Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

ELSEVIER

Comparative phylogenomics of pathogenic bacteria by microarray analysis

Nick Dorrell, Stewart J Hinchliffe and Brendan W Wren

DNA microarrays represent a powerful technology that enables whole-scale comparison of bacterial genomes. This, coupled with new methods to model DNA microarray data, is facilitating the development of robust comparative phylogenomics analyses. Such studies have dramatically increased our ability to differentiate between bacteria, highlighting previously undetected genetic differences and population structures and providing new insight into virulence and evolution of bacterial pathogens. Recent results from such studies have generated insights into the evolution of bacterial pathogens, the levels of diversity and plasticity in the genome of a species, as well as the differences in virulence amongst pathogenic bacteria.

Addresses
Department of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

Corresponding author: Wren, Brendan W (brendan.wren@lshtm.ac.uk)

Current Opinion in Microbiology 2005, **8**:620–626

This review comes from a themed issue on
Genomics
Edited by Gerhard Gottschalk and Stephan C Schuster

Available online 24th August 2005

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2005.08.012

Introduction

Traditional phylogenetic classification of bacteria to study evolutionary relatedness is based on the characterisation of a limited number of genes, rRNA or signature sequences. However, owing to the acquisition of DNA through lateral gene transfer, the differences between closely related bacterial strains can be vast. By contrast, whole-genome sequencing comparisons allow a multitude of genes to be compared. Already, several bacterial species have had more than a single representative sequenced (e.*g. Escherichia coli* [1], *Campylobacter jejuni* [2], *Helicobacter pylori* [3], *Yersinia pestis* [4], *Neisseria meningitides* [5], *Staphylococcus aureus* [6] and several *Chlamydia* species [7]). Nevertheless, whole-scale genome sequencing remains an expensive endeavour and such comparisons are limited to only a handful of strains.

DNA microarrays represent an alternative technology for whole-genome comparisons, enabling a 'bird's-eye view' of all the genes absent or present in a given genome compared to the reference genome on the microarray. Harnessing DNA microarray information through interrogative and robust algorithms has enabled a true 'comparative phlyogenomics' approach to be developed. Recent comparative phlyogenomics studies have been undertaken on increasingly large collections of strains of defined origin. One common feature of these studies is the unexpectedly large genetic diversity between strains of the same species, blurring our definition of species boundaries. Whole-genome comparisons typically identify sets of 'core genes', which are shared by all strains in a species, and 'accessory genes', which are present in one or more strains in a species and often result from gene acquisition. It is these differences that can often be used to identify genes and/or genetic islands related to 'gain-of-function traits' in pathogenic strains. Uncovering the mechanisms behind this variability is fundamental in understanding and ultimately counteracting the threats posed by our old adversaries.

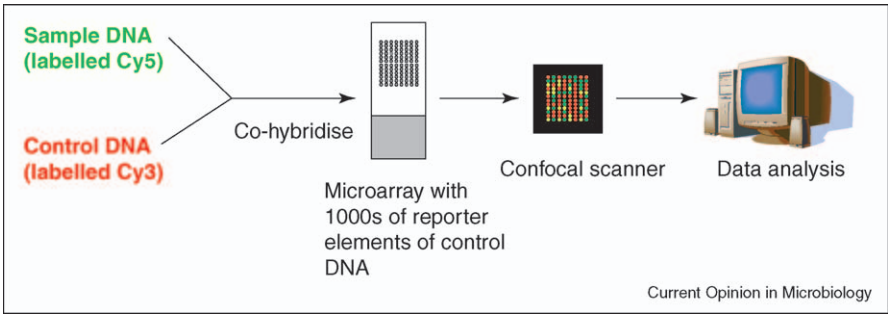
This review will describe selected recent examples in which comparative phylogenomics using microarray data has provided new and often unexpected insights into the evolution, ecology, virulence and genome diversity of bacterial pathogens.

