Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infection, Genetics and Evolution 5 (2005) 103–108
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Review

Exploring microbial microevolution with microarrays

Howard Ochman*, Scott R. Santos

Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA

Received 7 June 2004; received in revised form 1 September 2004; accepted 2 September 2004

Abstract

Gene arrays are typically employed to monitor gene expression and regulation, but they are finding additional applications in studying patterns of evolution in bacterial genomes. In particular, this approach has been applied to answer questions about the heterogeneity in full gene repertoires among bacterial strains and species without relying on more costly and time-consuming methodologies. In this review, we evaluate some of the evolutionary patterns and processes affecting bacterial genomes as detected with microarrays, and also delineate the limitations and conclusions stemming from such studies.

Keywords: Gene contents, Genome evolution, Microarrays

Contents

1. Introduction

Microevolution, the changes that occur within a population or species, has largely been examined on a gene-by-gene basis. As such, the analysis of each additional gene or trait across organisms has usually meant a commensurate and often substantial increase in research effort, thereby dashing hopes that such studies might be conducted at the level of whole genomes. Although the availability of the genome sequences of closely related strains has allowed scrutiny of microevolutionary processes within certain microbial species, we are still far from recovering such information for experiments of any scope or design. Fortunately, full genome sequences can serve as the basis for the fabrication of microarrays, which provide the potential to assay virtually every gene within a genome in a single hybridization experiment (Fitzgerald and Musser, 2001; Murray et al., 2001; Schoolnik, 2002).

Microarrays have most often been employed to monitor gene expression across all genes within a genome (RNA/cDNA-probe experiments), but there has been an increased

* Corresponding author. Tel.: +1 520 626 8355; fax: +1 520 621 3709.
E-mail address: hochman@email.arizona.edu (H. Ochman).

1567-1348/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2004.09.002

use of microarrays to assess gene inventories (DNA-probe experiments), particularly in microorganisms. From the start, we should mention that such DNA-probe experiments, although providing information about the full complement of genes from a particular genome, have some notable limitations. First, only the profile of genes that are shared by the reference and test strains, as well as the genes present only in the reference strain, can be identified. Next, such experiments do not recognize genes or regions of any length that are unique to the test strain, nor most structural rearrangements between the two organisms. Finally, non-specific hybridization, imaginable when probing several thousand spotted genes, will produce false positives, whereas highly diverged homologs might yield false negatives. Despite these shortcomings, microarrays offer a robust approach to the study of microevolutionary processes of bacterial genomes for a fraction of the expense and time of more traditional methods.

2. Defining species by gene composition

When it is neither practical nor necessary to produce complete genome sequences, microarrays provide an expedient means for approximating the gene repertoire. Microarrays have been employed to resolve numerous questions regarding the changes as well as the functional consequences of variation in genome composition among closely related bacterial strains and species. This approach has already defined the ''core'' (conserved) and ''dispensible'' (sporadically distributed) genes for numerous species (see Table 1). As anticipated, the percentage of genes classified as belonging to this ''core'' varies widely across bacterial species, not only due to the overall extent of diversity within the ''species'', but also to the relationships of the sampled genomes to the particular reference strain included on the array. In comparisons of 24 strains of *Salmonella enterica*, representing the six known subspecies, only 55% of the assayed ORFs were assigned as ''core'' genes (Chan et al., 2003). In contrast, analyses of *Helicobacter pylori* (Salama et al., 2000) and *Staphylococcus aureus* (Fitzgerald et al., 2001) each found that nearly 80% of ORFs were shared among the strains examined whereas over 94% of the genes were common to all strains in a survey of 100 clinical isolates of *Mycobacterium tuberculosis* (Tsolaki et al., 2004). In the extreme, nine strains of *Vibrio cholerae* isolated over the last century had the largest set of ''core'' genes, sharing approximately 99% of the ORFs present in the fully sequenced genome of seventh pandemic El Tor strain N16961 (Dziejman et al., 2002). It should also be noted that array quality (e.g., synthesis and/or hybridization conditions) can greatly influence analysis, and thus the conclusions regarding the number of ''core'' and ''dispensable'' genes, further highlighting the need for stringent experimental controls and cautious interpretation of data.

