The Pangenome: A Data-Driven Discovery in Biology

Duccio Medini, Claudio Donati, Rino Rappuoli, and Hervé Tettelin

Abstract An early example of Big data in biology: how a mathematical model, developed to address a practical question in vaccinology, transformed established concepts, opening biology to the “unbounded.”

Keywords Pangenome · Heaps’ law · Reverse vaccinology · Group B *Streptococcus* · Big data · Unbounded diversity

1 The Quest for a *Streptococcus agalactiae* Vaccine

In August of 2000, a collaboration between Rino Rappuoli’s team, including Duccio Medini, Claudio Donati, and Antonello Covacci at Chiron Vaccines in Siena, Italy, and Claire Fraser’s group, including Hervé Tettelin at the Institute for Genomic Research (TIGR) in Rockville, MD USA, was established to apply their recently pioneered reverse vaccinology approach (Pizza et al. 2000; Tettelin et al. 2000) to the problem of neonatal Group B *Streptococcus* (GBS, or *Streptococcus agalactiae*) infections (Fig. 1a). The collaboration also included Dennis Kasper, Michael Wessels, and colleagues, experts in GBS biology from the Boston Children’s Hospital, Harvard Medical School, Boston, MA USA.

GBS is a leading cause of neonatal life-threatening infections, despite the extensive application of antibiotic prophylaxis. Therefore, a vaccine was dearly needed to

D. Medini · R. Rappuoli
GSK Vaccines R&D, Siena, Italy
e-mail: duccio.x.medini@gsk.com

C. Donati
Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all’Adige, Italy

H. Tettelin (✉)
Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA
e-mail: tettelin@som.umaryland.edu

Fig. 1 Pangenome visuals. (a) 1999—Plymouth (NH, USA): Rino and Hervé in the woods around the time of initial discussions about the GBS collaboration. (b) 2004—Rockville (MD, USA): Pangenome early sketch and (Hervé the) genome in his pants. (c) Early 2005—Siena (Italy): Duccio and Claudio labor over the pangenome formula development. (d) 2018—Ellicott City (MD, USA): pangenome book editing, Hervé and Duccio locked in the basement

effectively prevent GBS infections. The manufacturing of a capsular polysaccharide-based vaccine was hindered by the existence and high incidence of at least five different disease-causing serotypes of GBS. Thus, the collaborative team embarked on the development of a GBS protein-based vaccine.

The concept was to use the *Streptococcus agalactiae* genome sequence information to predict proteins likely to be surface exposed and use these in experimental assays for antigenicity and antibody accessibility toward the development of a GBS vaccine via active maternal immunization [for details on GBS reverse vaccinology, see Maione et al. (2005)].

Unlike the case of *Neisseria meningitidis*, with which reverse vaccinology was pioneered right before the GBS project using a single genome, two GBS gap-free genomes were available when the project was initiated, and more genomes were generated early in the course of the project. Indeed, Tettelin et al. [TIGR (Tettelin et al. 2002)] and Glaser et al. [Pasteur Institute, France (Glaser et al. 2002)] independently reported the first two complete gap-free genome sequences of GBS in September of 2002.

At that time, sequencing multiple strains or isolates of the same species was far from commonplace. Both strains, serotype V 2603 V/R and serotype III NEM316, were clinical isolates. Glaser et al. compared their NEM316 genome to that of *Streptococcus pyogenes* (group A *Streptococcus*, GAS) and concluded that 50% of the GBS genes without an ortholog in GAS were located in 14 potential pathogenicity islands enriched in genes related to virulence and mobile elements. Tettelin et al. used a microarray-based comparative genomic hybridization (CGH) approach, whereby they hybridized the genomic

DNA of each of 19 GBS isolates of various serotypes onto a microarray of spotted 2603 V/R gene-specific amplicons, and identified several regions of genomic diversity among GBS isolates, including between isolates of the same serotype (see Fig. 2a).