Comparative phylogenomics microarray methodology

Comparative phylogenomics studies involve analysis of the relative hybridisation capacity of sample bacterial DNA to reporter DNA elements that are bound to a microarray surface. DNA microarrays consist of minute samples of DNA (known as reporter elements) that are arrayed in a grid formation on a suitable medium. Each predicted coding sequence within a target genome or genomes is represented on the array by at least one specific reporter element. All DNA reporter elements are designed to minimise cross-hybridisation (hybridisation of DNA from one gene to reporter elements that represent other genes owing to sequence similarity) between spots.

Sample genomic DNA from a test strain is labelled with a fluorescent marker (usually Cy3 or Cy5) and is competitively hybridised with genomic DNA from a control strain that has been labelled with a different fluorescent marker (Figure 1). The control strain is usually DNA from the strain used to design the array, as this should hybridize to all reporter elements on the microarray. After washing to remove any unbound DNA, microarrays are scanned by a

Figure 1



Principles of comparative phylogenomics using DNA microarrays. For further details, see main body of text.

high-resolution fluorescence scanner, and the image obtained is analysed by specific software packages that are not discussed in this review. Control spots on the array and DNA 'spikes' in the labelling reaction can be used to aid normalisation of array data. A comparison of the levels of test and control DNA bound to each reporter element on the array can be performed and statistical methods are utilized to determine if the coding sequence in question is present, absent or highly divergent in the test strain.

Which type of DNA microarray should be used?

Polymerase chain reaction product-based DNA microarrays

DNA microarrays often consist of gene fragments (reporter elements) amplified by polymerase chain reaction (PCR) that represent individual genes. Currently, up to 50 000 gene fragments can be spotted onto a single microscope slide using robotic technology. Advantages of this technology are: flexibility in the design of the array; the relative ease of production; and its relatively low cost. Multiple identical microarrays can be robotically printed in batches of over a hundred in a single run. Most of the cost in printing such arrays is caused by the synthesis of oligonucleotide primer pairs required for the amplification of target gene fragments. The use of clone libraries as the template for the amplification of PCR products can dramatically reduce this cost [8].

Oligonucleotide-based DNA microarrays

Oligonucleotide microarrays do not rely upon PCR amplification of gene targets, and areas of specificity within any gene sequence can be chosen for hybridisation analysis. However, the sequence of target DNA must be known before oligonucleotide synthesis. The two most frequently used formats are the Affymetrix and the Qiagen Operon systems. In the Affymetrix format (www.affymetrix.com), oligonucleotides are synthesized *in situ* on a derivatised glass surface using a combination of photolithography and combinatorial chemistry. The synthesized oligonucleotides are usually 20–25 bases in length. The GeneChip® *E. coli* Genome Array is the first Affymetrix microarray product for the analysis of gene expression in a prokaryotic organism. The Qiagen Operon format (www.operon.com) uses an optimised 70-mer oligonucleotide to represent each gene in a genome. Each 70-mer is designed to have optimal specificity for its target gene and is melting-temperature normalised.

Oligonucleotide-based microarrays offer many advantages over PCR-product-based microarrays including a reduction in cross-hybridisation and an increase in the differentiation of overlapping genes or highly homologous regions. For example, mutant alleles or single nucleotide polymorphisms can be detected using such oligonucleotide microarrays owing to the shorter probe size compared to PCR product-based microarrays.

Comparative phylogenomics to study the evolution of bacterial pathogens

Comparative phylogenomics using microarray data can be particularly useful for retrospective analyses of the recent (past hundred years) and the longer term (several thousand years) evolution of bacterial pathogens.

Recent evolution of bacterial pathogens usually involves lateral gene transfer and large-scale genomic deletions. Such genetic differences are more evident in the highly clonal species such as *Vibrio cholera*e, in which microarray studies showed that all strains tested were <1% divergent from each other over their entire genome [9]. These differences provide clues to the evolution of *V. cholerae* as a human pathogen and as an environmental organism. Historically, the first six *V. cholerae* pandemics between 1817 and 1923 were caused by classical biotype strains from the O1 serogroup. However, the emergence of the seventh pandemic El Tor biotype in 1961 eventually resulted in the global elimination of classical biotype strains as a cause of disease. Microarray analysis revealed chromosomal islands specific to all pandemic strains as well as two chromosomal islands that are uniquely present in the modern-day pandemic El Tor strain. This suggested that these regions might encode factors specifically associated with displacement of the pre-existing classical

strains in South Asia and might also promote the establishment of endemic disease in previously cholera-free locations [9].

