Identifying strains that contribute to complex diseases through the study of microbial inheritance

Jeremiah J. Faith\textsuperscript{a,1}, Jean-Frédéric Colombel\textsuperscript{b,1}, and Jeffrey I. Gordon\textsuperscript{c,1}

\textsuperscript{a}Immunology Institute and Institute for Genomics and Multiscale Biology and \textsuperscript{b}Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; and \textsuperscript{c}Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108

Edited by Inder M. Verma, The Salk Institute for Biological Studies, La Jolla, CA, and approved December 9, 2014 (received for review September 29, 2014)

It has been 35 y since Carl Woese reported in PNAS how sequencing ribosomal RNA genes could be used to distinguish the three domains of life on Earth. During the past decade, 16S rDNA sequencing has enabled the now frequent enumeration of bacterial communities that populate the bodies of humans representing different ages, cultural traditions, and health states. A challenge going forward is to quantify the contributions of community members to wellness, disease risk, and disease pathogenesis. Here, we explore a theoretical framework for studies of the inheritance of bacterial strains and discuss the advantages and disadvantages of various study designs for assessing the contribution of strains to complex diseases.

In a 1977 issue of PNAS, Woese and Fox defined the 16S rRNA gene as a molecular and evolutionary marker to distinguish members of the domain Archaea from those belonging to Bacteria (1). This report profoundly changed the course of microbiology and opened the door for culture-independent studies (2). With the advent of next-generation sequencing, there has been an explosive increase in the use of this phylogenetic marker to characterize microbial communities in the environment and those associated with the body habitats of myriad animal species, including our own. The increasing ease and economy of defining human microbial ecology within and across individuals has spawned many studies that have sought to correlate community membership with health status (3–9).

These studies have highlighted the variation in microbial diversity between body habitats in a given individual and interindividual variation in community membership within a given habitat over time, as well as substantial interpersonal differences. They have also revealed how a given species-level phylogenetic group in a given body habitat’s microbial community (microbiota) can be composed of multiple strains of that species, each harboring a set of shared genes, as well as genes that are variably represented across strains (10–12). This variation in a species “pan-genome” has yet to be systematically characterized for a large number of taxa recovered from different habitats of a given individual, or between large numbers of individuals. Defining this strain-level diversity and its biological significance requires moving beyond the 16S rRNA gene to genome-wide analyses.

Identifying the “effector” strains represented in a microbiota that are responsible for shaping specific facets of our human biology is a daunting challenge because there are a large number of possible combinations of component organisms that could operate to produce a given phenotype. In this Perspective, we discuss the proposal that characterizing acquisition, retention, and passage of strains (“microbial inheritance”) will be a useful element in the quest to identify effector taxa associated with health or disease (Fig. 1).

Using the human gut microbiota and its bacterial constituents as an illustration, we consider this proposal from the perspective of four concepts: (i) The majority of resident strains are acquired during the first three years of life; (ii) once acquired, the majority of strains are retained in an individual for decades; (iii) strain sharing is biased toward proximal microbe-rich sources (e.g., “family members”) with very little to no strain overlap in the microbiota of “nonfamily” members; and (iv) the effect of a strain’s residency may take decades to manifest itself.

The gut microbiota of humans is dominated by relatively few bacterial phyla compared with other body habitats (13); a notable feature of this community is its great strain-level diversity (10, 12, 13). The difficulty in identifying causative strains for complex diseases may relate to the fact that they represent species that are common in the gut microbiota of healthy individuals. With the advent of methods for culturing a significant amount of the bacterial diversity present in a person’s fecal microbiota (14, 15), performing whole-genome sequencing of members of clonally arrayed culture collections provides one way to identify the strain-level variation that an individual harbors as a function of time, health status, and other covariates (6, 10). Although “traditional” 16S rRNA sequencing lacks the resolution and depth to track the representation and persistence of bacterial strains within a microbiota, a method known as low error amplicon sequencing (LEA-Seq) uses redundant sequencing (∼10× coverage per 16S rRNA gene amplicon) to achieve this goal with high resolution in samples collected over time from subjects (10). For example, when applied to 37 healthy adults sampled for up to 5 y, LEA-Seq revealed that each person harbored ∼100 bacterial species and ∼200 bacterial strains in their fecal microbiota. On average 60% of strains were retained in

Author contributions: J.J.F., J.-F.C., and J.I.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This article is part of the special series of PNAS 100th Anniversary articles to commemorate exceptional research published in PNAS over the last century.

