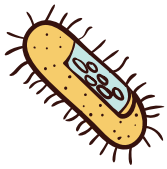
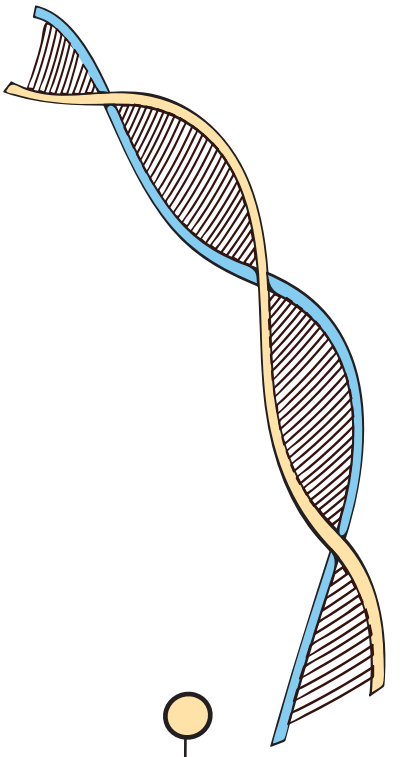
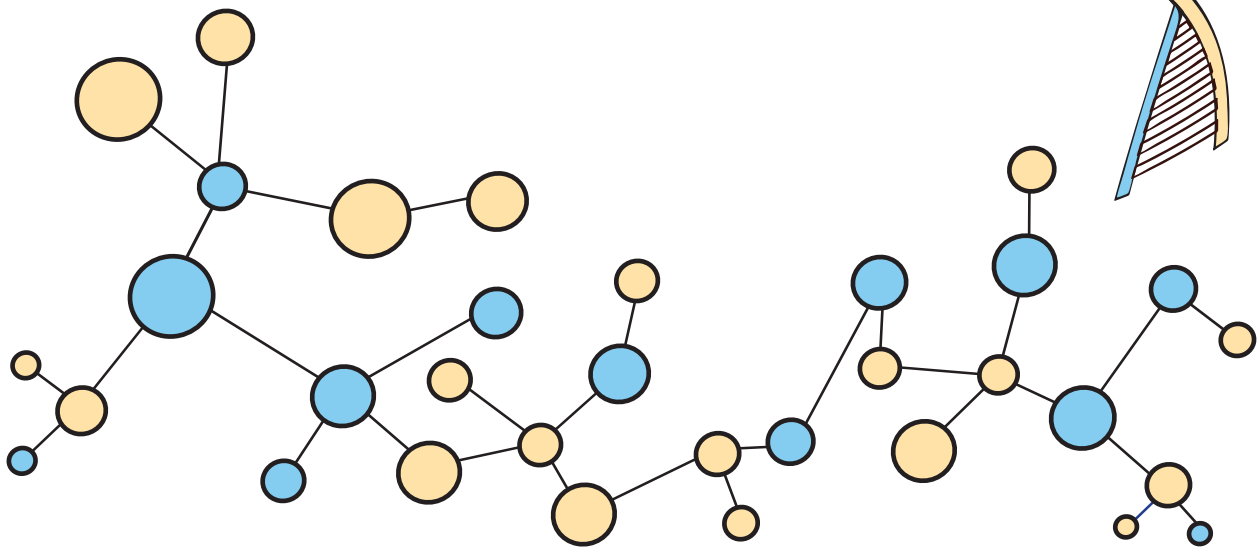
MODELING BIOLOGICAL SYSTEMS:

FROM GENOME SEQUENCES TO FUNCTIONAL INSIGHTS







LUANA PRESTA

DOTTORATO DI RICERCA IN BIOLOGIA
Curriculum Microbiologia e Genetica

CICLO XXX

Modelling biological systems: from genome sequences to functional insights

Settore Scientifico Disciplinare BIO18

Dottorando:
Luana Presta

Tutor:
Prof. Alessio Mengoni

Coordinatore:
Prof. Alberto Ugolini

Anni 2014/2017

PhD dissertation

Modelling biological systems: from genome sequences to functional insights

Luana Presta

Supervisor Professor Alessio Mengoni

October 31, 2017

Luana Presta
Modelling biological systems:
from genome sequences to functional insights
PhD dissertation, October 31, 2017
Supervisor: Professor Alessio Mengoni

University of Florence
Department of Biology
Computational Biology Group, COMBO
Via Madonna del Piano, 6
Sesto Fiorentino 50019, Firenze

Abstract

More than twenty years ago the first genome sequencing of an organism was seen as a revolution in the world of biological sciences. After a few years, Carl Woese suggested that in the long run, the real justification of genomics would have been genomic of prokaryotic microorganisms, due to the important implication for the study of biological evolution and the many biotechnological applications (spanning from medical, to agricultural, environmental and industrial). However, the challenge was then the possibility to computationally infer the biological properties of an organism on the simple basis of its genome sequence. Such challenge, still ongoing, relies on the possibility: i) to reconstruct the genome sequences of organisms from the (relatively short) sequence reads obtained on the various sequencing platforms; ii) to identify genes inside DNA sequences and assign functions; iii) to predict organisms' phenotypes. It is possible to imagine such challenges as a 1-D, 2-D and 3-D "-omics" reconstruction. In this thesis, the aim was to explore, by using specific case studies, such 1-D, 2-D and 3-D computational biology inference on prokaryotic genome sequences. Each chapter of the results section will provide data on bacterial genomes of relevant strains for various biotechnological applications. The overall results are presented according to the depth of functional (phenotypical) inference, from genome assembly and simple functional annotation to the powerful genome-scale metabolic models. The main focus will be centered to emphasize the high predictive value of genome-scale metabolic modeling for complex phenotypes and *in silico* prediction of gene essentiality. The main conclusion is concerned with an integrated use of the computational tools to assist systems biology-based computational inferences, to help whole predictive biological (genomic) sciences.

Sommario

Più di vent'anni fa il primo sequenziamento genico automatizzato di un organismo costituì una rivoluzione nel mondo delle scienze biologiche. Dopo qualche anno Carl Woese suggerì che, alla lunga, la reale giustificazione della genomica sarebbe stata la genomica dei microrganismi procarioti, a causa delle importanti implicazioni per lo studio dell'evoluzione biologica e le tante applicazioni biotecnologiche (da quelle mediche a quelle industriali, ambientali e agricole). Ad ogni modo, la sfida divenne poi la possibilità di inferire computazionalmente le proprietà biologiche di un organismo sulla sola base della sua sequenza genica. Tale sfida, ancora in corso, risiede nella possibilità i) di ricostruire le sequenze genomiche di un organismo a partire dalle (relativamente corte) reads ottentute da diverse piattaforme di sequenziamento; ii) di identificare geni all'interno delle sequenze di DNA e assegnare loro funzioni; iii) di predire i fenotipi degli organismi. E' possibile immaginare queste sfide come ricostruzioni -omiche 1-D, 2-D e 3-D. In questa tesi lo scopo era di esplorare, usando specifici casi studio, tali inferenze biologiche computazionali 1-D, 2-D e 3-D sulle sequenze genomiche procariotiche. Ciascun capitolo della sezione risultati presenterà dati su genomi batterici di ceppi importanti per svariate applicazioni biotecnologiche. I risultati complessivi sono presentati a seconda del grado di profondità dell'inferenza funzionale (fenotipica), dall'assemblaggio e semplice annotazione funzionale di un genoma fino ai potenti modelli metabolici genome-scale. Il focus principale sarà incentrato sull'enfatizzazione del valore predittivo del metabolic modeling su scala genomica per fenotipi complessi e predizione di geni essenziali *in silico*. La conclusione principale riguarda l'uso integrato dei tool computazionali per assistere le inferenze di systems biology, per aiutare tutte le scienze biologiche (genomiche) predittive.

Publications

International Journals

1. **Presta**, L., Bosi, E., Mansouri, L., Dijkshoorn, L., Fani, R., and Fondi, M. (2017). Constraint-based modeling identifies new putative targets to fight colistin-resistant *A. baumannii* infections. *Scientific Reports* 7 (1), 839, 2017. 2017

2. Maggini, V., **Presta**, L., Miceli, E., Fondi, M., Bosi, E., Chiellini, C., Fagorzi, C., Bogani, P., Di Pilato, V., Rossolini G.M., Firenzuoli, F., Mengoni, A and Fani, R. (2017). Draft Genome Sequence of *Pseudomonas* sp. Strain Ep R1 Isolated from *Echinacea purpurea* Roots and Effective in the Growth Inhibition of Human Opportunistic Pathogens Belonging to the *Burkholderia cepacia* Complex. *Genome Announcements*,5(20), e00351-17.

3. Miceli, E.*, **Presta**, L.*, Maggini, V., Fondi, M., Bosi, E., Chiellini, C., Fagorzi, C., Bogani, P., Di Pilato, V., Rossolini G.M., Firenzuoli, F., Mengoni, A and Fani, R. (2017). New Genome Sequence of an *Echinacea purpurea* Endophyte, *Arthrobacter* sp. Strain EpSL27, Able To Inhibit Human-Opportunistic Pathogens. *Genome Announcements*,5(25), e00565-17.

4. Fondi, M., Bosi, E., **Presta**, L., Natoli, D., and Fani, R. (2016). Modelling microbial metabolic rewiring during growth in a complex medium. *BMC Genomics*,17(1), 970.

5. **Presta**, L., Bosi, E., Fondi, M., Maida, I., Perrin, E., Miceli, E., Bogani, P., Di Pilato, V., Rossolini G.M., Firenzuoli, F., Mengoni, A and Fani,

R. (2017). Phenotypic and genomic characterization of the antimicrobial producer Rheinheimera sp. EpRS3 isolated from the medicinal plant *Echinacea purpurea*: insights into its biotechnological relevance. *Research in Microbiology*,168(3), 293-305.

6. **Presta**, L., Fondi, M., Perrin, E., Maida, I., Miceli, E., Chiellini, C., Bogani, P., Di Pilato, V., Rossolini G.M., F., Mengoni, A and Fani, R. (2016). *Arthrobacter* sp. EpRS66 and *Arthrobacter* sp. EpRS71: draft genome sequences from two bacteria isolated from Echinacea purpurea rhizospheric soil. *Frontiers in Microbiology*,7.

7. **Presta**, L., Inzucchi, I., Bosi, E., Fondi, M., Perrin, E., Miceli, E., Tutino, M.L., Lo Giudice, A., de Pascale, D., and Fani, R. (2016). Draft genome sequence of *Flavobacterium* sp. strain TAB 87, able to inhibit the growth of cystic fibrosis bacterial pathogens belonging to the *Burkholderia cepacia* complex. *Genome Announcements*,4(3), e00410-16.

8. **Presta**, L., Bosi, E., Fondi, M., Maida, I., Perrin, E., Miceli, E.,Chiellini, C., Bogani, P., Di Pilato, V., Rossolini G.M., F., Mengoni, A and Fani, R. (2016). Draft Genome Sequence of *Pseudomonas* sp. EpS/L25, Isolated from the Medicinal Plant *Echinacea purpurea* and Able To Synthesize Antimicrobial Compounds. *Genome Announcements*,4(3), e00346-16.

9. **Presta, L.**, Inzucchi, I., Bosi, E., Fondi, M., Perrin, E., Maida, I., Miceli, E., Tutino, M.L., Lo Giudice, A., de Pascale, D., and Fani, R.(2016). Draft Genome Sequences of the Antimicrobial Producers *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 Isolated from Antarctic Sediments. *Genome Announcements*,4(4), e00728-16.

10. Adessi, A., Spini, G., **Presta**, L., Mengoni, A., Viti, C., Giovannetti, L., Fani, R., and De Philippis, R. (2016). Draft genome sequence and overview of the purple nonsulfur bacterium *Rhodopseudomonas palustris* 42OL. *Standards in Genomic Sciences*,11(1), 24.

Books

- **Presta** L., Fondi M., Emiliani G., Fani R (2015) Molybdenum Cofactors and Their role in the Evolution of Metabolic Pathways. Springer, The Netherlands. ISBN 9789401799713. Editor and author

Submitted for publication:

- Zoledowska S.*, **Presta** L.*, Fondi M., Decorosi, F., Giovannetti, L., Łojkowska E., and Mengoni A. (2017), Metabolic modelling of the plant pathogenic bacterium Pectobacterium parmentieri provides insights into the metabolic pathways of plant-associated bacteria.

Contribution Summary

1. L.P. conceived the study, performed metabolic network reconstruction and modeling, and prepared the manuscript.

2. L.P. performed reads check, genome assembly and annotation.

3. L.P. performed reads check, genome assembly, annotation and assisted in the analysis.

4. L.P. participated to the model refinement step.

5. L.P. performed reads check, genome assembly, annotation, analyzed genome features and performed comparative genomics analysis, prepared and contributed writing the paper.

6. L.P. performed reads check, genome assembly, annotation and wrote the paper.

7. L.P. performed reads check, genome assembly, annotation and wrote the paper.

8. L.P. performed reads check, genome assembly, annotation and wrote the paper.

9. L.P. performed reads check, genome assembly, annotation and wrote the paper.

10. L.P. performed comparative and phylogeny analysis and contributed in writing the paper.

11. L.P. carried out the literature review, prepared the draft and contributed writing the book.

12. L.P. reconstructed the model and performed the computational experiments, analyzed the data and participated writing the paper.

She captured a feeling
Sky with no ceiling
The sunset inside a frame

She lived in her liquor
And died with a flicker
I'll always remember the flame

— To R.

Acknowledgements

As a child, I daydreamed the possible work that I would do as an adult. Among the many, my favorite fantasy was to became an astronaut and surfing between the stars. I had no idea that the journey that I would finish doing was in a different, tiny universe.
Expectations, engagement, dedication, hopefully, fulfillment and a nip of panic: when I started the doctoral program that's what I thought I was going to experience in my up-coming three years. It has possibly been even more. I learned not just by successes but especially from failures and loss. And, unexpectedly, by "major revisions" too.
I don't know to what extent this experience will influence my future life, however, whatever job I will finally get to do, now I know what I seek **to be**. Perhaps this awareness is the greatest achievement.
I am very grateful to all the coworkers who accompanied me in such a great experience and here I would like to acknowledge them.
Thanks to all the lab-people with whom I shared this years, especially Giovanni and Emanuele which have been for me both teachers and mates.
Thanks to Marco, for always being, in a single word, a mentor.
Thanks to Professor Renato Fani and Professor Alessio Mengoni, not just for supervising and encouraging my work, but especially for the trust in me. It is priceless.
Finally, my biggest acknowledgement is to my family: to my parents and my siblings for the never ending support and love since I was that little daydreaming child; to Simone as he is always by my side, believing in me more than I do and sharing the weight of each, single choice I make.

To them I would like to say: you are the people who made it possible.

Contents

Part I

INTRODUCTION

Preface

> *The whole is greater than the sum of its parts.*
>
> **— Aristotele**

The past twenty years of biological studies have been defined as the genomic era. Enormous progress in the experimental and computational techniques used to sequence genomes allowed an unprecedent level of knowledge. However, many questions are still uncovered. The main reason of these gaps is probably connected to a "reductionist" approach to the study of organisms, which historically privileged the "fraction" over the "system". This basically means that complex biological mechanisms have been studied under the lens of molecular biology, which focuses on searching below the large-scale manifestations of a system. However, to reach a deep understanding of such complex processes, it is necessary to examine the structure and dynamics of the cellular and organism functions, rather than the characteristics of their isolated parts. This is the leading idea of systems biology. Luckily, these two views are not intended to be set apart for ever: computational systems biology is emerging as the natural meeting point of both.
Molecular aspects can now be seen in a functional perspective thanks to the use of tools like genome-scale metabolic models. These implementations synthesize the bottom-up (data-driven) and the top-down (model-driven) scientific approaches, connecting these two in a loop of wet- and dry- experimental steps in which both result fundamental. The overall goal of this dissertation is to show how genomics easily flow in GEMs framework, thus enabling for functional estimations of organisms' properties, instead of purely descriptive ones. The works herein contained can be conceptually distinct in three main sections:

- genome sequencing, assembly and annotation;
- comparative and evolutionary biology;
- metabolic modeling.

However, this subdivision has the only scope to drive the logical progression of the studies, rather than actually separate them. On the contrary, the underlaying intent is to show their common scope, the possible interconnections and, hopefully, to sketch the auspicable path of future research field.

1 The genomic era

1.1 The first glimpse

The first DNA sequencing methods can be tracked back to 70s and 80s (Maxam-Gilbert sequencing and Sanger sequencing [1, 2]); however, it was only in 1995 that the first completely automated genome sequencing of an organism took place [3]. The chance to know the exact DNA sequence of an organism kick-started for deep-impact genomic studies and a new era in biological science begun.

At the early stages, the major implications of this innovation were in the field of bacteriology (see Figure 1.1), where it immediately revealed unexpected insights and transformed the way to discern bacteria evolution and functions. Since they were relatively small and easy to handle, sequencing technology boosted the study of microbial genomes. Initially, the priority was to sequence as many different organisms as possible, with great concern about the genome composition and organization. Publicly available databases and repositories collecting the sequenced data, like GenBank, were created. Moreover, software tools, as BLAST [4], made it easy to infer the function of novel, unknown sequences, based on comparison with those already stored into existing databases. Soon, thanks to the increasing number of available genomes, comparative and phylogenetic approaches revealed how deep the impact of horizontal gene transfer (HGT) was and how it shaped the evolution of Bacteria and Archea.

In April 2003, another milestone in human history and scientific discovery was reached when the Human Genome Project (HGP) [5] was completed. The ambitious, thirteen years long and three-billion-dollar project, raised the expectation of many to reveal the “secret of life” (and of disease?). However, information is only as good as the ability to use it, and indeed, on one side it enabled to answer to many questions, and on the other side, it also brought to light new challenges for researchers and scientists: how to read and interpret the contents written into the genome? The following years witnessed enthusiastic explorations and disappointing difficulties.

In 2004, a pioneering shotgun metagenomics sequencing project reported thousands of new genes [6]. Uncultivable species whose presence had always

been postulated were now finally detected. It would have impacted the whole field, spanning from ecological to evolutionary, from clinical to agricultural microbiology.

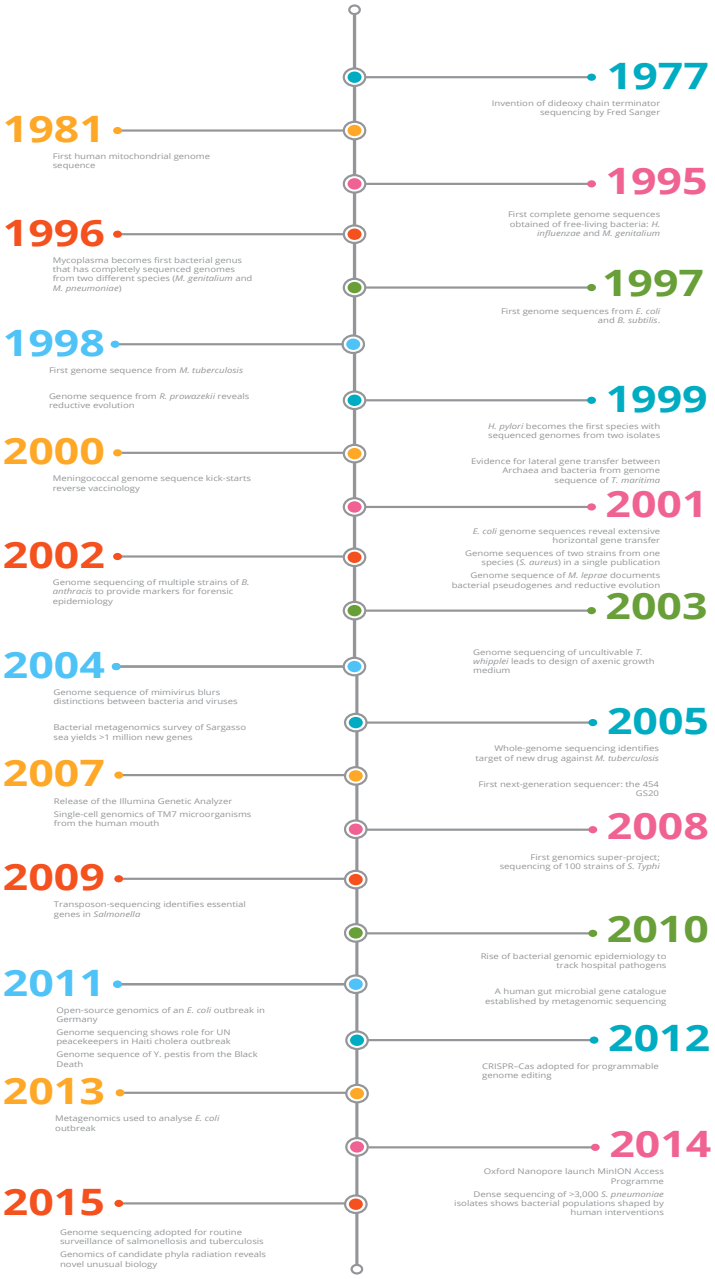


Fig. 1.1: Milestones in bacterial genome sequencing. Figure edited from [7]

1.2 The next-generation sequencing technology

A decade after the *H. influenzae* genome sequencing [3], society was ready for the first high-throughput sequencing technology. The new platforms replaced the onerous, expensive, biologically-based method with a new chemistry-based one. As a matter of fact, the new *in-vitro* method, allowed fast, high-throughput and relatively low-cost results. On the contrary, the first genome shotgun technology required the *in-vivo* generation of templates which was slow and less effective since it demanded for colonies selection and plasmids preparation. The switch to the new technology had two big consequences: it delivered the ability to perform genome-sequencing inside universities and small research centers and it also affected the quality of the whole system by producing shorter and less-refined reads (*i.e.* short copies of DNA fragments). This basically meant that scientific community witnessed a shift from complete genomes to draft genomes. In fact, the new technology made extremely easy to sequence a genome but in turn it required a big effort to assemble and annotate it. Moreover, the shorter reads affected the resolution of repeated regions. Paired-end (PE) sequencing [8, 9] was hence introduced in order to solve the problem. Sequencing both ends of the DNA fragments and aligning the forward and reverse reads as read-pairs enabled a regain in accuracy. However, it also doubled the amount of reads, which entailed demanding computational support. Thus, the methodological development was accompanied by bioinformatic advance. At that point indeed, given the small experimental effort needed, the fastness of the method and the low-cost of the sequencing, massive quantity of data started to be delivered, and the proportion of the phenomena was so huge (Fig.1.2 and Fig. 1.3) that the new bottleneck became how to analyze the raw data rather than how to generate them.

1.3 Bioinformatic blooming and the third generation of NGS

In such a context, bioinformatics became both the answer and the challenge to face-off. The mere availability of powerful sequencing platform would

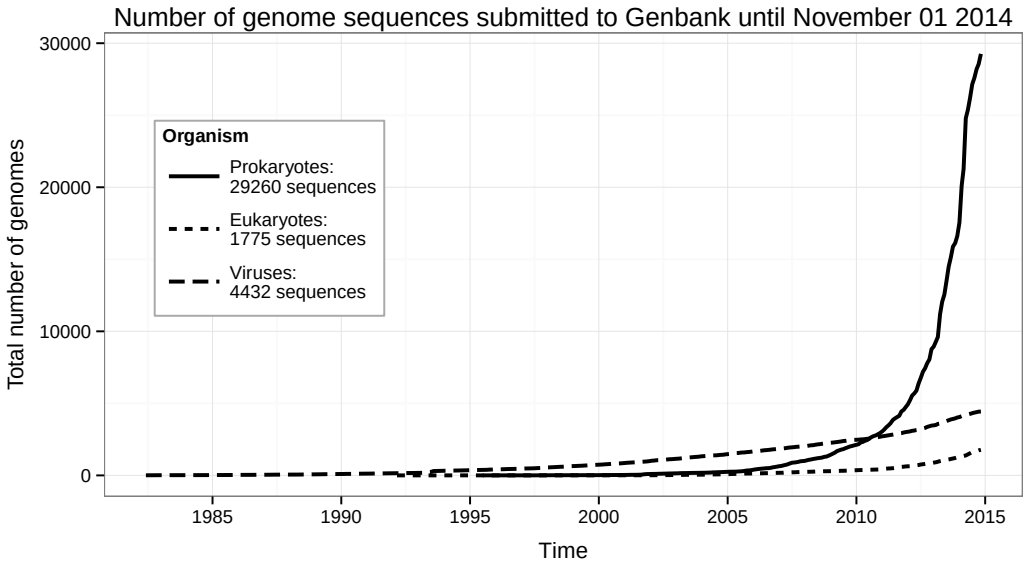


Fig. 1.2: The number of genome projects has increased as the technological improvements continue to lower the cost of sequencing. Exponential growth of genome sequences database GeneBank since 1995. Original source PhD dissertation http://hdl.handle.net/2158/986409

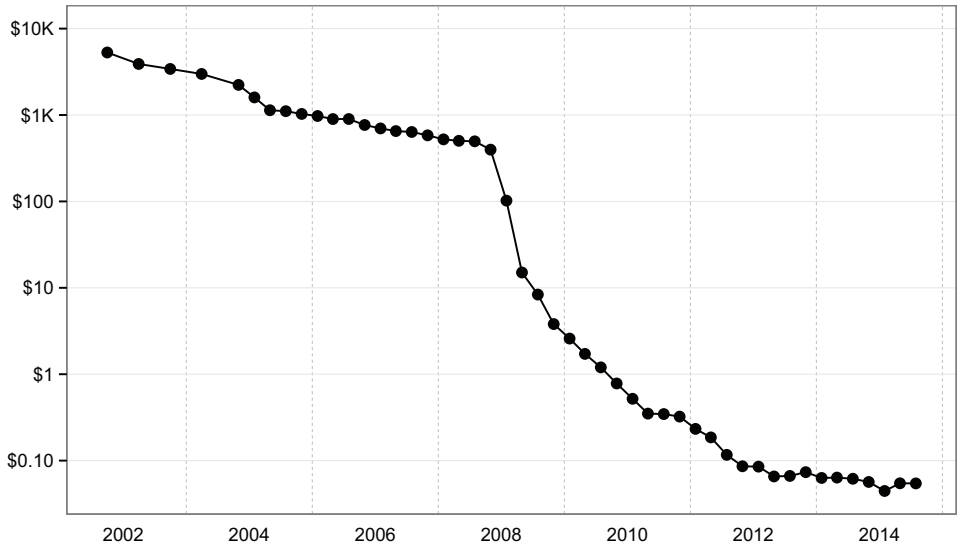


Fig. 1.3: Average cost for raw megabase of DNA sequence. Costs are reported using logaritmic scale beginning from 2001 until the end of 2014. Data are collected from the NHGRI website. Original source PhD dissertation http://hdl.handle.net/2158/986409

have been ineffective without raw-data processing. Similarly, the knowledge of an exact genome sequence alone would have been useless in absence of algorithms to analyze its content. Simple local alignment that could be performed by hand were already used for sometime. Nevertheless, as a larger amount of sequences could now be compared, it became necessary to

move to new, specific algorithms and parallel, mathematical computing to address the issue. BLAST was one of the first similarity search, heuristic-based methods to be developed and constituted the foundation of the successive algorithms designed for genome assembly (that in the meantime became very critical). New and powerful tools were released both for assembly, such as ABySS [10], SOAPdenovo [11] and Velvet [12], and for gene calling and annotation, like Phred, Phrap [13], Glimmer [14] and Artemis [15]. However, the biggest return of all these innovations was, probably, the adoption of a new mindset in the scientific community with an increasing interest in evolutionary turn-out and phenotype inferences. The joint development of wet-lab techniques and computational methods fostered the start and dissemination of new comparative genomics studies. The high-throughput screening of gene function now rapidly allows the investigation of basic biological similarities/differences and the estimation of the gene repertoire of various taxonomic group. Structural and genetic landmarks and evolutionary relationships between organisms can be targeted, sometimes influencing the concept of species as we knew it.

Today we are assisting in the third revolution in the genome sequencing practice. Different machineries, relying on different methodologies, have been commercially released. These provide a new opportunity to study genetic variation on an unprecedented scale by producing longer reads, like the first generation of sequencing machines, but with the high throughput introduced by the second one. The new platforms (e.g. long-read, single-molecule sequencing technology (SMRT) [16] and Nanopore sequencing [17]) restored the ability to produce complete genomes, a practice that was largely dismissed with the advent of the first NGS technologies. This enables high accuracy of the analysis, especially of repeat-rich areas of the genome, revealing large genomic complexity and, at the same time, reducing the annotation effort/cost.

Before moving to practical implications that all these technical improvements brought, the following chapters will cover some bioinformatics aspects about genome sequencing.

Bibliography

1. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* **74,** 560–564 (1977) (cit. on p. 5).
2. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94,** 441IN19447–446IN20448 (1975) (cit. on p. 5).
3. Fleischmann, R. D., Adams, M. D., White, O., *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *science,* 496–512 (1995) (cit. on pp. 5, 7).
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990) (cit. on p. 5).
5. Mural, R. J., Adams, M. D., Myers, E. W., *et al.* A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002) (cit. on p. 5).
6. Venter, J. C., Remington, K., Heidelberg, J. F., *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *science* **304,** 66–74 (2004) (cit. on p. 5).
7. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* **13,** 787–794 (2015) (cit. on p. 6).
8. Edwards, A., Voss, H., Rice, P., *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6,** 593–608 (1990) (cit. on p. 7).
9. Edwards, A. & Caskey, C. T. Closure strategies for random DNA sequencing. *Methods* **3,** 41–47 (1991) (cit. on p. 7).

10. Simpson, J. T., Wong, K., Jackman, S. D., *et al.* ABySS: a parallel assembler for short read sequence data. *Genome research* **19,** 1117–1123 (2009) (cit. on p. 9).

11. Li, R., Zhu, H., Ruan, J., *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20,** 265–272 (2010) (cit. on p. 9).

12. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18,** 821–829 (2008) (cit. on p. 9).

13. Rieder, M. J., Taylor, S. L., Tobe, V. O. & Nickerson, D. A. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic acids research* **26,** 967–973 (1998) (cit. on p. 9).

14. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic acids research* **27,** 4636–4641 (1999) (cit. on p. 9).

15. Rutherford, K., Parkhill, J., Crook, J., *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000) (cit. on p. 9).

16. Eid, J., Fehr, A., Gray, J., *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009) (cit. on p. 9).

17. Eisenstein, M. *Oxford Nanopore announcement sets sequencing sector abuzz* 2012 (cit. on p. 9).

2 From raw data to complete genome

Rapid progress in high-throughput sequencing technology would have been useless without the intense bioinformatic labour that accompanied its onset. The new sequencing method in fact, presented two major critical concerns: how to place back together the enormous amount of raw data produced (reads) to make up the original genome, and how to deal with repetitive sequences. However, the *de novo* assembly is a mathematically difficult problem, regardless of the sequencing technology used, for which no efficient solution is known, falling in the class of the computational NP-hard problems (nondetermistic polynomial-time hard problem) [1]. This chapter will cover the basic downstream data processing pipeline used to approach these issues.

2.1 A preliminary phase: raw data check

The first action required to reconstruct long stretches of DNA sequence (up to genome length) from the reads library is to check their quality. Sequencing in fact, is an error prone process during which different anomalies may occur (e.g. PCR duplications, poor base calling toward the 3' (terminal) ends of reads, inaccurate determination of simple sequence repeats, etcetera). To avoid misleading results, reads are pre-processed and the PHRED score (Q) [2, 3] associated to each nucleotide is estimated (it stands for the degree of confidence of the identification of the nucleobases and corresponds to the ASCII-encoded quality number). Other parameters like the GC content, repeats abundance and/or the proportion of duplicated reads are then taken into account to estimate the overall features makeup and individuate possible artifacts. To achieve this scope several tools may be adopted. One of the most famous is the user-friendly FASTQC software [4] that yields graphical visualization of basic statistical parameters. Based on the results of this preliminary analysis, reads can be trimmed. Trimming is a non-trivial process which aims at removing low quality portions while preserving the longest high quality part of a NGS read. During such a phase duplicates, contaminant

sequences, adapters and low quality reads are eventually filtered out. Here, the principal effort is to pick the optimal trimming cut-off in order to not reduce the coverage and to speed-up the following assembly process.

2.2 The assembly problem

Determining the correct order of reads seeks computational approach. An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. The sequencing depth influence the correct achievement of that task. Perfect duplicated repeats located in different regions of the genome may result indistinguishable, especially if the repeats are longer than the reads. Over-sampling the target DNA is thus the first requirement to satisfy minimum detectable overlap criteria (but may not be sufficient in case of short reads). However, high coverage increases complexity and intensifies computational issues. Estimation of the correct assembly is demanded to different algorithms, some of them aim to minimize mis-assemblies while others mainly focus on improving the contiguity. Beyond the possible strategies, they are all based on graph theory and they can be classified accordingly. Moreover, all assembly softwares rely on the notion of a K-mer (figure 2.1). This is a sequence of a number K calls of nucleic base, where K is any positive integer. The detection of shared K-mer in the reads reduces the computational cost of assembly. Overlap discovery is sensitive to settings of K-mer size, minimum overlap length, and minimum percent identity required for an overlap. These three parameters affect robustness. A brief description of the joining problem and the basic work flow of the algorithms is provided below.

2.2.1 Structuring the assembly as a graph

As mentioned, computational solutions represent the sequence of the reads in an assembly graph where the labels along a path of the graph encode a sequence. A graph is an abstraction widely used in computer science. It is a set of nodes connected by a set of edges. Nodes and edges may also be called vertices and arcs, respectively. If the edges are traversed in one direction, the graph is known as a directed graph, otherwise it is undirected. Importantly, each directed edge represents a connection from one source node to one sink node. Collections of edges form paths that visit nodes in

some order, such that the sink node of one edge forms the source node for any subsequent nodes. In modeling the assembly as a graph reads are nodes and the overlaps connect the nodes with edges. For the assembly problem the principal algorithms used are three:

- the Overlap/Layout/Consensus (OLC)
- the de Bruijn graph (DBG)
- the Greedy graph

The factors that influence the choice of algorithms include the quantity of data (as read length and coverage), quality of data and genome structure (*e.g.*, GC content and the number and size of repeated regions).

2.2.2 The Overlap/Layout/Consensus algorithm

Overlap/Layout/Consensus (OLC) method rely on an overlap graph and it is manly used for long reads. In an overlap graph assembling the reads into the genome is equivalent to finding a Hamiltonian path, which is a path that visits every node of the graph exactly once. Overlap discovery involves all-against-all, pairwise read comparison. The seed and extend heuristic algorithm is used for efficiency. The software precomputes K-mer content across all reads, selects overlap candidates that share K-mers, and computes alignments using the K-mers as alignment seeds. The second (layout) step is the construction of the graph, such that two vertices are connected with an edge if the corresponding reads overlap. However, due to this heuristics, the set of sequences obtained (contigs) may be biased and fragmented because of complex patterns in the graph that are generated by sequencing errors, genomic variants and repeats. Construction and manipulation of an overlap graph leads to an approximate read layout. The overlap graph need not include the sequence base calls, so large-genome graphs may fit into practical amounts of computer memory. Paths in overlap graphs converge at repeats longer than a read. Many widely used assembly programs adopted OLC, such as Phrap [5], Newbler [6], Arachne [7], etc.

2.2.3 The De Bruijn Graph algorithm

The De Bruijn graph (DBG) was generated for the "string reconstruction problem" *i.e.* finding the shortest superstring that contains as substrings all possible reads of a given length k over an arbitrary alphabet. Today it is now largely applied in genome assembly, manly in case of short reads produced by the second generation of sequencing technologies. The De Bruijn graph has the form of a K-mer graph for the search of a Eulerian path (see Figure 2.1). The nodes represent all possible fixed-length strings. The edges represent suffix-to-prefix perfect overlaps and its nodes all the fixed-length subsequences drawn from a larger sequence. that contains every edge exactly once is called an Eulerian path. K-mer graphs are more sensitive to repeats and sequencing errors and paths converge at perfect repeats of length K or more. Several tools have been developed based on DBG, such as Euler-USR [8], Velvet [[9], ABySS [10], AllPath-LG [11] and SOAPdenovo [12].

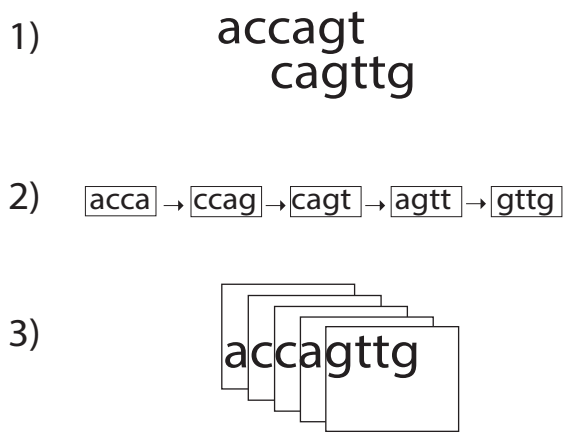


Fig. 2.1: K-mer graph based on pair-wise overlap. **(1)** Two reads have an error-free overlap of 4 bases. **(2)** One K-mer graph, with K=4, represents both reads.**(3)** Consensus sequence is easily reconstructed from the path.

2.2.4 The greedy graph algorithm

The greedy algorithms apply one basic operation: given any reads or contigs merge the two with the maximum overlap score at the time. The basic operation is performed at both the 3' and the 5' end of the read until no further extensions are possible until no more reads are left. Even in this case the problem is the search of an Hamiltonian path. Greedy approach drastically simplify the graph by considering only the high-scoring edges. Indeed, issues may arise since a local maxima may occur when the current contig is merged with reads that would help further contigs to grow even larger. As an optimization, they may actually instantiate just one overlap for each read end they examine, hence this strategy besides being very easy to implement, but it does not necessarily result in the optimal solution. Moreover, tools which rely on greedy approach have relatively long running times. Examples of softwares are SSAKE [13], SHARCGS [14] and VCAKE [15].

2.3 Growing in size: the scaffolding process

After reads are grouped in contigs (long stretches of DNA) they are merged to form scaffolds. Scaffolding is the process that links together a non-contiguous set of sequences establishing their order and orientation with a high confidence level. Analysing gene order and synteny, carrying out comparative or functional genomics or investigating patterns of recombination all rely heavily on obtaining an assembly with good continuity. However, the number of contigs depends on a combination of assembly tool used, sequence coverage, sequencing methodology and genome complexity. The more fragmented the assembly is, the harder the downstream analysis becomes. Algorithms used by assembly software are very diverse, however, even in this case, the problem can be formalised using graph theory, with contigs corresponding to nodes of the graph, and linking read pairs corresponding to edges. Gaps occur where reads from the two sequenced ends of at least one fragment overlap with other reads in two different contigs (as long as the arrangement is otherwise consistent with the contigs being adjacent). They may be of undefined or defined length (even if an estimation could always be

done), such length often has no relation to the true gap size. In several genomes, gaps are arbitrarily set to certain fixed lengths, most of the cases a series of 100 Ns connects two contigs or supercontigs. The reason for gaps in scaffolded assemblies is the presence of large, repetitive elements which short-read sequencing methods struggle to bridge. Thus, duplicated genes, pseudogenes, short tandem repeats, and many other structural genomic features are often unresolved in scaffolds. The uncertainties of gap sizes also result in an inability to understand the true spatial relationships of functional elements in genomes and is an underestimate of the actual extent of missing information. Paired-end information can be used to fill these gaps joining contigs into scaffolds. Most of the software based on such approach have several preparatory steps in which read and contig libraries are first converted to a specific format, then mapped against each other by means of an external aligner (e.g. BWA, [16] or BOWTIE, [17]). Currently available methods/software based on that pipeline include SOPRA [18], SCARPA [19], Opera [20], GRASS [21] and SSPACE [22]. An alternative and efficient approach to partially overcome gap issues may be that based on the use of one or more completes (closed) reference genomes to guide the ordering and the orientating of the contigs. Many software relying on that system exist: CONTIGuator [23], progressiveMauve [24], Projector2 [25], Ragout [26], MeDuSa [27].

2.4 Closing the genome

Gaps remaining after scaffolding can be closed only through supplementary refinement steps which are crucial, time-consuming and expensive but which in return result in complete genomes. Usually such refinements consist of amplifying the region corresponding to gaps followed by sequencing; alternatively, for more complex assemblies with a high number of gaps, a run with long reads, single-molecule sequencing platform is strongly suggested. After, the mapping approach can be adopted using the principal assembly as a backbone.

2.5 Genome annotation

The final operation of the modern genome sequencing work flow is the one that enables to find the location of each gene in the raw DNA sequence and labeling all the relevant features, *i.e.* the genome annotation process. This is properly the phase which actually treasures all the upstream effort and makes it possible to analyze genome contents. The information used to annotate genomes comes from three types of analysis: (i) *ab initio* gene finding programs, which are run on the DNA sequence to predict protein-coding genes; (ii) alignments of cDNAs and expressed sequence tags (ESTs), if available, from the same or related species; and (iii) translated alignments of the DNA sequence to known proteins. These types of evidence are abundant in various amounts depending on the organism; for less well-studied species, cDNA and EST evidence is often lacking, and annotators depend much more heavily on *ab initio* prediction programs, *i.e.* based only on the statistical parameters in the sequence. Many tools have been developed for such a scope, but generally proteins coding genes are annotated in two stages. First the coordinates of candidate genes are identified, then the putative gene product is described. The traditional way to predict what a gene codes for is to compare it with a large database of known sequences, usually at a protein sequence level, and transfer the annotation of the best significant match. Annotation can be performed through various online annotation servers (such as the NCBI Prokaryotic Genomes Automatic Annotation Pipeline service or RAST web server [28] for annotating bacterial and archaeal genomes, the fisrt has a turn-around time measured in days while the second provides annotation results in under a day) or through softwares which may be run locally (Prokka [29]). Importantly, identification of different genomic features may be demanded to different methods which are called by the software while running, for example Prokka uses Prodigal [30] to recognize Coding sequence (CDS), RNAmmer [31] for Ribosomal RNA genes (rRNA), Aragorn [32] for Transfer RNA genes, SignalP [33] for Signal leader peptides and Infernal [34] for Non-coding RNA.

Bibliography

1. Medvedev, P., Georgiou, K., Myers, G. & Brudno, M. *Computability of models for sequence assembly* in *WABI* **4645** (2007), 289–301 (cit. on p. 13).
2. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **8,** 175–185 (1998) (cit. on p. 13).
3. Richterich, P. Estimation of errors in "raw" DNA sequences: a validation study. *Genome Research* **8,** 251–259 (1998) (cit. on p. 13).
4. Andrews, S. FastQC. *A quality control tool for high throughput sequence data,* 13 (2010) (cit. on p. 13).
5. De la Bastide, M. & McCombie, W. R. Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics,* 11–4 (2007) (cit. on p. 15).
6. Margulies, M. Genome sequencing in microfabricated high-density picolitre reactors (Corrigendum). *NATURE-LONDON-* **1,** 502–502 (2006) (cit. on p. 15).
7. Batzoglou, S., Jaffe, D. B., Stanley, K., *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome research* **12,** 177–189 (2002) (cit. on p. 15).
8. Chaisson, M. J., Brinza, D. & Pevzner, P. A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research* **19,** 336–346 (2009) (cit. on p. 16).
9. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18,** 821–829 (2008) (cit. on p. 16).

10. Simpson, J. T., Wong, K., Jackman, S. D., *et al.* ABySS: a parallel assembler for short read sequence data. *Genome research* **19,** 1117–1123 (2009) (cit. on p. 16).

11. Gnerre, S., MacCallum, I., Przybylski, D., *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108,** 1513–1518 (2011) (cit. on p. 16).

12. Li, R., Zhu, H., Ruan, J., *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20,** 265–272 (2010) (cit. on p. 16).

13. Warren, R. L., Sutton, G. G., Jones, S. J. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23,** 500–501 (2006) (cit. on p. 17).

14. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome research* **17,** 1697–1706 (2007) (cit. on p. 17).

15. Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23,** 2942–2944 (2007) (cit. on p. 17).

16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009) (cit. on p. 18).

17. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** R25 (2009) (cit. on p. 18).

18. Dayarian, A., Michael, T. P. & Sengupta, A. M. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC bioinformatics* **11,** 345 (2010) (cit. on p. 18).

19. Donmez, N. & Brudno, M. SCARPA: scaffolding reads with practical algorithms. *Bioinformatics* **29,** 428–434 (2012) (cit. on p. 18).

20. Gao, S., Sung, W.-K. & Nagarajan, N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *Journal of Computational Biology* **18,** 1681–1691 (2011) (cit. on p. 18).

21. Gritsenko, A. A., Nijkamp, J. F., Reinders, M. J. & Ridder, D. d. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* **28,** 1429–1437 (2012) (cit. on p. 18).

22. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2010) (cit. on p. 18).

23. Galardini, M., Biondi, E. G., Bazzicalupo, M. & Mengoni, A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine* **6,** 11 (2011) (cit. on p. 18).

24. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* **5,** e11147 (2010) (cit. on p. 18).

25. Van Hijum, S. A., Zomer, A. L., Kuipers, O. P. & Kok, J. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic acids research* **33,** W560–W566 (2005) (cit. on p. 18).

26. Kolmogorov, M., Raney, B., Paten, B. & Pham, S. Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30,** i302–i309 (2014) (cit. on p. 18).

27. Bosi, E., Donati, B., Galardini, M., *et al.* MeDuSa: a multi-draft based scaffolder. *Bioinformatics* **31,** 2443–2451 (2015) (cit. on p. 18).

28. Aziz, R. K., Bartels, D., Best, A. A., *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9,** 75 (2008) (cit. on p. 19).

29. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014) (cit. on p. 19).

30. Hyatt, D., Chen, G.-L., LoCascio, P. F., *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11,** 119 (2010) (cit. on p. 19).

31. Lagesen, K., Hallin, P., Rødland, E. A., *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research* **35,** 3100–3108 (2007) (cit. on p. 19).

32. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research* **32,** 11–16 (2004) (cit. on p. 19).

33. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods* **8,** 785–786 (2011) (cit. on p. 19).

34. Kolbe, D. L. & Eddy, S. R. Fast filtering for RNA homology search. *Bioinformatics* **27,** 3102–3109 (2011) (cit. on p. 19).

The comparative approach 3

Although comparative genomics lays its root in the early 1980s with the comparison of viral genomes [1], it reached its top only later thanks to *in-silico* applications [2–4]. Indeed, during the first years of shot-gun genome sequencing there was a great concern about genome structures and major landmarks such as genome size, number of genes, and chromosome/replicons number became an entry point into comparative genomic analysis at different taxonomic levels (see Table 3.1). These focus rapidly led to one of the most striking findings of that time *i.e.* that genome size does not correlate with evolutionary status, nor is the number of genes proportionate to genome size. However, it was just the kick-off. From the middle 90s onwards, the increasing reservoir of available genomics data, did tickle the use of comparing basic biological similarities and differences, as well as evolutionary relationships between organisms, at a finer-resolution, *i.e.* by mean of sequences comparison. Such a chance results from the equation that if a function is evolutionarily conserved among different organisms then the DNA encoding for it should be conserved too. Likewise, different functions between species should be ascribed to divergent genetic material. Therefore, gene functions can be inferred by analyzing sequences.

Organisms	**Size (bp)**	**Chromosomes**	**Genes**
Human (*Homo sapiens*)	3 billion	46	25000
Mouse (*Mus musculus*)	2.9 billion	40	25000
Fruit fly (*Drosophila melanogaster*)	165 million	8	13000
Plant (*Arabidopsis thaliana*)	157 million	10	25000
Roundworm (*Caenorhabditis elegans*)	97 million	12	19000
Yeast (*Saccharomyces cerevisiae*)	12 million	32	6000
Bacteria (*Escherichia coli*)	4.6 million	1	3200

Tab. 3.1: Genomic features of different organisms

This approach, coupled with genome structure analysis, mobile elements clues and many other genomic observations, pushed the research-field at unprecedented levels by comparing fruit fly or yeast genomes to that of humans or plants and pinpointing less straightforward correlations like genes that are essential to life and genomic signals that control gene functions across many species.

Beside these informations may be used for very diverse applications, the first and probably the naturally closest one, was in evolutionary biology. Fresh-new challenges gave a brand new impetus to scientists which started combining molecular fingerprints with the more classic physical and physiological methods to understand how the biology of living organisms has changed over time.

3.1 From molecular evolution to clinical applications

Mathematical computing not only assisted the transition to new frontiers of study but actually it has been the device that made it possible. Indeed, sequences comparison was computational expensive, involving multi-alignment of DNA or protein sequences, score-matrix generation, phylogeny trees construction etc. By implementing new algorithms and softwares enabling to estimate the evolutionary distance between species, bioinformaticians delivered the capability to analyze entire genomes and to track the simplest genetic events, like point mutations, as very complex evolutionary marks, like, for instance, gene duplication, deletion or insertion and horizontal gene transfer (HGT). Moreover, the magnitude of the analysis was enormous, involving not only closely related organism but also very phylogenetically distant taxonomic species. Models and algorithms to infer such events, range from exact, heuristics, fixed parameter and approximation algorithms (for problems based on parsimony models) to Markov Chain Monte Carlo algorithms (for Bayesian analysis of problems based on probabilistic models).

3.1.1 *In-silico* taxonomy inference

Despite large experimental techniques availability (DNA-DNA hybridization for instance), today molecular phylogeny became almost a requirement to infer taxonomy. Bacterial phylogenies, for example, are inferred based on the molecular analysis of 16S RNA. The secondary structure of the small ribosomal subunit is the most conserved genomic region known across all the prokaryotic species and its degree of conservation varies widely only between hypervariable regions, which makes it perfect for taxonomic gathering purpose. High-conserved regions in fact, correlate to higher-level taxonomy

while less conserved regions correlate to lower levels, such as genus and species. It is also possible to deduce the phylogeny of a specific function based on homologous sequences set-up of many other genes or proteins. This often benefits of the comparison with the reference species phylogeny to estimate the frequency and the points of the HGT events. Recently, the horizontal transfer of 16S RNA has been argued [5] and, although this may be infrequent, the validity of 16S rRNA-based typing could be reevaluated. Indeed, the concept of species itself has been challenged in the genomic era. By providing evidence of extensive HGT, comparative genomics has in fact shown that the strains belonging to the same prokaryotic species may share a fraction only of their genome, thus, the concept of *pan-genome* analysis [6, 7] has been developed to accommodate such peculiar feature of prokaryotic genomes. Several definitions of prokaryotic species have been proposed (*e.g.* ecological, [8]); however, an operative definition for defining species affiliation of strains is based on the percentage of nucleotide identity within the shared (core) genome fraction, known as Average Nucleotide Identity (ANI, [9]).

3.1.2 The pangenome perspective

A *pangenome* [6, 7] is defined as the full complement of genes in a taxonomic group and is a tool to study the unique and cohesive genomic features of that class. In other words, it provides the ability to perform in depth analysis of intra-species diversity by taking advantage of an incredible accurate, though comprehensive, view. It consists of the determination of (see Figure 3.1):

- the "*core*" genome, *i.e.* those genes shared by all the strains of the studied group;
- the "*dispensable*" genome, *i.e.* those genes present in only two or more strains;
- and, finally, the "*unique*" genome, *i.e.* that repertoire of genes belonging to exclusively one strain of the examined group.

The size of each category (and pangenome itself) can result as a function of the number of genomes embedded into the analysis. As mentioned, different strains have different capacity to acquire exogenous DNA (due to different

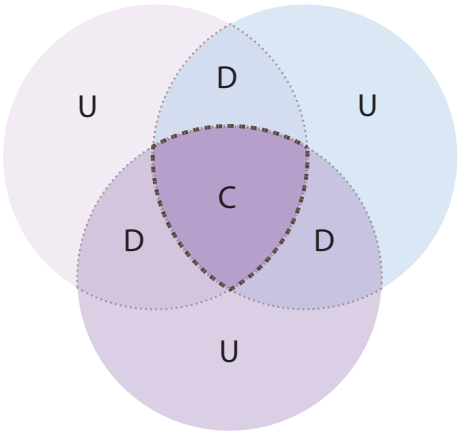


Fig. 3.1: Pangenome representation. The uninion of three different genomes results in core genome **C**, dispensable **D** and unique **U**.

HGT mechanisms and their occurrence); this influences the amount of the unique pangenome's segment. Because of that, it can be classified as "open" if its size increase as increasing the number of genomes analyzed. Conversely, if the size of the three categories is not influenced by the number of included genomes, then the pangenome is named "closed". Such definition stands for the ability of the species to maintain a certain set of genes, a property that is strictly related with the ecological niche occupied by the species. Closed pangenome in fact, are proper of those species that live in a narrow ecological niche where there is no real reason for a different adaptation and diversification of the strains, that would result in an open pangenome instead [10, 11].

Consequently, some considerations arise: genes belonging to core genomes usually encode for housekeeping functions of the species, therefore it can be used to identify those traits that are vertically inherited, while the other two identify lifestyle-related functions and strains' peculiarities, which are often horizontally acquired. Though evolutionary aspects are very important, far-reaching applications raise from the observations of genetic traits embedded in core, accessory and unique genomes. Reverse vaccinology [12], for example, is an emergent field that benefits from core features species insights to develop vaccine and pan-vaccine. Observations of dispensable or unique characters instead, cover great relevance in eco-evo studies and in antibiotic production or resistance mechanisms targeting.

NGS technology catalyzed the development of a plethora of comparative genomics methods. In addition to pangenome analysis in fact, several techniques are now used to characterize a species, or strains within species, at

various level: pulsed field gel electrophoresis [13], multilocus sequence typing (MLST) [14], multispacer sequence typing [15–17] or single nucleotide polymorphisms (SNPs) [18] represent a small subset used for various applications. The last, for instance, is a technique that finds a huge number of applications. It is used in forensic studies to match DNA samples and to infer phenotypic traits, in medicine to underline differences in human susceptibility to disease, in microbiology as one of the election methods to geo-typify relevant pathogens (covering great value in diagnostic and in outbreaks surveillance).

However, new perspective and applications of comparative genomics are awaited as whole genome shot-gun sequencing is likely to reach the mainstream settings, such as hospitals, where it promises to became the designated device even for routinely tasks.

Bibliography

1. Argos, P., Kamer, G., Nicklin, M. J. & Wimmer, E. Similarity in gene organization and homology between proteins of animal picomaviruses and a plant comovirus suggest common ancestry of these virus families. *Nucleic Acids Research* **12,** 7251–7267 (1984) (cit. on p. 25).
2. Koonin, E. V. & Galperin, M. Y. in *Sequence—Evolution—Function* 111–192 (Springer, 2003) (cit. on p. 25).
3. Koonin, E. V. & Galperin, M. Y. in *Sequence—Evolution—Function* 25–49 (Springer, 2003) (cit. on p. 25).
4. Koonin, E. V. & Galperin, M. Y. in *Sequence—Evolution—Function* 193–226 (Springer, 2003) (cit. on p. 25).
5. Badger, J. H., Eisen, J. A. & Ward, N. L. Genomic analysis of Hyphomonas neptunium contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and Caulobacterales. *International Journal of Systematic and Evolutionary Microbiology* **55,** 1021–1026 (2005) (cit. on p. 27).
6. Tettelin, H., Masignani, V., Cieslewicz, M. J., *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102,** 13950–13955 (2005) (cit. on p. 27).
7. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current opinion in genetics & development* **15,** 589–594 (2005) (cit. on p. 27).
8. Gevers, D., Cohan, F. M., Lawrence, J. G., *et al.* Re-evaluating prokaryotic species. *Nature Reviews Microbiology* **3,** 733–739 (2005) (cit. on p. 27).

9. Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of systematic and evolutionary microbiology* **64,** 346–351 (2014) (cit. on p. 27).

10. Georgiades, K. & Raoult, D. Defining pathogenic bacterial species in the genomic era. *Frontiers in microbiology* **1,** 151 (2011) (cit. on p. 28).

11. Diene, S. M., Merhej, V., Henry, M., *et al.* The rhizome of the multidrug-resistant Enterobacter aerogenes genome reveals how new "killer bugs" are created because of a sympatric lifestyle. *Molecular biology and evolution* **30,** 369–383 (2012) (cit. on p. 28).

12. Rappuoli, R. Reverse vaccinology. *Current opinion in microbiology* **3,** 445–450 (2000) (cit. on p. 28).

13. Amit, U., Porat, N., Basmaci, R., *et al.* Genotyping of invasive Kingella kingae isolates reveals predominant clones and association with specific clinical syndromes. *Clinical infectious diseases* **55,** 1074–1079 (2012) (cit. on p. 29).

14. Xiong, X., Wang, X., Wen, B., Graves, S. & Stenos, J. Potential serodiagnostic markers for Q fever identified in Coxiella burnetii by immunoproteomic and protein microarray approaches. *BMC microbiology* **12,** 35 (2012) (cit. on p. 29).

15. Arricau-Bouvery, N., Hauck, Y., Bejaoui, A., *et al.* Molecular characterization of Coxiella burnetii isolates by infrequent restriction site-PCR and MLVA typing. *Bmc Microbiology* **6,** 38 (2006) (cit. on p. 29).

16. Roest, H. I., Ruuls, R. C., Tilburg, J. J., *et al.* Molecular epidemiology of Coxiella burnetii from ruminants in Q fever outbreak, the Netherlands. *Emerging infectious diseases* **17,** 668 (2011) (cit. on p. 29).

17. Tilburg, J. J., Rossen, J. W., van Hannen, E. J., *et al.* Genotypic diversity of Coxiella burnetii in the 2007-2010 Q fever outbreak episodes in The Netherlands. *Journal of clinical microbiology,* JCM–05497 (2011) (cit. on p. 29).

18. Reuter, S., Harrison, T. G., Köser, C. U., *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a Legionella outbreak. *BMJ open* **3,** e002175 (2013) (cit. on p. 29).

4 A system-wide view

The described genome sequencing flow, combined with increasingly efficient annotation methods, set the stage for the systematic identification of most enzymes encoded in the genomes. Moreover, thanks to NGS technologies a huge variety of high-throughput experimental methodologies (the so-called -OMICS) rapidly emerged. These now routinely provide large-scale functional clues and snapshots of the physiological dynamics under specified conditions (and their underlying molecules). In a certain sense, the goal of the traditional studies of biological systems to isolate and measure their single involved components, reached the highest point. The technology enabled to successfully identify most of the cellular components and to describe their mechanisms at the molecular level. Unfortunately, despite this astonishing possibility, scientists often fail to capture multiple relevant biological features at the same time. The reason is that this is a reductionist approach which strives to describe the whole as the sum of its parts. The result is a lack of comprehensive, multi-layer knowledge whereas it is necessary to depict the long-range and complex interactions that are essential for the functionality. Since a cellular system is not just an ensemble of genes and proteins, its properties cannot be fully described by the merely depictions of isolated parts. To make a parallelism, identifying all the genes and proteins in an organism is like listing all the parts of an airplane. While such a list provides a catalog of the individual components, it is not sufficient to understand the complexity underlying the engineered object. It is necessary to know how these parts are assembled to form the structure of the airplane. Similarly, it is possible to draw diagrams of the principal molecular interconnections in a cell, but, although such diagrams represent an important first step, they are analogous to static road-maps, whereas what really matters are the traffic patterns, why such traffic patterns emerge, and how to control them (see Figure 4.1). Thus, despite an understanding of basic mechanism and molecular components continues to be important, the comprehension of system's structure and dynamics now moved on top of the scientific agenda. Here is where systems biology comes into play. Its final goal is to integrate biological and analytical systems to generate large-scale data about dynamic cellular processes. The election method used for that is to create a predictive, computational model. *In-silico* prediction approaches have received little attention in the past, but

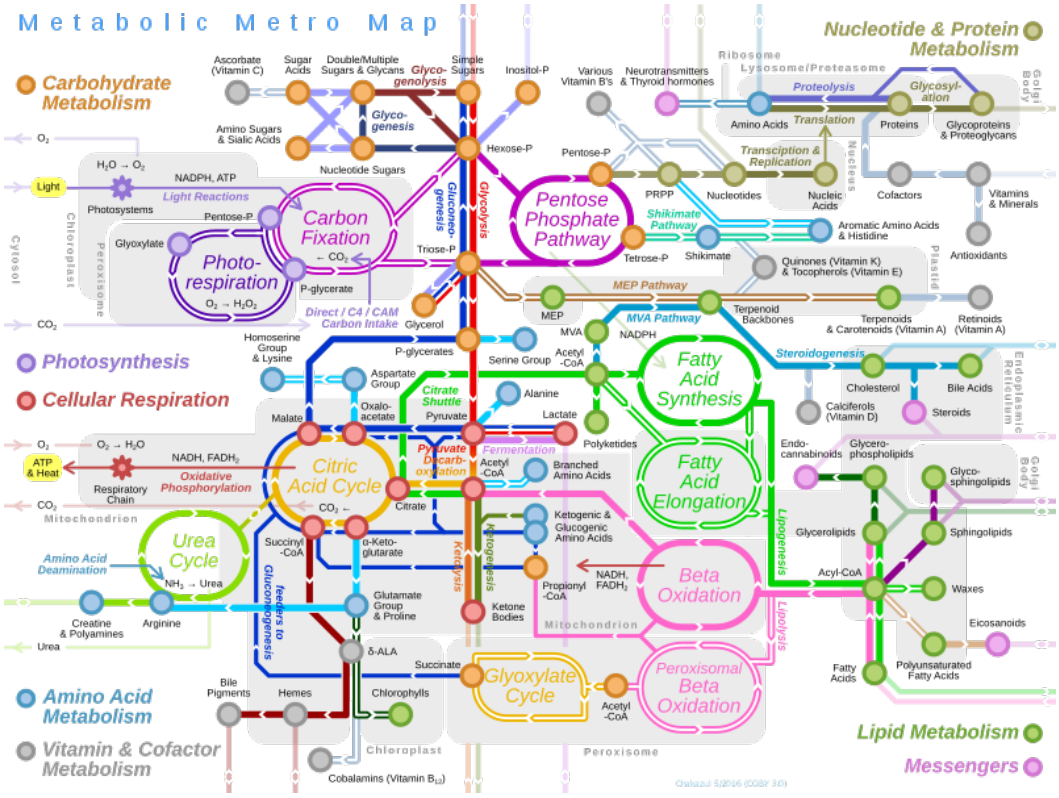


Fig. 4.1: **A:** Major metabolic pathways in metro-style map. Single lines: pathways common to most lifeforms. Double lines: pathways not in humans (occurs in *e.g.* plants, fungi, prokaryotes). Metabolic metro. Orange nodes: carbohydrate metabolism. Violet nodes: photosynthesis. Red nodes: cellular respiration. Pink nodes: cell signaling. Blue nodes: amino acid metabolism. Grey nodes: vitamin and cofactor metabolism. Brown nodes: nucleotide and protein metabolism. Green nodes: lipid metabolism. Original source: Wikipedia

it is now changing and great expectations are refueling the field. Substantial advances in software and computational power, in fact, have enabled the creation and analysis of reliable, yet intricate, biological models. These aim to predict the dynamics of systems so that the validity of certain assumptions can be tested. Detailed behaviors in computer-executable frames can be compared to experimental observations. Inconsistency at this stage stands for erroneous or incomplete assumptions. Conversely, models that survive the initial validation can then be used to make predictions and to design experiments, as well as to explore questions that are not amenable to experimental inquiry (carefully considering the purpose of models themselves, *i.e.* the scope and abstraction level of the models should always be defined before performing any analysis).

The choice of the used analytical methods often depends on the availability of biological knowledge to incorporate into the model. For instance, genome-

scale metabolic models (GEMs) [1–4] represent a type of systems modeling framework which (at a first stage) is based "only" on the genome annotation of the organism to be modeled. They are a consistent framework on which a mechanistic basis for the metabolic genotype-phenotype relationship can be formulated. In other words, GEMs are a bottom-up reconstructions of structured knowledge-bases that abstract pertinent information on the biochemical transformations taking place within specific target organisms. The conversion of such models into a mathematical format facilitates computational biological studies including evaluation of network content, hypothesis testing and generation, analysis of phenotypic characteristics, and metabolic engineering. For example, steady-state analysis can be performed on GEMs by just using the network structure (*i.e.* without any knowledge of the rate constants for a particular reaction). After a basic reconstruction, several high-throughput molecular data can be integrated. These lead to a very broad spectra of applications with high-confidence degree. A deeper description of such models will be now afforded.

4.1 Genome-scale metabolic models

GEMs represent very valuable resources but their applications have been limited due to the large manual effort required to develop them. The metabolic reconstruction, in fact, is a bottom-up operation based on genomic and bibliomic data, usually very labor and time intensive, depending on the targeted organism and the knowledge of its biology. The process is often iterative, consisting of multiple checks and, although a bunch of automated reconstruction methods exist, their output always require manual effort (see Figure 4.2 for an overlook of the reconstructions steps). The first stage consists of the generation of a draft network reconstruction only based on the genome annotation of the target organism. Software tools such as KBase [5], Seed [6], RAVEN [7] can be used to automatically generate it.

At this stage the network will represent a collection of genome encoded metabolic functions, some of which may be falsely included while other ones may be missing. Main issues are constituted by low-confidence annotations which can fail to correctly represent all the metabolic features of an organism. This results in lack of network connections and/or dead-end pathways. Thus, the quality and reliability of the genome annotation is crucial to the reconstruction quality. With the assistance of semi-automated algorithms

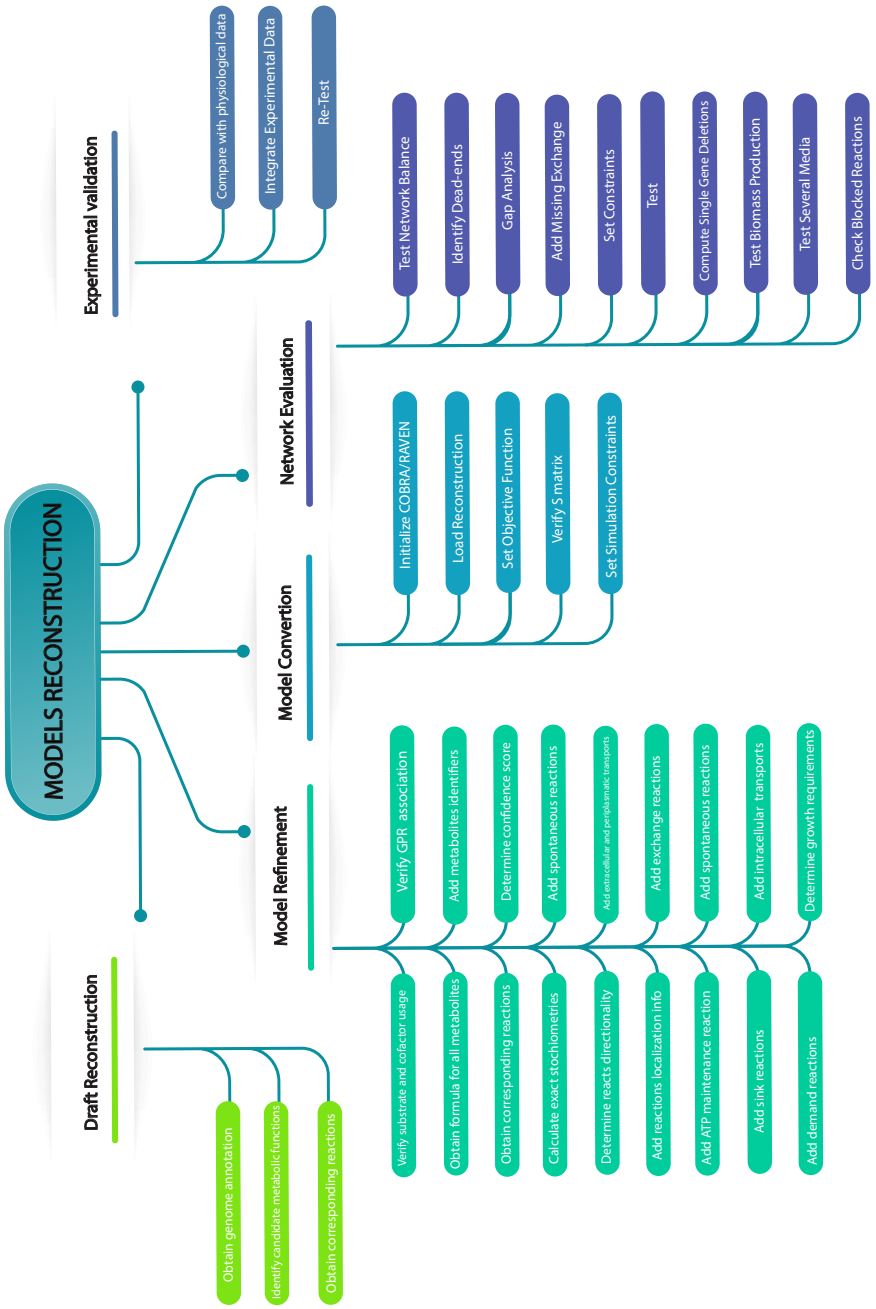


Fig. 4.2: Overview of the procedure to iteratively reconstruct metabolic models. The main steps are recovered from [8].

(tINIT, Task-driven Integrative Network Inference for Tissues algorithm in RAVEN [9], for example) these knowledge gaps can be filled-in by modifying the network to include additional biochemical reactions or by removing functions that were improperly added by previous annotators. However, manual inspection is mandatory. It mainly consists in verifying the connectivity of the network, identifying erroneous gene incorporation or possible gaps and, above all, retrieving experimental evidence for the presence of a gene product and its metabolic function. The step has a double relevance since it represents both a refinement and a validation process. Organism-specific literature must be collected and evaluated against model's behavior. If this lacks, it is possible to use phylogenetically close organisms to acquire the needed informations. For example, by finding ORFs encoding for enzymes orthologous to those that catalyze the same functions in other organisms, it is possible to create a higher quality metabolic model. Reference of publicly available databases like KEGG [10] and Brenda [11], is highly recommended. Data about biomass composition (that details all precursors and their fractional contributions to a cell's macromolecular composition), maintenance parameters, and growth conditions must be assessed during this phase too, since they deeply affect model's prediction credits. A final check of metabolite charge and reaction balance, as well as evaluation of reaction directionality and gene protein reaction rule (GPR) close the circle.

The standardized format used for models is the Systems Biology Markup Language (SBML). SBML file describing the model has to include at least the following information: stoichiometry, upper/lower bounds and objective function coefficients for each reaction. However, inclusion of further informations is strongly suggested in order to facilitate large reuse of models: reference of genes, metabolites and reactions should be added to be easily mapped to the majors existing catalogs (*i.e.* gene ontology (G.O.), KEGG, Seed, BiGG (Biochemical, Genetic and Genomic) [12], Chebi [13], Metanetx [14, 15]) and provide unique, unambiguous, standard-compliant and resolvable identifiers. At the end of the refinement, the model can be converted into a mathematical object to proceed with the simulation stage. COBRA [16, 17] and RAVEN toolboxes perform that conversion and come along with a suite of *ad hoc* useful functionalities that can be run locally by users in Matlab or Python environments. Various solver can be used in order to perform *in silico* tests, however, linear or quadratic solvers should be correctly chosen according to the computational problem to address.

4.2 Flux Balance Analysis

A widely used method (perhaps the most widely used) that exploits the transformation of genome-scale metabolic network in a mathematical object is the flux balance analysis (FBA) [18]. FBA calculates the flow of metabolites through the network structure allowing to make predictions on various biological aspects. The core features of FBA are:

1. a set of rules for gene-protein-reaction (GPR) relationships;

2. a mathematical representation of the model in the form of a numerical matrix (**S**-matrix).

GPR relationships represent conditional statements in Boolean logic between ORFs and their enzymatic functions while the **S**-matrix is comprised of biochemical reactions occurring in an organism: columns represent reactions and rows metabolites (see Figure 4.3). The coefficients of the substrates and products of each reaction are entered in the corresponding cell of the matrix.

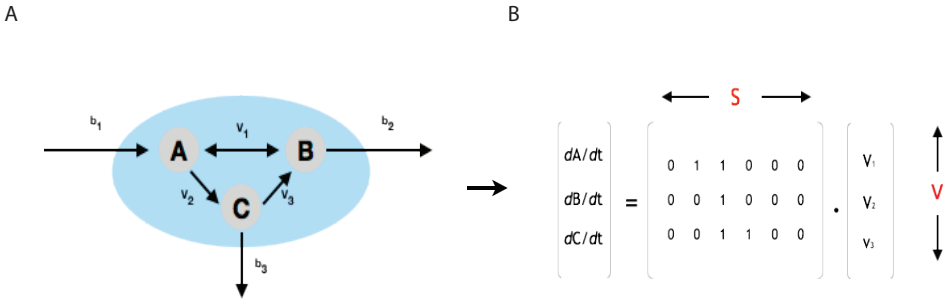


Fig. 4.3: **A:** A small metabolic network; reactions' flux and boundaries are represented as vectors. **B:** Stoichiometric matrix, S, of the network reaction list. The columns of S correspond to the network reactions, while the rows represent the network metabolites. The substrates in a reaction are defined to have a negative coefficient (they are consumed), while products have a positive value (they are produced). The metabolites participating in a reaction have non-zero entry in the **S**-matrix. All steady-states can be thus described by $\mathbf{S} \bullet \mathbf{v} = \mathbf{0}$, where **v** is a vector of reaction fluxes. Edited from [18]

The last impose constraints on the flow of metabolites through the network. Since all proper chemical equations can be described by stoichiometric coefficients, and since a set of chemical equations can be described by the **S**-matrix, all steady states of a network can be described by a simple linear equation, $\mathbf{S \cdot v = 0}$, where **v** is a vector of fluxes through chemical reactions. Thus, the computation of functional states of a network is enabled based on the known underlying chemistry. In FBA the restrains are represented in two ways, as equations that balance inputs and outputs and as inequalities that impose bounds on the system. The ensemble of reactions stoichiometries and systems boundaries identify condition-specific models defining the space of allowable flux distributions in a system and the rate at which every metabolite is consumed or produced by each reaction. In that conditions the flux distribution carries sufficient information to completely describe the state of the system (see Figure 4.4). Since the goal is to describe the joint operation of many metabolic reactions, it is convenient to define a flux distribution as a collection of reaction fluxes covering the entire system.

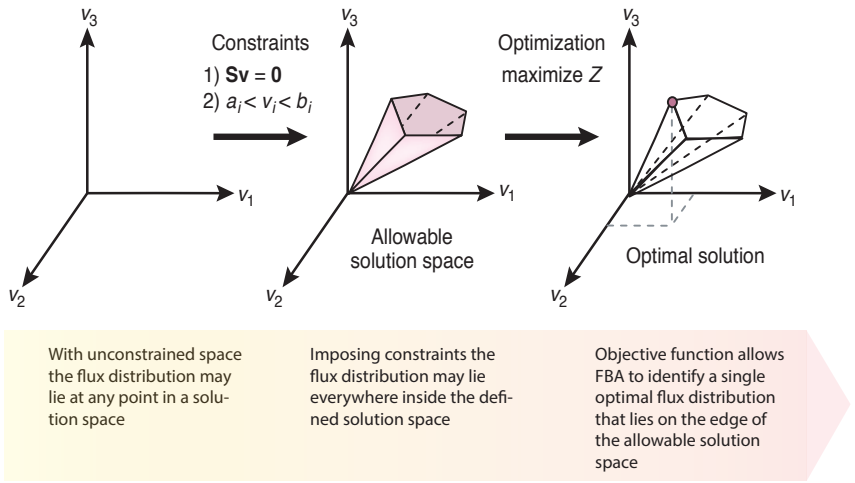


Fig. 4.4: The conceptual basis of constraint-based modeling. Edited from [18]

FBA can be applied in many contexts to analyze the phenotypic capabilities of organisms, in the background of different environmental and genetic perturbations by imposing external and internal pressures. Several and very different hypotheses can be computationally tested, such as by-product secretion, single and double gene deletion phenotypes, known (in)capabilities,

comparison of predicted physiological properties to known properties, quantitative evaluation of growth rate and so on. A reliable model could, in theory, be used in order to reproduce cells behavior during a variety of perturbations. However, when a cell is placed in a particular environment, it expresses a subset of its genes in response to environmental cues. The abundance of cellular components can be profiled using -OMICS methods. Such -OMICS data like transcriptomics, metabolomic or fluxomic data, can be analyzed using the model as a scaffold (see Figure 4.5 below) or turning the concept upside down, in a second modeling phase, -OMICS data can be mapped onto a network reconstruction to tailor it to the particular condition being considered. The result is that, when integrated in the computational framework, experimental outputs actually play as additional constraints that narrow the space solutions, giving raise to sharper predictions.

The union of the two methodologies (experimental and computational) creates a golden chance never reached before: to perform analysis at the genome scale while imposing the desired (real) external *stimuli* or genetic circuits with distinct properties. This is the core power of the mechanistic genotype-phenotype relationship for metabolism, *i.e.* it yields representation of specific metabolic conditions and derive non-trivial information about the underlying mechanism which are usually undetectable with the traditional approaches (see Figure 4.5).
The natural pipeline of model use should consist of feedback-loops between *in-silico* inferences and experimental assays (see Figure 4.6), where at each turn, better models allow better experiments and better results/hypothesis.

4.3 Metabolic systems biology perspectives

The ability of GEMs opens up new scientific directions to study fundamental biological phenomena. They were firstly used to predict the outcome of bacterial adaptation to new nutritional environments [19], even in the face of gene deletions [20], thereof the method has been extensively applied. Good computational models, in fact, accelerate design processes, and minimize prototyping, testing, and experimentation. Chapter 8, in the second part of this work, will offer two examples of development and usage of GEMs.

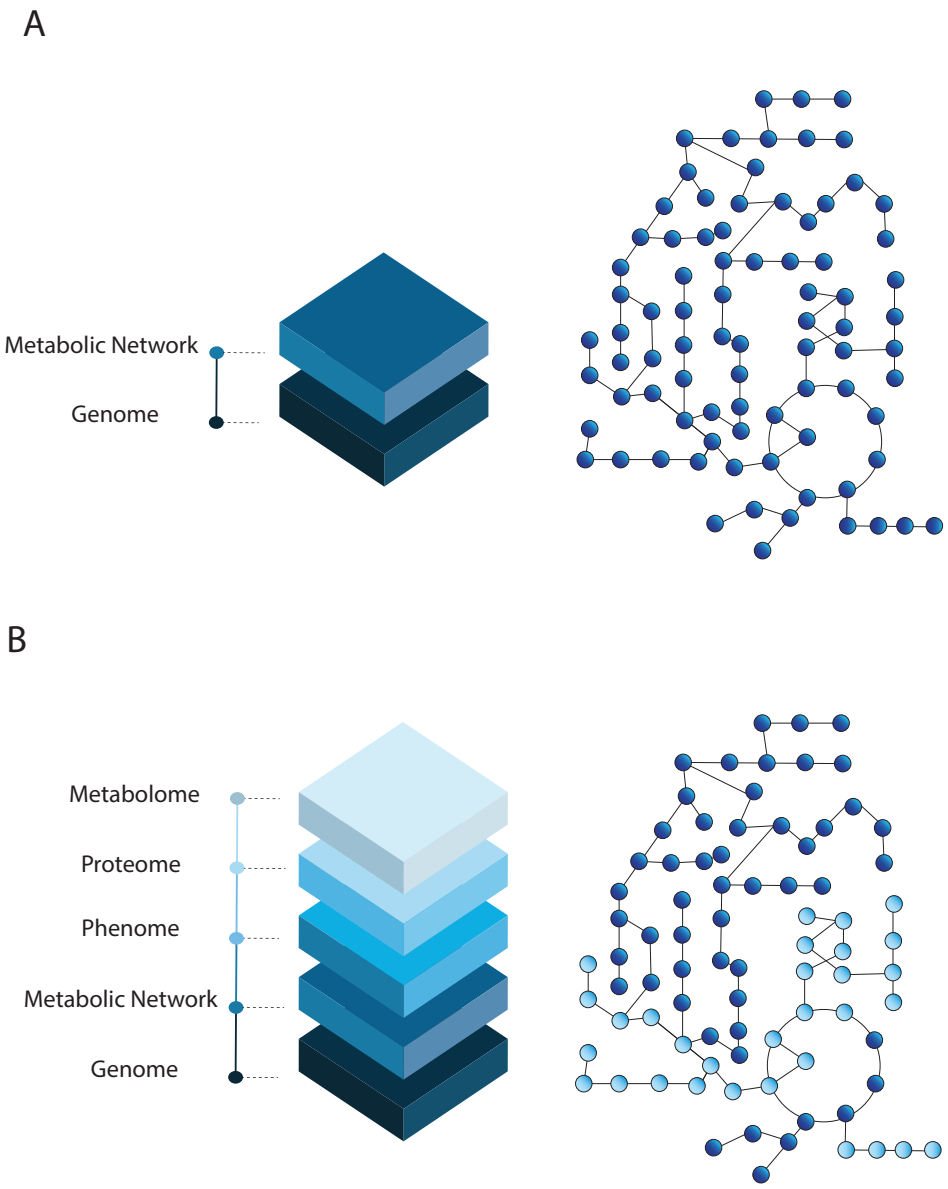


Fig. 4.5: **A:** The genomic and bibliomic data possessed are transferred in the corresponding metabolic network which can be converted in the scaffold metabolic model. **B:** Additional -OMICS data map the correct metabolic patterns and flux rates on the model which results in high confidence *in-silico* predictions.

There, thanks to the integrations of large-scale phenotypic data (Phenotype Microarray) and gene expression, functional clues on pathogenic microbial strains are made. However, the potential applications are enormous. The production of valuable metabolites or the degradation of pollutants and the

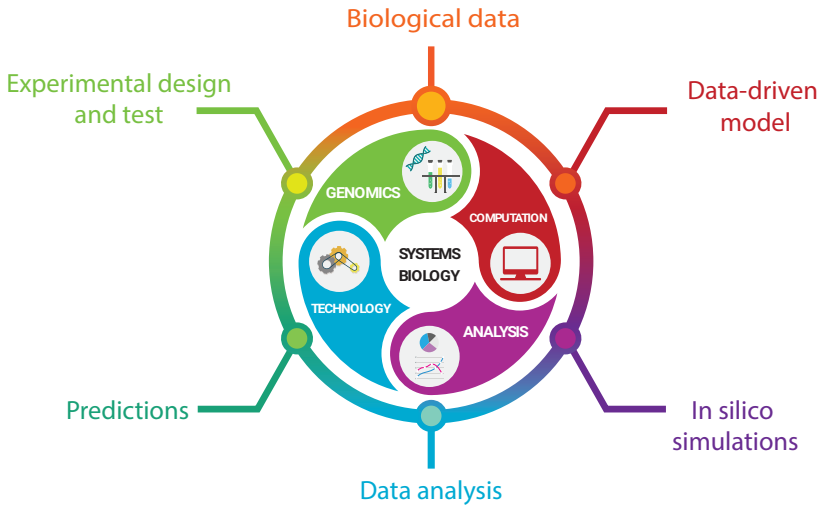


Fig. 4.6: A cycle of predictive model development begins with the selection of biological data which are transformed into a computable set of assumptions in the form of a model. Through the analysis of "dry" experiments results assumptions and predictions can be formulated. Based on that computer-derived hypothesis new experiments can be designed which at their turn lead to new experimental facts and-or thesis validation.

generation of renewable energy are just few of them [21–24]. About a decade since its appearance, genome-scale metabolic modeling aims to drive the use of microbial organisms for industrial purposes. Further applications can be found on microbial communities [25–27], plants [28, 29] and humans [1, 30, 31]. These developments currently represent an open avenue for highly complex systems evaluation. Whole-patient models for specific disease, such as obesity, diabetes and cancer offer a platform to explore personalized medicine and/or perform drug discovery [9, 32, 33]. Notably, the techniques and applications listed promise to have a broad impact on the life sciences in the next years and, hopefully, they are just at the early stages of their use.

Bibliography

1. Duarte, N. C., Becker, S. A., Jamshidi, N., *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences* **104,** 1777–1782 (2007) (cit. on pp. 35, 42).
2. Feist, A. M., Henry, C. S., Reed, J. L., *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology* **3,** 121 (2007) (cit. on p. 35).
3. Feist, A. M. & Palsson, B. Ø. The growing scope of applications of genome-scale metabolic reconstructions: the case of E. coli. *Nature biotechnology* **26,** 659 (2008) (cit. on p. 35).
4. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Molecular systems biology* **5,** 320 (2009) (cit. on p. 35).
5. Arkin, A. P., Stevens, R. L., Cottingham, R. W., *et al.* The DOE Systems Biology Knowledgebase (KBase). *bioRxiv,* 096354 (2016) (cit. on p. 35).
6. Henry, C. S., DeJongh, M., Best, A. A., *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28,** 977–982 (2010) (cit. on p. 35).
7. Agren, R., Liu, L., Shoaie, S., *et al.* The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS computational biology* **9,** e1002980 (2013) (cit. on p. 35).
8. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5,** 93 (2010) (cit. on p. 36).

9. Agren, R., Mardinoglu, A., Asplund, A., *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology* **10,** 721 (2014) (cit. on pp. 37, 42).
10. Kanehisa, M., Goto, S., Hattori, M., *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research* **34,** D354–D357 (2006) (cit. on p. 37).
11. Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I. & Schomburg, D. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Research* **35,** D511–D514 (2007) (cit. on p. 37).
12. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* **11,** 213 (2010) (cit. on p. 37).
13. Degtyarenko, K., De Matos, P., Ennis, M., *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research* **36,** D344–D350 (2007) (cit. on p. 37).
14. Ganter, M., Bernard, T., Moretti, S., Stelling, J. & Pagni, M. MetaNetX. org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* **29,** 815–816 (2013) (cit. on p. 37).
15. Moretti, S., Martin, O., Van Du Tran, T., *et al.* MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic acids research* **44,** D523–D526 (2016) (cit. on p. 37).
16. Becker, S. A., Feist, A. M., Mo, M. L., *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols* **2,** 727–738 (2007) (cit. on p. 37).
17. Schellenberger, J., Que, R., Fleming, R. M., *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nature protocols* **6,** 1290–1307 (2011) (cit. on p. 37).
18. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology* **28,** 245–248 (2010) (cit. on pp. 38, 39).
19. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology* **19** (2001) (cit. on p. 40).

20. Fong, S. S. & Palsson, B. Ø. Metabolic gene–deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nature genetics* **36,** 1056–1058 (2004) (cit. on p. 40).

21. Janssen, D. B., Dinkla, I. J., Poelarends, G. J. & Terpstra, P. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environmental Microbiology* **7,** 1868–1882 (2005) (cit. on p. 42).

22. Ro, D.-K., Paradise, E. M., Ouellet, M., *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 940–943 (2006) (cit. on p. 42).

23. Peng, R.-H., Xiong, A.-S., Xue, Y., *et al.* Microbial biodegradation of polyaromatic hydrocarbons. *FEMS microbiology reviews* **32,** 927–955 (2008) (cit. on p. 42).

24. Rittmann, B. E. Opportunities for renewable bioenergy using microorganisms. *Biotechnology and bioengineering* **100,** 203–212 (2008) (cit. on p. 42).

25. Stolyar, S., Van Dien, S., Hillesland, K. L., *et al.* Metabolic modeling of a mutualistic microbial community. *Molecular systems biology* **3,** 92 (2007) (cit. on p. 42).

26. Zhuang, K., Izallalen, M., Mouser, P., *et al.* Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME journal* **5,** 305–316 (2011) (cit. on p. 42).

27. Thiele, I., Swainston, N., Fleming, R. M., *et al.* A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31,** 419–425 (2013) (cit. on p. 42).

28. De Oliveira Dal'Molin, C. G. & Nielsen, L. K. Plant genome-scale metabolic reconstruction and modelling. *Current opinion in biotechnology* **24,** 271–277 (2013) (cit. on p. 42).

29. Poolman, M. G., Miguet, L., Sweetlove, L. J. & Fell, D. A. A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant physiology* **151,** 1570–1581 (2009) (cit. on p. 42).

30. Mardinoglu, A., Agren, R., Kampf, C., *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications* **5,** 3083 (2014) (cit. on p. 42).

31. Pornputtapong, N., Nookaew, I. & Nielsen, J. Human metabolic atlas: an online resource for human metabolism. *Database* **2015,** bav068 (2015) (cit. on p. 42).

32. Agren, R., Bordel, S., Mardinoglu, A., *et al.* Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology* **8,** e1002518 (2012) (cit. on p. 42).

33. Oberhardt, M. A., Yizhak, K. & Ruppin, E. Metabolically re-modeling the drug pipeline. *Current opinion in pharmacology* **13,** 778–785 (2013) (cit. on p. 42).

Part II

RESULTS

5 Genome assembly and annotation

Genome sequencing, assembly and annotation, with their challenges and their routines, can be considered the entry point of any modern biological study. These data collection, which is always interesting, may become crucial when dealing with relatively unknown strains. The set of insights brought to light by sequencing analysis in fact, often determines further investigation directions.

The genome projects reported in this chapter belong to strains collected in poorly characterized niche and in extreme environments. For this reason, the aim of the works was to reconstruct the DNA sequence of the organisms of interest. In the next chapters examples of usage of these data will be presented. Here, the assemblies produced allowed the basic annotation of the genetic features and the detection of potentially relevant metabolic abilities. The reconstructed sequences represented the input of comparative studies which provided structural overviews and relevant clues on the biology of the targeted organisms. For instance, inferences on resistome and secondary metabolites production have been derived and set the stage for future phenotypic characterization.

frontiers in Microbiology

DATA REPORT
published: 12 September 2016
doi: 10.3389/fmicb.2016.01417

Arthrobacter sp. EpRS66 and *Arthrobacter* sp. EpRS71: Draft Genome Sequences from Two Bacteria Isolated from *Echinacea purpurea* Rhizospheric Soil

*Luana Presta[1], Marco Fondi[1], Elena Perrin[1], Isabel Maida[1], Elisangela Miceli[1], Carolina Chiellini[1], Valentina Maggini[1], Patrizia Bogani[1], Vincenzo Di Pilato[2], Gian M. Rossolini[3, 4, 5, 6], Alessio Mengoni[1] and Renato Fani[1]**

[1] Department of Biology, University of Florence, Florence, Italy, [2] Department of Surgery and Translational Medicine, University of Florence, Florence, Italy, [3] Department of Medical Biotechnologies, University of Siena, Siena, Italy, [4] Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy, [5] Clinical Microbiology and Virology Unit, Careggi University Hospital, Florence, Italy, [6] Don Carlo Gnocchi Foundation, Florence, Italy

Keywords: endophyte, medicinal plant, plant growth promoting, antibiotics, resistance

Edited by:
Marina G. Kalyuzhanaya, San Diego State University, USA

Reviewed by:
Marla Trindade, University of the Western Cape, South Africa
Ramón Suárez-Rodríguez, Universidad Autónoma del Estado de Morelos, Mexico

****Correspondence:***
Renato Fani
renato.fani@unifi.it;
renato.fani@virgilio.it

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 August 2016*
Accepted: *26 August 2016*
Published: *12 September 2016*

Citation:
Presta L, Fondi M, Perrin E, Maida I, Miceli E, Chiellini C, Maggini V, Bogani P, Di Pilato V, Rossolini GM, Mengoni A and Fani R (2016) Arthrobacter sp. EpRS66 and Arthrobacter sp. EpRS71: Draft Genome Sequences from Two Bacteria Isolated from Echinacea purpurea Rhizospheric Soil. Front. Microbiol. 7:1417. doi: 10.3389/fmicb.2016.01417

INTRODUCTION

One of the most promising, but still overlooked fields of microbiological research is represented by endophytic microorganisms, i.e., those organisms living in the tissues of host plants and/or in their rhizosphere (Rosenblueth and Martínez-Romero, 2006; Reinhold-Hurek and Hurek, 2011). These microbes are emerging as a new potential source of secondary metabolites and products, for exploitation in medicine, agriculture, and industry.

From a biotechnological perspective, a controlled (engineered) colonization of plant's tissues by some bacteria may be desirable because of their ability to produce a variety of plant growth promoting (PGP) molecules, spanning from siderophores, nitrogenases, hormones, and so on. In addition, host-microbe interaction confers indirect advantage to the plant, resulting from the inhibition activity exerted by the associated microbial community toward potential pathogens.

Moreover, in the field of medicine, there are examples of very well-known molecules derived by endophytes like antibiotics, antimycotics, and anticancer drugs. Also, it is still unknown if plant-associated bacteria may enhance (or be responsible for) some of the effects exerted by the extracts of medicinal plants (essential oils) (Kloepper and Ryu, 2006; Hardoim et al., 2008).

In this regard, in October 2012, in Casola Valsenio (Italy), a collection of microorganisms was isolated from both internal tissues and the rhizospheric soil of the medicinal plant *Echinacea purpurea*, as reported in Chiellini et al. (2014). Among others, two strains sampled from the rhizosphere and belonging to *Arthrobacter* species were identified, i.e., *Arthrobacter* sp. EpRS66 and *Arthrobacter* sp. EpRS71. Based on their antibiotic resistance profile, reported in Mengoni et al. (2014), and on further tests performed on these two strains, they were selected as good candidates for genome sequencing analysis. The last, will constitute a resource to deeply investigate their genomic features and to perform comparative genomics analysis. Moreover, in the aim of new drugs discovery, the genome sequence will facilitate the identification of putative genes responsible for the production of bioactive compounds.

MATERIALS AND METHODS

DNA Extraction and Sequencing

Arthrobacter sp. EpRS66 and *Arthrobacter* sp. EpRS71 strains were inoculated overnight on TSB medium at 30°C. Their genomic DNA was then extracted using the CTAB method (Perrin et al., 2015). Furthermore, the authenticity of the genomic DNA was confirmed by 16S RNA gene sequencing.

MiSeq sequencing system (Illumina Inc., San Diego, CA) was used to perform the whole genome shot-gun of the two organisms. The method used a 2 × 300 bp paired-end approach, which produced a genome coverage of 246.0 x for *Arthrobacter* sp. EpRS66 and 91x for *Arthrobacter* sp. EpRS71.

Genome Assembly and Annotation

The quality of the obtained read pairs was evaluated by inspecting them with FastQC software package v. 0.52 (Kunde-Ramamoorthy et al., 2014). Poor quality bases were removed with StreamingTrim (Bacci et al., 2014). *De novo* assembly was performed by using SPAdes 3.5 software (Bankevich et al., 2012) with a k-mer length of 21, 33, and 55. After, those contigs with length inferior to 2000 bp were trimmed and the remaining (6 and 24 for *Arthrobacter* sp. EPRS66 and *Arthrobacter* sp. EPRS71, respectively) were launched in a multi-draft based analysis through MeDuSa scaffolder (Bosi et al., 2015), by using as references 5 *Arthrobacter* genomes retrieved at NCBI database (*Arthrobacter arilaitensis* Re117, *Arthrobacter* FB24, *Arthrobacter* Rue61a, *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6).

Automated annotation of the two draft genome sequences has then been performed with NCBI Prokaryotic Genome Annotation Pipeline.

RESULTS

The last version of *Arthrobacter* sp. EpRS66 genome has a total length of 3,707,708 bp and embeds only 2 scaffolds (L50 equal to 1), with a mean G+C content of 59.27%. The annotation analysis identified a total of 3485 genes, of which 3383 have been annotated as coding DNA sequences (CDS), 29 as pseudogenes, 4 as rRNAs, 68 as tRNAs, and 1 as ncRNA.

The draft genome sequence of *Arthrobacter* sp. EpRS71 24 is 4,849,450 bp long and its contigs are set-up in 10 scaffolds (L50 equal to 1). The G+C content is 61.60%, a value slightly higher than the previous but still perfectly comparable with that of other *Arthrobacter* genomes sequenced so far. The annotation of *Arthrobacter* sp. EpRS71 genome revealed the presence of 4515 genes. This total amount includes 4379 proteins coding sequences, 71 pseudogenes, and 62 RNA (6 rRNAs, 55 tRNAs, 1 ncRNA) coding sequences.

Both genome sequences have been deposited at NCBI database and are available in both fasta and GenBank format; the GenBank accession number of *Arthrobacter* sp. EPRS66 is LNUU00000000 and the version reported in this work was named LNUU01000000; the GenBank accession number of *Arthrobacter* sp. EPRS71 is LNUV00000000 and the version reported in this work is LNUV01000000.

AUTHOR CONTRIBUTIONS

This project was planned by RF and AM. The DNA extraction was performed by IM and EP. The DNA sequencing has been performed by GR and VD. The data processing has been performed by LP and MF. CC, VM, PB, and EM assisted substantially on the technical part of this work. All author contributed to writing and editing the present manuscript.

FUNDING

This work was supported financially by Ente Cassa di Risparmio di Firenze (Project 2013.0657).

REFERENCES

Bacci, G., Bazzicalupo, M., Benedetti, A., and Mengoni, A. (2014). StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies. *Mol. Ecol. Resour.* 14, 426–434. doi: 10.1111/1755-0998.12187

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M.-F., Lió, P., et al. (2015). MeDuSa: a multi-draft based scaffolder. *Bioinformatics* 31, 2443–2451. doi: 10.1093/bioinformatics/btv171

Chiellini, C., Maida, I., Emiliani, G., Mengoni, A., Stefano, M., Fabiani, A., et al. (2014), Endophytic and rhizospheric bacterial communities isolated from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia*. *Int. Microbiol.* 17, 165–174. doi: 10.2436/20.1501.01.219

Hardoim, P. R., van Overbeek, L. S., and Elsas van, J. D. (2008). Properties of bacterial endophytes and their proposed role in plant growth. *Trends Microbiol.* 16, 463–471. doi: 10.1016/j.tim.2008.07.008

Kloepper, J. W., and Ryu, C.-M. (2006). "Bacterial endophytes as elicitors of induced systemic resistance," in *Microbial Root Endophytes*, eds B. Schulz, C. Boyle, and T. N. Sieber (Berlin; Heidelberg: Springer), 33–52. doi: 10.1007/3-540-33526-9_3

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N. J., Harris, R. A., Xu, M., et al. (2014). Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res.* 42:e43. doi: 10.1093/nar/gkt1325

Mengoni, A., Maida, I., Chiellini, C., Emiliani, G., Mocali, S., Fabiani, A., et al. (2014). Antibiotic resistance differentiates *Echinacea purpurea* endophytic bacterial communities with respect to plant organs. *Res. Microbiol.* 165, 686–694. doi: 10.1016/j.resmic.2014.09.008

Perrin, E., Fondi, M., Maida, I., Mengoni, A., Chiellini, C., Mocali, S., et al. (2015). Genomes analysis and bacteria identification: the use of overlapping genes as molecular markers. *J. Microbiol. Methods* 117, 108–112. doi: 10.1016/j.mimet.2015.07.025

Reinhold-Hurek, B., and Hurek, T. (2011). Living inside plants: bacterial endophytes. *Curr. Opin. Plant Biol.* 14, 435–443. doi: 10.1016/j.pbi.2011.04.004

Rosenblueth, M., and Martínez-Romero, E. (2006). Bacterial endophytes and their interactions with hosts. *Mol. Plant. Microbe. Interact.* 19, 827–837. doi: 10.1094/MPMI-19-0827

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Draft Genome Sequence of *Pseudomonas* sp. EpS/L25, Isolated from the Medicinal Plant *Echinacea purpurea* and Able To Synthesize Antimicrobial Compounds

Luana Presta,[a] Emanuele Bosi,[a] Marco Fondi,[a] Isabel Maida,[a] Elena Perrin,[a] Elisangela Miceli,[a] Valentina Maggini,[a,b] Patrizia Bogani,[a] Fabio Firenzuoli,[b] Vincenzo Di Pilato,[c] Gian Maria Rossolini,[d,e,f,g] Alessio Mengoni,[a] Renato Fani[a]

Department of Biology, University of Florence, Florence, Italy[a]; Center for Integrative Medicine, Careggi University Hospital, University of Florence, Florence, Italy[b]; Department of Surgery and Translational Medicine, University of Florence, Florence, Italy[c]; Department of Medical Biotechnologies, University of Siena, Siena, Italy[d]; Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy[e]; Clinical Microbiology and Virology Unit, Careggi University Hospital, Florence, Italy[f]; Don Carlo Gnocchi Foundation, Florence, Italy[g]

We announce here the draft genome sequence of *Pseudomonas* sp. strain EpS/L25, isolated from the stem/leaves of the medicinal plant *Echinacea purpurea*. This genome will allow for comparative genomics in order to identify genes associated with the production of bioactive compounds and antibiotic resistance.

Received 15 March 2016 **Accepted** 16 March 2016 **Published** 5 May 2016

Citation Presta L, Bosi E, Fondi M, Maida I, Perrin E, Miceli E, Maggini V, Bogani P, Firenzuoli F, Di Pilato V, Rossolini GM, Mengoni A, Fani R. 2016. Draft genome sequence of *Pseudomonas* sp. EpS/L25, isolated from the medicinal plant *Echinacea purpurea* and able to synthesize antimicrobial compounds. Genome Announc 4(3):e00346-16. doi:10.1128/genomeA.00346-16.

Address correspondence to Renato Fani, renato.fani@unifi.it.

The genus *Pseudomonas* consists of a group of bacteria particularly relevant from both medical and biotechnological viewpoints (1). Thanks to their metabolic versatility, they successfully colonized several different niches, including water, soil, plants, and animals. Here, we present the draft genome sequence of *Pseudomonas* sp. EpS/L25, a strain close to *Pseudomonas oleovorans*, isolated from the stem/leaves of *Echinacea purpurea*, a medicinal plant whose essential oil possesses antimicrobial activity (2). The *E. purpurea* plants were collected in October 2012 (3) at the "Giardino delle Erbe," Casola Valsenio. Medicinal plants are known for their beneficial effects for humans (including their antibacterial activity), but, in spite of their high relevance, endophytic bacterial communities inhabiting their rhizosphere or internal tissues are almost totally unknown. Thus, it is still unknown if they contribute to the antimicrobial activity exerted by *E. purpurea* extracts.

Previous characterization of *Pseudomonas* sp. EpS/L25 revealed the ability of this strain to inhibit the growth of other *E. purpurea*-associated bacteria (4) and, more interestingly, some opportunistic bacterial pathogens belonging to the *Burkholderia cepacia* complex. Furthermore, it showed resistance to several antibiotic compounds (5). Due to these properties, it represents a good candidate for further molecular investigations on the genetic basis of such features, prompting for sequencing of its genome.

The genome sequence of *Pseudomonas* sp. EpS/L25 was determined by a 2 × 300-bp paired-end approach using the MiSeq sequencing system (Illumina Inc., San Diego, CA, USA). A total of 3,020,786 paired-end reads were obtained, representing approximately 158× coverage of the whole genome. *De novo* assembly was performed using SPAdes version 3.5 (6), which generated 300 contigs. Contigs with length less than 2,000 bp were discarded and the remaining ones used for a multi-draft-based analysis using 16 *Pseudomonas* genomes retrieved from the NCBI database (*Pseudomonas* ND6, *Pseudomonas* TKP, *Pseudomonas* VLB120, *P. aeruginosa* B136 33, *P. aeruginosa* UCBPP PA14, *P. brassicacearum* NFM421, *P. denitrificans* ATCC 13867, *P. entomophila* L48, *P. fluorescens* R124, *P. mendocina* NK 01, *P. poae* RE 1 1 14, *P. putida* BIRD 1, *P. putida* KT2440, *P. stutzeri* CCUG 29243, *P. syringae* B728a) through MeDuSa scaffolder (7). The final version of the genome embeds 18 scaffolds, the longest of which is 1,664,566 bp long. The draft genome assembly of *Pseudomonas* sp. EpS/L25 has a total length of 5,435,234 bp. The G+C content is 65.5%, similar to that of other *Pseudomonas* genomes. Automated annotation of the *Pseudomonas* sp. EpS/L25 draft genome sequence using NCBI Prokaryotic Genome Annotation Pipeline detected 4,690 protein coding genes, 76 RNA coding genes (5 complete rRNAs, 57 tRNAs, 14 ncRNAs), and 105 pseudogenes. Three CRISPR arrays were also identified.

Comparative genomics analysis confirmed the presence of antibiotic efflux pumps, some conferring specific resistance to beta-lactams (*pdc*), florfenicol (*cfrA*), and polymyxins (*arnA* and *pmrF*). Moreover, genes involved in the production of secondary metabolites with antimicrobial activity have also been detected (terpene, aryl-polyene, and two nonribosomal peptides).

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at GenBank under the accession number LNUP00000000. The version described in this paper is the first version, LNUP01000000.

REFERENCES:

1. **Nikel PI, Martínez-García E, de Lorenzo V.** 2014. Biotechnological domestication of pseudomonads using synthetic biology. Nat Rev Microbiol 12:368–379. http://dx.doi.org/10.1038/nrmicro3253.

2. **Hudson JB**. 2012. Applications of the phytomedicine *Echinacea purpurea* (purple coneflower) in infectious diseases. BioMed Res Int **2012:** http://dx.doi.org/10.1155/2012/769896.
3. **Chiellini C, Maida I, Emiliani G, Mengoni A, Mocali S, Fabiani A, Biffi S, Maggini V, Gori L, Vannacci A, Gallo E, Firenzuoli F, Fani R**. 2014. Endophytic and rhizospheric bacterial communities isolated from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia*. Int Microbiol **17:**165–174. http://dx.doi.org/10.2436/20.1501.01.219.
4. **Maida I, Chiellini C, Mengoni A, Bosi E, Firenzuoli F, Fondi M, Fani R**. 22 July 2015. Antagonistic interactions between endophytic cultivable bacterial communities isolated from the medicinal plant *Echinacea purpurea*. Environ Microbiol [Epub ahead of print.] http://dx.doi.org/10.1111/1462-2920.12911.
5. **Mengoni A, Maida I, Chiellini C, Emiliani G, Mocali S, Fabiani A, Fondi M, Firenzuoli F, Fani R**. 2014. Antibiotic resistance differentiates *Echinacea purpurea* endophytic bacterial communities with respect to plant organs. Res Microbiol **165:**686–694. http://dx.doi.org/10.1016/j.resmic.2014.09.008.
6. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA**. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
7. **Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, Crescenzi P, Fani R, Fondi M**. 2015. MeDuSa: a multi-draft based scaffolder. Bioinformatics **31:**2443–2451. http://dx.doi.org/10.1093/bioinformatics/btv171.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

New Genome Sequence of an *Echinacea purpurea* Endophyte, *Arthrobacter* sp. Strain EpSL27, Able To Inhibit Human-Opportunistic Pathogens

Elisangela Miceli,[a] Luana Presta,[a] Valentina Maggini,[a,b,c] Marco Fondi,[a] Emanuele Bosi,[a] Carolina Chiellini,[a] Camilla Fagorzi,[a] Patrizia Bogani,[a] Vincenzo Di Pilato,[d] Gian Maria Rossolini,[e] Alessio Mengoni,[a] Fabio Firenzuoli,[c] Elena Perrin,[a] Renato Fani[a]

Department of Biology, University of Florence, Florence, Italy[a]; Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy[b]; Center for Integrative Medicine, Careggi University Hospital, University of Florence, Florence, Italy[c]; Department of Surgery and Translational Medicine, University of Florence, Florence, Italy[d]; Clinical Microbiology and Virology Unit, Careggi University Hospital, Florence, Italy[e]

ABSTRACT We announce here the draft genome sequence of *Arthrobacter* sp. strain EpSL27, isolated from the stem and leaves of the medicinal plant *Echinacea purpurea* and able to inhibit human-pathogenic bacterial strains. The genome sequencing of this strain may lead to the identification of genes involved in the production of antimicrobial molecules.

Received 2 May 2017 **Accepted** 3 May 2017 **Published** 22 June 2017

Citation Miceli E, Presta L, Maggini V, Fondi M, Bosi E, Chiellini C, Fagorzi C, Bogani P, Di Pilato V, Rossolini GM, Mengoni A, Firenzuoli F, Perrin E, Fani R. 2017. New genome sequence of an *Echinacea purpurea* endophyte, *Arthrobacter* sp. strain EpSL27, able to inhibit human-opportunistic pathogens. Genome Announc 5:e00565-17. https://doi.org/10.1128/genomeA.00565-17.

Address correspondence to Renato Fani, renato.fani@unifi.it.

E.M. and L.P. contributed equally to this work.

Medicinal plants are well known and have been largely explored for centuries for their therapeutic properties (1). What is little known is that their therapeutic potential could be related to endophytic microorganisms inhabiting their tissues (2). Many bioactive molecules have been already extracted from endophytic bacteria (3). The promising potential of such organisms has led to the characterization of endophytic bacterial communities from medicinal plants, which are poorly known. Endophytic and rhizospheric bacterial communities from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia* have been characterized, highlighting the specific composition of such communities within plants' compartments (4). *Arthrobacter* sp. strain EpSL27, extracted from the stem and leaves of *E. purpurea*, has been evidenced as being resistant to a high level of oxidative stress (20 mM H_2O_2) and is able to degrade diesel fuel. Among such notable biotechnological potentialities, *Arthrobacter* sp. EpSL27 has also been found to show strong inhibition activity toward human-pathogenic bacteria from the *Burkholderia cepacia* complex (5), which are multidrug-resistant organisms able to induce serious infections in immunocompromised patients.

The intriguing information obtained by the above-cited analyses led to whole sequencing of the strain genome.

Arthrobacter sp. EpSL27 genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method (6), and its authenticity has been confirmed by 16S rRNA gene sequencing. Whole-genome shotgun sequencing was performed with a 2 × 300-bp paired-end approach using the MiSeq sequencing system (Illumina, Inc., San Diego, CA). The FastQC software package version 0.52 (7) was used to evaluate the quality of the obtained read pairs, and poor-quality bases were removed using StreamingTrim (8). Assembly was performed using the SPAdes 3.5 software (9), with k-mer lengths of 21, 33, and 55, generating 21 contigs. Those having a length shorter than 200 nucleotides were removed and the others launched for scaffolding through Medusa software (10), using the following genomes as references: *Arthrobacter arilaitensis* Re117 (11), *Arthrobacter* Rue61a (12), *Arthrobacter* sp. strain FB24 (13), *Arthrobacter*

aurescens TC1 (14), and *Arthrobacter chlorophenolicus* A6. The resulting scaffolds were then annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAAP) (15). The final version of the *Arthrobacter* sp. EpSL27 draft genome consists of 8 scaffolds, and its total length is 4,176,054 bp, with a coverage of 215.0×. The G+C content is about 67.8%, which reflects the characteristic high G+C content of the genus. The *Arthrobacter* sp. EpSL27 genome harbors 3,758 genes, 3,610 of which are protein-coding genes, 66 are RNA-coding genes (5 5S rRNA, 1 23S rRNA, 1 16S rRNA, 50 tRNAs, and 9 noncoding RNA [ncRNA]), and 91 are pseudogenes.

