Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Comparative genomics: the bacterial pan-genome

Hervé Tettelin[1], David Riley[1], Ciro Cattuto[2] and Duccio Medini[3]

Bacterial genome sequencing has become so easy and accessible that the genomes of multiple strains of more and more individual species have been and will be generated. These data sets provide for in depth analysis of intra-species diversity from various aspects. The pan-genome analysis, whereby the size of the gene repertoire accessible to any given species is characterized together with an estimate of the number of whole genome sequences required for proper analysis, is being increasingly applied. Different models exist for the analysis and their accuracy and applicability depend on the case at hand. Here we discuss current models and suggest a new model of broad applicability, including examples of its implementation.

Addresses
[1] Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, 20 Penn Street, Baltimore, MD 21201, USA
[2] Institute for Scientific Interchange Foundation, Viale S. Severo 65, 10133 Torino, Italy
[3] Novartis Vaccines and Diagnostics, Via Fiorentina 1, 53100 Siena, Italy

Corresponding authors: Tettelin, Hervé (tettelin@som.umaryland.edu) and Medini, ()Medini, Duccio (duccio.medini@novartis.com)

Current Opinion in Microbiology 2008, **12**:472–477

This review comes from a themed issue on
Genomics
Edited by Fiona Brinkman and Julian Parkhill

Available online 14th October 2008

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2008.09.006

Introduction

The advent of ultra-high throughput next generation sequencing technologies, for example, Roche-454 Life Sciences (www.roche-applied-science.com), Solexa-Illumina (www.illumina.com), and ABI-SOLiD (www.appliedbiosystems.com) and large-scale comparative genomics sequencing projects (e.g. http://www3.niaid.nih.gov/research/resources/mscs/, http://www.sanger.ac.uk/Projects/Microbes/, and http://genome.jgi-psf.org/mic_home.html) are leading to the availability of whole genome sequences for many strains of several bacterial species. Although until recently there has been a strong bias toward medically and environmentally relevant species, it is conceivable that in the near future many genome sequences will be available for most known bacterial species. Even unculturable species can now be tackled thanks to the emerging field of single cell genomics [1].

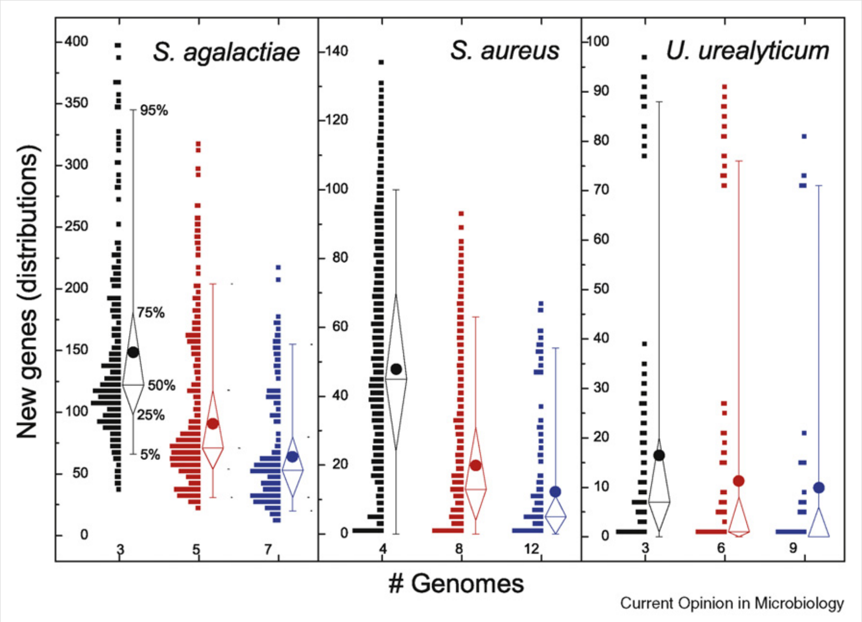
Comparative genomics analyses between multiple genomes of individual species have revealed extensive genomic intra-species diversity [2]. Given today's ease of generating draft whole genome sequences, it would be of value to know how many genomes should be sequenced for any given species to accurately represent its entire gene repertoire. A way to tackle this problem is to ask how many new genes are identified every time a new genome of the species of interest is sequenced. Tettelin *et al.* [3••] pioneered this approach using multiple genomes of *Streptococcus agalactiae*, followed by Hogg *et al.* [4••] who studied *Haemophilus influenzae* genomes. In both cases, the analyses resulted in the determination of a core genome that consists of genes shared by all the strains studied and probably encode functions related to the basic biology and phenotypes of the species. The striking feature of the studies was the realization that a significant percentage of each genome sequence was specific to each individual strain and therefore each new genome sequenced provided a number of new genes not previously characterized. Thus, the species' gene repertoire was significantly larger than that of any single strain of that species and a large number of genomes would have to be sequenced to characterize it. This led the authors to the concept of the bacterial pan-genome or supragenome, the topic of this review. The pan-genome is the sum of the above core genome and the dispensable genome that is composed of genes present in some but not all the strains studied as well as the strain-specific genes. The dispensable genome contributes to the species' diversity and probably provides functions that are not essential to its basic lifestyle but confer selective advantages including niche adaptation, antibiotic resistance, and the ability to colonize new hosts.

