

Discussion

Population genetics and population biology: what did they bring to the epidemiology of transmissible diseases? An e-debate

Una Morgan^a, Howard Ochman^b, François Renaud^c, Michel Tibayrenc^{c,*}

^a WHO Collaborating Centre for the Molecular Epidemiology of Parasitic Infections,
Murdoch University, Murdoch, WA, Australia

^b Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA

^c UMR CNRS/IRD 9926, Genetics of Infectious Diseases, IRD, BP 5045, 34032 Montpellier Cedex 1, France

Accepted 8 October 2001

1. First question

The definition of the CDC for molecular epidemiology is: “the various techniques derived from immunology, biochemistry, and genetics for typing or subtyping pathogens”. This definition is technology-based and does not make any reference to evolutionary genetic concepts. Do you think that this approach is misleading for medical and epidemiological applications, or do you think that molecular epidemiology can be performed based only on technology?

1.1. Response from Una Morgan

I do not believe that the CDC approach is “misleading for medical and epidemiological applications”. It is clear from the research being conducted at CDC that their definition of molecular epidemiology is an inclusive one. Within the framework of “molecular epidemiology” CDC laboratories use state-of-the-art technologies and concepts of evolutionary genetics in addressing the inter-related issues of When? Where? Why? and How? of pathogens. Investigations are never technique driven alone. Indeed, the majority of CDC studies are conducted in the context of clinical epidemiologic studies which therefore links genetics with clinical manifestations of illness. One can find examples of CDC’s work on many infectious agents including malaria, *Cryptosporidium*, *Cyclospora*, TB, and HIV in the current literature (Xiao et al., 2001; Biswas et al., 2000).

1.2. Comment from Michel Tibayrenc to Una Morgan’s response

I do know that the CDC researchers use wisely molecular epidemiology. Still the fact remains that many studies published in the literature rely only on the assumption that genetic identity based on a very limited set of markers implies recent clonal descent. Let us take a practical example: a set of *Staphylococcus aureus* stocks that prove to be identical with four RAPD primers (example encountered many times in the literature). Does it imply that the common ancestor of these stocks is only a few months old?

1.3. Response from Howard Ochman

Michel, my “before coffee” thoughts about the first question: Whoever wrote the original definition of molecular epidemiology did not want to pass judgement on the quality or utility of research performed under the guise of molecular epidemiology. How about this slight, albeit cumbersome, rewrite of the definition? Does anyone find this a bit more acceptable? Molecular epidemiology is: “the application of the various techniques derived from immunology, biochemistry, molecular biology and genetics for the purpose of identifying, typing, or determining the origins, ancestry or relationships of microorganisms, particularly pathogens”.

1.4. Comment from Michel Tibayrenc to Howard Ochman’s response

Your point is excellent. I would make it clear that: (i) the CDC definition cited is 6 years old; (ii) as Una Morgan pointed out, it is implicitly understood in this definition

* Corresponding author. Tel.: +33-467-41-6197; fax: +33-467-41-6299.
E-mail address: michel.tibayrenc@mpl.ird.fr (M. Tibayrenc).

that people will try to understand where, when and how pathogens spread over (what the CDC epidemiologists do). However, in its strict statement, the definition is technology-based, and I used it to illustrate the fact that many studies rely only on description/comparison of banding patterns. Please react on the comment I sent to Una.