The EpSL27 genome was analyzed using CARD (16) for the presence of genes conferring antibiotic resistance. The analysis has evidenced genes putatively involved in specific antibiotic resistance to isoniazid (*Mycobacterium tuberculosis kasA* mutant), fluoroquinolones (*mfd*), amynocoumarin (*Streptomyces rishiriensis parY* mutant), rifamycin (*rphB*), mupirocin (*Bifidobacterium* intrinsic *ileS*), and fosfomycin (*Chlamydia trachlomatis* intrinsic *murA*). antiSMASH (17) analysis for secondary metabolites with antimicrobial activities was also performed, revealing the presence of 5 clusters, with one cluster encoding nonribosomal peptide synthetase (NRPS), one cluster encoding type 3 polyketide synthase (T3pks), and another three clusters with an unspecified reference.

Accession number(s). The whole-genome shotgun project has been deposited at NCBI whole-genome sequencing (WGS) database under accession number LNUT00000000, and the version reported in this work is version LNUT00000000.1.

REFERENCES

1. Staub PO, Casu L, Leonti M. 2016. Back to the roots: a quantitative survey of herbal drugs in Dioscorides' de materia medica (ex Matthioli, 1568). Phytomedicine 23:1043–1052. https://doi.org/10.1016/j.phymed.2016.06.016.
2. Ryan RP, Germaine K, Franks A, Ryan DJ, Dowling DN. 2008. Bacterial endophytes: recent developments and applications. FEMS Microbiol Lett 278:1–9. https://doi.org/10.1111/j.1574-6968.2007.00918.x.
3. Shweta S, Bindu JH, Raghu J, Suma HK, Manjunatha BL, Kumara PM, Ravikanth G, Nataraja KN, Ganeshaiah KN, Uma Shaanker R. 2013. Isolation of endophytic bacteria producing the anti-cancer alkaloid camptothecin from *Miquelia dentata* Bedd. (Icacinaceae). Phytomedicine 20:913–917. https://doi.org/10.1016/j.phymed.2013.04.004.
4. Chiellini C, Maida I, Emiliani G, Mengoni A, Mocali S, Fabiani A, Biffi S, Maggini V, Gori L, Vannacci A, Gallo E, Firenzuoli F, Fani R. 2014. Endophytic and rhizospheric bacterial communities isolated from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia*. Int Microbiol 17:165–174. https://doi.org/10.2436/20.1501.01.219.
5. Chiellini C, Maida I, Maggini V, Bosi E, Mocali S, Emiliani G, Perrin E, Firenzuoli F, Mengoni A, Fani R. 2017. Preliminary data on antibacterial activity of *Echinacea purpurea*-associated bacterial communities against *Burkholderia cepacia* complex strains, opportunistic pathogens of cystic fibrosis patients. Microbiol Res 196:34–43. https://doi.org/10.1016/j.micres.2016.12.001.
6. Perrin E, Fondi M, Maida I, Mengoni A, Chiellini C, Mocali S, Cocchi P, Campana S, Taccetti G, Vaneechoutte M, Fani R. 2015. Genomes analysis and bacteria identification: the use of overlapping genes as molecular markers. J Microbiol Methods 117:108–112. https://doi.org/10.1016/j.mimet.2015.07.025.
7. Kunde-Ramamoorthy G, Coarfa C, Laritsky E, Kessler NJ, Harris RA, Xu M, Chen R, Shen L, Milosavljevic A, Waterland RA. 2014. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res 42:e43. https://doi.org/10.1093/nar/gkt1325.
8. Bacci G, Bazzicalupo M, Benedetti A, Mengoni A. 2014. StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies. Mol Ecol Resour 14:426–434. https://doi.org/10.1111/1755-0998.12187.
9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
10. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, Crescenzi P, Fani R, Fondi M. 2015. Medusa: a multi-draft based scaffolder. Bioinformatics 31:2443–2451. https://doi.org/10.1093/bioinformatics/btv171.
11. Monnet C, Loux V, Gibrat JF, Spinnler E, Barbe V, Vacherie B, Gavory F, Gourbeyre E, Siguier P, Chandler M, Elleuch R, Irlinger F, Vallaeys T. 2010. The *Arthrobacter arilaitensis* Re117 genome sequence reveals its genetic adaptation to the surface of cheese. PLoS One 5:e15489. https://doi.org/10.1371/journal.pone.0015489.
12. Niewerth H, Schuldes J, Parschat K, Kiefer P, Vorholt JA, Daniel R, Fetzner S. 2012. Complete genome sequence and metabolic potential of the quinaldine-degrading bacterium *Arthrobacter* sp. Rue61a. BMC Genomics 13:534. https://doi.org/10.1186/1471-2164-13-534.
13. Nakatsu CH, Barabote R, Thompson S, Bruce D, Detter C, Brettin T, Han C, Beasley F, Chen W, Konopka A, Xie G. 2013. Complete genome sequence of *Arthrobacter* sp. strain FB24. Stand Genomic Sci 9:106–116. https://doi.org/10.4056/sigs.4438185.
14. Mongodin EF, Shapir N, Daugherty SC, DeBoy RT, Emerson JB, Shvartzbeyn A, Radune D, Vamathevan J, Riggs F, Grinberg V, Khouri H, Wackett LP, Nelson KE, Sadowsky MJ. 2006. Secrets of soil survival revealed by the genome sequence of *Arthrobacter aurescens* TC1. PLoS Genet 2:e214. https://doi.org/10.1371/journal.pgen.0020214.
15. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O. 2008. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS 12:137–141. https://doi.org/10.1089/omi.2008.0017.
16. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother 57:3348–3357. https://doi.org/10.1128/AAC.00419-13.
17. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. 2015. antiSMASH 3.0–a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243. https://doi.org/10.1093/nar/gkv437.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Draft Genome Sequences of the Antimicrobial Producers *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 Isolated from Antarctic Sediments

Luana Presta,[a] Ilaria Inzucchi,[a] Emanuele Bosi,[a] Marco Fondi,[a] Elena Perrin,[a] Isabel Maida,[a] Elisangela Miceli,[a] Maria Luisa Tutino,[c] Angelina Lo Giudice,[d,e] Donatella de Pascale,[b] Renato Fani[a]

Department of Biology, University of Florence, Florence, Italy[a]; Institute of Protein Biochemistry, National Research Council, Naples, Italy[b]; Department of Chemical Sciences, University of Naples Federico II, Naples, Italy[c]; Institute for the Coastal Marine Environment, National Research Council (IAMC-CNR), Messina, Italy[d]; Department of Biological and Environmental Sciences, University of Messina, Messina, Italy[e]

L.P. and I.I. contributed equally to this article.

We report here the draft genome sequence of the *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 strains, isolated from Antarctic sediments during a summer campaign near coastal areas of Terra Nova Bay (Antarctica). Genome sequence knowledge allowed the identification of genes associated with the production of bioactive compounds and antibiotic resistance. Furthermore, it will be instrumental for comparative genomics and the fulfillment of both basic and application-oriented investigations.

Received 1 June 2016 **Accepted** 3 June 2016 **Published** 28 July 2016

Citation Presta L, Inzucchi I, Bosi E, Fondi M, Perrin E, Maida I, Miceli E, Tutino ML, Lo Giudice A, de Pascale D, Fani R. 2016. Draft genome sequences of the antimicrobial producers *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 isolated from Antarctic sediments. Genome Announc 4(4):e00728-16. doi:10.1128/genomeA.00728-16.

Address correspondence to Renato Fani, renato.fani@unifi.it.

Antarctica provides one of the largest unexplored sources of biodiversity. Here, the continuous environmental challenges led to extremely adapted living forms that may be sources of potentially novel, untapped gene functions. Particularly, it has been shown how microorganisms are claimed to be a reservoir of biotechnologically relevant molecules, such as antibiotics (1–6).

Here, we report the genome sequences of two *Pseudomonas* sp. strains, TAA207 and TAD18, isolated from Antarctic sediments during a summer campaign near the coastal areas of Terra Nova Bay (Antarctica). These bacteria have been screened for antimicrobial activity against human pathogens. The results obtained show how they totally inhibited 40 strains belonging to the *Burkholderia cepacia* complex (BCC), most of which are affiliated to the species *Burkholderia cenocepacia* and *Burkholderia multivorans*, two of the most important pathogens in immunocompromised patients affected by cystic fibrosis disease. Also, they produce antibiofilm molecules acting against *Staphylococcus aureus* and *Pseudomonas aeruginosa* (7).

Both genome sequences of *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 were determined through a paired-end approach using the Illumina/Solexa genome analyzer IIx platform at the Institute of Applied Genomics and IGA Technology Services Srl (University of Udine, Italy). A total of 11,007,120 and of 12,698,315 reads were obtained for *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18, respectively. Low-quality sequences were trimmed with StreamingTrim (8), and the remaining were assembled with SPAdes genome assembler version 3.6.1 (9). Only contigs longer than 1,000 bp were embedded in the final version of the draft genomes, which are 4,900,197-bp long for *Pseudomonas* sp. TAA207 (72 contigs, 453× average coverage, 57% GC content) and 4,917,586-bp long for *Pseudomonas* sp. TAD18 (82 contigs, average coverage 521×, 57.24% GC content).

Annotation was then performed using Prokka (10), which identified 4,379 and 4,403 genes for *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18, respectively. Among these, 4,028 are protein-encoding genes in the former organism, and 4,220 are in the latter one.

The genome sequences allowed comparative genomics analysis to check for the presence of genetic traits involved in secondary metabolite biosynthesis. The analysis was performed within the antiSMASH shell (11), revealing that both genomes harbor gene clusters encoding molecules involved in inhibitory activities. Particularly, the two strains embed gene clusters similar to those coding for aryl-polyene, terpene, bacteriocin, and nonribosomal peptide synthase. Additionally, *Pseudomonas* sp. TAA207 contains a cluster involved in microcin production.

Further, we investigated the possibility that both strains possess antibiotic resistance genes in their pool by probing their sequences in the Comprehensive Antibiotic Resistance Database (CARD). The outcome yields strong indications that both genomes have genes coding for general efflux pumps, alongside several genes conferring resistance to specific classes of antibiotics, including chloramphenicol, fluoroquinilone, beta-lactam, trimethoprim, tetracycline, polymyxin, aminoglycoside, and rifampin.

Nucleotide sequence accession numbers. The whole-genome shotgun projects of *Pseudomonas* sp. TAA207 and *Pseudomonas* sp. TAD18 have been deposited at GenBank under the accession numbers LLWJ00000000 and LLWI00000000, respectively. The

versions described in this paper are the first versions, LLWJ01000000 and LLWI01000000.

ACKNOWLEDGMENTS

This work was financially supported by grants from the Italian Cystic Fibrosis Research Foundation (grant FFC no. 12/2011), by two PNRA (Programma Nazionale per la Ricerca in Antartide) grants (PNRA 2013/B4.02 and PNRA 2013/AZ1.04), and by the EU KBBE Project Pharmasea 2012–2016 (grant agreement no. 312184).

REFERENCES

1. **Mangano S, Michaud L, Caruso C, Brilli M, Bruni V, Fani R, Lo Giudice AL**. 2009. Antagonistic interactions between psychrotrophic cultivable bacteria isolated from Antarctic sponges: a preliminary analysis. Res Microbiol **160:**27–37. http://dx.doi.org/10.1016/j.resmic.2008.09.013.
2. **Papaleo MC, Perrin E, Maida I, Fondi M, Fani R, Vandamme P**. 2010. Identification of species of the *Burkholderia cepacia* complex by sequence analysis of the *hisA* gene. J Med Microbiol **59:**1163–1170. http://dx.doi.org/10.1099/jmm.0.019844-0.
3. **Papaleo MC, Fondi M, Maida I, Perrin E, Lo Giudice AL, Michaud L, Mangano S, Bartolucci G, Romoli R, Fani R**. 2012. Sponge-associated microbial Antarctic communities exhibiting antimicrobial activity against *Burkholderia cepacia* complex bacteria. Biotechnol Adv **30:**272–293. http://dx.doi.org/10.1016/j.biotechadv.2011.06.011.
4. **Maida I, Fondi M, Papaleo MC, Perrin E, Orlandini V, Emiliani G, de Pascale D, Parrilli E, Tutino ML, Michaud L, Lo Giudice A, Romoli R, Bartolucci G, Fani R**. 2014. Phenotypic and genomic characterization of the Antarctic bacterium *Gillisia* sp. CAL575, a producer of antimicrobial compounds. Extremophiles **18:**35–49. http://dx.doi.org/10.1007/s00792-013-0590-0.
5. **Lo Giudice A, Fani R**. 2015. Cold-adapted bacteria from a coastal area of the Ross Sea (Terra Nova Bay, Antarctica): linking microbial ecology to biotechnology. Hydrobiologia **761:**417–441. http://dx.doi.org/10.1007/s10750-015-2497-5.
6. **Presta L, Inzucchi I, Bosi E, Fondi M, Perrin E, Miceli E, Tutino ML, Lo Giudice A, de Pascale D, Fani R**. 2016. Draft genome sequence of *Flavobacterium* sp. strain TAB 87, able to inhibit the growth of cystic fibrosis bacterial pathogens belonging to the *Burkholderia cepacia* complex. Genome Announc **4**(3): http://dx.doi.org/10.1128/genomeA.00410-16.
7. **Papa R, Selan L, Parrilli E, Tilotta M, Sannino F, Feller G, Tutino ML, Artini M**. 2015. Anti-biofilm activities from marine cold adapted Bacteria against staphylococci and *Pseudomonas aeruginosa*. Front Microbiol **6:**1333. http://dx.doi.org/10.3389/fmicb.2015.01333.
8. **Bacci G, Bazzicalupo M, Benedetti A, Mengoni A**. 2014. StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies. Mol Ecol Resour **14:**426–434. http://dx.doi.org/10.1111/1755-0998.12187.
9. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA**. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
10. **Seemann T**. 2012. Prokka: prokaryotic genome annotation system. Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria, Australia.
11. **Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R**. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res **39**(suppl 2):W339–W346. http://dx.doi.org/10.1093/nar/gkr466.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Draft Genome Sequence of *Flavobacterium* sp. Strain TAB 87, Able To Inhibit the Growth of Cystic Fibrosis Bacterial Pathogens Belonging to the *Burkholderia cepacia* Complex

Luana Presta,[a] Ilaria Inzucchi,[a] Emanuele Bosi,[a] Marco Fondi,[a] Elena Perrin,[a] Elisangela Miceli,[a] Maria Luisa Tutino,[c] Angelina Lo Giudice,[d,e] Donatella de Pascale,[b] Renato Fani[a]

Department of Biology, University of Florence, Florence, Italy[a]; Institute of Protein Biochemistry, National Research Council, Naples, Italy[b]; Department of Chemical Sciences, University of Naples Federico II, Naples, Italy[c]; Institute for the Coastal Marine Environment, National Research Council (IAMC-CNR), Messina, Italy[d]; Department of Biological and Environmental Sciences, University of Messina, Messina, Italy[e]

L.P. and I.I. contributed equally to this work.

We report here the draft genome sequence of the *Flavobacterium* sp. TAB 87 strain, isolated from Antarctic seawater during a summer campaign near the French Antarctic station Dumont d'Urville (60°40′S, 40°01′E). It will allow for comparative genomics and the fulfillment of both fundamental and application-oriented investigations. It allowed the recognition of genes associated with the production of bioactive compounds and antibiotic resistance.

Received 1 April 2016 **Accepted** 4 April 2016 **Published** 19 May 2016

Citation Presta L, Inzucchi I, Bosi E, Fondi M, Perrin E, Miceli E, Tutino ML, Lo Giudice A, de Pascale D, Fani R. 2016. Draft genome sequence of *Flavobacterium* sp. strain TAB 87, able to inhibit the growth of cystic fibrosis bacterial pathogens belonging to the *Burkholderia cepacia* complex. Genome Announc 4(3):e00410-16 doi:10.1128/genomeA.00410-16.

Address correspondence to Donatella de Pascale, d.depascale@ibp.cnr.it, or Renato Fani, renato.fani@virgilio.it.

Antarctic bacteria are cold-adapted microorganisms that have evolved peculiar features to overcome barriers for growth at low temperatures. They are driving scientific interest, both in the field of ecological sciences as they play a key role in maintaining proper ecosystem functions, and in the clinical setting, since they are known to produce molecules able to exert antibacterial activity in order to withstand strongly adapted competitors. Indeed, it was recently demonstrated that many Antarctic bacteria exhibited the ability to counteract the growth of other Antarctic strains (1) and, more intriguingly, of some human pathogens belonging to the *Burkholderia cepacia* complex (BCC) (2–5), which represent a serious threat among immunocompromised patients, especially those affected by cystic fibrosis (CF).

Here, we report the draft genome sequence of *Flavobacterium* sp. strain TAB 87, a Gram-negative bacterium belonging to the family *Flavobacteriaceae* (6). The strain was isolated from seawater during a summer campaign near the French Antarctic station Dumont d'Urville (60°40′ S, 40°01′ E). The genome analysis of this Antarctic strain enables both fundamental and application-oriented investigations. Indeed, this strain completely inhibited the growth of 40 BCC strains belonging to 18 different bacterial species, most of which belonged to the species *Burkholderia cenocepacia* and *Burkholderia multivorans*, two of the most important CF pathogens. Moreover, some of the antimicrobial compounds produced were volatile organic compounds (VOCs), according to previous observations (3, 4). The draft genome sequence of *Flavobacterium* sp. TAB 87 was determined by the Institute of Applied Genomics and IGA Technology Services Srl (University of Udine, Italy) through a paired-end approach using an Illumina (Solexa) Genome Analyzer II platform. A total of 19,040,534 paired-end reads (average coverage, 1,004×) were initially obtained, those with low quality were trimmed with Streaming Trim (7), and those remaining were assembled with SPAdes genome assembler version 3.6.1 (8), which generated a total of 5,056 contigs. Those contigs <1,000 bp were discarded, while the others were embedded in the final version of the draft genome, which is 3,827,405 bp long and harbors 38 contigs (the longest of which is 1,014,695 bp long). The G+C content is 65.5%, similar to that of other *Flavobacterium* genomes sequenced so far. Annotation was performed by using Prokka (9), which, among all the predicted genes (3,365), identified a total of 3,323 protein-coding genes, 3 rRNA-coding genes, and 39 tRNA-coding genes.

Moreover, we screened the genome sequence for the presence of genetic traits involved in secondary metabolite biosynthesis. The analysis was performed within antiSMASH shell (10), revealing that the *Flavobacterium* sp. TAB 87 genome harbors four interesting gene clusters: a type I and a type III polyketide synthase (PKS) and two terpene biosynthetic gene clusters. Additionally, the genome sequence was analyzed through CARD (11), leading to the identification of *cfrA* and *Staphylococcus aureus* parE, two genes conferring resistance to florfenicol and fluoroquinolones, respectively.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at GenBank under the accession no. LLWK00000000. The version described in this paper is version LLWK01000000.

ACKNOWLEDGEMENTS

We are grateful to Georges Feller, University of Liege, Belgium, for the gift of TAB 87 strain.

This work was financially supported by grants from the Italian Cystic Fibrosis Research Foundation (grant FFC# 12/2011), by two Programma

Nazionale per la Ricerca in Antartide (PNRA) grants (PNRA 2013/B4.02 and PNRA 2013/AZ1.04), and by the EU KBBE Project Pharmasea 2012–2016, grant agreement 312184.

FUNDING INFORMATION

This work was financially supported by grants from the Italian Cystic Fibrosis Research Foundation (grant FFC#12/2011), by two PNRA (Programma Nazionale per la Ricerca in Antartide) grants (PNRA 2013/B4.02 and PNRA 2013/AZ1.04), and by the EU KBBE Project Pharmasea 2012-2016 (grant agreement no. 312184).

REFERENCES

1. **Mangano S, Michaud L, Caruso C, Brilli M, Bruni V, Fani R, Lo Giudice AL**. 2009. Antagonistic interactions between psychrotrophic cultivable bacteria isolated from Antarctic sponges: a preliminary analysis. Res Microbiol **160:**27–37. http://dx.doi.org/10.1016/j.resmic.2008.09.013.
2. **Papaleo MC, Perrin E, Maida I, Fondi M, Fani R, Vandamme P**. 2010. Identification of species of the *Burkholderia cepacia* complex by sequence analysis of the *hisA* gene. J Med Microbiol **59:**1163–1170. http://dx.doi.org/10.1099/jmm.0.019844-0.
3. **Papaleo MC, Fondi M, Maida I, Perrin E, Lo Giudice AL, Michaud L, Mangano S, Bartolucci G, Romoli R, Fani R**. 2012. Sponge-associated microbial Antarctic communities exhibiting antimicrobial activity against *Burkholderia cepacia* complex bacteria. Biotechnol Adv **30:**272–293. http://dx.doi.org/10.1016/j.biotechadv.2011.06.011.
4. **Maida I, Fondi M, Papaleo MC, Perrin E, Orlandini V, Emiliani G, de Pascale D, Parrilli E, Tutino ML, Michaud L, Lo Giudice A, Romoli R, Bartolucci G, Fani R**. 2014. Phenotypic and genomic characterization of the Antarctic bacterium *Gillisia* sp. CAL575, a producer of antimicrobial compounds. Extremophiles **18:**35–49. http://dx.doi.org/10.1007/s00792-013-0590-0.
5. **Lo Giudice A, Fani R**. 2015. Cold-adapted bacteria from a coastal area of the Ross Sea (Terra Nova Bay, Antarctica): linking microbial ecology to biotechnology. Hydrobiologia **761:**417–441. http://dx.doi.org/10.1007/s10750-015-2497-5.
6. **Holmes B, Owen RJ, McMeekin TA**. 1984. Genus *Flavobacterium* Bergey, Harrison, Breed, Hammer and Huntoon. 1923:97^{AL}, p 353–361. *In* Krieg NR, Holt JG (ed), Bergey's manual of systematic bacteriology, vol 1. Williams & Wilkins, Baltimore, MD.
7. **Bacci G, Bazzicalupo M, Benedetti A, Mengoni A**. 2014. StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies. Mol Ecol Resour **14:**426–434. http://dx.doi.org/10.1111/1755-0998.12187.
8. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA**. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
9. **Seemann T**. 2012. Prokka: prokaryotic genome annotation system. Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria, Australia.
10. **Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R**. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res **39**(Suppl 2):W339–W346. http://dx.doi.org/10.1093/nar/gkr466.
11. **McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD**. 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother **57:**3348–3357. http://dx.doi.org/10.1128/AAC.00419-13.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

Draft Genome Sequence of *Pseudomonas* sp. Strain Ep R1 Isolated from *Echinacea purpurea* Roots and Effective in the Growth Inhibition of Human Opportunistic Pathogens Belonging to the *Burkholderia cepacia* Complex

Valentina Maggini,[a,b,c] Luana Presta,[a] Elisangela Miceli,[a] Marco Fondi,[a] Emanuele Bosi,[a] Carolina Chiellini,[a] Camilla Fagorzi,[a] Patrizia Bogani,[a] Vincenzo Di Pilato,[d] Gian Maria Rossolini,[b,e] Alessio Mengoni,[a] Fabio Firenzuoli,[c] Elena Perrin,[a] Renato Fani[a]

Department of Biology, University of Florence, Florence, Italy[a]; Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy[b]; Center for Integrative Medicine, Careggi University Hospital, University of Florence, Florence, Italy[c]; Department of Surgery and Translational Medicine, University of Florence, Florence, Italy[d]; Clinical Microbiology and Virology Unit, Careggi University Hospital, Florence[e], Italy

ABSTRACT In this announcement, we detail the draft genome sequence of the *Pseudomonas* sp. strain Ep R1, isolated from the roots of the medicinal plant *Echinacea purpurea*. The elucidation of this genome sequence may allow the identification of genes associated with the production of antimicrobial compounds.

Received 24 March 2017 **Accepted** 27 March 2017 **Published** 18 May 2017

Citation Maggini V, Presta L, Miceli E, Fondi M, Bosi E, Chiellini C, Fagorzi C, Bogani P, Di Pilato V, Rossolini GM, Mengoni A, Firenzuoli F, Perrin E, Fani R. 2017. Draft genome sequence of Pseudomonas sp. strain Ep R1 isolated from Echinacea purpurea roots and effective in the growth inhibition of human opportunistic pathogens belonging to the Burkholderia cepacia complex. Genome Announc 5:e00351-17. https://doi.org/10.1128/genomeA.00351-17.

Address correspondence to Renato Fani, renato.fani@unifi.it.

V.M. and L.P. contributed equally to this article.

Endophytic bacterial communities inhabiting the rhizosphere or internal tissues of the medicinal plants (MPs) may contribute to the therapeutic properties of these plants (1). Here we report on the draft genome sequence of *Pseudomonas* sp. strain Ep R1, a strain isolated from the roots of *Echinacea purpurea*, an MP with immunomodulant, antiviral, and antimicrobial activity (2). The *E. purpurea* bacterial endophytes were isolated and molecular and phenotypic characterizations were conducted (3). In particular, *Pseudomonas* sp. Ep R1 showed the ability to inhibit the growth of other *E. purpurea* endophytes (4) and of cystic fibrosis bacterial pathogens belonging to the *Burkholderia cepacia* complex (5). Moreover, it has been demonstrated to be highly (50 μg/ml) resistant to chloramphenicol and streptomycin (6).

The genome sequence of *Pseudomonas* sp. Ep R1 was determined by a 2- × 300-bp paired-end approach using the MiSeq sequencing system (Illumina Inc., San Diego, CA). A total of 1,148,852 paired-end reads were obtained, representing approximately 100× coverage of the whole genome. *De novo* assembly was performed using SPAdes 2.3 (7), which generated 363 contigs. Contigs with length less than 2,000 bp were discarded. The remaining contigs were used for a multidraft-based analysis using genome sequences of 13 *Pseudomonas* strains retrieved from the NCBI database (*P. aeruginosa* PAO1, *P. alkylphenolia* KL28, *P. denitrificans* ATCC 13867, *P. entomophila* L48, *P. fluorescens* F113, *P. fulva* 12-X, *P. knackmussii* B13, *P. mendocina* ymp, *P. protegens* CHA0, *P. putida* KT2440, *P. resinovorans* NBRC, *P. stutzeri* CGMCC, and *P. syringae* pv. tomato DC3000) through MeDuSa scaffolder (8). The final version of the draft genome assembly of *Pseudomonas* sp. Ep R1 is 6,797,087 bp long and embeds 158 contigs (the longest of which is 1,954,067 bp long). The G+C content is 65.5%, similar to that of other *Pseudomonas* genomes sequenced so far. Automated annotation of the *Pseudomonas* sp. Ep R1 draft genome sequence using the NCBI Prokaryotic Genome Annotation

Pipeline detected 6,001 protein-coding genes, 67 RNA-coding genes (7 complete rRNAs, 56 tRNAs, 4 noncoding RNAs [ncRNAs]), and 173 pseudogenes.

Genes involved in the biosynthesis of secondary metabolites with antimicrobial activity were searched. The analysis was performed within an antiSMASH shell (9), which revealed that the *Pseudomonas* Ep R1 genome harbors 6 clusters involved in the biosynthesis of streptomycin, stenothricin, pimaricin, type 3 polyketide synthase (T3PKS), siderophore (desferrioxamine B), and nonribosomal peptide synthetase (NRPS) (amychelin). Moreover, the genome sequence was analyzed through CARD (10), which led to the identification of several genes (*mexABEJKMNW*, *omrMN*, *katG*, *triC*, *mfd*, and *mdtC*) putatively involved in antibiotic resistance, some conferring specific resistance to fluoroquinolone, mupirocin, beta-lactam, aminocoumarin molecules, and others involved in regulatory or inactivating systems and efflux pumps.

Accession number(s). This whole-genome shotgun project has been deposited in GenBank under the accession no. MWTQ00000000. The version described in this paper is the version MWTQ00000000.1.

REFERENCES

1. Strobel GA. 2003. Endophytes as sources of bioactive products. Microbes Infect 5:535–544. https://doi.org/10.1016/S1286-4579(03)00073-X.
2. Hudson JB. 2012. Applications of the phytomedicine *Echinacea purpurea* (Purple Coneflower) in infectious diseases. J Biomed Biotechnol 2012: 769896. https://doi.org/10.1155/2012/769896.
3. Chiellini C, Maida I, Emiliani G, Mengoni A, Mocali S, Fabiani A, Biffi S, Maggini V, Gori L, Vannacci A, Gallo E, Firenzuoli F, Fani R. 2014. Endophytic and rhizospheric bacterial communities isolated from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia*. Int Microbiol 17:165–174. https://doi.org/10.2436/20.1501.01.219.
4. Maida I, Chiellini C, Mengoni A, Bosi E, Firenzuoli F, Fondi M, Fani R. 2016. Antagonistic interactions between endophytic cultivable bacterial communities isolated from the medicinal plant *Echinacea purpurea*. Environ Microbiol 18:2357–2365. https://doi.org/10.1111/1462-2920.12911.
5. Chiellini C, Maida I, Maggini V, Bosi E, Mocali S, Emiliani G, Perrin E, Firenzuoli F, Mengoni A, Fani R. 2017. Preliminary data on antibacterial activity of *Echinacea purpurea*-associated bacterial communities against *Burkholderia cepacia* complex strains, opportunistic pathogens of cystic fibrosis patients. Microbiol Res 196:34–43. https://doi.org/10.1016/j.micres.2016.12.001.
6. Mengoni A, Maida I, Chiellini C, Emiliani G, Mocali S, Fabiani A, Fondi M, Firenzuoli F, Fani R. 2014. Antibiotic resistance differentiates *Echinacea purpurea* endophytic bacterial communities with respect to plant organs. Res Microbiol 165:686–694. https://doi.org/10.1016/j.resmic.2014.09.008.
7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
8. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, Crescenzi P, Fani R, Fondi M. 2015. Medusa: a multidraft based scaffolder. Bioinformatics 31:2443–2451. https://doi.org/10.1093/bioinformatics/btv171.
9. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346. https://doi.org/10.1093/nar/gkr466.
10. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 57:3348–3357. https://doi.org/10.1128/AAC.00419-13.

6 Evolutionary genomics

The availability of completely sequenced genomes have opened a new opportunity for engaging in evolutionary studies. Evolutionary genomics, or molecular evolution as it is sometimes called, refers to the study of how features or components of a genome change both within and between species over evolutionary timescales. The interpretation of any changes in the structure or content of genomes is often made within a comparative phylogenetic context in order to infer evolutionary processes in organisms. This requires knowledge of the various biological mechanisms underling peculiar DNA structures and molecular signals, such as coding and regulatory regions in the genomes, mutations, transpositions, duplications, deletions and so on. The underlying noise in the data, hence, can be considered as the result of the variability of DNA sequences due to population genetic forces. This exact approach is the one I followed in the study of nitrogen fixation pathway evolution and on which I wrote an essay partially reported in this section.
Nitrogen fixation, or rather the biological convertion of atmospheric N_2 in to ammonia (NH_4^+), represents an excellent model to study the evolutionary interconnections that link different metabolic pathways and functional divergences of paralogous genes. The orgin of the process are still under debate but, it most likely represented a metabolic innovation playing an important role in first evolutionary steps, when the prebiotic substitutes of all nitrogen resources decreased. The growing number of primordial cells that thrived in the ancestral environment would have caused a progressive reduction of essential nutrients which, in turn, imposed an increasingly strong selective pressure. This favored those microorganisms able to synthesize those essential nutrients by themselves. In such a context, the ancestral pathway of the metabolic process may have been developed from a small number of genes encoding multifunctional and nonspecific enzymes capable of reacting with a variety of chemically related substrates. These primordial enzymes would have been likely responsible for the interconnection of nitrogen fixation with other metabolic pathways, such as bacterial photosynthesis and leucine/lysine biosynthesis [1]. After, duplication of genes and operons, gene recruitment, gene lengthening and intense horizontal transfer would have modeled the entire metabolic pathway.
Today this process represents the most important input of biologically avail-

able nitrogen in Earth's ecosystems, although being a metabolic ability possessed only by certain microbial species, defined diazotrophs. The last, occupy very diverse ecological niches, including terrestrial and water environments, having several different life-styles: photosynthetic (oxygenic and anoxygenic), chemolitotrophic, aerobic and anaerobic. Many of them are free-living organisms while some others are symbionts (endophytes). A similar heterogeneous arrangement appears particularly interesting, either from an evolutionary and biogeochemical point of view. In fact, since nitrogen fixation is an oxygen sensitive process, this trait may be correlated to the kind of atmosphere present on Earth when the process made the scene. Moreover, the dissemination of this molecular mechanism also stands for several adaptive strategies used by different living systems to bypass a common issue (the oxygen damage).

Nitrogen fixation is a very complex pathway, with an high energetic cost and requiring the contribution of many genes to properly functioning. The key-enzyme of the entire process is called nitrogenase. It is known in three isoforms, which mainly differ for the cofactor usage. Among the three, the best studied and the most widely-spread is the Mo-dependent nitrogenase, whose structural component are encoded by the genes *nifH*, *nifD* and *nifK* (often close in the genomes of diazotrophs organisms and organized as operon). The other two isoforms, usually reffered to as alternatives, are called Fe-V nitrogenase (dependent by Iron-Vanadium cofactor; encoded by vnfH, *vnfD*, *vnfG* and *vnfK* genes) and Fe-only nitrogenase (encoded by *anfH*, *anfD*, *anfG* and *anfK* genes; it does not need a cofactor). These two are omologous to the Mo-nitrogenase, with the only exception of possessing an additional subunit (G). Despite their odd metal content, the three kind of nitrogenases show an high degree of conservation at the aminoacidic level, and they are structurally, functionally and phylogenetically related [1–3].

As mentioned, the origin of this metabolic capacity is unclear. It has not yet been clarified whether it was acquired before the appearance of the Last Universal Common Ancestor (LUCA) and subsequently transferred vertically or rather, if it appeared posterior to LUCA in the branch of Bacteria or Archea and then handed down to the different phyletic lines by vertical and/or horizontal gene transfer (HGT) mechanisms. The first hypothesis is consistent with the geological theories that support a high availability of Mo in the primordial Earth. However, since the *nif* genes can be organized into operons and therefore more easily subject to HGT, their presence in both Archea and Bacteria, is not considered a clear demonstration of the antiquity of the metabolic pathway [3, 4]. Moreover, *nif* genes are those most widely dis-

tributed in today's diazotrophic populations but it does not necessarily reflect the genetic make up of early diazotrophs. The *vnf* or *anf* genes could have originated first, constituting a simpler and less efficient nitrogenase system from which, by means of gene duplication mechanisms, the *nif* genes could have evolved. Thus, by using the genome sequences collected from publicly available databases I investigated the molecular mechanisms responsible for the evolution of *nif*, *anf* and *vnf* genes, trying to reconstruct the evolutionary history of the nitrogen fixation process. In particular I was interested in:

1. nitrogen fixation ancestrality (did it appear before or after LUCA?);

2. which were the molecular mechanisms involved in the origin, evolution and development of the pathway;

3. which is the oldest nitrogenase system and

4. which is the criteria to *in-silico* distinguish diazotrophs from non-diazotrophs.

To answer these questions, the homologous sequences of Nif, Vnf and Anf protein of the model organism *Azotobacter vinelandii* were collected from NCBI repository. The Biderectional Best Hit (BBH) BLAST algorithm adopted as matching criteria, allowed to reconstruct a first draft table of nitrogen fixation distribution. By analyzing that, it was proposed a new method to positively recognize nitrogen fixing organisms from the others. According to a previous work [2], the minimal set of genes enabling to fix N_2 is the co-presence of NifHDKENB. However, if the alternative forms Vnf and Anf sequences are taken into account, this principle results too stringent, leaving outside many known diazotrophs. BBH results designate the genetic core-set of the process as the ensemble of *nifB* plus at least one among the homologous genes *nif*, *vnf*, *anf* HDKEN.

Afterwards, the sequences collected have been aligned and analyzed. Based on the computed molecular identity score (which represent how much sequences are conserved among different organisms), phylogenetic trees have been inferred. The comparison of the resulting specific phylogeny to the species reference phylogeny tree allowed the estimation of an evolutionary model for nitrogen fixation process. Particularly, the inconsistencies identified

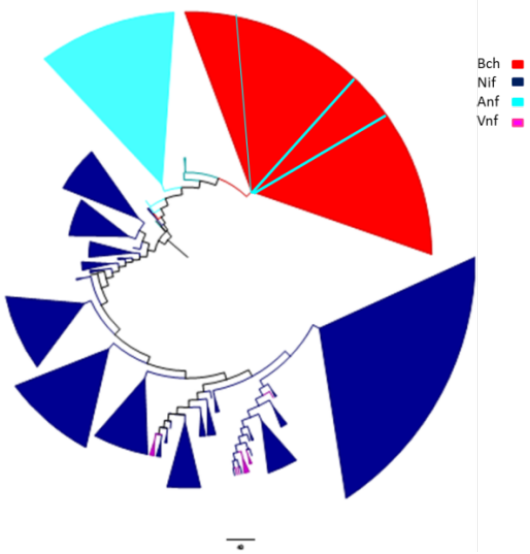


Fig. 6.1: Circular tree of Nif, Vnf and Anf H sequences. The color scheme indicate different taxonomic group.

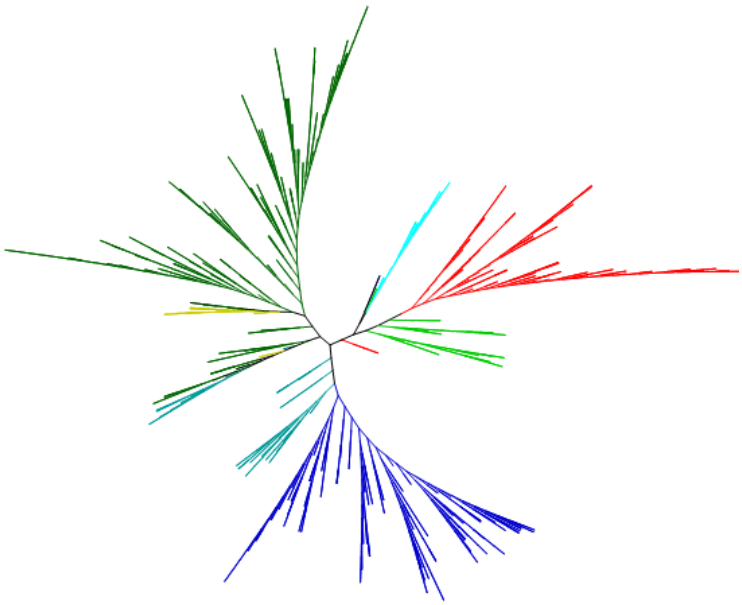


Fig. 6.2: Radial tree of Nif, Vnf and Anf DK sequences. The color scheme indicate different taxonomic group. **Red**: Bch, outgroup; **Light green and light blue**: Anf D and K, respectively; **Yellow and blue**: Vnf D and K, respectively; **Dark green and dark blue**: Nif D and K, respectively.

between these two highlight some HGT events that shaped the evolution and the spread of nitrogen fixation process among different bacterial species. The phylogenetic trees of H (see Figure 6.1) and DK (see Figure 6.2) sequences show that Anf sequences are those more closely related to the out-group used (*Bch* sequences), this meaning that *anf* encoded nitrogenase may be the most

ancient system. Additionally, DK tree trace back in the branch of *anf* genes family the event that gave birth to these two parologous genes (see Figure 6.2). According to that scenario, Mo-nitrogenase could be the last assembled system and it would have been enriched by many genes with regulatory and structural functions in various taxa. The following pages cover the relevance of Mo-based nitrogen fixation process (with a blink to biochemical details) and analyze this pathway distribution in present Bacteria and Archea and how it evolved. The general pipeline used to perform such kind of analysis is described in the second part.

Bibliography

1. Fani, R., Gallo, R. & Lio, P. Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes. *Journal of Molecular Evolution* **51,** 1–11 (2000) (cit. on pp. 63, 64).
2. Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C. & Dixon, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC genomics* **13,** 162 (2012) (cit. on pp. 64, 65).
3. Raymond, J., Siefert, J. L., Staples, C. R. & Blankenship, R. E. The natural history of nitrogen fixation. *Molecular biology and evolution* **21,** 541–554 (2004) (cit. on p. 64).
4. Shi, T. & Falkowski, P. G. Genome evolution in cyanobacteria: the stable core and the variable shell. *Proceedings of the National Academy of Sciences* **105,** 2510–2515 (2008) (cit. on p. 64).

Chapter 5
Nitrogen Fixation, a Molybdenum-Requiring Process

Abstract Nitrogen fixation is the most important input of biologically available nitrogen in Earth's ecosystems and is a metabolic ability possessed only by some Prokaryotes. To date four classes of nitrogenase enzymes have been characterized. Three nitrogenases are homologous enzymes with similar protein subunit composition and metal cofactor structure; these are the Mo-nitrogenase, V-nitrogenase, and Fe-only nitrogenase. How these three systems evolved and which of them first appeared on Earth is still under debate. The best studied system is the Fe-Mo-co based although several comparative analyses have been performed in past years.

In Chap. 4 only the class of pterin-based (Mo-co) molybdenum enzymes was discussed. However, there is an enzyme that uses a different form of Mo-cofactor (Fig. 5.1), the aforementioned nitrogenase, which catalyses the conversion of atmospheric nitrogen into ammonia, a process known as biological nitrogen fixation. Before entering nitrogenase's biochemical structure and the relative metabolic pathway, an overview of nitrogen fixation process will be provided.

5.1 Nitrogen Fixation in General

Nitrogen fixation is the most important input of biologically available nitrogen in Earth's ecosystems and is a metabolic ability possessed only by some Bacteria (Green Sulphur Bacteria, Firmicutes, Actinomycetes, Cyanobacteria and Proteobacteria) and Archaea, where it is mainly present in methanogens (Dixon and Kahn 2004). A recent analysis predicted that nearly 15 % of prokaryotic species with sequenced genomes are either known or potential diazotrophs (i.e. microbes able to fix nitrogen), a fraction much larger than commonly accepted (Dos Santos et al. 2012). Nitrogen fixation is a complex process with a high energetic cost and requiring the activity of several genes.

L. Presta et al., *Molybdenum Cofactors and Their role in the Evolution of Metabolic Pathways*, SpringerBriefs in Biometals, DOI 10.1007/978-94-017-9972-0_5

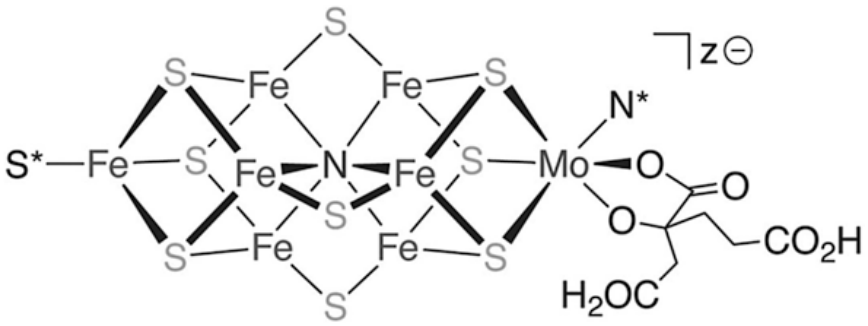


Fig. 5.1 Revised FeMo-cofactor structure derived from the most recent, high-resolution macromolecular structure determination. Reproduced from Lee and Holm (2003). With kind permission of © Proceedings of the National Academy of Sciences

To date four classes of nitrogenase enzymes have been characterized. Three nitrogenases are homologous enzymes with similar protein subunit composition and metal cofactor structure; these are the Mo-nitrogenase, V-nitrogenase, and Fe-only nitrogenase. The majority of present-day biological N_2 reduction is catalyzed by the Mo-dependent nitrogenase (encoded by *nif* genes), although V- and Fe-nitrogenase (encoded by *vnf* and *anf* genes, respectively) are important biological sources of fixed nitrogen in environments where Mo is present in limited amount. In fact the Mo-nitrogenase, which contains the iron-molybdenum cofactor or Fe-Mo-co, is the most commonly distributed nitrogenase; it is also the most efficient enzyme in the conversion of N_2 into NH_4^+. Alternative nitrogenases can be synthetized by diazotrophs, i.e. V-nitrogenase and/or Fe-only nitrogenase enzymes, which contain the FeV-co or Fe-co metalloclusters at their active sites, respectively. Regulation of alternative nitrogenases expression is dependent on the availability of molybdenum, vanadium or iron in the medium (Pau 2004). The three enzymes are oxygen-sensitive and they are irreparably damaged and deactivated by it. They show a high degree of conservation of structure, function and amino acid sequence across wide phylogenetic ranges. Interestingly, there are no reported diazotrophs lacking a Mo-nitrogenase and carrying uniquely an alternative nitrogenase. How these three systems evolved and which of them first appeared on Earth is still under debate. Several comparative analyses have been performed in past years although, to date, the best studied system is the Fe-Mo-co based.

The last type of nitrogenase is a Mo-nitrogenase phylogenetically unrelated to the other three classes and it has only been found in the bacterium *Streptomyces thermoautotrophicus*. This Mo-nitrogenase exhibits completely different biochemical features that consist of different protein composition, insensitivity to O_2, low Mg·ATP requirement, and a Mo-molybdopterin cytosine dinucleotide or Mo-MCD in the active site (Ribbe et al. 1997).

5.2 Biochemical Aspects

The overall reaction catalyzed by nitrogenase is usually depicted as follows:

$$\mathbf{N_2 + 8H^+ + 16MgATP + 8e^- \rightarrow 2NH_3 + H_2 + 16MgADP + 16Pi}\ \text{(inorganic phosphate)}.$$

This reaction not only represents the key entry point of reduced nitrogen into the global nitrogen cycle, but also embodies the formidable chemistry of breaking the triple bond of N_2 under ambient conditions (Hu and Ribbe 2011).

The FeMo-co nitrogenase (Fig. 5.2) is composed by two proteins, the iron (Fe) protein (encoded by *nifH*) and the molybdenum-iron (MoFe) protein (encoded by *nifD* and *nifK*). The homodimeric Fe protein is bridged by a single (4Fe–4S) cluster between the subunits and contains one ATP binding site within each subunit; whereas the $\alpha_2\beta_2$-tetrameric MoFe protein contains two unique clusters per αβ-dimer: the P-cluster (an (8Fe–7S) cluster), which is located at the α/β-subunit interface; and the Fe-Mo-co (a (Mo–7Fe–9S–X–homocitrate) cluster, where X is considered to be C, N or O), which is positioned within the α-subunit (Kim and Rees 1992). The catalysis of nitrogenase involves complex association/

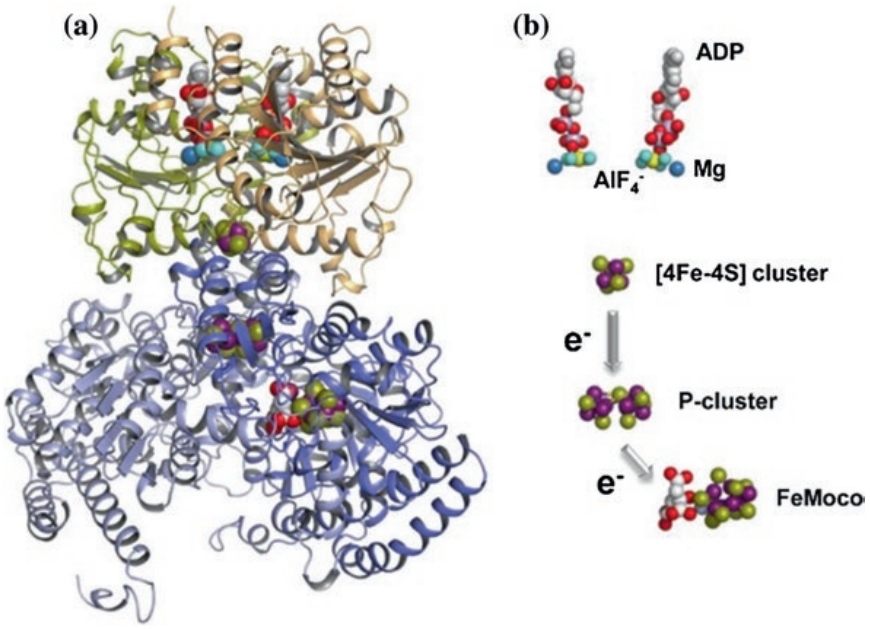


Fig. 5.2 X-ray crystal structure of half of the ADP·AlF_4^--stabilized Fe protein/MoFe protein complex (**a**) and the relative positions of components in the complex that are involved in the electron flow during catalysis (**b**). The identical subunits of Fe protein are shown in *yellow* and *orange*, and the α- and β-subunits of MoFe protein are shown in *light and dark blue*, respectively. The atoms of the components within the Fe protein/MoFe protein complex are colored as follows: Fe *purple*; S *yellow*; O *red*; C *gray*; Mg *dark blue*; Al *yellow*; F *light blue*. Reproduced from Hu and Ribbe (2011). With kind permission of © Coordination chemistry reviews

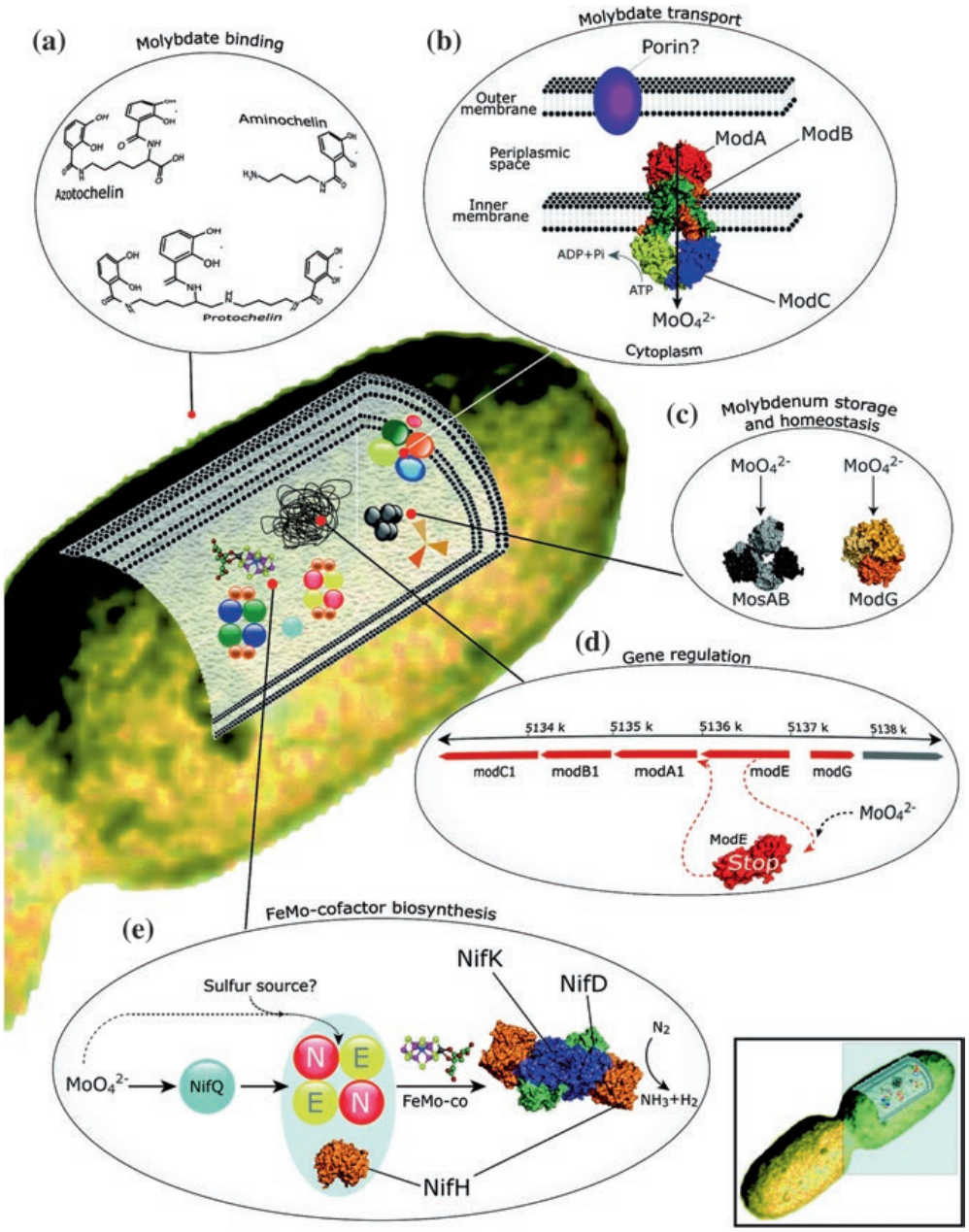


Fig. 5.3 Molybdenum trafficking for nitrogen fixation in the bacterium *Azotobacter vinelandii*. The figure shows a pathway of molybdenum towards the molybdenum-nitrogenase enzyme and the protein components involved in this pathway. Some of these proteins are exclusively dedicated to the nitrogenase biogenesis (e.g. NifQ and NifEN); some other have general roles in the metabolism of molybdenum (e.g. the ModABC molybdate transport system and the molybdenum-dependent transcriptional regulator ModE). Excreted siderophores with capacity to bind molybdate anions are also depicted. The *modABC* structure corresponds to the molybdate transport system of *Archaeoglobus fulgidus*. The ModE, NifH, and NifHDK structures shown are from *A. vinelandii*. Panel **e** shows a simplified FeMo-co biosynthetic pathway illustrating the two putative pathways for molybdenum incorporation into the Mo-nitrogenase cofactor. Reproduced from Hernandez et al. (2009). With kind permission of © Biochemistry

dissociation steps between the Fe protein and the MoFe protein, and the sequential, inter-protein transfer of electrons from the (4Fe–4S) cluster of the Fe protein, through the P-cluster, to the FeMoco of the MoFe protein, where substrate is reduced (Hu and Ribbe 2011).

The high level of complexity of nitrogenase's metallo clusters results in a (very) complex pathway for the assembly of nitrogenase and the insertion of the FeMo-co in the protein active site. The molybdenum trafficking for nitrogen fixation in the bacterium *A. vinelandii* is shown in Fig. 5.3.

