Resistance in *Staphylococcus aureus*. *The Yale journal of biology and medicine*, 90(2), 269-281

- McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. http://arxiv.org/abs/1802.03426

- Meftahi, N., Namouchi, A., Mhenni, B., Brandis, G., Hughes, D., & Mardassi, H. (2016). Evidence for the critical role of a secondary site *rpoB* mutation in the compensatory evolution and successful transmission of an MDR tuberculosis outbreak strain. *Journal of Antimicrobial Chemotherapy*, *71*(2), 324–332. https://doi.org/10.1093/jac/dkv345

- Mignot, T., Mesnage, S., Couture-Tosi, E., Mock, M., & Fouet, A. (2002). Developmental switch of S-layer protein synthesis in *Bacillus anthracis*. *Molecular Microbiology*, *43*(6), 1615–1627. https://doi.org/10.1046/j.1365-2958.2002.02852.x

- Moayeri, M., Leppla, S. H., Vrentas, C., Pomerantsev, A. P., & Liu, S. (2015). Anthrax Pathogenesis. In *Annual Review of Microbiology* (Vol. 69, Issue 1). https://doi.org/10.1146/annurev-micro-091014-104523

- Monaco, M., Pimentel de Araujo, F., Cruciani, M., Coccia, E. M., & Pantosti, A. (2017). Worldwide epidemiology and antibiotic resistance of *Staphylococcus aureus*. In *Current Topics in Microbiology and Immunology* (Vol. 409, pp. 21–56). Springer Verlag. https://doi.org/10.1007/82_2016_3

- Morlock, G. P., Metchock, B., Sikes, D., Crawford, J. T., & Cooksey, R. C. (2003). *ethA*, *inhA*, and *katG* Loci of Ethionamide-Resistant Clinical *Mycobacterium tuberculosis* Isolates. *Antimicrobial Agents and Chemotherapy*, *47*(12), 3799–3805. https://doi.org/10.1128/AAC.47.12.3799-3805.2003

- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

- Murray, C. J. L. (1996). Results of directly observed short-course chemotherapy in 112 842 Chinese patients with smear-positive tuberculosis. *Lancet*, *347*(8998), 358–362. https://doi.org/10.1016/S0140-6736(96)90537-1

- Nguyen, M., Wesley Long, S., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., Tyson, G. H., Zhao, S., & Davisa, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology*, *57*(2), 1–15. https://doi.org/10.1128/JCM.01260-18

- Niu, C., Yu, D., Wang, Y., Ren, H., Jin, Y., Zhou, W., Li, B., Cheng, Y., Yue, J., Gao, Z., & Liang, L. (2013). Common and pathogen-specific virulence factors are different in function and structure. *Virulence*, *4*(6), 473–482. https://doi.org/10.4161/viru.25730

- O'Neill, G. L., Murchan, S., Gil-Setas, A., & Aucken, H. M. (2001). Identification and characterization of phage variants of a strain of epidemic methicillin-resistant *Staphylococcus aureus* (EMRSA-15). *Journal of Clinical Microbiology*, *39*(4), 1540–1548. https://doi.org/10.1128/JCM.39.4.1540-1548.2001

- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x

- Palomino, J. C., Martin, A., Camacho, M., Guerra, H., Swings, J., & Portaels, F. (2002). Resazurin microtiter assay plate: simple and inexpensive method for detection of drug resistance in *Mycobacterium tuberculosis*. Antimicrobial agents and chemotherapy, 46(8), 2720–2722. https://doi.org/10.1128/AAC.46.8.2720-2722.2002

- Quainoo, S., Coolen, J. P. M., Sacha A. F. T. van Hijum, C., Martijn A. Huynen, c W. J. G. M., Willem van Schaik, E., & Wertheim, H. F. L. (2017). Whole-Genome Sequencing of Bacterial Pathogens : the Future of Nosocomial. *Clinical Microbiology Reviews*, *30*(4), 1015–1064.