The *Streptococcus agalactiae* pan-genome model

The first pan-genome analysis was conducted on *S. agalactiae*, a major cause of disease in newborns, infants, and the elderly [5,6]. On the basis of the first *S. agalactiae* genome sequence and its use as the reference strain in microarray-based comparative genomic hybridizations [7], it was determined that this species' genomic diversity was fairly extensive. Eighteen percent of the genes in the reference genome were absent in at least one of the 19 isolates tested on the microarray and most of them were located in genomic islands of five or more consecutive genes. The main limitation of the micro-

Box 1 The effect of strain selection and the choice of methods on pan-genome analyses.

The distributions of new gene values produced by the combinatorial procedure described in reference [3••] incorporate much more information than synthesized in their averages. In fact, in several cases their averages could even be an inappropriate descriptor, as described in the figure inset that shows the distributions of these values for three different, typical cases. The histograms for *S. agalactiae* are characterized by a dominant mode, located in the lowest part of the distribution, plus a smaller tail that accounts for ~25% of the combinations. The center of the principal mode is close to the median of the distribution (50th percentile), while the average is displaced upwards. This is due to the fact that averages, compared with medians, are much more affected by outliers in the distributions, and if the data display a broad distribution their meaning may be questionable. Even less informative in such cases is the standard deviation, while the 25–75 percentile of the distribution can be more safely used to weight the regression in the non-linear fit procedure. Fortunately, for *S. agalactiae* the distributions are reasonably narrow and the offset of the average vs. the median is approximately constant for different values of *N*, the number of genomes sequenced. The consequence is that a regression will give consistent results on averages or medians (power law exponents $\alpha = 1.02 \pm 0.6$ and 1.05 ± 0.09, respectively, both consistent with an open pan-genome). The case of *S. aureus* is different: Data are spread on a long tail that changes shape for different numbers of genomes, plus a peak at zero. Averages here have a limited ability to capture the central tendency of the data, and depart from medians increasingly with *N*. The two regressions give somewhat different results: Power law exponent $\alpha = 1.27 \pm 0.03$ for averages and 1.83 ± 0.15 for medians. A more extreme situation is shown in *U. urealyticum* data where one half of the values are 0 and another 25% are < 10. However, the presence of a strain whose genome data are highly fragmented introduces a separate cloud of artificially high values, so that the averages remain consistently above 10, in a region of the distribution with no data at all, while the medians correctly converge to zero. A regression performed on medians shows a closed pan-genome ($\alpha = 2.5 \pm 0.9$) while averages would give an extremely open pan-genome ($\alpha = 0.44 \pm 0.10$). By and large, these data highlight the crucial role of strain selection for sequencing in pan-genome studies, regardless of the model selected to interpret the data. So far strains have mostly been selected for practical reasons rather than by statistically and epidemiologically sound criteria. This results in an uneven distribution of diversity among isolates that must be accurately interpreted using available information on the species population structure. Therefore, the choice of a robust descriptor of centrality such as median rather than average is crucial but the final decision must be informed by the specifics of the case under study.

