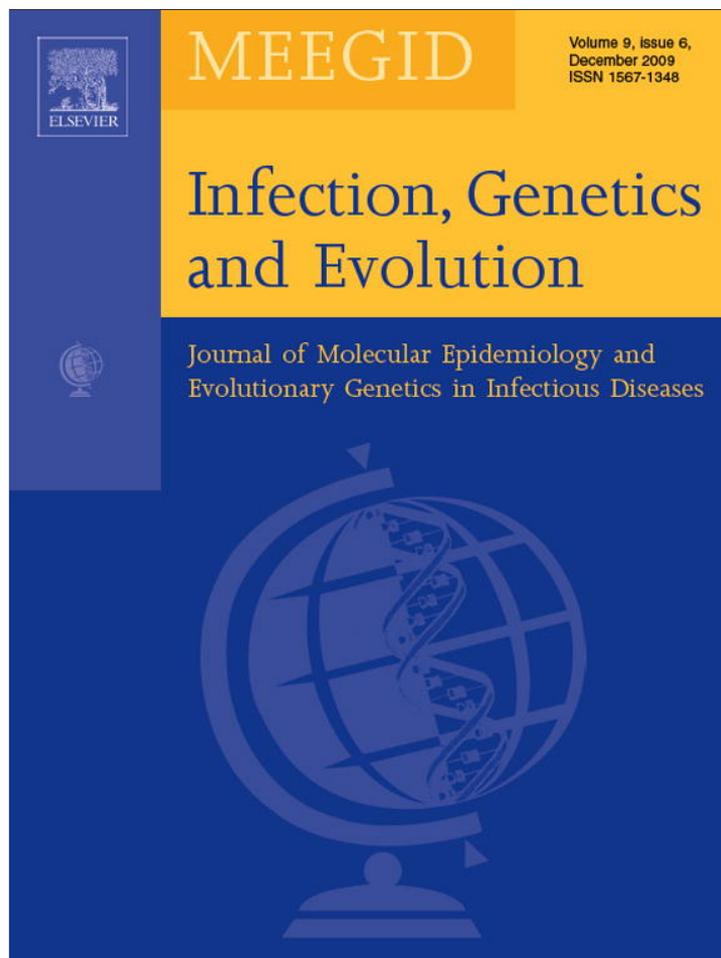


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

## Infection, Genetics and Evolution

journal homepage: [www.elsevier.com/locate/meegid](http://www.elsevier.com/locate/meegid)

## Discussion

## Linking questions to practices in the study of microbial pathogens: Sampling bias and typing methods

Elena Gómez-Díaz \*

*Génétique et Évolution des Maladies Infectieuses, UMR CNRS-IRD 2724, IRD, 911 Avenue Agropolis, BP 64501, F-34394 Montpellier, France*

## ARTICLE INFO

## Article history:

Received 20 June 2009

Received in revised form 21 August 2009

Accepted 21 August 2009

Available online 29 August 2009

## Keywords:

Sampling strategy

Population structure

Microbial evolution

Molecular epidemiology

Evolutionary ecology

Co-evolution

Multilocus sequence typing

## ABSTRACT

The importance of understanding the population genetics and evolution of microbial pathogens is increasing as a result of the spread and re-emergence of many infectious diseases and their impact for public health. In the last few years, the development of high throughput multi-gene sequence methodologies has opened new opportunities for studying pathogen populations, providing reliable and robust means for both epidemiological and evolutionary investigations. For instance, for many pathogens, multilocus sequence typing has become the “gold standard” in molecular epidemiology, allowing strain identification and discovery. However, there is a huge gap between typing a clinical collection of isolates and making inferences about their evolutionary history and population genetics. Critical issues for studying microbial pathogens such as an adequate sampling design and the appropriate selection of the genetic technique are also required, and will rely on the scale of study and the characteristics of the biological system (e.g., multi- vs. single-host pathogens and vector vs. food or air-borne pathogens). My aim here is to discuss some of these issues in more detail and illustrate how these aspects are often overlooked and easily neglected in the field. Finally, given the rapid accumulation of complete genome sequences and the increasing effort on microbiology research, it is clear that now more than ever integrative approaches bringing together epidemiology and evolutionary biology are needed for understanding the diversity of microbial pathogens.

© 2009 Elsevier B.V. All rights reserved.

