CURATED COLLECTION

Advances in Molecular Epidemiology of Infectious Diseases

Microbiology Spectrum
American Society for Microbiology Press

Advances in Molecular Epidemiology of Infectious Diseases: Definitions, Approaches, and Scope of the Field*

LEE W. RILEY[1] and RONALD E. BLANTON[2]

[1]Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, Berkeley, CA 94720; [2]Center for Global Health & Diseases, Case Western Reserve University, Cleveland, OH 44106

ABSTRACT Molecular epidemiology is a discipline that uses molecular microbiology tools to study the distribution and determinants of diseases in human populations and veterinary animals. Our understanding of epidemiology of infectious diseases has evolved with technological advancements made in molecular biology that refine our perception of the identity and dynamics of microorganisms. This review is an introduction to the *Microbiology Spectrum* Curated Collection: Advances in Molecular Epidemiology of Infectious Diseases that will discuss how these advancements have contributed to investigations of infectious disease outbreaks/epidemics, surveillance, transmission dynamics, risk factor identification, pathogenesis, and etiologic attribution of bacterial, viral, protozoan, and helminthic pathogens to a disease. Here we define "molecular epidemiology" and distinguish it from other disciplines that use many of the same molecular biology tools—taxonomy, phylogenetics, and molecular evolution of microorganisms. The Curated Collection will be spread throughout multiple issues of *Microbiology Spectrum* and will be divided into four general sections: (i) laboratory methods used to strain type microbial pathogens, (ii) methods used to analyze genotyping data, (iii) examples of molecular epidemiologic investigations of bacterial, viral, and parasitic diseases, and (iv) applications of molecular epidemiology to address new research questions in communicable and noncommunicable diseases. The major theme of this Curated Collection is to address the following question frequently asked by clinicians, clinical microbiologists, and public health professionals: what is the advantage or unique contribution of molecular epidemiology in solving infectious disease problems in the clinical and public health arenas? *This article is part of a curated collection.

INTRODUCTION

In 1980, Huang et al. published a paper in *New England Journal of Medicine* with the title "Molecular Epidemiology of Cytomegalovirus Infections in Women and Their Infants" (1). This is the first paper that is retrieved with the search term "molecular epidemiology of infectious diseases" queried against the NCBI PubMed bibliographic search engine. In 2016, the number of such "hits" peaked at 1,235. A 2004 book (*Molecular Epidemiology of Infectious Diseases: Principles and Practices*) made an attempt to "codify" the discipline of molecular epidemiology based on studies published through the early 2000s (2). Since 2004, new molecular biology technology and its applications have greatly expanded the scope and depth of epidemiologic

Received: 5 March 2018, **Accepted:** 18 June 2018, **Published:** 2 November 2018

Editor: Michael Sadowsky, BioTechnology Institute, University of Minnesota, St. Paul, MN

Citation: Riley LW, Blanton RE. 2018. Advances in molecular epidemiology of infectious diseases: definitions, approaches, and scope of the field. *Microbiol Spectrum* 6(6):AME-0001-2018. doi:10.1128/microbiolspec.AME-0001-2018.

Correspondence: Lee W. Riley, lwriley@berkeley.edu

Curated Collection: Advances in Molecular Epidemiology of Infectious Diseases.

investigations of infectious diseases. Today, molecular epidemiology is an integral part and a distinct branch of the discipline of epidemiology. This review defines molecular epidemiology and describes the scope of its investigative mandate. It presents an overview of how past and new molecular microbiology techniques are applied to address and solve various infectious disease epidemiology issues that were not possible to resolve by traditional non-laboratory- or laboratory-based methods.

The discipline of epidemiology focuses on populations (human or nonhuman animals). In epidemiology, study populations are dichotomized or stratified according to shared characteristics and the subgroups that are generated become units for comparative analyses. Molecular epidemiology just happens to use the tools of molecular microbiology to generate such subgroup data that can then be analyzed by observational and experimental techniques. As this series will emphasize, laboratory methods, while not always necessary to solve an epidemiologic problem, create an opportunity to identify more targeted intervention strategies. This series will demonstrate that molecular microbiology techniques can generate information that is more granular and therefore more informative than that produced by traditional laboratory or nonlaboratory methods to characterize infectious disease occurrence, temporal trend, pathogen reservoir, mode and pattern of disease transmission, risk factors, etiology attribution, and pathogens' genetic determinants of disease transmission. Genotype data are now also used to develop more predictive models of infectious disease transmission.

Molecular epidemiology, by definition, is laboratory based. Laboratory-based epidemiology constantly evolves as new tools are developed. The ability to culture bacteria in an artificial medium, first developed by Robert Koch in the late 1880s, made possible epidemiologic investigations of diseases such as tuberculosis and cholera (3, 4). It led to the discovery of the modes of transmission and risk factors for these infectious diseases. This knowledge, in turn, led to more focused public health intervention and prevention strategies that are still practiced today. The molecular microbiology methods developed more than 100 years later to genotype *Mycobacterium tuberculosis* and *Vibrio cholerae*, the agents of tuberculosis and cholera, respectively, have further advanced our epidemiologic knowledge of these diseases. Today we can estimate the proportion of new cases of tuberculosis in a community that result from recent transmission as opposed to those that result from infections that occurred many years earlier. We now know the natural reservoir of *V. cholerae* and how this reservoir can trigger cholera epidemics and pandemics. The ability today to rapidly sequence the entire genomes of multiple isolates of a pathogen helps us understand more precisely the transmission pathways of infectious agents. Molecular epidemiology is thus constantly evolving along with molecular microbiology knowledge, and this is why this discipline has become a distinct branch of epidemiology.

This review (i) defines molecular epidemiology and discusses the scope of investigations covered by this discipline, (ii) differentiates molecular epidemiology from other disciplines that use the same molecular microbiology methods, and (iii) reviews specific infectious disease epidemiologic issues addressed by the application of molecular microbiology tools.

DEFINITIONS AND SCOPE OF MOLECULAR EPIDEMIOLOGY

One textbook defines "epidemiology" as "the study of the distribution and determinants of diseases and injuries in human populations" (5). Veterinary researchers would, of course, replace "human populations" with "nonhuman animal populations." A simple way to define "molecular epidemiology" is to amend this standard definition as follows: "the study of the distribution and determinants of diseases and injuries in human and nonhuman animal populations using molecular microbiology methods." Molecular epidemiology studies could be descriptive or analytical. The latter type of studies always includes a comparison group (control), and it emphasizes identifying "determinants" of disease distribution.