In the free-living diazotroph *Klebsiella pneumoniae* at least 20 genes are involved in nitrogen fixation process (*nif* genes, Fig. 5.4) (Fani et al. 2000). In fact, apart from the catalytic components, additional gene's products are required to produce a fully functional enzyme. Indeed, several genes have been identified as

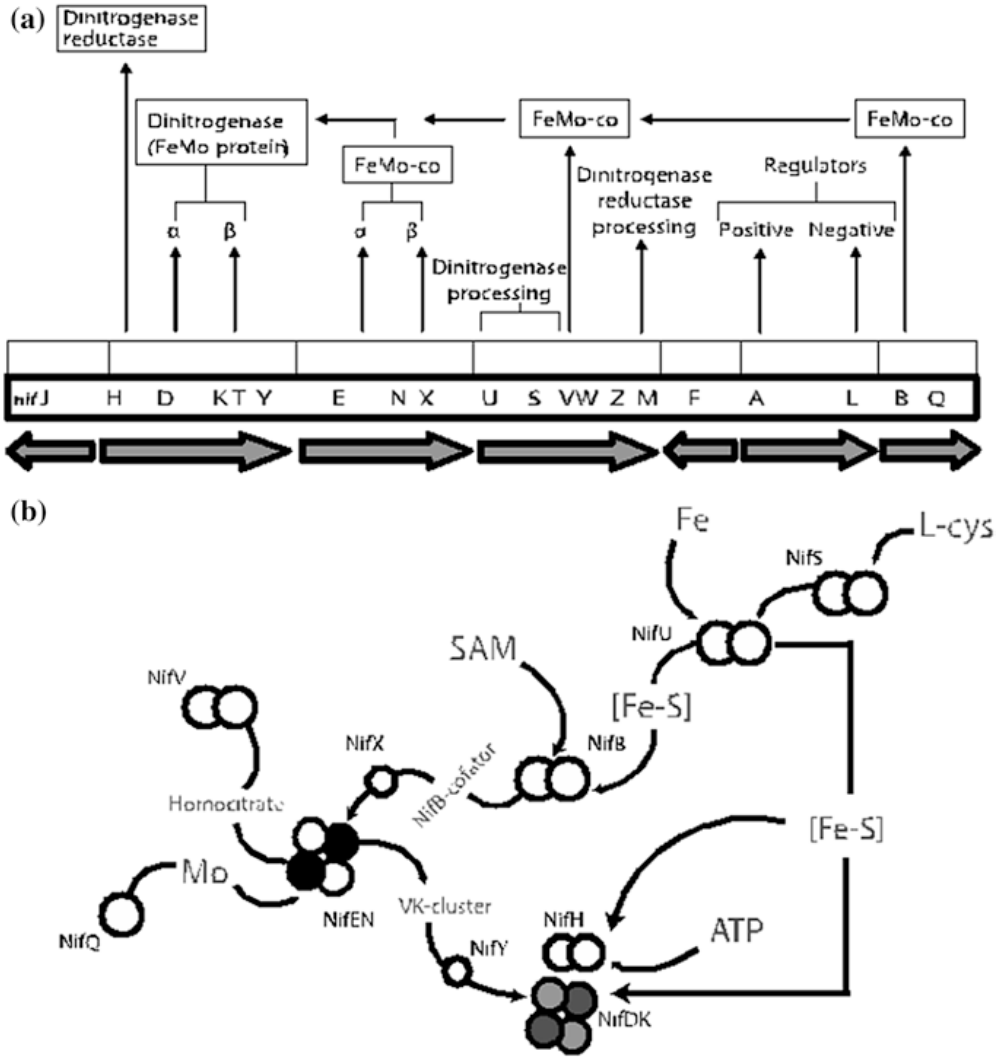


Fig. 5.4 **a** Organization of nif genes in *Klebsiella pneumonia* and **b** schematic representation of the nitrogen fixation process. Reproduced from Emiliani et al. (2010)

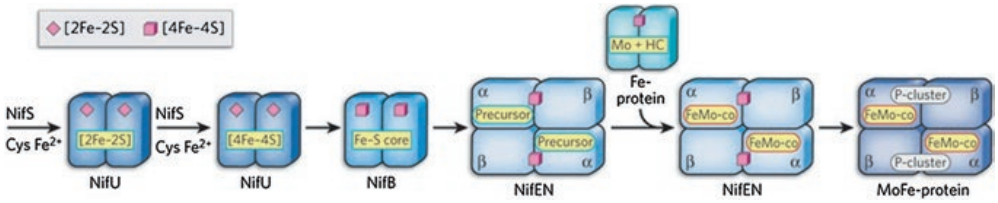


Fig. 5.5 Biosynthesis of FeMoco. NifS and NifU mobilize Fe and S for the sequential formation of $[Fe_2S_2]$ and $[Fe_4S_4]$ clusters, which are used as building blocks for the formation of a large Fe/S core on NifB. This Fe/S core is further processed into a molybdenum (Mo)- and homocitrate (HC)-free precursor, which can be converted to a mature FeMoco on NifEN upon Fe protein-mediated insertion of Mo and HC. Once the FeMoco is assembled on NifEN, it is delivered to its destined location in the MoFe protein. The permanent metal centers of the proteins are colored *gray* ($[Fe_2S_2]$ cluster *diamond*; $[Fe_4S_4]$ cluster *cube*; P-cluster *oval*), and the transient cluster intermediates are colored *yellow*. Reproduced from Schwarz et al. (2009). With kind permission of © Nature

being involved in this process (Fani et al. 2000 and references therein) although the number of proteins involved in the activation of nitrogenase seems to be species-specific and it varies according to the physiology of the organism and its environmental niche (Hamilton et al. 2011; Yan et al. 2010). In particular, FeMo-co biosynthesis requires, at least, the participation of *nifS*, *nifU*, *nifB*, *nifE*, *nifN*, *nifV*, *nifH*, *nifD* and *nifK* gene products. Based on genetic evidence, FeMo-co assembly is likely initiated by the actions of NifS and NifU (encoded by *nifS* and *nifU*) that generate small building blocks for FeMo-co assembly. NifS is a pyridoxal phosphate-dependent cysteine desulphurase, which forms a protein-bound cysteine persulphide that is subsequently donated to NifU for the sequential formation of small (2Fe–2S) and (4Fe–4S) clusters (Hu and Ribbe 2011). These small Fe–S fragments are then transferred to NifB (encoded by *nifB*) and processed into a FeMo-co core (with the help of NifV, Q and X). The last is then transferred to NifEN (encoded by *nifE* and *nifN*) and undergoes additional rearrangements before it is delivered to its target binding site in the MoFe protein through a protein-protein interaction (Hu and Ribbe 2011, Fig. 5.5).

5.3 How Did Nitrogen Fixation Evolve? Comparative Genomes Analyses Approach

The phylogenetic distribution of *nif* genes (i.e. the set of genes homologous to each of the 20 *nif* genes of another well-studied model organism, *Klebsiella pneumoniae*) in 842 completely sequenced prokaryotes (52 Archaea and 790 bacteria), revealed the presence of a common *core* of nitrogen fixation-related genes in a subset of 124 organisms (Emiliani et al. 2010). This *core* of genes is composed by only six *nif* genes (*nifHDKENB*), involved in nitrogenase, nitrogenase reductase and Fe–Mo Cofactor biosynthesis. Instead, all remaining *nif* genes display a

patchy phylogenetic distribution, thus revealing a complex evolutionary history of genes involved in the pathway. The finding of the same common *core* of genes has recently been confirmed by Dos Santos et al. (2012), who also proposed it as a novel in silico tool for the identification of additional diazotrophs in analyses performed in completely sequenced microbial genomes. Moreover, the finding of a common core of essential genes involved in the pathway might suggest that nitrogen fixation is an ancestral metabolic trait. Even if not proved, it is quite possible that only *nifHDKENB* genes were present in the genome of the LUCA (Last Universal Common Ancestor) community.

According to this idea, the *nifHDKENB* might represent a "universal core" for nitrogen fixation, whereas the other genes might have been differentially acquired during evolution in the different phylogenetic lineages. In fact, if nitrogen fixation had required other enzymes, their role might have been performed by others with low substrate specificity, in agreement with the Jensen hypothesis on the origin and evolution of metabolic pathways (Jensen 1976).

This idea is supported by analyses on the presence of *nif* genes homologs in the microbial domain revealing that most of the *nif* genes have in-paralogs (i.e. paralogs involved in the same pathway) and/or out-paralogs (i.e. paralogs involved in different pathways) as pointed out in Emiliani et al. (2010) (Fig. 5.6). The analysis did not retrieve any known paralogs for *nifW* (*nifO*), *nifT* (*fixU*), *nifQ* and *nifZ,* which are also missing from a large fraction of diazotrophs genomes. Eight *nif* genes (*nifAFHJLMSU*) are related, at a different extent, to proteins involved in other metabolic pathways (out-paralogs). NifS is related to a number of paralogs mainly involved in amino acid and carbon metabolisms. NifJ, a multidomain pyruvate:ferredoxin (flavodoxin) oxidoreductase, exhibited a large number of paralogs. Several proteins involved in Fe–Mo cofactor biosynthesis have paralogs in other cofactor biosyntheses.

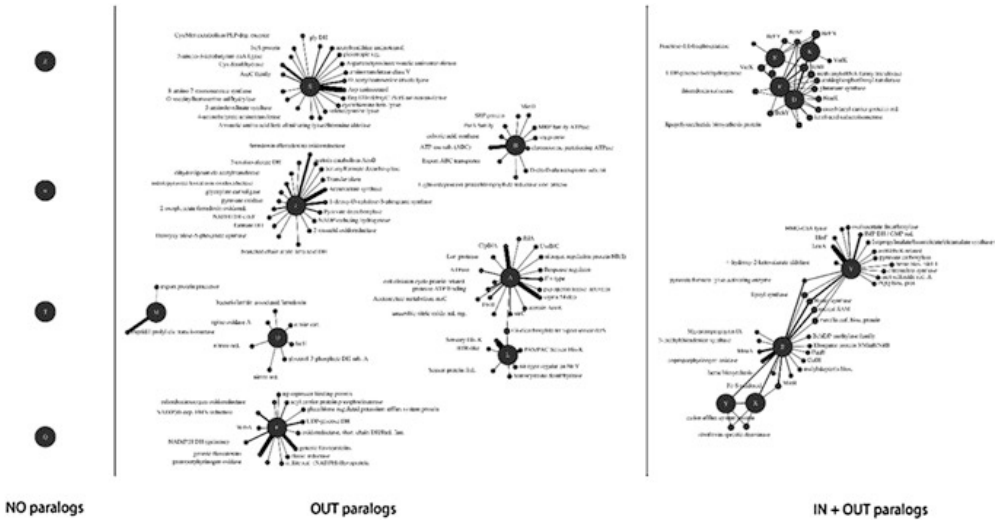


Fig. 5.6 In- and Out-paralogs network of nif genes. Nodes represent protein, links represent similarity values. Reproduced from Emiliani et al. (2010)

Eight Nif proteins share a significant degree of sequence similarity with proteins involved in other metabolic routes, and also with other *nif* genes products; this group can be further divided into two different clusters, the first of which includes *nifDKEN*, and the second being composed by *nifBXY* and *nifV*. NifB, NifX, and NifY share a common domain of about 90 aminoacids; moreover, *nifB* has an additional domain belonging to the S-adenosylmethionine (SAM) family, found in proteins that catalyze diverse reactions, including unusual methylations, isomerisation, sulphur insertion, ring formation, anaerobic oxidation and proteins' radical formation. Evidence exists that these proteins generate a radical species by reductive cleavage of SAM through an unusual Fe–S centre. The *nifV* and *nifB* genes are not directly linked although a sort of connection can be found in multidomain proteins sharing homology with NifV and NifB.

The evolutionary history of *nifDKEN* cluster has been investigate more in depth and more details are available to explain how it evolved (Fani et al. 2000).

Indeed, the two gene pairs *nifD-nifK* and *nifE-nifN* form a paralogous gene family, which code for nitrogenase and the tetrameric complex NifN2E2, respectively. Both *nifD-nifK* and *nifE-nifN* likely arose through duplications of an ancestral gene by a two-step model in which an ancestor gene underwent an in-tandem duplication event, giving rise to a bicistronic operon; this, in turn, duplicated, leading to the ancestors of the present-day *nifDK* and *nifEN* operons (Fani et al. 2000).

However, the role of these primordial enzymes is still under debate and it might have depended on the composition of the early atmosphere (Fig. 5.7). Accordingly,

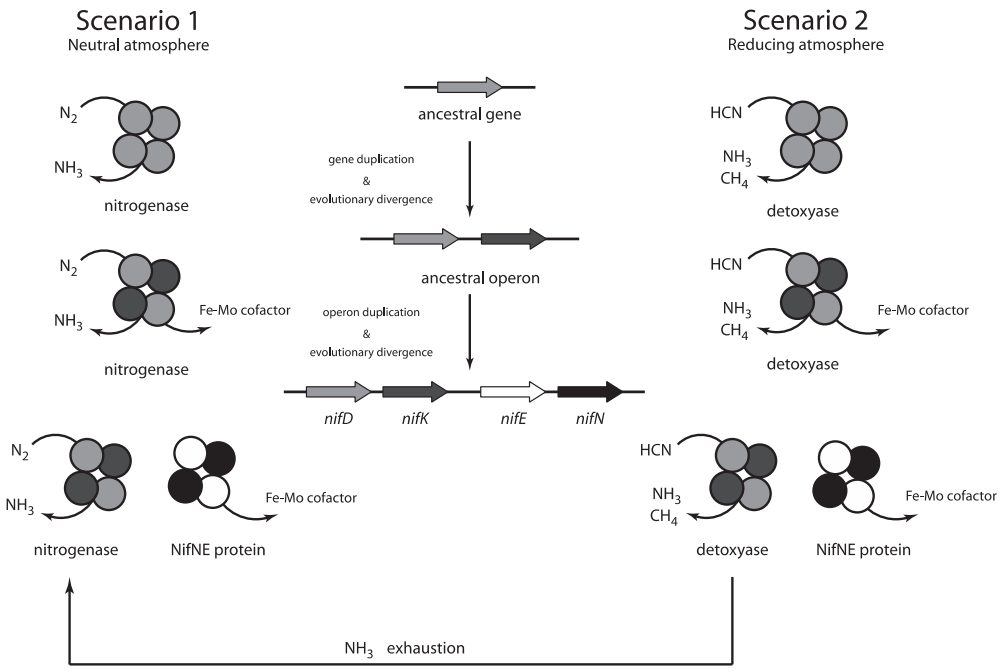


Fig. 5.7 Two possible scenarios depicted for the original function performed by the *nifDKEN* genes and their ancestor(s) gene(s). Reproduced from Fani et al. (2000). With kind permission of © Journal of Molecular Evolution

it can be accounted for by two alternative scenarios: (i) if we assume that O_2 was absent from the primordial atmosphere, an essential prerequisite for the appearance of (an ancestral) nitrogenase (since it is inactivated by free oxygen Fay 1992), then the appearance of nitrogenase would have represented a necessary event for the first cells, living in a planet whose atmosphere was neutral, containing dinitrogen, but not ammonia (first scenario). In fact, if the primitive microorganisms for their overall metabolism required ammonia, then its absence must have imposed a selective pressure favouring those cells that had evolved a system to synthesise ammonia from atmospheric dinitrogen.

Therefore, according to this scenario, the function of the ancestral enzyme might have been that of a slow, inefficient and with low substrate specificity "nitrogenase" thus being able to react with a wide range of compounds with a triple bond (Fani et al. 2000).

An alternative scenario would speculate an early reducing, free-ammonia rich atmosphere (Fig. 5.7). In those conditions, the evolution of a nitrogen fixation system would not have been a prerequisite because of the abundance of abiotically produced ammonia. Hence, why a nitrogenase in those days? The answer to this question relies in the catalytic properties of nitrogenase. In fact the enzyme is able to reduce also other molecules such as acetylene, hydrogen azide, hydrogen cyanide, or nitrous oxide, all of which contain a triple bond. Therefore, according to this second scenario (Fig. 5.7) the primitive enzyme encoded by the ancestor gene, would have been a detoxyase, an enzyme involved in detoxifying cyanides and other chemicals present in the primitive reducing atmosphere (Silver and Postgate 1973). This scenario implies that combined nitrogen progressive exhaustion would have imposed the refinement of the enzyme specificity. The last, very likely, should have modified and adapted to another triple-bond substrate, dinitrogen, and given to this change was selected and retained by some bacterial and archaeal lineages to enable survival in nitrogen-deficient environments. Finally, the decreasing of free ammonia and cyanides in the atmosphere triggered the evolution of the detoxyase toward nitrogenase, that might have been a common feature of all microbial life until photosynthesising cyanobacteria largely increased the oxygen concentration and burned cyanides.

Intriguingly, comparative genomes analyses also showed that genes coding for nitrogenase (*nifDK*) and nitrogenase reductase (*nifH*) are evolutionarily related to genes involved in bacteriochlorophyll biosynthesis (see below). Chlorophyll (Chl) and bacteriochlorophyll (Bchl) are the photochemically active reaction centre pigments for most of the extant photosynthetic organisms. During the synthesis of both Chl and Bchl, reduction of the tetrapyrrole ring system leads to the conversion of protochlorophyllide (Pchlide) into a chlorin. A second reduction, unique to the synthesis of Bchl, converts the chlorin into a bacteriochlorin. Two mechanisms for reducing the double bond in the fourth ring of protochlorophyllide are depicted. An enzymatic complex acts irrespective of the presence or absence of light and is thus termed "light-independent protochlorophyllide reductase". The second is a light-dependent reaction that utilizes the enzyme NADPH-protochlorophyllide oxidoreductase (Suzuki et al. 1997). In *Rhodobacter capsulatus*, the products of

three genes are required for each reduction: *bchL*, *bchN*, and *bchB* for the Pchlide reductase and *bchX*, *bchY*, and *bchZ* for the chlorin reductase (Burke et al. 1993b). Both enzymes are three-subunit complexes. Burke et al. (1993a, b) detected a significant degree of sequence similarity between BchlL, BchN, BchB, and BchX, BchY and BchZ, respectively, suggesting that the six genes represent two triads of paralogs and that the two enzymes are derived from a common three-subunit ancestral reductase. It was also found that the so-called "chlorophyll iron protein" subunits encoded by bchX, bchL, and chlL share a remarkable sequence similarity with the nitrogenase Fe proteins (Burke et al. 1993a). These findings led Burke et al. (1993b) to suggest that genes involved in bacteriochlorophyll biosynthesis and nitrogen fixation were related mechanistically, structurally and evolutionarily. Similarly to NifH protein, which serves as the unique electron donor for the nitrogenase complex, the products of *bchL* and *bchX* could serve as the unique electron donor into their respective catalytic subunits (BchB-BchN and BchY-BchZ). The idea of a common ancestry of *nifH*, *bchL* and *chlL* genes (Burke et al. 1993a; Fujita et al. 1993) has had an elegant experimental support by Cheng et al. (2005) who demonstrated in the photosynthetic eukaryote *Chlamydomonas reinhardtii* that NifH is able to partially complement the function of ChlL in the dark-dependent chlorophyll biosynthesis pathway.

Nitrogenases and carboxylases might have represented bacterial pre-adaptations, resulting in multigenic traits that were retained because leading to new selective advantages in altered environments. In fact, as depletion of abiotically produced organic matter occurred in early Earth environment, competition for the organic prerequisites for reproduction ensued. As the carboxylation and nitrogen-fixing functions were achieved, a new, abundant, and direct source of carbon and nitrogen for organic synthesis became available. The ability to take up atmospheric carbon and nitrogen would have been of great selective advantage (Margulis 1993). It is possible to propose a model (Fig. 5.8) for the origin and evolution of nitrogen fixation and bacterial photosynthesis based on multiple and successive paralogous duplications of an ancestral operon encoding an ancient reductase. The eight genes (*nifDKEN* and *bchYZNB*) are members of the same paralogous gene family, in that all of them are the descendants of a single ancestral gene. The model proposed posits the existence of an ancestral three-cistronic operon (Fig. 5.8) coding for an unspecific reductase. One might assume that this complex was (eventually) able to perform both carboxylation and nitrogen fixation. The following evolutionary steps might have been the duplication of the ancestral operon followed by an evolutionary divergence that led to the appearance of the ancestor of *nifH*, *nifDE*, and *nifKN* on one side, and *bchLX*, *bchNY* and *bchBZ* on the other one. In this way the two reductases narrowed their substrate specificity: one channelled toward nitrogen fixation and the other one toward photosynthesis. However, each of the two multicomplex proteins was able to perform at least two different reactions:

1. The ancestor of *nifDKEN*, was likely able to carry out the reduction of dinitrogen to ammonia and the synthesis of Fe–Mo cofactor (Fani et al. 2000).

2. The ancestor of protochlorophyllide- and chlorin-reductase performed both of the reactions that in the extant photosynthetic bacteria are carried out by two triads (BchLNB and BchXYZ, respectively).

The complete diversification of the function of the two heteromeric complexes was likely achieved thorough duplication of *nifDE* and *nifKN* ancestors and by the three-cistronic operon's duplication *bch*(LX)(NY)(BZ) followed by evolutionary divergence (Fig. 5.8). In our opinion, this idea may perfectly fit the Jensen's hypothesis.

Concerning the timing of the above reported evolutionary events (Fani et al. 2000) the two paralogous duplication events leading to *nifDK* and *nifEN* likely predated the appearance of the LUCA. Conversely, other authors (Raymond et al. 2004) have proposed a different scenario, according to which nitrogen fixation per se was invented by methanogenic Archaea and subsequently transferred, in at least three separate events, into bacterial lineages.

Differently from nitrogen fixation, tetrapyrrole-based photosynthesis occurs only in bacteria and bacterially derived chloroplasts, therefore it can be surmised that the appearance of photosynthesis should have not predated the divergence of Archaea and Bacteria.

Recently, a similar scenario has been proposed (Boyd et al. 2011). According to phylogenetic- and structure-based examinations of multiple nitrogenase proteins

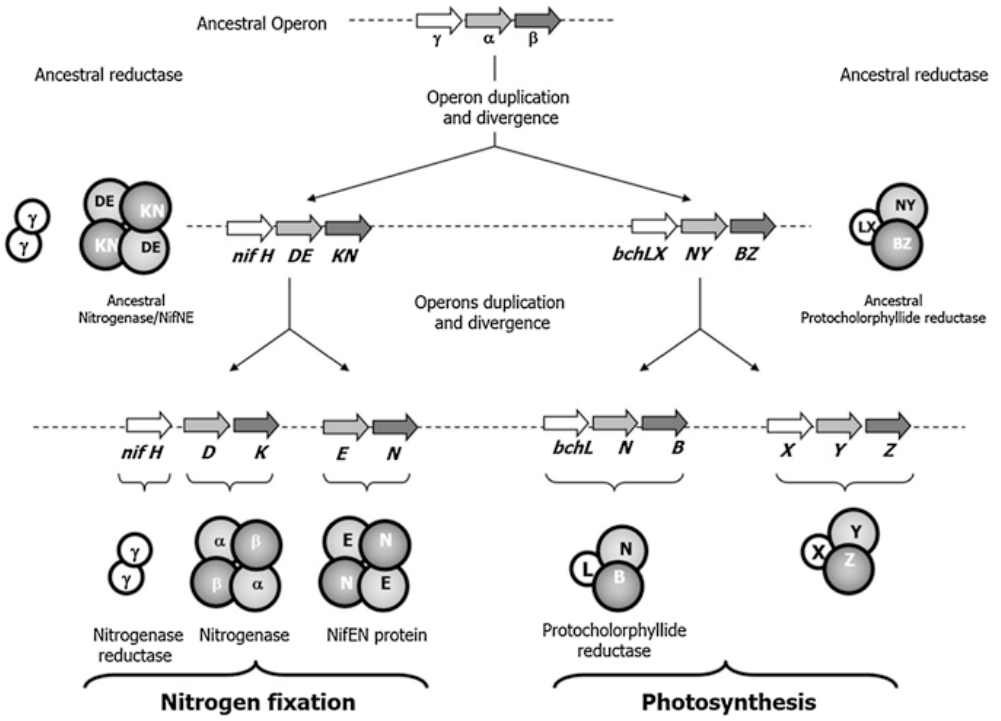


Fig. 5.8 Possible evolutionary model accounting for the evolutionary relationships between nif and bch genes. Reproduced from Emiliani et al. (2010)

these authors proposed a model for the emergence of nitrogenase whereby a gene encoding for an ancestral protein complex duplicated. The encoded protein should have possessed a cavity similar to that observed in the inferred structure of NflD (performing a nitrogenase-like reaction in Ni porphyrin F430 biosynthesis (Staples et al. 2007), the duplication led to the evolutionary precursor of BchN and NifD (including vanadium and Fe-only containing alternative nitrogenases, i.e. VnfD and AnfD). Serendipitously, metals (e.g., Fe) or metal clusters (e.g., 4Fe–4S) were bound in the cavity of the ancestor in a non-specific manner, resulting in an enzyme complex with altered reactivity, perhaps toward N_2 reduction (Boyd et al. 2011). In response to selective pressure of limited fixed nitrogen on early Earth, genes and associated gene's products were presumably recruited to improve the enzyme stepwise through the modification of the metal cofactor (Boyd et al. 2011).

5.4 Computational Studies on Mo-related Activities

In the last decade the availability of a large amount of data obtained from sequenced organisms has allowed the investigation of Mo utilization in prokaryotes and eukaryotes and its evolutionary changes. Comprehensive analyses of the occurrence and evolutionary trends of Mo-related traits can be carried out through bioinformatics tools by searching similar sequences of Mo uptake systems and Mo-dependent enzymes in the genomes of all so far sequenced organisms.

Recently several studies allowed the in silico identification of previously unknown features Mo-related such as the novel Mo-co-binding proteins that have been reported in both eukaryotes (mARC in pig mitochondria) (Havemeyer et al. 2006) and bacteria (YiiM and YcbX in *Escherichia coli*) (Kozmin et al. 2008).

These proteins share a significant degree of sequence homology with the C-terminal domain of eukaryotic Mo-co-sulfurase (MOSC) and show catalytic activity strictly dependent on Mo-co, suggesting that they may represent a previously unknown molybdoenzyme family. A study of 2011 by Yan Zhang et al. investigated the link between the MOSC motif and several known proteins highlighting interesting possible interaction between MO-systems and other pathways.

The MOSC domain is a superfamily of beta-strand-rich domains initially identified in the Mo-co sulfurase however it has subsequently been found in several other proteins from both prokaryotes and eukaryotes (Anantharaman and Aravind 2002). The MOSC domain of eukaryotic Mo-co sulfurase is involved in Mo-co binding with high affinity and its Mo-co carries a terminal sulfur ligand due to the catalytic activity of pyridoxal-5′-phosphate-dependent NifS-like domain (Wollers et al. 2008). On the other hand, Mo-co bound to the MOSC domain of mARC showed no terminal sulfur ligand. The function of the MOSC_N domain is unknown; however, it is predicted to adopt a beta barrel fold. Computational analysis allowed the distribution of these novel MOSC-containing molybdoproteins. It has been shown that all organisms containing MOSC-like proteins are Mo-co-utilizing organisms and that, in some organisms, genes for MOSC-like proteins are located close to

Mo-co biosynthesis components or molybdoenzymes such as MoaC and formate dehydrogenase. These findings indicate that MOSC-like proteins may be orthologs and they may serve as Mo-co chaperone involved in Mo-co transfer or storage. Further investigation may allow the complete understanding of interactions between Mo utilization and that of other trace elements, such as Fe and S.

The next chapter will focus on different methods and comparative genomics techniques through which computational studies can be carried out.

Bibliography

Anantharaman, V., & Aravind, L. (2002). MOSC domains: ancient, predicted sulfur-carrier domains, present in diverse metal–sulfur cluster biosynthesis proteins including Molybdenum cofactor sulfurases. *FEMS Microbiology Letters, 207*(1), 55–61.

Boyd, E. S., Hamilton, T. L., & Peters, J. W. (2011). An alternative path for the evolution of biological nitrogen fixation. *Frontiers in microbiology*, *2*.

Burke, D. H., Alberti, M., & Hearst, J. E. (1993a). The *Rhodobacter capsulatus* chlorin reductase-encoding locus, *bchA,* consists of three genes, *bchX, bchY,* and *bchZ. Journal of Bacteriology, 175*(8), 2407–2413.

Burke, D. H., Hearst, J. E., & Sidow, A. (1993b). Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. *Proceedings of the National Academy of Sciences, 90*(15), 7134–7138.

Cheng, Q., Day, A., Dowson-Day, M., Shen, G. F., & Dixon, R. (2005). *The Klebsiella pneumoniae* nitrogenase Fe protein gene (*nifH*) functionally substitutes for the *chlL* gene in *Chlamydomonas reinhardtii*. *Biochemical and biophysical research communications, 329*(3), 966–975.

Dixon, R., & Kahn, D. (2004). Genetic regulation of biological nitrogen fixation. *Nature Reviews Microbiology, 2*(8), 621–631.

Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., & Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics, 13*(1), 162.

Emiliani, G., Fondi, M., Liò, P., & Fani, R. (2010). *Evolution of metabolic pathways and evolution of genomes* (pp. 37–68). Netherlands: Springer.

Fani, R., Gallo, R., & Liò, P. (2000). Molecular evolution of nitrogen fixation: the evolutionary history of the *nifD*, *nifK*, *nifE*, and *nifN* genes. *Journal of Molecular Evolution, 51*(1), 1–11.

Fay, P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiological Reviews, 56*(2), 340.

Fujita, Y., Matsumoto, H., Takahashi, Y., & Matsubara, H. (1993). Identification of a *nifDK*-like gene (ORF467) involved in the biosynthesis of chlorophyll in the cyanobacterium *Plectonema boryanum*. *Plant and Cell Physiology, 34*(2), 305–314.

Hamilton, T. L., Ludwig, M., Dixon, R., Boyd, E. S., Dos Santos, P. C., Setubal, J. C., et al. (2011). Transcriptional profiling of nitrogen fixation in *Azotobacter vinelandii*. *Journal of bacteriology*, *193*(17), 4477–4486.

Havemeyer, A., Bittner, F., Wollers, S., Mendel, R., Kunze, T., & Clement, B. (2006). Identification of the missing component in the mitochondrial benzamidoxime prodrug-converting system as a novel molybdenum enzyme. *Journal of Biological Chemistry, 281*(46), 34796–34802.

Hernandez, J. A., George, S. J., & Rubio, L. M. (2009). Molybdenum trafficking for nitrogen fixation. *Biochemistry, 48*(41), 9711–9721.

Hu, Y., & Ribbe, M. W. (2011). Biosynthesis of nitrogenase FeMoco. *Coordination Chemistry Reviews, 255*(9), 1218–1224.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology, 30*(1), 409–425.

Kim, J., & Rees, D. C. (1992). Structural models for the metal centers in the nitrogenase molybdenum-iron protein. *Science, 257*(5077), 1677–1682.

Kozmin, S. G., Leroy, P., Pavlov, Y. I., & Schaaper, R. M. (2008). YcbX and yiiM, two novel determinants for resistance of *Escherichia coli* to N-hydroxylated base analogues. *Molecular Microbiology, 68*(1), 51–65.

Lee, S. C., & Holm, R. H. (2003). Speculative synthetic chemistry and the nitrogenase problem. *Proceedings of the National Academy of Sciences, 100*(7), 3595–3600.

Margulis, L. (1993). Symbiosis in cell evolution: Microbial communities in the Archean and Proterozoic eons.

Pau, R. N. (2004). Molybdenum uptake and homeostasis. In W. Klipp, B. Masepohl, J. R. Gallon, & W. E. Newton (Eds.), *Genetics and regulation of nitrogen fixation in free-living bacteria* (pp. 225–256). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Raymond, J., Siefert, J. L., Staples, C. R., & Blankenship, R. E. (2004). The natural history of nitrogen fixation. *Molecular Biology and Evolution, 21*(3), 541–554.

Ribbe, M., Gadkari, D., & Meyer, O. (1997). N2 Fixation by *Streptomyces thermoautotrophicus* involves a molybdenum-dinitrogenase and a manganese-superoxide oxidoreductase that couple N_2 reduction to the oxidation of superoxide produced from O_2 by a molybdenum-CO dehydrogenase. *Journal of Biological Chemistry, 272*(42), 26627–26633.

Schwarz, G., Mendel, R. R., & Ribbe, M. W. (2009). Molybdenum cofactors, enzymes and pathways. *Nature, 460*(7257), 839–847.

Silver, V. S., & Postgate, J. R. (1973). Evolution of asymbiotic nitrogen fixation. *Journal of Theoretical Biology, 56*, 340–373.

Staples, C. R., Lahiri, S., Raymond, J., Von Herbulis, L., Mukhophadhyay, B., & Blankenship, R. E. (2007). Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *Journal of Bacteriology, 189*(20), 7392–7398.

Suzuki, J. Y., Bollivar, D. W., & Bauer, C. E. (1997). Genetic analysis of chlorophyll biosynthesis. *Annual Review of Genetics, 31*(1), 61–89.

Wollers, S., Heidenreich, T., Zarepour, M., Zachmann, D., Kraft, C., Zhao, Y., & Bittner, F. (2008). Binding of sulfurated molybdenum cofactor to the C-terminal domain of ABA3 from *Arabidopsis thaliana* provides insight into the mechanism of molybdenum cofactor sulfuration. *Journal of Biological Chemistry, 283*(15), 9642–9650.

Yan, Y., Ping, S., Peng, J., Han, Y., Li, L., Yang, J., & Lin, M. (2010). Global transcriptional analysis of nitrogen fixation and ammonium repression in root-associated *Pseudomonas stutzeri A1501*. *BMC Genomics, 11*(1), 11.

Chapter 6
Protocols and Methods for the in Silico Reconstruction of the Origin and Evolution of Metabolic Pathways

Abstract It is possible to reconstruct the origin and evolution of metabolic pathways by inferring useful hints from the analyses of metabolic pathways existing in contemporary cells. Several bioinformatic tools allow the comparison of gene and genomes from organisms belonging to the three cell domains, giving rise to comparative genomics. Moreover the storage, the interpretation and the integration of several sources of information (gene structure and organization, gene regulation, protein-protein interactions) make possible to infer phylogenetic relationships among organisms, leading to a new approach referred to as phylogenomics. A short description of several tools is reported.

How can the origin ad evolution of metabolic pathways be studied and reconstructed? By assuming that useful hints may be inferred from the analysis of metabolic pathways existing in contemporary cells, important insights on the evolutionary development of microbial metabolic pathways can be obtained by the use of bioinformatic tools, allowing the comparison of gene and genomes from organisms belonging to the three cell domains (Archaea, Bacteria and Eukarya). This approach takes advantage of the availability of the phylogenetic relationships among (micro)organisms, and possibly on the existence of different structure and organization exhibited by orthologous genes. Beside, the more ancient is a pathway, the more information can be retrieved from this comparative analysis. Such kind of approach has become feasible since, in recent years, saw a dramatic increase in genomics data deriving from organisms belonging to all of the three known domains of life. By the way, the use of bioinformatic tools allowed the storage and the interpretation of several sources of information (gene structure and organization, gene regulation, protein-protein interactions) and, probably more importantly, their integration, a fundamental step for the global understanding of genomes properties and dynamics. This approach is usually referred to as comparative genomics. Combining data gained from comparative genomics with evolutionary studies of different species (i.e. phylogenetic inference), results in a new kind of approach, referred to as phylogenomics. A typical phylogenomics pipeline is illustrated in Fig. 6.1.

L. Presta et al., *Molybdenum Cofactors and Their role in the Evolution of Metabolic Pathways*, SpringerBriefs in Biometals, DOI 10.1007/978-94-017-9972-0_6

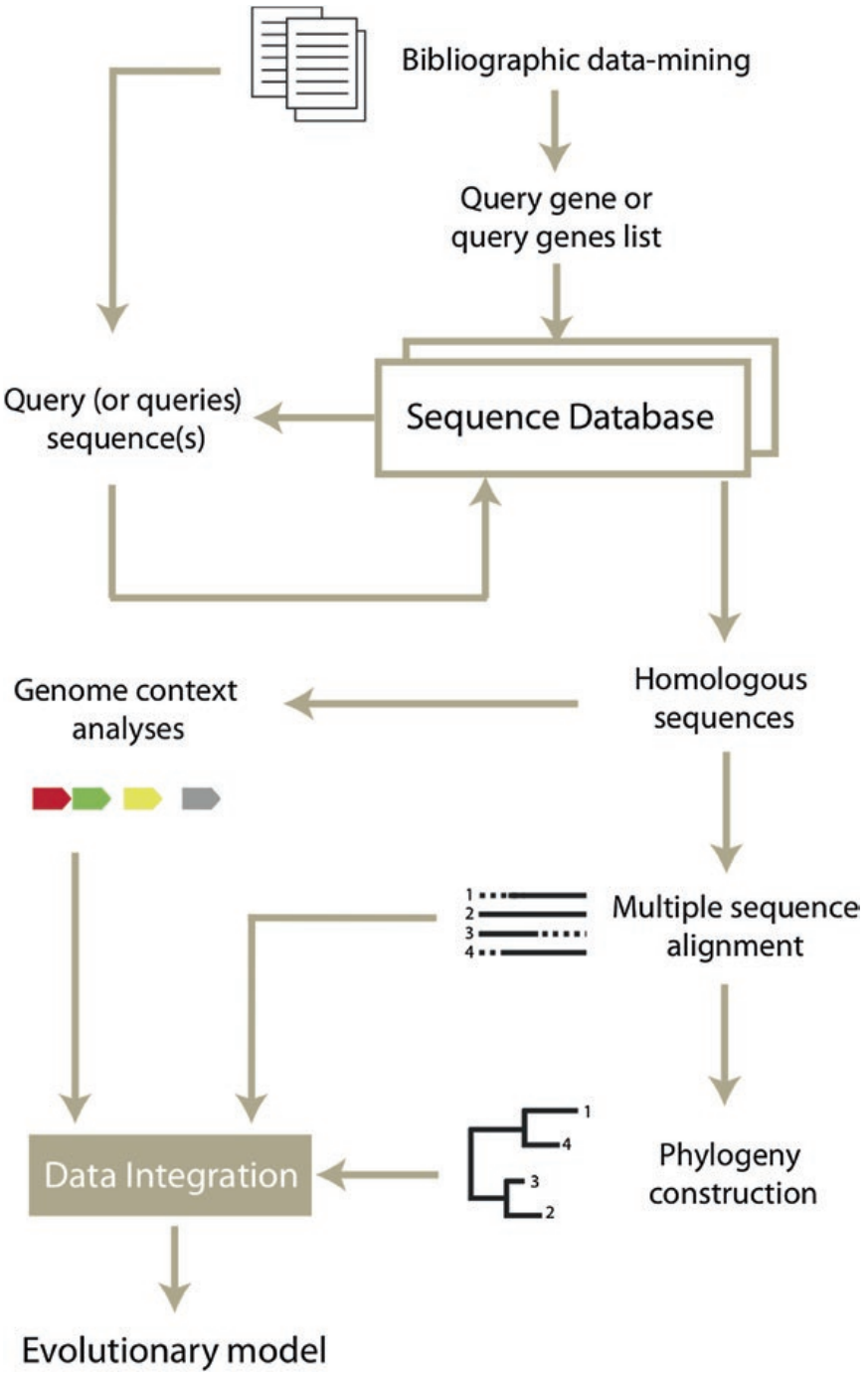


Fig. 6.1 A schematic representation of a typical phylogenetic pipeline

This way of investigating the evolutionary history of genes introduced several advantages; in fact, adopting a genome-scale approach theoretically overcomes incongruence derived from molecular phylogenies based on single genes mainly because (i) non-orthologous comparison (i.e. the comparison of those genes

erroneously defined as orthologous) is much more misleading when the analysis is performed on a single gene, whereas it is probably buffered in a multigene analysis and (ii) stochastic error naturally vanishes when more and more genes are considered. At the same time it is useful to recur to bioinformatics methods in order to detect new genes or proteins or to predict their function.

6.1 Browsing Microbial Genomes

At present, hundreds of microbial genomes have been sequenced, and hundreds more are currently in the pipeline. Furthermore, functional genomic studies have generated a large and growing body of experimental results for many different organisms belonging to the known domains of life. However, this whole body of data would reveal almost useless if not stored in a proper manner. To this purpose a growing number of public databases have been developed in recent years, usually providing also user-friendly tools for their interrogation. These tools, despite not allowing automatized large-scale phylogenomic analyses, often represent their first preliminary (and useful) step. This is the case for example of MicrobesOnLine [http://www.microbesonline.org (Alm et al. 2005; Dehal et al. 2010)], which embeds both structural and functional data on a large (almost 3000) dataset of completely sequenced genomes. These data are retrieved from a wide range of other specific databases (including KEGG, GeneOntology, RefSeq).

Interestingly, MicrobesOnLine also allows to interactively explore the neighborhood of any given gene, hence allowing, for example, a first analysis of the gene organization of a given metabolic pathway. Similarly, IMG (Chu et al. 2013) provides users some linked tools to support comparative microbial genes, genomes, and metagenomics analysis, including COG, KEGG, Pfam, InterPro, and the Gene Ontology. Consequently, thanks also to the graphical user interface IMG is particularly suited for nonexperienced bioinformaticians which want to perform comparative genome analyses (Checcucci and Mengoni 2015).

The same task can be pursued adopting also operonDB web service (http://odb.kuicr.kyoto-u.ac.jp/), (Pertea et al. 2009) aiming at collecting all known operons (derived from the literature and from publicly available database) in multiple species and to offer a system to predict operons by user definitions. Several other web sites and software tools have been described that assist in the annotation and exploration of comparative genomic data. The Prolinks (Bowers et al. 2004) and STRING (Jensen et al. 2009) databases offer convenient tools for browsing predicted functional associations among proteins. String, in particular imports protein association knowledge not only from databases of physical interactions, but also from databases of curated biological pathway knowledge. A number resources are included in the current release [MINT (Ceol et al. 2009), HPRD (Prasad et al. 2009), DIP (Xenarios et al. 2002), BioGRID (Stark et al. 2008), KEGG (Kanehisa and Goto 2000) and Reactome (Matthews et al. 2009) IntAct (Hermjakob et al. 2004), EcoCyc (Keseler et al. 2009)]. Furthermore, this set of previously known

and well-described interactions is then complemented by interactions that are predicted computationally, specifically for STRING, using a number of prediction algorithms (Jensen et al. 2009).

6.2 Orthologs Identification

Genomics data is a fundamental step for addressing the topic of the evolution of metabolic pathways, and strictly depends on a correct identification of orthologous proteins shared by different genomes. This field has been greatly developed in recent years and, paradoxically, the extant challenge seems not to be the lack of orthology predictions, but the right choice within the plethora of methods and databases that have been recently implemented (Gabaldon et al. 2009). The identification of orthologs between two genomes often relies on the so-called bidirectional best-hit (BBH) criterion, a reiteration of the BLAST algorithm (Altschul et al. 1997): two proteins, a and b, from genomes A and B respectively, are orthologs if a is the best-hit (i.e. the most similar) of b in genome A and vice versa. For three or more genomes, groups of orthologous sequences can be constructed by extending the BBH relationships with a clustering algorithm. This approach has led to the assembly of pre-compiled databases embedding groups of orthologous proteins, such as COG or KEGG-related systems (KOBAS and KAAS). Moreover, several other algorithms have been developed to fulfill this tasks, including Ncut (Abascal and Valencia 2002), Rio, (Zmasek and Eddy 2002), Outgroup Conditioned Score (OCS) (Cotter et al. 2002) or OrthoParaMap (Cannon and Young 2003). Recent advancements showed that clustering techniques applied to matrices storing pair-wise similarities perform quite well (Brilli et al. 2008). These algorithms work either on the grouping of weakly similar homologs or on the identification of protein domains. The most widespread are: (i) orthoMCL (Li et al. 2003), which adopts a Markov Clustering algorithm (previously implemented in tribeMCL (Enright et al. 2002), (ii) Ortholuge (Fulton et al. 2006) that aims at identifying orthologs by comparing proteins and species phylogenetic trees and, lastly, (iii) InParanoid (O'Brien et al. 2005) that relies on a similar flowchart. All these orthologs' identification methods have been recently tested on a dataset of proteins from different species previously characterized using functional genomics data, such as expression data and protein interaction data (Hulsen et al. 2006). Results have shown that InParanoid software seems the best orthologs' identification method in terms of identifying functionally equivalent proteins in different species (Hulsen et al. 2006).

6.3 Multiple Sequence Alignments

In a phylogenetic analysis workflow (but also when interested, for example, in structure modelling, functional site prediction and sequence database searching) a key step (usually following the correct orthologs retrieval procedure) consists in comparing those residues with inferred common evolutionary origin or structural/functional

equivalence in the whole sequence dataset. This task is fulfilled through multiple sequence alignment (MSA), that is arranging homolog protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived a single position in an ancestral sequence), superposable (in a rigid local structural alignment) or play a common functional role. Although these three criteria are essentially equivalent for closely related proteins, sequence, structure and function diverge over evolutionary time and different criteria may result in different alignments (Edgar and Batzoglou 2006). Many approximate algorithms have been developed for multiple sequence alignments, including the commonly used progressive alignment technique (Pei 2008). This greedy heuristic assembly algorithm involves estimating a guide tree (rooted binary tree) from unaligned sequences and then incorporating the sequences into the MSA with a pairwise alignment algorithm while following the tree topology. The scoring schemes used by the pairwise alignment algorithm are arguably the most influential component of the progressive algorithm. They can be divided in two categories, that is matrix- and consistency-based algorithms. Matrix-based algorithms such as ClustalW (Thompson et al. 2002), MUSCLE (Edgar 2004), and Kalign (Lassmann and Sonnhammer 2005) use a substitution matrix to assess the cost of matching two symbols or two profiled columns (Notredame 2007). Conversely, consistency-based schemes incorporate a larger share of information into the evaluation. This result is achieved by using an approach initially developed for T-Coffee (Notredame et al. 2000) and inspired by Dialign overlapping weights (Morgenstern et al. 1998; Subramanian et al. 2005). Its principle is to compile a collection of pairwise global and local alignments (primary library) and to use this collection as a position-specific substitution matrix during a regular progressive alignment. The aim is to deliver a final MSA as consistent as possible with the alignments contained in the library. Many extant algorithms are based on this approach such as PCMA (Pei et al. 2003), ProbCons (adopting a Bayesian framework) (Do et al. 2005), MUMMALS (Pei and Grishin 2007). Sequence and structural databases are expanding rapidly owing to genome sequencing projects and structural genomics initiatives, offering helpful sources to further improve multiple protein sequence alignments. Structural additional information, for example known 3-dimensional (3D) structures, can be exploited in some multiple alignment methods. In fact, since structures are generally more conserved than sequences, structural information is also valuable for aligning sequences. Several MS algorithm have started implementing this source of information, and they include 3DCoffee (Poirot et al. 2004) and FUGUE (Shi et al. 2001). Recently, the Expresso server (Armougom et al. 2006) extended the 3DCoffee method by automatically identifying highly similar 3D structural templates for target sequences and using structural alignments for consistency-based alignments.

6.4 Phylogeny Reconstruction

Understanding microbial evolution is essential for gathering information on the most ancient events in the history of Life on our planet (Gribaldo and Brochier 2009) as well as on the extant relationships between whole microbial communities. This task

implies the use of molecular phylogeny techniques that is the study of phylogenies and processes of evolution by the analysis of DNA or amino acid sequence data (Whelan and Goldman 2001). Although parsimony and distance-based methods are widely used, the most statistically robust approach is to consider the problem in a likelihood framework and use accurate models of evolution (Brilli et al. 2008). It is known (Whelan and Goldman 2001), in fact, that disadvantages of distance methods include the inevitable loss of evolutionary information when a sequence alignment is converted to pairwise distances, and the inability to deal with models containing parameters for which the values are not known a priori. Concerning maximum parsimony (MP), this approach selects and outputs the tree (or trees) that require the fewest evolutionary changes and is reasonably confident when the number of changes per sequence position is relatively small (Steel and Penny 2000). However, as more-divergent sequences are to be analyzed, the degree of homoplasy (i.e. parallel, convergent, reversed or superimposed changes) increases and MP tree reconstruction might be misleading since this method has no adequate means to deal with this (Whelan and Goldman 2001). Conversely, Maximum likelihood (ML) approaches take the hypothesis (the tree topology) that maximizes the likelihood of the data (the sequence alignment) in the light of an evolutionary model. A great attraction of this approach is the ability to perform robust statistical hypothesis tests and to use modern statistical techniques such as hidden Markov models, Markov chain Monte Carlo and Bayesian inference (Ewens and Grant 2005; Shoemaker et al. 1999). The ML framework also allows each site of the alignment to evolve with different replacement patterns, and with different substitution rates in all branches of the tree (Whelan and Goldman 2001) as in real proteins, where slowly evolving sites are generally functionally or structurally constrained, while variable sites are likely to be less important for protein function. The ML approach (including its variants as the Bayesian framework) has been included in a number of different packages, such as Phylip (http://evolution.gs.washington.edu/phylip.html) PAUP* (http://paup.csit.fsu.edu/) MEGA http://www.megasoftware.net/mega.html, (Tamura et al. 2008), PAML [http://abacus.gene.ucl.ac.uk/software/paml.html, (Yang 1997)], mrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) and phyML (Guindon and Gascuel 2003).

Bibliography

Abascal, F., & Valencia, A. (2002). Clustering of proximal sequence space for the identification of protein families. *Bioinformatics, 18*(7), 908–921.

Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., & Arkin, A. P. (2005). The MicrobesOnline web site for comparative genomics. *Genome Research, 15*(7), 1015–1022.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389–3402.

Armougom, F., Moretti, S., Keduas, V., & Notredame, C. (2006). The iRMSD: A local measure of sequence alignment accuracy using structural information. *Bioinformatics, 22*(14), e35–e39.

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., & Eisenberg, D. (2004). Prolinks: A database of protein functional linkages derived from coevolution. *Genome Biology, 5*(5), R35.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Liò, P., & Fani, R. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics, 9*(1), 551.

Cannon, S. B., & Young, N. D. (2003). OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics, 4*(1), 35.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., & Cesareni, G. (2009). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, gkp98.

Checcucci, A., & Mengoni, A. (2015). The integrated microbial genome resource of analysis in bacterial pangenomics (pp. 289–295). Springer, New York.

Chu, K., Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Szeto, E., Pillay, M., & Kyrpides, N. C. (2013). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*, gkt963.

Cotter, P. J., Caffrey, D. R., & Shields, D. C. (2002). Improved database searches for orthologous sequences by conditioning on outgroup sequences. *Bioinformatics, 18*(1), 83–91.

Dehal, P. S., Joachimiak, M. P., Price, M. N., Bates, J. T., Baumohl, J. K., Chivian, D., et al. (2010). MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Research, 38*(suppl 1), D396–D400.

Do, C. B., Mahabhashyam, M. S., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research, 15*(2), 330–340.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792–1797.

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology, 16*(3), 368–373.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research, 30*(7), 1575–1584.

Ewens, W. J., & Grant, G. R. (2005). *Statistical methods in bioinformatics: An introduction* (Vol. 746867830). New York: Springer.

Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M., & Brinkman, F. S. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics, 7*(1), 270.

Gabaldón, T., Silla-Martínez, J. M., & Capella-Gutiérrez, S. (2009). TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972–1973.

Gribaldo, S., & Brochier, C. (2009). Phylogeny of prokaryotes: Does it exist and why should we care? *Research in Microbiology, 160*(7), 513–521.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology, 52*(5), 696–704.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., & Apweiler, R. (2004). The HUPO PSI's molecular interaction format—A community standard for the representation of protein interaction data. *Nature Biotechnology, 22*(2), 177–183.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754–755.

Hulsen, T., Huynen, M. A., de Vlieg, J., & Groenen, P. M. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology, 7*(4), R31.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009). STRING 8—A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research, 37*(suppl 1), D412–D416.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27–30.

Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., & Karp, P. D. (2009). EcoCyc: A comprehensive view of Escherichia coli biology. *Nucleic Acids Research, 37*(suppl 1), D464–D470.

Lassmann, T., & Sonnhammer, E. L. (2005). Kalign—An accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics, 6*(1), 298.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research, 13*(9), 2178–2189.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research, 37*(suppl 1), D619–D622.

Morgenstern, B., Frech, K., Dress, A., & Werner, T. (1998). DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics, 14*(3), 290–294.

Notredame, C. (2002). Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics, 3*(1), 131–144.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology, 3*(8), e123.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217.

O'Brien, K. P., Remm, M., & Sonnhammer, E. L. (2005). Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Research, 33*(suppl 1), D476–D480.

Pei, J. (2008). Multiple protein sequence alignment. *Current Opinion in Structural Biology, 18*(3), 382–386.

Pei, J., & Grishin, N. V. (2007). PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics, 23*(7), 802–808.

Pei, J., Sadreyev, R., & Grishin, N. V. (2003). PCMA: Fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics, 19*(3), 427–428.

Pertea, G., Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., et al. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology, 10*(4), R42.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E., & Notredame, C. (2004). 3DCoffee@ igs: A web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Research, 32*(suppl 2), W37–W40.

Prasad, T. K., Kandasamy, K., & Pandey, A. (2009). Human protein reference database and human proteinpedia as discovery tools for systems biology. In *Reverse Chemical Genetics* (pp. 67–79). Humana Press.

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics, 19*(12), 1572–1574.

Shi, J., Blundell, T. L., & Mizuguchi, K. (2001). FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology, 310*(1), 243–257.

Shoemaker, J. S., Painter, I. S., & Weir, B. S. (1999). Bayesian statistics in genetics: A guide for the uninitiated. *Trends in Genetics, 15*(9), 354–358.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., et al. (2008). The BioGRID interaction database: 2008 update. *Nucleic Acids Research, 36*(suppl 1), D637–D640.

Steel, M., & Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution, 17*(6), 839–850.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545–15550.

Tamura, K., Kumar, S., Nei, M., & Dudley, J. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics, 9*(4), 299–306.

Thompson, J. D., Gibson, T., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics,* 2–3.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution, 18*(5), 691–699.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research, 30*(1), 303–305.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS, 13*(5), 555–556.

Zmasek, C. M., & Eddy, S. R. (2002). RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics, 3*(1), 14.

7 Comparative genomics

The large amount of genomic data deeply impacted on our understanding of the biology, diversity and evolution of bacteria. Applications of genome sequencing analysis techniques and computational approaches to genome comparison are the topic of the following chapter. Notably, two diverse case study are presented. Both show how analysis of genetic features through bioinformatic approaches can pave the way to wet-lab experimental procedures, revealing relevant clues and features deserving further investigations. In the first work a deep genetic characterization of a biotechnological strain, *Rhodopseudomonas palustris* 42OL, is presented. This strain, a facultative anaerobe, has a wide metabolic versatility which allowed it to spread throughout a variety of habitats. Thanks to its ability to efficiently produce hydrogen, *R. palustris* 42OL has been used in mixed culture for wastewater treatment. However, since it is able to survive in multiple conditions (it is autotrophic, heterotrophic, organotrophic, litotrophic, chemotrophic and phototrophic), it could be particularly suitable for a broad spectrum of biotechnological fields. For instance, its biomass was evaluated for SCP accumulation and amino acid composition [1] and used as a bio-sorbent for metal removal from wastewaters [2, 3]. Additionally it has been used for antibiotic delivery though liposomes formed with its lipids [4]. Given this long list of addresses and its versatility it was considered a good candidate for DNA sequencing. The availability of the brand new genome allowed for functional inquiry, structural sinteny analysis and phylogenetics studies, as reported in the article embedded in this chapter. The computational analysis confirmed the genetic features allowing for hydrogen and poly-beta-hydroxy-butyrate production. Such genetic marks of metabolic capabilities may help the genetic editing for large-scale industrial use.