Because such analyses focus on rather restricted subsets of the diversity within a species and, additionally, cannot account for the often substantial numbers of genes unique to the test strains, the actual proportion of the genome devoted to ''core'' genes can not be assessed. Due to lateral gene transfer, the amount of ''dispensable'' DNA embraced by a species is likely to be enormous; but based on broad surveys of *Campylobacter jejuni* (Dorrell et al., 2001), *Escherichia coli* (Dobrindt et al., 2003), *H. pylori* (Salama et al., 2000), *M. tuberculosis* (Tsolaki et al., 2004), *S. enterica* (Chan et al., 2003) and *S. aureus* (Fitzgerald et al., 2001), it seems at least 50% of the genes in a particular strain are distributed in all other members of the species.

The fraction of the genome devoted to core genes is governed by several factors, including bacterial lifestyle and the opportunity for gene acquisition. For example, the gene repertoire of *Buchnera aphidicola*, the bacterial endosymbionts of aphids, has been static over hundreds of millions of years (Tamas et al., 2002), whereas closely related strains of *E. coli* can contain upwards of 30% unique DNA (Ochman and Jones, 2000; Welch et al., 2002). Among the most important aspects of these genome comparisons relates to the size of the regions whose occurrence is polymorphic among strains within a species. Although it was known that gene acquisition events can encompass numerous genes, it was thought that large-scale deletions were rare because

Table 1
Array-based estimates of ORFs shared within selected bacterial species

Species	# ORFs[a]	# Strains	Shared[b] (%)	Reference
Campy lobacter jejuni	1654	11	79	Dorrell et al. (2001)
Escherichia coli	4290	5	88	Ochman and Jones (2000)
Helicobacter pylori	1660	15	78	Salama et al. (2000)
Mycobacterium tuberculosis	3924	19	98	Kato-Maeda et al. (2001)
Mycobacterium tuberculosis	3924	100	94	Tsolaki et al. (2004)
Salmonella enterica	4169	24	54	Chan et al. (2003)
Staphylococcus aureus	2817	36	78	Fitzgerald et al. (2001)
Streptococcus pyogenes	2137	36	90	Smoot et al. (2002)
Vibrio cholerae	3632	9	99	Dziejman et al. (2002)

In *E. coli*, this value is underestimated because lone ORFs missing from one or more strains were not considered.

[a] Total number of different genes included on the array.

[b] Computed as the proportion of total ORFs occurring in all strains tested.

most would include essential genes and be deleterious to the organism. However, the large proportion of dispensable genes within a genome, as well as the large number of multi-gene regions that are absent from any given genome, argues that large deletions have played a major role in the evolution and diversification of many bacterial species (Mira et al., 2001, Moran and Mira, 2001).

3. Functional aspects of gene content variation

Because the microarrays themselves are based on fully annotated genomes, there is usually a wealth of information about the function of genes that are shared by, or missing from, specific strains. As expected, the core set of genes encodes housekeeping functions, whereas dispensable regions are likely to confer properties that are variable among strains. In *S. enterica*, dispensable regions specific to subspecies 1 are likely responsible for the ability to infect mammals and birds exclusively (Porwollik et al., 2002). Likewise, the presence of the HHGI1 island in pathogenic strains of *H. hepaticus* is associated with liver disease in mice (Suerbaum et al., 2003). Unfortunately, the more usual case is that strains differ by more than a single gene or region, which, when coupled with the unknown influence of any point mutational or regulatory variation, complicates identification of the precise changes responsible for a particular phenotype. For example, no unique genomic regions were detected in those *C. jejuni* isolates associated with Guillian–Barre syndrome in comparisons with enteritis-related strains (Leonard et al., 2004). And similarly, no particular strain-specific regions were identified among the strains of *Pseudomonas aeruginosa* most commonly recovered from opportunistic human infections (Wolfgang et al., 2003) or among *S. enterica* serotype Typhimurimum isolates with restricted host ranges (Andrews-Polymenis et al., 2004). The lack or loss of particular genes or chromosomal regions often has no discernable effect on phenotype or pathogenicity. In *M. tuberculosis*, over 5% of the genome (224 genes spanning the major functional categories, Camus et al., 2002) was found to be dispensable, with strains still capable of causing infection despite lacking up to 50 of these genes (Tsolaki et al., 2004).