In general, infectious disease epidemiology deals with descriptively and quantitatively characterizing the following concerns: (i) disease occurrence and distribution in time and place, (ii) reservoir of infectious agents, (iii) modes and pattern of disease transmission, (iv) the setting of disease transmission, (v) pathogen-related biologic factors that influence transmission, (vi) host-related factors (demographic, behavioral, clinical, genetic, and microbiota) that influence transmission, (vii) environmental factors (socioeconomic, anthropologic, and ecologic) that influence transmission, and (viii) assigning an etiologic role of a microbe to a newly recognized disease or a disease previously not recognized to be associated with an infectious agent. Molecular microbiology techniques have been applied to address each of these areas of investigation for a wide variety of infectious diseases, and such investigations would be classified as molecular epidemiology studies.

Some may question this definition of molecular epidemiology. Just because tools applied to the analysis of an organism's DNA are used to characterize infectious disease pathogens in epidemiologic investigations, why should this approach be called "molecular epidemiology"? After all, when biochemical tests or immunological tests are used, we do not refer to such studies as "biochemical epidemiology" or "immunological epidemiology." Nor do we call them molecular, although molecules are involved. Be that as it may, the term "molecular epidemiology" has become established, just as the term "biostatistics" has become established, when it is in fact no different from statistics.

MOLECULAR EPIDEMIOLOGY VERSUS TAXONOMY, PHYLOGENETICS, AND MOLECULAR EVOLUTION

When the NCBI bibliographic search engine is queried with the search term "molecular epidemiology of infectious diseases," more than 8,680 publications appear (as of February 2018), going all the way back to 1980. Many of these publications only describe genetic characteristics of microbes. Many such publications with the words "molecular epidemiology," even in the title, describe genetic characteristics of microbes without reference to any human or animal characteristics. Epidemiology of infectious disease deals with human subjects (or nonhuman animals) that become exposed to and acquire infectious agents. Thus, to describe genetic analyses of microbes as epidemiology is as absurd as stating that infectious agents acquire disease. Such studies should be more appropriately described as "molecular taxonomy," "phylogenetics," or "molecular evolution." Taxonomy is the science of classification of organisms into naturally related groups based on a factor common to each. Phylogenetics is the study of lines of descent or evolutionary history of individual or a group of organisms (2). Molecular evolution is phylogenetics primarily based on analyses of nucleic acid sequences to infer evolutionary relationships of organisms (6).

The disciplines of molecular epidemiology, taxonomy, and phylogenetics/molecular evolution all share the same molecular microbiology laboratory tools, but each discipline has its distinct purpose. In taxonomy, phylogenetics, and molecular evolution, phenotypic and genotypic data of microbes are compared to each other (Fig. 1A), and they are then separated into discrete groups. The relatedness of the organisms is compared numerically according to similarity coefficient measures based on operational taxonomic units (numerical taxonomy) or inferred from an evolutionary model (cladistics or evolutionary phylogenetics). Thus, the main goal of these disciplines is to identify relationships among collections of microbes based on their genetic and phenotypic characteristics. In evolutionary phylogenetics, changes in nucleic acid sequences observed over time in microbes are assumed to be inherited events. The relationships of these observed genetic changes are inferred, based on some evolutionary model that may not be empirically validated. An evolutionary model may be built upon a set of assumptions and observations made with microbial DNA or RNA samples that happen to be available or accessible in the present time (e.g., microbes or microbial nucleic acid trapped in fossil, amber, arctic ice, or permafrost). The model may change as more microbial samples become available.

In the molecular epidemiology of infectious diseases, the genotyped microbes are compared to each other also, but the microbe relationship data are linked to the hosts from which these microbes are isolated. The hosts belong to a larger group of hosts (population) in a particular environment or a transmission pathway in which the microbes disseminate (Fig. 1B). The main goal of molecular epidemiology is thus to identify or rule out an epidemiologic relationship among hosts based on genotypic characteristics of the microbes interacting with these hosts within a defined environment. Hosts are the main target of analysis. Just as host characteristics, such as sex, age, ethnicity, education, or occupation, can be used to identify risk factors for a disease, microbes isolated from hosts can be genetically characterized to identify risk factors for transmission, disease manifestation, and disease progression. Most importantly, what distinguishes molecular epidemiology from the other disciplines is that it is grounded in the motivation to identify and then maximize an intervention or disease prevention strategy to solve a public health problem.

Despite the differences pointed out above, however, it is important to emphasize that in molecular epidemiology, many of the computational tools and statistical inference methods developed under phylogenetics/molecular evolution are routinely used to show evidence for epidemiologic relationships of the causalities of infectious diseases. Just as molecular microbiology tools have evolved, so have the analytical and computational tools of molecular phylogenetics, especially with the advent of next-generation sequencing (NGS) technologies, and they have now become vital for molecular epidemiologic investigations. These phylogenetic principles and applications will be discussed in more detail in the other reviews of this series.