These separate studies provided the first evidence that a significant amount of genomic information or gene content was variable among closely related streptococcal isolates, challenging the commonly accepted notion that the genome of a single isolate of a given species was sufficient to represent the genomic content of that species. Based on this understanding, the collaborative team decided to generate an additional 6 GBS genomes (Tettelin et al. 2005), selecting isolates from the five major disease-causing serotypes known at the time. The genome of the serotype Ia strain A909 was sequenced to completion in collaboration with the group of Craig Rubens at Children's Hospital and Regional Medical Center, Seattle, WA, USA. The other five strains—515 (serotype Ia), H36B (serotype Ib), 18RS21 (serotype II), COH1 (serotype III), and CJB111 (serotype V)—were sequenced as draft genomes, i.e., no attempt was made to manually close the gaps existing between contigs of the genome assemblies.¹ Comparison of the eight GBS whole-genome sequences confirmed the presence of the regions of genomic diversity previously identified by CGH (see Fig. 2b).

Surprisingly for the time, the shared backbone, or core set of genes present in each of the eight genomes, amounted to only about 80% of any individual genome's gene coding potential. Within these eight genomes, there was no pair that was nearly identical. Instead, each genome contributed a significant number of new strain-specific genes not present in any of the other genomes sequenced. Other sets of genes were shared by some but not all of the genomes.

This large amount of genomic diversity, which was not correlated to GBS serotypes, did not fail to stun members of the investigative team, including the experts in GBS biology. It also prompted an important question that formed the foundation of the pangenome concept: "How many genomes from isolates of the GBS species do we need to sequence to be confident that we identified all of the genes that can be harbored by GBS as a whole?"

This question, motivated by the need to identify all potential vaccine candidates for the species, led to active discussions among the collaborators, the drawing of highly accurate and inspirational scientific sketches (see Fig. 1b), and the decision to develop a mathematical model to determine how many other strains should have been sequenced.

2 When Data Amount and Complexity Exceed What Can Be Done Without Mathematics

The question was clear: "how many genomes..." i.e., the answer had to be a number. And a clear question is always a great way to start.

¹It should be noted that the COH1 genome, a representative of the highly prevalent disease-causing CC17 clonal complex, was later released as a gap-free genome (NCBI BioProject: [PRJEB5232](https://www.ncbi.nlm.nih.gov/bioproject/PRJEB5232)).

Fig. 2 Group B *Streptococcus* (GBS) genome diversity data that led to the pangenome discovery. (a) Comparative genome hybridization (CGH) provided a first hint about the high degree of genomic diversity within the GBS species. This circular representation of the GBS 2603 V/R

When the team in Siena was asked to figure out how to come up with an answer, they were faced with two assumptions, implicit in the question itself. First, the number was expected to be larger than eight, as the presence of specific genes in each of the eight isolates already sequenced suggested. Second, such a finite number was expected to exist.

The whole concept of biological species, a cornerstone of classical cladistics textbooks, had been evolving already toward the “species genome” concept thanks to the genomic revolution. The common knowledge, though, still held a 1:1 relationship between the species and the genome concepts. Consequently, a well-defined genetic repertoire for a bacterial species was the most natural assumption, implying that a finite—and hopefully small—number of genome sequences would be sufficient to exhaust it.

Genomic data had already introduced complexity and size in biology a decade before, when substantial mathematical work had been required to succeed in assembling tens of thousands of Sanger reads into a reconstructed chromosomal sequence (Sutton et al. 1995).

Here complexity and size were growing again, as the population scale of a species was being explored. More mathematical modeling was needed to translate the comparison among genomic data into a number.

Any modeling work starts with arbitrary choices. The first choice—that would remain a cornerstone of pangenome pipelines in the decades to come—was to adopt a reference-free approach.

Population genomics had been explored to that point mostly through cDNA microarrays (CGH), where the experimental design favors the physical comparison of DNA from many isolates with a reference one, usually a well-known laboratory strain used worldwide by the scientific community.

This approach has benefits also for *in silico* comparative genomics, because the number of comparisons to be performed scales linearly with the number of genome sequences to be compared, i.e., for any new isolate, one more comparison is performed. Also, the high-quality annotation of a well-studied genome can be easily transferred onto the others. However, the reference-based approach introduces strong limitations biasing the comparisons versus one specific individual of the species, which usually has no other ecological merit than having been around in microbiology labs for decades.