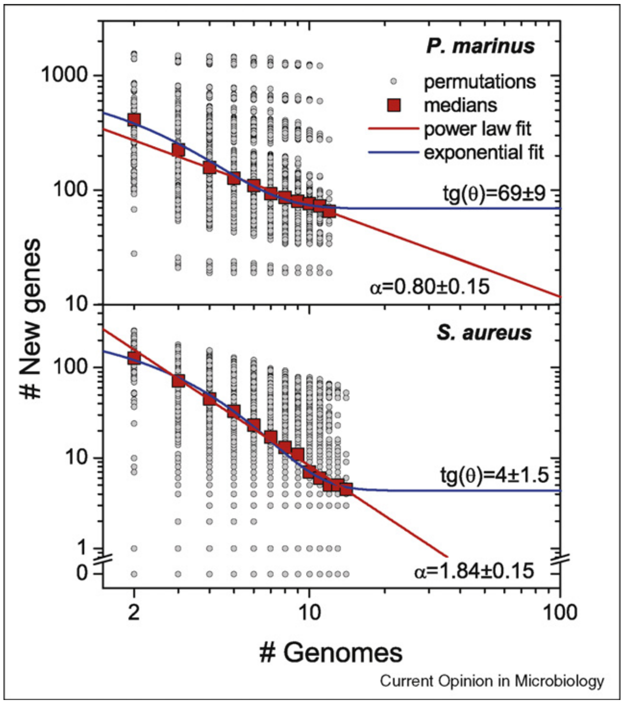


Distributions of new genes values: averages vs. medians. Vertical histograms represent the distributions of new genes values obtained for three bacterial species: *S. agalactiae*, *S. aureus*, and *U. urealyticum*. The pan-genome analysis described in reference [3••] was applied for three values of the number *N* of genomes sequenced. For each value of *N*, a box-plot of the distribution indicating the median (or 50th percentile, horizontal bar at the center of the diamond), the 25th and 75th percentile (vertexes of the diamond), the 5th and 95th percentile (tips of the whiskers), and the average (full circle) are shown.

array experiment resided in the fact that only the genes found in the reference genome were interrogated while the genes specific to the query strains were not identified. To circumvent this problem, additional genome sequences were generated such that a pan-genome analysis could be conducted on three complete genomes (free of gaps) and the sets of contigs resulting from five draft whole genome sequences where no gap closure was attempted [3••].

Genes predicted from the eight genomes were compared by means of all vs. all highly sensitive sequence alignments. An iterative procedure was used to estimate the rate of new genes discovered per additional genome sequenced. The number of new genes provided by the fourth genome, for instance, depends on the selection of both the fourth genome itself and the previous three considered. In order to randomize the data to the extent possible, all the combinations of $N = 1,2,\ldots,8$ genomes

Figure 1



Current Opinion in Microbiology

Power law and exponential regressions for new genes discovered with the availability of additional genome sequences. The numbers n of new genes found according to the pan-genome analysis described in [3••] are plotted for increasing values of the number N of genomes sequenced. Medians of the distributions are indicated by red squares (see Box 1). Blue curves are least squares fit of the exponential decay $n = \kappa \exp[-N/\tau] + \mathrm{tg}(\theta)$ to medians as in the original pan-genome model. The value of tg(θ) shown in the figure represents the number of new genes asymptotically predicted for further genome sequencing. Red curves are least squares fit of the power law $n = \kappa N^{-\alpha}$ to medians. The exponent α determines whether the pan-genome is open ($\alpha \leq 1$) or closed ($\alpha > 1$). The top panel shows data for an open pan-genome species, *P. marinus*; the bottom panel for a closed pan-genome species, *S. aureus*.