A comprehensive comparative genomics study of pathogenic *S. aureus* revealed that ~22% of the genome is composed of accessory genes [10]. These are mainly grouped in large chromosomal regions of difference (RDs), and are often associated with phage genes, integrases or transposases. Of the 18 RDs identified, 10 contained putative virulence determinants (including enterotoxins and the methicillin-resistance gene *mec*) that appear to have been acquired independently into the chromosome background of several phylogenetically distinct strains. This demonstrated that methicillin-resistant strains have evolved multiple independent times, rather than from a single ancestral strain. It has also helped to resolve the long-standing controversy that the 1970s epidemic of toxic shock syndrome was caused by a change in the host environment (i.e. the use of certain tampons), rather than by the rapid geographic dissemination of a new hypervirulent *S. aureus* strain [10]. Interestingly, the epidemic methicillin-resistant *S. aureus*-15 (EMRSA-15) clone, one of the most prevalent MSRA strains in the UK, appears to have resulted from a large-scale recombination event [11]. Other microarray studies of this kind have recently identified specific RDs and coding sequences that might be responsible for hypervirulence of certain strains in animal models of muscoskeletal infections [12].

Yersinia pestis, the causative agent of plague, diverged from *Yersinia pseudotuberculosis*, an enteric pathogen, an estimated 1500–20 000 years ago [13,14]. Genetic characterization of these closely related organisms represents a useful model to study the rapid emergence of bacterial pathogens that threaten mankind. Studies using a *Y. pestis*-specific microarray identified 11 *Y. pestis* DNA loci that were absent or highly divergent in all 22 strains tested. Four were regions of phage origin, whereas the other seven included genes that encode a vitamin B12 receptor and the insect toxin SepC. These studies help predict the genome content of the single *Y. pseudotuberculosis* strain that evolved into the *Y. pestis* (sub)species [15]. Microarray-based studies to determine the evolution of *Y. pestis* have shown that recombination between the frequent insertion sequence elements scattered around the chromosome has resulted in large-scale deletions in most strains [15,16•]. Recent studies on *Y. pestis* isolates from Brant's vole plague foci in China have shown these strains (renamed *Yersinia microtus*) were attenuated in the human host. Microarray and sequencing analysis of these strains shows a specific pattern of loss of particular regions of the genome including five regions that are present in all virulent strains tested [16•]. Additionally, several specific mutations were identified that are present in all *Y. microtus* strains but absent in human virulent strains. This study has therefore successfully identified regions of the *Y. pestis* genome responsible for *Y. microtus* niche-adaptation and for avirulence in humans, as well as mapping the accumulation of deletions in fully virulent *Y. pestis* strains as they spread among vole populations throughout China.

Comparison of the virulent *Mycobacterium tuberculosis* reference strain H37Rv with virulent *Mycobacterium bovis* strains and various Bacille Calmette-Guerin (BCG) daughter strains identified 11 regions (91 open reading frames [ORFs]) of H37Rv that were found to be absent from one or more *M. bovis* strains. Also, 5 additional regions (38 ORFs) were identified to be present in *M. bovis* strains but absent in some or all of the BCG strains [17]. This is indicative of ongoing genetic changes in BCG strains and offers a possible explanation for the loss of protective efficacy since their original derivation in 1921. Further studies have suggested that the common ancestor of the tubercle bacilli resembled *M. tuberculosis* or *Mycobacterium canettii* and might have been a human pathogen already, which contradicts the hypothesis that *M. tuberculosis* evolved from *M. bovis* [18].

Comparative phylogenomics to study genome diversity and plasticity

Many of the early comparative genomic studies investigated the presence or absence of genes in test strains compared to the sequenced strain represented on the array. Such studies have tended to be more informative for non-clonal populations of bacteria. Genome diversity refers to strains of the same species possessing many different accessory genes in different combinations, whereas genome plasticity refers to gain or loss of specific genes resulting in phenotypic changes.

