Pathogens **2014**, *3*, 93-108; doi:10.3390/pathogens3010093

OPEN ACCESS

pathogens

ISSN 2076-0817

www.mdpi.com/journal/pathogens

Article

Live Genomics for Pathogen Monitoring in Public Health

Giuseppe D'Auria [1,2,*], Maria Victoria Schneider [3] and Andrés Moya [1,2,4]

[1] Área de Genómica y Salud, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO-Salud Pública), Avenida de Cataluña 21, 46020 Valencia, Spain; E-Mail: andres.moya@uv.es

[2] CIBER en Epidemiología y Salud Pública (CIBEResp), C/ Melchor Fernandez Almagro 3-5, Madrid, Spain

[3] The Genome Analysis Centre, Norwich Research Park, Norwich, UK; E-Mail: vicky.sg@tgac.ac.uk

[4] Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universitat de València, C / Catedrático José Beltrán 2, 46980 Paterna-Valencia, Spain

* Author to whom correspondence should be addressed; E-Mail: dauria_giu@gva.es; Tel.: +34-961-92-5929; Fax: +34-961-92-5703.

Received: 26 September 2013; in revised form: 16 December 2013 / Accepted: 7 January 2014 / Published: 21 January 2014

Abstract: Whole genome analysis based on next generation sequencing (NGS) now represents an affordable framework in public health systems. Robust analytical pipelines of genomic data provides in a short lapse of time (hours) information about taxonomy, comparative genomics (pan-genome) and single polymorphisms profiles. Pathogenic organisms of interest can be tracked at the genomic level, allowing monitoring at one-time several variables including: epidemiology, pathogenicity, resistance to antibiotics, virulence, persistence factors, mobile elements and adaptation features. Such information can be obtained not only at large spectra, but also at the "local" level, such as in the event of a recurrent or emergency outbreak. This paper reviews the state of the art in infection diagnostics in the context of modern NGS methodologies. We describe how actuation protocols in a public health environment will benefit from a "streaming approach" (pipeline). Such pipeline would include NGS data quality assessment, data mining for comparative analysis, searching differential genetic features, such as virulence, resistance persistence factors and mutation profiles (SNPs and InDels) and formatted "comprehensible" results. Such analytical protocols will enable a quick response to the needs of locally circumscribed outbreaks, providing information on the causes of

resistance and genetic tracking elements for rapid detection, and monitoring actuations for present and future occurrences.

Keywords: pathogens outbreaks; pan-genome; comparative genomics; bioinformatics; resistance; public health

1. Introduction

1.1. Following Microbes in Public Health Microbiology

Care units such as oncology and surgery, where patients are in most cases under conditions of immunodepression, are known for the presence of "the usual suspects" such as multi resistant *Pseudomonas aeruginosa*, *Escherichia coli* ESBL and *Staphylococcus aureus* MRSA. These are often the last obstacle to the clinical evolution of the patients. Around 0.1% of patients suffer sepsis every year whereas 20%–40% of these die in hospital. Without a deep knowledge of the organisms causing the sepsis, empirical antibiotic treatment is the first practice applied to stop the infection [1]. Guidelines for empirical therapy have to take into account the epidemiology of microbes isolated in care units. Intensive care units and hospitals are major reservoirs for pathogenic opportunistic organisms. Succeeding in eradicating the infection is mainly a race against time coupled with the selection of the proper empirical antibiotic treatment and the capabilities of bacteria in exchanging or evolving a variety of factors such as resistance, virulence and persistence [2]. Although the wide use of antibiotics contributed to the eradication of many diseases, continuous changes in trends of antibiotic resistance are observed [3]. With the advent of NGS the challenge is now to provide the public health sector with tools for fast and robust characterisation of pathogenic organisms, particularly for those cases in which difficulties in eradication emerge.

Resistant bacteria can emerge by a selective process in a particular population, fixing mutation conferring antibiotic resistance or by colonisation or infection with drug-resistant organisms already present in the surrounding environment [4]. Antibiotic susceptibility is routinely tested in *in vitro* assays after obtaining the isolates. The common accepted method to test antibiotic susceptibility is still officially based on minimum inhibitory concentration. Break points for antimicrobial susceptibility are periodically revised by specific organisms, for more information see "The European Committee on Antimicrobial Susceptibility Testing" [5]. In intensive care units, the first data about bacteria susceptibility are provided within 48 h. This time frame allows antibiotic treatment adjustments to eventually be made. Several cases showed that the emergence of antibiotic resistant bacteria leads to ineffective treatments [6,7]. Rapid antibiotic resistance profiles are thought to highly improve the quality of therapies, reducing, on the other hand, side effects such as commensal over-infections or the generation of new resistant strains [8]. Exceptions where cohort based studies do not show a significant association between the application of the appropriate empiric antimicrobial therapy and in-hospital post-infection length of stay or mortality have also been observed [9,10]. However, when eradication of a given infection is delayed or when an outbreak is extended, a more detailed gathering of information can help in resolving the emergency.

1.2. Comparative Genomics in Public Health

Modern public health microbiology laboratories have means to isolate strains of interest, since most of the micro-organisms under surveillance can be accounted for with standardised isolation methodologies. At present, it is possible to store, organize and maintain the organism's genetic/genomic data and its associated metadata in bio-banks [11]. Historically, public health microbiology units base their daily work on classical microbiology practices. Several metabolite based kits and gene PCR-based systems allow identification with good approximation of the presence of specific and/or most common groups of bacteria. Current trends show how some of these extend to also recognising specific resistance factors [12–14]. These kits can detect the presence of the bacteria or the antibiotic resistance factor directly from blood avoiding the time lapse for cultivation. The limitation lies in the lack of continuity in updating the recognition power in terms of new organisms or resistance factors appearing, not only in terms of time, but also in terms of geographic spaces. Ideally, a new outbreak or a failure in antibiotic treatment is due to a change in microbial resistance or persistence profiles. This should be considered as a temporal and local (geographical) factor. Commonly observed is the appearance of dangerous nosocomial outbreaks which are geographically or temporally defined [15–17]. For these, deeper studies based on the whole genome sequencing causing the outbreak will be suitable for gaining insights into the infection and resistance mechanism adopted by the pathogen. Whole genome based monitoring during outbreaks as in the case of *Legionella pneumophila*, *Mycobacterium tuberculosis*, *E. coli*, *etc.* point out the importance of working with whole genome data in a comparative framework, highlighting taxonomy relationships, mutation based clustering, orthologues and accessory genes distribution. The latter is considered to be a main source for resistance, virulence and persistence factors [18–21].