See companion article, “Phylogenetic structure of the prokaryotic domain: The primary kingdoms” on page S088 in issue 11 of volume 74, and see Inner Workings on page 641.

1To whom correspondence may be addressed. Email: jeremiah.faith@mssm.edu, jean-frederic.colombel@mssm.edu, or jgordon@wustl.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1418781112

PNAS | January 20, 2015 | vol. 112 | no. 3 | 633–640
A Model of Microbial Inheritance

Although the logic and dynamics of microbial inheritance will require experimental and theoretical refinement, a simple probabilistic model of host colonization provides a structure for understanding key variables and the effect of modifications to these variables. In this simplified model, we do not account for effects of host genotype on selection of particular taxa. This exclusion is not to imply that host genetics play no role. Well-controlled quantitative trait loci mapping studies in mice indicate that microbiota composition is affected by host genetic factors, including those involved in immune responses, as well as by environmental exposures (18, 19). Although studies of dizygotic and monozygotic twins suggest that the representation of members of the gut community is impacted by genotype (11, 20), genome-wide surveys are needed to address the important question of the contributions of host genetics versus environment on microbiota structure and function. In our simplified model, we assume the probability that the (gut) microbiota of any individual is colonized with strain \( i \) (i.e., \( P(\text{colonized}) \)) has the following form

\[
P(\text{colonized}) = (P(\text{transmission}) \times P(\text{access}) \times (1 - P(\text{resistance})))
\]

where \( P(\text{transmission}) \) is the probability of transmission of strain \( i \), \( P(\text{access}) \) is the probability of access to strain \( i \), and \( P(\text{resistance}) \) is the resistance of the host to colonization by strain \( i \). The probability of transmission for a particular gut strain will depend on factors such as its ability to survive environmental insults (e.g., aerotolerance, spore-forming capacity) and its competitiveness (ability to stably establish itself in one or more available niches). Access will depend on the microbe’s abundance in the source community as well as the hygiene, frequency of contact with, and the health status of people in near proximity to the recipient. Resistance will depend on the current composition of the potential recipient’s microbiota, niche availability in the recipient’s microbiota for strain \( i \), and the hygiene of the recipient, as well as other factors including the recipient’s immune status and genotype (e.g., features of their innate immune system).

According to our model, microbial inheritance occurs preferentially between family members relative to unrelated individuals because \( P(\text{access}) \) is far higher for microbes harbored by individuals we most frequently encounter and because \( P(\text{resistance}) \) is lower during the first 2–3 y of postnatal life when microbial community assembly is occurring (Figs. 1 and 2). For example, comparison of draft genomes of strains of \( i \) a prominent saccharolytic bacterium in the human gut microbiota, \( Bacteroides thetaiotaomicron \) (10), and \( ii \) the principal human gut archaeon, \( Methanobrevibacter smithii \), isolated from the fecal microbiota of several families (11) disclosed that the same strains were shared between a majority of adult family members. Viewed in this light, microbial inheritance, where a proportion of our microbial

---

Fig. 1. Human genetic inheritance and microbial inheritance. (A) In human genetic inheritance, information transmission is vertical, with 50% of each child’s genome content inherited from the father and 50% from the mother. Note that birth order is from left to right. (B) In microbial inheritance, transmission occurs largely in the first 3 y of life from proximal microbe-rich sources, notably family members, and in particular between near-birth siblings who provide a higher degree of access to their microbes. However, microbial inheritance is not limited to family members and can include other environmental sources, including family friends, nearby surfaces, and the local water supply.
A

infant 3 year adult age

high low

P(access,)
P(resistance,)

B

P(colonized,)

C

P(colonized,)

Fig. 2. Consequences of microbial inheritance. (A) As we age, the likelihood of transmitting and receiving microbial strains changes. The increased incidence of vomiting, diarrhea, and acute respiratory illnesses at early ages, combined with less established hygiene practices, likely enables young children to provide higher access to their microbial inhabitants. In addition, the less established community of microbes harbored in and on their bodies provides less resistance to invasion and establishment of new organisms. (B) The consequences of these age-associated changes in access and resistance are that the probability that a young child is colonized by a given microbial strain is higher when both parents harbor the strain than when only one parent does (i.e., \( P(\text{access}) \) is higher because there are two low-access reservoirs of the strain rather than one), and higher if a near-birth sibling and a parent are colonized by the strain than if both parents are colonized (i.e., \( P(\text{access}) \) is higher because there is one low-access reservoir and one high-access reservoir of the strain rather than two low-access reservoirs). (C) In the context of multiple siblings, access is highest in near-birth siblings.