- Raabe, V. N., & Shane, A. L. (2019). Group b streptococcus (*Streptococcus agalactiae*). *Gram-Positive Pathogens*, *7*(2), 228–238. https://doi.org/10.1128/9781683670131.ch14

- Rajagopal, L. (2009). Understanding the regulation of Group B Streptococcal virulence factors. *Future Microbiology*, *4*(2), 201–221. https://doi.org/10.2217/17460913.4.2.201

- Rau, A., Marot, G., & Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, *15*(1), 1–10. https://doi.org/10.1186/1471-2105-15-91

- Ravenni, N., Rota-Stabelli, O., Venturini, E., Montagnani, C., Galli, L., Armanini, F., Dolce, D., Mengoni, A., Pasolli, E., Taccetti, G., Campana, S., Manara, S., Grandi, G., Segata, N., & Asnicar, F. (2018). Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of *Staphylococcus aureus* strains in a paediatric hospital. *Genome Medicine*, *10*(1), 1–19.

- Read, T. D., Petit, R. A., Yin, Z., Montgomery, T., McNulty, M. C., & David, M. Z. (2018). USA300 S*taphylococcus aureus* persists on multiple body sites following an infection. *BMC Microbiology*, *18*(1), 1–12. https://doi.org/10.1186/s12866-018-1336-z

- Risse, J., Thomson, M., Patrick, S., Blakely, G., Koutsovoulos, G., Blaxter, M., & Watson, M. (2015). A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *GigaScience*, *4*(1), 1–7. https://doi.org/10.1186/s13742-015-0101-6

- Rivera, F. E., Miller, H. K., Kolar, S. L., Stevens, S. M., Jr, & Shaw, L. N. (2012). The impact of CodY on virulence determinant production in community-associated methicillin-resistant *Staphylococcus aureus*. *Proteomics*, 12(2), 263–268. https://doi.org/10.1002/pmic.201100298

- Sára, M., & Sleytr, U. B. (1996). Crystalline bacterial cell surface layers (S-layers): from cell structure to biomimetics. *Progress in biophysics and molecular biology*, 65(1-2), 83–111. https://doi.org/10.1016/s0079-6107(96)00007-7

- Schaefers, M. M., Breshears, L. M., Anderson, M. J., Lin, Y. C., Grill, A. E., Panyam, J., Southern, P. J., Schlievert, P. M., & Peterson, M. L. (2012). Epithelial proinflammatory response and curcumin-mediated protection from staphylococcal toxic shock syndrome toxin-1. *PLoS ONE*, *7*(3), 1–9. https://doi.org/10.1371/journal.pone.0032813

- Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

- Sepehri, Z., Mirzaei, N., Sargazi, A., Sargazi, A., Mishkar, A. P., Kiani, Z., Oskoee, H. O., Arefi, D., & Ghavami, S. (2017). Essential and toxic metals in serum of individuals with active pulmonary tuberculosis in an endemic region. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, *6*, 8–13. https://doi.org/10.1016/j.jctube.2017.01.001

- Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C. C., & Chain, P. S. G. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific Reports*, *10*(1), 1–15. https://doi.org/10.1038/s41598-020-58356-1

- Shopsin, B., Eaton, C., Wasserman, G. A., Mathema, B., Adhikari, R. P., Agolory, S., Altman, D. R., Holzman, R. S., Kreiswirth, B. N., & Novick, R. P. (2010). Mutations in agr do not persist in natural populations of methicillin-resistant *Staphylococcus aureus*. *The Journal of infectious diseases*, 202(10), 1593–1599. https://doi.org/10.1086/656915

- Sonenshein, A. L. (2005). CodY, a global regulator of stationary phase and virulence in Gram-positive bacteria. *Current Opinion in Microbiology*, *8*(2), 203–207. https://doi.org/10.1016/j.mib.2005.01.001

- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

- Stenz, L., Francois, P., Whiteson, K., Wolz, C., Linder, P., & Schrenzel, J. (2011). The CodY pleiotropic repressor controls virulence in gram-positive pathogens. *FEMS Immunology and Medical Microbiology*, *62*(2), 123–139. https://doi.org/10.1111/j.1574-695X.2011.00812.x

- Sunuwar, J., & Azad, R. K. (2021). A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains. *Briefings in Bioinformatics*, *22*(6), 1–8. https://doi.org/10.1093/bib/bbab179

- Sylvestre, P., Couture-Tosi, E., & Mock, M. (2003). Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. *Journal of Bacteriology*, *185*(5), 1555–1563. https://doi.org/10.1128/JB.185.5.1555-1563.2003

- Tacconelli, E., & Pezzani, M. D. (2019). Public health burden of antimicrobial resistance in Europe. *The Lancet Infectious Diseases*, *19*(1), 4–6. https://doi.org/10.1016/S1473-3099(18)30648-0

- Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*(1), 33–36. https://doi.org/10.1093/nar/28.1.33

- Tola, H. H., Tol, A., Shojaeizadeh, D., & Garmaroudi, G. (2015). Tuberculosis treatment non-adherence and lost to follow up among TB patients with or without HIV in developing countries: A systematic review. *Iranian Journal of Public Health*, *44*(1), 1–11.