Nowadays, whole genome sequencing protocols are relatively easy to apply in order to solve daily problems. Probably, the most pragmatic approach relies on the proposal of surveillance units of candidate organisms to be sequenced. When this actuation plan starts, the same microbiology unit can proceed through DNA extraction of the target organism and easily pass the DNA sample(s) to a specialised service for sequencing through the most appropriate NGS methods. We have to keep in mind that an alert from a surveillance unit can justify the expense for a whole genome sequencing project, the costs for which are predicted to be continuously reduced due to the further development of sequencing technologies. When the sequencing unit returns the obtained sequences, these can flow through pipelines for data mining and extraction of the information required by surveillance and microbiology units for further decisions and actuations.

In this paper, we describe a "live" frame-work in microbial genome sequencing in which data mining from comparative genomics continuously populate a relational database with genetic information, allowing the extraction of useful differential data of interest such as virulence, resistance persistence factors, SNPs and InDels. Such approaches make data consequently suitable for immediate designing of new diagnostics systems for early predictions of organisms bringing potential dangerous features. Future advances in public health microbiology would entail services based on provision of large scale comparative analysis of organisms belonging to the same species.

For instance, the Global Microbial Identifier (GMI) initiative has started work in this direction promoting whole genome sequencing of organisms with public health relevance. GMI aims to store

data, compare genomes and identify genes which could be of interest in outbreak characterisation as well as describing emerging pathogens [22].

Theoretical and technological advances are ready to support all processes from strain collection, DNA purification and storing. We are at the turn of microbiology and genomics research towards providing multiple genomic information from most similar organisms, analysing commonly shared features (or core genome) and additional characters (or disposable genome), in other words, we are in the era of the "pan-genome".

1.3. Approaching Species Definition in the Genomics Era

One of the most discussed issues in microbiology is the species definition in taxonomy. Since the early 1970s, molecular methods and DNA sequencing techniques have been adopted to provide objective criteria in defining bacterial species. At first, whole genome DNA-DNA hybridisation was used to determine when two strains were homologous. Johnson (1973) [23] determined that strains from the same species nearly always shared 70% or more of their genomes. However, variation in gene content and the presence of polymorphisms among strains assigned by DNA-DNA hybridization to the same species have led some to consider that such a "species concept" is far too broad, compared to those in organisms of higher complexity than bacteria [24]. Other methods have been proposed to define microbial species based on diverse cut-off values studying one or more genes. 16S taxonomy, Multi Locus Sequence Typing (MLS or MLST) and biochemical characterisations are some of the most accepted classification methods [25,26]. Whereas none of these methods describe the complete genetic repertoire, whole genome analysis represents the gold standard in comparative genomics always permitting *a-posteriori* single or multiple gene-based characterisations [27]. Pathogenic organisms of interest could be tracked at the genomic level, monitoring at the same time its expansion, pathogenicity, resistance to antibiotics, virulence and persistence factors, mobile elements, adaptation features, all in a geographic context. Though several reference genomes are available, it is worth mentioning the efforts needed for sequencing and assembling *de-novo* without a reference backbone [28,29]. The American Center for Disease Control and Prevention (CDC) promotes whole genome sequencing for real time epidemiology, where results are expected to provide prospective information about outbreak evolution. The European Centre for Disease Prevention and Control (ECDC) is working in this direction, focusing on *how public health can benefit from the rapidly evolving NGS technology in molecular microbiology* [30]. While more and more reference genomes are available, comparative analysis represents a must in public health microbiology, offering the possibility to discover differential traits involved in the pathogenicity of a given organism.

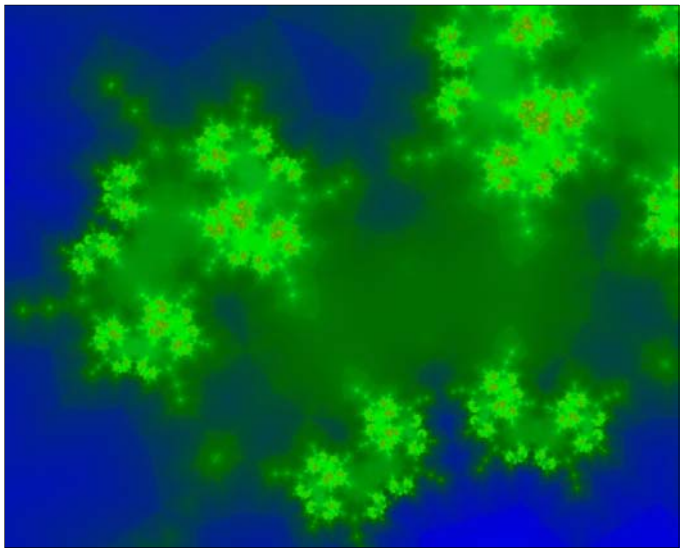
2. Pan-Genome

Tracking and comparing genomes entails studying of the genomic difference among compared strains. The term "pan-genome" refers to *pan* (from Greek παν, whole) and *genome* (genome) referring to the inclusion of the core and the dispensable genome [31,32]. While originally microbial expansion was thought as clonal, now it is well known that also the overnight culture bacteria go through substitutions, insertions and deletions due to polymerases errors, genomic rearrangements (e.g. mobile elements) and viral interactions. Bacteria isolated from the same environment can show important

genomic differences due to the accumulation of variations during their natural life cycles [33]. The differences in accumulation of variations are attributed to the different speeds and scales for all organisms in a given environment as a "*fractal pattern*" in evolution (Figure 1).

This continuous evolution process, jointly with the genetic enrichment provided by horizontal gene transfer events, prevent the genome of a bacteria isolated from a given environment overlapping with other genomes of other organisms defined as being taxonomically the same (in terms of 16S rDNA, MLST or DNA/DNA hybridisation profiles), which may have been isolated at different times or from different locations [34,35]. Central metabolic genes are mostly found as orthologues in all genomes of bacteria belonging to the same taxa, (common genome, shared genome or, most used, core-genome). Additional genes (dispensable genome) seem to be the ones making the difference [36]. For instance, antibiotic resistance factors, toxin/anti-toxin systems or phage-resistance clusters are considered in bacteria as arsenals for maintenance, evolution and transferring, making them a challenge in public health due to how difficult they are to track and control [37–40].