## 1. Introduction

The huge diversity of microbial pathogens represents an equally huge diversity of life styles, reproductive modes and population structures. These biological characteristics are indeed tightly linked phenomena with crucial consequences on pathogen's evolution. Understanding the diversity and population genetics of microbial pathogens is central to disease management; from the evolution of drug resistance and virulence, to vaccine design and the emergence and re-emergence of important diseases. At this regards one important challenge to infectious disease research has been to type the existing diversity of pathogens, which up to now continues to be the focus of much work (Gevers et al., 2005; Achtman and Wagner, 2008). In recent years however, the field has moved forward in investigating the implications of pathogen genetic variation to disease epidemics (Galvani, 2003). Nonetheless, the link between epidemic and population genetic processes remains poorly understood. And several authors have claimed the need to perform such a multidisciplinary endeavor (Grenfell et al., 2004; Tibayrenc, 2005; Archie et al., 2009).

Various genetic mechanisms including point mutations, homologous recombination (i.e. sexual reproduction) and selection contribute to a great or less extent to the evolution of pathogenic microbes (Morschhäuser et al., 2000). The relative frequency of those sexual and non-sexual processes results in a wide spectrum of population structures and evolutionary trajectories which, even for a single microbe, can be dynamic and vary in space and time and at different hierarchical levels (i.e. from genes to populations) (Feil and Spratt, 2001; Halkett et al., 2005; Heitman, 2006). In this context, evaluating the extent and pattern of genetic variation in natural pathogen populations is often challenging. Indeed, owing to the spatial and temporal complexity of pathogen populations and the intrinsic characteristics of the biological system, biases in estimating genetic parameters can easily arise. Furthermore, common techniques applied to the study of pathogen populations show important limitations, and their suitability and performance depends to a large extent on the biological model and the rationale of the study.

Here I discuss general problems associated with the study of microbial pathogens including both common biases associated with the sampling design and the typing strategy applied. To address these issues I advocate the need of linking questions to practices in the study of pathogenic microbes, which in my point of view requires of more integrative and multidisciplinary

\* Tel.: +33 4 67 41 62 54; fax: +33 4 67 41 62 99.

E-mail address: [elena.gomez-diaz@mpl.ird.fr](mailto:elena.gomez-diaz@mpl.ird.fr).

approaches all better examined under a co-evolutionary framework.

### 1.1. The sampling problem

Sampling design is probably the most important challenge in the study of microbial pathogens, potentially compromising any inference on their evolution and population structure. A number of factors influence the sampling strategy. For instance, the fraction of the population sampled, sample size, should be based on levels of genetic diversity for a given pathogen. In principle, the sampling effort should be greater for pathogens displaying high levels of genetic variability than for those genetically more uniform, so the genetic diversity sampled is representative of the whole population. In the case of microbes with little sequence diversity, once a certain (low) threshold of isolates is reached, increasing sample size does not necessarily improve the genetic estimates. In those cases, the variability of the loci can be a more important factor, and thus searching for alternative genetic markers with more discriminatory power more rewarding.

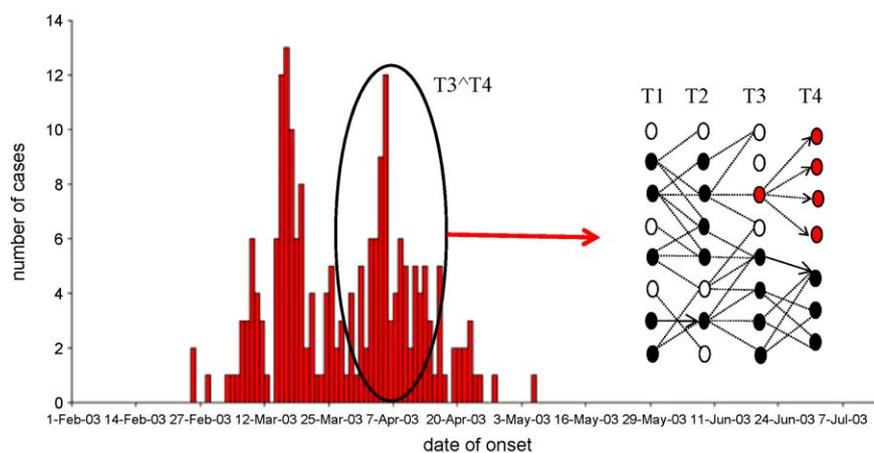
Pathogen populations can be both spatially and temporally structured and genetic variation can be partitioned at many different scales (Grenfell et al., 2004). In such situations, hidden population sub-structure in the sample such as temporal shifts from clonal to non-clonal reproduction, ecological niche specialization and geographic population subdivisions (Suerbaum and Achtman, 2004; Gagneux et al., 2006; Henriques-Normark et al., 2008); can have important consequences on genetic parameter estimates (“Wahlund effect” diploids (De Meeûs et al., 2006), or “admixture linkage disequilibrium” haploids (Falush et al., 2003)). Treating pathogens as being homogeneous will preclude any accurate inference on the population structure because it will be strongly dependent on the “sampling window” (Fraser et al., 2005) (Fig. 1). An example of this can be found in the yeast *Candida albicans* which has been extensively studied but yielded controversial results; from little or no linkage disequilibrium to extreme clonality (Tibayrenc, 1996). Recently, Nébavi et al. (2006) showed that in *Candida*, the reproduction mode was clonal and genetic variation occurred at the intrapopulation level (i.e. all pathogen isolates in an individual hosts). Thus, any design combining samples from different host individuals or different points in time would yield spurious results relative to the true extent of clonal reproduction. Similar examples are found in various protozoa, such as *Plasmodium falciparum*, which shows common shifts from sexual to non-sexual reproduction. In this case, an inadequate sampling, usually at a wider scale than the relevant one (i.e. intrapopulations (Annan et al., 2007)), have led to strong confounding results among studies (Tibayrenc and Ayala, 2002). Therefore, the ideal sampling scheme should apportion the genetic variance at all relevant levels (i.e. within individuals, between individuals and between populations and/or at different points in time) (Levin et al., 1999; Halkett et al., 2005).