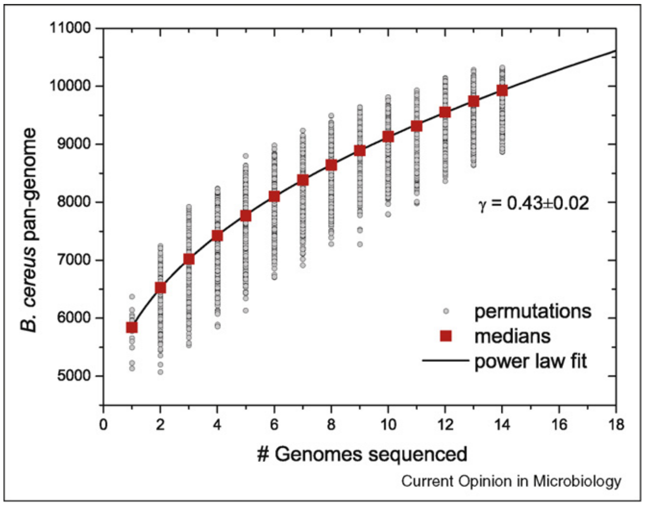
were calculated. A plot of the values obtained (similar to those shown in Figure 1) revealed clouds of points with a clear decay with N increasing. Intuitively, the more the genomes analyzed, the fewer the new genes discovered; but the accurate fit of an exponential decay function to the data indicated that the averages of the clouds converged to a finite number. In other words, analysis of the *S. agalactiae* pan-genome revealed that on average each additional genome sequence would reveal 33 previously uncharacterized genes, implying an unbounded or open pan-genome. As a consequence, a very large number of genome sequences would be required to characterize the entire gene repertoire to which the species has access. The exponential regression model with three parameters – the decay rate, the asymptotic number of new genes discovered, and a multiplicative constant – was used as a simple phenomenological descriptor of the experimental trend without further theoretical modeling. The averages of the gene numbers in each permutation were found to be reasonable descriptors of the distributions, and the same regression performed on all values rather than on the averages led to analogous results. This procedure for the analysis of a pan-genome has been applied to other bacterial species including *Streptococcus pyogenes* [8•], *Neisseria meningitidis* [9•], *Escherichia coli* [10•,11•], and *Prochlorococcus* [12•]. It has also provided the basic framework for identification and characterization of novel vaccine candidates using reverse vaccinology [13,14].

The *Haemophilus influenzae* pan-genome model

Hogg *et al.* [4••] further developed the pan-genome analysis by proposing a new model for the analysis of 13 *H. influenzae* genomes. The model takes into account the way in which the individual genes are distributed among the different genomes. Rather than trying to extrapolate the trend for new genes discovered by means of a simple regression, they probabilistically assign genes to classes that represent the proportion of strains in the population that possess the gene. Then they generate hypothetical genomes made of genes selected randomly from these classes, with a probability proportional to the frequency of occurrence of the class in the population. The parameters of the model are optimized in order to make the theoretical pan-genome as close as possible to the data collected for the real, sequenced genomes. Results of this study are intriguing, especially from a theoretical point of view. When trained on a subset of eight genomes the model predicts a pan-genome size of ~3000 genes (vs. an average genome size of ~1800 genes). When trained on all 13 genomes the predicted pan-genome size nearly doubles, growing to ~5200 genes, including a much larger number of rare genes, that is, genes present in a small fraction of the population. Besides several technical differences compared with the original model, the most relevant novelty is that this model assumes by design a finite number of genes in the species' pan-genome. This seems to be in radical contradiction with the original model of reference [3••], that allows the pan-genome size to be open. However, the more the genomes used to train the model, the larger the predicted size of the pan-genome, mainly due to the contribution of rare genes. Thus, although starting from radically different assumptions, Hogg *et al.* obtain exactly what the original model would define as an open pan-genome species for *H. influenzae*.

The approach gave similar results when applied to 17 *Streptococcus pneumoniae* isolates [15•], where the authors also showed a practical utility for modeling gene frequencies. A pan-genome might be extremely large and require a huge number of genomes sequences to be generated, but the genes with frequency >1% in the population might constitute a fairly small fraction of the gene pool, and if one is not interested in very rare genes then the amount of genomes required to characterize the species drops significantly.