1.5. Response from François Renaud to the first question

I agree with my two colleagues, and I believe that the CDC does not restrict molecular epidemiology to a technological concept. But, I found that the question of Michel “stirs up” the crucial debate on the species concept, which is particularly sharp in ecology and evolution of parasitism, specifically for the knowledge of epidemiological foci. So, if we try to analyze the parameters responsible for parasitic disease epidemiology, we have first to answer the following question: Who transmits what, where and when? In this context, the knowledge of population biology and species determination are essential to understand “Who transmits what”, and it is clear that if we had at disposal a non ambiguous definition of what is a species, we would use it since a long time. I perform the proposed definition for molecular epidemiology in this way, and it is unquestionable that various techniques derived from immunology, biochemistry and genetics are essential tools to identify different pathogen genomes. But, it only constitutes a first stage. Indeed, the real problem for parasitologists (*sensu lato*) is to understand why a parasite is a pathogen, more specifically why and when it becomes a pathogen in different molecular and physiological host environments. For example, why and when an *Escherichia coli* or a *Plasmodium* present different levels of pathogenicity? Why the same genotype is or is not pathogen in the environment where it develops? These questions must be solved at both individual and population scales (i.e. expression of pathogenicity, evolution of resistance and virulence ...). We are confronted to the interactions between genotypes and environments which define phenotypes. Because selection acts on phenotypes, and because phenotypes are the expression of genotypes in the different environments where phenotypic plasticity play a key role, the knowledge of disease epidemiology should be investigated from molecules to ecosystems ... But what could we do now, and what should we do in this crucial domain of evolutionary parasitology?

1.6. Comment from Michel Tibayrenc

In his interesting answer, François anticipated on a future e-debate I am planning on the species concept in microbiology. I feel sorry I did not formulate very well my first question and I am obviously trapped with my reference to the CDC definition of molecular epidemiology. Before I launch the second question, possibly you could comment a little bit on these many studies which rely only on band counting and empirical comparison of patterns. I do see a real danger of misleading conclusions with such an approach (see my

response to Una). Example: identical patterns with a limited set of markers is often taken as evidence for common clonal descent in molecular epidemiology. Now even in a sexual organisms and panmictic populations, the probability of observing identical genotypes can be non negligible, and increases when the number of loci surveyed and their variability decrease. It is only one example among many of the traps caused by empirical interpretation of genetic data.

1.7. Response from Howard Ochman

This brings up a question that we are too often asked, which is: What is the best technique for typing (or identifying or classifying) samples? Certainly, the answer depends on the particular question being investigated, but also on the technology and analytical tools (and also the time, funds and personnel) available to the investigator. In the types of studies that you mentioned, do you think that the problem is linked to insufficient data, or to insufficient expertise in interpreting the data. Both of these problems are curable, but I think that researchers (and, sorry to sound philistine here, funding sources) embrace new technology much more readily than new ways to analyze, interpret or augment old data.

1.8. Response from François Renaud

I think that there are two main ways in science. The first one concerns formal or “hard” science (i.e. physique for example) where processes are defined under global equations, and where predictions follow the different parameters of these equations. The second concerns the non-formal or “soft” sciences (i.e. medicine, biology, ecology and evolution *sensu lato*) where processes are so complex that they cannot be defined under equations, or requires thousands of them. In this case, it is unquestionable that we need to obtain the “optimal information”. But, what is the optimal information in disease epidemiology? At a first stage, we have to answer the question asked before: “Who transmits what”? To answer this first question we should have at disposal (i) the largest possible number of characters (morphology, anatomy, genetics ...) and (ii) the most appropriate sampling. Knowing that we cannot obtain all these informations from pathogens (specifically morphology and anatomy), the relevance of our results depends on a compromise between the two points. So, identical patterns obtained with a limited set of markers are not relevant! Indeed, we should analyse a large number of pathogens, sampled in different hosts and localities, on different variable and neutral molecular markers. Just one example, I find that most of studies realised on population biology of *Plasmodium falciparum* failed to respond to these criteria because (i) antigenic determinants generally used are under selection; (ii) their number and their variability are weak and (iii) samples are most often obtained from free clinics where people come from different undetermined locations. But I do accept that I am unable

to give you the minimum number of markers and samples required for a relevant analysis.