Characterizing isolates from disease outbreaks, antibiotic-resistant or highly pathogenic (i.e. *Staphylococcus aureus*), can be another important source of bias. Indeed, such samples often represent a small proportion of the population, thus leading to false conclusions about the extent of apparent clonality (Smith et al., 1993; Fraser et al., 2005). For example, *Neisseria meningitidis* is a weakly clonal pathogen, but clinical collections show apparent clonal population structure. Recent studies have demonstrated an extensive genetic diversity of isolates from carriers, in contrast to a minority of clonal complexes associated with invasive disease. The so-called hyperinvasive lineages can be over-represented in disease isolate collections by as much as two orders of magnitude, relative to their prevalence in asymptomatic carriage (Caugant and Maiden, 2009). A few studies have already begun to address some

of these biases (Jolley et al., 2000; Feil et al., 2003; Ruimy et al., 2008), yet covering the gaps only to some extent (e.g., single vs. multiple populations or carriers vs. disease cases), and still representing a minority in microbiology research. Nonetheless, far beyond the bias posed, the genome comparison of these different sampling frames, disease and carriage isolates, is shedding important light into the evolution of virulence in *Neisseria* and similar pathogens with crucial consequences for vaccine development and disease control measures (Maiden, 2008).

All together, understanding the evolution and global population structure in a given pathogen relies on adequate sampling designs which in turns requires knowledge on how the pattern and the extent of genetic variation changes with the temporal and spatial scales. This should be evaluated case by case as there are not general applicable guidelines. A reasonable start may be analyzing a random sample of isolates from several individuals on geographically distinct populations and in different years, so most of the spatial and temporal variability is taken into account. Then, several genetic methods are available to assess the level of population subdivision in a sample from where we can delineate our smallest and most appropriate reference sampling unit (i.e.  $F_{IS}$  and Wahlund effect, repeated genotypes and linkage disequilibrium analyses, see Halkett et al., 2005; De Meeûs et al., 2006). In those pathogens for which there is some previous knowledge on their genetic structuring, such an exploratory step may be not needed, but the *a priori* knowledge upon which we base our sampling strategy should be carefully evaluated according to the above criteria.

The biology of the pathogen is also crucial to sampling design, although often dismissed. The life cycle characteristics and transmission dynamics of a given pathogen will determine its ecological niche, thus defining the most appropriate sampling frame (Barrett et al., 2008). Indeed, many pathogens can infect more than one host species (i.e. multi-host pathogens), in each of which the pathogen population can show different population sizes, can be differently structured and undergo different selection pressures (Woolhouse et al., 2001). In this case, the extent and pattern of pathogen population structure can change among sub-samples coming from different host types. Hence, simply analyzing human isolates or mixing isolates from different host types can result in a false picture of pathogen genetics and epidemics. Furthermore, in vector-borne pathogens with complex natural cycles (i.e. *Borrelia*, *Yersinia*, *Bartonella*, *Rickettsia*, or *Anaplasma* spp), the life include both different hosts and vectors; the population biology of the pathogen can therefore only be fully understood with a consideration of each host/vector component of the system. This is particularly important since many emerging human pathogens are vector-borne (Woolhouse, 2002). Examples are the agent of plague, *Yersinia pestis*, or the Lyme disease agent *Borrelia burgdorferi*. Transmission is arthropod-borne via fleas or ticks respectively, which can be both vector and bacterial strain specific for different vertebrate host types and geographical areas (Staszewski et al., 2008; Gómez-Díaz et al., submitted); yet conflicting genetic patterns have been obtained and their evolutionary history still remain unclear (Achtman et al., 2004; Richter et al., 2006; Margos et al., 2008). In both cases, the role of local transmission dynamics, and host and vector population structure is required to understand the epidemiology of the disease (Perry and Fetherston, 1997; Kurtenbach et al., 2006). This is linked with the idea of co-evolution and the local adaptation of pathogens to their hosts, where local pathosystems may function independently in space and time and show heterogeneous patterns of co-adaptation across the landscape (i.e. pathogen infectivity and host resistance) (Gandon and Michalakis, 2002; Thompson and Cunningham, 2002). Indeed, in vector-borne pathogens with