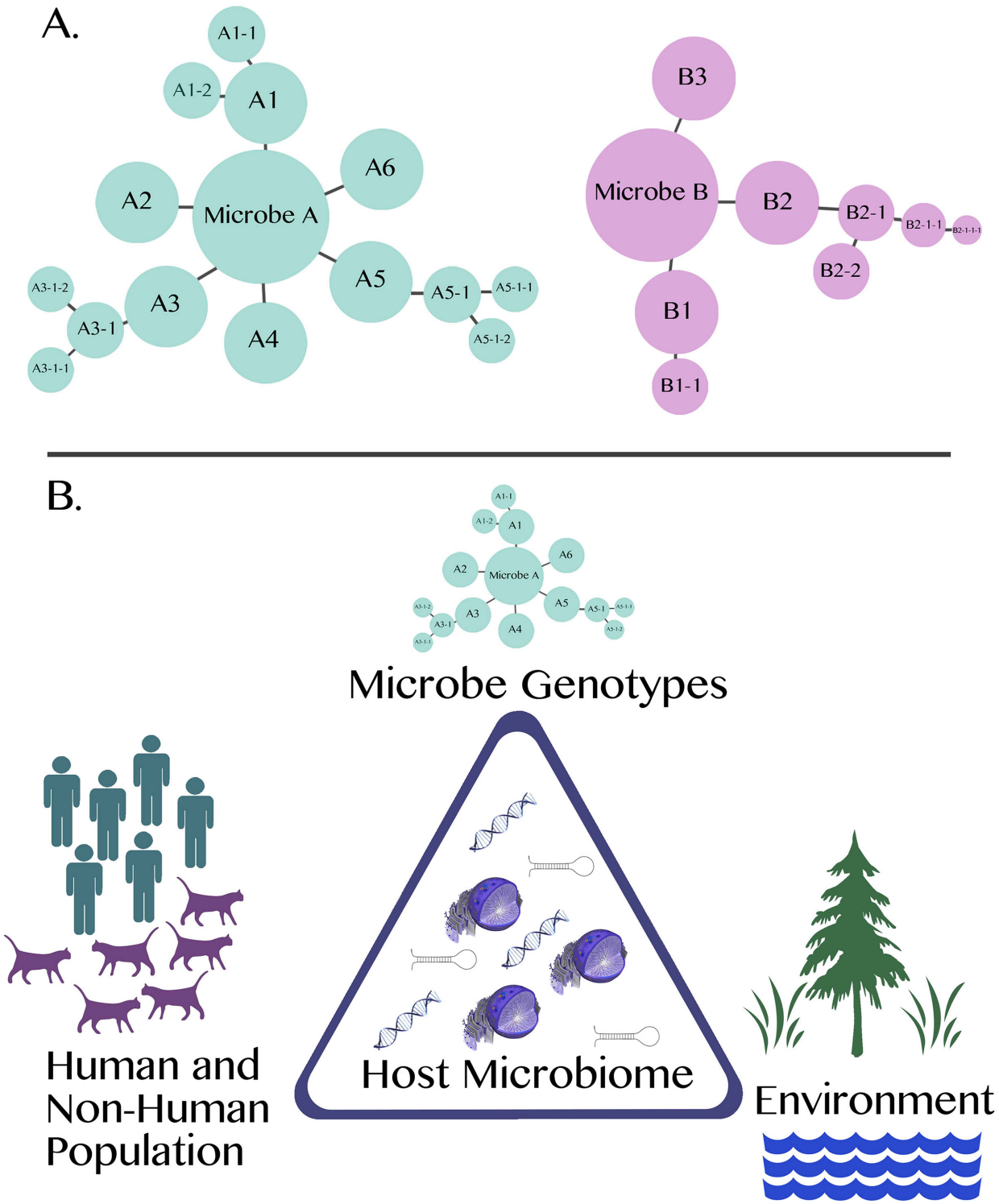

FIGURE 1 Differentiating molecular epidemiology from taxonomy, phylogenetics, and molecular evolution. **(A)** In taxonomy, phylogenetics, and molecular evolution, microbes' genetic profiles and relatedness are compared to each other. The main goal of these disciplines is to identify biologic relationships among collections of microbes based on their genetic or phenotypic characteristics. **(B)** In infectious disease molecular epidemiology, the genotyped microbes are also compared to each other, but the microbial relationship data are linked to a population of hosts from which these microbes are isolated in a particular environment or transmission pathways of these microbes. Recent observations highlight the important contribution of microbiomes in the host-microbe-environment relationship. (Illustrated by Paolo Harris Paz.)

INFECTIOUS DISEASE EPIDEMIOLOGIC ISSUES ADDRESSED BY MOLECULAR EPIDEMIOLOGY

The range of infectious disease epidemiologic issues that can be addressed by the application of molecular microbiology tools is dependent on the type of information that can be generated by the tools. Prior to the period of nucleic acid sequencing technology, most genotyping methods relied on comparison of electrophoretic band patterns of DNA or RNA fragments.

Electrophoretic band pattern analyses provided an opportunity to address a variety of epidemiologic issues. However, the later introduction of PCR technology and then nucleic acid sequencing technology expanded the range of investigative opportunities beyond what was possible by electrophoretic band pattern analyses. Below we discuss categories of epidemiologic issues of infectious disease addressable by molecular epidemiology methods.

Tracking Geographic and Temporal Distribution of Infectious Disease Agents

The study of the distribution of disease by time and place is part of the definition of epidemiology. A large proportion of published literature on molecular epidemiology of infectious diseases describes geographic and temporal distribution of distinct strains of microbes characterized genetically. These studies are largely descriptive. Consequently, factors that contribute to strain distribution (host, strain, and environment-related determinants of transmission) are often unknown, not described, or simply conjectured. That is, the determinants of strain distribution are not empirically characterized in such studies.

For example, in South America, a large proportion of tuberculosis cases are caused by a strain of *Mycobacterium tuberculosis* belonging to the Latin American-Mediterranean (LAM) lineage, a genotype characterized by a method called spoligotyping (7–9). As the name implies, this genotype is also widely distributed in the Mediterranean countries. One speculation to explain the geographic distribution of the LAM lineage is the high frequency of migration of Europeans and South Americans to and from these regions that began during the colonial period (7, 8). Another *M. tuberculosis* lineage—Beijing clade—is most commonly found in East and Southeast Asia (10, 11). It is distributed in many Northern Hemisphere countries, including countries in Western Europe and North America, but is uncommon in Southern Hemisphere countries, except in South Africa (11). Again, this pattern of distribution is speculated to have resulted from distinct human population migration patterns in the recent or remote past (10, 12).

Tracking strains across different geographic regions and time and speculating on how such distributions occur constitute an interesting topic for academic conversations, but from a public health or clinical perspective, such studies do not answer this question posed by clinicians, clinical microbiologists, and public health professionals: what is the advantage or unique contribution of molecular epidemiology in solving infectious disease problems in the clinical and public health arenas? Strain distribution studies rely on phylogenetic models; hence, as discussed above, the interpretations of the data are inferred from isolates that happen to be available. The generated information is rarely usable for infectious disease control and has not contributed to saving human or animal lives. As mentioned above, one major goal of epidemiology is to identify and maximize appropriate intervention or disease prevention opportunities. For those who wish to engage in such application of molecular epidemiology, the following categories of infectious disease issues need to be addressed.

Distinguishing Epidemic from Endemic Disease Occurrence

Infectious diseases can occur as part of an outbreak/epidemic or as part of seemingly unrelated sporadic/endemic number of cases. An outbreak is defined as occurrence of a disease in excess of what is expected at a particular place and time. In August 1978, a single case of cholera was diagnosed in Abbeville, LA (13). This single case would be considered an outbreak since no other cases of cholera had been reported from Louisiana since 1873 (13).

While outbreaks of foodborne diseases, such as salmonellosis, can be detected by surveillance systems, most reported cases of salmonellosis are not clearly linked to any recognized outbreaks. In fact, less than 10% of reported cases of salmonellosis come from recognized outbreaks in the United States (14, 15). The rest are reported from so-called sporadic cases. Cases of salmonellosis exhibit seasonal fluctuation, with predictably increased number of cases during the summer months in the temperate zone. Such cases of salmonellosis are considered to exhibit an endemic pattern of occurrence. Molecular epidemiologic investigations would reveal that many of the communicable infectious disease occurrences that seemingly exhibit an endemic pattern of occurrence are, in fact, comprised of multiple small outbreaks. This idea will be discussed in a later review of this series.