Fig. 2 (continued) genome shows predicted ORFs in the two outermost rims and those variable (blue bars) or absent (red bars) in the 19 genomic DNAs hybridized onto the 2603 V/R gene amplicon microarray. Regions of diversity are numbered 1–15 [for details, see (Tettelin et al. 2002)]. **(b)** *In silico* comparative genomic analysis of 8 GBS genomes confirmed CGH results and revealed additional regions of diversity using each genome as a reference. In this display, genes are arbitrarily color-coded by position in their genome along a gradient from yellow to blue. Genes are then depicted above their ortholog in the reference genome using the color they have in their home genome. Breaks in the color gradient reveal rearrangements and white regions reveal genomic regions absent in query genomes when compared to the reference. Each panel corresponds to each of the eight genomes used as the reference [for details, see Tettelin et al. (2005)]. Copyright 2002, 2005 National Academy of Sciences

Looking for a holistic assessment of a species diversity, the reference-free approach was natural, but it came with the disadvantage of scaling quadratically, i.e., any new genome would have to be compared to all the genomes already considered, leading to significant computational challenges.²

The second modeling choice was to use the gene as a unit of comparison or, more precisely, the open reading frames (ORFs) bioinformatically predicted on each genome sequence. Consequently, the analysis focused on an arbitrary subset of the genetic material, ignoring noncoding sequences whose relevance would have been increasingly appreciated in the years to come. Also, it implied accepting a certain number of nucleotide-level polymorphisms as not relevant for the diversity they were trying to model: allelic variants of the same gene would be considered as the same entity, as the problem was not to characterize microevolution—that strains accumulate mutations was well known—but to quantify the amount of “novel” genetic material contributed by each new sequence.

Intuitively, the more genomes analyzed, the fewer new genes (ORFs not observed with sufficient similarity in any other genome) should be identified. To answer the original question (“how many genomes. . .”) the team decided to determine the pace at which new genes would decrease with increasing numbers of genomes sequenced, in order to extrapolate the trend toward the number of genomes corresponding to no new genes identified.

As the number of new genes identified in the n -th genome depends on the selection of both the n -th genome itself and the previous $n - 1$ genomes considered, for each n from 1 to 8 we considered all the $8! / [(n - 1)! \cdot (8 - n)!]$ possible combinations to avoid bias, i.e., a total of 1024 pairwise, whole genome vs. whole-genome comparisons, i.e., ~2 billion gene vs. gene comparisons.

For each n from 1 to 8, we obtained a cloud of values and, following the same approach, the number of core genes (ORFs observed with sufficient similarity in all other genomes) was also measured.

Both new and core gene averages showed the expected decreasing trends, with the number of core genes for GBS decreasing exponentially toward the asymptotic value of 1806.

Surprisingly, though, the decreasing number of new genes was not trending toward zero in any way. Rather, the trend was reasonably reproduced by an exponential decay converging to a fixed value of 33, significantly greater than zero (see Fig. 3a).

In summary, mathematical extrapolation of the trend observed with the first eight genomes indicated that, for every new genome sequenced, new genes would have been discovered, even after a large number of genomes had been sequenced.

The extrapolation had two immediate implications: (i) no number of sequenced genomes would have assured a complete sampling of the GBS species pangenome, because (ii) the genetic repertoire of the species had to be considered as an unbounded entity.

²This would have been mitigated a few years later by the introduction of an unbiased, random sampling adjustment (Tettelin et al. 2008).

Fig. 3 Mathematical models revealing the “unbounded” pangenome. **(a)** *The first GBS pangenome* (Tettelin et al. 2005), copyright 2005 National Academy of Sciences. The number of specific genes is plotted as a function of the number n of strains sequentially added. The blue curve is the least-squares fit of the exponential pangenome function to the data. The extrapolated average number of strain-specific genes is shown as a dashed line. (Inset) Size of the GBS pangenome as a function of n . The red curve is the calculated pangenome size with values of the parameters obtained from the fit of the pangenome function to data. **(b)** *The refined power-law pangenome model* (Tettelin et al. 2008). Pangenome of *Bacillus cereus* using medians and a power-law fit. The total number of genes found with the pangenome analysis is shown for increasing numbers of genomes sequenced. Medians of the distributions are indicated by red squares. The curve is a least-squares fit of the power-law pangenome function to medians. The exponent $\gamma > 0$ indicates an open pangenome species

Understandably, the conclusion elicited in the group reactions comprised between complacent irony and the gentle suggestion to redo the work and find the mistake (see Fig. 1c).