SARS epidemic outbreak in 2003 in Singapore. Image source: Ministry of Health, Singapore.

**Fig. 1.** Sampling bias associated with the scale of the study as there may be temporal fluctuations on the pathogen population structure for a sample taken at different points in time. In a first time period (T1<sup>T2</sup>), the population is non-clonal with horizontal genetic exchange. There is a point in time (T3), when a lineage (red cell) with an increased capacity to cause disease emergence (epidemic clone) evolves. At T4 individuals are isolates recovered from disease where this lineage is highly over-represented. Analysis of isolates from T4 may therefore show linkage disequilibrium due to the sampling bias introduced by the recent emergence of the epidemic clone. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

complex transmission cycles, host–vector–pathogen interactions can take place across very different scales. In this context, investigating the genetic structure of pathogen populations at the different levels of interaction (vector and hosts) is of utmost importance to understanding virulence/resistance evolution and its implications for disease dynamics. Nonetheless, in epidemiological research co-evolutionary thinking is only just beginning to emerge (Woolhouse et al., 2002).

In summary, the sampling design should be based on both; previous knowledge on the distribution of pathogen genetic diversity as well as on the spatial and the temporal aspects of the population structure. But most pathogens are zoonotic and display complex life cycles involving several hosts and/or vectors. Under such complex co-evolutionary scenarios, the ecology and life history characteristics of these pathosystems need to be carefully considered.

### 1.2. Microbial typing: promises and pitfalls of current practices

Traditional methods in studying microbial diversity include DNA–DNA hybridization and 16S ribosomal RNA sequencing. These techniques pose however important limitations regarding data standardization and time and laboratory costs. In the last years, the development of high throughput multi-gene sequence methodologies coupled with advances in statistical modeling have opened new venues in the study of pathogen populations, providing reliable and robust means for both epidemiological and evolutionary investigations. DNA microarrays, multilocus enzyme electrophoresis (MLEE), multilocus sequence typing (MLST), pulse-field electrophoresis (PFGE), and single nucleotide polymorphisms (SNPs); are commonly applied typing techniques in the study of microbial pathogens (Glossary, see Van Belkum et al., 2001; Medini et al., 2008, for a more complete review). High degree of reproducibility, portability, low cost-time consuming and high discriminatory power characterize the optimal typing technique. But these parameters are not fixed and the suitability of each method should however be evaluated case by case as it will critically depend on the biology of the pathogen and the rationale of the study. My aim here is not to provide a detailed catalogue of all existing typing methods, rather to discuss strengths and weaknesses of the widely applied MLST together with recent advances and future directions in microbial typing.

Among those various methods available, *Multilocus sequence typing* (MLST) has become the “gold standard” for the molecular epidemiology of many pathogens. MLST is an extremely standardized typing method that relies on (usually) 7 housekeeping genes of ~450 bp each. Based on the sequence information at each locus, allele sequences are either identical or non-identical so all unique sequences for a given locus are assigned an allele number in order of discovery. Isolates that share the same combinations of alleles are referred to as sequence types (ST) that each consists of a unique allelic profile. The advantages are clear as it provides a portable and reproducible typing method, using data that are amenable to population genetic analysis and which can be made readily accessible in public web-based databases (Maiden et al., 1998; Spratt, 1999; Maiden, 2006). The main area of application is the strain typing which consist in identifying the relationships of unknown isolates to described strains (i.e. genetic variants). But accurate identification requires a comprehensive molecular database against which unknowns can be compared. Thus, the second goal of MLST is to facilitate the strain discovery process of those novel genetic variants. In addition, because multilocus sequence typing has been applied to a large number of microbes, it can be a powerful tool in comparative-genomic studies or in exploratory data analyses on the degree of pathogen genetic diversity. Far beyond these basic applications, multilocus sequence typing is of great applied interest and can be, and is, commonly used to identify variants of genes encoding virulence traits, antibiotic resistance and antigenic variability (Maiden, 2006). Despite the invaluable advantages of MLST approaches in epidemiological typing, there are some major shortcomings about using MLST data for making inferences about the evolutionary history and population genetics of microbial pathogens. First, strict applications of allele information, as originally described for MLST data (Maiden et al., 1998), are generally not appropriate for inferring accurate evolutionary relationships in highly clonal microbes (Feil et al., 2004; Achtman and Wagner, 2008). Genetic relationships are established based on the allele definition and considers alleles to be equally distinct regardless whether they differ at one or several nucleotides. The degree of sequence diversity and the underlying genetic information is ignored. This approach may however be valid under certain circumstances, in particular in pathogens displaying intermediate levels of recombination. Indeed, allele based methods would buffer against the