In some cases, array-based genome comparisons have pointed to regions having a potential role in the diversification of strains and whose functions have subsequently been confirmed by experimental or analytical studies. For example, an attenuated strain of *H. pylori* lacked a segment of the *cag* pathogenicity island required for full virulence (Israel et al., 2001) and only methicillin-resistant strains of *S. aureus*, although phylogenetically diverse, contained portions of the 50 gene *mec* cassette (Fitzgerald et al., 2001). Microarray comparisons of 36 isolates of *Staphalococcus aureus* recovered from humans, sheep and cows identified 18 large chromosomal regions that differed with respect to strain COL (Fitzgerald et al., 2001), and 10 of these regions contained putative virulence factors or proteins mediating antibiotic resistance (Fitzgerald et al., 2001). An in depth sequence analysis of one of the sporadically distributed regions of the *S. aureus* chromosome (RD13) revealed that several mechanisms, including events of gene acquisition, recombination, purifying selection, and deletions, have led to present-day configurations of genes within this region (Fitzgerald et al., 2003).

In laboratory populations of *E. coli* propagated at near lethal temperatures, Riehle et al., (2001) detected a duplication that arose independently at the same chromosome location in three of the evolved lines. Although the convergence of traits by independent lineages can be viewed as strong circumstantial evidence of adaptive evolution, the specific genes within this duplication, such as *rpoS* and *pcm*, which both function in stress response, provide additional support for its role in the adaptation to higher temperatures.

4. Chronicling changes in genome composition

In a study that identified which portions of the *M. tuberculosis* H37Rv genome were absent from 19 clinical isolates, the test strains lacked, on average, 13 kb present in the H37Rv reference strain, and there was also a overall reduction in certain virulence attributes associated with the amount of missing DNA (Kato-Maeda et al., 2001) [but see above]. In addition, a study of an additional 100 *M. tuberculosis* strains suggest that mobile elements are deleted at a significantly higher rate than expected by chance (Tsolaki et al., 2004). These analyses presumed that all of the observed variation resulted from deletions, implying that the present gene composition of *M. tuberculosis* H37Rv is the ancestral state. However, the polarity of many events cannot be assigned unambiguously, and it is likely that some of the differences in gene contents, such as those regions associated with phage or other mobile elements, were due to the appropriation of sequences by H37Rv relative to the various test strains.

By knowing the genealogy of strains, such that the ancestral state of a character (in this case, presence of a particular gene) can be deduced, it is possible to trace the history of deletions that occurred within a bacterial lineage. Deletions present in multiple strains of *M. tuberculosis* can be traced back to their occurrence in a single ancestor (Hirsh et al., 2004); however, many of the same or similar deletions have occurred independently in different lineages, suggests that some regions of the chromosome may be more susceptible than others to deletions or that these particular deletions are favored by natural selection (Tsolaki et al., 2004). A phylogenetic approach has also been applied to investigate the variation among the attenuated strains of *M. bovis* that are used as vaccines against tuberculosis (termed BCG), which were derived from a vaccine strain produced by Calmette and Guerin (Behr et al., 1999). Current vaccine