Tracking microbes through families where members have a disease that is not monogenic, and where an intersection of host genetic as well as environmental factors are posited to contribute to pathogenesis, offers an opportunity to identify important microbial contributors to risk, prevention, and/or treatment of that “complex disease.” In an illustrative hypothetical example, consider gastric cancer, a disease correlated with the decades-long presence of some strains of \( H. \text{ pylori} \). In our example, monozygotic twins have a father who harbors an \( H. \text{ pylori} \) strain containing the cag pathogenicity island (\( \text{cag}^+ \)) and a mother who is colonized with a \( \text{cag}^- \) strain. During the first 2 y of life, each sibling may acquire an \( H. \text{ pylori} \) strain that retains its membership in their gastric microbiota for decades, or for the remainder of their lives. The strain they acquire is critically important because acquisition of the \( \text{cag}^+ \) strain will increase their risk of gastric cancer roughly twofold (29). The probabilistic model emphasizes the importance of which strain is transmitted first to one of the cotwins; the increased number of reservoirs, combined with the increased contact between siblings who are members of a twin pair, will increase the chance that the second twin will be colonized by the same strain harbored by the first cotwin to be colonized (Fig. 2), and thus help determine their concordance for this disease.

Microbial inheritance could involve acquisition of one or a number of effector strains that individually or collectively modulate the risk of developing complex diseases. Just as polygenic disorders are far more common than monogenic disorders, we posit that complex diseases are more likely to reflect the contributions of multiple microbial strains that both mediate and mitigate pathogenesis in the context of a specific host genotype. Thus, the dynamics of their exchange (inheritance) would be more complex than in the simplified example presented above and, in the case of twins, help define the degree of concordance or discordance for disease. Crohn’s disease, rheumatoid arthritis, and multiple sclerosis are examples of complex diseases where microbial agents have been hypothesized to have etiologic roles (30–36). Concordance for disease in monozygotic twins is below 40% in all three diseases (37–39). Genome-wide association studies have identified overlapping loci between the three diseases, including loci involved in immune regulation as well as microbial recognition and defense (40–42). Antibiotics are used to treat all three diseases (43–45) but with limited efficacy. Mouse models mimicking features of these diseases have shown a significantly lower incidence of pathology (or no pathology)
when they are reared germ-free compared with their conventionally raised counterparts (i.e., mice that acquire a microbiota from their environment beginning at birth) (46–48).

**Family Studies**

**Transmission Among Spouses/Cohabiting Adults.** The lack of transmission of complex disease between spouses is cited as evidence against microbial agents playing a role in pathogenesis or at least as evidence that a shared environment later in life is not an important etiologic factor. Although we are unaware of any evidence beyond case reports of an increase in relative risk for the spouse of an individual with multiple sclerosis, a small but significant increase risk of rheumatoid arthritis has been observed (49). In an analysis of two cohorts living in Belgium and France, the frequency of Crohn’s disease after marriage to an individual with this disease was significantly increased relative to what would be expected by chance alone given the population size studied (50). In both rheumatoid arthritis and inflammatory bowel disease, relative risk was extremely small (1.17 and 1.82, respectively). These small effect sizes may be related to \( P(\text{resistance}) \) being high in adulthood either because existing members of the microbiota prevent new strains from establishing themselves or because the partner has already developed immune mechanisms to eliminate the strains.

Although increased relative risk upon cohabitation with an affected spouse may be small, the relative risk of Crohn’s disease (51), rheumatoid arthritis (49), or multiple sclerosis (52) is highly increased when a sibling is affected (6.3, 4.6, and 7.1, respectively). In the context of microbial inheritance, the shared risk factors are stably colonized microbial strains inherited between family members that increase (or decrease) the probability of complex disease over a lifetime.