After discussion with my colleague Thierry de Meeüs, who works on population genetics, he suggests that:

- less than five loci will lead to inconclusive results, seven loci is a good minimum and 10 loci is really good but 30 loci would be the top;
- 5–10 alleles per locus, each one with a frequency over 0.05 are nice;
- the number of individuals per sample is much more critical than the number of samples: 60 individuals per sample is ideal, but 30 individuals per sample is good providing all samples have 30 individuals (balanced sampling);
- two samples is not appropriate for any investigations in population biology; thus three is the minimum number required; but in order to analyse accurately the genetic structure of natural populations, 10 samples seem to be a good minimum number.

I just want to add that these conditions are seldom gathered in all works we could read in the literature!

1.9. Comment from Michel Tibayrenc

Thank you, François. In reaction to your response, two comments:

1. In medical microbiology, ideal samples are sometimes very difficult to gather. One has to make do with what is available.
2. Sampling depends on the question under study. The goal is not always to get a comprehensive knowledge of a pathogen's genetic diversity and population structure on its whole geographical range. In the type of studies performed by the CDC (and IRD), the goal is rather to follow the route of an epidemic (example: which *Escherichia coli* clone contaminated a stock of meat in a fast-food chain). In this case, the sample will be stricky focused on its target. If one analyses mating system and population structure, and the linkage disequilibrium is extremely strong, you may have fair presumptions of it with few individuals and few loci. It has been the case with our early studies on *Trypanosoma cruzi*, the agent of Chagas disease, dealing with four isoenzyme loci and no more than 70 individuals. The conclusions have been fully corroborated by more sophisticated samples.

2. Second question

You all know the strong implications of a pathogen's mating system on its population structure, and hence, on the possibility to use its multilocus genotypes as stable epidemiological tracers. In a recent paper, Levin et al. (1999) have proposed that the mating system of a pathogen is highly flexible and depends on the ecological cycle considered. Example: *Plasmodium falciparum* would be panmictic

in high transmission places and clonal in low transmission cycles. Others (like me) think rather that this parameter is strongly driven by in-built biological properties and is only moderately modulated by the environment. According to your personal experience, what is your opinion on this important point?

2.1. Response and comment from François Renaud

First, I just want to notify that my previous answer was a little bit provocative, because it is clear that all investigations depend on the question under study as underlined by Michel. Anyway, I think that we have to keep in mind that sampling constitutes often the "Achilles-heel" of many studies focusing on population biology. For my opinion, disease epidemiology should be investigated in the context of population biology!

Response to the present question: this is another great question: "When and why parasites or pathogens do sex?". The problem can be easily solved for asexual parasites like *Trypanosoma cruzi* where sex seems to be very sporadic. For these organisms, multilocus genotypes can be considered as epidemiological tracers. There are two cases for sexual parasites (i) gonochorism and (ii) hermaphroditism. For gonochoric parasites (male and female functions separated in different individuals like *Schistosoma*), the mating system is clear, and demography plays a main role for partner meeting inside the host. Multilocus genotypes cannot be used as stable epidemiological tracers because genetic recombination often breaks them. The problem is sharper for hermaphrodite parasites (male and female functions in the same individual) which can invest differentially in one of the two functions. In this case, I agree with Levin et al. that mating system could highly depends from host environmental conditions. *Plasmodium* are hermaphrodite parasites! In a very interesting paper, Paul et al. (2000) showed that, at least for *P. gallinaceum*, the investment for an individual in the male and the female function depends on the evolution of the infection inside the host. So, the parasite is able to control the quality of its gamete production in front of different environmental parameters. Elsewhere, we showed (Trouvé et al., 1996, 1999) that differential sex allocation (selfing vs outcrossing) of an *Echinostoma* (Trematoda) depends on parameters (number and origin of partners) encountered inside the rodent host. These observations emphasize the consequences that such a phenomenon could have on parasite population structures, and the difficulty to use multilocus genotypes as stable epidemiological tracers when sex occurs.