Figure 1. Fractal evolution model. Artwork describing ideally a fractal evolution model showing the outcome of new offspring (light green islands) with fitness advantages which are fixed and explode (darker islands nuclei), although the evolutionary pattern is maintained (fractal periodicity). On the other hand, small generations could be slower in their evolution or disappear in the time lapse. This model is established by representing a continuum among all organisms inhabiting a given environment.



2.1. Core Genome

By analysing bacteria belonging to the same species at a whole genome level, comprehensive comparative genomics can be carried out. Shared genes among multiple strains are mostly related to house-keeping genes or central metabolic processes, most of the structural information and main genotypic features. The core genome could be thought as the number of shared features in a pool of genomes. The size of the core genome decreases, increasing the number of genes added to the pool. While it is possible to link the core genome to common tracts among considered bacteria, it is worth

mentioning that such a calculation also depends technically on the number of genomes available for the computation. In this review, we propose a comparative analysis example of some organisms considering all genomes currently deposited in GenBank of *Campylobacter jejuni* (nine genomes), *Streptococcus suis* (13 genomes), *L. pneumophila* (10 strains), and *Staphylococcus aureus* (31 genomes, see SI1 for a strain list).

2.2. Dispensable-Genome

As long as new genes (by definition as part of dispensable compartment of the pan-genome) are added in the computation to the pool, the volume of the pan-genome increases. The increase of the pan-genome size has been observed to be either faster or slower. This has created another concept of the open or closed pan-genome [41]. A species pan-genome is considered closed, when as many new genomes are added to the pool and no new genetic information appears. For example, *Streptococcus agalactiae* pan-genome can exceed at least three-fold the average genome size [32]. On the other hand, highly adapted bacteria, especially those characterised by living in very restricted environmental niches, such as host specific pathogens such as *Salmonella paratyphi* or *Bacillus anthracis,* show a closed pan-genome [36,42].

The dispensable genome, also defined as "accessory" or "adaptive genome" [32], includes genes conferring adaptive advantages to the strain in order to survive in a specific environment. In most cases, these factors are linked to antibiotic resistance, virulence, capsular serotype, adaptation, and might reflect the organisms predominant lifestyle [40,41]. Being aware of differential traits in terms of presence/absence of genes and their annotations means having knowledge about the versatility or pathogenicity of a given organism separate from the pure taxonomic position. A recent study by den Bakker and collaborators (2011) offers a good example of comparative genomics applied to the identification of evolutionary clades of *S. enterica* subsp. *enterica.* Here, the authors provide a population genetic framework for studying the virulence and propagation of this pathogen. In their work, 46 complete genomes of *S. enterica*, 16 new genomes sequenced using SOLiD^{TM} system and 30 genomes already present in GenBank were analysed. *S. enterica*'s pan-genome was calculated, and common and different genomic traits were spotted by identifying in the two clades what differed in terms of metabolic capabilities, adhesion and colonization properties. Studying two clades of *S. enterica* subsp. en*terica* at the level of its pan-genome highlighted the existence of conserved pathogenicity islands and a virulence gene repertoire [43].

In 2010, D'Auria *et al.* described the pan-genome of *L. pneumophila* (five genomes at that time) revealing strain-specific and common traits including anti-drug resistance systems; a system for transport and secretion of heavy metals; three systems related to DNA transfer; two CRISPR systems, known to provide resistance against phage infections; and seven islands of phage-related proteins, five of which seem to be strain-specific and two shared among compared genomes [40].

3. SNPs/InDels Profiles

Whole genome sequencing allows having at one time the whole SNPs and InDels profiles for each gene in a multi-genome context [44]. Whole genome taxonomy was reviewed in depth by Rannala and Yang in 2008 [45]. Phenotype identification and genotypic typing techniques were mentioned as the

basis for infectious disease epidemiology, providing profiles that are of use, not only for taxonomic reconstruction, but also for tracking strains during outbreaks [46]. The *Yersinia pestis* plague, characterised by a genetic uniformity, made it easy to be traced at the global level. Multi-genome SNPs profiles allowed to define its origin in or around China migrating from East to West. SNPs lineages were then traced highlighting the radiation to Europe, Africa and South-east Asia, while North American radiation originated from a single point [47]. The availability of such fine tools for lineage tracking is of great interest. In a public health frame-work, it is probably not useful, nor possible, to sequence all genomes of organisms from an outbreak episode, but having the genome of some representative strain would help towards obtaining a deeper knowledge about the factors responsible for resistance complicating infection eradication. While data mining processes allow the identification of statistically relevant SNPs/InDels, applied science will provide the basis to develop new tracking PCR-based systems. PCR detection systems have been successfully applied with several genes of interest and are often included in commercial test kits [48–50]. In this context, a "live" or continuously growing SNPs/InDels database would highly contribute to the prediction of polymorphisms suitable for test kits development, not only with globally applicable purposes, but even more interestingly, at small local scale (hospitals, villages, city, *etc.*). In other words, having polymorphisms' profiles for bacteria of interest belonging to a specific outbreak would help to elaborate a short time test for its immediate identification and tracking.

4. Automatic Pipelines

While it seems impossible to afford the sequencing and perform the bioinformatics tasks in a routine daily work time frame for a microbiology unit in public health, automatic pipelines represent the solution. These will help automatising several of the necessary steps and tasks providing the user with the final data required in a round time compatible with the work beat. While the number of software for NGS data analysis is continuously growing [51–53], existing pipelines can be integrated and extended according to the microbiology unit needs. In terms of reducing human intervention, pipelines can be designed to cover as much of the analytical tasks as possible, not only for quality assessment and ancillary data production, but also for data mining and visualisation. Figure 2 reports a schematic pipeline for NGS data production and data mining, starting from sequencing of organisms of interest to gathering of useful data.