The second case study deals with the genetic and phenotypic characterization of *Rheneimera* sp. EpRS3, an endophytic bacterial strain isolated from the medicinal plant *Echinacea purpurea*. Today there is an increasing interest in the microbiota of such plants, either from an ecological and pharmacological point of view [5]. Medicinal plants, in fact, harbor highly complex bacterial communities with peculiar genetic features to survive the strong competition undergoing inside the tissues and in the rhizosphere. Some of these members may promote plants growth through various mechanisms [6].

Some other can exploit active ecological roles, producing molecules involved in plant's defense mechanisms and structuring the communities of specific organs. The last, by exhibiting antimicrobial activity, may also be medically relevant strains. This was the case of *Rheneimera* sp. EpRS3 [7, 8]. The outcomes of its genome sequencing project, reported in the second article of the chapter, highlighted potential, functional clues that guided wet-lab tests. The investigation of its resistome and secondary metabolite production, through comparative methods, revealed the presence of genetic traits suitable for experimental inquiry. The analysis through antiSMASH suite [9], revealed that the genome harbors 111 genes, split in 8 clusters, associated with the synthesis of various molecules exhibiting antimicrobial activity. In particular, *Rheinheimera* sp EpRS3 has:

1. two different putative bacteriocin clusters (a class of extracellular small peptides exhibiting bactericidal activity) with a comparable number of genes (6 and 7, respectively);

2. two clusters associated with the synthesis of lantipeptides, harboring 11 and 12 genes each (lantipeptides are a class of polycyclic peptides characterized by the presence of the thioether-cross-linked amino acids meso-lanthionine (Lan) and (2S, 3S, 6R)-3-methyllanthionine);

3. three biosynthetic clusters involved in production of different compounds: resorcinol, non-ribosomal peptides (NRPS), polyketides (PKS).

The presence of PKS synthases and NRPS synthases have a topical relevance for future biotechnological exploitation of this strain, in that some of the most successful antibiotics (*i.e.* vancomycin, avermectin, erythromycin) are produced by these enzymes.

Bibliography

1. De Philippis, R., Ena, A., Guastiini, M., Sili, C. & Vincenzini, M. Factors affecting poly-β-hydroxybutyrate accumulation in cyanobacteria and in purple non-sulfur bacteria. *FEMS Microbiology Letters* **103,** 187–194 (1992) (cit. on p. 93).
2. Colica, G., Caparrotta, S. & De Philippis, R. Selective biosorption and recovery of Ruthenium from industrial effluents with Rhodopseudomonas palustris strains. *Applied microbiology and biotechnology* **95,** 381–387 (2012) (cit. on p. 93).
3. Bertini, I., Dikiy, A., Luchinat, C., *et al.* An NMR study of the 7Fe-8S ferredoxin from Rhodopseudomonas palustris and reinterpretation of data on similar systems. *Biochemistry* **36,** 3570–3579 (1997) (cit. on p. 93).
4. Bianchi, L., Mannelli, F., Viti, C., Adessi, A. & De Philippis, R. Hydrogen-producing purple non-sulfur bacteria isolated from the trophic lake Averno (Naples, Italy). *International Journal of Hydrogen Energy* **35,** 12216–12223 (2010) (cit. on p. 93).
5. Maida, I., Chiellini, C., Mengoni, A., *et al.* Antagonistic interactions between endophytic cultivable bacterial communities isolated from the medicinal plant Echinacea purpurea. *Environmental microbiology* **18,** 2357–2365 (2016) (cit. on p. 93).
6. Lugtenberg, B. & Kamilova, F. Plant-growth-promoting rhizobacteria. *Annual review of microbiology* **63,** 541–556 (2009) (cit. on p. 93).
7. Chiellini, C., Maida, I., Emiliani, G., *et al.* Endophytic and rhizospheric bacterial communities isolated from the medicinal plants Echinacea purpurea and Echinacea angustifolia. *Int Microbiol* **17,** 165–174 (2014) (cit. on p. 94).

8. Mengoni, A., Maida, I., Chiellini, C., *et al.* Antibiotic resistance differentiates Echinacea purpurea endophytic bacterial communities with respect to plant organs. *Research in microbiology* **165,** 686–694 (2014) (cit. on p. 94).
9. Weber, T., Blin, K., Duddela, S., *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research* **43,** W237–W243 (2015) (cit. on p. 94).

Adessi et al. Standards in Genomic Sciences (2016) 11:24
DOI 10.1186/s40793-016-0145-y

Standards in Genomic Sciences

SHORT GENOME REPORT Open Access

Draft genome sequence and overview of the purple non sulfur bacterium *Rhodopseudomonas palustris* 42OL

Alessandra Adessi[1,2], Giulia Spini[1], Luana Presta[3], Alessio Mengoni[3], Carlo Viti[1], Luciana Giovannetti[1], Renato Fani[3] and Roberto De Philippis[1,2*]

Abstract

Rhodopseudomonas palustris strain 42OL was isolated in 1973 from a sugar refinery waste treatment pond. The strain has been prevalently used for hydrogen production processes using a wide variety of waste-derived substrates, and cultured both indoors and outdoors, either freely suspended or immobilized. *R. palustris* 42OL was suitable for many other applications and capable of growing in very different culturing conditions, revealing a wide metabolic versatility. The analysis of the genome sequence allowed to identify the metabolic pathways for hydrogen and poly-β-hydroxy-butyrate production, and confirmed the ability of using a wide range of organic acids as substrates.

Keywords: *Rhodopseudomonas palustris* 42OL, Purple non-sulfur bacteria, Hydrogen production, Wastewater treatment, PHB accumulation

Introduction

Rhodopseudomonas palustris is a PNSB belonging to the class *Alphaproteobacteria*. According to Imhoff et al. [1], the term PNSB is used to indicate a physiological group of anoxygenic phototrophic bacteria, affiliated to both *Alphaproteobacteria* and *Betaproteobacteria*, containing photosynthethic pigments and able to carry out anoxygenic photosynthesis.

Strains of *R. palustris* have been isolated from a variety of different environments, from eutrophic lagoons to moist soils, from freshwater ponds to marine coastal sediments [2–4]. The very wide spread of *R. palustris* throughout a variety of habitats is due to its extreme metabolic versatility, with all modes of metabolism represented (autotrophic, heterotrophic, organotrophic, litotrophic, chemotrophic and phototrophic); moreover, the organism is a facultative anaerobe [5].

All PNSBs are characterized by the ability of carrying out anoxygenic photosynthesis; in the presence of oxygen, photosynthesis is inhibited and a number of PNSBs are able to carry out respiration [4]. Under anaerobic conditions, and subject to light irradiation, PNSBs are able to fix nitrogen via nitrogenase; hydrogen is produced as a by-product of nitrogen fixation. Among PNSBs, *R. palustris* is considered a model organism for studying biological hydrogen production, due to its capacity of efficiently producing hydrogen during organic wastes degradation [6].

R. palustris 42OL has been used previously for hydrogen production processes under various conditions [7–22], i.e., with different substrates, and cultured indoors and outdoors, using freely suspended or immobilized cells. However, its first application was in mixed culture in wastewater treatment [23]. Its biomass was evaluated for SCP accumulation and amino acid composition [24]. The accumulation of PHB and its connection to hydrogen production were investigated [8, 11, 25]. More recently, the biomass of *R. palustris* 42OL was also used as a biosorbent for metal removal from wastewaters [26, 27]. Furthermore, the NMR and X-ray structures of its 7Fe-8S ferredoxin and cytochrome c_2

* Correspondence: roberto.dephilippis@unifi.it
[1]Department of Agrifood Production and Environmental Sciences, University of Florence, via Maragliano 77, 50144 Firenze, Italy
[2]Institute of Chemistry of Organometallic Compounds, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy
Full list of author information is available at the end of the article

were studied [28–30]. The latest application of the strain was for antibiotic delivery though liposomes formed with its lipids [31].

The long history and the versatility of this particular strain render it a very good candidate for further investigating the basis of its ability to acclimate to very different culturing conditions.

Organism information

Classification and features

R. palustris 42OL was isolated in 1973 from a catch-basin collecting the effluents of a sugar refinery waste treatment pond, in Castiglion Fiorentino (AR), Italy. The enrichment was carried out aimed at selecting waste degrading phototrophs. The isolated microorganism has been stored since 1973 at CSMA Collection (WDCM number 147) under the collection name CSMA73/42, growing anaerobically on solid RPN medium [32] with malate 2 g L^{-1} as the carbon source and supplemented with 0.4 g L^{-1} of yeast extract. The general features of the isolate are reported in Table 1.

The isolate 42OL was firstly assigned morphologically to *R. palustris*. Phylogenetic analysis performed subsequently (unpublished results) by comparing 16S rRNA gene sequences revealed that the isolate might be indeed affiliated to the species *R. palustris*. With the present work, a further phylogenetic analysis was conducted and, as shown in the phylogenetic tree in Fig. 1, confirms its affiliation.

R. palustris 42OL is a Gram-negative rod shaped bacterium, of 0.6–1.2 μm (see Fig. 2a). It replicates by polar budding (Fig. 2b) and new cells present one single *flagellum* that is lost in the subsequent phases of cell cycle [33, 34]. Its photosynthetic apparatus is located on lamellar ICMs, clearly visible in Fig. 2c, d, as characteristic of this species [2]. The major carotenoid molecules that

Table 1 Classification and general features of *Rhodopseudomonas palustris* 42OL, according to MIGS standards [45]

MIGS ID	Property	Term	Evidence code[a]
	Classification	Domain *Bacteria*	TAS [46]
		Phylum *Proteobacteria*	TAS [47]
		Class *Alphaproteobacteria*	TAS [48, 49]
		Order *Rhizobiales*	TAS [49, 50]
		Family *Bradyrhizobiaceae*	TAS [49, 50]
		Genus *Rhodopseudomonas*	TAS [1, 51, 52]
		Species *Rhodopseudomonas palustris*	TAS [51, 53, 54]
		strain: 42OL *(CSMA73/42)*	
	Gram stain	Negative	NAS
	Cell shape	Rod	IDA
	Motility	Motile only during first part of cell cycle	NAS
	Sporulation	Non sporulating	
	Temperature range	mesophilic	NAS
	Optimum temperature	28–30 °C	IDA
	pH range; Optimum	6.0–8.0; 6.8	TAS [32]
	Carbon source	VFA, CO_2	IDA
MIGS-6	Habitat	Sugar refinery waste pond	IDA
MIGS-6.3	Salinity	Not determined	
MIGS-22	Oxygen requirement	Facultatively anaerobic	IDA
MIGS-15	Biotic relationship	Free-living	NAS
MIGS-14	Pathogenicity	Non-pathogen	NAS
MIGS-4	Geographic location	Castiglion Fiorentino, AR, Italy	IDA
MIGS-5	Sample collection	1973	IDA
MIGS-4.1	Latitude	43° 19' 30.054"	IDA
MIGS-4.2	Longitude	11° 53' 18.4518"	IDA
MIGS-4.4	Altitude	248 m	IDA

[a]Evidence codes - *IDA* inferred from direct assay, *TAS* traceable author statement (i.e., a direct report exists in the literature), *NAS* non-traceable author statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [55]

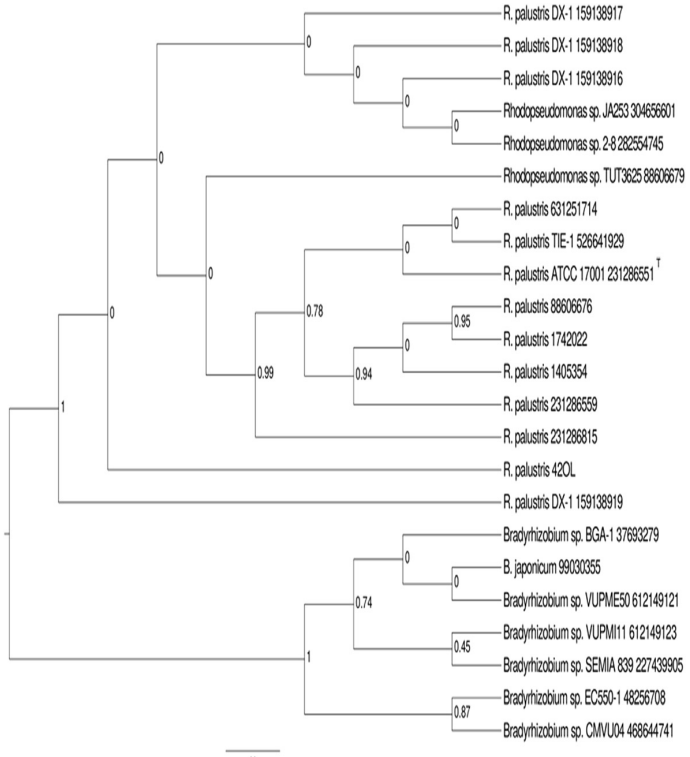


Fig. 1 Maximum Likelihood dendrogram based on 16S rRNA gene; Bar = 1.1 indicates the nucleotides substitution rate. Numbers at the nodes indicate bootstrap values after 500 random replicates. Numbers after strain name indicate the GI code. Sequences have been selected after BLAST Explorer [56] search for most similar sequences present in GenBank database. Dendrogram has been constructed by using the Maximum Likelihood algorithm with default options present in phylogeny.fr web server [56]. Strains TIE-1 and DX-1 have completely sequenced genomes; ATCC17001 is the type strain and is indicated as^{T}

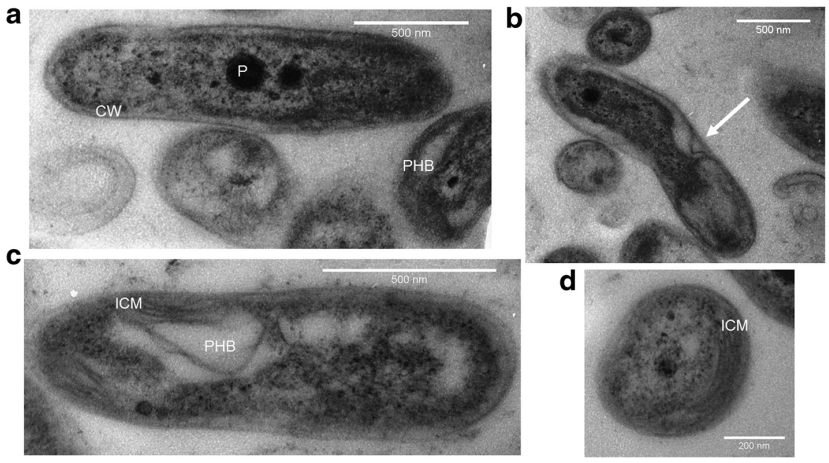


Fig. 2 Electron micrographs of *Rhodopseudomonas palustris* 42OL grown on RPN medium; **a** whole cell, longitudinal section; **b** cell during polar budding (white arrow), longitudinal section; **c** whole cell containing PHB granules. **d** lamellar ICMs in whole cells, transversal section; PHB, poly-β-hydroxybutyrate granules; ICM, intra-cytoplasmic membranes; P, polyphosphate granules; CW, cell wall

are present in the photosynthetic apparatus of this strain are spirilloxanthin, rhodopin, rhodovibrin, anhydrorhodovibrin and lycopene (our unpublished results).

The first characterization of *R. palustris* 42OL was in terms of protein accumulation and amino acid composition on different carbon and nitrogen sources. Malate and ethanol were tested as carbon sources, both under nitrogen fixing (N_2 sparged) and non-fixing (NH_4^+ supplied) conditions. A significantly lower amount of protein was produced in presence of ethanol, while the nitrogen source did not have any effect. However, both nitrogen and carbon sources significantly influenced the amino acid composition of the protein biomass [24].

The carbon metabolism of the strain was investigated in terms of substrates that could be used for growth and hydrogen production. Short chain fatty acids such as acetate, pyruvate, lactate, malate and succinate were found to be photodissimilated by the strain with substrate conversion efficiency of 40, 52, 61, 56, and 67 %, respectively [35]. Butyrate was found to be suitable for growth and hydrogen production but with the significantly lower substrate-to-hydrogen conversion efficiency of 9 % (unpublished data).

Vincenzini et al. [36] characterized nitrogenase activity of *R. palustris* 42OL in terms of hydrogen production and acetylene reduction with different atmosphere composition and at different pH values. The optimal conditions for hydrogen production were under Argon gas as the atmosphere, for early logarithmic stage cells, at pH 6.8. The authors also demonstrated the presence of a hydrogenase enzyme, recycling the hydrogen produced during late logarithmic and light-limited stage.

Evidences of its suitability for the treatment of wastes combined with hydrogen production were reported [35], using wastewaters deriving from a sugar refinery and a paper mill. The same strain was shown to grow and produce hydrogen on different substrates such as vegetable wastes [15], olive mill wastewaters [12, 13] and dark fermentation saline effluents [14].

PHB is synthetized as a reservoir for reducing equivalents by the strain, in this way competing with hydrogen production [25], especially when grown on acetate [25], or when subject to phosphorus starvation [11]; in this condition, *R. palustris* 42OL could accumulate up to 18 % w/w of PHB on cell dry weight. *R. palustris* 42OL accumulates PHB in large amorphous granules, as shown in Fig. 2. Glycogen is synthetized as well as carbon and energy reserve [25].

Another relevant characteristic of this strain is the possibility of cultivation outdoors, under light/dark cycles both for biomass [37, 38] and hydrogen production [10, 11, 22], with an impressive capability of the photosystem to take advantage of the high light conditions that take place during the central hours of the day [21].

Genome sequencing information

Genome project history

The organism was selected for genome sequencing on the basis of its metabolic versatility and biotechnological relevance, as witnessed by its long history and by the diversity of applications. Project information is available from the Genomes OnLine Database [39], under the GOLD study ID Gs0114708. The WGS sequence is deposited in GenBank (LCZM00000000).

Growth conditions and genomic DNA preparation

R. palustris 42OL (CSMA73/42) was maintained anaerobically on solid RPN medium [32] with malate 2 g L^{-1} as the carbon source and supplemented with 0.4 g L^{-1} of yeast extract. For the extraction of genomic DNA a single colony of cells grown on agar plate was harvested and cultured anaerobically on the same liquid medium in 20 mL sealed glass tubes, at room temperature with a light irradiance of 80 μmol of photons m^{-2} s^{-1}. Cultures were then transferred into 100 ml round bottles and the headspace was exchanged with Argon gas for anaerobiosis. Cells were harvested at an $OD_{660} = 0.5$, in mid-logarithmic phase, pelleted and stored at −20 °C. DNA was isolated from the cells using a CTAB bacterial genomic DNA isolation method, and checked on agarose gel. The genomic DNA purity was assessed by spectrophotometric measurements [40].

Genome sequencing and assembly

The draft genome sequence was generated using the Illumina technology. A Nextera XT DNA library was constructed and sequenced using Illumina MiSeq platform which generated 23,625,870 reads. After trimming, a total of 7,574,912 paired end reads were obtained and assembled into 308 high quality contigs (larger than 5419 bp each) using Abyss 1.0.0 software present on the Galaxy Orione server [41]. A summary of the project information is shown in Table 2.

Genome annotation

Genes were identified using the prokaryotic genome annotation software Prokka 1.4.0 [42] (Galaxy Orione server [40]). For gene finding and translation, Prokka makes use of the program Prodigal [43]. Homology searching (BLAST, hmmscan) was then performed using the translated protein sequences as queries against a set of public databases (CDD, PFAM, TIGRFAM) as well as custom databases that come with Prokka. Additional gene prediction analysis and functional annotation were performed within the CBS Bioinformatics Tools platform developed by the Technical University of Denmark (Table 3).

Table 2 Project information

MIGS ID	Property	Term
MIGS 31	Finishing quality	High-quality Draft
MIGS-28	Libraries used	Paired-end Nextera XT DNA
MIGS 29	Sequencing platforms	Illumina MiSeq
MIGS 31.2	Fold coverage	366 ×
MIGS 30	Assemblers	Abyss version 1.0.0 (Galaxy/CRS4 Orione server)
MIGS 32	Gene calling method	Prokka version 1.4.0 (Galaxy/CRS4 Orione server)
	Locus Tag	AB661
	Genbank ID	LCZM00000000
	GenBank Date of Release	5 June 2015
	GOLD ID	Gs0114708
	BIOPROJECT	PRJNA283573
MIGS 13	Source Material Identifier	CSMA73/42
	Project relevance	Metabolic versatility (hydrogen production), Biotechnology

Genome properties

The genome of *R. palustris* 42OL resulted to be 5,128,858 bp in length with a GC content of about 65.74 % (Table 3). It was predicted to contain 4767 genes, 4715 of which coded for proteins and 52 for RNA (tRNA and rRNA). The majority of the predicted genes (68.74 %) could be assigned to one of of 25 functional COG categories whilst the 8.42 % of the remaining genes were annotated as hypothetical and 38.9 % as unknown function proteins. The distribution of genes into COGs functional categories is presented in Table 4.

Table 3 Genome statistics*

Attribute	Value	% of total
Genome size (bp)	5,128,858	100.00
DNA coding (bp)	4,388,835	85.00
DNA G + C (bp)	3,369,731	65.74
DNA scaffolds	1	100.00
Total genes	4767	100.00
Protein coding genes	4715	98.91
RNA genes	52	1.09
Pseudo genes	NA	NA
Genes in internal clusters	NA	NA
Genes with function prediction	3277	68.74
Genes assigned to COGs	3660	76.78
Genes with Pfam domains	3312	69.48
Genes with signal peptides	449	9.41
Genes with transmembrane helices	1212	25.42
CRISPR repeats	1	0.09

*NA, not available

Table 4 Number of genes associated with general COG functional categories

Code	Value	% age	Description
J	170	3.61	Translation, ribosomal structure and biogenesis
A	0	0.00	RNA processing and modification
K	218	4.62	Transcription
L	144	3.05	Replication, recombination and repair
B	1	0.02	Chromatin structure and dynamics
D	25	0.53	Cell cycle control, Cell division, chromosome partitioning
V	57	1.21	Defense mechanisms
T	192	4.07	Signal transduction mechanisms
M	215	4.56	Cell wall/membrane biogenesis
N	80	1.70	Cell motility
U	37	0.78	Intracellular trafficking and secretion
O	165	3.50	Posttranslational modification, protein turnover, chaperones
C	267	5.66	Energy production and conversion
G	169	3.58	Carbohydrate transport and metabolism
E	358	7.59	Amino acid transport and metabolism
F	59	1.25	Nucleotide transport and metabolism
H	145	3.08	Coenzyme transport and metabolism
I	239	5.07	Lipid transport and metabolism
P	242	5.13	Inorganic ion transport and metabolism
Q	97	2.06	Secondary metabolites biosynthesis, transport and catabolism
R	397	8.42	General function prediction only
S	383	8.12	Function unknown
–	1055	22.38	Not in COGs

The total is based on the total number of protein coding genes in the genome

Insights from the genome sequence

The genome of *R. palustris* 42OL contained, as expected, genes related to nitrogen fixation (*nif* H, D, K, E, N, B, U, X, Q, W, Z), genes involved in carbon fixation (RubisCO), the complete tricarboxylic acid cycle, the glyoxylate shunt, a Embden-Meyerhof pathway, and a pentose phosphate pathway. Genes coding for the synthesis of glycogen and poly-β-hydroxyalkanoates as carbon storage polymers were also found, as well as genes related to the photosynthetic apparatus, similarly to all the other *R. palustris* strains so far sequenced.

The genome of *R. palustris* 42OL was analyzed in terms of synteny with other strains sequenced. The genome was found to be highly syntenic with those of other strains of *R. palustris* (data not shown). Exclusive reactions were then mapped on KEGG with respect to other *R. palustris* strains sequenced so far (BisA53, BisB18, BisB5, CGA009, Haa2, TIE1) by using DuctApe v 0.17.2 software [44]. Data obtained are reported in Table 5. The proteome size ranged between 4392 and 5242 protein

Table 5 DuctApe analysis report

Strain ID	Proteome size	Mapped to KEGG	Reactions	Exclusive
420L	4715	2383	2775	0
BisA53	4851	2297	2442	5
BisB18	4864	2452	2523	19
BisB5	4392	2266	2787	4
CGA009	4811	2542	3012	0
Haa2	4680	2446	2881	31
TIE1	5242	2566	2984	0

coding genes, corresponding respectively to strains BisB5 and TIE1. The total number of reactions ranged between 2442 and 3012, respectively for strains BisA53 and CGA009. Strain Haa2 resulted to have the highest number of unique reactions.

Conclusions

In this study, we characterized the genome of *R. palustris* strain 42OL isolated from a wastewater pond of a sugar refinery in 1973. Along the last four decades, this strain has been successfully used in a wide number of applications, from hydrogen production on wastewaters (its major application) to PHB production. The present genome analysis supported those findings.

Abbreviations

CSMA: Centro Studi Microrganismi Autotrofi; CTAB: Cetyl trimethyl ammonium bromide; ICM: intra-cytoplasmic membrane; PHB: poly-β-hydroxy butyrate; PNSB: purple non sulfur bacterium; SCP: single cell protein.

Competing interests

None of the authors has any competing interests in the manuscript.

Authors' contributions

AA coordinated the study and the manuscript drafting. AA cultured the strain and analyzed electron micrographs. GS extracted the DNA. GS, AM, and LP performed all the bioinformatics analysis. AA, GS, AM and LP drafted the manuscript. All Authors discussed and revised the manuscript. All Authors commented on the manuscript before submission. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge Professor Massimo Vincenzini, University of Florence, for sharing his experience and for having first elaborated the wide metabolic potentials of *R. palustris* 42OL strain.
The genome sequencing was performed by Francesca Crucianelli at Polo GGB, Perugia, Italy. The electron micrographs were performed by Maria Cristina Salvatici at CE.M.E., CNR, Florence, Italy.
Authors acknowledge the contribution given to this study by *Ente Cassa di Risparmio di Firenze (ECRF)* that supported AA with a fellowship in the frame of Florence Hydrolab[2] Project.

Author details

[1]Department of Agrifood Production and Environmental Sciences, University of Florence, via Maragliano 77, 50144 Firenze, Italy. [2]Institute of Chemistry of Organometallic Compounds, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy. [3]Department of Biology, University of Florence, Via Madonna del Piano 6, 50019 Sesto Fiorentino, Italy.

Received: 8 July 2015 Accepted: 3 November 2015
Published online: 09 March 2016

References

1. Imhoff JF, Trüper HG, Pfennig N. Rearrangements of the species and genera of the phototrophic "purple nonsulfur bacteria. Int J Syst Bacteriol. 1984;34:340–3.
2. Imhoff JF, Hiraishi A, Süling J. Anoxygenic Phototrophic Purple Bacteria. In: Bergey's Manual® of Systematic Bacteriology. US: Springer; 2005. p. 119–32.
3. Imhoff JF, Trüper HG. The genus *Rhodospirillum* and related genera. In: Balows A, Trüper HG, Dworkin M, Harder W and Schleifer KH, editors. The Prokaryotes. A Handbook on the Biology of Bacteria. Ecophysiology, Isolation, Identification, Applications. 2nd ed. New York: Springer; 1992. p. 2141–55.
4. Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, Land ML, Pelletier DA, Beatty JT, Lang AS, Tabita FR, Gibson JL, Hanson TE, Bobst C, Torres y Torres JL, Peres C, Harrison FH, Gibson J, Harwood CS. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nat Biotechnol. 2004;22:55–61.
5. McKinlay JB, Harwood CS. Photobiological production of hydrogen gas as a biofuel. Curr Opin Biotechnol. 2010;21:244–51.
6. Vincenzini M, Balloni W, Mannelli D, Florenzano G. A bioreactor for continuous treatment of waste waters with immobilized cells of photosynthetic bacteria. Experientia. 1981;37:710–2.
7. Vincenzini M, Materassi R, Tredici MR, Florenzano G. Hydrogen production by immobilized cells—I. light dependent dissimilation of organic substances by *Rhodopseudomonas palustris*. Int J Hydrog Energy. 1982;7:231–6.
8. Vincenzini M, Materassi R, Tredici MR, Florenzano G. Hydrogen production by immobilized cells—II. H_2-photoevolution and waste-water treatment by agar-entrapped cells of *Rhodopseudomonas palustris* and *Rhodospirillum molischianum*. Int J Hydrog Energy. 1982;7:725–8.
9. Vincenzini M, Materassi R, Sili C, Florenzano G. Hydrogen production by immobilized cells. III—Prolonged and stable H_2 photoevolution by *Rhodopseudomonas palustris* in light-dark cycles. Int J Hydrog Energy. 1986;11:623–6.
10. Vincenzini M, Marchini A, Ena A, De Philippis R. H_2 and poly-β-hydroxybutyrate, two alternative chemicals from purple non sulfur bacteria. Biotechnol Lett. 1997;19:759–62.
11. Pintucci C, Giovannelli A, Traversi ML, Ena A, Padovani G, Carlozzi P. Fresh olive mill waste deprived of polyphenols as feedstock for hydrogen photo-production by means of *Rhodopseudomonas palustris* 42OL. Renew Energy. 2013;51:358–63.
12. Padovani G, Pintucci C, Carlozzi P. Dephenolization of stored olive-mill wastewater, using four different adsorbing matrices to attain a low-cost feedstock for hydrogen photo-production. Bioresour Technol. 2013;138:172–9.
13. Dipasquale L, Adessi A, d' Ippolito G, Rossi F, Fontana A, De Philippis R. Introducing capnophilic lactic fermentation in a combined dark-photo fermentation process: a route to unparalleled H_2 yields. Appl Microbiol Biotechnol. 2014;99:1001–10.
14. De Philippis R, Bianchi L, Colica G, Bianchini C, Peruzzini M, Vizza F. From vegetable residues to hydrogen and electric power: Feasibility of a two step process operating with purple non sulfur bacteria. J Biotechnol. 2007;131:S122.
15. Carlozzi P, Lambardi M. Fed-batch operation for bio-H_2 production by *Rhodopseudomonas palustris* (strain 42OL). Renew Energy. 2009;34:2577–84.
16. Carlozzi P. The effect of irradiance growing on hydrogen photoevolution and on the kinetic growth in *Rhodopseudomonas palustris*, strain 42OL. Int J Hydrog Energy. 2009;34:7949–58.
17. Carlozzi P, Pintucci C, Piccardi R, Buccioni A, Minieri S, Lambardi M. Green energy from *Rhodopseudomonas palustris* grown at low to high irradiance values, under fed-batch operational conditions. Biotechnol Lett. 2010;32:477–81.
18. Carlozzi P, Buccioni A, Minieri S, Pushparaj B, Piccardi R, Ena A, Pintucci C. Production of bio-fuels (hydrogen and lipids) through a photofermentation process. Bioresour Technol. 2010;101:3115–20.
19. Carlozzi P. Hydrogen photoproduction by *Rhodopseudomonas palustris* 42OL cultured at high irradiance under a semicontinuous regime. J Biomed Biotechnol. 2012;2012:1–8.
20. Adessi A, Torzillo G, Baccetti E, De Philippis R. Sustained outdoor H_2 production with *Rhodopseudomonas palustris* cultures in a 50 L tubular photobioreactor. Int J Hydrog Energy. 2012;37:8840–9.
21. Adessi A, Fedini A, De Philippis R. Photobiological hydrogen production with *Rhodopseudomonas palustris* under light/dark cycles in lab and outdoor cultures. J Biotechnol. 2010;150:14–5.
22. Balloni W, Florenzano G, Materassi R. Lineamenti di un nuovo metodo di trattamento biologico degli effluenti di zuccherificio. Ind Sacc Ital. 1976;6:137–44.

23. Paoletti C, Citernesi U, Pushparaj B. Lo spettro aminoacidico di un ceppo di *Rhodopseudomonas* sp. azotofissatore. Rivista Ital Delle Sostanze Grasse. 1975;LI:42–3.
24. De Philippis R, Ena A, Guastini M, Sili C, Vincenzini M. Factors affecting poly-β-hydroxybutyrate accumulation in cyanobacteria and in purple non-sulfur bacteria. FEMS Microbiol Rev. 1992;9:187–94.
25. Colica G, Caparrotta S, Bertini G, De Philippis R. Gold biosorption by exopolysaccharide producing cyanobacteria and purple nonsulphur bacteria. J Appl Microbiol. 2012;113:1380–8.
26. Colica G, Caparrotta S, De Philippis R. Selective biosorption and recovery of Ruthenium from industrial effluents with *Rhodopseudomonas palustris* strains. Appl Microbiol Biotechnol. 2012;95:381–7.
27. Bertini I, Dikiy A, Luchinat C, Macinai R, Viezzoli MS, Vincenzini M. An NMR study of the 7Fe-8S ferredoxin from *Rhodopseudomonas palustris* and reinterpretation of data on similar systems. Biochemistry (Mosc). 1997;36:3570–9.
28. Garau G, Geremia S, Randaccio L, Vaccari L, Viezzoli MS. Crystallization and preliminary X-ray analysis of two pH-dependent forms of cytochrome c_2 *from Rhodopseudomonas palustris*. Acta Crystallogr D Biol Crystallogr. 2000;56:1699–701.
29. Geremia S, Garau G, Vaccari L, Sgarra R, Viezzoli MS, Calligaris M, Randaccio L. Cleavage of the iron-methionine bond in c-type cytochromes: Crystal structure of oxidized and reduced cytochrome c2 from *Rhodopseudomonas palustris* and its ammonia complex. Protein Sci Publ Protein Soc. 2002;11:6–17.
30. Colzi I, Troyan AN, Perito B, Casalone E, Romoli R, Pieraccini G, Škalko-Basnet N, Adessi A, Rossi F, Gonnelli C, Ristori S. Antibiotic delivery by liposomes from prokaryotic microorganisms: *similia cum similis* works better. Eur J Pharm Biopharm. in press.
31. Bianchi L, Mannelli F, Viti C, Adessi A, De Philippis R. Hydrogen-producing purple non-sulfur bacteria isolated from the trophic lake Averno (Naples, Italy). Int J Hydrog Energy. 2010;35:12216–23.
32. Whittenbury R, McLee AG. *Rhodopseudomonas palustris* and *Rh. viridis*—photosynthetic budding bacteria. Arch Für Mikrobiol. 1967;59:324–34.
33. Westmacott D, Primrose SB. Synchronous growth of *Rhodopseudomonas palustris* from the swarmer phase. J Gen Microbiol. 1976;94:117–25.
34. Vincenzini M, Tredici M, Ena A, Florenzano G. Produzione di idrogeno con *Rhodopseudomonas palustris* immobilizzato. Estr Atti 19 Congr Naz Della Soc Ital Microbiol. 1980;425–38.
35. Vincenzini M, Materassi R, Sili C, Balloni W. Evidence for an hydrogenase dependent H_2-producing activity in *Rhodopseudomonas palustris*. Ann Microbiol Ed Enzimologia. 1985;35:155–64.
36. Carlozzi P, Sacchi A. Biomass production and studies on *Rhodopseudomonas palustris* grown in an outdoor, temperature controlled, underwater tubular photobioreactor. J Biotechnol. 2001;88:239–49.
37. Carlozzi P, Pushparaj B, Degl'Innocenti A, Capperucci A. Growth characteristics of *Rhodopseudomonas palustris* cultured outdoors, in an underwater tubular photobioreactor, and investigation on photosynthetic efficiency. Appl Microbiol Biotechnol. 2006;73:789–95.
38. Pagani I, Liolios K, Jansson J, Chen I-M, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2012;40:D571–9.
39. Protocols and sample preparation information. [http://jgi.doe.gov/collaborate-with-jgi/pmo-overview/protocols-sample-preparation-information/] last accessed in December 2014.
40. Galaxy Orione Server. [https://orione.crs4.it/] last accessed in March 2015.
41. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
42. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
43. Galardini M, Mengoni A, Biondi EG, Semeraro R, Florio A, Bazzicalupo M, Benedetti A, Mocali S. DuctApe: a suite for the analysis and correlation of genomic and OmniLogTM Phenotype Microarray data. Genomics. 2014;103:1–10.
44. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26:541–7.
45. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87:4576–79.
46. The proteobacteria (part C). In: Garrity GM, Brenner DJ, Krieg NR, Staley JT, editors. Bergey's Manual of Systematic Bacteriology, Volume 2. 2nd ed. New York: Springer; 2005. p. 1.
47. Garrity GM, Bell JA, Lilburn TG, et al. Class I. Alphaproteobacteria class. nov. In: Bergey's Manual of Systematic Bacteriology, vol. 2. 2nd ed. New York: Springer; 2005. p. 1–574.
48. List Editor. Validation List No. 107. List of new names and new combinations previously effectively, but not validly, published. Int J Syst Evol Microbiol. 2006;56:1–6.
49. Garrity GM, Bell JA, Family LT, VII. Bradyrhizobiaceae fam. nov. In: Garrity GM, Brenner DJ, Krieg NR, Staley JT, editors. Bergey's Manual of Systematic Bacteriology, vol. 2. 2nd ed. Springer, New York: Part C; 2005. p. 438.
50. Skerman VBD, McGowan V, Sneath PHA. Approved lists of bacterial names. Int J Syst Bacteriol. 1980;30:225–420.
51. Czurda V, Maresch E. Beitrag zur Kenntnis der Athiorhodobakterien-Gesellschaften. Arch Mikrobiol. 1937;8:99–124.
52. Ramana W, Chakravarthy SK, Raj PS, Kumar BV, Shobha E, Ramaprasad EVV, Sasikala Ch, Ramana CV. Descriptions of *Rhodopseudomonas parapalustris* sp. nov., *Rhodopseudomonas harwoodiae* sp. nov. and *Rhodopseudomonas pseudopalustris* sp. nov., and emended description of *Rhodopseudomonas palustris*. Int J Syst Evol Microbiol. 2012;62:1790–8.
53. van Niel CB. The culture, general physiology, morphology, and classification of the non-sulfur purple and brown bacteria. Bacteriol Rev. 1944;8:1–118.
54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
55. Dereeper A, Audic S, Claverie JM, Blanc G. BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol. 2010;10:8.
56. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36:W465–9.

Research in Microbiology 168 (2017) 293–305
www.elsevier.com/locate/resmic

Original Article

Phenotypic and genomic characterization of the antimicrobial producer *Rheinheimera* sp. EpRS3 isolated from the medicinal plant *Echinacea purpurea*: insights into its biotechnological relevance

Luana Presta [a], Emanuele Bosi [a], Marco Fondi [a], Isabel Maida [a], Elena Perrin [a], Elisangela Miceli [a], Valentina Maggini [a,b], Patrizia Bogani [a], Fabio Firenzuoli [b], Vincenzo Di Pilato [c], Gian Maria Rossolini [d,e,f,g], Alessio Mengoni [a], Renato Fani [a,*]

[a] *Dept of Biology, University of Florence, Via Madonna del Piano 6, I-50019 Sesto F.no, FI, Italy*
[b] *Center for Integrative Medicine, Careggi University Hospital, University of Florence, Florence, Italy*
[c] *Department of Surgery and Translational Medicine, University of Florence, Florence, Italy*
[d] *Department of Medical Biotechnologies, University of Siena, Siena, Italy*
[e] *Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy*
[f] *Clinical Microbiology and Virology Unit, Careggi University Hospital, Florence, Italy*
[g] *Don Carlo Gnocchi Foundation, Florence, Italy*

Received 5 April 2016; accepted 14 November 2016
Available online 22 November 2016

Abstract

In recent years, there has been increasing interest in plant microbiota; however, despite medicinal plant relevance, very little is known about their highly complex endophytic communities. In this work, we report on the genomic and phenotypic characterization of the antimicrobial compound producer *Rheinheimera* sp. EpRS3, a bacterial strain isolated from the rhizospheric soil of the medicinal plant *Echinacea purpurea*. In particular, EpRS3 is able to inhibit growth of different bacterial pathogens (Bcc, *Acinetobacter baumannii*, and *Klebsiella pneumoniae*) which might be related to the presence of gene clusters involved in the biosynthesis of different types of secondary metabolites. The outcomes presented in this work highlight the fact that the strain possesses huge biotechnological potential; indeed, it also shows antimicrobial effects upon well-described multidrug-resistant (MDR) human pathogens, and it affects plant root elongation and morphology, mimicking indole acetic acid (IAA) action.

Keywords: Endophytes; Antibiotic resistance; Antibiotic production; Biotechnological applications; Genome analysis

1. Introduction

A diverse range of bacteria, including parasites, commensals and mutualists, can colonize the rhizosphere of plants, growing around their roots [1] and eventually colonizing plant internal tissues and surfaces (as endophytes or epiphytes). Despite medicinal plant relevance, very little is known about their plant-associated bacteria. In recent years, there has been increasing interest in medicinal plant microbiota; numerous efforts have been made to explore both endophytic and rhizospheric diversity [2–9]. Those studies showed that medicinal plants harbor highly complex bacterial communities, whose structure is possibly influenced by the plant organ colonized and secondary compounds with medicinal properties (e.g. plant essential oils) [7,8]. In particular, in a recent work on *Echinacea purpurea* [7], different antibiotic resistance phenotypes have been observed for rhizospheric and

* Corresponding author.
E-mail addresses: renato.fani@unifi.it, renato.fani@virgilio.it (R. Fani).

http://dx.doi.org/10.1016/j.resmic.2016.11.001

endophytic bacterial communities, suggesting the presence of a plethora of defense mechanisms and differential production of antimicrobial compounds by members of these bacterial communities.

Rheinheimera sp. EpRS3, a Gammaproteobacterium, belonging to the family of *Chromatiales*, is one such member. It was isolated from a sample of rhizospheric soil of *E. purpurea* [7,8], a medicinal plant with documented antimicrobial activity [6]. *Rheinheimera* sp. EpRS3 showed resistance to several antibiotic compounds and was able to inhibit growth of various bacteria isolated from both *E. purpurea* rhizospheric soil and plant tissues [7–9].

Therefore, the aim of this work was to characterize the *Rheinheimera* sp. EpRS3 strain from different viewpoints so as to evaluate its possible biotechnological potential, in terms of both the ability to affect plant growth and to interfere with growth of human bacterial pathogens. For this purpose, it was characterized at a phenotypic level by testing different traits (especially those related to production of antimicrobials and those involved in plant growth promotion); moreover, the genome sequence was determined and analyzed in order to correlate phenotypic data with the genetic repertoire.

2. Materials and methods

2.1. Bacterial strains, isolation and growth conditions

Rheinheimera sp. EpRS3 was isolated from the medicinal plant *E. purpurea* as described by Chiellini et al. [5]. This strain belongs to a collection of isolates from a pool of five *E. purpurea* plants that were grown in a common garden at the "Giardino delle Erbe", Casola Valsenio (Italy).

Rheinheimera sp. EpRS3 was grown either in tryptic soy agar (TSA) or in minimal medium Davis (MMD) (1.0 g l^{-1} of $(NH_4)_2SO_4$, 7.0 g l^{-1} of K_2HPO_4, 2.0 g l^{-1} of KH_2PO_4, 0.5 g l^{-1} of Na_3-citrate $2H_2O$, 0.1 g l^{-1} of $MgSO_4 \cdot 7H_2O$) and in minimal medium Venetia (MMV) (1.0 g l^{-1} of $MgSO_4 \cdot 7H_2O$, 0.7 g l^{-1} of KCl, 2.0 g l^{-1} of KH_2PO_4, 3.0 g l^{-1} of Na_2HPO_4, 1.0 g l^{-1} of NH_4NO_3, and 24.0 g l^{-1} of NaCl in deionized water) containing 0.4% diesel fuel or 1% glucose as the sole carbon and energy source. Diesel fuel (Esso Italiana) was previously filtered through a 0.2 μm-pore-size filter (Sartorius) for sterilization and particle removal. Bacterial cultures were incubated at 30 °C.

Burkholderia strains were grown in TSA medium at 37 °C.

2.2. Response to oxidative stresses and heavy-metals resistance

Resistance to oxidative stress was tested by evaluating *Rheinheimera* sp. EpRS3 growth on TSA medium supplemented with hydrogen peroxide, zinc, copper, cadmium or nickel (heavy metals). The strain was grown for 48 h (or until satisfactory growth) in each of the supplemented TSA media at 30 °C, and an isolated colony was then suspended in 100 μl of saline solution (0.85% NaCl). The cell suspension was then streaked onto TSA plates supplemented with different concentrations of: hydrogen peroxide (1-5-10-20 mM); zinc ($ZnSO_4$ 5-10-15-25 mM); copper ($CuSO_4$ 0.25-0.5-1-2.5-5-10-15-25 mM); cadmium ($Cd(NO_3)2$ 5-10-15-25 mM); nickel ($NiCl_2$ 5-10-15-25 mM). Strain growth was rated as absence of growth (i.e. sensitivity to exposure) or growth (i.e. tolerance at the tested oxidative stress). No quantitative assessment of the amount of growth was evaluated.

2.3. Extracellular enzymatic activity (EEA)

Extracellular amylolytic, proteolytic, phospholipase and lipase enzyme activities were assessed as follows. Amylolytic activity was evaluated by streaking *Rheinheimera* sp. EpRS3 onto 10% TSA medium enriched with 1% starch from potatoes (Fluka Analytical). After incubation at 30 °C for 48 h, 2 ml of Lugol's iodine solution was added to the plate. A positive reaction was defined as the observation of a pale yellow zone around the colony, indicating starch degradation [7].

Proteolytic activity was evaluated in 10% TSA supplemented with 1% skim milk powder (Fluka Analytical). After incubation for 48 h at 30 °C, a clear halo around the colony indicated milk degradation and thus a positive reaction [7].

Phospholipase activity was evaluated in 10% TSA supplemented with 2% egg yolk emulsion (Fluka Analytical). The cleavage of the phosphate ester bonds forms water-insoluble lipids. After incubation for 48 h at 37 °C, enzyme activity can be observed as a halo of opalescence in the opaque medium around the colony.

The presence of extracellular lipase was evaluated either in 10% TSA enriched with 1% (v/v) tributyrin (Sigma Aldrich) or in 10% TSA enriched with 1% (v/v) Tween 80 (Thermo Scientific). After incubation at 37 °C for 48 h, a positive reaction is indicated by a clear halo around the colony in opaque tributyrin agar medium, and by a precipitate around the colony in Tween 80 agar medium, respectively. Each experiment was performed in triplicate.

2.4. Antimicrobial activity assay *vs.* Burkholderia cepacia *complex,* Acinetobacter baumannii *and* Klebsiella pneumoniae *strains*

Antibacterial activity of *Rheinheimera* sp. EpRS3 (defined as a "tester strain") was tested using the cross-streak method as described by Papaleo et al. [8] against six *A. baumannii* (two of which, RUH 875 and RUH 134, are well-described MDR strains of the respective international clones II and I [9],) eleven *K. pneumoniae* and thirty *Burkholderia* strains, listed in Table 1 and referred to as "target strains". Moreover, in susceptibility testing, some *A. baumannii* and *K. pneumoniae* strains used in this work showed a colistin resistance phenotype (see Table 1).

In the assay against *Burkholderia* strains, the tester strain was streaked across two different half Petri dishes containing TSA medium either with or without a central septum separating the tester strain from target strains. Such an expedient has been used to test whether any antimicrobial effect occurring is detected only when molecules can reach the target by

Table 1
List of bacterial strains belonging to the *Burkholderia cepacia* complex (Bcc), *A. baumannii* and *K. pneumoniae* used in this work. Abbreviations: CF, Cystic Fibrosis; AI, Animal Infection; E, Environment; C, Clinical.

Position on plate	Taxonomy	Strain	Origin
1	*B. ambifaria*	LMG 19182	E
2	*B. anthina*	LMG 20980	E
3	*B. arboris*	LMG 24066	E
4	*B. cenocepacia*	LMG 16656	CF
5	*B. cepacia*	LMG 1222	E
6	*B. contaminans*	LMG 23361	AI
7	*B. diffusa*	LMG 24065	CF
8	*B. dolosa*	LMG 18943	CF
9	*B. lata*	LMG 22485	E
10	*B. latens*	LMG 24064	CF
11	*B. metallica*	LMG 24068	CF
12	*B. multivorans*	LMG 13010	CF
13	*B. pseudomultivorans*	LMG 26883	CF
14	*B. pyrrocinia*	LMG 14191	E
15	*B. seminalis*	LMG 24067	CF
16	*B. stabilis*	LMG 14294	CF
17	*B. ubonensis*	LMG 20358	E
18	*B. vietnamiensis*	LMG 10929	E
19	*B. cenocepacia*	FCF 12	CF
20	*B. cenocepacia*	FCF 13	CF
21	*B. cenocepacia*	FCF 14	CF
22	*B. cenocepacia*	FCF 15	CF
23	*B. cenocepacia*	FCF 18	CF
24	*B. cenocepacia*	FCF 19	CF
25	*B. cenocepacia*	LMG 16656	CF
26	*B. multivorans*	FCF 5	CF
27	*B. multivorans*	FCF 6	CF
28	*B. multivorans*	FCF7	CF
29	*B. multivorans*	FCF 8	CF
30	*B. multivorans*	FCF 9	CF

Position on plate	Taxonomy	Strain	Origin	Genotype	Phenotype
1	*A. baumannii*	RHU 134	C	–	–
2	*A. baumannii*	RHU 875	C	–	–
3	*A. baumannii*	ATCC17978	C	–	–
4	*A. baumannii*	O8C29	C	*blaOXA-23/blaOXA-58*	Beta-lactams R
5	*A. baumannii*	N5O	C	*blaOXA-24*	Beta-lactams R, col-R
6	*A. baumannii*	YMCR363	C	*blaVIM-2*	Beta-lactams R
1	*K. pneumoniae*	BO 1	C	bla_{KPC}	Resistant to: beta-lactams, fluoroquinolones, gentamycin; colistin S
2	*K. pneumoniae*	BO 4	C	bla_{KPC}	Resistant to: beta-lactams, fluoroquinolones, gentamycin; colistin R
3	*K. pneumoniae*	B1	C	bla_{KPC}	Resistant to: beta-lactams, fluoroquinolones, gentamycin; colistin S
4	*K. pneumoniae*	B2	C	bla_{KPC}	Resistant to: beta-lactams, fluoroquinolones, gentamycin; colistin R
5	*K. pneumoniae*	ATCC 700603	C	–	–
6	*K. pneumoniae*	KP397	C	$bla_{CTX\text{-}M\text{-}65}$/fosA	Resistant to: extended-spectrum cephalosporins, aminoglycosides, fluoroquinolones,
7	*K. pneumoniae*	KP477	C	fosA	Resistant to: extended-spectrum cephalosporins, aminoglycosides, fluoroquinolones, fosfomycin
8	*K. pneumoniae*	KPC249	C	bla_{KPC}/*armA*	Resistant to: beta-lactams, aminoglycosides
9	*K. pneumoniae*	KPC261	C	bla_{KPC}/*armA*	Resistant to: beta-lactams, aminoglycosides
10	*K. pneumoniae*	KPC284	C	bla_{KPC}/*armA*	Resistant to: beta-lactams, aminoglycosides
11	*K. pneumoniae*	NTUHK2044	C	–	–

diffusing into the medium, or if the antimicrobial effect is carried by a volatile compound that hence can also exert its effect when the tester and target are separated by a barrier. In tests to probe antimicrobial activity against *A. baumannii* and *K. pneumoniae*, we used the plates without septum.

All plates were then incubated at 30 °C for 48 h. Next, target strains were streaked perpendicularly to the initial streak in the empty half of the plate; plates with *Burkholderia* strains were incubated at 30 °C and 37 °C for an additional 48 h, respectively, while the other strains were incubated at 37 °C for 24 h. The antagonistic effect was indicated by growth inhibition of target strains in the confluence area. All experiments were performed in triplicate and in parallel with a positive control to verify viability of target strains.

2.5. Determination of fosfomycin minimum inhibitory concentration

The minimum inhibitory concentration (MIC) of fosfomycin was determined by the agar dilution reference method according to CLSI guidelines [10]. Mueller-Hinton agar plates, supplemented with 25 μg/ml of glucose-6-phosphate, were added with the following concentrations of the antibiotic: 0.064 μg/ml, 0.125 μg/ml, 0.256 μg/ml, 0.5 μg/ml, 1 μg/ml, 2 μg/ml, 4 μg/ml, 8 μg/ml, 16 μg/ml, 32 μg/ml, 64 μg/ml, 128 μg/ml, 256 μg/ml and 512 μg/ml. Starting from a 0.5 McFarland suspension of *Rheinehimera* EpRS3, a 1:10 dilution in saline solution was prepared and agar plates were inoculated with 5 μl spots from the diluted McFarland inoculum. Plates were then incubated at 37 °C overnight. *Escherichia coli* ATCC 25922 and *Pseudomonas aeruginosa* ATCC 27853 strains were used as quality control strains.

2.6. Colorimetric estimation of 3-indole acetic acid production

Three ml of tryptic soy broth (TSB), Biorad solution 1:10, supplemented with 1 mg/ml L-tryptophan, were inoculated with 200 μl of *Rheinheimera* sp. EpRS3 liquid culture. According to the assay described in [20], after incubation overnight at 30 °C, 50 μl of $FeCl_3$-$HClO_4$ reagent were added to 50 μl of medium. Absorbance was measured after 30 min at 530 mμ.

2.7. Plant growth conditions

Nicotiana tabacum cv. Xanthi seeds were surface-sterilized for 20 min in 5% NaOCl solution, followed by three washes with sterile distilled water. Sterilized seeds were then transferred onto Petri dishes containing Lindsmaier and Skoog medium (LS, Lab Associates BV, the Netherlands) supplemented with 1% sucrose and 0.7% plant agar. Plates were incubated in a growth chamber at 24 ± 1 °C, 80% humidity and in the dark until root differentiation. Seeds with primary roots were then placed in the light at 1500 lux and with a 16 h light/8 dark photoperiod regimen. Ten days from germination, *N. tabacum* seedlings were used for a vertical agar plate (VAP) assay [11].

2.8. Dual culture vertical agar plate assay

The endophytic *Rheinheimera* sp. EpRS3 strain and *E. coli* DH5α (used as control) were removed from −80 °C storage, streaked onto plates containing TSA and incubated at 30 °C for 48 h. Single colonies were then inoculated into tryptic soy broth (TSB) and grown in a shaker at the same temperature overnight. The bacterial suspensions were then adjusted to 10^8 cfu/ml ($OD_{600} = 1$). Twenty *N. tabacum* seedlings of the same age and dimension were grown on 15 cm Petri dishes containing LS basal medium; 100 μl of each of the OD_{600} suspension cultures or 100 μl culture filtrates were inoculated onto a sterilized filter paper disc placed 1 cm below the root tips of the seedlings, approximately at the center of the line of plants. Control treatments were made with 100 μl of TSB culture medium. Plates were incubated vertically in the growth chamber at 24 ± 1 °C and scored for root growth and morphology after 7 and 14 days from treatments. Root growth was reported as root length fold increase (mm) measured as (fl−il)/il where fl was the length of primary root after 7 or 14 days of culture and il was the initial length of primary roots. Each experiment was performed in duplicate.

2.9. Statistical analysis

All statistical analyses were performed using the PAST program, version 3.12 [12]. The analysis of variance between treated and untreated *N. tabacum* plants was conducted using One-way ANOVA ($P < 0.05$). Mean separations were performed using the method of Tukey. Student's *t* test was used for comparing two averages.

2.10. DNA isolation and whole genome sequencing

The *Rheinheimera* sp. EpRS3 strain was grown at 30 °C on TSB medium; genomic DNA was extracted using the CTAB method [13]. Authenticity of genomic DNA was confirmed by 16S rRNA gene sequencing. Whole genome shotgun sequencing was performed with a 2 × 300 bp paired-end approach using the MiSeq sequencing system (Illumina Inc., San Diego, CA), which produced a total of 1,245,634 high-quality reads.