Extensive genomic diversity between pathogenic *E. coli* and *Shigella* strains has been shown by way of microarray analysis [19•,20]. The core genes for *E. coli* consist of about 2800 ORFs, and unique relationships between enteroinvasive *E. coli* strains and *Shigella* and between uropathogenic and enteropathogenic *E. coli* strains have been identified [19•]. Similar studies for *Salmonella* species have also been reported, which suggests that there is a high level of gene gain and gene loss as well as rapid divergence along all lineages [21,22].

Analysis of genetic diversity amongst different clinical isolates of the human gastric pathogen *H. pylori* has demonstrated that ~22% of the *H. pylori* genome is composed of accessory genes [23]. Although the core genes encode most metabolic and cellular processes, the accessory genes include those that might be important for long-term infection of a genetically diverse host. Indeed *H. pylori* has been shown to exhibit genetic diversity within the stomach of a single human host [24]. Comparison of two clinical isolates that produced different levels of inflammation in the gerbil gastritis

model identified a large deletion in the *cag* pathogenicity island in the less proinflammatory strain [25].

The initial observation on genetic variability in the gastrointestinal pathogen *C. jejuni* [8] has been expanded by several recent studies [26–28], which suggest that *C. jejuni* exhibits high levels of genome diversity and low levels of genome plasticity [28]. Unfortunately, it has proved difficult to use microarray data to identify genetic markers that can predict the source or outcome of infection. Infection with *C. jejuni* can result in the development of the post-infectious neurological disorder Guillian-Barré syndrome (GBS). A comparison of isolates from patients who developed GBS with isolates from patients with uncomplicated gastrointestinal infection failed to identify specific GBS genes or regions [29].

A large chromosomal region (RD13), common to all strains of *S. aureus*, encodes a family of proteins designated staphylococcal exotoxin-like (SET) proteins. Extensive genetic diversity and variation has been shown in the *set* gene content [30]. This is indicative of multiple episodes of *set* gene deletion in different clonal lineages in parallel. There was evidence that horizontal gene transfer and recombination have also contributed to the diversification of RD13.

By contrast, not all pathogens exhibit extensive genetic diversity. Surprisingly, microarray analysis of serotype M18 group A *Streptococcus* (GAS) strains that are associated with acute rheumatic fever showed little or no variation in gene content [31]. In fact, the data indicated that M18 strains recovered during two acute rheumatic fever outbreaks that occurred 12 years apart were genetically identical.

Analysis of the genomes of 100 epidemiologically well-characterised *M. tuberculosis* clinical isolates showed that deletions were not distributed randomly, but were tightly aggregated throughout the genome [32•]. Some aggregations appear to be caused by a genetic event specific to a single mycobacterial lineage, whereas others reveal regions of genomic vulnerability throughout the species. The majority of deletions are likely to be slightly deleterious to *M. tuberculosis*. However, some deletions could offer short-term advantages of escape from the hosts immune system, and others could confer strong advantages such as antibiotic resistance. An example of the latter was seen through microarray analysis of clinical isolates from an outbreak of tuberculosis in Leicester, UK [33].

Comparative phylogenomics applied to decipher bacterial virulence and host specificity

Genome comparisons between pathogenic and non-pathogenic bacteria within a species (or genus) are particularly useful for the identification of determinants that are likely to be important for virulence, transmission and host specificity. Whole-genome analysis of strains with documented clinical histories, or of those of known host origin, is crucial for deciphering genetic clues related to virulence or host specificity. In terms of identifying potential virulence factors, the comparison of isolates from patients with a spectrum of disease (including asymptomatic carriage) is particularly useful for pathogens in which a suitable animal model of disease is unavailable.

Comparison of clinical and commensal *Staphylococcus epidermidis* strains revealed high genetic variability in the genome and identified several markers of invasiveness for this nosocomial pathogen [34]. A proline-betaine transporter homolog gene was more frequent among invasive strains, which suggests a previously unexpected role for osmoprotection factors during infection. By contrast, a similar study demonstrated that *Pseudomonas aeruginosa* strains possess a highly conserved genome that encodes genes important for survival in numerous environments, thus allowing this opportunistic pathogen to cause a variety of human infections [35]. However, clusters of strain-specific genes were also identified, termed variable segments, which appear to highlight the presence of preferential sites for integration of novel genetic material into the *P. aeruginosa* genome.

