REVIEW ARTICLE

GENOMIC MEDICINE
W. Gregory Feero, M.D., Ph.D., and Alan E. Guttmacher, M.D., *Editors*

Microbial Genomics and Infectious Diseases

David A. Relman, M.D.

From the Departments of Medicine and of Microbiology and Immunology, Stanford University, Stanford; and the Veterans Affairs (VA) Palo Alto Health Care System, Palo Alto — both in California. Address reprint requests to Dr. Relman at VA Palo Alto Health Care System, 154T, Bldg. 101, Rm. B4-185, 3801 Miranda Ave., Palo Alto, CA 94304, or at relman@stanford.edu.

N Engl J Med 2011;365:347-57.

The pace of technical advancement in microbial genomics has been breathtaking. Since 1995, when the first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was published,[1] 1554 complete bacterial genome sequences (the majority of which are from pathogens) and 112 complete archaeal genome sequences have been determined, and more than 4800 and 90, respectively, are in progress.[2] A total of 41 complete eukaryotic genome sequences have been determined (19 from fungi), and more than 1100 are in progress. Complete reference genome sequences are available for 2675 viral species, and for some of these species, a large number of strains have been completely sequenced. Nearly 40,000 strains of influenza virus[3] and more than 300,000 strains of human immunodeficiency virus (HIV) type 1 have been partially sequenced.[4] However, the selection of microbes and viruses for genome sequencing is heavily biased toward the tiny minority that are amenable to cultivation in the laboratory, numerically dominant in particular habitats of interest (e.g., the human body), and associated with disease.

In 2006, investigators reported in-depth metagenomic sequence data from a human mixed microbial community[5]; in 2007 more than 1000 genes from single cells of cultivation-resistant bacteria were identified.[6] Since then, a flood of such data has ensued (Fig. 1).[7-9] Individual investigators can now produce a draft sequence of a bacterial genome containing 4 million base pairs in about a day.[10-12] The revolution in DNA-sequencing technology has to a large extent democratized microbial genomics and altered the way infectious diseases are studied.[11] However, gene annotation and error correction still take time and effort. Today, the major challenges in microbial genomics are to predict the function of gene products and the behavior of organisms and communities from their sequences and to use genomic data to develop improved tools for managing infectious diseases.

Genomic Diversity

The human body contains remarkable microbial taxonomic richness, with thousands of symbiont species and strains per individual host. Of these, an estimated 90% have not yet been cultivated in the laboratory.[13] Differences between closely related strains and species are responsible for virulence, host-species adaptation, and other aspects of lifestyle and account for the individualized nature of the human microbiota. For example, the gene content of pathogenic and nonpathogenic strains of *Escherichia coli*, as well as different pathogenic types, varies by as much as 36%.[14,15] Comparisons of complete genome sequences from multiple strains of the same bacterial species reveal a set of core genes that are common to all strains and a set of dispensable genes that are absent in at least one strain.[16] The sum of these genes (i.e., those represented in at least one strain) constitutes the species pangenome.

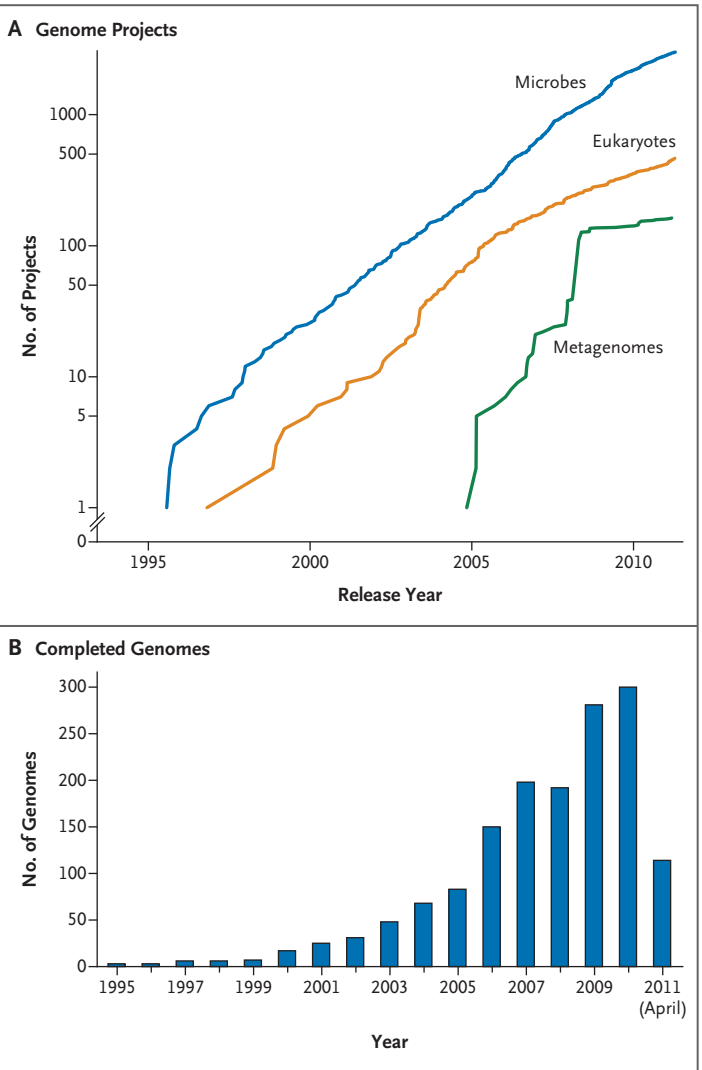

Figure 1. Genome Projects and Completed Genomes since 1995.