There are other infectious diseases that are not readily recognized to occur as part of an outbreak or an epidemic. Diseases such as community-acquired urinary tract infection (UTI) and pneumococcal pneumonia are rarely described to occur as outbreaks. Molecular epidemiologic analyses, however, would reveal that such infections can and do occur as epidemics. The application of molecular microbiology tools may unmask an epidemic from what appears to be sporadic, epidemiologically unrelated cases of an infectious disease. The recognition that endemic occurrence of an infectious disease may represent a collection of several small outbreaks is one of the major discoveries made by molecular epidemiologic investigations.

Stratification of Data To Refine Study Designs

As mentioned above, one of the first steps taken in an epidemiologic investigation is to subgroup the population in which the epidemic occurs. This can be done by dichotomizing the target population into those with and without the disease or by further subgrouping them according to some shared characteristics regardless of whether they have the disease. If no laboratory tools are available, a target population can be subdivided according to characteristics such as sex, age, ethnicity, national origin, occupation, education, or some observable clinical manifestation. For example, in an outbreak of a diarrheal illness, cases could be defined as those with diarrhea in a population that occurred within a defined period. The groups separated into diarrhea (cases) versus no diarrhea (controls) groups can then be examined for possible association with the above host-related characteristics (case-control analysis). In contrast, if these subgroups based on shared host-related characteristics are each and separately analyzed for the occurrence of diarrhea or no diarrhea, a subgroup characteristic associated with diarrhea may also be discovered (cohort analysis). In both types of analysis, the target population is subgrouped or stratified either by disease state (outcome) or by host, environment, or microbe characteristics (exposure). The discovered association with disease will be limited to one or more of the subgroup characteristics initially included in the analysis. Thus, in the above example, an association with diarrhea that may be discovered would be one or more of the above host-related characteristics. Intervention options based on this finding would be somewhat limited. Not much can be done about these host characteristics to prevent diarrhea, other than to just state that these are risk factors for diarrhea.

If, however, an investigator assumes that the diarrhea was caused by some contaminated food product, the target population can be further stratified according to their recent food intake history. If assumed correctly, the investigator may be able to implicate a contaminated food product. However, if the diarrhea was caused by an agent that is predominantly transmitted person to person (e.g., norovirus), or if analyses did not include the implicated food item in the survey instrument, no association may be discovered.

If the target population is further characterized clinically and with laboratory tools, additional subgroups or strata could be created. If most of the diarrhea cases were, for example, discovered to have been caused by *Salmonella enterica* serovar Enteritidis, a recognized agent of foodborne disease, an investigator could design a more detailed and focused survey instrument to assess recent food intake history. Knowledge about the types of food product previously shown to be associated with *S.* Enteritidis (e.g., poultry and eggs) would prompt an investigator to ask about such exposures. In a large target population, diarrhea may be caused by a variety of agents. Thus, in this example, the diarrhea cases could be further stratified into those with *S.* Enteritidis in stool versus other or no organisms. *S.* Enteritidis may also be isolated from individuals who did not have any diarrhea symptoms. Thus, a new case definition could be created, based on the presence of *S.* Enteritidis in stool, regardless of the symptoms. This case definition refinement may enable the investigator to implicate a specific contaminated food vehicle. Thus, the additional clinical and laboratory-generated data provide another level of data stratification that makes it possible to identify a more specific risk factor for an illness in question.

In public health, it is not sufficient to just implicate a contaminated vehicle. To prevent recurrence of such contamination, one must also be able to demonstrate how an implicated food product was contaminated in the first place. *S.* Enteritidis is one of the most commonly reported *Salmonella* serotypes that cause salmonellosis in the United States. An outbreak of *S.* Enteritidis diarrhea can occur from contamination of a food product that occurred locally at the time of food preparation (e.g., preparation of tiramisu with contaminated raw egg at a local Italian restaurant) or distally at a poultry farm or slaughterhouse. Application of a genotyping tool may enable an investigator to further stratify the isolates of *S.* Enteritidis. The genotype information may enable an investigator to trace the strain back to a source where the contamination occurred, correct the problem that contributed to the contamination, and prevent recurrence.

Thus, as these examples illustrate, each level of data refinement creates an opportunity to make a new case

definition, which could then be used to conduct a case-control or cohort analysis to identify a more specific association or risk factor. The identification of a specific association creates an opportunity to devise a more targeted intervention strategy. Molecular microbiology tools are just one way to stratify data used to refine study designs to identify the most optimal intervention options. This concept will be further illustrated with specific examples in later reviews.

Distinguishing Pathovars versus Commensal Organisms or Saprophytes

One unfortunate consequence of our inability to stratify microbes into finer discrete units is the confusion that arises in our attempt to characterize a pathogen versus a commensal or saprophytic organism belonging to the same species (species as defined by a standard taxonomic procedure). For example, we recognize that *Escherichia coli* is a member of the commensal flora of intestines of warm-blooded animals. Commensal organisms, by definition, do not cause disease. Yet we recognize that *E. coli* can cause diarrhea, UTI, bloodstream infection, sepsis, wound infection, and meningitis. Do all these *E. coli* strains—commensal and disease-causing strains—belong to a single taxonomic group?

E. coli is a taxonomic unit based on biochemical characterization of a discrete group of Gram-negative bacterial organisms. In most clinical and public health laboratories, no further differentiation (other than drug susceptibility testing or serotyping) is made. *E. coli* isolated from a "sterile" body site (blood, urine, or cerebrospinal fluid) associated with clinical manifestation consistent with an infection is considered a pathogenic variant or type (pathovar or pathotype) of *E. coli*. *E. coli* isolated from stool from a healthy, asymptomatic person is considered a commensal organism. The problem arises when *E. coli* is isolated from stool of someone with diarrhea—how does one differentiate *E. coli* that represents the commensal flora from *E. coli* that causes diarrhea? The need to separate *E. coli* isolates into pathovars versus nonpathovars stems from an artificial constraint of our having to use the term *E. coli* to refer to a group of bacteria that happen to share common biochemical characteristics used in clinical settings to distinguish them from other infectious disease-causing bacterial organisms.

If the current molecular microbiology tests happened to have been used before these groups of bacteria were taxonomically classified as *E. coli*, we would most likely be referring to pathovar and nonpathovar *E. coli* groups as belonging to two or even more distinct taxonomic groups. Since this did not happen, we are stuck with the term *E. coli*; therefore, we need to introduce terms like "pathovar" and "pathotype."

Fortunately, because we can now use molecular microbiology tools to differentiate *E. coli* strains that cause diarrhea (referred to in this review as intestinal pathogenic *E. coli* [IPEC]) from commensal *E. coli*, we have learned a lot about the epidemiology of IPEC. This will be discussed further in a different review of this series.