confounding effects of recombination as they will only count the “number of events” (recombination or mutation) no matter the number of base changes introduced; a property that may indeed be particularly interesting in some kind of analysis (i.e. divergence times). Nonetheless, over the 10 years since its definition, multilocus sequence based analyses (also named *MLSA*, see *Glossary*) have been far more prevalent for population and evolutionary studies. But even in that case the suitability of the *MLSA* should be carefully evaluated as sequence based methods, contrary to the strict *MLST* approach, can be more easily confounded by convergent evolution (homoplasy), homologous recombination and lateral gene transfer (Achtman and Wagner, 2008; Pearson et al., 2009).

Another important concern regarding *MLST* methodologies relies on the general application of a particular set of loci, ubiquitous core genes, to several distinct and often unrelated microbes. However, increasing evidence suggest that the pattern of genetic variability and ultimately the value of a particular set of *MLST* genes can be pathogen-specific (Pérez-Losada et al., 2006). Housekeeping genes are by definition core parts of the genome required for metabolic function that are under strong selective constraints, and thus often fit neutrality expectations. But *MLST* genes under certain circumstances or in some pathogens can be under selection and/or undergo recombination (Chattopadhyay et al., 2009). Conversely, because they are highly conserved and selective constraint, their utility may be limited in pathogens which typically show very little genetic variability (i.e. genetically monomorphic pathogens: *Bacillus anthracis*, *Yersinia pestis*, or *Burkholderia mallei*) (Achtman, 2008). In these cases, one complementary and perhaps more powerful strategy, involves analyzing genes linked to the adaptive potential of the pathogen (i.e. *virulence genes*). Whereas neutral core housekeeping genes would serve to define basic clonal assignments, sequences from adaptive dispensable loci can be used to ‘zoom in’ on specific clones. This complementary approach may be especially useful in fine scale epidemiological typing when *MLST* typing provides no discrimination (Miller et al., 2005; Dingle et al., 2008). Likewise, genetic information from adaptive loci can be particularly interesting for inferring pathogen evolution in those cases where the evolution of pathogen populations is driven by co-evolutionary interactions with the vector(s), the host(s) or both (non-neutral selective forces) (Gupta and Maiden, 2001). However, virulence genes can show contrasting results when compared with those of core *MLST* genes. In such cases, interpreting the population genetics and evolutionary history may be difficult and need some caution as we know little about the evolutionary rules or rates of change in these genes.

### 1.3. Future prospects and new challenges

Far beyond the *MLST* revolution, the advent of whole genome sequencing has increasingly become a hot topic in microbial typing and may dominate the field in the upcoming years. More than 13 years have passed since the first complete microbial genome was sequenced (Fleischmann et al., 1995). Nowadays, coupled with the development of new high throughput sequencing technologies, more than 700 microbial genomes are available (Fournier et al., 2007; Medini et al., 2008). Genome analysis of microbial pathogens is providing unique insights into their virulence, host adaptation and evolution with both fundamental and applied interest. We can now obtain answers to each detailed aspect about the complete sets of genes of a given microbe. However, behind the promise, whole genome sequencing poses several challenges to microbiology research. For instance, these technologies still face important limitations in terms of sequence quality and cost-effectiveness. New advances and high throughput platforms are being developed