Before identification of *H. pylori* as the etiologic agent, familial aggregation was cited as evidence supporting genetic factors over environmental factors in the pathogenesis of peptic ulcer disease. In a 1950 analysis of familial risk, Doll and Buch (53) stated that “the possibility that similar environments may have been responsible for the results cannot be excluded. It is unlikely, as an excess of ulcers occurred in two generations of the relatives. For if environmental influences were responsible they would have to have operated in childhood, and it is difficult to believe that the fathers have been equally subjected to the same influences in their childhood”. In light of our current knowledge of the persistence and familial transmission of *H. pylori* (23, 54–57), an increased incidence in peptic ulcer disease across two generations can just as easily be ascribed to transmission of a pathogenic strain as it can to the genetic origins suggested by Doll and Buch.

**Transmission Among Siblings.** Given that \( P(\text{access}) \) for effector strains will be higher when proximal environmental sources have less developed hygiene practices, more episodes of shedding, increased number of reservoirs (e.g., multiple sibling carriers), and increased contact with the recipient (Fig. 2), one would also expect an increased relative risk of disease in the context of affected siblings and clustering of cases between consecutively born children. Blaser et al. found a statistically significant correlation between higher birth order and gastric ulcer (58).

In the context of cag\( ^{+} \) *H. pylori* strains, the risk of developing gastric cancer was twice as high for individuals in sibships of more than seven individuals compared with individuals in sibships in one of three individuals (59). In a study of 102 Crohn’s disease sibships with at least two affected siblings and one healthy sibling, we observed a statistically significant increase in affected siblings in consecutive births relative to nonconsecutive births (60). In a study of multiple sclerosis in 370 Canadian twin pairs, a trend toward increased disease susceptibility was observed with dizygotic twins, with a proband-wise concordance of 5.4 ± 2.8% compared with 2.9 ± 0.6 for their non-twin siblings (37).

**Spatial Clusters**

Spatial clusters characterized by a statistically significant increase in disease prevalence in a particular geographic region provide compelling evidence for an environmental contribution, including a microbial one. Although such clusters are often disputed as statistical outliers, or dismissed for other reasons, the abundance of observed clusters in different geographic regions can cumulatively provide evidence for a shared environmental risk factor in disease pathogenesis. Disease clusters and outbreaks are systematically tracked for many viruses and foodborne illnesses whereas clusters of complex disease are more commonly identified from large cohorts (61) or through citizens and local physicians who notice and report a surprising number of cases in a given locale (62, 63).

In the probabilistic model of microbial inheritance, a disease cluster would form when \( P(\text{transmission}) \) is high and \( P(\text{access}) \) is high across a geographic region. Although most of our inherited microbes come from other humans with whom we interact, broad access to an effector strain or strains would likely require a nonhuman source. Public water supplies, swimming pools, rivers, lakes, and ponds are often hypothesized to be the source of regional clusters. These clusters also imply effector strains that can survive (and ideally multiply) outside of a human host, which is uncommon for the large proportion of taxa colonizing the gut. Although we are unaware of any documented geographic clusters of rheumatoid arthritis, numerous clusters of Crohn’s disease have been identified throughout the world, including Mankato, MN (64) and Blockley Parish, Gloucestershire, England (65). With large-cohort databases, geographic clusters can be detected systematically in an unbiased manner using spatial scan statistics that search large regions for variations in disease incidence and reduce the chance of spurious clusters due to sampling biases (66). We recently applied these statistical methods to the Registre Epidemiologique des Maladies de l’Appareil Digestif inflammatory bowel disease (IBD) cohort and identified 18 significant Crohn’s disease geographic clusters, the largest with 61 cases among a population of only 5,703 (10 times higher than expected) (61). Spatial techniques also identified geographic clusters of IBD in Manitoba, Canada (67). Multiple sclerosis has a large number of proposed disease clusters, including the Faroe Islands (36) and Key West, FL (63). Advances in spatial epidemiology plus geographical information systems and global positioning systems are already being applied to the study of infectious diseases (68), and their forthcoming application to chronic diseases should enable more rigorous and sensitive identification of spatial variation in complex diseases (62, 69).

**Microbial Inheritance as a Tool to Identify Effector Strains**

Several study designs and strategies can be envisioned for quantifying the risk associated with stably colonized and transmitted members of the (gut) microbiota. The key to such studies is to generate a “microbe type” equivalent to a genotype. This microbe type requires high-resolution tools to identify the unique set of microbes in each individual at the strain level.