strains of *M. bovis* are highly polymorphic due to the fact that until methods allowing preservation of permanent stocks were developed in the 1960s, cultures had been under continuous propagation since the 1920s. This resulted in BCG strains that experienced perhaps a thousand separated passages and in vaccines that confer different levels of immunity. Testing *M. bovis* BCG strains against gene arrays based on the closely related but virulent *M. tuberculosis* H37Rv identified 16 regions that are absent from one or more strains of *M. bovis*. Nine regions, ranging from 2 to 12 kb were missing from all *M. bovis* strains tested and are a likely source of the phenotypic differences between *M. tuberculosis* and *M. bovis*. Furthermore, one 9-kb region was absent from all BCG strains of *M. bovis* and thus lost during the original derivation of the attenuated strain by Calmette and Guerin. Aside from showing the patterns of genome evolution, the incidence of sporadically distributed deletions among BCG strains identified candidate regions that might be responsible for the lineage-specific differences in vaccine efficacy (Behr et al., 1999).

It is also possible to distinguish insertions (into the reference strain) from deletions (out of the test strain) by tracing the occurrence of ORFs along a pre-established phylogeny, such that the ancestral state of polymorphic regions can be inferred. Because *E. coli* have been subject to large amounts of gene acquisition but its genome size is not ever expanding, this method has been used to estimate the relative amounts of gene gain and gene loss occurring over the evolution of the species (Ochman and Jones, 2000). Based on the distribution of the 4290 *E. coli* MG1655 ORFs among natural and laboratory strains of *E. coli* of known phylogenetic relationships, insertions and deletion were each found to span several kilobases; but, on average, the insertions detected by this technique were larger, more numerous and more variable in base composition than were deletions.

In *Streptomyces coelicolor*, a screening of 21 laboratory isolates derived from strain A3(2) found 10 strains possessing an additional 1.06 Mb relative to the completely sequenced genome of *S. coelicolor* M145 (Weaver et al., 2004). This single insertion accounts for an 11% increase in genome size (9.7 Mb compared to 8.6 Mb) between isolates and was localized to an inverted repeat sequence in the right end of the *S. coelicolor* linear chromosome (Weaver et al., 2004). Since the histories of the strains were well documented, Weaver et al. (2004) determined that the duplication represented the chromosomal state of the original isolate and was subsequently truncated in strains that lacked it. A similar change in chromosome organization was observed in archived cultures of *S. enterica* serovar Typhimurium LT2: after >40 years of storage, one of 14 isolates possessed an insertion encompassing nearly 4% of the chromosome, which resulted from a duplication and translocation event mediated by recombination at *rrn* operons (Porwollik et al., 2004).

5. Compositional profiling across taxa

Having a complete genomic sequence, much less a gene array, for any organism of choice is still a rare luxury. To date, microarrays have been synthesized for only 10–20% of the 100 + sequenced bacteria, and even fewer are readily available. Among the first commercially distributed gene arrays were the complete set of annotated *E. coli* ORFs, and several studies have used these filters to explore variation in gene distribution or expression beyond *E. coli*.

Most often, the *E. coli* arrays have been used to assess the gene composition of other bacteria that are sufficiently closely related to *E. coli* to allow the unambiguous cross-hybridization of homologous genes. The amount and quality of information obtained from such interspecies comparisons depend on the genome size of the test organism, its phylogenetic proximity to *E. coli*, and level of sequence divergence. For example, approximately 3000 of the *E. coli* ORFs are present in the maize endophyte *Klebsiella pneumoniae* (Dong et al., 2001), slightly lower than the number of genes common to *E. coli* and *Salmonella enterica*, which share a more recent common ancestor. [Based of comparisons of fully sequenced genomes, nearly 2500 genes are shared among all enterics (McClelland et al., 2000, 2001)]. Unlike full sequence determination, microarray studies are rapid, but, as already discussed, they provide no information about the genes that are unique to the test strain. Given that the genome size of this strain of *K. pneumoniae* is 4.8 Mb, we estimate that this technique revealed >70% of the coding capacity of this organism and that at least 1500 *Klebsiella* genes remain unidentified.