to meet this need and suggest whole genome sequencing will soon supersede current typing methods (Shendure and Ji, 2008). Yet, next-generation sequencing had to overcome the inertia of the field, so it is not clear how many years will pass before it becomes the large-scale, routinely applicable affair that the *MLST* has become. A common concern is that whole genome sequencing might be economically justified only for human pathogens. Indeed, it is difficult to think about the feasibility of large-scale genome projects for population level studies which usually require the sequencing of several pathogen isolates from each single population, particularly for the vast amount of neglected diseases of no economical or human-health interest (Molyneux, 2004). Moreover, related to current available data on thousands of microbial pathogens, it is not clear how large-scale genome data will correspond to the diversity characterized by other typing approaches such as the *MLST*, that is the cross-application of methods, as well as the integration of whole genome and partial genome genotyping data (i.e. SNPs). But perhaps the most critical issue for the application of whole genome approach involves our ability to manage and interpret such large amounts of data, which at present, is not obvious. In other words, microbial genomes may provide such tremendous amount of information so we are not able to capture the information we need (i.e. phylogeographic patterns or basic population genetic parameters). Nevertheless, all these issues are relatively new and remain poorly understood by many researchers. Importantly, new generation sequencing poses important challenges to comparative genomics and bioinformatics research (Pop and Salzberg, 2008). Yet these two fields still remain unlinked, at least in terms of user-friendly approaches to data management and analysis, and whether current computational resources can keep up with large-scale genome sequencing technologies still remains unanswered.

#### 1.3.1. Concluding remarks

The rationale of the sampling design represents an important challenge to the study of microbial pathogens. The study scale and the characteristics of the biological model are critical issues to consider in the sampling strategy. In addition, the choice of the typing method should rely on the level of resolution required and the type of answer expected. Overall, improving our understanding of microbial diversity and evolution requires more integrative approaches incorporating ecology, epidemiology and host–parasite co-evolution. Despite the popularity of multilocus sequence approaches in molecular epidemiology, a re-examination of their limitations and the alternatives available is needed for their general use in population genetics and evolution. In this context, the advent of easily and cheaply available whole genome sequencing technologies offer new and promising opportunities but also poses important challenges to this rapid evolving field. Important aspects such as the cross-application of different typing methods as well as the integration and interpretation of large-scale genomic data still remain unanswered. In this context, now more than ever, integrative approaches incorporating ecology, epidemiology and host–parasite co-evolution are strongly required. Despite the difficulty of performing such a multidisciplinary endeavor, these efforts will undoubtedly provide a much better basis for understanding the evolutionary biology and epidemiology of microbial pathogens.

#### Acknowledgments

Thanks to Bryan G. Spratt, Karen D. McCoy and Thierry de Meeûs and to three anonymous referees for their valuable comments on the manuscript. E.G.-D. was supported by a Marie Curie fellowship (no. PIEF-GA-2008-221243) and a contract by the Agence National de la Recherche (ANR-06-JCJC-0095-01).