So the team did, adding different alignment algorithms, running accurate sensitivity analyses on the thresholds adopted for sequence alignment, applying the same pipeline to other bacterial species known to be less variable as negative control and rechecking every line of the code. Eventually, the team agreed that the extrapolation was correct and novel genes belonging to the GBS species would be found even after sequencing a very large number of genomes. At this point, the team realized the need for a new entity in the genome world to account for those genes that belong to the species but are not present in some genomes. After long discussions, the team agreed on the pangenome concept and described the pangenome of each species by three differentiated components: its core genome, i.e., the genes present in each isolate of the species; its accessory genome, also called initially dispensable, i.e., the genes present in several but not all isolates; and, finally, its strain-specific genes, sampled in one isolate only.

As it would become apparent a few years later, when more genome sequences became available, and for multiple species, a much more accurate description of the trend of new genes would have been provided by a power-law function (derived from the Heaps' law, see Fig. 3b) actually decreasing to zero, as described in more detail in the next section.

But for *S. agalactiae* and some other species, the exponent of the power law was smaller than a critical value, i.e., the decrease of the number of new genes observed with new genomes was so slow, that the size of the pangenome remained an increasing and unbounded function of the number of genomes considered, as is the number of new words discovered in text corpora written in a live language (Heaps 1978). In other words, although the initial modeling work was still incomplete, the conclusion was already correct.

Another critical element that would have gained relevance over time in pangenome analyses, was the heterogeneity of the population sampling. As in any population-modeling exercise, the conclusions at the population scale are heavily dependent on the randomness of the sample, particularly if small, and can be seriously affected by the presence of structure in the population. If only a few, related isolates would be sequenced in an otherwise heterogeneous population, the sample would underestimate the population's diversity. Conversely, if in a population characterized by a few groups of highly similar isolates, we would assess only one genome per group, by extrapolating the measurement to the whole population we would largely overestimate the overall diversity. An effective, albeit incomplete, mitigation of the sampling bias was obtained by replacing the mean of the permutations with medians, which are more robust indicators of centrality.

However, in the original analysis of the eight *S. agalactiae* isolates, one of the more surprising results for the experts of the species' biology, was the lack of any specific relatedness among isolates belonging to the same serotypes, indicating that the phenotypic criteria used to classify the species thus far had no direct relationship with the genomic repertoire of the isolates. From a molecular perspective, this is

explained by the fact that genes encoding GBS capsular polysaccharides are part of a single locus, and this locus can be transferred across isolates by lateral gene transfer, showing how the repertoire of dispensable or strain-specific genes can, under specific circumstances, become available to any strain of a given species.

All in all, the answer to the question “How many isolates do we need to sequence to identify all the GBS genes?” was: “there is no such number, the GBS pangenome is open.”

The very idea of an unbounded genomic repertoire for a bacterial species was opening the microbiology community to a new way of looking at bacterial species and their anatomy.

While the core-genome remains substantially stable after a few tens of isolates are properly sampled, confirming the genomic consistency of the bacterial species concept, the more isolates are sequenced the more strain-specific genes merge into the accessory genome, expanding the pangenome size.

The underpinning mechanisms and ecological consequences of these dynamics of novelty-generation, spanning the scales of individual mutation, horizontal gene transfer promoted by phage transduction, bacterial conjugation or natural transformation, and population effects would become the object in the years to come of ever-increasing attention of the scientific community (see Fig. 4). A recent example was the observation that the majority of the metabolic innovations in the evolution of *Escherichia coli* arose through the horizontal transfer of single DNA segments (Pang and Lercher 2019).