An initial challenge with microbe typing is that the organisms in a community are present at different abundances and that their genomes have differing degrees of homology to one another. Therefore, a sample of that community needs to be assayed to a sufficient depth to detect all of the relevant organisms, and the assay must be of sufficient resolution to dereplicate the sequences into the unique set of strains.
present [high-abundance organisms will be assayed many orders of magnitude more frequently than low-abundance organisms, and assay noise (e.g., sequencing errors) could overinfluence the microbe type with false positives]. The most accurate method to microbe type would be to acquire the whole genome sequence of each member of the community, but this is not practical at the present time. Currently, high-throughput whole-genome microbe typing is largely limited to culturable organisms. For example, using this approach, we recently sequenced over 500 microbial genomes, generating microbe types for six individuals (10). By quantifying the fraction of DNA alignment between each pair of bacterial genomes (coverage score) (70), we found taxa from the same genus and species had coverage scores of 0.30 ± 0.20 and 0.77 ± 0.12 (mean ± SD), respectively. Strains of bacteria isolated from individuals over time and those shared between family members had a coverage score of >0.96. These findings led to a provisional proposal that a threshold for dereplication at the strain level would be to cluster isolates with genome coverage scores of >0.96 (10).

Microbe typing by bacterial isolation (culture) and genome sequencing is laborious and expensive. Therefore, finding alternative methods is desirable. Matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry, a culture-dependent method that identifies microbes from their protein spectra, provides a cost-effective and rapid alternative for dereplication at the species level and perhaps at the strain level (assuming further algorithmic and sample extraction optimizations in the future) (71, 72). Sequencing 16S rRNA amplicons and shotgun sequencing of total community DNA currently represent the two most commonly used culture-independent metagenomic technologies but suffer from low resolution and difficult dereplication, respectively.

One highly anticipated outcome of microbe-typing improvements is a critical assessment of how well the definition of “strain” [based on a highly accurate 16S rDNA (or larger rRNA locus) amplicon marker sequence, or other methods including MALDI-TOF] is indicative of a given gene content, or whether cultured isolates with the same marker sequence have varying gene content (73). Improvements in read length, read number (sampling depth), and read quality could help address this challenge by (i) enabling sequencing of several components of or the entire 16S rRNA-ITS-23S rRNA region with low-error amplicon methods (10, 74), and (ii) facilitating the dereplication/assemble of shotgun reads generated from community DNA. Improvements in the throughput and quality of single-cell DNA sequencing would provide an ideal solution with its culture independence and genome-wide resolution. Continued improvements in algorithms for the assembly (75) and quantitative analysis of metagenomics data (76–78), combined with improved microbe reference databases (79, 80) and a growing understanding of pan-genome variation within species, could spawn additional methods to track strains. These microbe-typing advances will need to be validated experimentally with defined artificial microbial communities composed of collections of sequenced organisms with differing degrees of genome similarity and different abundance distributions within these “mock” communities.

Across human populations, the specific microbial genes responsible for effects on health and disease pathogenesis may be present in the genomes of different microbial species and strains. The advantage of strain tracking is that all of the genetic information is “linked” in the genome sequence of each organism being passed between individuals: by tracking the presence or absence of strains with >96% shared genome content, statistical tests can be performed to identify strains that modulate disease risk without the need to initially identify the specific genes responsible; this linkage drastically reduces the number of variables and statistical tests. Alternatively, mapping shotgun sequencing datasets generated from microbiomes to reference gene collections (80) or reference microbial genomes (79) provides the possibility of directly identifying the genes or genomic variants that explain disease variation. However, these approaches are currently complicated by our lack of knowledge of the resolution at which pathogenic effects might occur (e.g., a rare, highly penetrant, disease-specific microbial gene versus a low penetrance polymorphism in a commonly occurring horizontally transferable microbial gene).

Study Designs

As in human genetics, several study designs can be used to attempt to identify strains with significant odds ratios between healthy and affected individuals. In the case-control design (Fig. 3A), microbe types are obtained from groups of affected and unaffected individuals with the goal of identifying strains that are significantly overrepresented in one of the two groups. This design is unlikely to be effective in identifying stably colonized effector strains because, unlike our human genomes, there seems to be very little to no strain overlap in the microbiota of nonfamily members. However, this important point needs more rigorous testing: e.g. whereas we did not observe a shared microbial strain (coverage score >0.96) in surveys of 21 unrelated people living in the United States and 578 cultured isolates representing 76 species that were cultured from their fecal microbiota, surveys of more isolates from a given individual and more individuals representing different lifestyles are needed. If the observation of little to no strain overlap in unrelated individuals holds, a case-control study would require sampling a very large population, ideally from a small geographic location with little immigration.