## References

- Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A.J., Wagner, D.M., Allender, C.J., Easterday, W.R., Chenal-Francisque, V., Worsham, P., Thomson, N.R., Parkhill, J., Lindler, L.E., Carniel, E., Keim, P., 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. *PNAS* 101, 17837–17842.
- Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* 62, 53–70.
- Achtman, M., Wagner, M., 2008. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* 6, 431.
- Annan, Z., Durand, P., Ayala, F.J., Arnathau, C.I., Awono-Ambene, P., Simard, F.d.r., Razakandrainibe, F.G., Koella, J.C., Fontenille, D., Renaud, F.o., 2007. Population genetic structure of *Plasmodium falciparum* in the two main African vectors. *Anopheles gambiae* and *Anopheles funestus*. *PNAS* 104, 7987–7992.
- Archie, E.A., Luikart, G., Ezenwa, V.O., 2009. Infecting epidemiology with genetics: a new frontier in disease ecology. *TREE* 24, 21.
- Barrett, L.G., Thrall, P.H., Burdon, J.J., Linde, C.C., 2008. Life history determines genetic structure and evolutionary potential of host–parasite interactions. *TREE* 23, 678.
- Caugant, D.A., Maiden, M.C.J., 2009. Meningococcal carriage and disease—population biology and evolution. *Vaccine* 27, B64–B70.
- Chattopadhyay, S., Weissman, S.J., Minin, V.N., Russo, T.A., Dykhuizen, D.E., Sokurenko, E.V., 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *PNAS* 106, 12412–12417.
- De Meeûs, T., Lehmann, L., Balloux, F., 2006. Molecular epidemiology of clonal diploids: a quick overview and a short DIY (do it yourself) notice. *Infect. Genet. Evol.* 6, 163.
- Dingle, K.E., McCarthy, N.D., Cody, A.J., Peto, T.E.A., Maiden, M.C.J., 2008. Extended sequence typing of *Campylobacter* spp., United Kingdom. *Emerg. Infect. Dis.* 14, 1620–1622.
- Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.
- Feil, E.J., Spratt, B.G., 2001. Recombination and the population structures of bacterial pathogens. *Annu. Rev. Microbiol.* 55, 561–590.
- Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., Moore, C.E., Day, N.P.J., 2003. How clonal is *Staphylococcus aureus*? *J. Bacteriol.* 185, 3307–3316.
- Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* 186, 1518–1530.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.
- Fournier, P.-E., Drancourt, M., Raoult, D., 2007. Bacterial genome sequencing and its use in infectious diseases. *Lancet Infect. Dis.* 7, 711.
- Fraser, C., Hanage, W.P., Spratt, B.G., 2005. Neutral microepidemic evolution of bacterial pathogens. *PNAS* 102, 1968–1973.
- Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS* 103, 2869–2873.
- Galvani, A.P., 2003. Epidemiology meets evolutionary ecology. *TREE* 18, 32–139.
- Gandon, S., Michalakis, Y., 2002. Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. *J. Evol. Biol.* 15, 451–462.
- Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., de Peer, Y.V., Vandamme, P., Thompson, F.L., Swings, J., 2005. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733.
- Gómez-Díaz, E., Doherty, P.F., Duneau, D., McCoy, K.D., submitted for publication. Cryptic population structure can mask vector-specific patterns of infection: an example from the marine cycle of Lyme borreliosis.
- Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., Holmes, E.C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332.
- Gupta, S., Maiden, M.C.J., 2001. Exploring the evolution of diversity in pathogen populations. *Trends Microbiol.* 9, 181.
- Halkett, F., Simon, J.-C., Balloux, F., 2005. Tackling the population genetics of clonal and partially clonal organisms. *TREE* 20, 194.
- Heitman, J., 2006. Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* 16, R711.
- Henriques-Normark, B., Blomberg, C., Dagerhamn, J., Battig, P., Normark, S., 2008. The rise and fall of bacterial clones: *Streptococcus pneumoniae*. *Nat. Rev. Microbiol.* 6, 827.
- Jolley, K.A., Kalmusova, J., Feil, E.J., Gupta, S., Musilek, M., Kriz, P., Maiden, M.C.J., 2000. Carried Meningococci in the Czech Republic: a diverse recombining population. *J. Clin. Microbiol.* 38, 4492–4498.
- Kurtenbach, K., Hanincova, K., Tsao, J.L., Margos, G., Fish, D., Ogden, N.H., 2006. Fundamental processes in the evolutionary ecology of Lyme borreliosis. *Nat. Rev. Microbiol.* 4, 660–669.
- Levin, B.R., Lipsitch, M., Bonhoeffer, S., 1999. Population biology, evolution, and infectious disease: convergence and synthesis. *Science* 283, 806–809.
- Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS* 95, 3140–3145.
- Maiden, M.C.J., 2006. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 60, 561–588.
- Maiden, M.C.J., 2008. Population genomics: diversity and virulence in the *Neisseria*. *Curr. Opin. Microbiol.* 11, 467–471.
- Margos, G., Gatewood, A.G., Aanensen, D.M., Hanincova, K., Terekhova, D., Vollmer, S.A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M.A., Feil, E.J., Fish, D., Casjens, S., Wormser, G.P., Schwartz, I., Kurtenbach, K., 2008. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *PNAS* 105, 8730–8735.
- Medini, D., Serruto, D., Parkhill, J., Relman, D.A., Donati, C., Moxon, R., Falkow, S., Rappuoli, R., 2008. Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* 6, 419.
- Miller, W.G., On, S.L.W., Wang, G., Fontanoz, S., Lastovica, A.J., Mandrell, R.E., 2005. Extended multilocus sequence typing system for *Campylobacter coli*, *C. lari*, *C. upsaliensis*, and *C. helveticus*. *J. Clin. Microbiol.* 43, 2315–2329.
- Molyneux, P.D.H., 2004. “Neglected” diseases but unrecognized successes—challenges and opportunities for infectious disease control. *Lancet* 364, 380.
- Morschhäuser, J., Köhler, G., Ziebuhr, W., Blum-Oehler, G., Dobrindt, U., Hacker, J., 2000. Evolution of microbial pathogens. *Phil. Trans. Lond. B: Biol. Sci.* 355, 695–704.
- Nébavi, F., Ayala, F.J., Renaud, F., Bertout, S., Eholié, S., Moussa, K., Mallié, M., de Meëüs, T., 2006. Clonal population structure and genetic diversity of *Candida albicans* in AIDS patients from Abidjan (Côte d’Ivoire). *PNAS* 103, 3663–3668.
- Pearson, T., Okinaka, R.T., Foster, J.T., Keim, P., 2009. Phylogenetic understanding of clonal populations in an era of whole genome sequencing. *Infect. Genet. Evol.* 9, 1010.
- Pérez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidi, R.P., Crandall, K.A., 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect. Genet. Evol.* 6, 97.
- Perry, R.D., Fetherston, J.D., 1997. *Yersinia pestis*—etiologic agent of plague. *Clin. Microbiol. Rev.* 10, 35–66.
- Pop, M., Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142.
- Richter, D., Postic, D., Sertour, N., Livey, I., Matuschka, F.R., Baranton, G., 2006. Delineation of *Borrelia burgdorferi* sensu lato species by multilocus sequence analysis and confirmation of the delineation of *Borrelia spielmanii* sp nov. *Int. J. Syst. Evol. Microbiol.* 56, 873–881.
- Ruimy, R., Maiga, A., Armand-Lefevre, L., Maiga, I., Diallo, A., Koumare, A.K., Ouattara, K., Soumare, S., Gaillard, K., Lucet, J.-C., Andremont, A., Feil, E.J., 2008. The carriage population of *Staphylococcus aureus* from Mali is composed of a combination of pandemic clones and the divergent panton-valentine leukocidin-positive genotype ST152. *J. Bacteriol.* 190, 3962–3968.
- Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135.
- Smith, J.M., Smith, N.H., O’Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? *PNAS* 90, 4384–4388.
- Spratt, B.G., 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. *Curr. Opin. Microbiol.* 2, 312.
- Staszewski, V., McCoy, K.D., Boulinier, T., 2008. Variable exposure and immunological response to Lyme disease *Borrelia* among North Atlantic seabird species. *Proc. R. Soc. Lond. B: Biol. Sci.* 275, 2101.
- Suerbaum, S., Achtman, M., 2004. *Helicobacter pylori*: recombination, population structure and human migrations. *Int. J. Med. Microbiol.* 294, 133.
- Thompson, J.N., Cunningham, B.M., 2002. Geographic structure and dynamics of coevolutionary selection. *Nature* 417, 735.
- Tibayrenc, M., 1996. Towards a unified evolutionary genetics of microorganisms. *Annu. Rev. Microbiol.* 50, 401–429.
- Tibayrenc, M., Ayala, F.J., 2002. The clonal theory of parasitic protozoa: 12 years on. *Trends Parasitol.* 18, 405.
- Tibayrenc, M., 2005. Bridging the gap between molecular epidemiologists and evolutionists. *Trends Microbiol.* 13, 575–580.
- Van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H., Tibayrenc, M., 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clin. Microbiol. Rev.* 14, 547–560.
- Woolhouse, M.E.J., Taylor, L.H., Haydon, D.T., 2001. Population biology of multithost pathogens. *Science* 292, 1109–1112.
- Woolhouse, M.E.J., 2002. Population biology of emerging and re-emerging pathogens. *Trends Microbiol.* 10, s3.
- Woolhouse, M.E.J., Webster, J.P., Domingo, E., Charlesworth, B., Levin, B.R., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* 32, 569–577.