2.11. Phylogenetic affiliation

Eighteen 16S rRNA gene sequences belonging to *Rheinheimera* species (representing all the species available for this genus) were selected from the Ribosomal DataBase Project (RDP) (http://rdp.cme.msu.edu/) [14] and from NCBI ftp site. Moreover, the 16S rRNA gene sequence from *Gallaecimonas pentaromativorans* CEE 131 was included as an outgroup in

phylogenetic analysis. The program Muscle v3.8.31 [15] was used to align 16S rRNA coding sequences; poorly aligned positions and divergent regions were deleted using Gblocks software v.0.91b [16]. The phylogenetic tree was inferred by using PhyML software v.3.0 [17], which computes genetic distances using a maximum likelihood method, with a GTR substitution model, 4 substitution rate categories and an estimated gamma shape parameter.

2.12. Genome assembly and annotation

FastQC software package v. 0.52 [18] was used to evaluate the quality of the obtained read pairs; poor quality bases were then removed using StreamingTrim [19]. Assembly was performed using SPAdes 3.5 software [20] with a k-mer length of 21, 33 and 55. The resulting contigs were annotated using the NCBI Automated Genome Annotation Pipeline.

The whole-genome shotgun project has been deposited at NCBI WGS database under accession number LNQS00000000; the version reported in this work was named LNQS01000000.

2.13. Comparative genomics

All *Rheinheimera* representative genomes sequenced thus far (i.e. *Rheinheimera* A13L, *Rheinheimera nanhaiensis* E407-8, *Rheinheimera perlucida* DSM 18276, *Rheinheimera baltica* DSM 14885, *Rheinheimera texasensis* DSM 17496, *R.* sp. KL1, *Rheinheimera* sp. F8, *Rheinheimera* sp. IITR-13) were collected from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/assembly/) and, alongside the *Rheinheimera* sp. EpRS3 genome, were analyzed using the dgenome module of the Ductape suite [21], to identify shared orthologous and strain-specific genes. Furthermore, the COG database [22] was consulted to classify them.

2.14. Secondary metabolite search parameters

The stand-alone version of antiSMASH software [23] for genome-wide identification, annotation and analysis of secondary metabolite biosynthetic gene clusters was used to scan the *Rheinheimera* sp. EpRS3 genome sequence. The homology relationship between clusters of the same families were inferred via First Best Hit (FBH) BLAST analysis [24] using a threshold e-value of $1e^{-20}$.

Table 2
Rheinheimera sp. EpRS3 genome features.

Attribute	Value
Genome size (bp)	4,396,207
DNA G+C (%)	49.3
DNA scaffolds	14
Total genes	3961
Protein coding genes	3868
RNA genes	62
Pseudo genes	31
ncRNAs	1
Genome coverage	82.0×

3. Results and discussion

3.1. The genome of Rheinheimera sp. EpRS3

After assembling the reads, a *Rheinheimera* sp. EpRS3 draft genome embedding 14 contigs with an overall length of 4,396,207 bp was obtained. The EpRS3 genome was annotated to identify the functional elements and their putative biological role, revealing the presence of 3961 genes, 3868 of which are protein-coding genes, 62 are RNA encoding genes, one is an ncRNA (non-coding RNA), and 30 are pseudogenes. The main features of the *Rheinheimera* sp. EpRS3 genome are reported in Table 2.

The phylogenetic tree (Fig. 1) inferred from the 16S rRNA gene sequence of the strain shows strain *Rheinheimera* sp. EpRS3 within the other members of the genus *Rheinheimera*. In particular, the *Rheinheimera* sp. EpRS3 strain proved to be close to *R. perlucida* DSM 18276 (99% of sequences identity).

Comparative genomics analysis was then performed between *Rheinheimera* sp. EpRS3 and *Rheinheimera* genomes available in NCBI databases (i.e. *Rheinheimera* A13L, *R. nanhaiensis* E407-8, *R. perlucida* DSM 18276, *R. baltica* DSM 14885, *R. texasensis* DSM 17496, *R.* sp. KL1, *Rheinheimera* sp. F8, *Rheinheimera* sp. IITR-13). Fig. 2 shows the size of unique, accessory and core genomes possessed by these organisms as a framework to estimate the genomic diversity of the dataset. The core genome of the genus *Rheinheimera* proved to be relatively small (13.8%, 1413 genes), while the ensemble of unique and accessory genome fraction of *Rheinheimera* was large, 51.2% (5237 genes) and 35.0% (3587 genes), respectively. Such large dispensable genomes could be related to the different ecological niches colonized by members of this genus, including water, soil and plant tissues [24–26], as seen for other bacterial genera [27]. Additionally, as shown in Fig. 2, the pangenome is open, a feature very common when bacteria belonging to the same taxonomic group live with many partners in very diverse environments, since gene gain events are common and lead to a heterogeneous gene pool among the members of the same genus.

In order to obtain functional characterization of the *Rheinheimera* genus, the open-reading frames (ORFs) were mapped to the COG [22] database; the resulting categories assigned to each gene are reported in Table 3. The abundance of core, accessory and unique genes was plotted and is compared in Fig. 3. As expected, the core genome displayed predominance in COG categories generally associated with housekeeping functions, such as categories C (energy production and conversion, 719 genes), E (amino acid transport and metabolism, 1119 genes), T (signal transduction mechanisms, 777 genes) and J (translation ribosomal structure and biogenesis, 856 genes) (Table 3). However, the core genome also contains a large number of genes with unknown function. Conversely, the accessory genome was enriched for the COG categories M (cell wall/membrane/envelope biogenesis, 960 genes), T (signal transduction mechanisms, 871 genes) and K (transcription, 897 genes) (Table 3). However, the largest part

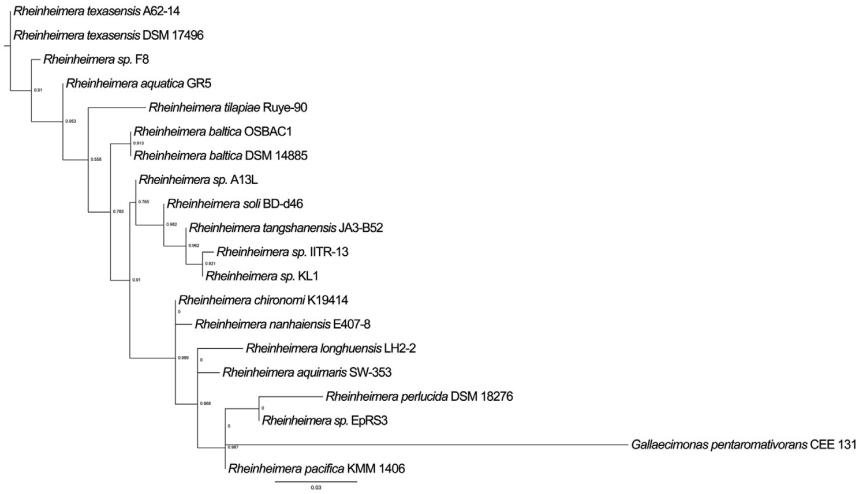


Fig. 1. Phylogenetic tree of the genus *Rheinheimera* based on 16S RNA sequences.

of both the accessory and unique genome was constituted by genes without a characterized functional class (X).

3.2. Antibiotics resistance

In a previous work [4], *Rheinheimera* sp. EpRS3 was demonstrated to be unable to grow in the presence of tetracycline (MIC 0.5 mg/l), rifampicin (MIC 5 mg/l), chloramphenicol (MIC 1 mg/l) or ciprofloxacin (MIC 0.5 mg/l). On the other hand, it showed resistance to some aminoglycoside antibiotics, including streptomycin and kanamycin at concentrations $\geq$10 mg/l and 5 mg/l, respectively. In an effort to identify the genetic determinants responsible for such resistance profiles, the genome of *Rheinheimera* sp. EpRS3 was annotated using the Resistance Gene Identifier (RGI) [28] tool of the Comprehensive Antibiotic Resistance Database (CARD). This led to identification of several elements usually associated with antibiotic resistance (original output obtained from CARD database inquiry are reported in Supplementary materials). The majority of these genes encode only minor components of efflux pump systems (like AcrB (belonging to the AcrAB/TolC system), MexI (belonging to MexGHI-OpmD complex), or MdtB and MdtC (that forms a heteromultimer complex, a subunit of MdtABC-TolC efflux pump)), but others encode complete RND-types systems (such as the complete set of genes encoding for the multidrug transporter MexEF-OprN complex and its transcriptional regulator *mexT*, an LysR-type transcriptional activator that positively regulates expression). Some of these efflux systems might be involved in resistance to kanamycin and streptomycin [29], which has been observed in this strain [5]. Additionally, such resistance could be due to another genetic element found in this investigation, i.e. a gene homologous to *acrD*, which is known to encode an aminoglycoside efflux pump in *E. coli* [30].

Also, inspection of the *Rheinheimera* sp. EpRS3 genome through the CARD database led to identification of unexpected genetic features that could potentially confer resistance to other drugs. For example, we found a gene coding for the Cfr 23S ribosomal RNA methyltransferase, which catalyzes methylation of the 23S rRNA subunit at A2503, suggesting its potential involvement in resistance to many drugs that target protein synthesis [36]. Moreover, the *Rheinheimera* sp. EpRS3 genome harbors genes coding for FosC2 and for MurA, two enzymes that usually confer resistance to fosfomycin [31,32,39]. Given the latter hint, we investigated the capacity of *Rheinheimera* sp. EpRS3 to grow in the presence of fosfomycin by MIC testing (as described in Section 2). Data obtained revealed that *Rheinheimera* sp. EpRS3 was able to grow up to 16 μg/ml, suggesting that the FosC2 and the MurA gene products could be involved in poor susceptibility to fosfomycin. Interestingly, such complex resistance profiles, joined with its ability to inhibit other strains inhabiting host-tissues [4], highlights the importance of *Rheinheimera* sp. EpRS3 from an ecological perspective.

3.3. Genome mining and inhibition of human pathogen strains

It has been previously reported that the *Rheinheimera* sp. EpRS3 strain is able to produce antibacterial compounds capable of inhibiting growth of many other bacteria isolated from the same plant [5] (although the molecules responsible for such inhibition have not yet been identified). Therefore, to gain insight into the genetic basis of the observed antimicrobial activity spectrum, the genome of *Rheinheimera* sp. EpRS3 was analyzed using the antiSMASH suite, allowing for identification and annotation of genes encoding secondary metabolites. As reported in Table 5, this approach revealed that the *Rheinheimera* sp. EpRS3 genome harbors a considerable number of genes with putative inhibitory activity. Indeed, we found a total of 111 genes split in 8 clusters, associated with synthesis of different molecules exhibiting antimicrobial

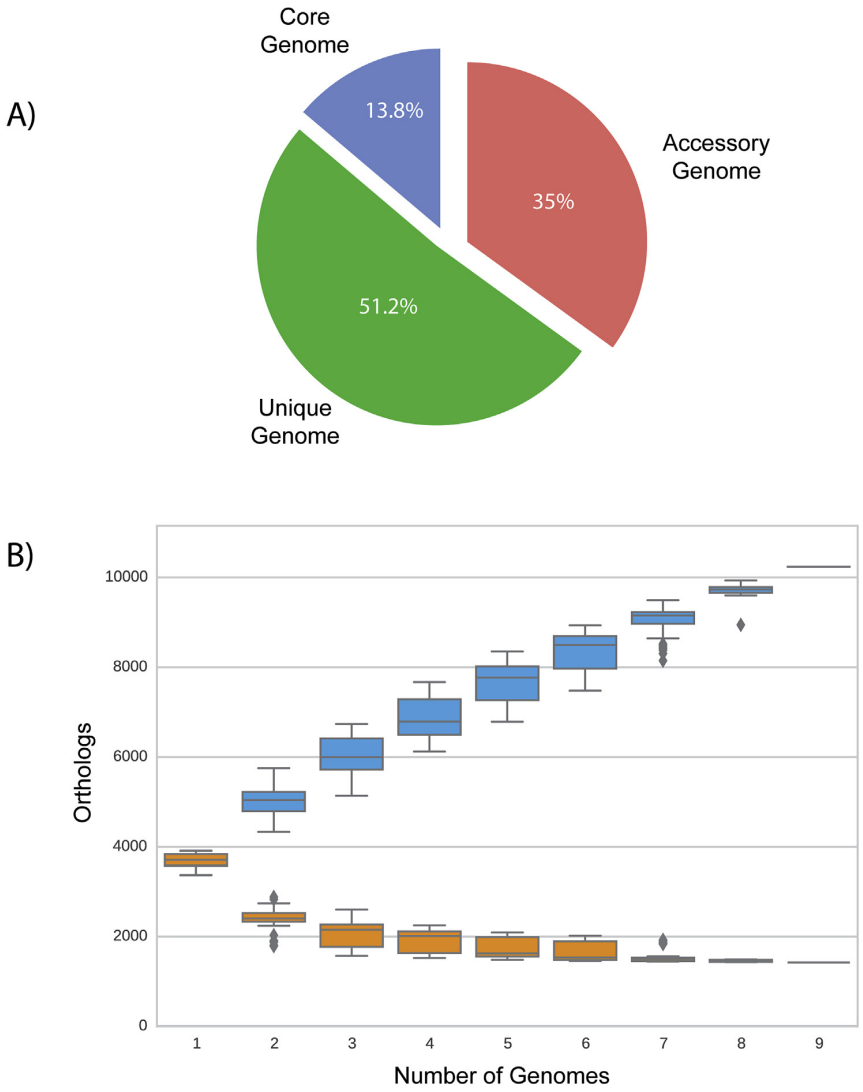


Fig. 2. The *Rheinheimera* genus PanGenome. 1 A: Pangenome shape. B: number of orthologs genes vs. number of genomes orange and blue represent core genome genes and pangenome size, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

activity. In particular, we found two different putative bacteriocin clusters (a class of extracellular small peptides exhibiting bactericidal activity) with a comparable number of genes (6 and 7). Two clusters were found to be associated with the synthesis of lantipeptides (harboring 11 and 12 genes each), a class of polycyclic peptides characterized by the presence of the thioether-cross-linked amino acids *meso*-lanthionine (Lan) and (2*S*, 3*S*, 6*R*)-3-methyllanthionine (MeLan; [33]). The latter three biosynthetic clusters are involved in production of different compound classes (resorcinol, non-ribosomal peptides, polyketides). In particular, the presence of polyketide synthases (PKSs) and non-ribosomal peptide synthases (NRPSs) have a topical relevance for future biotechnological exploitation of this strain, in that some of the most successful antibiotics (i.e. vancomycin, avermectin, erythromycin) are produced by these enzymes.

The finding that the *Rheinheimera* sp. EpRS3 strain was able to inhibit growth of other bacteria, especially Gram-negative ones, stimulated us to evaluate its inhibitory effects against human pathogens. Hence, we tested *Rheinheimera* sp. EpRS3 through the cross-streaking method against a panel of 30 (target) strains belonging to the *B. cepacia* complex (Bcc), a group of opportunistic pathogens able to cause severe infections in immune-compromised subjects, such as cystic fibrosis (CF) patients [34]. As shown in Table 1, Bcc strains were of different origin (environmental, animal infection or clinical). Data from cross-streaking experiments are shown in Fig. 4A and Table 4. As reported, the assay revealed that *Rheinheimera* sp. EpRS3 is able to interfere with growth of most target Bcc strains. In particular, we observed that it is strongly effective against all tested *Burkholderia multivorans* and *Burkholderia cenocepacia* strains. Moreover, since it has

Table 3
Distribution of *Rheinheimera* sp EpRS3 genes in COG functional category.

COG functional category	Description	No. of genes	%
J	Translation ribosomal structure and biogenesis	175	4.52
A	RNA processing and modification	1	0.03
K	Transcription	203	5.24
L	Replication recombination and repair	133	3.43
B	Chromatin structure and dynamics	3	0.08
D	Cell cycle control cell division chromosome partitioning	33	0.85
Y	Nuclear structure	0	0
V	Defense mechanisms	69	1.78
T	Signal transduction mechanisms	206	5.32
M	Cell wall/membrane/envelope biogenesis	209	5.39
N	Cell motility	134	3.46
Z	Cytoskeleton	0	0
W	Extracellular structures	0	0
U	Intracellular trafficking secretion and vesicular transport	43	1.11
O	Posttranslational modification protein turnover chaperones	138	3.56
C	Energy production and conversion	165	4.26
G	Carbohydrate transport and metabolism	163	4.21
E	Amino acid transport and metabolism	222	5.73
F	Nucleotide transport and metabolism	62	1.6
H	Coenzyme transport and metabolism	114	2.94
I	Lipid transport and metabolism	97	2.5
P	Inorganic ion transport and metabolism	157	4.05
Q	Secondary metabolites biosynthesis transport and catabolism	45	1.16
R	General function prediction only	335	8.65
S	Function unknown	296	7.64
X	No functional class found	865	22.5

been previously demonstrated that growth of Bcc strains is inhibited by microbial volatile organic compounds (mVOCs) synthesized by Antarctic bacteria [8], we performed the same cross-streaking experiments on Petri dishes with a central septum separating the tester strain from the target strains. Data obtained revealed that *Rheinheimera* sp. EpRS3 is still able to inhibit Bcc strains, though to a lesser extent. The latter finding suggests that at least some of the molecules exhibiting antimicrobial activity vs. Bcc strains might be volatile organic compounds (VOCs). Thus, it is quite possible that the inhibitory activity exhibited by the *Rheinheimera* sp. EpRS3 strain relies on a combination of both volatile and diffusible compounds, reminiscent of the inhibitory activity that many Antarctic bacteria possess vs. Bcc strains [8]. Chemical profiling is currently ongoing to characterize such molecules.

In order to check whether the antimicrobial compounds synthesized by *Rheinheimera* sp. EpRS3 might also be effective against other human clinically relevant pathogens, we performed cross-streaking experiments using a panel of MDR strains belonging to *A. baumannii* and *K. pneumoniae* species as targets (Table 1). Data obtained are shown in Fig. 4 and demonstrate that *Rheinheimera* sp. EpRS3 completely inhibits *A. baumanni* N50 and *A. baumannii* YMCR363 strains (Table 1, Fig. 4B) and partially inhibits *K. pneumoniae* KPC284, *K. pneumoniae* B04, *K. pneumoniae* B2 (Table 1, Fig. 4C). Interestingly, some of these strains are resistant to a specific class of antibiotics, i.e. polymyxins, that are encoded by NRPSs. More in detail, in *K. pneumoniae* B04 and B2 strains, such resistance depends on modifications of the lipopolysaccharide (LPS), while in *K. pneumoniae* KPC284 and *A. baumanni* N50, the precise mechanism responsible for such a phenotype is currently unknown. Notably, colistin resistance in *A. baumannii* may occur due to mutations in the PmrAB two-component system or in *lpxA*, *lpxC*, and *lpxD* genes (as in *K. pneumoniae* strains); the two systems, respectively, lead to the modification and loss of LPS (as in *K. pneumoniae* strains) [35]. Thus, we can argue that *Rheinheimera* sp. EpRS3 is probably strongly effective against the colistin-resistant organisms tested since, due to their (known or putative) LPS-associated differences, they may be the only ones in which the antimicrobial molecules produced by the tester are able to enter into the target cells. We are completely aware that, at this stage, this scenario represents a hypothesis that will require more in-depth investigation. Nevertheless, it could represent an important resource in the treatment of infections sustained by MDR pathogens, particularly those resistant to last-line treatment with colistin.

3.4. *Identification of genetic elements in* Rheinheimera *sp. EpRS3 associated with its ecological role in rhizospheric soil*

Different bacterial genera are involved in a plethora of biotic activities of the soil ecosystem, such as biocontrol of microbial communities and/or control of nutrient dynamic turn-over [39],[40]. Several mechanisms come into play when dealing with plant growth promotion, both directly and indirectly. Among the most well-known, we can cite indole acetic acid (IAA) production (that increases the root surface) [36], nodulation and nitrogen-fixing ability, siderophore production

Table 4
Inhibition level of the *Rheinheimera* sp RS3 (tester) versus a panel of 30 Bcc members (target), 6 A. baumannii and 11 K. pneumoniae. The numbers indicate the different inhibition level from 0 to 3 as: complete (3), strong (2), weak (1), and absent (0). Abbreviations: E, Environmental; A, Animal Infection; CF, Cystic Fibrosis, C, Clinical. Plates with and without septum.

Target			Tester strain	Inhibition level	
Position on plate	Species	Strain		Without Septum	With septum
1	*B.ambifaria*	LMG 19182		2	2
2	*B. anthina*	LMG 20980		2	2
3	*B. arboris*	LMG 24066		1	0
4	*B. cenocepacia*	LMG 16656		3	3
5	*B. cepacia*	LMG 1222		0	0
6	*B. contaminans*	LMG 23361		1	1
7	*B. diffusa*	LMG 24065		1	0
8	*B. dolosa*	LMG 18943		1	1
9	*B. lata*	LMG 22485		1	1
10	*B. latens*	LMG 24064		2	3
11	*B. metallica*	LMG 24068		0	0
12	*B. multivorans*	LMG 13010		1	1
13	*B. pseudomultivorans*	LMG 26883		0	0
14	*B. pyrrocinia*	LMG 14191		0	0
15	*B. seminalis*	LMG 24067	*Rheinheimera* sp. EpRS3	0	0
16	*B. stabilis*	LMG 14294		1	1
17	*B. ubonensis*	LMG 20358		0	0
18	*B. vietnamiensis*	LMG 10929		1	1
19	*B. cenocepacia*	FCF 12		3	3
20	*B. cenocepacia*	FCF 13		3	3
21	*B. cenocepacia*	FCF 14		3	3
22	*B. cenocepacia*	FCF 15		3	3
23	*B. cenocepacia*	J2315		3	2
24	*B. cenocepacia*	FCF 18		3	2
25	*B. cenocepacia*	FCF 19		3	2
26	*B. multivorans*	FCF 5		3	1
27	*B. multivorans*	FCF 6		3	1
28	*B. multivorans*	FCF 7		3	0
29	*B. multivorans*	FCF 8		3	0
30	*B. multivorans*	FCF 9		3	0
1	*A. baumannii*	RHU 134	*Rheinheimera* sp. EpRS3	0	–
2	*A. baumannii*	RHU 875		0	–
3	*A. baumannii*	ATCC17978		0	–
4	*A. baumannii*	O8C29		0	–
5	*A. baumannii*	N5O		3	–
6	*A. baumannii*	YMCR363		3	–
1	*K. pneumoniae*	BO 1	*Rheinheimera* sp. EpRS3	0	–
2	*K. pneumoniae*	BO 4		1	–
3	*K. pneumoniae*	B1		0	–
4	*K. pneumoniae*	B2		1	–
5	*K. pneumoniae*	ATCC 700603		0	–
6	*K. pneumoniae*	KP397		0	–
7	*K. pneumoniae*	KP477		0	–
8	*K. pneumoniae*	KPC249		0	–
9	*K. Pneumoniae*	KPC261		0	–
10	*K. pneumoniae*	KPC284		1	–
11	*K. pneumoniae*	NTUHK2044		0	–

and 1-aminocyclopropane-1-carboxylate (ACC) deaminase activity. Therefore, the *Rheinheimera* sp. EpRS3 genome was investigated for the presence of genes involved in such metabolic abilities. Nodulation-related genes (*nod)* and several siderophore encoding genes have been detected. Also, data obtained revealed the absence of genetic elements related to ACC deaminase biosynthesis; moreover, the *Rheinheimera* sp. EpRS3 genome did not encounter the in silico criteria necessary to be considered a diazotroph (i.e. the co-presence of *nifHDKENB* genes in its genetic makeup [37]). On the other hand, we detected genetic traits responsible for production of enzymes involved in IAA biosynthesis (i.e. indole-3-acetaldehyde dehydrogenase (IpyA/TAM), indole-3-pyruvate decarboxylase (IpyA), tryptophan monooxydase (IAM),

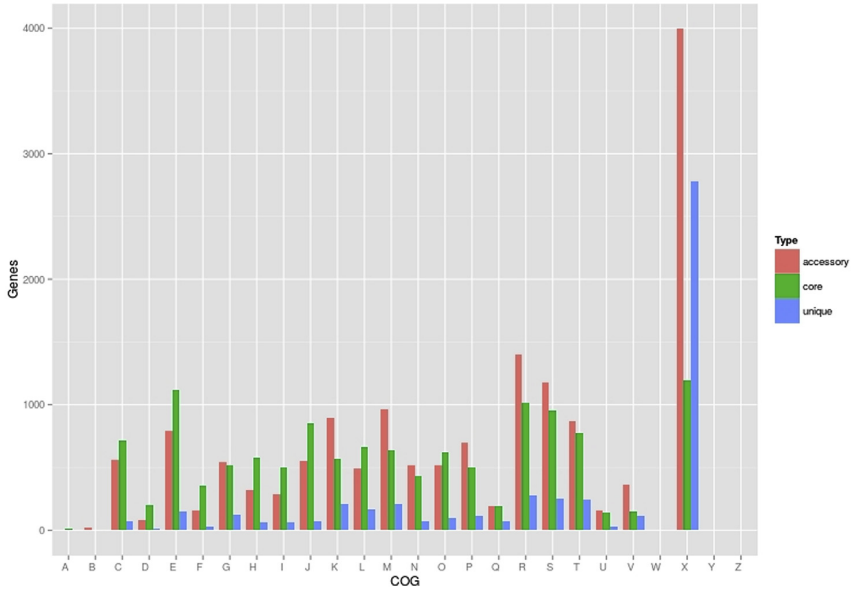


Fig. 3. Distribution of *Rheinheimera* genus, core, 1 accessory and unique gene plot according to COG functional category. Abbreviations: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control cell division chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, co-enzyme transport and metabolism; I, lipid transport and metabolism; J, translation ribosomal structure and biogenesis; K, transcription; L, replication recombination and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, post-translational modification protein turnover chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking secretion and vesicular transport; V, defense mechanisms; W, extracellular structures; X, no functional class found; Y, nuclear structure; Z, cytoskeleton.

tryptophanase). Colorimetric assay to estimate production of IAA was performed as described in Section 2 to confirm in silico data; tests gave positive results ($OD_{530} = 0.225$). On this track, we decided to investigate whether IAA production by the *Rheinheimera* sp. EpRS3 strain could affect plant growth, as already reported for several endophytic strains in different plant systems [36,38], or whether, inside the plant rhizosphere, such ability is not expressed. Hence, we monitored elongation of primary roots of a test plant, namely *N. tabacum* cv. Xanthi, after inoculation with *Rheinheimera* sp. EpRS3 cells. As described in Section 2, we used the dual culture vertical agar plate approach previously used in other systems by other authors [38]. Results on the effect of different concentrations of indole acetic acid (IAA) on inhibition of the length of primary roots in *N. tabacum* seedlings after 7 days of treatment are shown in Supplementary Material SM1. Data obtained revealed that different concentrations of IAA induce a significant decrease ($p < 0.001$), compared to the control in primary root length of *N. tabacum* seedlings in a dose-dependent manner. In particular, inhibition of primary root length seemed to be strongly affected by the distance of the inoculated paper disc from the seedlings. In fact, seedlings grown at a distance of more than 2 cm (>2 cm) from the paper disc showed a lower decrease in root elongation in contrast to seedlings placed within 2 cm (<2 cm) of the disc. In detail, statistical analysis showed p values, respectively, of <0.01 for treatment of plants with 200 μM IAA and $p < 0.05$ for plants subjected to 20 μM IAA (SM1). The same trend was observed after 14 days of treatment, in this case differences between the two groups of plants (>2 cm or <2 cm) were not significant (data not shown). The effect of bacterial inoculations of tobacco seedlings with either *E. coli*, both DH5α (known IAA overproducer used as control) or *Rheinheimera* sp. EpRS3 strains, was to inhibit primary root elongation, with *E. coli* DH5α having a higher activity than *Rheinheimera* sp. EpRS3 (see Supplementary Material SM2, SM3). In particular, the inoculation with DH5α cells or culture filtrate registered a significant (respectively $p < 0.01$ and $p < 0.001$) decrease of root elongation compared to the control. The same was true only in the case of culture filtrate ($p < 0.01$) of *Rheinheimera* sp. EpRS3 strain. For both the strains we have not observed

Table 5
Gene clusters involved in secondary metabolites biosynthesis in the *Rheinheimera* sp. RS3 genome; the type of secondary metabolite produced and the number of genes embedded in the clusters are reported.

Biosynthetic cluster	Type	No. of genes
Cluster 1	Resorcinol	22
Cluster 2	Lantipeptide	12
Cluster 3	Lantipeptide	11
Cluster 4	Hserlactone	9
Cluster 5	Bacteriocin	6
Cluster 6	Nrps	7
Cluster 7	Hserlactone	11
Cluster 8	Bacteriocin	7

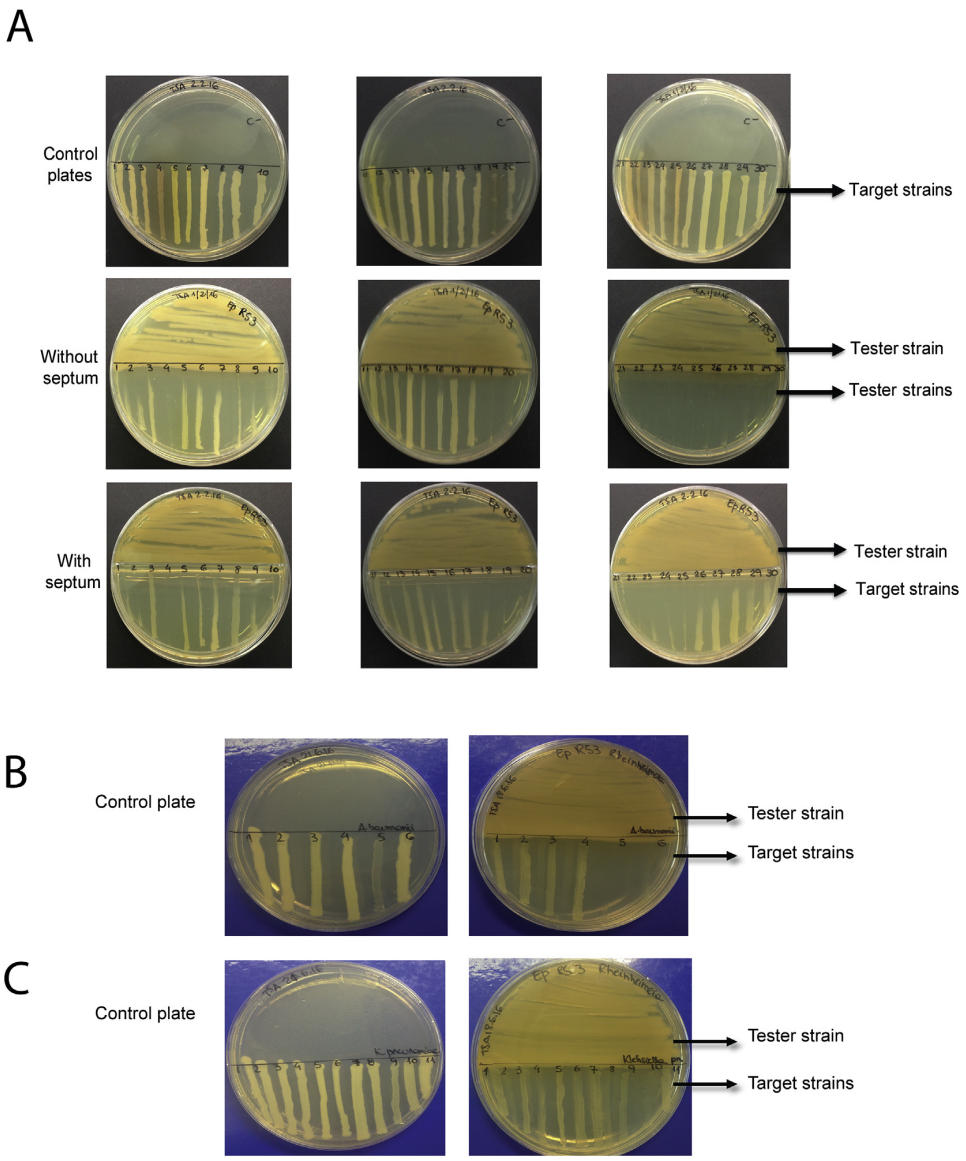


Fig. 4. Petri dishes of cross-streaking experiments against: A) BCC strains (numbers refer to different organisms from 1 to 10, 11–20, 21–30); B) *A. baumannii* strains (numbers refer to different organisms from 1 to 6); C) *K. pneumoniae* strains (numbers refer to different organisms from 1 to 11).

any significant difference between groups of plants classified as >2 cm or <2 cm (Fig. SM2). Moreover, the treatments of plants with bacterial suspensions induced changes of primary root morphology concerning either the increase in root hairs formations and lateral root induction compared to the negative control (SM3).

Results concerning both the inhibition of the primary root elongation and morphological changes of roots of tobacco seedlings suggested a possible relationship between the IAA production by *Rheinheimera* sp. EpRS3 strain and these growth modifications. Experiments are currently ongoing to demonstrate a possible role of the strain as plant growth promoting bacteria (PGPB).

3.5. *Other phenotypic traits of* Rheinheimera *sp. EpRS3*

Since the ability of some soil bacteria to degrade organic contaminants is promising for the development of bio-treatment systems to counteract soil pollution, we analyzed whether *Rheinheimera* sp. EpRS3 is able to degrade complex hydrocarbon molecules or to resist to oxidative stresses, like those occurring in presence of heavy metals. To this purpose the *Rheinheimera* sp. EpRS3 was tested for its ability to grow on two different minimal media (MMV and MMD), differing in salt concentration, supplemented with either glucose or diesel fuel as the sole carbon source. Data obtained revealed that *Rheinheimera* sp. EpRS3 was not able to grow on diesel

fuel as carbon source. The only condition supporting growth was MMD supplemented with glucose as the sole carbon and energy source, suggesting the inability to degrade hydrocarbon compounds. The absence of genes involved *n*-alkanes degradation (i.e. *alk* genes) in *Rheinheimera* sp. EpRS3 genome was in agreement with the experimental data.

Moreover, since plant root tissues can produce reactive oxygen species (ROS) in response to different *stimuli*, we tested the resistance of the strain to oxidative stress by using hydrogen peroxide or a heavy metal (i.e. zinc, copper, cadmium and nickel), revealing that *Rheinheimera* sp. EpRS3 is sensitive to zinc, cadmium, nickel and hydrogen peroxide, while it exhibits a tolerance to low concentrations of $CuCl_2$ (0.25, 0.5 and 1 mM). The genome analysis results, however, were in disagreement with this conclusion, in that we found a number of genes (*katA*, *B*, *C* and *sodB*) involved in the response to oxidative stress. Indeed, the disagreement between experiments and genomic analyses is puzzling and prompts for further investigation. A possible explanation is that heavy metals may somehow impair the proper expression of these genes or the functionality of their products. On the other hand, we have no data regarding the expression of these genes, which might even be not functional.

Lastly, we performed experimental assays to test whether EpRS3 is able to secrete enzymes into the extracellular space in order to hydrolyze macromolecules into constituents that can be imported for microbial nutrition. Data obtained suggest that the *Rheinheimera* sp. EpRS3 strain possesses extracellular lipases, phospholipase and proteolytic activities, while no amylolytic activity towards starch has been detected (data not shown).

In conclusion, the aim of this work was to characterize newly isolated rhizospheric strain *Rheinheimera* sp. EpRS3, isolated from rhizospheric soil of the medicinal plant *E. purpurea*, to test its ability to influence plant growth and to synthesize antimicrobial compounds that might be effective against bacterial human pathogens. For this purpose, a set of phenotypic parameters was tested and data obtained were correlated with the gene repertoire of the *Rheinheimera* sp. EpRS3 genome. Cross-streaking experiments revealed that *Rheinheimera* sp. EpRS3 is able to inhibit growth of different bacterial pathogens (Bcc, *A. baumanni*, and *K. pneumoniae*), most of which exhibit a multi-drug-resistance phenotype. This ability was very likely due to the presence of gene clusters involved in different types of secondary metabolites. However, it cannot be excluded that some of these molecule are VOCs, as shown by cross-streaking experiments performed with Petri dishes harboring a central septum. It is particularly intriguing that the *Rheinheimera* sp. EpRS3 strain synthesize molecules able to "bypass" the resistance of *K. pneumoniae* and *A. baumanni* strains to colistin.

Moreover, in silico analysis indicated the presence in the genome of various components involved in multidrug-efflux transporter and cation efflux system biosynthesis, alongside traits usually responsible for specific resistance.

These data underline how endophytic and/or rhizospheric bacteria may play an important ecological role inside the host plant; it is possible that both antibiotic resistance and synthesis of antimicrobial compounds might help in shaping the structure of bacterial communities living near or inside the plant, as recently suggested [4].

If this is so, the possibility that such bacterial strains might also influence the growth of the host plant should be taken into account. Data obtained in this work (i.e. the presence of a gene encoding triptophanase, involved in IAA production, as well as results from IAA production and effects on plant growth) are in agreement with this idea. Lastly, it is worth mentioning that these bacteria may directly synthesize or influence host synthesis of some bioactive molecules identified in extracts/essential oils of medicinal plants, as recently suggested by a work performed on rhizospheric and endophytic bacteria isolated from the medicinal plant *Lavandula angustifolia* [39].

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was financially supported by the Ente Cassa di Risparmio di Firenze Project 2013.0657 "Herbiome: nuove molecole antibiotiche da batteri endofiti di piante medicinali".

We are very grateful to two anonymous referees for their helpful suggestions in improving the manuscript. Pei-Fang Hsieh and Jin-Town Wang are gratefully acknowledged to have provided G.M.R with the NTUH-K2044 strain.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.resmic.2016.11.001.

References

[1] Danhorn T, Fuqua C. Biofilm formation by plant-associated bacteria. Annu Rev Microbiol 2007;61:401–22. http://dx.doi.org/10.1146/annurev.micro.61.080706.093316.

[2] Taghavi S, van der Lelie D. Genome sequence of the plant growth-promoting endophytic bacterium *Enterobacter* sp. 638. Mol Microb Ecol Rhizosph 2013;2:899–908. http://dx.doi.org/10.1002/9781118297674.ch84.

[3] Maida I, Chiellini C, Mengoni A, Bosi E, Firenzuoli F, Fondi M, et al. Antagonistic interactions between endophytic cultivable bacterial communities isolated from the medicinal plant *E chinacea purpurea*. Environ Microbiol 2015;18:2357–65. http://dx.doi.org/10.1111/1462-2920.12911. n/a-n/a, Thematic Issue: Thematic Issue on Microbial Mutualism and Symbiosis.

[4] Mengoni A, Maida I, Chiellini C, Emiliani G, Mocali S, Fabiani A, et al. Antibiotic resistance differentiates *Echinacea purpurea* endophytic bacterial communities with respect to plant organs. Res Microbiol 2014;165:686–94. http://dx.doi.org/10.1016/j.resmic.2014.09.008.

[5] Chiellini C, Maida I, Emiliani G, Mengoni A, Stefano M, Fabiani A, et al. Endophytic and rhizospheric bacterial communities isolated from the medicinal plants *Echinacea purpurea* and *Echinacea angustifolia* n.d. http://revistes.iec.cat/index.php/IM/article/viewFile/136531/135273. [Accessed 17 March 17 2016].

[6] Hudson JB. Applications of the phytomedicine *Echinacea purpurea* (purple coneflower) in infectious diseases. J Biomed Biotechnol 2012; 2012. http://dx.doi.org/10.1155/2012/769896.
[7] Corry JEL. Handbook of biological media, vol. 22; 1994. http://dx.doi.org/10.1016/0168-1605(94)90011-6.
[8] Papaleo MC, Fondi M, Maida I, Perrin E, Lo Giudice A, Michaud L, et al. Sponge-associated microbial Antarctic communities exhibiting antimicrobial activity against *Burkholderia cepacia* complex bacteria. Biotechnol Adv 2012;30:272–93. http://dx.doi.org/10.1016/j.biotechadv.2011.06.011.
[9] Diancourt L, Passet V, Nemec A, Dijkshoorn L, Brisse S. The population structure of *Acinetobacter baumannii*: expanding multiresistant clones from an ancestral susceptible genetic pool. PLoS One 2010;5. http://dx.doi.org/10.1371/journal.pone.0010034.
[10] Clsi. Performance standards for antimicrobial susceptibility testing; Twenty-second informational supplement, vol. 32; 2013.
[11] Zhang H, Forde BG. An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. Science 1998;279:407–9. http://dx.doi.org/10.1126/science.279.5349.407.
[12] Hammer Ø, Harper DAT, Ryan PD. Paleontological statistics software package for education and data analysis. Palaeontol Electron 2001;4: 9–18. http://dx.doi.org/10.1016/j.bcp.2008.05.025.
[13] Perrin E, Fondi M, Maida I, Mengoni A, Chiellini C, Mocali S, et al. Genomes analysis and bacteria identification: the use of overlapping genes as molecular markers. J Microbiol Methods 2015;117:108–12. http://dx.doi.org/10.1016/j.mimet.2015.07.025.
[14] Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 2009;37. http://dx.doi.org/10.1093/nar/gkn879.
[15] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7. http://dx.doi.org/10.1093/nar/gkh340.
[16] Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 2000;17:540–52. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334.
[17] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59: 307–21. http://dx.doi.org/10.1093/sysbio/syq010.
[18] Kunde-Ramamoorthy G, Coarfa C, Laritsky E, Kessler NJ, Harris RA, Xu M, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res 2014; 42. http://dx.doi.org/10.1093/nar/gkt1325.
[19] Bacci G, Bazzicalupo M, Benedetti A, Mengoni A. StreamingTrim 1.0: a Java software for dynamic trimming of 16S rRNA sequence data from metagenetic studies. Mol Ecol Resour 2014;14:426–34. http://dx.doi.org/10.1111/1755-0998.12187.
[20] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77. http://dx.doi.org/10.1089/cmb.2012.0021.
[21] Galardini M, Mengoni A, Biondi EG, Semeraro R, Florio A, Bazzicalupo M, et al. DuctApe: a suite for the analysis and correlation of genomic and OmniLog™ phenotype microarray data. Genomics 2014; 103:1–10. http://dx.doi.org/10.1016/j.ygeno.2013.11.005.
[22] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28:33–6. http://dx.doi.org/10.1093/nar/28.1.33.
[23] Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 2011;39:W339–46. http://dx.doi.org/10.1093/nar/gkr466.
[24] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
[25] Brettar I, Christen R, Manfred GH. *Rheinheimera baltica* gen. nov., sp. nov., a bluecoloured bacterium isolated from the central Baltic Sea. Int J Syst Evol Microbiol 2002;52:1851–7. http://dx.doi.org/10.1099/ijs.0.02151-0.
[26] Zhang X, Sun L, Qiu F, McLean RJC, Jiang R, Song W. *Rheinheimera tangshanensis* sp. nov., a rice root-associated bacterium. Int J Syst Evol Microbiol 2008;58:2420–4. http://dx.doi.org/10.1099/ijs.0.65605-0.
[27] Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 2008;11:472–7. http://dx.doi.org/10.1016/j.mib.2008.09.006.
[28] McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 2013;57:3348–57. http://dx.doi.org/10.1128/AAC.00419-13.
[29] Moore RA, Deshazer D, Reckseidler S, Weissman A, Woods DE. Efflux-mediated aminoglycoside and macrolide resistance in *Burkholderia pseudomallei*. Antimicrob Agents Chemother 1999;43:465–70.
[30] Rosenberg EY, Ma D, Nikaido H. AcrD of Escherichia coli is an aminoglycoside efflux pump. J Bacteriol 2000;182:1754–6. http://dx.doi.org/10.1128/JB.182.6.1754-1756.2000.
[31] Lee SY, Park YJ, Yu JK, Jung S, Kim Y, Jeong SH, et al. Prevalence of acquired fosfomycin resistance among extended-spectrum??-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* clinical isolates in korea and IS26-composite transposon surrounding fosA3. J Antimicrob Chemother 2012;67:2843–7. http://dx.doi.org/10.1093/jac/dks319.
[32] Takahata S, Ida T, Hiraishi T, Sakakibara S, Maebashi K, Terada S, et al. Molecular mechanisms of fosfomycin resistance in clinical isolates of *Escherichia coli*. Int J Antimicrob Agents 2010;35:333–7. http://dx.doi.org/10.1016/j.ijantimicag.2009.11.011.
[33] Knerr PJ, van der Donk WA. Discovery, biosynthesis, and engineering of Lantipeptides. 2012.
[34] Döring G, Flume P, Heijerman H, Elborn JS. Treatment of lung infection in patients with cystic fibrosis: current and future strategies. J Cyst Fibros 2012;11:461–79. http://dx.doi.org/10.1016/j.jcf.2012.10.004.
[35] López-Rojas R, McConnell MJ, Jiménez-Mejías ME, Domínguez-Herrera J, Fernández-Cuenca F, Pachón J. Colistin resistance in a clinical *Acinetobacter baumannii* strain appearing after colistin treatment: effect on virulence and bacterial fitness. Antimicrob Agents Chemother 2013; 57:4587–9. http://dx.doi.org/10.1128/AAC.00543-13.
[36] Spaepen S, Vanderleyden J, Remans R. Indole-3-acetic acid in microbial and microorganism-plant signaling. FEMS Microbiol Rev 2007;31: 425–48. http://dx.doi.org/10.1111/j.1574-6976.2007.00072.x.
[37] Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 2012;13:162. http://dx.doi.org/10.1186/1471-2164-13-162.
[38] Shi C-L, Park H-B, Lee JS, Ryu S, Ryu C-M. Inhibition of primary roots and stimulation of lateral root development in *Arabidopsis thaliana* by the rhizobacterium Serratia marcescens 90–166 is through both auxin-dependent and -independent signaling pathways. Mol Cells 2010;29: 251–8. http://dx.doi.org/10.1007/s10059-010-0032-0.
[39] Emiliani G, Mengoni A, Maida I, Perrin E, Chiellini C, Fondi M, et al. Linking bacterial endophytic communities to essential oils: clues from *Lavandula angustifolia* mill. Evidence-Based Complement Altern Med 2014;2014. http://dx.doi.org/10.1155/2014/650905.

8 Metabolic modeling

The works presented herein offer practical examples about systems-fashion uses of the genomic data collected through the technology and works covered in the previous sections of this thesis. The reconstruction of metabolic models, based on genome annotation, phenotypic data and omics analysis integration, allowed for systems analysis of a human- and a plant-pathogenic strain, respectively *Acinetobacter baumannii* ATCC 19606 and *Pectobacterium parmentieri* SCC3193. Thanks to such *in silico* methods, functional data may be derived at the systems level allowing hypothesis drawing and tests.

A. baumannii is an opportunistic, multi-drug resistant, bacterial pathogen associated with a large number of deadly infections. The therapeutic solutions are very few and strains able to resist also to the last-line antibiotic treatment (colistin antibiotic) have been reported. New antimicrobial agents and/or strategies are urgently required to control *A. baumannii* infections and to eradicate antibiotic resistant strains from affected patients. However, the traditional approach to the study of this organism keeps failing these fundamental goals. To address new possible solutions, a system-level study of antibiotic-response in *A. baumannii* is required. In the work attached to this chapter, computational metabolic modeling offered a relatively quick and effective method to identify valuable and non-trivial potential drug targets in *A. baumannii*. After reconstructing the genome scale metabolic model of the type strain ATCC 19606, the system-level metabolic consequences of *A. baumannii* exposure to colistin were explored through integration of gene expression data in that condition [1]. The analysis revealed the metabolic reprogramming occurring in this strain following the establishment of the stressful condition and offered a glimpse on gene essentiality (EG). The last being pretty relevant. The identification of EGs, in fact, is one of the key-steps in a drug discovery pipeline (see Figure 8.1). In the case study, it was achieved by simulating genome-scale gene knock-out in multiple, specific conditions. At a first stage different, environmental niches were simulated by changing the set of available nutrients. Thereafter, increasingly demanding situations, like antibiotic exposure and mutations, were introduced in the model (using available expression data in similar conditions, for instance). The different *scenario* tested were useful to recognize both general and condition-specific EGs [2]. The formers are required to sustain life un-

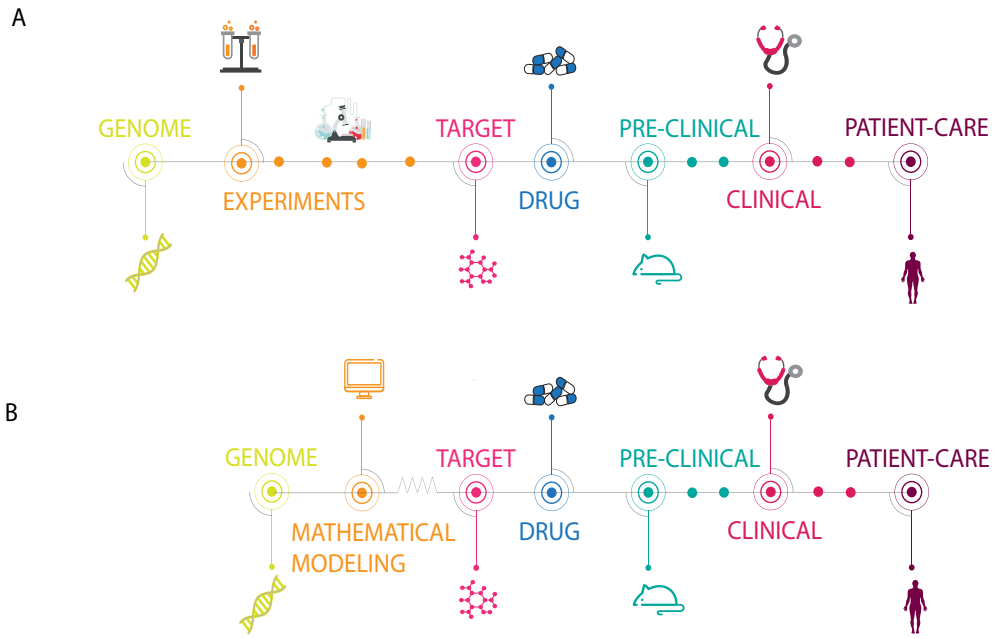


Fig. 8.1: **A:** The classic drug discovery pipeline. **B:** The use of mathematical modeling may abbreviate the identification of drug targets.

der virtually all growth bottom line; conversely, under specific constraints, changes of central metabolism may occur, leading either to a variation in flux distribution throughout the network and to noise in gene essentiality. For example, testing the system in two different media (one rich and one minimal), highlighted how nutrients availability affects cell metabolism and, interestingly, how different environmental pressures influence gene essentiality. Particularly, 67 genes were predicted to be indispensable under both the tested growth conditions, while 71 are prone to become like that only when limited nutrients availability force the cell to reprogram its metabolic behavior. Similar trends have been obtained mapping expression records on the network (*i.e.* up-regulation and down-regulation genes ratios) on a normal and on *LPS*$^-$ version of the model (deficiency of lipopolysaccharides leads to colistin resistance). Beside each singularity of gene expression patterns found, one of the significant outcome emerging from the work is that the genes designated to be EGs are likely to be conserved at the *A. baumannii* species level. The general trend observed was that more than 90% of the genomes scanned (*i.e.* all the 1099 *A. baumannii* genomes sequenced to date) possessed the searched queries (EGs sequences with identity treshold $> 30\%$). Also, the analysis shows that this tendency is kept almost unchanged even when imposing values of identity threshold up to 90%. Although these genes

may not be essential in other strains, this preliminary result is encouraging, since it expresses the possibility that the target genes indicated for *A. baumannii* ATCC 19606 are probably common targets in most of *A. baumannii* type infections. A second evidence to pinpoint is the one emerging from the comparison of data obtained through metabolic modeling approach and wet-lab experiments. This in fact, largely supports the strength of metabolic modeling performances. The results yielded by the study in fact, stand out for comprehensiveness, highly predictive power and quick response (in the order of seconds) while the classic expensive and laborious experimental gene knock-out assays lack wide coverage and, sometimes, reproducibility.

The second metabolic reconstruction features the assessment of metabolic adaptation in *P. parmentieri* SCC3193. The strain causes severe crop diseases worldwide and has been included among the 10 most important bacterial plant pathogens based on economic impact [3]. To develop new preventive strategies against the spread of plant infections and to perform sustainable agriculture, an understanding of host-microbe interactions is crucial. This interplay is connected with the complexity of the environment in which bacteria persist: soil, rhizosphere or plant tissues. Constraint-based metabolic modeling provides an engaging way to investigate the adaptation to different ecological niches and may give insights into the metabolic versatility of plant pathogenic bacteria. Given that, a metabolic model of the strain was first reconstructed and extensively manually curated. High-throughput Phenotype Microarrays data were used to asses model's performances and accuracy. After, by means of flux balance analysis, the metabolic adjustment to soil and rhizosphere were predicted. These two different ecological niches were chosen as they are relevant for both the bacterium persistence and the plant colonization.

The simulation of environmental switch modified the flux of about 20% of the reactions embedded in the network. Moreover, a shift in sugars utilization was observed: from pentose phosphate and hexose to amino sugars metabolism, which are supposed to be more abundant in rhizosphere [4]. This shows an inherent robustness of the metabolic network, which can easily accommodate for nutrient variability when moving from soil to rhizosphere. Such robustness may imply that several, still unknown, plant species (together with their soil and rhizosphere) can be a reservoir for this pathogenic bacterium. Genome-scale gene deletions provided evidence for the presence of a compact core of essential genes in *P. parmentieri* that could explain why bacteria of this species are able to persist on plant residuals without interacting with host plant for long periods of time.

Bibliography

1. Henry, R., Crane, B., Powell, D., *et al.* The transcriptomic response of Acinetobacter baumannii to colistin and doripenem alone and in combination in an in vitro pharmacokinetics/pharmacodynamics model. *Journal of Antimicrobial Chemotherapy* **70,** 1303–1313 (2015) (cit. on p. 119).
2. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences* **109,** E1121–E1130 (2012) (cit. on p. 119).
3. Mansfield, J., Genin, S., Magori, S., *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular plant pathology* **13,** 614–629 (2012) (cit. on p. 121).
4. Sørensen, J., Jensen, L. E. & Nybroe, O. Soil and rhizosphere as habitats for Pseudomonas inoculants: new knowledge on distribution, activity and physiological state derived from micro-scale and single-cell studies. *Plant and soil* **232,** 97–108 (2001) (cit. on p. 121).