Neisseria gonorrhoeae causes an infection that is generally localized to the genitourinary tract, whereas *Neisseria meningitidis* can enter the bloodstream and cross the blood–brain barrier to invade the meninges. Genomic comparisons of different clinical isolates of both pathogens and of *Neisseria lactamica*, a commensal of the nasopharanx, highlighted genes that are present in both pathogenic *Neisseria* species but are absent from *N. lactamica*, and also identified genes specific to *N. meningitidis* strains [36]. These pathogenic-specific sequences were of relatively small size and scattered throughout the genome rather than being clustered in pathogenicity islands.

Non-O1, non-O139 *V. cholerae* can cause gastroenteritis and extraintestinal infections, but, unlike O1 and O139 strains of *V. cholerae*, little is known about the virulence gene content of non-O1, non-O139 strains and their phylogenetic relationship to other pathogenic *V. cholerae*. In a recent comparative genomic microarray analysis of four pathogenic *V. cholerae* strains (non-O1/non-O139), it was shown that these strains are quite divergent from O1 and O139 strains. Genomic sequence analysis of a non-O1/non-O139 strain that appeared to be particularly pathogenic in experimental animals suggested that this strain carries a type III secretion system (TTSS). The genes for this *V. cholerae* TTSS system appear to be present in many clinical and environmental non-O1/

non-O139 strains, including at least one clone that is globally distributed. The presence of the TTSS in some pathogenic non-O1/non-O139 *V. cholerae* might be involved in the virulence and environmental fitness of these strains [37•].

Mycobacterum microti, a member of the highly clonal *M. tuberculosis* complex, causes vole tuberculosis but is at least partially attenuated in other mammalian species including humans. Genomic comparisons of *M. microti* with *M. tuberculosis* identified a total of 13 deletions present in at least one of the twelve strains of *M. microti*, including ten deletions that were previously described for the human avirulent *M. bovis* [38]. Although the RD1 deletion is crucial for attenuation in *M. bovis* [39], only the RD3 deletion was present in all *M. microti* strains tested. The variation in the number of deletions in each strain might play a role in the different host ranges of the individual *M. microti* strains.

The molecular basis of the emergence of new virulent bacterial subclones is poorly understood and should prove a fruitful area for microarray-based research. For example, M3 GAS strains cultured from patients with invasive infection over an 11 year period that included two distinct infection peaks were analysed using genome-wide investigative methods including microarrays [40•]. All variation in gene content could be attributed to the acquisition or loss of prophages, resulting in strains that had unique combinations of virulence genes. Distinct serotype M3 genotypes experienced rapid population expansion and caused infections that differed significantly in character and severity.

Limitations of microarray technology for comparative genome analyses

Despite the high throughput capability and potential applications of microarrays for comparative genome analysis, they have inherent problems as they are limited by the genetic information on the array. Most microarrays that are currently in use represent the genome of a single sequenced strain only. However, changes in the virulence potential of bacterial species are largely influenced by horizontally transferred DNA, therefore standard microarrays are unable to identify acquired DNA in diverse strains. As more genomes are sequenced, coupled with techniques such as suppression subtractive hybridisation, the interrogative capacity of arrays will be increased further and information on arrays could be used to investigate a given bacterial genus or mixture of genomes. The *S. aureus* pan-array currently produced by BμG@S (http://bugs.sgul.ac.uk/) represents the genome of seven sequenced strains.

Even with gene-specific microarrays, for some microbial species potential problems will occur with cross-hybridisation between gene fragments that represent different but very similar genes, meaning that genetic rearrangements, insertions, inversions and duplications are difficult to detect using spotted PCR products. The move towards using oligonucleotides as reporter elements rather than larger PCR products will increase the specificity of arrays for the individual target gene.