Panel A shows a cumulative plot of the number of genome projects — involving microbial (bacterial and archaeal), eukaryotic, and viral genomes — and metagenome projects, according to the release year at the National Center for Biotechnology Information since 1995. Panel B shows the number of completed microbial genome sequences according to year. (The most recent data were collected on April 21, 2011.)

As compared with the genomes of plants and animals, genomes of microbes are small and usually contain one or two chromosomes, as well as a variable number of plasmids (see Glossary). Yet, approximately 90% of a typical microbial genome encodes proteins or structural RNAs,[17] whereas only about 1.1% of the human genome is coding sequence.[18] As a result, some complex bacteria have more genes than some simple eukaryotes.

Microbial diversification and adaptation have been accompanied by gene loss and genome reduction, genome rearrangement, horizontal gene transfer, and gene duplication.[19,20] The first two of these processes are especially evident in human-specific pathogens, such as *Bordetella pertussis* (the causative agent of whooping cough),[21,22] *Tropheryma whipplei* (the agent of Whipple's disease),[23] and *Yersinia pestis* (the agent of bubonic plague). A total of 3.7% of *Y. pestis* genes appear to be inactive, especially those associated with enteropathogenicity.[24] The genome of *Mycobacterium leprae*, the cause of leprosy, provides an even more dramatic example of reductive evolution. Protein-coding genes account for less than half of its genome, whereas inactive and fragmented genes account for most of the remainder.[25]

Genomic islands are discrete clusters of contiguous genes found in bacterial chromosomes and plasmids, usually between 10,000 and 200,000 base pairs in length with features that suggest a history and origin distinct from other segments of the genome (see Glossary).[26,27] Some islands are stably assimilated into the genome; others appear to have been acquired recently and may still be mobile. Genomic islands enhance the fitness of the recipient by providing new, accessory functions, such as pathogenicity, drug resistance, or catabolic functions.

One of the most dramatic examples of short-term genome evolution can be seen in the CRISPR (clustered regularly interspaced short palindromic repeat) loci of bacteria and archaea. CRISPRs serve as a defense against invading phages and plasmids, in a manner akin to adaptive immunity.[28] These genomic loci contain segments of phage and plasmid sequences captured from previous encounters. These segments are stored within the CRISPR loci as spacer sequences and are expressed as small RNAs, which then interfere with replication of newly encountered phages and plasmids that bear the same sequences.

POPULATION STRUCTURE, EVOLUTION, AND MOLECULAR EPIDEMIOLOGY

Differences in the sequence and structure of genomes from members of a microbial population reflect the composite effects of mutation, recombination, and selection. With the increasing availability of genome sequences, these effects have become better characterized and more effectively exploited so as to understand the history and evolution of microbes and viruses and their sometimes

Glossary

DNA microarray: A technology that is used to study many genes or other sequences at once. Thousands of sequences are placed in known locations on a glass slide, silicon chip, or other surface. A sample containing DNA or RNA is deposited on the slide, which is sometimes referred to as a gene chip. The binding of complementary base pairs from the sample and the sequences on the chip can be measured with the use of fluorescence to detect the presence (and determine the amount) of specific sequences in the sample. In addition, when conserved sequences on the chip are used to capture any member of a family of related sequences in the sample, the bound DNA can be removed and its sequence determined.

Effector protein: A protein that is secreted by microbes directly into a host cell to alter physiological processes in the host. Pathogens use these proteins to subvert host defenses and hence to enhance infection, promote their survival, and produce disease.

Genome reduction: A decrease in the genome size during evolution of an organism, as measured by the total number of nucleotides or by the number of genes. Genome reduction in pathogens and symbionts often results from the deletion of genes that are no longer needed by or are disadvantageous to the microbe as it adapts to a host and becomes restricted to fewer habitats.

Genomic islands: Regions of a genome with distinct nucleotide composition or with clusters of genes that encode specialized functions, such as virulence attributes. These discrete genomic regions are believed to be acquired from other organisms by horizontal gene transfer and are flanked by direct repeats or phage attachment sites.

Horizontal (or **lateral**) **gene transfer:** The exchange of genetic material between contemporaneous, extant organisms, as compared with vertical inheritance of genetic material from an ancestor. Mechanisms of horizontal gene transfer include uptake of naked DNA (transformation), transfer mediated by plasmids or by pieces of DNA that promote their own transposition to new sites in a genome (conjugation), and transfer mediated by viruses (transduction).

Metagenomic analysis: Genomic analysis that is performed directly on a mixture of heterogeneous organisms, genomes, or genes.