Among the data which a microbiology unit needs to know in a "live" context, we also suggest including properties linked to genes or to specific mutations, in other words, the pan-genome relationship of a given isolate within its species and the mutational (SNPs/InDels) profile. This kind of automatic pipeline provides the user, in a short frame of time, with differential genes data (dispensable genome) as well as with mutation profiles which allow positioning of the studied strain in a phylogenetic context.

In our example, an automatic pipeline was applied to simulated NGS data obtained using Illumina error profiles [54] on complete genomes of *Campylobacter jejuni*, *Streptococcus suis*, *L. pneumophila* and *S. aureus* strains (Table 1 and SI1).

The applied pipeline starts from sequence data going step-by-step from data cleaning and quality assessment through the production of useful intermediate data characterising, on one hand gene-related data in a pan-genomic context, and on the other hand, the mutation profiles.

Figure 2. Automatic data mining pipeline schema. Picture shows proposed schema from strains selection to data reading for actuation plans. All steps over blue path are automatic and do not need user supervision.

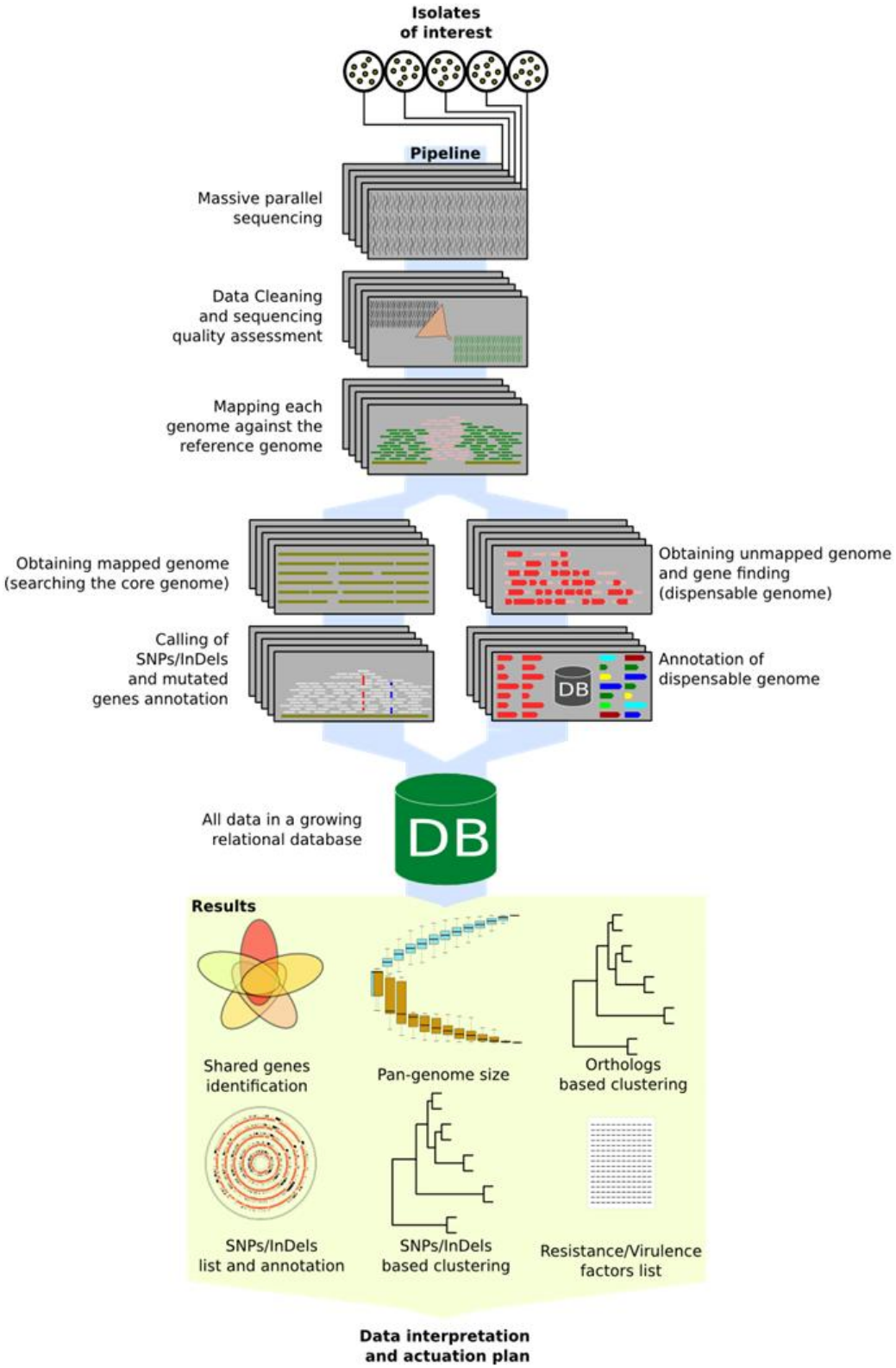


Table 1. Used genomes and main genomic features. SI1 reports accession numbers for each genome of each species. Orthologues' assignation to define pan-genome size was carried out by clustering all genes of each organism dataset with at least 60% similarity (amino acid) and 60% sequences overlapping using CD-HIT program [55].

	C. jejuni	***S. suis***	***L. pneumophila***	***S. aureus***
Number of genomes	9	13	10	31
Average genome length	1,678,553.8 bp	2,090,478.5 bp	3,302,389.7 bp	2,894,586.7 bp
Pan-genome size	953	1,125	1,933	1,765

4.1. Pipeline—A Little Bioinformatics

Pipelines are thought to reduce human intervention, maintaining analysis as robust and reproducible. Almost all of the necessary steps can be run consecutively one after the other, feeding the next step with the output of the previous one. The applied pipeline goes through several concatenated steps through the quality assessment process using "PRINSEQ" program [56], mapping using "SMALT" program [57], consensus definition and data conversions by the use of SAMtools [58], *de-novo* assembly of unmapped reads by MIRA program [59], gene finding using GLIMMER3 [60], automatic annotation using Hidden Markov Models algorithm searching in PFAM database [61–63], SNPs/IndDels calling by the use of VarScan program [64], and almost all required statistics have been performed in R environment [65,66]. The actual data report can be obtained using appropriate software which allows script execution and which could be concatenated with data-mining steps.