3 The Vocabulary of Life: Heaps’ Law and Pangenomes

In the initial work on *S. agalactiae* (Tettelin et al. 2005), the authors used a decreasing exponential to model the number of new genes discovered in each new genome sequenced. This mathematical function converges asymptotically to a constant value (Fig. 3a and blue curves in Fig. 5). The openness of the pangenome followed from the fact that the best fit of the exponential function to data indicated an asymptotic value significantly higher than zero, i.e., a fixed number of new genes to be discovered in each new genome after the first eight sequenced. Although comforted by the biological diversity observed, such a conclusion was theoretically disturbing because it indicated that, no matter how exhaustively the species would have been sampled, the amount of novelty discovered per new isolate would have remained, on average, constant. A possibility extremely unusual across a wide variety of sampling problems.

In the subsequent work on *H. influenzae* (Hogg et al. 2007), the authors proposed a different approach, focused on the frequency distribution of genes and on the more conservative assumption of a mathematically closed pangenome. However, an increasing number of genomes used to train their model, led to larger predicted size of the pangenome, pointing again toward pangenome openness.

Fig. 4 Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pangenome, and metagenome (Medini et al. 2008). Intra-species (a), inter-species (b), and population dynamic (c) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their interrelationships. Multiple genome sequences are needed to describe the pangenome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (d)

Fig. 5 Power-law regression for new genes (Tettelin et al. 2008). The numbers n of new genes are plotted for increasing values of the number N of genomes sequenced. Medians of the distributions are indicated by red squares. Blue curves are least-squares fit of the exponential function, as in the original pangenome model. Red curves are least-squares fit of the power-law function. The exponent α determines whether the pangenome is open ($\alpha \leq 1$) or closed ($\alpha > 1$). The top panel shows data for an open pangenome species, *P. marinus*; the bottom panel for a closed pangenome species, *S. aureus*

The collaboration with **Ciro Cattuto** from the Institute for Scientific Interchange (ISI) Foundation in Turin offered the opportunity to recognize that determining the size of a pangenome was a problem analogous to many similar sampling problems, already addressed when dealing with macroscopic characteristics of complex systems, including human languages.

Before delving into the analogy between genomics and linguistics that allowed to mathematically solve the pangenome problem, a short diversion into the origins of the science of *complex systems* may be useful.

Since the 1970s, a few brilliant minds from disparate academic backgrounds, realized that challenges and opportunities posed by contemporaneity bear a level of complexity exceeding the capacity of established scientific paradigms (Ledford 2015). In 1984, a small group of Nobel laureates and eminent scientists from

Physics, Economics, and other disciplines founded the Santa Fe Institute (Santa Fe Institute) with the visionary ambition of creating a novel science called *complexity* (Waldrop 1993).

That original intuition is at the basis of today's widespread concept of *complex system*, adopted ubiquitously to deal with biological, ecological, economic, technological, and societal systems that cannot be effectively described by linear, inductive approaches, because of the nature of the interactions among system's components, and between the system itself and its environment.

The inductive approach of empirical sciences (i) observes the detailed phenomenology of a system to (ii) infer its underlying dynamics and (iii) uses the inferred laws to describe deterministically the macroscopic properties of the system. For example, (i) observe the movements of planet Earth, Moon and of the Sun to (ii) infer the laws of gravitation and (iii) predict the future trajectory of the planets in the Solar system (Newton 1687).

The approach proposed from the pioneers of complex systems was, in a way, the opposite: (i) start by observing macroscopic, statistical properties shared by multiple systems, (ii) identify a characteristic common to the disparate systems sharing the same property, and (iii) infer generative models, based on that characteristic, capable of accounting for the macroscopic properties observed. For example, (i) observe that in social networks, such as Facebook, few individuals have many connections, and many individuals have few connections, i.e., the frequency “y” of the degree “x” of the network nodes follows a power law “ $y = x^\alpha$ ” for some value of the exponent alpha; (ii) confirm that the frequency of words in human languages, of genes in genomes and of inhabitants in cities, all share the same property described for social networks, and all these systems are “modular”, i.e., composed of discrete, connected elements; (iii) show that the “preferential attachment” mechanism—according to which the more an element is frequent, the higher the likelihood its frequency will further increase—can be used to generate systems showing the power-law property observed above (Albert and Barabasi 2002).