What about *E. coli* strains that cause extraintestinal infections, such as UTI, bloodstream infection, and sepsis (referred to in this review as extraintestinal pathogenic *E. coli*)? Are they true pathovars like IPEC, or are they commensal *E. coli* strains that happen to breach a sterile niche to cause disease? What about other organisms that colonize nonsterile niches of the human or animal host or that are present as saprophytes in the environment? *Staphylococcus aureus* is a common colonizer of the human anterior nares. Pneumococci colonize the pharynx. *Helicobacter pylori* resides in the stomach. *Klebsiella pneumoniae*, *Acinetobacter baumanii*, and *Pseudomonas aeruginosa* can occur as saprophytes in the environment. Yet these organisms cause disease. Can these organisms be distinguished genetically as pathovars versus nonpathovars? Among protozoa, *Entamoeba histolytica* (pathogen) cannot easily be distinguished from the commensal ameba *Entamoeba dispar* (nonpathogen) except by PCR. The ability to distinguish them as pathovars versus nonpathovars facilitates appropriate treatment decision making and public health intervention. This is one important application of molecular epidemiology that will be explored further in the other reviews of the series.

Identifying New Modes of Transmission of Infectious Diseases

Communicable infectious diseases result from transmission of microbes from person to person or animal to person (zoonosis) or via contaminated food, water, or fomites, intermediate vectors such as mosquitoes, exposure to an environmental reservoir (such as soil containing *Bacillus anthracis* spores), laboratory exposure, and iatrogenic procedures. However, there are many infectious diseases in which the mode of transmission is not obvious. To this day, we do not have a clear understanding of how and where *Helicobacter pylori* is acquired. And is community-acquired UTI—the most common bacterial infectious disease of women—a transmissible infectious disease?

The ability to genotype microbes has helped to identify modes of transmission of infectious agents that have

not been previously described or known. One recent example is the discovery that Zika virus, which has long been known to be vector borne, can be transmitted sexually (16–20). Until recently, Ebola virus was recognized to be transmitted only by direct contact with blood of a patient with Ebola virus disease (EVD); we now know that it too can be transmitted sexually (21, 22). Another review will provide molecular epidemiology evidence that a large proportion of community-acquired UTIs may be foodborne. Identifying new modes of transmission is an important component of molecular epidemiology investigation which can contribute to devising new prevention and intervention strategies.

Studying Epidemiology of Health Care- or Institution-Associated Infectious Diseases

In the United States, 400,000 to more than 720,000 episodes of health care-associated infections (HAIs) are estimated to occur each year (23, 24). Drug-resistant infectious agents greatly complicate the clinical management of these HAIs. Several important epidemiologic questions arise in addressing the problem of infections recognized in health care settings, including the following. (i) What proportion of drug-resistant microbes detected in health care settings are indeed due to the selective pressure of antimicrobial agents used in such settings? (ii) How does one distinguish hospital outbreaks from endemic occurrence of HAIs? (iii) How does one differentiate contamination versus disease for a commensal organism isolated from a sterile body site? (iv) Are disease clusters observed in a particular unit of a hospital or institution caused by a single lineage of a pathogen or by different strains of opportunistic pathogens in hosts who have immunosuppression or were exposed to broad-spectrum antibiotics? (v) How does one determine the direction of transmission of an infectious agent when it is isolated from multiple patients or also found in the hospital environment or on a medical device? In many ways, the field of molecular epidemiology progressed due to advancements made in molecular microbiology tools designed to address the questions posed above (25–32).

Surveillance and Monitoring Response to Intervention against Infectious Diseases

Surveillance is defined as "a continuous and systematic process of collection, analysis, interpretation and dissemination of descriptive information for monitoring health problems" (33). The collected information is meant to be used to devise an intervention or prevention strategy. Surveillance is an invaluable tool for epidemiologic investigation of infectious diseases. It facilitates systematic collection of stratified data from populations over time in a defined geographic setting. As described above, the utility of infectious disease surveillance for public health or clinical intervention is enhanced by greater resolution of the data that are collected. Therefore, laboratory-based surveillance, especially that which uses molecular microbiology techniques, provides the greatest resolution of data and therefore opportunities to devise the most focused and targeted intervention. Surveillance is also used as a reference to validate new genotyping tests that can be applied to conduct epidemiologic investigations. Finally, it provides an opportunity to assess the impact of interventions that resulted from use of a molecular microbiologic subtyping test. The application of molecular epidemiology in infectious disease surveillance will be discussed in a later review.

Analyzing Population Dynamics of Parasitic Pathogens

Parasitic organisms—protozoa and helminths—are eukaryotes. That is, they have a membrane surrounding the nucleus and carry two sets of chromosomes (diploid) as opposed to a single, non-membrane-bound chromosome (haploid) found in prokaryotic organisms. As described above, bacterial organisms reproduce clonally; thus, the basic unit of analysis of an epidemiologic investigation of bacterial infections is the clone. Since many parasitic organisms reproduce sexually at some point in their development, clonal characterization cannot be an objective of molecular epidemiologic studies of these organisms. Rather, it is the population of parasitic organisms that becomes the epidemiologic unit of analysis. Genetic markers, such as microsatellites, are used to assess and compare allele frequencies in a population to characterize changes in population structure over time or place to study or infer epidemiologic events. While the basic concepts of molecular epidemiology still apply to parasites, their differences from prokaryotic organisms in reproductive cycle and genetic composition must be taken into consideration to characterize their epidemiology. These new concepts will be described in separate reviews and applied to studies of schistosomiasis and malaria as examples.

Analyzing Microbiomes To Study Noninfectious Disease Epidemiology

NGS technology makes it possible, in principle, to identify and quantitate all known organisms in an ecologic niche sample, collectively termed the microbiota, whether it is in the human body or in the environment.

This technology has engendered a new discipline called metagenomics. Applied to the human microbiota, metagenomics is a study of the totality of nucleic acids belonging to all microbes in or on the human body. Until recently, it had been often stated that the total number of bacterial cells in the human microbiota was 10-fold greater than the total number of human cells (34, 35). A more systematic calculation has determined that the total number of bacterial cells in an average 70-kg male human body is of the same order (3.8×10^{13}) as the total number of human cells (3.7×10^{13}) (36, 37). That is, humans are only half human. All of the cells—bacterial and human—compete for the same energy source and nutrients ingested via the human mouth. Changes, therefore, in the "one half" (microbiota) will have a profound effect on the "other half" (human body). Metagenomics studies have revealed that changes in the gut microbiota can dramatically affect body physiology (38) and contribute to major public health concerns, such as obesity and its associated chronic, noncommunicable diseases, including diabetes, hypertension, and dyslipidemia (39, 40). Transfer of the gut microbiota from one person to another (fecal microbiota transplantation) has been shown to have a beneficial effect, such as in the treatment of *Clostridium difficile* colitis (41), but also a potentially harmful effect, such as weight gain (42). Thus, while the target of analysis in metagenomics is the commensal flora, the information that is generated can be used to characterize the epidemiology of noninfectious diseases. This is a new chapter in molecular epidemiology, and examples of such application are discussed in a separate review.