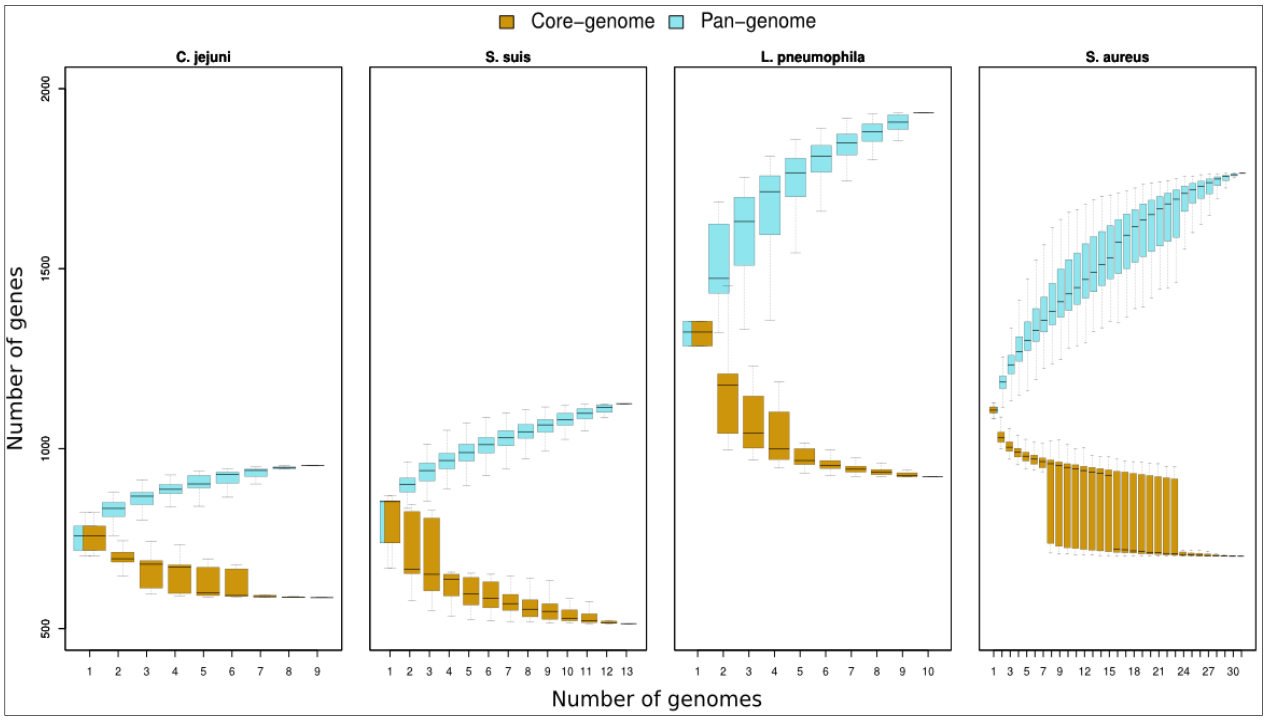
4.2. Pan-genome Data Mining Pipeline

The interest in pan-genomic data relies on discovering additional features of a given organism. In this context, a closed or open pan-genome can provide an idea about the versatility of the studied strain. Generally, we face organisms with a reduced ecological niche such as *C. jejuni* (Figure 3, left panels), where the increase in pan-genome size slows down within the nine genomes considered (almost closed pan-genome) or with organisms such as *L. pneumophila* or *S. aureus* (Figure 3, right panels), inclined to horizontal gene transfer and characterised by having an open pan-genome.

Further, orthologue distributions provide the user with gene profiles for each organism, identifying differential tracts among organisms. When the mapping step of the pipeline runs by identifying reads overlapping with reference genome, unmapped reads can be tracked and associated to parts of the genome not present in the reference genome. These reads can be *de-novo* assembled and annotated, contributing towards revealing what makes the strain special with respect to previous known genomes. Annotating differential genes highlights the presence of pathogenicity factors. Clustering organisms by their differences in genes presence/absence and frequency allows stratifying annotation information, in terms of differences in versatility and pathogenicity. Figure 4 shows orthologues based dendrograms for the four species used as examples. Below each dendrogram, blue marks indicate some of the differential gene features related to antibiotics resistance factors highlighted by the data mining pipeline. For instance, several strains among all genomes reported resistance factors to betalactamases with different

mechanisms that only a fast comparative analysis can bring to light. SI2 reports the complete table of differential genes encountered among all considered strains of the four pan-genomes.

Figure 3. Core genome distributions. Graph shows boxplots of pan-genome and core genome contents for increasing pools of genomes belonging to *C. jejuni*, *S. suis*, *L. pneumophila* and *S. aureus*. In *S. aureus* graph, we considered 31 genomes; due to the elevated number of possible combinations of genomes pools from n = 8 to n = 23 the boxes describe sampling of 2000 random combinations.