Further potential problems can occur when using mixed bacterial species for microarray analysis. Because of the range of G+C content in bacterial genes and genomes, hybridisation conditions have to be carefully considered as organisms with different G+C content will have differing hybridisation efficiencies to a given microarray. Additionally, microarrays that use amplified products will not identify point mutations or single nucleotide polymorphisms, although suitably designed high-density oligonucleotide arrays (eg: Affymetrix GeneChips®) are able to distinguish such mutations. However, oligonucleotide arrays remain expensive and, in the case of the Affymetrix system, the technology itself is proprietary. Currently, there are only a limited number of whole microbial genome arrays in this format. As the number of oligonucleotide arrays for different microorganisms increases and competition reduces the cost of these arrays, this technology platform should predominate.

Another stumbling block to reaping the benefits of DNA microarrays is the lack of globally accepted standardised bioinformatic methods to analyse the data generated. Standardised analyses of the data are crucial for interpretation and comparison across different studies. Simple concepts such as the presence or absence of coding sequences within a particular strain rely on how cut-offs are determined, a process that differs between laboratories. In the future, sustained improvements in software and computing speed, coupled with the application of improved mathematical models and Bayesian-based algorithms, will dramatically increase the scale of problems we tackle using microarray-based technology to understand the basic biology and evolution of bacterial pathogens.

Conclusions

Microarray analysis is a powerful enabling technology that allows global comparative analysis of the gene content between different strains in a given species and even between species in a genus. Comparative phylogenomics that use standard microarrays from a range of bacterial pathogens, such as *H. pylori*, *C. jejuni*, *E. coli* and *S. aureus*, clearly demonstrate the diversity and adaptability of these specialized groups of organisms. Studies have revealed much evidence of lateral gene transfer and recombination. This supports an evolutionary scenario involving vertical diversification by mutagenesis, punctuated by frequent lateral gene transfer, resulting in a global mosaic genome structure. By contrast, *M. tuberculosis*, *V. cholerae* and *Y. pestis* appear to exhibit limited genome diversity,

but nevertheless are adaptable enough for pandemic strains to emerge.

Recent advances in high-density oligonucleotide arrays have allowed for the development of high throughput resequencing techniques. Resequencing arrays are designed to cover the entire genome by overlapping oligonucleotides. Multiple versions of each oligonucleotide are spotted on the array to represent the four possible base combinations (A, T, G and C) for each nucleotide position. Competitive hybridisation of test and control strains to the array allows the identification of single nucleotide polymorphisms and small deletions or insertions that would be missed by standard microarrays. To date, this technique has only been applied to *Bacillus anthracis*, with 56 strains being resequenced using a custom-designed resequencing array [41••]. As this technique becomes more robust, many more bacterial species will be investigated in this way. The same technique has also been used to track the evolution of the severe acute respiratory syndrome coronavirus [42].

The development of a database of nucleotide differences among strains should allow the design of a universal microbial pathogen microarray that would have wide applications in the study of the epidemiology, population genetics, molecular phylogeny and evolution of microbial pathogens, as well as having diagnostic applications. Currently we are developing an 'active surveillance pathogen' microarray that consists of genes from mobile elements such as pathogenicity islands, phage and plasmid sequences; this will have multiple applications. The active surveillance pathogen array could be used as an 'early warning system' to allow identification of the emergence of highly transmissible or virulent strains that could, for example, be traced back to an individual, or a flock or herd, and then eliminated, thus averting the spread of an emerging virulent strain. No doubt the knowledge garnered from these and other microarray studies will be applied to well-designed intervention strategies to help reduce the burden of infectious disease.

Acknowledgements

We wish to acknowledge the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC), the Defence Science Technology Laboratory (DSTL) and the Department for Environment, Food and Rural Affairs (DEFRA) for funding related research in this area in our laboratory.

References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al.*: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***. *Proc Natl Acad Sci USA* 2002, **99**:17020-17024.

2. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, Brinkac LM, DeBoy RT, Parker CT, Daugherty SC *et al.*: **Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species**. *PLoS Biol* 2005, **3**:e15.

3. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL *et al.*: **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori***. *Nature* 1999, **397**:176-178.

4. Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H *et al.*: **Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans**. *DNA Res* 2004, **11**:179-197.

5. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T *et al.*: **Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491**. *Nature* 2000, **404**:502-506.

6. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M *et al.*: **Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain**. *J Bacteriol* 2005, **187**:2426-2438.

7. Read TD, Myers GS, Brunham RC, Nelson WC, Paulsen IT, Heidelberg J, Holtzapple E, Khouri H, Federova NB, Carty HA *et al.*: **Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the *Chlamydiaceae***. *Nucleic Acids Res* 2003, **31**:2134-2147.

8. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV *et al.*: **Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity**. *Genome Res* 2001, **11**:1706-1715.

9. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ: **Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease**. *Proc Natl Acad Sci USA* 2002, **99**:1556-1561.

10. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM: **Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic**. *Proc Natl Acad Sci USA* 2001, **98**:8821-8826.

11. Saunders NA, Underwood A, Kearns AM, Hallas G: **A virulence-associated gene microarray: a tool for investigation of the evolution and pathogenic potential of *Staphylococcus aureus***. *Microbiology* 2004, **150**:3763-3771.

12. Cassat JE, Dunman PM, McAleese F, Murphy E, Projan SJ, Smeltzer MS: **Comparative genomics of *Staphylococcus aureus* musculoskeletal isolates**. *J Bacteriol* 2005, **187**:576-592.

13. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: ***Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis***. *Proc Natl Acad Sci USA* 1999, **96**:14043-14048.

14. Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR *et al.*: **Microevolution and history of the plague bacillus, *Yersinia pestis***. *Proc Natl Acad Sci USA* 2004, **101**:17837-17842.

15. Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, Nichols RA, Oyston PC, Hinds J, Titball RW, Wren BW: **Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis***. *Genome Res* 2003, **13**:2018-2029.

16. • Zhou D, Han Y, Song Y, Tong Z, Wang J, Guo Z, Pei D, Pang X, Zhai J, Li M *et al.*: **DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation**. *J Bacteriol* 2004, **186**:5138-5146.

This study suggests that modern plague originated in China owing to biovar orientalis that arose from biovar antiqua rather than from biovar mediaevalis.

17. Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM: **Comparative genomics of BCG vaccines by whole-genome DNA microarray**. *Science* 1999, **284**:1520-1523.

18. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K *et al.*: **A new evolutionary scenario for the *Mycobacterium tuberculosis* complex**. *Proc Natl Acad Sci USA* 2002, **99**:3684-3689.

19. • Fukiya S, Mizoguchi H, Tobe T, Mori H: **Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray**. *J Bacteriol* 2004, **186**:3911-3921.
Evidence for unique relationships between enteroinvasive *E. coli* and *Shigella*, and between uropathogenic and enteropathogenic strains of *E. coli*.

20. Anjum MF, Lucchini S, Thompson A, Hinton JC, Woodward MJ: **Comparative genomic indexing reveals the phylogenomics of *Escherichia coli* pathogens**. *Infect Immun* 2003, **71**:4674-4683.

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

22. Porwollik S, Wong RM, McClelland M: **Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis**. *Proc Natl Acad Sci USA* 2002, **99**:8956-8961.

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

24. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM Jr: ***Helicobacter pylori* genetic diversity within the gastric niche of a single human host**. *Proc Natl Acad Sci USA* 2001, **98**:14625-14630.

25. Israel DA, Salama N, Arnold CN, Moss SF, Ando T, Wirth HP, Tham KT, Camorlinga M, Blaser MJ, Falkow S *et al.*: ***Helicobacter pylori* strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses**. *J Clin Invest* 2001, **107**:611-620.

26. Leonard EE II, Takata T, Blaser MJ, Falkow S, Tompkins LS, Gaynor EC: **Use of an open-reading frame-specific *Campylobacter jejuni* DNA microarray as a new genotyping tool for studying epidemiologically related isolates**. *J Infect Dis* 2003, **187**:691-694.

27. Pearson BM, Pin C, Wright J, I'Anson K, Humphrey T, Wells JM: **Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays**. *FEBS Lett* 2003, **554**:224-230.

28. Taboada EN, Acedillo RR, Carrillo CD, Findlay WA, Medeiros DT, Mykytczuk OL, Roberts MJ, Valencia CA, Farber JM, Nash JH: **Large-scale comparative genomics meta-analysis of *Campylobacter jejuni* isolates reveals low level of genome plasticity**. *J Clin Microbiol* 2004, **42**:4566-4576.