Identifying Direction and Chain of Transmission of an Infectious Agent

By classical epidemiologic approaches, it is not straightforward to characterize the direction of transmission of an infectious agent, especially when the transmission involves multiple person-to-person transmission events, contaminated fomites, or a contaminated environment. When two or more individuals are infected by a pathogen transmitted by direct human contact in a congregate setting or in a social network, how does one show the direction or chain of transmission using traditional or pre-NGS molecular microbiology techniques? This is an issue that comes up frequently in health care settings. In a hospital setting, how does one show that a health care provider is the source or the recipient of an infectious agent when multiple patients are found to be infected with the same agent or the same genotype of this agent? Contamination of a fomite (e.g., bedsheet, sink, or floor) or a medical device (e.g., bronchoscope) from infected patients can occur, but such items may also serve as a source for patient infections. For example, without the nucleic acid sequence information, a medical device may be blamed as the source of patient infections and may be subjected to an expensive decontamination procedure that may not contribute to interrupting an outbreak. In a health care setting, a detailed comparison of sequential single nucleotide changes revealed by whole-genome sequencing (WGS) of a collection of bacterial or viral isolates from different patients, fomites, and devices, together with detailed epidemiologic information, may be used to more precisely track both the direction and chain of transmission (29).

Identifying Hidden Social Networks To Unmask Infection Transmission Links

A WGS analysis of sexually transmitted pathogens may reveal a social network that was not previously recognized. However, the identification of such a network requires isolation of a recognized pathogen from individuals with a disease. Often, this is not possible because an individual may already have been treated with an antibiotic, was cultured late in the course of an illness, or may just not be available to be tested.

In another type of disease, such as EVD, the incubation period following exposure to blood of a diseased individual may be as long as 2 weeks. It is often difficult to determine who should be monitored and quarantined after possible exposures. A rapid identification of a social network of exposed individuals is vital for diseases like EVD, since the disease is associated with high mortality. Can the network be determined before the expected end of the incubation period without detecting the virus from possibly exposed individuals?

Another situation in which the determination of a social network is useful is in identifying a cause of a cluster of a disease of unknown etiology. For example, cancer clusters are frequently reported from various locations and if found to represent a true cluster (i.e., number in excess of what is expected at that location and time), possible environmental, chemical, occupational, drug, or infection etiologies are suggested (43, 44). If, somehow, these clusters can be shown to represent a distinct social network, an etiologic transmissible agent may be considered and then identified.

In each of the above-mentioned situations, some way to identify a social network or close social relationships in the absence of information about etiology is needed. This can be done, of course, by interviewing individuals, but information obtained from such a method is often

not reliable. NGS technology can be applied to address this issue by targeting commonly shared human commensal microbes—e.g., oral microbiota—as a surrogate for some suspected transmissible agent. The human mouth microbiota contains a distinct set of microbes shared by everyone. It may be compared among people in a disease cluster to unmask a hidden social network or to estimate degrees of social contact (45). Once such a network is identified, this information may be used to identify all individuals in a transmission pathway of a sexually transmitted disease, triage individuals for quarantine during an EVD epidemic, or search for a causative transmissible agent in a cluster of a disease of unknown etiology. Such an application of the NGS technology is completely new, and it will be discussed in another review.

Identifying Genetic Determinants of Disease Transmission

One new obligation of molecular epidemiology involves the study of biological determinants of disease transmission. Every communicable agent has evolved to ensure its survival in a community of hosts. Human enteric pathogens, such as nontyphoidal *Salmonella* spp., must reside and replicate inside a nonhuman animal host intestine (e.g., in cattle, poultry, or reptiles) to perpetuate themselves. They have successfully adapted to colonize these host reservoirs without causing disease. They cannot reside or survive long term outside of such hosts. The disease that occurs in humans infected with such *Salmonella* organisms, therefore, is an accident. Thus, their genetic determinants of transmission must be examined at the level of nonhuman animal reservoirs.

On the other hand, *Salmonella enterica* serovar Typhi is uniquely adapted to the human host. It has evolved to establish an asymptomatic chronic carriage state only in the human host, but it also must be able to escape the human host to transmit to other humans. This is done via exiting from the intestine. Mechanisms of human-specific adaptation, prolonged colonization (asymptomatic infection), and exit from the intestine are genetically determined. These host-adapted features of *Salmonella* spp. are critical for transmission to their respective new hosts and define their distinct epidemiologic behavior.

Unlike *Salmonella* spp., pathogens such as *Mycobacterium tuberculosis* absolutely require symptomatic disease (tuberculosis) to facilitate their transmission to new hosts. As with *S.* Typhi, humans are the only natural reservoir of *M. tuberculosis*. Latent tuberculosis infection (LTBI) is the most common phase of *M. tuberculosis* infection, but *M. tuberculosis* cannot transmit itself to new hosts from those with LTBI. The bacterium is most efficiently transmitted to new hosts if the disease occurs in the lungs. That is, the organism has evolved to remain latent most of the time, especially when its infected host is immunologically intact and healthy, but it must ultimately escape the host to enter a new host for it to survive long term. Tuberculosis most often occurs in a host whose immunity wanes (e.g., with old age, diabetes, cancer, or AIDS). Somehow, *M. tuberculosis* can recognize this disruption of host immunity. The disease required for transmission occurs in the lungs, which elicits cough. Cough will allow *M. tuberculosis* to escape the diseased host via aerosol droplets. These aerosol droplets, if inhaled by another host, will ensure long-term survival of that strain of *M. tuberculosis*. All of these steps—adaptation to the human host, latent infection during a period of intact host immunity, recognition of waning host immunity, and tropism for lungs—are all genetically determined as part of *M. tuberculosis*'s strategy to maintain its transmission in a community of human hosts. A study of these genetic determinants of transmission and how they work is part of molecular epidemiology research.

CONCLUDING REMARKS

As can be surmised from the discussion above, the review series to follow will be structured to cover genetics of microbes and how such information is applied to conduct epidemiologic investigations. Of course, molecular epidemiology of infectious disease can just as well cover host genetics. Host susceptibility genes can certainly determine infectious disease outcome and transmission, and therefore, they should serve as targets of molecular epidemiology studies. However, it is beyond the scope of this series to cover host genetics of infections, and the reader is referred to other excellent sources for this topic (46, 47).