The gene inventories of two other bacteria, *Wigglesworthia glossinidia* and *Sodalis glossinidus*, the primary and secondary endosymbionts of tsetse flies, respectively, have been assessed on *E. coli* microarrays. These symbionts have reduced genomes, which increased the probability that their constituent genes are present on the *E. coli* gene arrays. In both cases, the authors reported that a substantial fraction of genes were catalogued with the *E. coli* arrays (Akman and Aksoy, 2001; Akman et al., 2001); however, subsequent sequencing of the *Wigglesworthia* genome (Akman et al., 2002) suggests that many of these original assignments were incorrect. Based on the phylogenetic distance between *E. coli* and *Wigglesworthia* (Lerat et al., 2003) as well as the extreme differences in their base compositions, it is not surprising that such heterologous hybridizations might miss large numbers of genes and produce some false signals. Recent studies confirm that gene arrays are most accurate when levels of sequence identity are over 80% (Evertsz et al., 2001), and this value is rarely exceeded in comparisons of *E. coli* and *Wigglesworthia* homologs. Although gene arrays sometimes offer the only opportunity to assay the gene inventories of a bacterial strain or species, such results suggest that this

approach should be applied prudently and in cases where results can be confirmed by other methods.

References

Akman, L., Aksoy, S., 2001. A novel application of gene arrays: *Escherichia coli* array provides insight into the biology of the obligate symbiont of tsetse flies. Proc. Natl. Acad. Sci. U.S.A. 98, 7546–7551.

Akman, L., Rio, R.V.M., Beard, C.B., Aksoy, S., 2001. Genome size and coding capacity of *Sodalis glossinidius*, an enteric symbiont of tsetse flies, as revealed by hybridization analysis to *Escherichia coli* gene arrays. J. Bacteriol. 185, 4517–4525.

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., et al., 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies *Wigglesworthia glossinidia*. Nat. Genet. 32, 402–407.

Andrews-Polymenis, H.L., Rabsch, W., Porwollik, S., McClelland, M., Rosetti, C., Adams, L.G., Baumler, A.J., 2004. Host restriction of *Salmonella enterica* serotype Typhimurium pigeon isolates does not correlate with loss of discrete genes. J. Bacteriol. 186, 2619–2628.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., Small, P.M., 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284, 1520–1533.

Camus, J.C., Pryor, M.J., Medigue, C., Cole, S.T., 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiol. 148, 2967–2973.

Chan, K., Baker, S., Kim, C.C., Detweiler, C.S., Dougan, G., Falkow, S., 2003. Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovars Typhimurium DNA microarray. J. Bacteriol. 185, 553–563.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., et al., 2003. Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. J. Bacteriol. 185, 1831–1840.

Dong, Y., Glasner, J.D., Blattner, F.R., Triplett, E.W., 2001. Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte, *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. Appl. Environ. Microbiol. 67, 1911–1921.

Dorrell, N., Mangan, J.A., Laing, K.G., Hinds, J., Linton, D., Al-Ghusein, H., et al., 2001. Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. Genome. Res. 11, 1706–1715.

Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., Mekalanos, J.J., 2002. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc. Natl. Acad. Sci. U.S.A. 99, 1556–1561.

Evertsz, E.M., Au-Young, J., Ruvolo, M.V., Lim, A.C., Reynolds, M.A., 2001. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. Biotechniques 31, 1182–1192.

Fitzgerald, J.R., Musser, J.M., 2001. Evolutionary genomics of pathogenic bacteria. Trends Microbiol. 9, 547–553.