- Tong, S. Y. C., Davis, J. S., Eichenberger, E., Holland, T. L., & Fowler, V. G. (2015). *Staphylococcus aureus* infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clinical Microbiology Reviews*, *28*(3), 603–661. https://doi.org/10.1128/CMR.00134-14

- Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K.

S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

- Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., & Malorny, B. (2020). Typing methods based on whole genome sequencing data. *One Health Outlook*, *2*(1), 1–19. https://doi.org/10.1186/s42522-020-0010-1

- Uplekar, S., Rougemont, J., Cole, S. T., & Sala, C. (2013). High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis. Nucleic Acids Research*, *41*(2), 961–977. https://doi.org/10.1093/nar/gks1260

- Valsesia, G., Rossi, M., Bertschy, S., & Pfyffer, G. E. (2010). Emergence of SCC *mec* Type IV and SCC *mec* Type V Methicillin-Resistant *Staphylococcus aureus* Containing the Panton-Valentine Leukocidin Genes in a Large Academic Teaching Hospital in Central Switzerland: External Invaders or Persisting Circulators? *Journal of Clinical Microbiology*, *48*(3), 720–727. https://doi.org/10.1128/JCM.01890-09

- Van Hoek, A. H. A. M., Mevius, D., Guerra, B., Mullany, P., Roberts, A. P., & Aarts, H. J. M. (2011). Acquired antibiotic resistance genes: An overview. *Frontiers in Microbiology*, *2*(SEP), 1–27. https://doi.org/10.3389/fmicb.2011.00203

- Van Wamel, W. J. B., Rooijakkers, S. H. M., Ruyken, M., Van Kessel, K. P. M., & Van Strijp, J. A. G. (2006). The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of *Staphylococcus aureus* are located on β-hemolysin-converting bacteriophages. *Journal of Bacteriology*, *188*(4), 1310–1315. https://doi.org/10.1128/JB.188.4.1310-1315.2006

- Vilchèze, C., Hartman, T., Weinrick, B., Jain, P., Weisbrod, T. R., Leung, L. W., Freundlich, J. S., & Jacobs, W. R. (2017). Enhanced respiration prevents drug tolerance and drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), 4495–4500. https://doi.org/10.1073/pnas.1704376114

- Vogler, A. J., Busch, J. D., Percy-Fine, S., Tipton-Hunton, C., Smith, K. L., & Keim, P. (2002). Molecular analysis of rifampin resistance in *Bacillus anthracis* and *Bacillus cereus*. *Antimicrobial Agents and Chemotherapy*, *46*(2), 511–513. https://doi.org/10.1128/AAC.46.2.511-513.2002

- Vysakh, P. R., & Jeya, M. (2013). A comparative analysis of community acquired and hospital acquired methicillin resistant *Staphylococcus aureus. Journal of Clinical and Diagnostic Research*, *7*(7), 1339–1342. https://doi.org/10.7860/JCDR/2013/5302.3139

- Wang, Y. T., Missiakas, D., & Schneewind, O. (2014). GneZ, a UDP-GlcNAc 2-epimerase, is required for s-layer assembly and vegetative growth of *Bacillus anthracis*. *Journal of Bacteriology*, *196*(16), 2969–2978. https://doi.org/10.1128/JB.01829-14

- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484

- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature biotechnology, 37(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

- World Health Organization. Global Tuberculosis report 1991. Geneva: WHO; 2021.

- Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6). https://doi.org/10.1371/journal.pcbi.1005595

- Wright, G. D. (2010). Q&A: Antibiotic resistance: Where does it come from and what can we do about it? *BMC Biology*, *8*. https://doi.org/10.1186/1741-7007-8-123

- Zampieri, M., Szappanos, B., Buchieri, M. V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., Papp, B., & Sauer, U. (2018). High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Science Translational Medicine*, *10*(429). https://doi.org/10.1126/scitranslmed.aal3973