A similar thinking process led to the solution of the pangenome problem. The rapid accumulation of tens of genome sequences for multiple species had clearly shown that the number of new genes discovered per new genome sequenced follows a decreasing power law, rather than a decreasing exponential trend (see Fig. 5). A similar behavior, for the number of new words discovered upon analyzing increasing numbers of instance texts written in English, had been observed decades earlier by the mathematical linguist Gustav Herdan (1960) and then generalized by Harold Stanley Heaps in the context of information theory as the Heaps' law (Heaps 1978).

When the number of new genes (or words) discovered is a power law of the increasing number of genomes (or text corpora), the overall size of the pangenome (or vocabulary) is also a power law, and the mathematical function depends only on two parameters: the power-law exponent and a proportionality constant. The rate of discovery of new genes is predicted to decrease always toward zero, but the speed of the decrease varies by species. With open pangenomes, such a number is just not decreasing fast enough for the cumulated number of observed genes to level off. Thus, a power-law behavior for the observed number of specific genes allows the

possibility of having an open pangenome without requiring that a fixed number of new genes be discovered for each new genome.

In order to complete the approach proposed by the pioneers of complexity, extensive work has been dedicated in recent years to the search for generative models that would account for the macroscopic properties of pangenomes and similar complex systems, including preferential attachment (Albert and Barabasi 2002), self-organized criticality (Bak et al. 1987; Mora and Bialek 2011), and random group formation (Baek et al. 2011). The Heaps' law, however, is only one of such properties displayed by genome data, the other two notable ones being the Zipf's law for the frequency distribution of gene family sizes in complete genomes (Huynen and van Nimwegen 1998) and the "U-shaped" gene frequency distribution (Haegeman and Weitz 2012). The generative models proposed so far could generate some of the observed macroscopic characteristics, but not all at the same time. More recently, a novel mechanism based on a sample space-reducing process (Corominas-Murtra et al. 2015) was proposed, and shown to reproduce naturally the three major properties of pangenomes at once (Mazzolini et al. 2018). Generative processes show how a certain system can be built ("generated") following a pre-defined rule or mechanism; for example, by choosing the elements of the system from an infinite pool of possible components, one after the other randomly. The idea behind the sample space-reducing process for the generation of a certain realization (genome, book) is that when a component (gene, word) is chosen, that choice restricts the space of the possible elements than can be chosen thereafter, permitting only certain other components—but not all—to be added. This assumption seems particularly relevant for genomic and linguistic systems, where the functioning (for genomes) or meaningfulness (for texts) depends on ordered combinations of multiple elements (genes in operons, words in sentences) that are not random (after a restriction enzyme, only a methylation gene produces a restriction-modification system; after a subject, only a verb produces a proposition). For this reason, and considering the relative simplicity of its mathematical implementation, the sample space-reducing process bears promise in the quest for a deeper understanding of the fundamental mechanisms responsible for the generation of pangenomes.

4 Pangenome Vaccinology

The existence of species with an open pangenomes has a profound effect on the selection of potential vaccine candidates identified by a reverse vaccinology approach. Indeed, the accessory genome was found to be an important contributor to protein antigens (Mora et al. 2006) implying that, for many bacterial species, a protein-based universal vaccine would only be possible by including a combination of antigens from the core and the accessory genomes.

The pathogen population structure and dynamics became a key element of vaccine research, paving the way for a modern approach to vaccine discovery known as *pangenomic reverse vaccinology* (Donati et al. 2010; Mora et al. 2006;

Budroni et al. 2011). The key principles of this approach, that expands the reverse vaccinology paradigm based on a single genome sequence (Rappuoli 2000), include reducing bias in isolate selection for genome sequencing (to the extent possible, e.g., carriage vs. invasive isolates, or commensal vs. pathogenic isolates) based on epidemiology, followed by defining the population genomic structure of the species, including its pangenome.

Reverse vaccinology pipelines are then applied to predict the antigenic potential of proteins based on a collection of desired (and undesired features) that they carry, for a recent review on reverse vaccinology pipelines, see Dalsass et al. (2019). Top-ranked vaccine candidate proteins can then be taken through the experimental portion of the vaccine development phase whereby their accessibility and antigenicity are assayed, for instance starting with antigen-based serological typing [for a review on this experimental phase and subsequent phases, please see (Del Tordello et al. 2017)].