2.2. Response from Howard Ochman

Perhaps I am sounding a bit ornery, but I believe the topic of the evolution and maintenance of sex to be among the largest source of "just-so" stories in biology. We can all think of cases where a successful pathogen is clonal and

non-recombining, and others where it is virtually panmictic. Everyone can pick their favorite organism to favor whichever view they wish to support that day. Naturally, we tend to study the successful (as opposed to the extinct) pathogens, so we are forced to come up with reasons why a particular organism has prospered as a result of (or in spite of) its mating system. Based on the phrasing of the question, and the types of responses given so far, I cannot tell whether Levin et al. meant that individual pathogens (or pathogenic species) have flexible mating systems, or whether they meant that pathogens, in general, show a diverse array of mating systems. In any case, I would bet that Levin et al. tend to agree with Michel's statement about moderate environmental modulation.

2.3. Response from Una Morgan

In my experience, I tend to agree with Michel that mating systems of pathogens are influenced much more by in-built biological properties than by their environment. Take *Cryptosporidium* as an example. It is a ubiquitous protozoan parasite that is transmitted by the faecal-oral route and by contamination of water supplies. Since the infamous 1993 Milwaukee outbreak in which 400,000 individuals contracted cryptosporidiosis as a result of drinking contaminated water supplies, there has been extensive research into this pathogen. *Cryptosporidium* has a highly complex life cycle that involves both sexual and asexual stages, yet despite the numerous loci that have been examined by sequence analysis, there is very little evidence for genetic recombination. Genetic evidence to date has established the existence of numerous distinct "genotypes" or discrete typing units that have remained stable over time and across widespread geographic areas (Morgan et al., 1999; Xiao et al., 2000). This appears to be independent of whether infections were symptomatic or asymptomatic or whether epidemiological studies were conducted in areas where levels of transmission were high or low. Therefore the evidence to date with *Cryptosporidium* strongly supports the concept of an essentially clonal population structure that appears to be largely independent of environmental conditions.

3. Third question

Gene phylogenies have a sharp resolution due to the considerable amount of information conveyed by sequences. However, the phylogeny of one gene or of a limited number of genes may not be representative of the overall phylogeny of the organism under study ("species phylogeny"). On the other hand, multilocus phylogenies based on markers such as MLEE or RAPDs have drawbacks such as lack of resolution and/or homoplasy. Multilocus sequence typing (MLST) (Maiden et al., 1998) is supposed to combine the advantages of gene phylogeny and multilocus typing. According to your own experience, do you consider MLST as a considerable

progress in molecular epidemiology and population genetics of pathogens?

3.1. Response from Una Morgan

Traditionally multilocus phylogenies have been based on MLEE markers and more recently RAPD markers. One of the most serious technical disadvantages of these markers is lack of reproducibility between laboratories, particularly RAPD markers. Another drawback of RAPD data is the amplification of non-specific products which can result in unreliable phylogenies. MLST is a development of multilocus enzyme electrophoresis in which the alleles at multiple house-keeping loci are assigned directly by nucleotide sequencing, rather than indirectly from the electrophoretic mobilities of their gene products. A major advantage of MLST is that sequence data are unambiguous and electronically portable, allowing molecular typing of infectious agents via the Internet. MLST has confirmed both the existence of clones and the high rates of recombination for several bacterial pathogens (Smith et al., 2000). A recent study of *Streptococcus pyogenes* examined the nucleotide sequences of internal fragments of seven selected housekeeping loci for 212 isolates. A total of 100 unique combinations of housekeeping alleles (allelic profiles) were identified. The MLST scheme was highly concordant with several other typing methods (Enright et al., 2001). One disadvantage of MLST methods is that they are laborious and time-consuming however, semiautomated methods for MLST using a 96-well format liquid handler and an automated DNA sequencer have been developed for pathogens such as *Neisseria meningitidis* which should greatly speed up the process (Clarke et al., 2001).