SCIENTIFIC REPORTS

OPEN

Constraint-based modeling identifies new putative targets to fight colistin-resistant *A. baumannii* infections

Received: 14 September 2016
Accepted: 26 April 2017
Published online: 16 June 2017

Luana Presta[1], Emanuele Bosi[1], Leila Mansouri[1], Lenie Dijkshoorn[2], Renato Fani[1] & Marco Fondi[1]

Acinetobacter baumannii **is a clinical threat to human health, causing major infection outbreaks worldwide. As new drugs against Gram-negative bacteria do not seem to be forthcoming, and due to the microbial capability of acquiring multi-resistance, there is an urgent need for novel therapeutic targets. Here we have derived a list of new potential targets by means of metabolic reconstruction and modelling of *A. baumannii* ATCC 19606. By integrating constraint-based modelling with gene expression data, we simulated microbial growth in normal and stressful conditions (*i.e.* following antibiotic exposure). This allowed us to describe the metabolic reprogramming that occurs in this bacterium when treated with colistin (the currently adopted last-line treatment) and identify a set of genes that are primary targets for developing new drugs against *A. baumannii*, including colistin-resistant strains. It can be anticipated that the metabolic model presented herein will represent a solid and reliable resource for the future treatment of *A. baumannii* infections.**

Bacteria of the genus *Acinetobacter* were long considered harmless, environmental organisms, but from the 1960s onward, an increasing number of reports have documented the emergence of *Acinetobacter* strains of this genus among severely ill, hospitalized patients. These strains showed unusually high levels of resistance to antibiotics that could be used at the time. Also, they gave rise to cross-infections and outbreaks among patients[1]. Recently, resistance to antibiotics in *A. baumannii* has risen to worrisome proportions (as reviewed in ref. 2), from susceptible prior to the 1960s, to multidrug-resistant (MDR) (end 1970s), and extended- and pan-drug resistant (XDR, PDR) today. Currently, *A baumannii* is one of the most prominent organisms that are both antibiotic resistant and involved in health associated infections, the so-called ESKAPE organisms (that include *Enterococcus faecium*, *Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa*, and *Enterobacter* spp.[3]).

One of the last-line treatments for MDR *A. baumannii* is colistin, a positively charged molecule that, by interacting with the lipid A moiety of lipopolysaccharide (LPS), causes disorganization of the outer membrane. Unfortunately, colistin resistance in *A. baumannii* has also been reported, thus highlighting the urgency of finding new molecules to face this threat[4]. Although careful monitoring, antimicrobial stewardship and measures to prevent spread in health care institutions are important for controlling *A. baumannii* infections, new antimicrobial agents and/or strategies are urgently required to eradicate antibiotic resistant strains from affected patients.

To address new possible solutions, a system-level study of antibiotic-response in *A. baumannii* is required. Metabolic network reconstruction and its conversion to a mathematical framework has become a cornerstone for studying the systems biology of metabolism[5], allowing the examination of the connection between phenotype and genotype and driving biological discoveries. In particular, constraint-based tools (such as Flux Balance Analysis, FBA) enable the estimation of the rate that metabolites' flow through a metabolic network and to compute cellular phenotypes for various growth conditions[6]. Interestingly, by identifying those genes whose deletion is predicted to impair cellular growth, this *in silico* technique can be used to predict essential genes (EGs) at a genome-scale. Following a metabolic modelling approach, several EGs datasets have already been derived

[1]Department of Biology, University of Florence, Florence, Italy. [2]Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands. Correspondence and requests for materials should be addressed to M.F. (email: marco.fondi@unifi.it)

Figure 1. Comparison between *in silico* and wet-lab experimental outcomes.

for important pathogens such as *Helicobacter pylori*[7], *Pseudomonas aeruginosa*[8], *Mycobacterium tuberculosis*[9], and *Staphylococcus aureus*[10]. Usually, such predictions are performed simulating growth in an arbitrarily defined medium, accounting for the main nutrients used by the microbe and without imposing any additional constraint to the model. Indeed, the search space of essential genes predicted can be narrowed by imposing additional constraints on the model. One possibility consists of modulating the flux admissible across each reaction on the basis of the expression values of the corresponding genes. By doing so, it is possible to generate context-specific models that reflect the actual set of reactions employed[11]. This approach promises to reduce i) the gap between the predicted and real cellular metabolic landscapes, and ii) the number of false positives/negatives in EGs predictions. Additionally, it might reveal hints for the synergistic use of antibiotics and, in particular, to the possible additional targets that might arise from the adaptation/response of a microbe's metabolism to a single antibiotic. Indeed, changes in gene expression might redirect the cellular metabolic fluxes in such a way that novel and untapped essential reactions may emerge, representing good candidates for a synergic antibiotic. Despite that the use of antibiotics in combination is sometimes questionable, this approach can be considered in cases of severe infections and it has been shown to be effective in the case of *Pseudomonas* and *Acinetobacter* spp.[12,13].

Here, we explored the system-level metabolic consequences of *A. baumannii* exposure to colistin. We integrated gene expression data during exposure to colistin[14] with a newly reconstructed genome scale metabolic model, allowing for constraint-based modelling of the type strain ATCC 19606. Our data revealed the metabolic reprogramming that occurred in this strain following the establishment of a stressful condition such as the presence of an antibiotic. Furthermore, the metabolic reconstruction provided here represents an important resource for the future understanding of *A. baumannii* metabolism and for the detection and identification of novel drug targets.

Results and Discussion

Genome-scale *A. baumannii* ATCC 19606 model is consistent with large scale phenotypic data. A preliminary draft reconstruction of the *A. baumannii* ATCC 19606 metabolic model was obtained through the Kbase server (http://kbase.us). This was manually curated as described in Methods. Afterwards, we used previously published large-scale phenotypic data[15] to validate our reconstruction over a large set of experimental tests. Manual curation was performed by comparing FBA outcomes with such auxotrophies data (determined through Phenotype Microarray (PM) technology). During this process, the capability of our model to represent the observed phenotypes was tested.

Growth rates were firstly estimated *in silico* in simulated Simmons minimal medium (a standard bacteriological medium that contains only essential inorganic salts) under aerobic conditions by iteratively probing each C-source used in PM plates. During these simulations, biomass optimization was selected as the model objective function (O.F.). Results of the simulations (either "growth" or "no growth", *i.e.* the estimated flux value across biomass assembly reaction) were compared with the activity directly measured during an experimental phenotype microarray experiment, and discrepancies identified between the *in silico* and experimental data were manually adjusted as possible (such as by filling in missing transport reactions or metabolic gaps).

Following this procedure, we reached an overall agreement of about 84% between the *in silico* and experimental data: out of the 67 *in silico* screened metabolites, 24 were correctly found to be carbon and energy sources for *A. baumannii* ATCC 19606 (true positives) and 32 not (true negatives) while only 10 disagreements remained, 5 false negatives and 5 false positives. All data are briefly summarized in Fig. 1, and a detailed description of the outcomes of the comparison is reported in Supplementary Material S1, Supplementary Table 1.

The current version of the *A. baumannii* ATCC 19606 genome-scale metabolic model (named iLP844 according to the current naming convention[16]) contains 1628 reactions (162 exchange reactions), 1509 metabolites,

COG Functional Category	Description	N. of genes
J	Translation, ribosomal structure and biogenesis	28
A	RNA processing and modification	0
K	Transcription	11
L	Replication, recombination and repair	21
B	Chromatin structure and dynamics	0
D	Cell cycle control, cell division, chromosome partitioning	1
Y	Nuclear structure	0
V	Defense mechanisms	4
T	Signal transduction mechanisms	5
M	Cell wall/membrane/envelope biogenesis	60
N	Cell motility	0
Z	Cytoskeleton	0
W	Extracellular structures	0
U	Intracellular trafficking, secretion, and vesicular transport	0
O	Posttranslational modification, protein turnover, chaperones	20
C	Energy production and conversion	120
G	Carbohydrate transport and metabolism	61
E	Amino acid transport and metabolism	182
F	Nucleotide transport and metabolism	57
H	Coenzyme transport and metabolism	82
I	Lipid transport and metabolism	97
P	Inorganic ion transport and metabolism	71
Q	Secondary metabolites biosynthesis, transport and catabolism	16
R	General function prediction only	32
S	Function unknown	12
X	No Functional Class Found	15

Table 1. Number of genes in the model per COG categories.

and 844 genes (~23% of all ORFs present in this organism, see Supplementary Material S2). Importantly, this proportion is comparable with the coverage of *Escherichia coli* K12 model, iAF1260[17] (27%), *A. baylyi* ADP1 model, iAbaylyi[18] (22%), and *A. baumanni* AYE model, AbyMBEL891[19] (17%). A COG classification of all the genes embedded in the model is provided in Table 1. The final *A. baumannii* ATCC 19606 model is available as Supplementary Material in SBML validated format and embedding cross-references to several databases (Supplementary Material S3).

Constraint-based modelling identifies metabolic EGs. As already mentioned, the identification of EGs is one of the key-steps in a drug discovery pipeline. Indeed, both general and condition-specific EGs can be distinguished[20]. The formers are required to sustain life under virtually all growth conditions; conversely, under specific constraints, changes of central metabolism may occur leading not only to a change in flux distribution throughout the network, but also to changes in gene essentiality and to the appearance of condition-specific essential genes. Hence, we systematically evaluated relevant switches in both unconstrained and constrained models (different *scenarios*), by imposing the necessary constraints to the metabolic reconstruction. Accordingly, screens for EGs were performed for multiple specific conditions: starting by simply changing the set of available nutrients (*i.e.* simulating different environmental niches) and then by simulating stressful situations such as antibiotic exposure and mutations (using available expression data in such conditions).

Nutrient availability influences identified metabolic Egs. As we were interested in modelling the system in a gradually constrained manner, we initially simulated an arbitrary rich medium, allowing our model to have virtually all the nutrients needed – as likely happens inside of a host[19]. To do so, we set the lower bound of exchange reactions as described in methods. Then we performed *in silico* gene deletions for each gene in the model. Accordingly, each gene was defined as essential if its elimination destroyed the network's ability to synthesize at least one key biomass molecules (*i.e.* the model predicts no-growth); otherwise, the gene was considered to be dispensable. Gene essentiality analysis was performed through both FBA and MOMA approaches (see Methods),

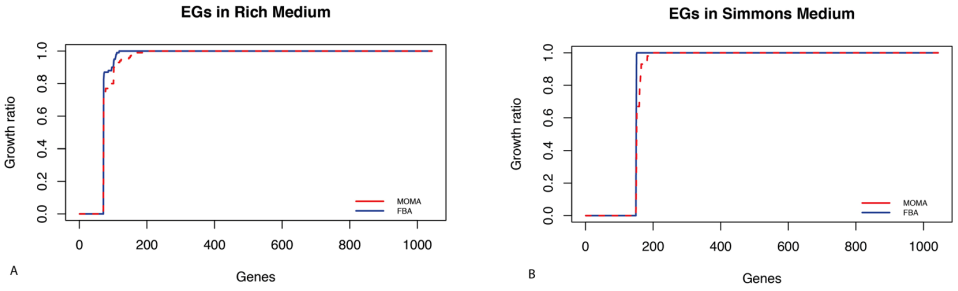


Figure 2. GR_{ratio} value for each gene deletion in rich (**A**) and minimal (**B**) media. Blue and red lines represent MOMA and FBA predictions, respectively. Please note that, in order to make the analysis more comprehensive, also gap-filling genes (i.e. those virtually coding for gap-filling reactions) were included, leading to a total of 1043 simulated knock-outs.

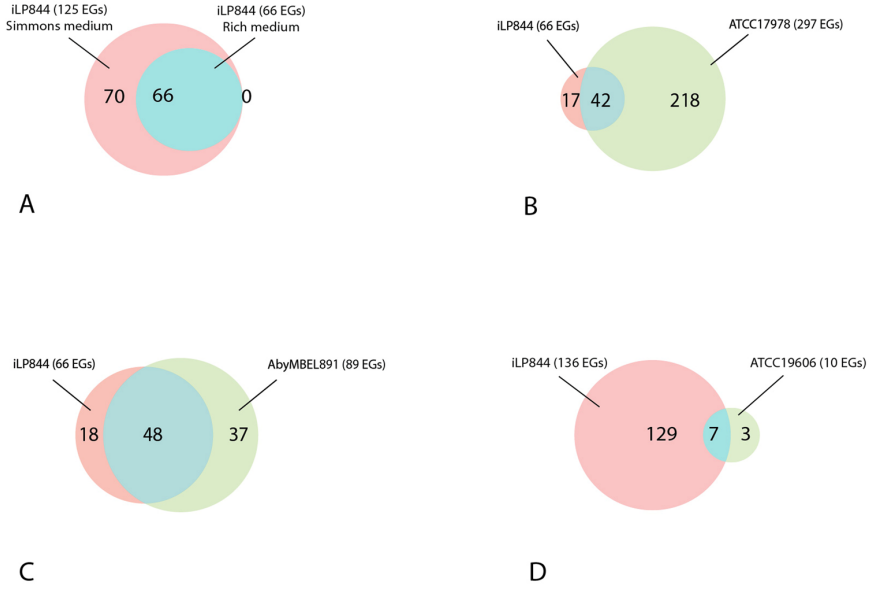


Figure 3. (**A**) A venn diagram proportionally showing EGs predicted only in Simmons medium (pink), EGs predicted only in rich medium (green), and EGs predicted by both (blue). (**B**) A venn diagram proportionally showing EGs predicted *in silico* only by iLP844 (pink), EGs obtained only by wet-lab experiment in ATCC 17978 (green), and EGs predicted by both methods (blue). (**C**) A venn diagram proportionally showing essential reactions predicted in iLP844 (pink), essential reactions predicted only in *A. baumannii* AYE model (AbyMBEL891) (green), and essential reactions predicted by both (blue). (**D**) A venn diagram proportionally showing EGs predicted only in iLP844 (pink), EGs predicted only by wet-lab experiment in *A. baumannii* ATCC 19606 cell (green), and EGs predicted by both (blue).

which both lead to the identification of 67 EGs. Fig. 2A shows the values of the ratio between the predicted growth rate of the gene knock-out mutant and the wild type strain (GR_{ratio}, see Methods). The complete list of EGs and their functions is reported in Supplementary Material S4.

Next, we repeated EGs prediction by simulating growth in a minimal medium (Simmons medium, as defined in Supplementary Material S1, Supplementary Table 2). As shown in Fig. 2B, this approach labelled a total of 138 genes as indispensable for growth in this condition (see Supplementary Material S4 for the complete list). Differences emerging from these two simulations highlight how nutrient availability affects cell metabolism and, interestingly, how different environmental pressures influence gene essentiality. Particularly, 67 genes were predicted to be essential under both the tested growth conditions, while 71 are likely to become essential only when limited nutrient availability force the cell to reprogram its metabolic behaviour (*i.e.* in Simmons medium, Fig. 3A).

Predicted EGs are consistent with available experimental datasets. A large body of data exists concerning *A. baumannii* gene essentiality. Here we used such information both to validate our EGs prediction and to understand whether the identified EGs sets are particular to *A. baumannii* ATCC 19606.

First we compared the EGs dataset obtained in the arbitrary rich medium to that obtained with an *in vivo* experiment on *A. baumannii* ATCC 17978 pathogenesis[21]. By using LB medium (a well-known bacteriological rich medium), Wang *et al.*[21] labelled 481 genes as essentials for that strain. However, not all of them were comparable with our predictions since a large fraction was neither metabolic or possessed an orthologous gene in *A. buamannii* ATCC 19606. In both cases these genes are absent in iLP844. For the same reason, not all the 67 EGs found through our simulation were comparable with the reported experiment. After performing all these necessary restrictions, we reached the result shown in Fig. 3B, *i.e.* 42 genes have been predicted to be essential by both approaches (*in silico* and wet-lab) for the two *A. baumannii* strains considered (Fig. 3B). A complete description of these EGs is provided in Supplementary Material S4 and represents an experimentally validated dataset in the context of *A. baumannii* drug target identification. Nevertheless, the two experiments show large discrepancies. The most likely reason for such inconsistency is strains genomic diversity, as previously reported for *E. coli* strains[22].

Furthermore, our predictions in arbitrary rich medium were compared to those achieved performing the same analysis on the AbyMBEL891 model, an existing model of *A. baumannii* AYE[19]. In order to implement the simulation, it was necessary to perform a preliminary editing step on the AbyMBEL891 model, as the entire set of gene-reaction-rules was missing from the main reconstruction file. This difficulty in running the analysis highlights the need for a common protocol to be adopted during metabolic reconstruction and a standard to be reached in order to facilitate model re-use and data sharing among research groups. Nevertheless, after including the genes in the Abymbel891 model, we carried out single gene deletion analysis on both models, as described in the methods. As shown in Fig. 3C, 48 genes were predicted to be essential in both models, whereas 18 and 37 EGs were specific for *A. baumannii* ATCC 19606 and AYE, respectively. Information about the gene function are reported in Supplementary Material S4.

Comparisons were also carried out between our *in silico* predictions and wet-lab results in minimal (Simmons) medium. Specifically, we compared our EGs set to that obtained by Dorsey *et al.* through insertional mutagenesis experiments with *A. baumannii* ATCC 19606[23], where the metabolic deficiency of insertion derivatives was subsequently confirmed, identifying essentiality of 10 disrupted genes. Repeating the assay *in silico*, our model correctly represented the phenotypes of the *A. baumannii* mutants, with 8 out of the 10 genes predicted as essential by Dorsey and colleagues also shown to be essential in iLP844 (Fig. 3D, Supplementary Material S4). Additionally, in 6 out of the 7 cases, *A. baumannii* ATCC 19606 model growth was correctly restored (as done in the corresponding wet-lab experiments) by adding to the minimal medium the metabolite(s) whose production was affected by the mutation.

Antibiotic treatment defines condition-specific models. Although a large fraction of the predictions was supported by previous experimental data, a possible source of error, using the methodology described above, stems from the observation that not all the reactions of the model will be active during growth in a given physiological condition. In particular, changes in gene expression are likely to influence the activity rate of the corresponding cellular metabolic reactions, leading to the observation that a given reaction can be considered 'turned on' or 'off' on the basis of the expression levels of the encoding gene(s). Using available computational methodologies, it is possible to modulate the flux across each reaction on the basis of the expression values of the corresponding genes. This permits to take a picture of the current metabolic state and tightening up the predictive capabilities of the model itself. Accordingly, as the dynamic changes of metabolic reprogramming are likely mirrored by changes in gene essentialities, a possible solution for avoiding or reducing false positives is to merge transcriptomics data of the tested *scenario* into the genome-scale model.

Arguably, one of the most interesting physiological conditions of *A. baumannii* strains is the exposure to antibiotics and to colistin in particular[14]. Importantly, both the (metabolic) consequences and the occurrence of targets to be used in a synergic treatment are, currently, almost untapped. In order to study the dynamic changes of the metabolic network following antibiotic exposure and to derive a more realistic picture of gene essentiality patterns in a real scenario (antibiotic treatment), we used available transcriptomic data for *A. baumannii* ATCC 19606 in response to colistin treatment[14]. Up-regulation and down-regulation ratios (and corresponding *P*-values) of genes were combined with the iLP844 by using MADE (Metabolic Adjustment by Differential Expression)[24]. Briefly, MADE uses statistically significant changes in gene expression measurements to determine binary expression states (highly and lowly expressed reactions) *i.e.* reactions are turned on and off depending on the changes in mRNA transcript levels. Thus, by mapping gene expression data into the model, the *in silico* metabolic predictions are more consistent with the actual physiological state of the cell.

In the experiment by Henry *et al.*[14], *A. baumannii* was grown in two different media, *i.e.* with and without 2 mg/L of colistin, and then sampled at 15 and 60 minutes after exposure. Following the described approach, we integrated the available transcriptomic data regarding all the metabolic genes embedded in our *in-silico* reconstruction (*i.e.* about 80 genes). Accordingly, we obtained four distinct models, each representing the predicted functional metabolic state of the cell at both 15 and 60 minutes, treated and untreated with colistin. These models differ in that some of their reactions are (completely) 'turned on' or 'off' according to the measured levels of their corresponding genes. Afterwards, optimization of the four models was performed, allowing the analysis of flux distribution in the network and the occurring metabolic reshape.

Colistin exposure changes predicted metabolic fluxes in central *A. baumannii* pathways. In order to highlight changes in the overall metabolic behaviour and to identify changes on the metabolic rewiring occurring after antibiotic exposure, we compared flux distributions at the two time-points by calculating the flux

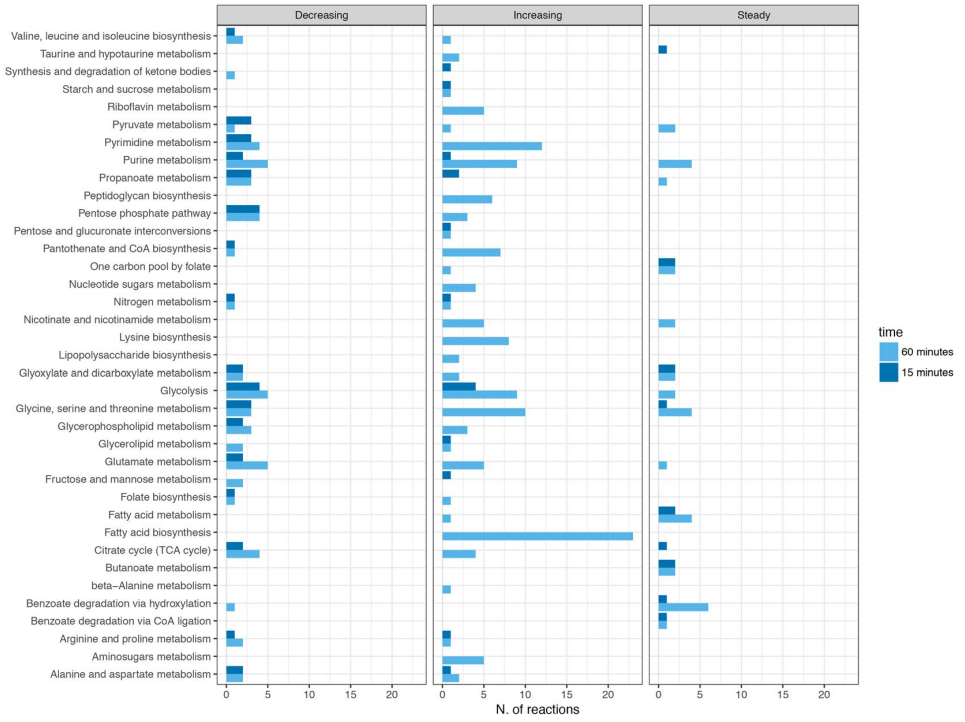


Figure 4. Abundance plot of reactions affected by colistin treatment at 15 (blue) and 60 (pink) minutes arranged according to three categories: 'steady', 'increasing', 'decreasing'. Pathways which they belong to are reported.

ratio (RF_{ratio}, see Methods) of treated *vs*. untreated models, for all the reactions. However, FBA only provides one of the possible optimal solutions out of many alternative (and feasible) cellular flux distributions. Hence, in order to correctly predict metabolic changes following antibiotic exposure we restricted the feasible solution space by performing Flux Variability Analysis (FVA)[25]. This approach allows estimating the minimum and maximum flux admissible across each reaction (under the same constraints as in FBA) and hence it can be used to estimate the correctness and accuracy of FBA predictions (see Methods).

As shown in Supplementary Material S5, according to FVA, the range of admissible flux is sometimes very large (spanning from the minimum −1000 to maximum 1000 $mmol/g*h^{-1}$ in some cases), revealing the lack of accuracy in some of FBA-derived predictions. We here used FVA outcomes (as described in Methods) to filter out those reactions whose fluxes display little variation. In other words, each reaction was considered for downstream analyses only if both the maximum and minimum FVA predicted fluxes did not differ from the FBA predicted flux by more than 20%. Consequently, we were left with 901 reactions at 15 minutes and 970 reactions at 60 min. It is worth noting that several intervals of admissible flux ranges were tested and we report in Supplementary Materials S1, Supplementary Figure 1 the number of reactions filtered for each set of intervals. After carrying out this preliminary step, we observed the effects of the treatment at the metabolic level (for each reaction) by comparing the flux values in the untreated *vs*. treated condition.

Both qualitative and quantitative flux changes were analysed by dividing the reactions into three categories ('steady', 'increasing', and 'decreasing', see Fig. 4) according to their trends in the examined experimental conditions. Also, we report a survey of the pathways in which they are involved in and their relative abundance for each category. Reactions' fluxes were considered 'steady' if their values did not change in the two conditions, otherwise they were defined to be 'increasing' or 'decreasing' according to the corresponding trend.

As shown in Fig. 4, at both 15 and 60 min time-points there is an increase in flux in most of the reactions. Interestingly, such change in flux mainly occurs in three biosynthetic pathways: fatty acid, peptidoglycan, and lysine biosynthesis. On the other hand, under the given constraints, there is a change in flux in some catabolic pathways (mainly involved in sugars and nucleotide metabolism). In our opinion, such a finding could be related to the rearrangement of the external membrane layer, a well-known effect of colistin treatment. If this is true, it is possible that the cell reacts to the antibiotic treatment by trying to repair the damage established by colistin while at the same time redirecting a certain amount of LPS components to catabolic processes.

Also, we would like to point-out that, although a down-regulation of certain genes involved in fatty acids biosynthesis was detected by Henry *et al*., here our data suggest that it does not necessarily imply a turning-off of the pathway. On the contrary, in our simulation fatty acid biosynthesis registers an increase in flux, probably as a side-effect of LPS disassembly as stated above.

Locus Tag	Function	15′ without colistin	15′ with colistin	60′ without colistin	60′ with colistin	LPS− without colistin	LPS− with colistin
HMPREF0010_00435	phosphopyruvate hydratase	yes	no	yes	no	no	no
HMPREF0010_00813	fructose-bisphosphate aldolase class II Calvin cycle subtype	yes	no	yes	no	no	no
HMPREF0010_00815	phosphoglycerate kinase	yes	no	yes	no	no	no
HMPREF0010_00975	amino acid ABC transporter periplasmic protein	yes	no	yes	no	no	no
HMPREF0010_01733	PAP2 superfamily protein	yes	no	yes	no	no	no
HMPREF0010_01995	23-bisphosphoglycerate-independent phosphoglycerate mutase	yes	no	yes	no	no	no
HMPREF0010_02140	fructose-16-bisphosphatase	yes	no	yes	no	no	no
HMPREF0010_03273	glucose-6-phosphate isomerase	yes	no	yes	no	no	no
HMPREF0010_00382	phosphotransferase system fructose-specific EI/HPr/EIIA component	yes	no	yes	no	no	no
HMPREF0010_03275	UTP-glucose-1-phosphate uridylyltransferase	yes	no	yes	no	no	no
HMPREF0010_01305	alphaalpha-trehalose-phosphate synthase (UDP-forming)	yes	no	yes	no	no	no
HMPREF0010_01353	glutamine-dependent NAD+ synthetase	yes	no	no	no	no	no
HMPREF0010_00342	ornithine carbamoyltransferase	no	yes	no	no	no	yes
HMPREF0010_00949	malonate decarboxylase epsilon subunit	no	yes	no	yes	no	no
HMPREF0010_01149	acetyl-CoA carboxylase biotin carboxylase	no	yes	no	no	no	no
HMPREF0010_01969	argininosuccinate lyase	no	yes	no	yes	no	no
HMPREF0010_02047	carbamoyl-phosphate synthase large subunit	no	yes	no	no	no	yes
HMPREF0010_02048	carbamoyl-phosphate synthase small subunit	no	yes	no	no	no	yes
HMPREF0010_02972	argininosuccinate synthase	no	yes	no	no	no	yes
HMPREF0010_03445	34-dihydroxy-2-butanone 4-phosphate synthase	no	yes	no	no	no	no
HMPREF0010_00048	sugar kinase	no	yes	no	no	no	no
HMPREF0010_02330	glutamine synthetase type I	no	no	yes	no	no	no
HMPREF0010_01216	quinolinate synthetase complex A subunit	no	no	no	yes	no	no
HMPREF0010_02175	thymidylate kinase	no	no	no	yes	no	yes
HMPREF0010_03295	nicotinate-nucleotide diphosphorylase (carboxylating)	no	no	no	yes	no	no
HMPREF0010_01705	ketol-acid reductoisomerase	no	no	no	no	yes	no
HMPREF0010_00392	ornithine-oxo-acid transaminase	no	no	no	no	no	yes
HMPREF0010_00419	N-acetyl-gamma-glutamyl-phosphate reductase	no	no	no	no	no	yes
HMPREF0010_00840	methionine adenosyltransferase	no	no	no	no	no	yes
HMPREF0010_01215	ArgJ protein	no	no	no	no	no	yes
HMPREF0010_01382	acetylglutamate kinase	no	no	no	no	no	yes
HMPREF0010_01506	methionine synthase	no	no	no	no	no	yes
HMPREF0010_01669	3-methyl-2-oxobutanoate hydroxymethyltransferase	no	no	no	no	no	yes
HMPREF0010_01887	2-heptaprenyl-14-naphthoquinone methyltransferase	no	no	no	no	no	yes
HMPREF0010_01969	argininosuccinate lyase	no	no	no	no	no	yes
HMPREF0010_02002	riboflavin synthase alpha subunit	no	no	no	no	no	yes
HMPREF0010_02344	adenosylhomocysteinase	no	no	no	no	no	yes
HMPREF0010_0275979	glutamate racemase	no	no	no	no	no	yes
HMPREF0010_03444	67-dimethyl-8-ribityllumazine synthase	no	no	no	no	no	yes

Table 2. Complete set of condition specific EGs found. "Yes" and "no" refet to gene essentiality in the corresponding condition.

Colistin exposure changes gene essentiality patterns. According to the new constraints taken into account, gene essentiality was re-evaluated by calculating growth ratios (FBA and MOMA) at both 15 and 60 minutes after exposure to colistin. As for the analysis involving nutrient availability, shifts in gene essentiality emerged following antibiotic stress. The complete sets of the predicted EGs for each condition have been reported in Supplementary Material S4. As with the previous case, we can easily recognize genes likely to be essential in both conditions (treated and not) and, more interestingly, genes that emerged as essential only after the treatment. Specifically, following 15 minutes of colistin exposure, a total of 87 EGs were predicted: 66 were required both in presence and absence of colistin, but an additional 21 EGs were marked as condition-specific: 12 related to the non-treated model and 9 related to the treated one, reported in Table 2 and in Fig. 5A.

The same outline has been depicted in the second time-point condition (60 minutes): we identified 12 and 5 condition-specific EGs in the absence and in the presence of the antibiotic, respectively, see Table 2. Moreover, we found the same set of 66 EGs mentioned above (see Fig. 5B), suggesting that this represents a functionally relevant set of genes for sustaining growth in *A. baumannii* ATCC 19606. Interestingly, however, some genes switch from the 'essential' condition to the 'non-essential' one, following the exposure to colistin.

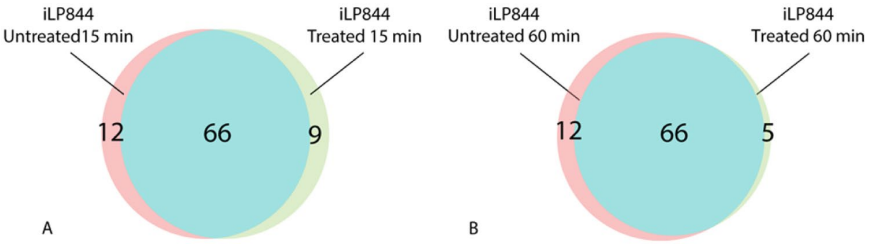


Figure 5. Venn diagrams proportionally showing EGs predicted only in absence of colistin (pink), EGs predicted only in presence of colistin (green), and EGs predicted by both (blue), at 15 minutes (**A**) and 60 minutes (**B**).

The two new sets of EGs show how changes in gene expression induced by the presence of the antibiotic might influence gene essentiality patterns in the strain ATCC 19606 and provide additional, nontrivial targets for drug design in such an organism. We also performed additional robustness analyses in order to test whether nutrients depletion occurring in treated and untreated samples during the *in vivo* experiments could impact the set(s) of predicted EGs. Specifically, the robustness of the number of predicted EGs in each of these conditions (*i.e.* treated *vs.* untreated samples) in respect to possible variations in the medium composition was assessed via random permutation. We tested up to 1,000 different nutritional compositions as described in detail in Supplementary Material S1, Supplementary Figure 2). The results showed that possible changes in the nutritional environment had only minor implications for the set of predicted EGs.

Further, with the aim of discriminating whether the products of all the identified hypothetical EGs are *A. baumannii* specific or have orthologs in *Homo sapiens*, meaning that they would not represent good candidates for antibiotic treatment development, the sequences of the identified potential EGs were used to probe the human genome. Based on this BLAST[26] search (see Methods), we excluded from further studies those genes presenting more than 30% sequence identity with their human counterparts. Targeting of such genes is non-ideal since they may cause potential side-effects by perturbing critical components in the human body. All BLAST results are reported in Supplementary Material S6.

Among all the queries, we identified 46 (out of 73) general EGs and 4 (out of 12) condition-specific EGs that do not have any human orthologous. Thus, the 46 EGs represent valuable targets for further development of brand new drugs against *A. baumannii* ATCC 19606 infections. However, it is relevant to remark that, while these 46 general EGs could have been detected in several conditions, the other 4 condition-specific EGs are the result of specific constraints integrated in the model (gene expressions data). Hence, as already mentioned, they are nontrivial detections and they could represent a suitable horizon in the field of colistin-coupled treatment. The four genes, named HMPREF0010_00949, HMPREF0010_02972 and HMPREF0010_03445, HMPREF0010_01216 respectively encode a malonate decarboxylase (epsilon subunit), an arginine succinate synthase, a 3-4-dihydroxy-2-butanone-4-phopsphate synthase and a quinolate synthase. Interestingly, malonate decarboxylase epsilon subunit has already been characterized in the closely related organism *Pseudomonas putida* and labelled as an indispensable component of the enzyme for the cyclic decarboxylation ofmalonate[27]. However, to the best of our knowledge, no therapies targeting this protein have been developed to date. The product of 3-4-dihydroxy-2-butanone-4-phopsphate synthase is an intermediate in the biosynthesis of riboflavin. The enzyme requires a divalent cation, preferably Mg_2^+, to be active. The step becomes essential after colistin treatment as the antibiotic is predicted to cause an increase in flux through this pathway, probably following the shutdown of other parts of the network due to the down regulation of the corresponding genes. The arginine succinate synthase, is an enzyme catalysing the penultimate step in arginine biosynthesis (urea-cycle): the ATP-dependent ligation of citrulline to aspartate in order to form arginino-succinate, AMP, and pyrophosphate. The last EGs encodes for a bacterial specific enzyme that catalyses the second step in the *de novo* biosynthesis of NAD+ from aspartate.

EGs in colistin resistant *A. baumannii*. Up to now, we have presented how metabolic reconstruction and mathematical modelling can be used to explore the strain's metabolic response during colistin treatment and how it can lead to the identification of novel potential drug targets. Our last attempt is now to illustrate how, starting from the same available experimental data, the model can be employed as a ready-to use blueprint in order to test new hypothesis.

As it was reported by Moffatt *et al.*[28], the mechanism responsible for colistin resistance is linked to LPS. Specifically, mutations in the *lpxA*, *lpxC*, and *lpxD* genes have been reported as the main cause of LPS loss, thus abolishing the initial charge-based interaction with the antibiotic. Hence, to simulate an *A. baumannii* LPS (LPS^-) deficient and colistin-resistant strain, we removed this component from the biomass formulation in our genome-scale model. Then, to determine which genes are central for the cell's survival in such a condition, we used the transcriptomic data of the mutant strain in the presence/absence of colistin at 60 minutes[14] and mapped the data onto the new LPS^- model. After this, we repeated the EGs prediction pipeline described above.

The analysis yielded a total of 55 and 72 EGs in the untreated and treated condition, respectively. Even in this case, the two sets share several elements (54 EGs) that remain mandatory for the cell in the two conditions (listed in Supplementary Material S4). Additionally, it is possible to observe that 18 genes become essential (reported

in Table 2) only after antibiotic exposure: 9 of them were already found to be EGs in the wild type strain while 9 represent specific EGs of the mutant. Since the latter are non-trivial EGs (obtained through gene expression integration and introduction of an antibiotic resistance promoting mutation into the model) they have been re-used as seed for an additional BLAST search against the human genome (see Supplementary Material S6). The search led to the identification of 5 genes that do not have orthologs in humans: HMPREF0010_01215 encoding for glutamate-N-acetyltransferase (member of the ornithine acetyltransferase, OAT, family), HMPREF0010_00419, encoding for N-acetyl-gamma-glutamyl-phosphate reductase, HMPREF0010_01382 encoding for N-acetyl-L-glutamate-kinase, all of which are involved in the arginine biosynthesis pathway, HMPREF0010_0275979 that encodes for a glutamate-racemase and HMPREF0010_02344 that encodes for an adenosylhomocysteinase. This group of 4 EGs represents a potential achievement obtained from this work as it suggests specific targets to be taken into consideration when developing therapies in combination with colistin in an LPS- strain.

Predicted EGs are common in A. baumannii. Finally, we checked the distribution of EGs predicted for the strain ATCC 19606 within the entire *A. baumannii* species. The sets of predicted EGs were searched in all of the 1099 *A. baumannii* genomes sequenced to date, as described in the Methods. The overall result is shown in Supplementary Material S1, Supplementary Figure 3. The general trend observed was that more than 90% of the genomes analysed possessed the searched queries (identity >30%). Also, our analysis shows that this tendency is kept almost unchanged even when imposing an identity threshold greater than 50%, 70% and 90%. Accordingly, it can be stated that the possible target genes are broadly distributed and their sequence is conserved at the *A. baumannii* species level. Although we do not have any information about the EGs at such a wide level, this preliminary result is encouraging, since it expresses the possibility that the target genes we indicated for *A. baumannii* ATCC 19606 are probably common targets in most of *A. baumannii* type infections.

Conclusions

In this work, we have reconstructed and validated a genome-scale metabolic model of *A. baumannii* ATCC 19606. The model is comprehensive and accurate, as it covers ~23% of all CDSs in the genome of this microorganism and it was shown to have 84% agreement with Phenotype Microarray growth experiments. Based on the model's reliability, we applied constraint-based modelling to derive a global understanding of the behaviour of this metabolic system. By integrating gene expression data with constraint-based modelling we described the metabolic reprogramming occurring after colistin-exposure in *A. baumannii* and the changes in the pattern of gene essentiality during this stress condition. All the sets of condition-specific putative target genes that we propose have been compared (and partially validated) with the results obtained from experiments found in the literature. Some of these genes, although not yet experimentally validated, might represent primary targets for future research on the treatment of both the wild type and LPS-mutant (*i.e.* colistin resistant) strains. Our results have practical implications for the identification of new therapeutics as the identified essential genes can be used in drug-design pipelines. Moreover, we showed that the sequences of predicted EGs for the type strain ATCC 19606 are shared by most of the members of *A. baumannii* species, encouraging further research to check whether they are valuable drug targets for a larger number of strains than currently known. Finally, it can be anticipated that the iLP844 model illustrated herein represents a reliable and solid platform for further developments and the system-level understanding of the physiology of *A. baumannii* representatives and for the treatment of their infections.

Methods

Draft model reconstruction. We obtained a draft metabolic model of *A. baumannii* ATCC 19606 based on the genome annotation using Kbase automated reconstruction method (https://kbase.us/)[29]. This reconstruction was then thoroughly inspected and refined by integrating data from additional functional databases (MetaNetX[30], Bigg[31], Seed[32], KEGG[33]). Further integration was performed by searching for orthologous genes (genes likely having an identical biological function in a different organism) in closely related organisms (*Acinetobacter baumanii* AYE, *Acinetobacter baylyi* ADP1, and *Escherichia coli*) through a BBH (Bidirectional Best Hit) approach (inParanoid[34]). Information regarding transport proteins was obtained probing the Transporter Classification Data Base (TCDB[35]) and transportDB[36].

In order to predict proper phenotypes, the general biomass producing reaction of Gram negative bacteria automatically generated by Kbase was substituted with a more accurate one that takes into account strain's specific components, which was recovered from the previously reported model of the related strain *A. baumannii* AYE, AbyMBEL891[19].

Metabolic modelling. The reconstructed model was analysed using CoBRApy-0.4.1 COnstraints-Based Reconstruction and Analysis for Python[37] and COBRAToolbox-2.0[38] in MATLAB® R2016a (Mathworks Inc.). Gurobi 6.5.0 (www.gurobi.com) and GLPK 4.32 (http://www.gnu.org/software/glpk/) solvers were used for computational simulations presented. A MATLAB® script to obtain all the results shown in this manuscript is provided as Supplementary Material S7.

Two growth media were considered during the *in silico* simulations:

Rich medium. Lower bounds of salts uptake reactions were set to −1000 mmol/g*h^{-1} in order to mimic non-limiting conditions. Carbon sources uptake reactions were set to −100 mmol/g*h^{-1}.

Simmons medium. Lower bounds of all uptake reactions accounting for the nutrients present in Simmons medium[39] (see Supplementary Material S1, Supplementary Table 2), were set to −1000 mmol/g*h^{-1}, to mimic non limiting conditions, only the C-source (citrate) was set to −5.

FVA. FVA analysis allows the determination of the span of possible flux variability (*i.e.* the maximum and minimum values of all the fluxes that satisfy the given constraints) while keeping the same optimal objective value.

This approach has been used in this work in order to impose bounds to FBA flux predictions (f_{FBA}), which are notably non-unique. In fact, for any optimal solution found through FBA there may exist alternate flux distribution patterns yielding the same growth rate. Hence, the space of reliable FBA-flux predictions has been restricted by selecting only those that occur in the interval defined as follows:

$$\text{with } f_{FBA} < 0 \qquad f_{FVA,\,min} \geq 1.2 * f_{FBA} \wedge f_{FVA,\,max} \leq 0.8 * f_{FBA} \tag{1}$$

$$\text{with } f_{FBA} > 0 \qquad f_{FVA,\,min} \geq 0.8 * f_{FBA} \wedge f_{FVA,\,max} \leq 1.2 * f_{FBA} \tag{2}$$

Gene essentiality and flux ratios calculation. Gene essentiality testing was performed by simulating deletion of each gene within the metabolic network and hence setting the associated reactions to carrying no flux (according to the corresponding Gene-Protein-Reaction (GPR) rule). To predict the growth of the mutant strain and determine the set of EGs, we used two different approaches, FBA and MOMA[40]. The main difference between them is that while the first predicts growth yield and metabolic fluxes based on the biological assumption of optimal growth, the second does not assume optimality of growth but approximates metabolic phenotype by performing distance minimization in flux space. The second approach has been shown to be more accurate in predicting lethal phenotypes[40]. The knocked-out gene was defined as 'essential' according to the results obtained computing the ratio (GR_{ratio}) between the simulated knocked out strain growth rate (μ_{KO}) and the one predicted for the wild type strain (μ_{WT}). Formulated as:

$$GR_{ratio} = \mu_{KO}/\mu_{WT} \tag{3}$$

Following this approach, if $GR_{ratio} = 0$, then the knocked out gene is labelled as essential. Conversely, in case $GR_{ratio} = 1$, the removal of the gene has no effect on the growth phenotype. Finally, when $0 < GR_{ratio} < 1$, the deleted gene was labelled as fitness-contributing gene, *i.e.* its removal partially affects the capability of the cell to produce biomass.

As MOMA and FBA predictions may lead to different essential gene sets[41,42,40], we used both approaches to compute essential genes in all the conditions tested in this work. Although no major differences were observed, results obtained with both methods are presented throughout the manuscript.

In order to evaluate the range of the change in the carried flux of each reaction in the model following colistin exposure, we compute the ratio between the predicted flux in the treated *vs*. the untreated conditions as follows:

$$RF_{ratio} = (\varphi_{Treated}/\varphi_{Untreated}) \tag{4}$$

Values of RF_{ratio} equal to one indicate that no changes in the activity of the corresponding reactions were observed when simulating growth in the treated *vs*. the untreated conditions. Conversely, values of RF_{ratio} between 0 and 1 or values greater than 1 will indicate a reduced or increased activity of the corresponding reactions in the treated condition, respectively. Finally, negative values of RF_{ratio} will indicate those reactions whose directionality is predicted to change after the treatment.

Transcriptomics data integration and data visualization. In order to add transcriptional regulatory rules to the metabolic model, we imported the model from COBRA Toolbox into TIGER-1.2.0 0 (Toolbox for Integrating Genome-scale metabolism, Expression, and Regulation) framework (12). Then, the up- and down-regulation ratios of gene expression were mapped into the *A. baumannii* ATCC 19606 metabolic model by using MADE (Metabolic Adjustment by Differential Expression)[24]. The program uses significant statistical changes in gene or gene expression to create functional metabolic models. By adopting an optimization approach that applies Boolean rules, MADE connects reactions to the binary expression states of associated genes. The four arrays of genes to be switched-off yielded by MADE have been reported in Supplementary Material S8.

EGs BLAST searches in *H. sapiens* and *A. baumannii* species. Protein sequences of the corresponding EGs found in *A. baumannii* ATCC 19606 were probed against the human proteome to test their validity as potential drug target in infections with this pathogen, *i.e.* to exclude any cross-interactions between the drug used for the treatment and human proteome elements.

Queries were aligned to the protein sequences of *H. sapiens* using the default search parameters of the NCBI BLASTP online tool (BLOSUM62 matrix and gap costs equal to Existence 11, Extension 1). Results were considered positive (orthologous sequences found) if their sequence identity score value was equal to/greater than 30.

In addition, the global distribution of EGs was evaluated at the *A. baumannii* species level by probing them against all of sequenced genomes retrieved at NCBI ftp site, *i.e.* 1099 genomes. Particularly, the focus was centred on EGs found in Simmons medium and in rich medium, as well as for those found after 15 and 60 minutes of colistin exposure. BLAST search parameters and analysis of the results were performed as described above.

References

1. Dijkshoorn, L., Nemec, A. & Seifert, H. An increasing threat in hospitals: multidrug-resistant Acinetobacter baumannii. *Nat. Rev. Microbiol.* **5**, 939–951, doi:10.1038/nrmicro1789 (2007).
2. Gonzalez-Villoria, A. M. & Valverde-Garduno, V. Antibiotic-Resistant *Acinetobacter baumannii* Increasing Success Remains a Challenge as a Nosocomial Pathogen. *J. Pathog* **2016**, 1–10, doi:10.1155/2016/7318075 (2016).
3. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J. Infect. Dis.* **197**, 1079–81 (2008).

4. Cai, Y., Chai, D., Wang, R., Liang, B. & Bai, N. Colistin resistance of Acinetobacter baumannii: Clinical reports, mechanisms and antimicrobial strategies. *J. Antimicrob. Chemother.* **67**, 1607–1615, doi:10.1093/jac/dks084 (2012).
5. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121, doi:10.1038/nprot.2009.203 (2010).
6. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8, doi:10.1038/nbt.1614 (2010).
7. Thiele, I., Vo, T. D., Price, N. D. & Palsson, B. Ø. Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): An in silico genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol.* **187**, 5818–5830, doi:10.1128/JB.187.16.5818-5830.2005 (2005).
8. Oberhardt, M. A. *et al.* Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320, doi:10.1038/msb.2009.77 (2009).
9. Raman, K., Rajagopalan, P. & Chandra, N. Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs. *PLoS Comput. Biol.* **1**, 0349–0358, doi:10.1371/journal.pcbi.0010046 (2005).
10. Lee, D.-S. *et al.* Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. *J. Bacteriol.* **191**, 4015–24, doi:10.1128/JB.01743-08 (2009).
11. Machado, D. & Herrgård, M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Comput. Biol.* **10**, e1003580, doi:10.1371/journal.pcbi.1003580 (2014).
12. Tängdén, T. Combination antibiotic therapy for multidrug-resistant Gram-negative bacteria. *Ups. J. Med. Sci.* **9734**, 149–153, doi:10.3109/03009734.2014.899279 (2014).
13. Motaouakkil, S. *et al.* Colistin and rifampicin in the treatment of nosocomial infections from multiresistant Acinetobacter baumannii. *J. Infect.* **53**, 274–278, doi:10.1016/j.jinf.2005.11.019 (2006).
14. Henry, R. *et al.* The transcriptomic response of Acinetobacter baumannii to colistin and doripenem alone and in combination in an *in vitro* pharmacokinetics/pharmacodynamics model. *J. Antimicrob. Chemother.* **70**, 1303–1313, doi:10.1093/jac/dku536 (2014).
15. Peleg, A. Y. *et al.* The success of acinetobacter species; genetic, metabolic and virulence attributes. *PLoS One* **7**, e46984, doi:10.1371/journal.pone.0046984 (2012).
16. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of Escherichia coli K-12 (i JR904 GSM/GPR). *Genome Biol.* **4**, R54, doi:10.1186/gb-2003-4-9-r54 (2003).
17. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121, doi:10.1038/msb4100155 (2007).
18. Durot, M. *et al.* Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.* **2**, 85, doi:10.1186/1752-0509-2-85 (2008).
19. Kim, H. U., Kim, T. Y. & Lee, S. Y. Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen Acinetobacter baumannii AYE. *Mol. Biosyst.* **6**, 339–348, doi:10.1039/b916446d (2010).
20. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. USA* **109**, E1121–30, doi:10.1073/pnas.1113065109 (2012).
21. Wang, N., Ozer, E. A., Mandel, M. J. & Hauser, A. R. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. *MBio* **5**, e01163–14, doi:10.1128/mBio.01163-14 (2014).
22. Monk, J. M. *et al.* Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments.* **110**, 20338–43, doi:10.1073/pnas.1307797110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1307797110 (2013).
23. Dorsey, C. W., Tomaras, A. P. & Actis, L. a. Genetic and Phenotypic Analysis of Acinetobacter baumannii Insertion Derivatives Generated with a Transposome System Genetic and Phenotypic Analysis of Acinetobacter baumannii Insertion Derivatives Generated with a Transposome System. *Appl. Environ. Microbiol.* **68**, 6353–6360, doi:10.1128/AEM.68.12.6353-6360.2002 (2002).
24. Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27**, 541–7, doi:10.1093/bioinformatics/btq702 (2011).
25. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276, doi:10.1016/j.ymben.2003.09.002 (2003).
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10, doi:10.1016/S0022-2836(05)80360-2 (1990).
27. Chohnan, S., Kurusu, Y., Nishihara, H. & Takamura, Y. Cloning and characterization of mdc genes encoding malonate decarboxylase from Pseudomonas putida. *FEMS Microbiol. Lett* **174**, 311–9, doi:10.1111/fml.1999.174.issue-2 (1999).
28. Moffatt, J. H. *et al.* Colistin resistance in Acinetobacter baumannii is mediated by complete loss of lipopolysaccharide production. *Antimicrob. Agents Chemother.* **54**, 4971–4977, doi:10.1128/AAC.00834-10 (2010).
29. Davenport, K. W. *et al.* Draft Genome Assembly of Acinetobacter baumannii ATCC 19606. *Genome Announc* **2**, e00832–14, doi:10.1128/genomeA.00832-14 (2014).
30. Moretti, S. *et al.* MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res* **44**, D523–D526, doi:10.1093/nar/gkv1117 (2016).
31. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* **44**, D515–D522, doi:10.1093/nar/gkv1049 (2016).
32. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–982, doi:10.1038/nbt.1672 (2010).
33. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361, doi:10.1093/nar/gkw1092 (2017).
34. Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–52, doi:10.1006/jmbi.2000.5197 (2001).
35. Saier, M. H., Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181–D186, doi:10.1093/nar/gkj001 (2006).
36. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* **35**, D274–D279, doi:10.1093/nar/gkl925 (2007).
37. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. (2013).
38. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6**, 1290–1307, doi:10.1038/nprot.2011.308 (2011).
39. James, S. & Simmons., A. Culture Medium for Differentiating Organisms of Typhoid-Colon Aerogenes Groups and for Isolation of Certain Fungi. *J. Infect. Dis* **39**, 209–214, doi:10.1093/infdis/39.3.209 (1926).
40. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–7, doi:10.1073/pnas.232349399 (2002).
41. Snitkin, E. S. *et al.* Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome Biol.* **9**, R140, doi:10.1186/gb-2008-9-9-r140 (2008).
42. Wang, Z. *et al.* Why Is the Correlation between Gene Importance and Gene Evolutionary Rate So Weak? *PLoS Genet.* **5**, e1000329, doi:10.1371/journal.pgen.1000329 (2009).

Author Contributions

M.F. and L.P. conceived the study and prepared the first draft of the manuscript. L.P. performed metabolic network reconstruction. L.P. and M.F. performed the simulations with the model. L.M. performed model gap-filling and participated in the modelling step. M.F., L.P., R.F., L.D. and E.B. discussed the results and participated in the writing process.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03416-2

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Metabolic modeling of *Pectobacterium parmentieri* SCC3193 provides insights into metabolic pathways of plant pathogenic bacteria

Sabina Zoledowska[1,+], Luana Presta[2,+], Marco Fondi[2,], Francesca Decorosi[3,], Luciana Giovannetti[3,], Alessio Mengoni[2,*], and Ewa Lojkowska [1,]

[1]Department of Biotechnology, Intercollegiate Faculty of Biotechnology, University of Gdansk and Medical University of Gdansk, 58 Abrahama Street, 80-307 Gdansk, Poland
[2]Department of Biology, University of Florence, Madonna del Piano 6, 50019 Sesto Fiorentino, Florence, Italy
[3]Department of Agrifood Production and Environmental Science, University of Florence, Florence, Italy
[*]Corresponding author: alessio.mengoni@unifi.it
[+]these authors contributed equally to this work

ABSTRACT

Understanding the plant-microbe interactions are crucial for improving plant productivity and plant protection. The latter aspect is particularly relevant for sustainable agriculture and for development of new preventive strategies against the spread of plant diseases. Constraint-based metabolic modeling is providing one of the possible ways to investigate the adaptation to different ecological niches and may give insight into the metabolic versatility of plant pathogenic bacteria. In this study, we present a fully curated, through high-throughput phenotype testing with Phenotype Microarrays™, a metabolic model of the emerging plant pathogenic bacterium *Pectobacterium parmentieri* SCC3193. The produced model includes: 1245 genes (covering approximately 28% of the all coding sequences in the genome), 2182 reactions, and 2080 metabolites. By means of flux balance analysis (FBA), we predict the metabolic adaptation to two different ecological niches, relevant for the persistence and the plant colonization by this bacterium: soil and rhizosphere. We also performed in silico gene deletions to predict the set of core genes essential for this bacterium to grow in laboratory minimal medium, soil and rhizosphere. The changes of environmental conditions modified the flux of about 10% of the reactions embedded in the model. Nearly 20% of genes were designated essential (251 in M9 medium, 250 in soil and 245 in the rhizosphere). Most of these essential genes were shared in the three environmental conditions. However, ten of these genes resulted fundamental in one or two conditions only, suggesting the presence of specific genes for metabolic adaptation. We anticipate that our metabolic model will be a valuable element for defining a set of metabolic targets to control infection and spreading of this plant pathogen and a scaffold to interpret future -omics datasets for this bacterium.