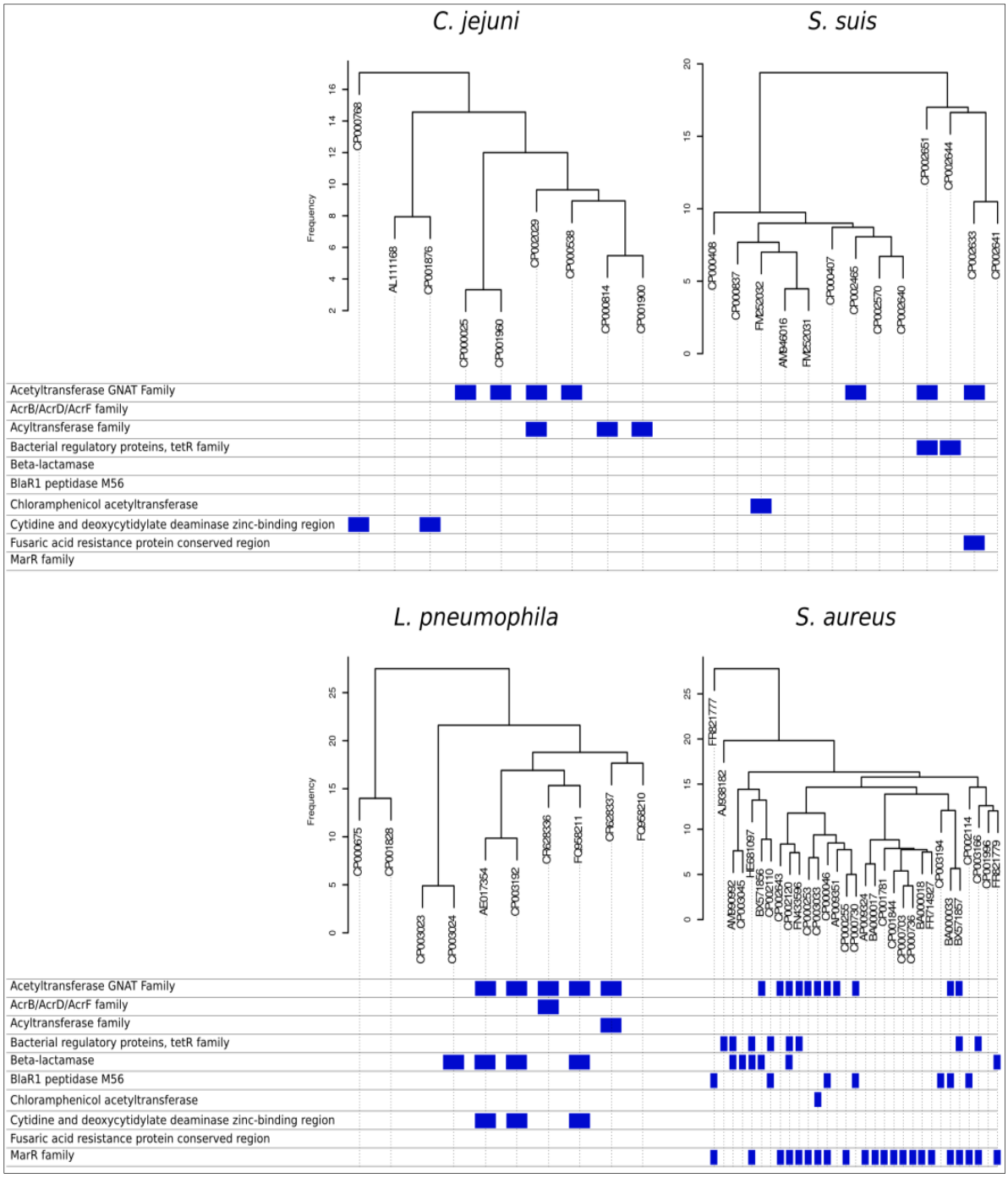


5. Conclusions

NGS coupled with automatic data mining pipelines represent nowadays the future for promptly definition of actuation plans responding to outbreaks or recurrent infections in public health systems. This framework is even more useful in small scale situations where a specific and proper action is needed. The speed and robustness of NGS methodologies and strategies now make possible the production of genetic and mutation profiles within a couple of days, making the automatic data mining process compatible with the immediate need of information in emergency cases. In the case of genetic data, differential gene profiles obtained comparing the strain(s) of interest with the ones already present in databases allows the definition of its pan-genome, thus data mining pipelines can reveal to the users which genes are making the difference in terms of antibiotic resistance or environmental persistence factors. Moreover, mutation profiles provide users with the correct information for taxonomic identification of the proposed strain on a clonal scale as well as with possible targets for the fast development of tracking systems kits, based, for example, on PCR methods. An important technical challenge is surely represented by the data storage and data mining process in such a growing frame-work. The experience from information and communication technology will probably be of help

in optimising automatic pipeline processes and data flows avoiding redundancies and focussing the attention mainly on differential tracts labelling and flagging. Finally, having these data in relational databases interconnected among multiple centres of public health system in the future will be key to creating the "live" context for modern genome-based microbiology.

Figure 4. Antibiotic resistance factors. Dendrograms shows distribution among strains of *C. jejuni*, *S. suis*, *L. pneumophila* and *S. aureus* based on presence/absence of orthologues genes. For each genome, the presence of dispensable genes related with antibiotic resistance is identified, and by applying automatic pan-genome analysis, the pipeline is reported.



Acknowledgments

This work was funded by grant CP09/00049 Miguel Servet, Instituto de Salud Carlos III, Spain to GD; by projects SAF2009-13032-C02-01, SAF2012-31187 from the Spanish Ministry for Science and Innovation (MCINN), FU2008-04501-E from Spanish Ministry for Science and Innovation (MCINN) in the frame of ERA-Net PathoGenoMics, and Prometeo/2009/092 from Conselleria D'Educació Generalitat Valenciana, Spain, to AM. VS thanks The Genome Analysis Centre (TGAC, Norwich, UK) and the Biotechnology and Biological Sciences Research Council (BBSRC, UK).

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Freifeld, A.G.; Bow, E.J.; Sepkowitz, K.A.; Boeckh, M.J.; Ito, J.I.; Mullen, C.A.; Raad, I.I.; Rolston, K.V; Young, J.-A.H.; Wingard, J.R. Clinical practice guideline for the use of antimicrobial agents in neutropenic patients with cancer: 2010 update by the Infectious Diseases Society of America. *Clin. Infect. Dis.* **2011**, *52*, e56–e93.
2. Björkman, J.; Andersson, D.I. The cost of antibiotic resistance from a bacterial perspective. *Drug Resist. Updates* **2000**, *3*, 237–245.
3. McCormick, J.B. Epidemiology of emerging/re-emerging antimicrobial-resistant bacterial pathogens. *Curr. Opin. Microbiol*. **1998**, *1*, 125–129.
4. Fraimow, H.S.; Tsigrelis, C. Antimicrobial resistance in the intensive care unit: Mechanisms, epidemiology, and management of specific resistant *pathogens. Crit. Care Clin.* **2011**, *27*, 163–205.
5. The European Committee on Antimicrobial Susceptibility Testing. Available online: http://www.eucast.org/ (accessed on 26 September 2013).
6. Blot, S.I.; Vandewoude, K.H.; Hoste, E.A.; Colardyn, F.A. Outcome and attributable mortality in critically ill patients with bacteremia involving methicillin-susceptible and methicillin-resistant *Staphylococcus aureus. Arch. Intern. Med.* **2002**, *162*, 2229–2235.
7. Cosgrove, S.E.; Qi, Y.; Kaye, K.S.; Harbarth, S.; Karchmer, A.W.; Carmeli, Y. The impact of methicillin resistance in Staphylococcus aureus bacteremia on patient outcomes: mortality, length of stay, and hospital charges. *Infect. Contr. Hosp. Epidemiol.* **2005**, *26*, 166–174.
8. Barenfanger, J.; Drake, C.; Kacich, G. Clinical and financial benefits of rapid bacterial identification and antimicrobial susceptibility testing. *J. Clin. Microbiol.* **1999**, *37*, 1415–1418.
9. Osih, R.B.; McGregor, J.C.; Rich, S.E.; Moore, A.C.; Furuno, J.P.; Perencevich, E.N.; Harris, A.D. Impact of empiric antibiotic therapy on outcomes in patients with Pseudomonas aeruginosa bacteremia. *Antimicrob. Agents Chemother.* **2007**, *51*, 839–844.
10. Thom, K.A.; Schweizer, M.L.; Osih, R.B.; McGregor, J.C.; Furuno, J.P.; Perencevich, E.N.; Harris, A.D. Impact of empiric antimicrobial therapy on outcomes in patients with *Escherichia coli* and *Klebsiella pneumoniae* bacteremia: A cohort study. *BMC Infect. Dis.* **2008**, *8*, 116.