3.2. Response from Howard Ochman

The question of "considerable progress" has two parts: the first pertains to the technical aspects and utility of MLST; but more importantly, we should realize that researchers using MLST, as well as those reading papers where MLST has been applied, are putting and understanding these data in an evolutionary/phylogenetic/population genetic context. With regard to the first issue, I think that the standardization and portability of data offered by MLST is a great improvement over MLEE; and moreover, the utility of these data for purposes other than epidemiological tracing is very appealing. I see that I have already begun to segue into the second part of my response without mentioning any of the impressions of the different methodologies. There are many papers explaining the advantages of MLST over previous methods (e.g. Enright and Spratt, 1999), so there is little need to repeat those here. A quick search through my (perhaps not up-to-date) citation manager yields less than 20 original papers on MLST, and a few of these were applying new analytic approaches to previously published data. Moreover, the authors of these papers display a somewhat

high coefficient of co-publication relatedness, though we might expect there to be a lag before other labs apply MLST to their organism. On the positive side, many of the types of questions that interest me rely on data from MLST studies (though I must admit that a 450-bp fragment is less interesting than having the complete sequence of a gene). Having a large standardized dataset on diverse pathogens introduces cohesion to the field, promotes comparative analyses, and allows us to think beyond our own individual studies.

4. Last question

Give as briefly as possible your definition of the term “strain” for pathogens.

4.1. Response from François Renaud to third and last question

It is clear that technical advances during these last years have lead to substantial progresses for the knowledge of biology and evolution of organisms. In molecular epidemiology, MLST represents one relevant tool compared to MLEE or RAPD because (i) it provides direct information on different gene sequences distributed within the genome; (ii) results can be compared and used in different research laboratories worldwide. But in spite of this important technical aspect, we have to think about the relevance of phylogeny in epidemiology of pathogens. The classical goal of phylogeny is to reconstruct, or at least to give a proposal on, the relationships between taxa during time. “Phylogeneticists” try to understand the evolutionary connections between species which are not supposed to exchange genetic information since different time scales (often a long time). But the problem for epidemiology is quite different because we are most often confronted by pathogen genotypes belonging to one taxon, evolving in one ecosystem at one time and supposed to have the possibility to exchange genes. What is a phylogeny and how is it relevant in this context? I think that it is more appropriate to talk about genetic relationships between genotypes based on the molecular parameters analyzed, than phylogeny, and this is not a semantic problem. If for pathogens which almost present an asexual or a selfing mode of reproduction the different genotypic lineages can be considered as genetic entities (“strains”) where phylogenetic approaches reflect their evolution, the situation is totally different when gene flow occurs between individuals and/or lineages in one ecosystem. The life time expectancy of one genotype can be very short, and phylogeny could only represent an instantaneous situation which is not relevant for our understanding of disease epidemiology. This remark lead to the last question on the definition of the term “strain” for pathogens. I am sorry, but I have no clear definition because we are confronted to the same problem than the species concept. What is the strain concept in the context of population biology parameters (i.e. mutation, selection,

migration, genetic drift, gene flow, genetic recombination . . .)? The debate is still open!

4.2. Response from Howard Ochman

The shortest answer that I can come up with, based on the use of the term in a wide range of papers, is: “a microbial variant”. The term has been rather loosely applied to distinguish among isolates that differ sufficiently with respect to the trait in question, and, hence, no set amount of difference (genetic or otherwise) distinguishes one strain from another. Am I satisfied with this? Yes, but only to the degree that an investigator knows (and can state) the basis for distinguishing among strains. I can conceive of cases where a single point mutation might warrant classification as a different strain, and others where the gain or loss of several genes (such as those encoding some antibiotic resistance determinants) would still be viewed as constituents of a single strain.

4.3. Response from Una Morgan

This term has engendered much confusion as there have been numerous definitions of “strains” in the literature. Amongst parasitic protozoologists, the term has been defined as a homogenous population possessing a set of defined genetic and biological characters (Thompson and Lymbery, 1990). The emphasis is on a combination of both genetic and biological characteristics since reliance on genotype alone may confer significance on a feature of little biological relevance. Thus a “strain” can be defined as an artificial term of relevance mostly in epidemiological studies that refers to a group of organisms that are genetically different in gene frequencies and which share one or more characters of epidemiological significance.