Introduction

Plant-bacteria interplay have been studied over long periods of time, mainly in terms of pathogenic and beneficial (symbiotic) interactions. Various details are now known concerning the molecular basis of such interactions[1]. For example, biological studies of plant pathogenic bacteria allowed understanding the modulation of bacterial recognition by the plants and revealed important aspects of plant immune responses[2]. Furthermore, additional investigations have confirmed that plant pathogenic bacteria exploit high flexibility in utilization of different kinds of sugar, nitrogen and phosphorus resources while adapting to the new environment, *e.g.* bacterial plant pathogen *Pseudomonas syringae* pv. *Tomato* specifically employs amino acid and sugar transporters to gain access to nutrients present in its environment. Subsequently, during infection processes of tomato plant, *P. syringae* pv. *Tomato* uses resources within the host, specifically from apoplast fluid[3].
In order to access the nutrients present in the plant tissues, pathogenic bacteria colonize, invade and, later on, establish chronic infections within host plants. During the infection process, they enter plant tissues either through wounds or natural openings and occupy the apoplast of tissues or the xylem, where they multiply and spread. Notably, phytopathogenic microorganisms cause damage and often impair plant growth and reproduction. On the other hand, to defend themselves against microbiological invasion, plants rely on two kinds of innate immunity, *i.e. via* pathogen triggered immunity (PTI)[4] and effector-triggered immunity (ETI).
Hence, to achieve a compatible interaction, microorganisms at first need to overcome plant's defenses that could abort the infection. Plant pathogenic bacteria, like *Pectobacterium parmentieri*, combine numerous strategies to accomplish that goal, *e.g.* they rely on quorum sensing system, hence the probability of weakening of plant defenses is higher due to high quantity of

bacteria in plant environment[5]. *P. parmentieri* is a pectinolytic, bacterium belonging to *Pectobacteriaceae* family (known as Soft Rot *Pectobacteriaceae*, SRP)[6]. It is a newly established species due to the transfer of all potato-originating isolates of *Pectobacterium wasabiae* to *P. parmentieri* sp. nov. (Ppa) on the basis of *in-silico* calculated DDH, gANI, and ANI values[7]. These Gram(-), rod-shaped bacteria are necrotrophs, that are able to destroy plant tissue components through the activity of PCWDE such as pectinases, cellulases, and proteases, further secreted via Type I or II secretion system[8,9]. Nonetheless, bacteria belonging to this species can exhibit a different level of activities of above-mentioned enzymes[10]. Pectinolytic bacteria, and among them *P. parmentieri*, require favorable environmental conditions to cause disease symptoms; however, they can reside inside plant tissues as endophytes for a long time[11]. Bacteria from the genus *Pectobacterium* are causative agents of soft rot in economically important crops such as potato, tomato or maize. Also, they are responsible for the blackleg disease, so far reported only on potato plants[11,12]. It is worth to mention that bacteria from the genus *Pectobacterium* have been included among 10 most important bacterial plant pathogens based on their economic impact[13] since crop losses caused by phytopathogenic microorganisms can reach up to 20% of total yield[14]. There is still limited knowledge regarding cascade of genes being expressed before and during infection process in *P. parmentieri*. It was previously reported, that initialization of infection progress is controlled by quorum sensing in closely related *Pectobacterium atrosepticum*[15]. Moreover, massive production of butanediol during plant infection by bacteria of the genera *Dickeya* and *Pectobacterium* was reported[16]. However, metabolic pathways important for promoting bacterial multiplication before and during plant infection connected with carbon and other compounds utilization in *P. parmentieri* have received little attention so far. Given the complexity of bacterial-host relationships, they cannot be adequately investigated only by means of classical microbiological and molecular methods; rather, the coupled use of Phenotype Microarrays™, computational, large-scale and systemic frameworks are advisable. Metabolic modeling is now a promising way to interpret puzzling, heterogeneous bacterial phenotypes[17], especially those related to bacterial-host interaction[18]. Constraints-based approaches, and in particular Flux Balance Analysis (FBA)[19] have been shown to be accordingly predict growth phenotypes and are claimed to provide a systems biology view on multi-omics data, possibly allowing to predict physiological changes and evolution of bacterial populations[20,21]. Recently, genome-scale metabolic model (GEM) reconstruction and FBA have been used for deciphering the metabolic adaption of environmental microbes following ecological parameters variation[22], ecological niche shift[23], as well as for providing insights into the metabolic adaptation in human and bacterial plant pathogens[24,25]. To the best of our knowledge only in a few cases, FBA has been applied in understanding the metabolic adaptation of specific plant bacterial pathogens *e.g.* studies performed on *R. solanacearum* showed that trade-off between virulence factor production and bacterial proliferation is controlled by the quorum-sensing-dependent regulatory protein PhcA[26].

The aim of this work is the reconstruction of a highly curated metabolic model of the plant pathogenic bacterium *P. parmentieri* SCC3193 and the usage of this model to putatively identify the metabolic pathways relevant for *P. parmentieri* fitness in two different ecological niches, soil and rhizosphere. We show that niche switching may lead to a metabolic reassessment in carbon-related pathways in *P. parmentieri* SCC3193 and we spot the core-set of essential genes in the two examined conditions. Moreover, we anticipate that the model itself will represent a valuable element which will pave the way to both, knowledge-base of strain's biology and novel, applied technologies, like genetic engineering and synthetic biology experiments.

Results

Reconstruction and validation of *P. parmentieri* SCC3193 genome-scale metabolic model.

An initial draft model of *P. parmentieri* SCC3193 was obtained through the KBase server (http://kbase.us) then it has been reviewed and manually curated as described in the methods. The model was augmented by mean of orthologous gene search in closely related *E. coli* strains. Following this approach, we hence expanded our reconstruction by adding 93 genes and 383 reactions from *E. coli* model[27]. Most of the added genes encode for transport reactions taking place in the membrane, where compounds are often modified and then transferred into the cytosol. This led to the necessity to add another compartment in *P. parmentieri* model, namely the periplasm, which improved descriptive model capability.

Model validation was carried-out by exploiting different approaches. Firstly, a comparison model's performance to physiological data obtained in high throughput experiment with the use of PM Biolog plates was performed. FBA has been employed to test if the model could accurately predict the ability of *P. parmentieri* to produce biomass on utilized carbon sources. In particular, we tested 91 compounds, previously used in PM experiment and embedded in the metabolic network. The model was hence first gap-filled by iteratively correcting the inconsistencies between *in silico* predictions and PM outcomes. Where possible, missing reactions and transporters were added according to indications obtained from functional databases (MetaNetX[28], Bigg[29], Seed[30], KEGG[31] and transportDB[32]) otherwise, gap-filling genes or reactions have been added and named after the missing components. Precisely, in twenty-two cases, the initial model's prediction was refined thanks to experimental validations (further details on the predictions of metabolites usage before and after this stage are given in Supplementary Data SD5). The refined model displayed agreement with PM in 83 out of 91 tested carbon substrates. As summarized in Figure 1, sensitivity, specificity, precision, accuracy, negative predictive value and F-score (calculated as described in[33], see materials and methods)

reached very high scores, suggesting a good reliability of the model. The final *P. parmentieri* model was termed iLP1245 in accordance with the nomenclature standard[34], and includes 1245 genes (covering 28% of the total number of coding sequences in the genome, 4449), 2182 reactions, and 2080 metabolites. A description of the model and of the genetic features captured within are reported in Supplementary Information SI1.The Systems Biology Markup Language (SBML) file of the model was validated by the online SBML validator tool (http://sbml.org/Facilities/Validator/) and is available online as Supplementary Information SI3. All the metabolites embedded in the model were annotated by using identifiers.org and MIRIAM[35] registry in order to facilitate model reuse and search strategies, by providing unique, unambiguous, perennial, standard-compliant and directly resolvable identifiers.

Phenotypic characterization of *P. parmentieri* SCC3193 and metabolic model deep validation

Phenotypical profiling with the use of Biolog Plates PM1, PM2A, PM3 and PM4 revealed that *P. parmentieri* SCC3193 is able to utilize all common sugar components of plant cell walls at high levels *e.g.*: sucrose, tartaric acid, D-cellobiose, stachyose and most importantly pectin (Poly(1,4-alpha-D-galacturonide)) (SD3). Model's predictions (SD5) are in agreement with in vivo obtained results. Precisely, this bacterium was very effectively exploiting D-Glucosamine and its derivatives (N-Acetyl-D-Glucosamine, N-Acetyl-D-Galactosamine) together with xanthine. Intriguingly, 70.5% of phosphorus and sulfur sources used in experiment were effectively metabolized by *P. parmentieri* SCC3193 (SD1).
Cross-check of PM Microarray obtained results was achieved by applying EnVisionTM analyses and measuring bacterial mass after 24 h of incubation in M9 media supplemented with randomly selected carbon sources: alpha-D-glucose, D-trehalose, and D-xylose. Data obtained in the nutrients metabolization assay enabled to determine growth curves of *P. parmentieri* SCC3193 in M9 with three different carbon sources: alpha-D-glucose, D-trehalose, and D-xylose (Figure 2). Concerning bacterial growth in alpha-D-glucose, we observe rapid logarithmic phase and plateau of bacterial culture (Figure 2). Interestingly, in case of D-trehalose and D-xylose a longer lag phase occurs followed by an intermediate logarithmic phase. Comparison of EnVision^{TM} experiment results to models' predictions revealed agreement in case of all 3 tested carbon sources (Figure 2).

Differential metabolic adaptation to soil and rhizosphere environment

By analyzing flux changes in response to simulated environmental conditions (*e.g.* M9, soil, rhizosphere) it is possible to observe whether any significant metabolic rewiring occurs. Loopless FBA and loopless FVA (see materials and methods) were used in order to estimate niche-specific metabolic adaptations in soil and rhizosphere. The results obtained through FBA were cross-checked with FVA (see Supplementary Data SD3 for results). Possible fluxes variation occurring during the transition from soil to rhizosphere was interpreted as a metabolic shift between the two niches. The magnitude of the variation was described according to different cutoff, *i.e.* as a variation greater than 10%, 20%, 30%, 40% and 50% of the initial flux value. Results indicate that the number of the reactions whose fluxes changed significantly is slightly affected by the stringency of the cutoff, vouching for robustness of the analysis. Accordingly to the previous report[23], we focused only on the results for which variation is higher than 50% in rhizosphere compared to soil. Finally, the total amount of reactions taken into account was 208, corresponding to $\tilde{1}$0% of those embedded in the reconstruction. These were further classified as reactions whose flux increase or decrease and reactions turned on or off during the environmental stress (see Figure 3). Based on that classification, it appears that several pathways keep being active in both examined conditions, though recruiting a different set of reactions. The main shift seems to occur in sugars metabolic pathways, tracking a switch from pentose phosphate and hexose to amino sugars metabolism. Also, some peculiar systems seem to be turned on in concomitance of such niche change, as nitrogen, butanoate, galactose, and propanoate metabolism, and biosynthetic pathways, including those of steroids and folate, attesting a specific adaptive response of *P. parmentieri* SCC3193 to this environmental (nutritional) switch.

In silico genes deletion provides insight into fitness relevance of metabolic modules

Simulation of single genes deletion is a very powerful *in silico* method, as it not only enables to estimate gene's knock-out fatality but also, allows predictions on a genome-scale. We performed such test on iLP1245 with two different approach (see materials and methods), FBA and MOMA. As shown in Figure 4, both methods labeled the same number of genes as essential, according to the three media examined, hence from now on we will refer only to FBA outcomes.

Specifically, the number of essential genes found is 251 in M9 medium, 250 in soil and 245 in the rhizosphere. The locus tag of such genes, alongside with their corresponding encoded protein, are reported in Supplementary information SI4. The Venn's diagram in Figure 5 shows the overlap among the three different conditions. Results indicate that, a huge core of genes is likely to be mandatory in all the *in silico* tested conditions, while only a small number stand-out as essential just in one or two out of the three media. Notably, in *in silico* gene deletions predicted the gene W5S-RS13875 encoding for DNA starvation/stationary phase protection protein (WP-014700482.1) as specifically required during growth in soil and rhizosphere, the gene W5S-RS15765 encoding for an ammonium transporter (WP-014700821.1) essential in M9 and rhizosphere, while eight genes appear to be specifically essential during growth in soil and M9 but not in rhizosphere.

Seven of them belong to the thiamine and sulfur metabolism pathway: W5S-RS00965 cystathionine gamma-synthase, WP-014698476.1), W5S-RS01140 (thiazole synthase ThiG, WP-012822036.1), W5S-RS01145 (sulfur carrier protein ThiS, WP-014698484.1), W5S-RS01150 (adenylyltransferase ThiF, WP-014698485.1), W5S-RS01155 (thiamine phosphate synthase ThiE, WP-014698486.1), W5S-RS01160 (phosphomethylpyrimidine synthase ThiC, WP-014698487.1), W5S-RS05940 (hydroxymethylpyrimidine/phosphomethylpyrimidine kinase ThiD, WP-014698998.1), while the last one, W5S-RS18250 (diaminopimelate decarboxylase, WP-043899153.1) is an enzyme involved in secondary metabolite production. Interestingly, the simulated rhizosphere growth medium contains thiamine which, conversely, is absent in soil and M9. Moreover, three condition-specific genes have been predicted, one during growth in M9 medium, W5S-RS19605 encoding for class II fructose-bisphosphate aldolase (WP-014701539.1) and two during growth in the rhizosphere, W5S-RS19610 and W5S-RS06525, respectively encoding for a phosphoglycerate kinase (WP-005973111.1) and a long-chain fatty acid transporter (WP-014699104.1).

Discussion

Plant-microbe interactions have been under intensive investigation in recent years, regardless of the nature of the interaction itself: pathogenic or symbiotic. Severity to understand this interplay is connected with the complexity of the environment in which bacteria persist: soil, rhizosphere or plant tissues. Metabolic modeling allows predicting and examining biochemical reactions involved in adaptation to above-mentioned ecological niches as well as predicting the phenotypic outcomes of gene deletions. In this paper, for the first time we report a high throughput experimental validation on metabolic capability of *P. parmentieri* SCC3193, and a highly curated metabolic model of such plant pathogenic bacterium (iLPI245). iLPI245 is highly reliable, as *in silico* obtained results overlap in 91% with experimentally obtained data on carbon utilization phenotypes, a value that perfectly fits with the currently accepted standard for genome-scale metabolic reconstructions[36,37]. For example, a previously described genome-scale metabolic reconstruction of *Pectobacterium aroidearum* PC1, a monocotyledonous plant pathogenic bacterium, showed the agreement of 80.4% between *in silico* simulations and Phenotype Microarray™ (Biolog) experiments[25]. This difference in accuracy of the models was probably related to the fact, that the metabolic model of *P. aroidearum* PC1 was constructed on the template of the latest version of *E. coli* K12 MG1655 metabolic model iJO1366, whereas in our case *E. coli* genome was used only for biomass estimation, and identification of orthologues genes, in consequence producing a more *P. parmentieri*-specific model. The iLP1245 model of *P. parmentieri* SCC3193 was then used to assess the relative importance of single metabolic pathways and genes in adaptation to growth under laboratory (M9 medium) and *in silico* simulated field conditions (soil and rhizosphere). Soil and rhizosphere conditions were chosen since they represent two key environments were *P. parmentieri* is abundant, and for which the information on chemical composition is readily available and reliable[23]. While performing FBA on *P. parmentieri* SCC3193, we observed a shift in sugars metabolic pathways while conditions were changed from soil to rhizosphere. Namely, switch from pentose phosphate and hexose to amino sugars metabolism, which are supposed to be more abundant in rhizosphere[38]. We can consequently suggest, that adapting to rhizosphere environment involves utilization of this latter carbon compounds. Interestingly, a down-regulation of genes expression involved in amino sugars and nucleotide sugar metabolism has been associated with starvation stress response in *P. atrosepticum*[39]. Then, our prediction of metabolic fluxes implies, that rhizosphere represents a rich environment for *P. parmentieri*, where these bacteria can thrive in. Moreover, butanoate, propanoate, steroid metabolism and, folate biosynthetic pathways are turned on while we *in silico* shift environment from soil to the rhizosphere. These compounds are precursors for volatiles compounds (VOCs) production or are comprised in VOC metabolic pathways; *e. g.* butanoate is an ester of butyric acid, which is among the most frequently secreted compounds[16,40]. We can hypothesize that interbacterial communication and possibly plant-bacteria interaction could be mediated by VOCs, which may then have a role in pathogenesis and later on, in developing strategies for biocontrol. All those volatiles are strictly connected either with encountering other bacteria growing in rhizosphere or with virulence of plant pathogenic bacteria[16,40]. These data were compared with those obtained from the plant symbiotic nitrogen-fixing bacterium *Sinorhizobium meliloti*[23]. We observed that only 10% of reactions changed flux (at least by 50%) in *P. parmentieri*, while the shift from soil to rhizosphere conditions changed the flux of more than 20% of reactions (including reactions which reversed direction) in *S. meliloti*, though the number of reactions present in the two models (iLP1245 and iGD1575) was similar (2182 and 1.825 reactions, respectively). Moreover, *in S. meliloti* 13% of active reactions were specific to just one of the environments, while only the 5.3% of active reactions were environment specific in *P. parmentieri*. We can hypothesize that the smaller and more compact genome of *P. parmentieri* compared to *S. meliloti* (4449 vs. 6204 protein-coding genes for *P. parmentieri* and *S. meliloti*, respectively, including a multipartite genome organization in the latter species) allows a reduced metabolic redundancy in *P. parmentieri* compared to *S. meliloti*[41] and a more generalist *vs.* specialist metabolic network (*i.e.* most reactions are not changing while environment fluctuates). To test this hypothesis, we performed MOMA and FBA simulations of gene deletions (see Supplementary Data SD6). Both analyses revealed that *P. parmentieri* SCC3193 possesses an essential gene core composed of 241 metabolic genes, for growth in both soil and the rhizosphere. Here, a set of 8 genes only was found to be essential in soil but not in rhizosphere. This is in large contrast to *S.*

meliloti, where 66 genes were found as essential for growth in the same simulated rhizosphere environment23, and supports the previously proposed hypothesis of a robust and versatile metabolic network of *P. parmentieri* SCC3193, which may allow the strain to rapidly accommodate with relevant changes in environmental nutrient sources. Additionally, we can assume that compact core of essential genes important for *P. parmentieri* is a reason why bacteria from this species are able to persist on plant residuals without interacting with host plant (potato) for long periods of time, and as well are cosmopolites in the environment[12]. To summarize our findings, we established a functional metabolic model of plant pathogenic bacterium *P. parmentieri* SCC3193, showing an inherent robustness of the metabolic network, which can easily accommodate for nutrient variability when moving from soil to rhizosphere growth conditions. Such robustness may imply that several, still unknown, plant species (together with their soil and rhizosphere) can be a reservoir for this pathogenic bacterium.

Methods

Metabolic network reconstruction and model refinement

A draft metabolic model was build using the KBase Narrative Interface (www.kbase.com), later it was expanded based on the comparison of identical functions (orthologs genes) in the closely related strain *Escherichia coli* K-12 MG1655 (through a Bidirectional Best Hit approach, inParanoid[42]), for which a reliable metabolic reconstruction is available[27]. Further model refinement was then performed through experimental validation, *i.e.* by iteratively comparing model's growth outcomes to the real strain's growth during Phenotype Microarrays experiment (Omnilog^{TM}). Additional informations on such validation step is reported are Supplementary Information SI1.

Metabolic modeling

Metabolic modeling was performed using COBRApy Toolbox version 0.6.129 and the Gurobi 7.0.2 solver (www.gurobi.com). Scripts enabling to run all the *in silico* analysis performed in this work are available in Supplementary Information SI2. No comprehensive description of the macromolecular composition of the *P. parmentieri* biomass is available in the literature. However, such data are available for *E. coli*[27], therefore we approximated the *P. parmentieri* gross biomass composition to that one of this closely related species. The complete biomass composition is given in Supplementary Information SI1. The biomass reaction was set as the objective function for growth in all the media used in this work.

Bacterial strains and culture conditions

Bacterial strain used in this study is *P. parmentieri* reference strain SCC3193 isolated from potato tuber in Finland[7,43]. For high-throughput phenotypic characterization, bacteria were grown on TSA medium at 28 C for 24 h. For EnVision^{TM} experiment bacteria were first grown in LB at 28 C for 24 h with constant agitation (120 RPM), later on in M9 for 24 h with constant agitation (130 RPM).

Experimental high-throughput phenotypic characterization on *P. parmentieri* SCC3193.

Biolog Plates PM1, PM2A, PM3 and PM4 were utilized for high-throughput phenotypic characterization of *P. parmentieri* SCC3193. Overnight bacterial culture was transferred from TSA medium to 5 ml of 0.85% NaCl and bacterial suspension was adjusted to OD600 equaling 0.1 Later on, 1 ml of bacterial suspension was transferred to 11 ml of Minimal Salts medium (M9-C: 0.6% Na2HPO4, 0.3% KH2PO4, 0.05% NaCl, NH4Cl, 0.005% Yeast Extract) supplemented with 120 ul of Biolog A dye. To inoculate wells in PM plates 100 ul of described bacterial suspension was used. The measurement was carried out in OmniLog^{TM} for 46h. Results were analyzed with DuctApe[44]. All the results are reported in Supplementary Data SD1. Nutrients metabolization assay with the use of EnVision^{TM} plate reader was performed to cross-check high-throughput phenotypic characterization. M9 media supplemented with 20% of selected carbon sources: alpha-D-glucose, D-xylose, and D-trehalose were prepared. *P. parmentieri* SCC3193 was grown in LB medium overnight at 28C with constant agitation (120 RPM). Afterwards, overnight bacterial cultures were centrifuged and washed twice in sterile Ringer Buffer and later OD600 of inoculum was adjusted to 0.1. To establish growth curves 50 ul of inoculum was transferred to 450 ul of M9 supplemented with different carbon sources. Bacteria were cultured with agitation set at 120 RPM in 28C in 24-well plate in EnVision^{TM} plate reader. Optical density measurement at 600 nm was performed every 30 min.

OmniLog^{TM} data processing and analysis

PM1 and PM2A obtained data analysis was performed with DuctApe[44]. Activity index (AV) values were calculated following subtraction of the value obtained for blank well from that of inoculated wells, whereas plots of the growth curves are of the unblanked data. Bacterial growth with each compound was considered positive if the AV value was ≥ 3. Growth phenotypes were defined as negative if the AV value was ≤ 2, and following a manual inspection of the unblanked curves. All the results are reported in Supplementary Data SD2.

In silico environmental representations

In silico representations of the nutritional composition of the soil, and rhizosphere were derived from a previously published paper[23]. The composition of M9, soil, and rhizosphere media is reported in Table 1.

Flux Balance Analysis and Flux Variability Analysis

Flux distribution predictions were assessed by performing FBA in M9, soil and rhizosphere media. Moreover, to avoid *in silico* artifacts, such as loops (which looks unreal to happen *in vivo*), we performed loopless-FBA, which instead identifies the closest (to the reference) thermodynamically consistent flux state avoiding loops. Moreover, since FBA only predicts one flux distribution among all possible solutions, we also performed loopless-FVA (Flux Variability Analysis) and filtered out those reactions whose fluxes were outside the following criteria:

(1)

$$if fFBA < 0 fFVA, min 1.2 * fFBA \wedge fFVA, max 0.8 * fFBA$$

(2)

$$if fFBA > 0 fFVA, min 0.8 * fFBA \wedge fFVA, max 1.2 * fFBA$$

Loopless FBA and FVA predictions for each reaction can be retrieved in Supplementary data SD3 or by running Supplementary information SI2.

Model's predictive value estimation

M9 growth medium was simulated *in silico*, *i.e.* by constraining the lower bound of import reactions for each of the compounds present in the medium (as reported in Table 1). When reproducing PM experiment in silico, model's performances were considered as true positives (TP) if growth was obtained both *in silico* and *in vivo*, true negatives (TN) in case of non-growth both *in silico* and *in vivo*, false positives (FP) if growth was obtained *in silico* but not *in vivo* and false negatives (FN) if *vice versa*. Reliability of the obtained predictions was then estimated according to the following parameters:

(3)

$$Sensitivity = TP/(TP + FN)$$

(4)

$$Specificity = TN/(TN + FP)$$

(5)

$$Precision(PPV) = TP/(TP + FP$$

(6)

$$Negativepredictedvalue(NPV) = TN/(TN + FN)$$

(7)

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

(8)

$$F - score = 2(precision * sensitivity)/(precision + sensitivity)$$

Single gene deletion analysis

Using genome-scale metabolic networks (GEMs) genes knockout can be simulated in order to identify those genes whose removal is likely to impair the organism's growth. Specifically, it is possible to simulate mutants by deleting each gene included in the metabolic reconstruction and testing the predicted effects on the microbe's growth. Through this strategy, it is possible to calculate the growth ratio (GRratio) between the growth rate of the mutant model (uKO) and the one of the wild-type (uWT) as:

(9)

$$GRratio = uKO/uWT$$

This measure can provide hints on the essentiality of the knocked out gene. In particular, in our work the knocked out gene was considered essential when $GRratio < 0.9$. Furthermore, when $0.9 < GRratio < 1 and GRratio = 1$ we considered the knocked out gene as semi-essential and essential, respectively. Minimization of Metabolic Adjustment (MOMA[45]) and FBA were the algorithms chosen to perform such analyses.

COG analyses

The WebMGA webserver33 was used to provide functional Cluster of Orthologous Genes (COG) annotations (p-value cutoff of 0.001) to each gene in the model. The COG annotation for each gene associated with variable reactions during the transition between two niches was extracted from the WebMGA server. Biases were determined after standardizing by the number of genes in each class of variable genes. Statistical significance was determined using Pearson's Chi-squared tests. The complete list of COG annotations is available as Supplementary Data SD4.

References

1. Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.* **57**, 233–266 (2006).
2. Cui, H., Xiang, T. & Zhou, J.-M. Plant immunity: a lesson from pathogenic bacterial effector proteins. *Cell. microbiology* **11**, 1453–1461 (2009).
3. Rico, A. & Preston, G. M. Pseudomonas syringae pv. tomato dc3000 uses constitutive and apoplast-induced nutrient assimilation pathways to catabolize nutrients that are abundant in the tomato apoplast. *Mol. plant-microbe interactions* **21**, 269–282 (2008).
4. Chisholm, S. T., Coaker, G., Day, B. & Staskawicz, B. J. Host-microbe interactions: shaping the evolution of the plant immune response. *Cell* **124**, 803–814 (2006).
5. Põllumaa, L., Alamäe, T. & Mäe, A. Quorum sensing and expression of virulence in pectobacteria. *Sensors* **12**, 3327–3349 (2012).
6. Adeolu, M., Alnajar, S., Naushad, S. & Gupta, R. S. Genome-based phylogeny and taxonomy of the 'enterobacteriales': proposal for enterobacterales ord. nov. divided into the families enterobacteriaceae, erwiniaceae fam. nov., pectobacteriaceae fam. nov., yersiniaceae fam. nov., hafniaceae fam. nov., morganellaceae fam. nov., and budviciaceae fam. nov. *Int. journal systematic evolutionary microbiology* **66**, 5575–5599 (2016).
7. Khayi, S. *et al.* Transfer of the potato plant isolates of pectobacterium wasabiae to pectobacterium parmentieri sp. nov. *Int. journal systematic evolutionary microbiology* **66**, 5379–5383 (2016).
8. Charkowski, A. *et al.* The role of secretion systems and small molecules in soft-rot enterobacteriaceae pathogenicity. *Annu. review phytopathology* **50**, 425–449 (2012).
9. Ma, B. *et al.* Host range and molecular phylogenies of the soft rot enterobacterial genera pectobacterium and dickeya. *Phytopathol.* **97**, 1150–1163 (2007).
10. Zoledowska, S., Motyka, A., Zukowska, D., Sledz, W. & Lojkowska, E. Population structure and biodiversity of pectobacterium parmentieri isolated from potato fields in temperate climate. *Plant Dis.* (2017).
11. Pérombelon, M. Potato diseases caused by soft rot erwinias: an overview of pathogenesis. *Plant Pathol.* **51**, 1–12 (2002).
12. Toth, I. K., Bell, K. S., Holeva, M. C. & Birch, P. R. Soft rot erwiniae: from genes to genomes. *Mol. plant pathology* **4**, 17–30 (2003).
13. Mansfield, J. *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. plant pathology* **13**, 614–629 (2012).
14. Czajkowski, R., Perombelon, M. C., van Veen, J. A. & van der Wolf, J. M. Control of blackleg and tuber soft rot of potato caused by pectobacterium and dickeya species: a review. *Plant pathology* **60**, 999–1013 (2011).
15. Liu, H. *et al.* Quorum sensing coordinates brute force and stealth modes of infection in the plant pathogen pectobacterium atrosepticum. *PLoS pathogens* **4**, e1000093 (2008).
16. Effantin, G., Rivasseau, C., Gromova, M., Bligny, R. & Hugouvieux-Cotte-Pattat, N. Massive production of butanediol during plant infection by phytopathogenic bacteria of the genera dickeya and pectobacterium. *Mol. microbiology* **82**, 988–997 (2011).
17. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. systems biology* **5**, 320 (2009).
18. Duan, G., Christian, N., Schwachtje, J., Walther, D. & Ebenhöh, O. The metabolic interplay between plants and phytopathogens. *Metab.* **3**, 1–23 (2013).
19. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. biotechnology* **28**, 245–248 (2010).

20. Großkopf, T. *et al.* Metabolic modelling in a dynamic evolutionary framework predicts adaptive diversification of bacteria in a long-term evolution experiment. *BMC evolutionary biology* **16**, 163 (2016).

21. Angione, C. & Lió, P. Predictive analytics of environmental adaptability in multi-omic network models. *Sci. reports* **5** (2015).

22. Giudice, A. L. & Fani, R. Antimicrobial potential of cold-adapted bacteria and fungi from polar regions. In *Biotechnology of Extremophiles:*, 83–115 (Springer, 2016).

23. Checcucci, A. *et al.* Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in sinorhizobium meliloti. *Nat. communications* **7** (2016).

24. Mithani, A., Hein, J. & Preston, G. M. Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in pseudomonas. *Mol. Biol. Evol.* **28**, 483–499 (2010).

25. Wang, C. *et al.* Construction of a genome-scale metabolic network of the plant pathogen pectobacterium carotovorum provides new strategies for bactericide discovery. *FEBS letters* **589**, 285–294 (2015).

26. Peyraud, R., Cottret, L., Marmiesse, L., Gouzy, J. & Genin, S. A resource allocation trade-off between virulence and proliferation drives metabolic versatility in the plant pathogen ralstonia solanacearum. *PLoS pathogens* **12**, e1005939 (2016).

27. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol. systems biology* **3**, 121 (2007).

28. Moretti, S. *et al.* Metanetx/mnxref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic acids research* **44**, D523–D526 (2016).

29. King, Z. A. *et al.* Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research* **44**, D515–D522 (2015).

30. Overbeek, R. *et al.* The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research* **42**, D206–D214 (2013).

31. Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).

32. Ren, Q., Chen, K. & Paulsen, I. T. Transportdb: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic acids research* **35**, D274–D279 (2006).

33. Mendoza, S. N., Cañón, P. M., Contreras, A., Ribbeck, M. & Agosin, E. Genome-scale reconstruction of the metabolic network in oenococcus oeni to assess wine malolactic fermentation. *Front. microbiology* **8** (2017).

34. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of escherichia coli k-12 (i jr904 gsm/gpr). *Genome biology* **4**, R54 (2003).

35. Juty, N., Le Novère, N. & Laibe, C. Identifiers. org and miriam registry: community resources to provide persistent identification. *Nucleic acids research* **40**, D580–D586 (2011).

36. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. protocols* **5**, 93–121 (2010).

37. Chan, S. H., Cai, J., Wang, L., Simons-Senftle, M. N. & Maranas, C. D. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinforma.* (2017).

38. Sørensen, J., Jensen, L. E. & Nybroe, O. Soil and rhizosphere as habitats for pseudomonas inoculants: new knowledge on distribution, activity and physiological state derived from micro-scale and single-cell studies. *Plant soil* **232**, 97–108 (2001).

39. Gorshkov, V. *et al.* Global gene expression analysis of cross-protected phenotype of pectobacterium atrosepticum. *PloS one* **12**, e0169536 (2017).

40. Kanchiswamy, C. N., Malnoy, M. & Maffei, M. E. Chemical diversity of microbial volatiles and their potential for plant growth and productivity. *Front. plant science* **6** (2015).

41. Finan, T. M. *et al.* Genetic redundancy is prevalent within the 6.7 mb sinorhizobium meliloti genome. *Mol. genetics genomics* **290**, 1345–1356 (2015).

42. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. molecular biology* **314**, 1041–1052 (2001).

43. Nykyri, J. *et al.* Revised phylogeny and novel horizontally acquired virulence determinants of the model soft rot phytopathogen pectobacterium wasabiae scc3193. *PLoS Pathog.* **8**, e1003013 (2012).

44. Galardini, M. *et al.* Ductape: a suite for the analysis and correlation of genomic and omnilog^{TM} phenotype microarray data. *Genomics* **103**, 1–10 (2014).

45. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.* **99**, 15112–15117 (2002).

Acknowledgements

This work was supported by the National Science Centre in Poland via grant no. 2014/14/M/NZ8/00501 awarded to EL and by the University of Gdansk in Poland via grant no. 538-M031-B187-16 awarded to SZ.

Author contributions statement

Must include all authors, identified by initials, for example: A.M. conceived the experiment(s), L.P. reconstructed the model and performed the computational experiments, S.D. and F.D. conducted the wet-lab experiments, M.F., A.M., L.P. and S.Z. analysed the results. All authors reviewed the manuscript.

Additional information

The authors declare that they have no competing interests.

1 SUPPLEMENTARY FILES

Supplementary Information 1. Includes the Supplementary Text, Supplementary Tables, and Supplementary Figures. File type: PDF document **Supplementary Information 2**. Script enabling to perform all the analysis in the manuscript **Supplementary Information 3**. The sbml file of the model. File type: XML formatted file **Supplementary Information 4**. **Supplementary Data 1**. All Phenotype MicroArrayTM data generated in this study and plots representing results from Phenotype MicroArrayTM analysis with the use of DuctApe program. File type: Compressed zip archive **Supplementary Data 2**. The raw Phenotype MicroArrayTM data, in the form of .csv files, obtained in this study. File type: Compressed zip archive **Supplementary Data 3**. Reactions showing different characteristics during growth in soil versus the rhizosphere. File type: Excel document **Supplementary Data 4**. The COG annotations for all genes included in the model, as generated by WebMGA. File type: Excel document **Supplementary Data 5**. Prediction on usage of metabolites before and after gap filling process. File type: Excel document **Supplementary Data 6**. All essential genes found with their encoded protein and function.

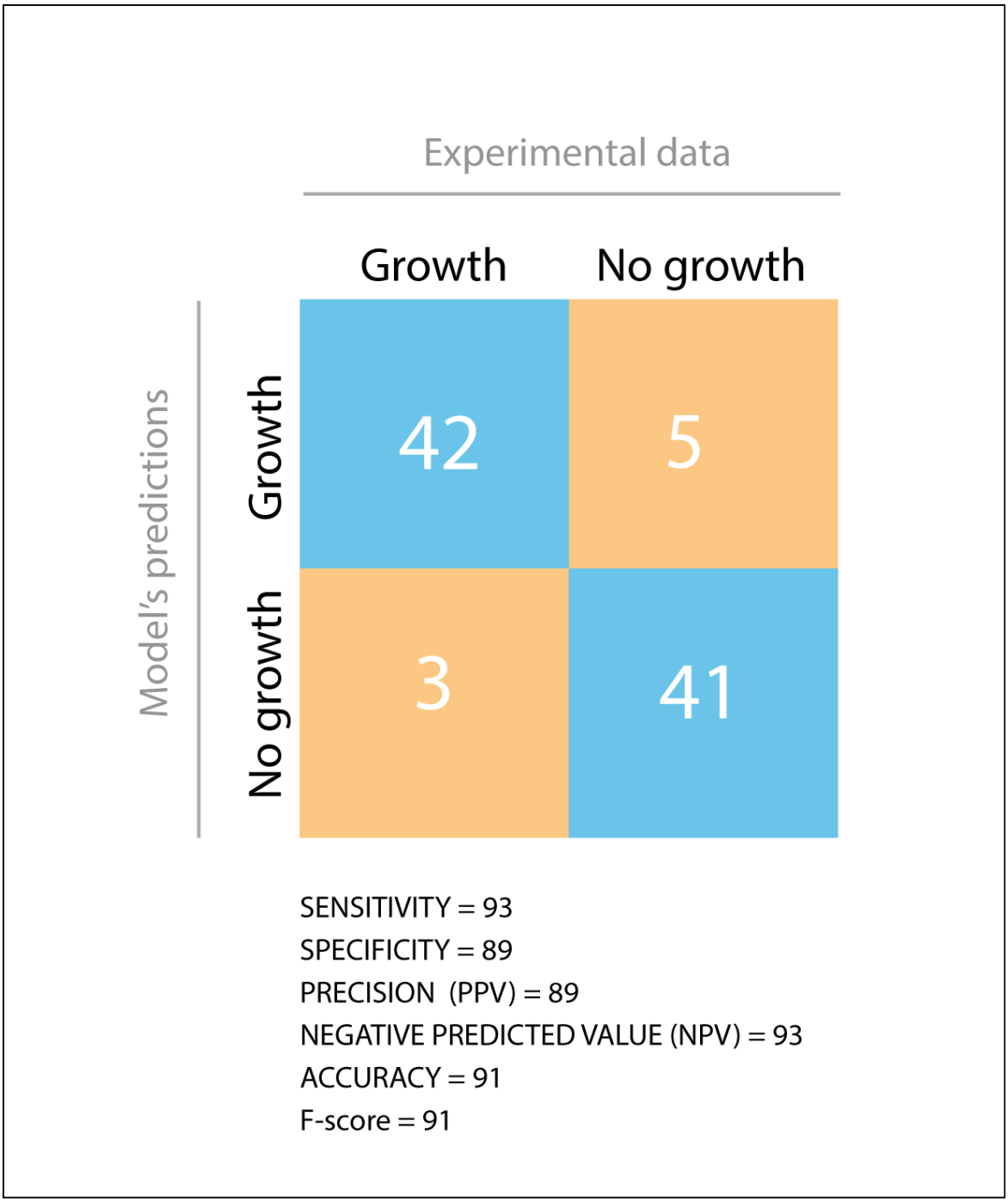


Figure 1. Comparison between Phenotype Microarray data and model's predictions. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. Statistical parameters were calculated as described in materials and methods.

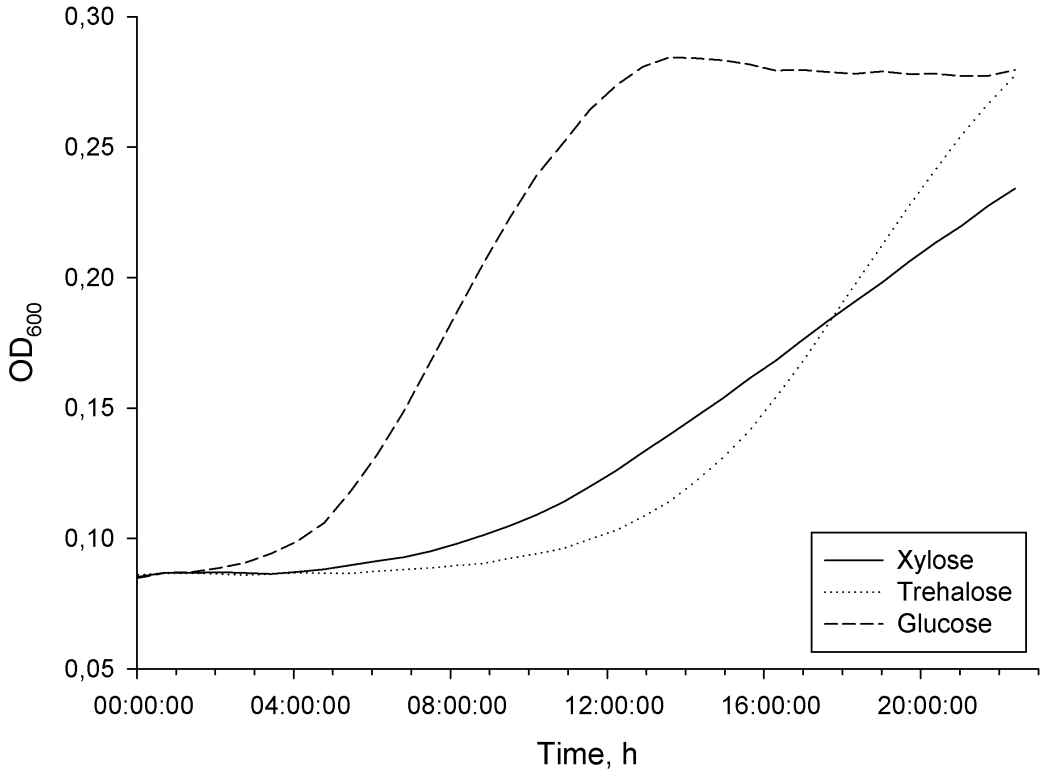


Figure 2. Growth curves of *P. parmentieri* SCC3193 in M9 supplemented with different carbon sources.

Figure 3. Biochemical reactions which fluxes change, while bacterial transition from soil to the rhizosphere.

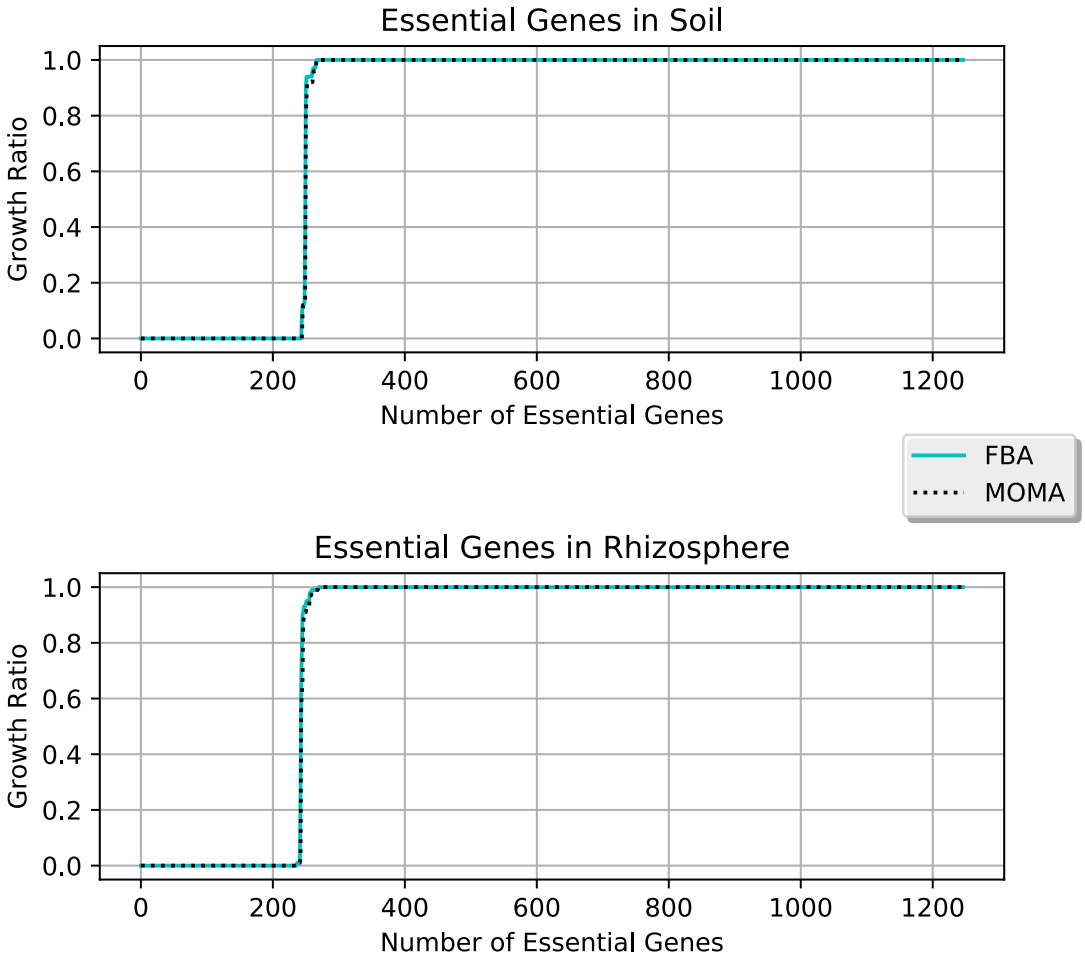


Figure 4. GRratio value for each gene deletion in rhizosphere and soil media according to FBA and MOMA predictions.

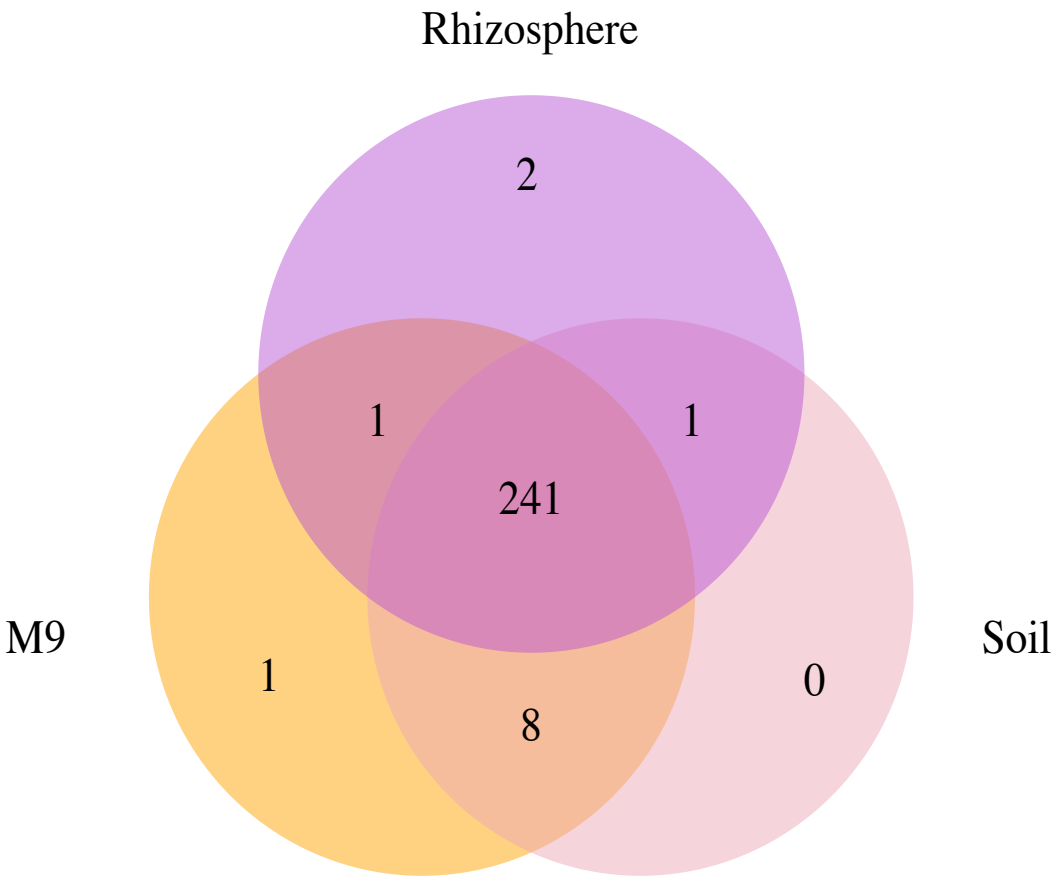


Figure 5. Venn's diagram showing the amount of shared/unique essential genes for each of the examined conditions (M9, rhizosphere, and soil).

9 Conclusions

The last two decades delivered the great opportunity to study the biology of organisms at unprecedented level. Thanks to technological and experimental progresses, the molecular basis of biological activity have been deeply investigated and finally unraveled in several fields, under the common denominator of genomics. Comparative approaches and evolutionary techniques have been developed to examine the genetic relationships and the structural landmarks among different groups of organisms. The challenge of the -OMICS era is now to integrate such informations in a functional contest. This challenge, still ongoing, relies on the possibility

1. to reconstruct the genome sequences from the reads obtained on the various sequencing platforms;

2. to identify genes inside these sequences and assign functions;

3. to predict organism's phenotypes (based on functional annotations).

The overall content of this thesis aims to describe the logical progression between these three tasks. The results herein reported, beside referring to various fields of genomics, all intend to generate or use genomic data to draw predictions of functional abilities (at different depth) on the examined subjects. Particularly, this scope can be synthesized by the global genome-scale metabolic model tools, which combine in a single framework the several encountered topics. GEMs take the described predictive challenges: they computationally infer the biological properties of an organism on the simple basis of its genome sequence and return functional, systems clues. However, although being a synthesis, they do not represent the conclusive step of research. Conversely, they embody the starting point of new experimental loops, where molecular aspects and overall mechanisms finally met, in a system-wide fashion.
Such novelties possibly represent the future of research progress but still need a huge tuning in order to improve the method. In particular, one of the main stumbling-block is the gap existing between genome sequence data (which are growing at explosive rates) and the experimental-based knowledge about

genes function (which is collected slowly). Metabolic modeling can assist the process linking biological functions to ORFs by generating hypotheses and biological reasoning, but it is not sufficient. The body of experimental data must be increased to achieve accurate functional annotations. The generation of new high-throughput experimental platforms to provide these data would accelerate the process, like the setting-up of sequencing machineries made the pace of scientific discoveries during the last twenty years.

Still, it looks like the integration of experimental evidences and computational solutions will be the main key-driver of research progress, gently molding biology in a quantitative science. -OMICS can provide data for thousands of cellular components across a variety of scales; GEMs can reduce dimensionality across the wide spectrum. However, that achievement requires to find solutions at technical issues. For example, the integration of RNA-seq data faces unique challenges like experimental and inherent biological noise, variation among experimental platforms, detection bias, and the unclear relationship between gene expression and reaction flux. At the moment, different tools (GIMME [1], MADE [2], iMAT [3]) deal with these aspects by applying consistent, although different, mathematical strategies. In the next years higher resolutions may be offered at the upstream level, the experimental one.

Abundance and characterization of biomass components is also problematic. Deep awareness is only possessed for few strains, generally model organisms. *Escherichia coli* and *Saccaromices cerevisiae* are brilliant examples of how a great knowledge baggage has been exploited by the metabolic modelers community. Yeast modeling, for instance, is a long, successful history which brought to the genetic editing of several species in order to be used as cell-factories for sustainable production of energy and of a plethora of molecules [4–7]. This kind of biochemical and molecular profiling is really auspicable also for other species. In the next future improvements of empirical measurements and mathematic applications may both suits the scope.

Also, new devices embodying kinetics laws may be delivered, giving birth to a 2.0 generation of metabolic models [8]. Availability of transcriptomics, proteomics and metabolomics data will grow crucial as the combination of computational and experimental fields will definitely take place. For this, facilities collecting and sourcing similar informations are mandatory.

Standardization is another knot. Scientific community should benefits of works achieved all over the globe. However, re-usability of computational models (and biological data in general) are often challenged by lack of formality. Guidelines describing the need to unambiguously and perennially

identify model components are arising [9–11]. Information regarding model origin and development (metadata) are sometimes as relevant as the core data. Controlling these parameters will allow accurate and efficient research strategies and models comparison (or format conversion). Addressing all these issues will drastically impact models and, in a larger perspective, whole biology.

Despite GEMs are still young, they have already offered relevant insights, especially in microbial systems biology. Here they also promise to open the field to evolutionary aspects [12] and community models [13–17]. However, the greatest expectations are probably placed on to human GEMs and their possible impact on personalized medicine [17–22]. Modeling of disease states like cancer, obesity and diabetes already started [23–25]. These applications will facilitate drug discovery, and possibly, therapeutic solutions will be tailored to individual needs.

Given these incredible chances, the concluding remark of this dissertation is to underline the importance to drive genomics and scientific efforts in such a direction. Computational tools and systems-biology should be completely integrated in daily research to assist biological (genomic) sciences becoming predictive. This will fundamentally transform not just biology and the way we learn but the entire society.

Bibliography

1. Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS computational biology* **4,** e1000082 (2008) (cit. on p. 152).
2. Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27,** 541–547 (2010) (cit. on p. 152).
3. Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology* **26,** 1003–1010 (2008) (cit. on p. 152).
4. Yu, T., Zhou, Y. J., Wenning, L., *et al.* Metabolic engineering of Saccharomyces cerevisiae for production of very long chain fatty acid-derived chemicals. *Nature communications* **8,** 15587 (2017) (cit. on p. 152).
5. Wei, Y., Gossing, M., Bergenholm, D., Siewers, V. & Nielsen, J. Increasing cocoa butter-like lipid production of Saccharomyces cerevisiae by expression of selected cocoa genes. *AMB Express* **7,** 34 (2017) (cit. on p. 152).
6. Zhu, Z., Zhou, Y. J., Krivoruchko, A., *et al.* Expanding the product portfolio of fungal type I fatty acid synthases. *Nature Chemical Biology* **13,** 360–362 (2017) (cit. on p. 152).
7. Tippmann, S., Scalcinati, G., Siewers, V. & Nielsen, J. Production of farnesene and santalene by saccharomyces cerevisiae using fed-batch cultivations with rq-controlled feed. *Biotechnology and bioengineering* **113,** 72–81 (2016) (cit. on p. 152).
8. Sánchez, B. J., Zhang, C., Nilsson, A., *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology* **13,** 935 (2017) (cit. on p. 152).

9. Taylor, C. F., Field, D., Sansone, S.-A., *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology* **26,** 889–896 (2008) (cit. on p. 153).

10. Smith, B., Ashburner, M., Rosse, C., *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* **25,** 1251–1255 (2007) (cit. on p. 153).

11. Le Novère, N., Finney, A., Hucka, M., *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* **23,** 1509–1515 (2005) (cit. on p. 153).

12. Monk, J. M., Charusanti, P., Aziz, R. K., *et al.* Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences* **110,** 20338–20343 (2013) (cit. on p. 153).

13. Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., *et al.* Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell metabolism* **22,** 320–331 (2015) (cit. on p. 153).

14. Shoaie, S., Karlsson, F., Mardinoglu, A., *et al.* Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific reports* **3,** 2532 (2013) (cit. on p. 153).

15. Stolyar, S., Van Dien, S., Hillesland, K. L., *et al.* Metabolic modeling of a mutualistic microbial community. *Molecular systems biology* **3,** 92 (2007) (cit. on p. 153).

16. Zhuang, K., Izallalen, M., Mouser, P., *et al.* Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME journal* **5,** 305–316 (2011) (cit. on p. 153).

17. Thiele, I., Swainston, N., Fleming, R. M., *et al.* A community-driven global reconstruction of human metabolism. *Nature biotechnology* **31,** 419–425 (2013) (cit. on p. 153).

18. Ghaffari, P., Mardinoglu, A., Asplund, A., *et al.* Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. *Scientific reports* **5** (2015) (cit. on p. 153).

19. Ghaffari, P., Mardinoglu, A. & Nielsen, J. Cancer metabolism: a modeling perspective. *Frontiers in physiology* **6** (2015) (cit. on p. 153).

20. Mardinoglu, A., Agren, R., Kampf, C., *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications* **5,** 3083 (2014) (cit. on p. 153).

21. Duarte, N. C., Becker, S. A., Jamshidi, N., *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences* **104,** 1777–1782 (2007) (cit. on p. 153).

22. Pornputtapong, N., Nookaew, I. & Nielsen, J. Human metabolic atlas: an online resource for human metabolism. *Database* **2015,** bav068 (2015) (cit. on p. 153).

23. Agren, R., Mardinoglu, A., Asplund, A., *et al.* Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology* **10,** 721 (2014) (cit. on p. 153).

24. Agren, R., Bordel, S., Mardinoglu, A., *et al.* Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology* **8,** e1002518 (2012) (cit. on p. 153).

25. Oberhardt, M. A., Yizhak, K. & Ruppin, E. Metabolically re-modeling the drug pipeline. *Current opinion in pharmacology* **13,** 778–785 (2013) (cit. on p. 153).

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Firenze, October 31, 2017

Luana Presta