Molecular epidemiology is now an established branch of epidemiology. As with other tools, such as biostatistics, molecular epidemiology evolves constantly as new analytical and quantitative tools are developed. Each step of technical advancement contributes to further refinement and new ways to conduct analytical epidemiologic investigations. The reviews in this series will expand on the discussion above with specific examples to illustrate these concepts. These examples will be preceded by reviews that present an overview of laboratory and analytical techniques that are used to conduct epidemiologic investigations. They are included to provide readers with basic familiarity with microbiologic,

phylogenetic, and computational biology concepts and methods relevant to the discussion of these infectious disease examples.

CURATED COLLECTION

Click here to read other articles in this collection.

REFERENCES

1. Huang ES, Alford CA, Reynolds DW, Stagno S, Pass RF. 1980. Molecular epidemiology of cytomegalovirus infections in women and their infants. *N Engl J Med* **303:**958–962. http://dx.doi.org/10.1056/NEJM198010233031702.

2. Riley LW. 2004. *Molecular Epidemiology of Infectious Diseases: Principles and Practices*. ASM Press, Washington, DC. http://dx.doi.org/10.1128/9781555817688.

3. Hitchens AP, Leikind MC. 1939. The introduction of agar-agar into bacteriology. *J Bacteriol* **37:**485–493.

4. Brock TD. 1988. *Robert Koch. A Life in Medicine and Bacteriology*. Science Tech Publishers, Madison, WI.

5. Mausner JS, Bahn AK. 1974. *Epidemiology. An Introductory Text*. WB Saunders Co, Philadelphia, PA.

6. Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York, NY.

7. Lopes JS, Marques I, Soares P, Nebenzahl-Guimaraes H, Costa J, Miranda A, Duarte R, Alves A, Macedo R, Duarte TA, Barbosa T, Oliveira M, Nery JS, Boechat N, Pereira SM, Barreto ML, Pereira-Leal J, Gomes MG, Penha-Goncalves C. 2013. SNP typing reveals similarity in *Mycobacterium tuberculosis* genetic diversity between Portugal and Northeast Brazil. *Infect Genet Evol* **18:**238–246. http://dx.doi.org/10.1016/j.meegid.2013.04.028.

8. Ritacco V, Iglesias MJ, Ferrazoli L, Monteserin J, Dalla Costa ER, Cebollada A, Morcillo N, Robledo J, de Waard JH, Araya P, Aristimuño L, Díaz R, Gavin P, Imperiale B, Simonsen V, Zapata EM, Jiménez MS, Rossetti ML, Martin C, Barrera L, Samper S. 2012. Conspicuous multidrug-resistant *Mycobacterium tuberculosis* cluster strains do not trespass country borders in Latin America and Spain. *Infect Genet Evol* **12:**711–717. http://dx.doi.org/10.1016/j.meegid.2011.06.006.

9. Cardoso Oelemann M, Gomes HM, Willery E, Possuelo L, Batista Lima KV, Allix-Béguec C, Locht C, Goguet de la Salmonière YO, Gutierrez MC, Suffys P, Supply P. 2011. The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. *PLoS One* **6:**e18256. http://dx.doi.org/10.1371/journal.pone.0018256.

10. Wan K, Liu J, Hauck Y, Zhang Y, Liu J, Zhao X, Liu Z, Lu B, Dong H, Jiang Y, Kremer K, Vergnaud G, van Soolingen D, Pourcel C. 2011. Investigation on *Mycobacterium tuberculosis* diversity in China and the origin of the Beijing clade. *PLoS One* **6:**e29190. http://dx.doi.org/10.1371/journal.pone.0029190.

11. Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, Sola C, Rastogi N. 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol* **41:**1963–1970. http://dx.doi.org/10.1128/JCM.41.5.1963-1970.2003

12. Mokrousov I, Ly HM, Otten T, Lan NN, Vyshnevskyi B, Hoffner S, Narvskaya O. 2005. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. *Genome Res* **15:**1357–1364. http://dx.doi.org/10.1101/gr.3840605.

13. Blake PA, Allegra DT, Snyder JD, Barrett TJ, McFarland L, Caraway CT, Feeley JC, Craig JP, Lee JV, Puhr ND, Feldman RA. 1980. Cholera—a possible endemic focus in the United States. *N Engl J Med* **302:**305–309. http://dx.doi.org/10.1056/NEJM198002073020601.

14. Centers for Disease Control and Prevention. 2000. Surveillance for foodborne-disease outbreaks—United States, 1993–1997. *MMWR Surveill Summ* **49:**1–62.

15. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV. 1999. Food-related illness and death in the United States. *Emerg Infect Dis* **5:**607–625. http://dx.doi.org/10.3201/eid0505.990502.

16. Musso D, Roche C, Robin E, Nhan T, Teissier A, Cao-Lormeau VM. 2015. Potential sexual transmission of Zika virus. *Emerg Infect Dis* **21:**359–361. http://dx.doi.org/10.3201/eid2102.141363.

17. Turmel JM, Abgueguen P, Hubert B, Vandamme YM, Maquart M, Le Guillou-Guillemette H, Leparc-Goffart I. 2016. Late sexual transmission of Zika virus related to persistence in the semen. *Lancet* **387:**2501. http://dx.doi.org/10.1016/S0140-6736(16)30775-9.

18. Foy BD, Kobylinski KC, Chilson Foy JL, Blitvich BJ, Travassos da Rosa A, Haddow AD, Lanciotti RS, Tesh RB. 2011. Probable non-vector-borne transmission of Zika virus, Colorado, USA. *Emerg Infect Dis* **17:**880–882. http://dx.doi.org/10.3201/eid1705.101939.

19. Davidson A, Slavinski S, Komoto K, Rakeman J, Weiss D, Centers for Disease Control and Prevention. 2016. Suspected female-to-male sexual transmission of Zika virus—New York City, 2016. *MMWR Morb Mortal Wkly Rep* **65:**716–717. http://dx.doi.org/10.15585/mmwr.mm6528e2.

20. Hills SL, Russell K, Hennessey M, Williams C, Oster AM, Fischer M, Mead P, Centers for Disease Control and Prevention. 2016. Transmission of Zika virus through sexual contact with travelers to areas of ongoing transmission—continental United States, 2016. *MMWR Morb Mortal Wkly Rep* **65:**215–216. http://dx.doi.org/10.15585/mmwr.mm6508e2.