4.4. Concluding remarks from Michel Tibayrenc

This was a tricky question. At the MEEGID II congress in Montpellier, 1997, we had organized a whole roundtable on the notion of strain. We collected roughly as many different responses as the number of participants. The less disputable proposal was “a collection of stocks that share the same multilocus genotype”. Una’s definition adds a relevant notion of biomedical/epidemiological specificity.

I wish to thank you for your very valuable contribution to this debate.

References

- Biswas, S., Escalante, A., Chaiyaroj, S., Angkasekwinai, P., Lal, A.A., 2000. Prevalence of point mutations in the dihydrofolate reductase and dihydropteroate synthetase genes of *Plasmodium falciparum* isolates from India and Thailand: a molecular epidemiologic study. *Trop. Med. Int. Health* 5, 737–743.
- Clarke, S.C., Diggle, M.A., Edwards, G.F., 2001. Semiautomation of multilocus sequence typing for the characterization of clinical isolates of *Neisseria meningitidis*. *J. Clin. Microbiol.* 39, 3066–3071.

- Enright, M.C., Spratt, B.G., 1999. Multilocus sequence typing. Trends Microbiol. 7, 482–487.
- Enright, M.C., Spratt, B.G., Kalia, A., Cross, J.H., Bessen, D.E., 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between emm type and clone. Infect. Immun. 69, 2416–2427.
- 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.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. Proc. Nat. Acad. Sci. U.S.A. 95, 3140–3145.
- Morgan, U.M., Monis, P.T., Fayer, R., Deplazes, P., Thompson, R.C., 1999. Phylogenetic relationships among isolates of *Cryptosporidium*: evidence for several new species. Intern. J. Parasitol. 29, 1733–1751.
- Paul, R., Coulson, T., Raibaud, A., Brey, P., 2000. Sex determination in malaria parasites. Science 287, 128–131.
- Smith, J.M., Feil, E.J., Smith, N.H., 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 22, 1115–1122.
- Thompson, R.C.A., Lymbery, A.J., 1990. Intraspecific variation in parasites—what is a strain? Parasitol. Today 6, 345–348.
- Trouvé, S., Renaud, F., Durand, P., Jourdane, J., 1996. Selfing and outcrossing in a parasitic hermaphrodite helminth (*Trematoda, Echinostomatidae*). Heredity 77, 1–8.
- Trouvé, S., Renaud, F., Jourdane, J., Durand, P., Morand, S., 1999. Differential sex allocation in a simultaneous hermaphrodite. Evolution 53, 1599–1604.
- Xiao, L., Morgan, U.M., Fayer, R., Thompson, R.C., Lal, A.A., 2000. *Cryptosporidium* systematics and implications for public health. Parasitol. Today 16, 287–292.
- Xiao, L., Bern, C., Limor, J., Sulaiman, I., Roberts, J., Checkley, W., Cabrera, L., Gilman, R.H., Lal, A.A., 2001. Identification of 5 types of *Cryptosporidium* parasites in children in Lima. Peru. J. Infect. Dis. 183, 492–497.



Una Morgan is Australian. She works at Murdoch University with Professor Andrew Thompson, who is the co-editor of our journal to Australia. Una was until recently at the Centers for Disease Control and Prevention in Atlanta for a short stay. She is a specialist of molecular epidemiology and works more specifically on *Cryptosporidium*.



Howard Ochman is American. Professor at the University of Tucson (Arizona), he is a specialist of bacterial evolution and the evolution of virulence. He has been the author of pioneering works on bacterial population structure with Bob Selander.



François Renaud is French. A specialist of parasite evolution, he has been trained at the Institut des Sciences de l'Evolution, University of Montpellier, and is presently in my unit of research where he is the head of an autonomous group. His main researches deal with ecology and evolution of host/parasite systems. He presently has a big programme on *Plasmodium falciparum* population genetics.