11. De Paoli, P. Bio-banking in microbiology: from sample collection to epidemiology, diagnosis and research. *FEMS Microbial. Rev.* **2005**, *29*, 897–910.
12. Mencacci, A.; Leli, C.; Cardaccia, A.; Meucci, M.; Moretti, A.; D'Alò, F.; Farinelli, S.; Pagliochini, R.; Barcaccia, M.; Bistoni, F. Procalcitonin predicts real-time PCR results in blood samples from patients with suspected sepsis. *PloS One* **2012**, *7*, e53279.
13. Lehmann, L.E.; Hunfeld, K.P.; Emrich, T.; Haberhausen, G.; Wissing, H.; Hoeft, A.; Stüber, F. A multiplex real-time PCR assay for rapid detection and differentiation of 25 bacterial and fungal pathogens from whole blood samples. *Med. Microbial. Immunol.* **2008**, *197*, 313–324.
14. Endimiani, A.; Hujer, K.M.; Hujer, A.M.; Kurz, S.; Jacobs, M.R.; Perlin, D.S.; Bonomo, R.A. Are we ready for novel detection methods to treat respiratory pathogens in hospital-acquired pneumonia? *Clin. Infect. Dis.* **2011**, *52 Suppl 4*, S373–S383.
15. Siširak, M.; Hukić, M. An outbreak of multidrug-resistant Serratia marcescens: The importance of continuous monitoring of nosocomial infections. *Acta medica academica* **2013**, *42*, 25–31.
16. Fernández, J.A.; López, P.; Orozco, D.; Merino, J. Clinical study of an outbreak of Legionnaire's disease in Alcoy, Southeastern Spain. *Eur. J. Clin. Microbiol. Infect. Dis.* **2002**, *21*, 729–735.
17. Warny, M.; Pepin, J.; Fang, A.; Killgore, G.; Thompson, A.; Brazier, J.; Frost, E.; McDonald, L.C. Toxin production by an emerging strain of Clostridium difficile associated with outbreaks of severe disease in North America and Europe. *Lancet* **2005**, *366*, 1079–1084.
18. Reuter, S.; Harrison, T.G.; Köser, C.U.; Ellington, M.J.; Smith, G.P.; Parkhill, J.; Peacock, S.J.; Bentley, S.D.; Török, M.E. A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. *BMJ Open* **2013**, *3, e002175*.
19. Gardy, J.L.; Johnston, J.C.; Ho Sui, S.J.; Cook, V.J.; Shah, L.; Brodkin, E.; Rempel, S.; Moore, R.; Zhao, Y.; Holt, R.; *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New Engl. J. Med.* **2011**, *364*, 730–739.
20. Sabat, A.J.; Budimir, A.; Nashev, D.; Sá-Leão, R.; van Dijl, J.M; Laurent, F.; Grundmann, H.; Friedrich, A.W. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Eurosurveillance* **2013**, *18*, 20380.
21. Rohde, H.; Qin, J.; Cui, Y.; Li, D.; Loman, N.J.; Hentschke, M.; Chen, W.; Pu, F.; Peng, Y.; Li, J.; *et al.* Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *New Engl. J. Med.* **2011**, *365*, 718–724.
22. Global Microbial Identifier. Available online: www.globalmicrobialidentifier.org (accessed on 26 September 2013).
23. Johnson, J.L. Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. *Int. J. Syst. Bacteriol* **1973**, *23*, 308–315.
24. Staley, J.T. Biodiversity: are microbial species threatened? *Curr. Opin. Biotechnol.* **1997**, *8*, 340–345.
25. Spratt, B.G. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Curr. Opin. Microbiol.* **1999**, *2*, 312–316.
26. Maiden, M.C.; Bygraves, J.A.; Feil, E.; Morelli, G.; Russell, J.E.; Urwin, R.; Zhang, Q.; Zhou, J.; Zurth, K.; Caugant, D.A.; *et al*. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 3140–3145.