Fitzgerald, J.R., Reid, S.D., Ruotsalainen, E., Tripp, T.J., Liu, M.Y., Cole, R., et al., 2003. Genomic diversification in *Staphylococcus aureus*: molecular evolution of a highly variable chromosomal region encoding the staphylococcal exotoxin-like family of proteins. Infect. Immun. 71, 2827–2838.

Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., Musser, J.M., 2001. Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. Proc. Natl. Acad. Sci. U.S.A. 98, 8821–8826.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc. Natl. Acad. Sci. USA 101, 4871–4876.

Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S., Peek Jr., R.M., 2001. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. Proc. Natl. Acad. Sci. USA 98, 14625–14630.

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., Small, P.M., 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11, 547–554.

Leonard II, E.E., Tompkins, L.S., Falkow, S., Nachamkin, I., 2004. Comparison of *Campylobacter jejuni* isolates implicated in Guillain-Barre syndrome and strains that cause enteritis by a DNA microarray. Infect. Immun. 72, 1199–1203.

Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLOS Biol. 1, 101–109.

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latrelille, P., Courtney, L., et al., 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413, 852–856.

McClelland, M., Florea, L., Sanderson, K.E., Clifton, S.W., Parkhill, J., Churcher, C., et al., 2000. Comparison of the *Escherichia coli* K-12 genome with sampled genomes of *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium Typhi and Paratyphi. Nucl. Acids Res. 28, 4974–4986.

Moran, N.A., Mira, A., 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. 2, 0054. 1–0054.12.

Mira, A., Ochman, H., Moran, N.A., 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 10, 589–596.

Murray, A.E., Lies, D., Li, G., Nealson, K., Zhou, J., Tiedje, J.M., 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Natl. Acad. Sci. U.S.A. 98, 9853–9858.

Ochman, H., Jones, I.B., 2000. Evolutionary dynamics of full genome content in *Escherichia coli*. EMBO J. 19, 6637–6643.

Porwollik, S., Mei-Yi Wong, R., McClelland, M., 2002. Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. Proc. Natl. Acad. Sci. U.S.A. 99, 8956–8961.

Porwollik, S., Mei-Yi Wong, R., Helm, R.A., Edwards, K.K., Calcutt, M., Eisenstark, A., et al., 2004. DNA amplification and rearrangements in archival *Salmonella enterica* serovars Typhimurium LT2 cultures. J. Bacteriol. 186, 1678–1682.

Riehle, M.M., Bennett, A.F., Long, A.D., 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 98, 525–530.

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., Falkow, S., 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. U.S.A. 97, 14668–14673.

Schoolnik, G.K., 2002. Functional and comparative genomics of pathogenic bacteria. Curr. Opin. Microbiol. 5, 20–26.

Smoot, J.C., Barbian, K.D., Van Gompel, J.J., Smoot, L.M., Chaussee, M.S., Sylva, G.L., et al., 2002. Genome sequence and comparative microarray analysis of serotype Ml8 group A Streptococcus strains associated with acute rheumatic fever outbreaks. Proc. Natl. Acad. Sci. U.S.A. 99, 4668–4673.

Suerbaum, S., Josenhans, C., Sterzenback, T., Drescher, B., Brandt, P., Bell, M., et al., 2003. The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. Proc. Natl. Acad. Sci. U.S.A. 100, 7901–7906.

Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., et al., 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296, 2376–2379.

Tsolaki, A.G., Hirsh, A.E., DeRiemer, K., Enciso, J.A., Wong, M.Z., Hannan, M., et al., 2004. Functional and evolutionary genomics oi *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. Proc. Natl. Acad. Sci. U.S.A. 101, 4865–4870.

Weaver, D., Karoonuthaisiri, N., Tsai, H.H., Huang, C.H., Ho, M.L., Gai, S., et al., 2004. Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. Mol. Microbiol. 51, 1535–1550.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., et al., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 99, 17020–17024.

Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., et al., 2003. Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. Proc. Natl. Acad. Sci. U.S.A. 100, 8484–8486.