Figure 2



Pan-genome of *Bacillus cereus* using medians and a power law fit. The total number of genes found according to the pan-genome analysis described in reference [3••] is shown for increasing values of the number N of *B. cereus* genomes sequenced. Medians of the distributions are indicated by red squares (see Box 1). The curve is a least squares fit of the power law $n = \kappa N^{\gamma}$ to medians. The exponent $\gamma > 0$ indicates an open pan-genome species.

Heaps' law and a new model for open pan-genomes

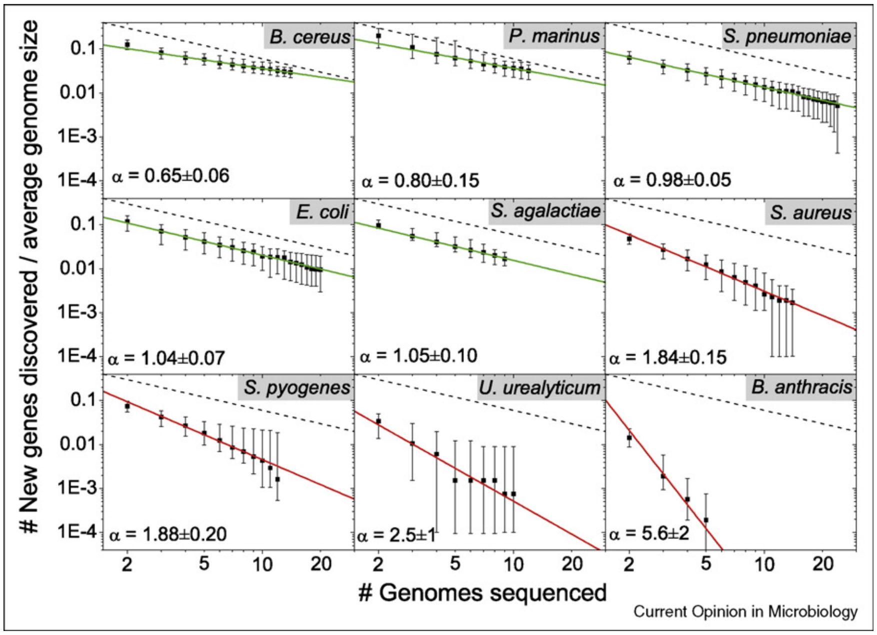
From a fundamental point of view, measuring the size of the pan-genome is one instance of a general class of measurements where, given a collection of 'entities' and their 'attributes', the number of distinct attributes that have been observed is monitored as a function of the number of entities considered. For the case of the pan-genome, 'entities' are genomes and their 'attributes' are genes. In many cases of practical interest it is known that the number n of distinct attributes grows according to a sub-linear power law of the number N of entities considered. That is, $n \sim N^{\gamma}$, with $0 < \gamma < 1$. This empirical law, originally formulated in the domain of information retrieval, is known as Heaps' law [16•]. Figure 2 shows a Heaps' law fit of the overall number of genes (pan-genome) obtained as in [3••] for *Bacillus cereus*. As an example, Heaps' law is known to hold for corpora of natural language: As one considers more and more instance text, the number of different words that have been observed grows as a power law of the number of scanned words. This is equivalent to say that the rate at which new attributes (words or genes) are found decreases as one considers more and more entities (instance text or genomes), as this rate is proportional to $N^{(\gamma-1)} = N^{-\alpha}$, with $\alpha = 1 - \gamma$. That is, as sampling proceeds, discovering a new attribute becomes increasingly harder. Qualitatively, this is what we observe for the number of new genes discovered as one considers more and more genomes.