Studies of large families where multiple members manifest a complex disease provide a way to identify effector strains significantly associated with health status (Fig. 3B) because, as noted above, microbial strains are shared among family members. Powering such studies depends on the penetrance of each effector strain and the number of affected and unaffected individuals in the family; increased power requires large families with high disease frequency (e.g., for an effector strain with 80% penetrance, at least four affected and nine total members are needed to obtain P value < 0.05 by Fisher’s exact test). Although such methods could identify a single organism that best explains a condition (e.g., H. pylori in peptic ulcer disease), they should also be capable of identifying and quantifying the combined influence, both beneficial and detrimental, of multiple microbial community members. Powering studies to quantify the combined influence of multiple strains, each contributing a small amount of disease risk or prevention, requires additional sampling. However, there are statistical advantages with strain tracking compared with human genotyping; not all strains will be acquired by vertical (parental) inheritance, thereby reducing the number of candidate effector strains; transmission involves a highly unique microbial strain that is far less likely to cooccur between two unrelated individuals by chance alone compared with SNPs, which can manifest one of four nucleotide states. Identification of candidate effector strains would set the stage for development of strain-specific assays for screening larger populations (e.g., more distant relatives and family friends), for more accurate calculations of the risk associated with harboring a given strain, and for aiding identification of genetic regions in each strain that confer altered disease risk.

Spatial clusters, if caused by microbial inheritance, provide the sample source with the highest statistical power (Fig. 3C). Identification of shared microbial strains between affected unrelated individuals in a region would be a highly significant event. Studies of
Strains that are weighted by the probability that such an event would occur. A strain-level database in states or countries with complementary geographical information systems in their health departments would provide an opportunity to track potential effector strains moving through the population (68, 81, 82).

Gnotobiotic animal models could help us to understand the dynamics of strain transmission in a highly controlled manner because male and female mice colonized separately with defined cultured isolates from a human father’s microbiota and a human mother’s microbiota could be mated. Cohousing male and female animals could provide a way to study microbial transmission of different strains in the context of established adult microbial communities in animals given diets similar to that consumed by the human microbiota donors. The resulting litters or subsequent rounds of litters could be compared with estimates of the proportion of strains inherited maternally and paternally and the similarity and differences between litters from the same parents. “Humanized” gnotobiotic mouse models represent a facilitated test of transmissibility; their coprophagy ensures that $P(\text{access})$ is high and relatively constant across studies.

The equation and study designs described above aim to quantify the risk or benefit associated with strains that do or do not stably engrave in an affected population relative to a control population. Persistent colonization with an effector strain in peptic ulcer disease, gastric cancer, Whipple’s disease, leprosy, and tuberculosis provides precedent for the paradigm that a chronic progressive inflammatory condition requires chronic colonization. The epidemiologic studies described above provide evidence that such effects can be detected in population data as, for example, unexpected deviations in sibling disease risk. Thus, a static model provides the simplest starting point for understanding the microbiota’s contribution to disease risk at the strain level. However, as strain-level microbe-typing technologies improve to enable longitudinal studies, the equation could be modified by including terms for the probability of a strain’s extinction $[P(\text{extinction})]$, or to enable $P(\text{transmission})$, $P(\text{access})$, and $P(\text{resistance})$ of each strain $i$ to change as a function of time. These additions would facilitate identification of transient microbes that imprint a beneficial or deleterious trajectory on their host that manifests some time after their extirpation.

Spatial clusters should include unaffected, unrelated controls to ensure that strains identified as enriched are not enriched in all humans in the community due to shared environmental sources. If shared strains are identified between affected unrelated individuals, microbe typing a significant proportion of a small town with a complex disease cluster would also provide a wealth of information about microbial transmission and the degree of sharing of the town’s collective microbiota.