It should be noted that the actual transcription, translation, and exposure of a set of selected vaccine protein candidates may vary with the environment, including colonization or infection of various organs, and niches within these organs. The pangenome can inform on these specificities by including isolates with a propensity to target certain organs/niches vs. others (e.g., skin vs. throat isolates of group A *Streptococcus*). Interactions of antigens with host moieties are also key to designing successful candidates.

Ultimately, a combination of pangenomic reverse vaccinology with other multi-omics approaches in the context of host–pathogen interactions will better inform the rational design of next-generation vaccine targets and will lead to the most promising formulations to test in vivo.

5 Discussion

In 1946, even before the very discovery of DNA’s structure (Watson and Crick 1953), Joshua Lederberg and Edward L. Tatum had demonstrated the existence of “sexual” genetic recombination in bacteria (Lederberg and Tatum 1946), a discovery that granted them the Nobel Prize in 1958.

Horizontal DNA exchange in bacteria has been the subject of intense research ever since so, by the time the pangenome came along, the concept of diversity in bacterial species had been ingrained in the scientific community for half a century already. However, possibly because of the hubris induced by the breakthrough of first-generation genome sequencing technologies, for more than a decade the community had inadvertently reverted to a “pre-Lederberg-Tatum” mindset, considering each of the few genomes generated at the time as representatives of the respective species’ genetic blueprint. Of note, the name *pangenome* came to life after many, long discussions on how to name this new concept possibly reflecting the paradigm-shift that was required, at that time, to recognize the simple evidence of facts.

Fig. 6 Pangenome bibliometric data: overall number of pangenome publications since 2005. Blue curve: Number of pangenome publications in PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>). Query: “pan-genome” [Title/Abstract] OR “pangenome” [Title/Abstract] OR “pan genome” [Title/Abstract]. Orange curve: Number of scientific publications in Google Scholar (<https://scholar.google.it>) mentioning the pangenome. Query: Exact match to anywhere in the text: “pan-genome” OR “pangenome” OR “pan genome”

The pangenome discovery, at first sight, brought the scientific community back to Earth, to realize that a single genome was far from describing a whole species and that, as a side consequence particularly relevant for the genome pioneers of the time, genome sequencing was there to stay as a flourishing business for decades.

At the same time, though, the pangenome introduced a new dimension in microbiology that could hardly be associated with already established awareness: the concept that the genetic repertoire of a defined biological entity, such as a species, could be unbounded. In a way, the pangenome introduced the infinite in biology, with some humble analogy to what Theodosius Dobzhansky had done 30 years before, much more fundamentally, through the explanatory light of evolution (Dobzhansky 1973).

This could partly explain why, over a relatively short period of time, pangenomics became a discipline in itself (see Fig. 6). From a more practical stance, the impressive acceleration of bacterial genome sequencing was generating high numbers of genes that would not map to species’ reference genomes. The new concept offered a conceptual framework to accommodate the wealth of new data, becoming rapidly a must have for any microbial sequencing project.

Thirteen years later (see Fig. 1d), however, two further elements could be identified, that contributed to transforming a specific, empirical question, into a discovery that opened the scientific community to a new research field.

First, a concrete challenge motivated by a burning, unmet medical need, had gathered together people with very different backgrounds, spanning from Biology and Medicine to Physics and Engineering. This collision model, extensively used in modern science and business, promotes ideas that challenge the status quo by facilitating cross-fertilization and lateral thinking. Questioning the serotypes, the team discovered the pangenome. Simple in hindsight but challenging the established paradigm of biological species.

Second, pioneering technological breakthroughs at the bleeding edge, as it was for genome sequencing and assembly at the time, frequently unveils new, unexpected horizons. Not always, though: a critical condition remains the osmotic collaboration between scientists and technology experts mastering the data generation process, to bring in the team awareness of the limitations intrinsic to the data, reducing the risk of hasty misinterpretations, as well as the frustration of missed opportunities.