As a consequence of the above general observations, a power law (see [17] for a thorough discussion of power law properties and applications) is a natural candidate for the functional dependence of the number of new genes discovered as a function of the number of genomes considered. The functional form of the power law depends on two parameters only, the exponent α and a proportionality constant. For $\alpha > 1$ ($\gamma < 0$), the size of the pan-genome approaches a constant as more genomes are sampled, that is, the pan-genome is closed. Conversely, for $\alpha < 1$ ($0 < \gamma < 1$), the size of the pan-genome is an increasing and unbounded function of the number of genomes considered, that is, the size of the pan-genome follows Heaps' law and the pan-genome is open. Finally, if the exponent α is exactly equal to 1 the pan-genome size follows a logarithmic trend, that is, grows very slowly but it is still technically unbounded. Overall, if $\alpha \leq 1$ the data are consistent with an open pan-genome.

It is important to remark that the rate of discovery of new genes is also decreasing in this case, and no asymptotic plateau needs to exist for the number of new genes discovered. The 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 pan-genome without requiring that a fixed number of new genes be discovered for each new genome.

A comparison between exponential and power law regression of the new genes data is shown in Figure 1 for two bacterial species. Unfortunately, a sound statistical comparison of different functional forms would require more data [18]. Nevertheless, the log–log plots in Figure 1 facilitate the evaluation of the asymptotic behaviors (several genomes, few new genes) of the two models. By power law regression, *Prochlorococcus marinus* has an open pan-genome with $\alpha = 0.80 \pm 0.15$. The same conclusion is reached when applying the exponential regression (horizontal asymptote = 69 ± 9 genes) but the asymptotic behavior of that function is hardly justifiable on the basis of the available data as opposed to the power law trend that provides a more natural extrapolation. Conversely, for *Staphylococcus aureus* the exponential regression converges to a small but finite number of asymptotically discovered new genes (4 ± 1.5), while the power law predicts a closed pan-genome ($\alpha = 1.84 \pm 0.15$). Given the small number of genomes, it seems really hard to maintain a prediction of open pan-genome for such a species on the basis of the exponential regression results, and the power law model seems more adherent to the available data and their biological implications.

Figure 3



Current Opinion in Microbiology

Power law regression for species with open and closed pan-genomes. The medians of the number n of new genes discovered for increasing values of the number N of genomes sequenced, normalized to the average genome size of the species, are displayed along with their 25–75 percentile intervals. Solid curves show the power law $n = \kappa N^{-\alpha}$ least squares fit to data for $N \geq 3$, weighted for the 25–75 percentile interval (error bar values lower than 10^{-4} have been rescaled to 10^{-4} for graphical reasons). Red curves indicate closed pan-genomes, green curves indicate open ones. In each box a dashed guide to the eye shows the borderline power law $n \sim N^{-1}$ to facilitate the comparison of slopes. For $N > 5$, values of *B. anthracis* fall below the scale limits. The estimated value of $\alpha \sim 0.6$ for *B. cereus* is in good agreement with the value of $\gamma \sim 0.4$ obtained from Heaps' law regression on total number of genes (see Figure 2).

Applying Heaps' law to a number of species

A subset of nine bacterial species for which nine or more whole genome sequences were available with annotation in Genbank (either in the complete genome or the whole genome shotgun WGS sections) were selected for application of the pan-genome model with power law regression using medians (Figure 3). In all cases the model fitted the data well. The analysis revealed five species with an open pan-genome: *S. agalactiae*, *S. pneumoniae*, *E. coli*, *B. cereus*, and *P. marinus*. The diversity in lifestyle between these species with an open pan-genome probably indicates that a number of very different environments require high levels of adaptability or promote a lot of lateral exchange of genetic material. On the contrary, four species display a closed pan-genome: *S. aureus*, *S. pyogenes*, *Ureaplasma urealyticum*, and *Bacillus anthracis*. The latter was already known for its very limited gene content diversity, and it still represents the most extreme closed pan-genome. While the *U. urealyticum* pan-genome is also somewhat limited, the remaining two are closed but their gene repertoire is predicted to be quite large.