27. Larsen, M.V.; Cosentino, S.; Rasmussen, S.; Friis, C.; Hasman, H.; Marvig, R.L.; Jelsbak, L.; Sicheritz-Pontén, T.; Ussery, D.W.; Aarestrup, F.M.; *et al.* Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **2012**, *50*, 1355–1361.
28. Kisand, V.; Lettieri, T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics* **2013**, *14*, 211.
29. Perkins, T.T.; Tay, C.Y.; Thirriot, F.; Marshall, B. Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes. *PloS One* **2013**, *8*, e67539.
30. Palm, D.; Johansson, K.; Ozin, A.; Friedrich, A.W.; Grundmann, H.; Larsson, J.T.; Struelens, M.J. Molecular epidemiology of human pathogens: how to translate breakthroughs into public health practice, Stockholm, November 2011. *Eurosurveillance* **2012**, *17*, 20054.
31. Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594.
32. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, S. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955.
33. Blount, Z.D.; Barrick, J.E.; Davidson, C.J.; Lenski, R.E. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* **2012**, *489*, 513–518.
34. Cuadros-Orellana, S.; Martin-Cuadrado, A.-B.; Legault, B.; D'Auria, G.; Zhaxybayeva, O.; Papke, R.T.; Rodriguez-Valera, F. Genomic plasticity in prokaryotes: The case of the square haloarchaeon. *ISME J* **2007**, *1*, 235–245.
35. Morowitz, M.J.; Denef, V.J.; Costello, E.K.; Thomas, B.C.; Poroyko, V.; Relman, D.A.; Banfield, J.F. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc. Natl. Acad. Sci. USA* **2010**, *108***,** 1128–1133.
36. Tettelin, H.; Riley, D.; Cattuto, C.; Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **2008**, *11*, 472–477.
37. Novais, A.; Comas, I.; Baquero, F.; Cantón, R.; Coque, T.M.; Moya, A.; González-Candelas, F.; Galán, J.-C. Evolutionary trajectories of beta-lactamase CTX-M-1 cluster enzymes: predicting antibiotic resistance. *PLoS Pathogens* **2010**, *6*, e1000735.
38. Sorek, R.; Kunin, V.; Hugenholtz, P. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **2008**, *6*, 181–186.
39. Geric, B.; Johnson, S.; Gerding, D.N.; Grabnar, M.; Rupnik, M. Frequency of binary toxin genes among Clostridium difficile strains that do not produce large clostridial toxins. *J. Clin. Microbiol.* **2003**, *41*, 5227–5232.
40. D'Auria, G.; Jiménez-Hernández, N.; Peris-Bondia, F.; Moya, A.; Latorre, A. Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genom.* **2010**, *11*, 181.
41. Mira, A.; Martín-Cuadrado, A.B.; D'Auria, G.; Rodríguez-Valera, F. The bacterial pan-genome:a new paradigm in microbiology. *Int. Microbiol.* **2010**, *13*, 45–57.
42. Liang, W.; Zhao, Y.; Chen, C.; Cui, X.; Yu, J.; Xiao, J.; Kan, B. Pan-genomic analysis provides insights into the genomic variation and evolution of Salmonella Paratyphi A. *PloS One* **2012**, *7*, e45346.

43. Den Bakker, H.C.; Switt, A.M.; Govoni, G.; Cummings, C.A.; Ranieri, M.L.; Degoricija, L.; Hoelzer, K.; Rodriguez-Rivera, L.D.; Brown, S.; Bolchacova, E.; *et al*. Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of Salmonella enterica. *BMC Genom.* **2011**, *12*, 425.
44. Verma, M.; Lal, D.; Kaur, J.; Saxena, A.; Kaur, J.; Anand, S.; Lal, R. Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. *Res. Microbiol.* **2013**, *164*, 718–728.
45. Rannala, B.; Yang, Z. Phylogenetic inference using whole genomes. *Annu Rev Genom Hum Genet* **2008**, *9*, 217–231.
46. Parkhill, J.; Wren, B.W. Bacterial epidemiology and biology—lessons from genome sequencing. *Genome Boil.* **2011**, *12*, 230.
47. Morelli, G.; Song, Y.; Mazzoni, C.J.; Eppinger, M.; Roumagnac, P.; Wagner, D.M.; Feldkamp, M.; Kusecek, B.; Vogler, A.J.; Li, Y.; *et al*. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat. Genet.* **2010**, *42*, 1140–1143.
48. Hudson, C.R.; Quist, C.; Lee, M.D.; Keyes, K.; Dodson, S.V.; Morales, C.; Sanchez, S.; White, D.G.; Maurer, J.J. Genetic relatedness of Salmonella isolates from nondomestic birds in Southeastern United States. *J. Clin. Microbiol*. **2000**, *38*, 1860–1865.
49. Lamoth, F.; Jaton, K.; Prod'hom, G.; Senn, L.; Bille, J.; Calandra, T.; Marchetti, O. Multiplex blood PCR in combination with blood cultures for improvement of microbiological documentation of infection in febrile neutropenia. *J. Clin. Microbiol.* **2010**, *48*, 3510–3516.
50. Gordon, D.M. Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology* **2001**, *147*, 1079–1085.
51. Zhao, Y.; Wu, J.; Yang, J.; Sun, S.; Xiao, J.; Yu, J. PGAP: Pan-genomes analysis pipeline. *Bioinformatics* **2012**, *28*, 416–418.
52. Altmann, A.; Weber, P.; Bader, D.; Preuss, M.; Binder, E.B.; Müller-Myhsok, B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* **2012**, *131*, 1541–1554.
53. Kislyuk, A.O.; Katz, L.S.; Agrawal, S.; Hagen, M.S.; Conley, A.B.; Jayaraman, P.; Nelakuditi, V.; Humphrey, J.C.; Sammons, S.A.; Govil, D.; *et al*. A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* **2010**, *26*, 1819–1826.
54. Sherman—A tool to simulate FastQ files for high-throughput sequencing experiments. Available online: http://www.bioinformatics.babraham.ac.uk/projects/sherman/ (accessed on 26 September 2013).
55. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006** *22*, 1658–1659.
56. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864.
57. Martin, J.; Sykes, S.; Young, S.; Kota, K.; Sanka, R.; Sheth, N.; Orvis, J.; Sodergren, E.; Wang, Z.; Weinstock, G.M.; *et al.* Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS One* **2012**, *7*, e36427.
58. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.

59. Chevreux, B.; Wetter, T.; Suhai, S. Genome sequence assembly using trace signals and additional sequence information. In *Proceedings of German Conference on Bioinformatics*; Hannover: Germany, 1999; pp. 45–56.
60. Delcher, A.L.; Harmon, D.; Kasif, S.; White, O.; Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*. 1999, *27*, 4636–4641.
61. Zhang, Y.; Sun, Y. HMM-FRAME: Accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics* **2011**, *12*, 198.
62. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37.
63. Sonnhammer, E.L.; Eddy, S.R.; Birney, E.; Bateman, A.; Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **1998**, *26*, 320–322.
64. Koboldt, D.C.; Chen, K.; Wylie, T.; Larson, D.E.; McLellan, M.D.; Mardis, E.R.; Weinstock, G.M.; Wilson, R.K.; Ding, L. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **2009**, *25*, 2283–2285.
65. R Core Team. R: A language and environment for statistical computing; R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0.
66. Morgan, M.; Anders, S.; Lawrence, M.; Aboyoun, P.; Pagès, H.; Gentleman, R. ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **2009**, *25*, 2607–2608.