In conclusion, the pangenome is an early example of mathematical modeling applied to biological Big data: a serendipitous, data-driven discovery from a human health challenge, fostered by technological breakthroughs and people with different backgrounds willing to challenge the status quo. We are deeply grateful to the many investigators worldwide who took the pangenome concept well beyond what could be envisioned at the time, perfected and expanded techniques and applications, and ignited the fascinating evolution of discoveries that the reader now has the opportunity to explore in the remainder of this book.

References

- Albert M, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47. <https://doi.org/10.1103/RevModPhys.74.47>
- Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. *New J Phys* 13:043004
- Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of the $1/f$ noise. *Phys Rev Lett* 59:381–384
- Budroni S, Siena E, Dunning Hotopp JC, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV et al (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A* 108(11):4494–4499
- Corominas-Murtra B, Hanel R, Thurner S (2015) Understanding scaling through history-dependent processes with collapsing sample space. *Proc Natl Acad Sci U S A* 112:5348–5353. <https://doi.org/10.1073/pnas.1420946112>
- Dalsass M, Brozzi A, Medini D, Rappuoli R (2019) Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. *Front Immunol* 10:113. <https://doi.org/10.3389/fimmu.2019.00113>
- Del Tordello E, Rappuoli R, Delany I (2017) Reverse vaccinology: exploiting genomes for vaccine design. In: Modjarrad K, Koff WC (eds) *Human vaccines, emerging technologies in design and*

- development. Academic, Amsterdam, pp 65–86. <https://doi.org/10.1016/B978-0-12-802302-0.00002-9>
- Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. *Am Biol Teach* 35:125–129
- Donati C, Medini D, Rappuoli R (2010) Pangenomic reverse vaccinology. In: Sintchenko V (ed) *Infectious disease informatics*. Springer, New York, pp 203–221
- Glaser P et al (2002) Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol Microbiol* 45:1499–1513
- Haegeman B, Weitz JS (2012) A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* 13:196. <https://doi.org/10.1186/1471-2164-13-196>
- Heaps HS (1978) *Information retrieval - computational and theoretical aspects*. Academic, Orlando, FL
- Herdan G (1960) *Type-token mathematics*. Mouton & Co., The Hague
- Hogg JS et al (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 8:R103. <https://doi.org/10.1186/gb-2007-8-6-r103>
- Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* 15:583–589. <https://doi.org/10.1093/oxfordjournals.molbev.a025959>
- Lederberg J, Tatum EL (1946) Gene recombination in *Escherichia coli*. *Nature* 158:558
- Ledford H (2015) How to solve the world's biggest problems. *Nature* 525:308–311. <https://doi.org/10.1038/525308a>
- Maione D et al (2005) Identification of a universal group B streptococcus vaccine by multiple genome screen. *Science* 309:148–150
- Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M (2018) Zipf and Heaps laws from dependency structures in component systems. *Phys Rev E* 98:012315. <https://doi.org/10.1103/PhysRevE.98.012315>
- Medini D et al (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol* 6:419–430
- Mora T, Bialek W (2011) Are biological systems poised at criticality? *J Stat Phys* 144:268–302
- Mora M, Donati C, Medini D, Covacci A, Rappuoli R (2006) Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Curr Opin Microbiol* 9(5):532–536
- Newton I (1687) *Philosophiae naturalis principia mathematica*. Jussu Societatis Regiae ac Typis Josephi Streater, Londini. <https://www.loc.gov/resource/rbc0001.2013gen20872/>
- Pang TY, Lercher MJ (2019) Each of 3,323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. *Proc Natl Acad Sci U S A* 116:187–192. <https://doi.org/10.1073/pnas.1718997115>
- Pizza M et al (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287:1816–1820
- Rappuoli R (2000) Reverse vaccinology. *Curr Opin Microbiol* 3:445–450
- Santa Fe Institute Complexity Science. <https://www.santafe.edu/about/history>
- Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci Tech* 1:9–20
- Tettelin H et al (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287:1809–1815
- Tettelin H et al (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* 99:12391–12396
- Tettelin H et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci U S A* 102:13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 11:472–477

- Waldrop MM (1993) Complexity: the emerging science at the edge of order and chaos. Simon and Schuster, New York
- Watson JD, Crick FH (1953) The structure of DNA. Cold Spring Harb Symp Quant Biol 18:123–131

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.