Conclusions

Bacterial intra-species diversity keeps unveiling its depth and secrets as the power and speed of whole genome sequencing technologies increases. Several groups have made use of pan-genome models in an attempt to better characterize the breadth of the gene repertoire accessible to individual microbes and understand the amount of additional genomic data required for proper characterization of this repertoire. Defining the pan-genome of a bacterium sheds light on its biology and life style and has implications for the definition of the species itself [19••,20]. The field of metagenomics tackles entire communities using genomics and it is conceivable that pan-genome approaches could be used to characterize the gene sets available to organisms living in a given environment, defining what could be called the pan-microbiome of a specific niche.

From a theoretical perspective, it is now possible to move pan-genome modeling toward a more population-based perspective. Casting the problem of the pan-genome size in the mathematical language of complex systems science – where generative models for power laws have been extensively explored – may pave the way to modeling the statistical properties of the pan-genome and gaining insight into the mechanisms that underlie the observed properties, such as spontaneous mutations, intra-species recombination, and lateral gene transfer.

Acknowledgements

We are grateful to Joshua Orvis, Anup Mahurkar, and Dave Kemeza for their dedication in pushing the intensive pan-genome computes through the IGS computer grid.

References and recommended reading

Papers of particular interest published within the period of review have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Walker A, Parkhill J: **Single-cell genomics**. *Nat Rev Microbiol* 2008, **6**:176-177.

2. Pallen MJ, Wren BW: **Bacterial pathogenomics**. *Nature* 2007, **449**:835-842.

3. •• Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'**. *Proc Natl Acad Sci U S A* 2005, **102**: 13950-13955.

The first description of a pan-genome analysis, the exponential regression was used to fit the data.

4. •• Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD: **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* 2007:8 doi: 10.1186/gb-2007-1188-1186-r1103.

The second pan-genome model that takes into account the distribution of individual genes among the genomes studied.

5. Doran KS, Nizet V: **Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy**. *Mol Microbiol* 2004, **54**:23-31.

6. Schuchat A, Wenger JD: **Epidemiology of group B streptococcal disease. Risk factors, prevention strategies, and vaccine development**. *Epidemiol Rev* 1994, **16**:374-402.

7. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD *et al.*: **Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae***. *Proc Natl Acad Sci U S A* 2002, **99**:12391-12396.

8. • Lefebure T, Stanhope MJ: **Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition**. *Genome Biol* 2007, **8**:R71.

Application of the pan-genome analysis.

9. • Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O *et al.*: **Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitides***. *Proc Natl Acad Sci U S A* 2008, **105**:3473-3478.

Application of the pan-genome analysis.

10. • Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW: **Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray**. *Genome Biol* 2007, **8**:R267.

Application of the pan-genome analysis.

11. • Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sperandio V, Ravel J: **The pan-genome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates**. *J Bacteriol* 2008.

Application of the pan-genome analysis.

12. • Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J *et al.*: **Patterns and implications of gene gain and loss in the evolution of Prochlorococcus**. *PLoS Genet* 2007, **3**:e231.

Application of the pan-genome analysis.

13. Mora M, Donati C, Medini D, Covacci A, Rappuoli R: **Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach**. *Curr Opin Microbiol* 2006, **9**:532-536.

14. Tettelin H, Medini D, Donati C, Masignani V: **Towards a universal group B Streptococcus vaccine using multistrain genome analysis**. *Expert Rev Vaccines* 2006, **5**:687-694.

15. • Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J *et al.*: **Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome**. *J Bacteriol* 2007, **189**:8186-8195.

Application of the pan-genome analysis.

16. • Heaps HS: *Information Retrieval—Computational and Theoretical Aspects*. Orlando, FL: Academic Press; 1978.

Heaps' law.

17. Newman MEJ: **Power laws, Pareto distributions and Zipf's law**. *Contemp Phys* 2005, **46**:323-351.

18. Clauset A, Shalizi CR, Newman MEJ: **Power-law distributions in empirical data**. *arXiv:0706.1062v1*; 2007.

19. •• Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Curr Opin Genet Dev* 2005, **15**:589-594.

A review of the application of the first pan-genome model (exponential regression) and its impact on comparative genomics, bacterial populations, and the definition of species.

20. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era**. *Nat Rev Microbiol* 2008, **6**:419-430.