## Glossary

**Wahlund effect and heterozygote deficit:** Whenever a sample consists of individuals that were sampled from genetically differentiated sub-populations, this leads to artificial departures from panmictic expectations as the loss of heterozygosity.

**Admixture linkage disequilibrium:** The non-random association of genetic variants (i.e. alleles at a given loci) due to admixture (gene flow) between genetically distinct sub-populations.

**DNA microarray:** DNA chips or microarrays are composed of large numbers of DNA strands (i.e. genes) fixed on a solid support that function as probes for the DNA of

the organism under study. In a single hybridization experiment, the absence/presence status of all genes of a given microbe against the reference set of genes (plotted on the chip) can be examined.

*Multilocus enzyme electrophoresis (MLEE)*: Enzyme polymorphisms between strains are detected on the basis of differing electrophoretic mobilities of the encoded proteins on a starch gel.

*Multilocus sequence analysis (MLSA)*: It is a generalization of the MLST approach but in this case rather than using allele information, MLSA compares the primary DNA

sequences from multiple conserved protein-coding loci which may not necessarily be housekeeping genes.

*Pulsed-field gel electrophoresis (PFGE)*: This typing method detects genetic variation between strains using rare-cutting restriction endonucleases followed by separation of the resulting large genomic fragments on an agarose gel.

*Single nucleotide polymorphisms*: SNP denotes point nucleotide changes that result from the comparison of the genomic sequences from a number of related strains.