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

30. Fitzgerald JR, Reid SD, Ruotsalainen E, Tripp TJ, Liu M, Cole R, Kuusela P, Schlievert PM, Jarvinen A, Musser JM: **Genome diversification in *Staphylococcus aureus*: molecular evolution of a highly variable chromosomal region encoding the *Staphylococcal* exotoxin-like family of proteins**. *Infect Immun* 2003, **71**:2827-2838.

31. Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, Sturdevant DE, Ricklefs SM, Porcella SF, Parkins LD *et al.*: **Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks**. *Proc Natl Acad Sci USA* 2002, **99**:4668-4673.

32. • Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM: **Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains**. *Proc Natl Acad Sci USA* 2004, **101**:4865-4870.
The application of comparative genomics to investigate the mechanisms and consequences of genomic deletions in *Mycobacterium tuberculosis*.

33. Rajakumar K, Shafi J, Smith RJ, Stabler RA, Andrew PW, Modha D, Bryant G, Monk P, Hinds J, Butcher PD *et al.*: **Use of genome level-informed PCR as a new investigational approach for analysis of outbreak-associated *Mycobacterium tuberculosis* isolates**. *J Clin Microbiol* 2004, **42**:1890-1896.

34. Yao Y, Sturdevant DE, Villaruz A, Xu L, Gao Q, Otto M: **Factors characterizing *Staphylococcus epidermidis* invasiveness determined by comparative genomics**. *Infect Immun* 2005, **73**:1856-1860.

35. Wolfgang MC, Kulasekara BR, Liang X, Boyd D, Wu K, Yang Q, Miyada CG, Lory S: **Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa***. *Proc Natl Acad Sci USA* 2003, **100**:8484-8489.

36. Perrin A, Bonacorsi S, Carbonnelle E, Talibi D, Dessen P, Nassif X, Tinsley C: **Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species**. *Infect Immun* 2002, **70**:7063-7072.

37. • Dziejman M, Serruto D, Tam VC, Sturtevant D, Diraphat P, Faruque SM, Rahman MH, Heidelberg JF, Decker J, Li L *et al.*: **Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system**. *Proc Natl Acad Sci USA* 2005, **102**:3465-3470.
The first evidence of virulence-associated genes in non-01 and non-0139 strains of *V. cholerae*.

38. Frota CC, Hunt DM, Buxton RS, Rickman L, Hinds J, Kremer K, van Soolingen D, Colston MJ: **Genome structure in the vole bacillus, *Mycobacterium microti*, a member of the *Mycobacterium tuberculosis* complex with a low virulence for humans**. *Microbiology* 2004, **150**:1519-1527.

39. Pym AS, Brodin P, Brosch R, Huerre M, Cole ST: **Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti***. *Mol Microbiol* 2002, **46**:709-717.

40. • Beres SB, Sylva GL, Sturdevant DE, Granville CN, Liu M, Ricklefs SM, Whitney AR, Parkins LD, Hoe NP, Adams GJ *et al.*: **Genome-wide molecular dissection of serotype M3 group A *Streptococcus* strains causing two epidemics of invasive infections**. *Proc Natl Acad Sci USA* 2004, **101**:11833-11838.
Analysis of serotype M3 invasive isolates collected over an 11 year period identifies rapid population expansion resulting in infections that differed significantly in character and severity.

41. •• Zwick ME, McAfee F, Cutler DJ, Read TD, Ravel J, Bowman GR, Galloway DR, Mateczun A: **Microarray-based resequencing of multiple *Bacillus anthracis* isolates**. *Genome Biol* 2004, **6**:R10.
First microarray-based resequencing study of a bacterial species. *B. anthracis* strains were shown to be derived by a recent rapid clonal expansion from a single ancestral strain, with no inter-strain recombination or exchange of plasmids.

42. Wong CW, Albert TJ, Vega VB, Norton JE, Cutler DJ, Richmond TA, Stanton LW, Liu ET, Miller LD: **Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays**. *Genome Res* 2004, **14**:398-405.