21. Christie A, Davies-Wayne GJ, Cordier-Lassalle T, Blackley DJ, Laney AS, Williams DE, Shinde SA, Badio M, Lo T, Mate SE, Ladner JT, Wiley MR, Kugelman JR, Palacios G, Holbrook MR, Janosko KB, de Wit E, van Doremalen N, Munster VJ, Pettitt J, Schoepp RJ, Verhenne L, Evlampidou I, Kollie KK, Sieh SB, Gasasira A, Bolay F, Kateh FN, Nyenswah TG, De Cock KM, Centers for Disease Control and Prevention. 2015. Possible sexual transmission of Ebola virus—Liberia, 2015. *MMWR Morb Mortal Wkly Rep* **64:**479–481.

22. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, Christie A, Schroth GP, Gross SM, Davies-Wayne GJ, Shinde SA, Murugan R, Sieh SB, Badio M, Fakoli L, Taweh F, de Wit E, van Doremalen N, Munster VJ, Pettitt J, Prieto K, Humrighouse BW, Ströher U, DiClaro JW, Hensley LE, Schoepp RJ, Safronetz D, Fair J, Kuhn JH, Blackley DJ, Laney AS, Williams DE, Lo T, Gasasira A, Nichol ST, Formenty P, Kateh FN, De Cock KM, Bolay F, Sanchez-Lockhart M, Palacios G. 2015. Molecular evidence of sexual transmission of Ebola virus. *N Engl J Med* **373:**2448–2454. http://dx.doi.org/10.1056/NEJMoa1509773.

23. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R, Maloney M, McAllister-Hollod L, Nadle J, Ray SM, Thompson DL, Wilson LE, Fridkin SK, Emerging Infections Program Healthcare-Associated Infections and Antimicrobial Use Prevalence Survey Team. 2014. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med* **370:**1198–1208. http://dx.doi.org/10.1056/NEJMoa1306801.

24. Zimlichman E, Henderson D, Tamir O, Franz C, Song P, Yamin CK, Keohane C, Denham CR, Bates DW. 2013. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA Intern Med* **173:**2039–2046. http://dx.doi.org/10.1001/jamainternmed.2013.9763.

25. Elwell LP, Inamine JM, Minshew BH. 1978. Common plasmid specifying tobramycin resistance found in two enteric bacteria isolated

from burn patients. *Antimicrob Agents Chemother* **13:**312–317. http://dx.doi.org/10.1128/AAC.13.2.312.

26. Sadowski PL, Peterson BC, Gerding DN, Cleary PP. 1979. Physical characterization of ten R plasmids obtained from an outbreak of nosocomial *Klebsiella pneumoniae* infections. *Antimicrob Agents Chemother* **15:**616–624. http://dx.doi.org/10.1128/AAC.15.4.616.

27. Schaberg DR, Tompkins LS, Falkow S. 1981. Use of agarose gel electrophoresis of plasmid deoxyribonucleic acid to fingerprint gram-negative bacilli. *J Clin Microbiol* **13:**1105–1108.

28. Tompkins LS, Plorde JJ, Falkow S. 1980. Molecular analysis of R-factors from multiresistant nosocomial isolates. *J Infect Dis* **141:**625–636. http://dx.doi.org/10.1093/infdis/141.5.625.

29. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA, NISC Comparative Sequencing Program Group. 2012. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* **4:**148ra116. http://dx.doi.org/10.1126/scitranslmed.3004129.

30. Singh A, Goering RV, Simjee S, Foley SL, Zervos MJ. 2006. Application of molecular techniques to the study of hospital infection. *Clin Microbiol Rev* **19:**512–530. http://dx.doi.org/10.1128/CMR.00025-05.

31. Carle GF, Frank M, Olson MV. 1986. Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. *Science* **232:**65–68. http://dx.doi.org/10.1126/science.3952500.

32. Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR, Cookson BT. 2015. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *J Clin Microbiol* **53:**1072–1079. http://dx.doi.org/10.1128/JCM.03385-14.

33. Rothman KJGS, Lash TL. 2008. *Modern Epidemiology*, 3rd ed. Wolters Kluwer/Lippincott Williams and Wilkins, Philadelphia, PA.

34. Savage DC. 1977. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* **31:**107–133. http://dx.doi.org/10.1146/annurev.mi.31.100177.000543.

35. Luckey TD. 1972. Introduction to intestinal microecology. *Am J Clin Nutr* **25:**1292–1294. http://dx.doi.org/10.1093/ajcn/25.12.1292.

36. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol* **14:**e1002533. http://dx.doi.org/10.1371/journal.pbio.1002533.

37. Sender R, Fuchs S, Milo R. 2016. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell* **164:**337–340. http://dx.doi.org/10.1016/j.cell.2016.01.013.

38. Harris K, Kassis A, Major G, Chou CJ. 2012. Is the gut microbiota a new factor contributing to obesity and its metabolic disorders? *J Obes* **2012:**879151.

39. Blaser M. 2014. *Missing Microbes: How the Overuse of Antibiotics Is Fueling Our Modern Plagues*. Henry Holt & Company, Inc, New York, NY.

40. Riley LW, Raphael E, Faerstein E. 2013. Obesity in the United States—dysbiosis from exposure to low-dose antibiotics? *Front Public Health* **1:**69. http://dx.doi.org/10.3389/fpubh.2013.00069.

41. Brandt LJ. 2012. Fecal transplantation for the treatment of *Clostridium difficile* infection. *Gastroenterol Hepatol (N Y)* **8:**191–194.

42. Alang N, Kelly CR. 2015. Weight gain after fecal microbiota transplantation.*Open Forum Infect Dis* **2:**ofv004. http://dx.doi.org/10.1093/ofid/ofv004.

43. Caldwell GG. 1990. Twenty-two years of cancer cluster investigations at the Centers for Disease Control. *Am J Epidemiol* **132**(Suppl)**:**S43–S47. http://dx.doi.org/10.1093/oxfordjournals.aje.a115787.

44. Thun MJ, Sinks T. 2004. Understanding cancer clusters. *CA Cancer J Clin* **54:**273–280. http://dx.doi.org/10.3322/canjclin.54.5.273.

45. Francis SS, Plucinski MM, Wallace AD, Riley LW. 2016. Genotyping oral commensal bacteria to predict social contact and structure. *PLoS One* **11:**e0160201. http://dx.doi.org/10.1371/journal.pone.0160201.

46. Harrison L, Griffin DE. 1993. *Infectious Diseases*. Academic Press, Inc, San Diego, CA. http://dx.doi.org/10.1016/B978-0-08-092566-0.50016-1.

47. Khoury MJBT, Cohen BH. 1993. *Fundamentals of Genetic Epidemiology*. Oxford University Press, New York, NY.