**Improving Tools and Resources for Microbial Inheritance Studies**

Microbe typing healthy families from different regions of the world with diverse dietary and cultural traditions would provide baseline transmissibility information, detailing what proportion of the (gut) microbiota is inherited from family members versus other environmental sources and which strains are more or less transmissible. A transmissibility database would facilitate development of statistical models for identifying effector
data collection. Our goal with the current model is to provide a starting point for assessing the implications of a microbiota acquired early in life and stably maintained for decades.

The search for microbial community contributions to complex human disease is increasingly focused on identifying potentially pathogenic states: i.e., a “dysbiosis.” Although numerous searches for organisms that cause complex disease have failed, we have lacked tools with sufficient resolution to determine whether, among common species in communities, there was a unique disease-associated strain or strains. Whether a given disease is mediated by a single or multiple organisms, we are still attempting to solve a community-wide problem in the sense that the goal is to identify key elements of the community, be they individual strains, metabolic states, or thus far unknown characteristics that are causally related to, rather than just the effect of, that disease. If the community state is the primary mediator of disease, the potential for a pathogenic conformation of the community is likely influenced or determined by the strain-level composition of the community.

Assuming that some diseases are mediated by multiple strains in a community, we must carefully quantify what this assumption entails. In the context of microbial inheritance, the probability that unrelated individuals share a single member of their microbiota at the strain level is small. If microbes were inherited independently, even a disease that is only mediated by two strains would represent the product of two small probabilities and would be extremely rare, unless the causative organisms developed non-independent transmission mechanisms, in which case such organisms should be identifiable by strain-level cooccurrence analyses of the microbiota of individuals with the conditions of interest.

The cases of multiple organisms mediating disease could represent multiple causative genes harbored in microbes across diverse phylogenetic groups yet these genes must be common enough in the global metagenome to allow the joint probability of their occurrence in a single affected host. As we move from one causative organism or gene to two and beyond, the probabilities quickly become vanishingly small unless these organisms or genes are common or have developed mechanisms to transmit non-independently through the human population. A major advantage of microbial inheritance, and the study designs described here, is that it is agnostic to the number of organisms and forges the need to initially search for causative genes by taking advantage of the linkage of genes to the strains that are being passed between hosts. Enrichment or depletion of one strain or many strains in affected individuals should be detectable.

Assuming that microbial inheritance plays a role in complex disease, the largest hurdle to quantifying this influence is our ability to consistently detect the same strains in and across samples. False positives in microbe-type analyses will likely be rare, particularly if whole-genome sequencing is used. However, false negatives (the inability to detect or isolate a particular organism in an individual’s microbiota) are likely to be an unwanted source of variance until better strain-level identification methods are available. One could argue that many “patterns” could be found that correlate well with disease in large samples by simple chance alone. However, these correlations should be rare because only a fraction of strains are shared even between related individuals. An additional complicating issue is where to search for the organisms of interest. The gut and mouth, which seem to be stable sources of microbial life (10, 83–85), represent attractive initial places to microbe type. However, effector strains that increase the probability of disease might also be found in higher abundances at local sites of disease. In addition, microbial typing should by no means be needed, but limited to bacteria because inherited effector strains could also be members of Archaea or Eukarya (or their viruses).

If disease develops from a stably colonized strain or collection of strains, or a lack of these strains, microbial inheritance provides a set of strategies to facilitate detection of such organisms. Studies of microbial inheritance can be applied to a broad range of complex diseases occurring in varied geographic and cultural settings, including, for example, undernourishment of children in low-income countries where the size of extended families can be large and spatial scan statistics may reveal previously unappreciated clusters of diseases that are long standing and common in the population or of diseases emerging coincident with Westernization. Finally, one hoped-for outcome of studies of microbial inheritance would be new strategies for disease treatment and prevention that involve addition to, or elimination of, strains from an individual’s microbiota. The anticipation is that one day we will look back at the history of diagnosing and treating complex diseases that once were progressive and refractory to cure and marvel why we were not able to see their microbial roots.

ACKNOWLEDGMENTS. Work cited from the J.J.F., J.-F.C., and J.I.G. laboratories was supported by grants from the NIH and the Crohn’s and Colitis Foundation of America.


Bureau of Environmental Health, Community Assessment Program (2012) Review of the Prevalence of Type 1 Diabetes Among Children and Adolescents in Weston, Wellesley and Newton (Massachusetts Department of Public Health, Boston, MA).