Plasmid: An extrachromosomal, self-replicating piece of DNA. Plasmids are usually circular and transferrable between cells and sometimes carry genes that provide accessory functions, including drug resistance and virulence.

Pseudogene: A mutated form of a gene that is no longer functional, either because parts of the coding region are missing or altered or because it is no longer transcribed.

Ribosomal RNA: Noncoding ribonucleic acid that binds proteins to form the two subunits of the ribosome. All ribosomal RNAs (rRNAs) are named on the basis of their sedimentation rate, which is a reflection of their size and shape. In bacteria, the small subunit rRNA is called the 16S rRNA.

Shotgun sequencing: An approach in which thousands or millions of short, random fragments of a DNA sample are sequenced simultaneously and then reassembled with the use of computer algorithms on the basis of matching overlapping ends. This technique can be applied to metagenomic analysis.

intimate relationships with humans. The resulting insights have practical importance for epidemiologic investigations, forensics, diagnostics, and vaccine development.[29]

Y. pestis, the cause of the Black Death, arose from a more genetically diverse ancestor that was related to *Y. pseudotuberculosis*, through genome reduction and gene loss. By analyzing approximately 1200 single-nucleotide polymorphisms (SNPs) and a worldwide collection of strains, the origins of this monomorphic pathogen have been placed between 2600 and 28,000 years ago in China, from which it spread to other areas of the world, giving rise to country-specific lineages.[30] All the *Y. pestis* strains that are found in the United States today are descendants of a single import that probably arrived in San Francisco in 1899. As another example, patterns of early human migration have been traced by comparing genome sequences from contemporary isolates of the chronic gastric pathogen, *Helicobacter pylori*.[31] Transmission of the pathogen is primarily from mother or other household members to baby, and colonization is usually lifelong; thus, pathogen sequences are reasonable markers of host ancestry and host migration. Sequence data for the *H. pylori* genome indicate the sequential timing and directionality of two distinct waves of human migration into the Pacific region.[32] Population mosaicism in *H. pylori* gene sequences has been used to infer the history of social interactions in human populations.[31]

The power of full-genome sequencing to discriminate between closely related strains and track real-time evolution of disease-associated clonal isolates offers the possibility of tracing person-to-person transmission and identifying point sources of outbreaks. Using this approach, investigators established a previously unrecognized link among five patients with the same clonal strain of methicillin-resistant *Staphylococcus aureus* from a hospital in Thailand.[33] A study of *Vibrio cholerae* genome sequences from the October 2010 cholera outbreak

in Haiti suggested that the Haitian strains were clonal and more closely related to strains from Bangladesh that were isolated in 2002 and 2008 than to strains isolated in Peru in 1991 and in Mozambique in 2004. The authors concluded that the Haitian outbreak may have originated with the introduction of a *V. cholerae* strain from South Asia as a result of human activity rather than climatic events or other local environmental factors.[12] However, the source of this outbreak has not been fully resolved; genome sequences of environmental strains and additional clinical isolates from Haiti may provide further insight.

A major challenge is the prediction of patterns of evolution and emergence of disease agents. The antigenic evolution of influenza virus is known to follow a punctuated equilibrium model in which periods of relative virus stability around the globe are followed by periods of rapid change, requiring modification of the influenza vaccine. However, it was not clear whether variants arise first in East and Southeast Asia and then seed other geographic regions or whether strains persist locally and evolve simultaneously in a similar fashion. An analysis of the gene encoding hemagglutinin (the major antigenic determinant) from more than 1000 human influenza A (H3N2) isolates that were collected worldwide from 2002 through 2007 produced strong support for external seeding, rather than local persistence, and suggested that the source of seeding is East and Southeast Asia.[34] On the basis of whole-genome sequence analysis, the novel 2009 human H1N1 influenza strain was thought to have entered the human population in January of that year after arising from multiple swine virus progenitors that had probably been circulating in swine populations undetected for at least a decade.[35] Work of this type will help target efforts regarding influenza virus surveillance more effectively, refine the selection of vaccine strains, and improve predictions of future antigenic characteristics.[36] Similar approaches will assist in anticipating the emergence and spread of antibiotic and antiviral resistance.

PATHOGENESIS AND SYMBIOSIS

Pathogens have received most of the attention in microbial genomics, despite their relative rarity in the microbial world.[17,19] As a result, we now have a more complete and deeper understanding of how microbes cause disease and of pathogen emergence, host adaptation, and spread in human populations. The study of microbial genomes reveals four themes with respect to virulence.

First, horizontal gene transfer (see Glossary) has had a major role in the acquisition of genes associated with virulence. Most genes that encode virulence factors are physically segregated in clusters and located within mobile genetic elements. In *S. aureus*, these genes often occur within phage-related chromosomal islands and encode a variety of superantigens, including the toxin associated with the toxic shock syndrome and staphylococcal enterotoxin B, and encode factors that mediate antibiotic resistance, biofilm induction, and other virulence-associated properties (Fig. 2).[37] Genomic islands with similar features occur in other gram-positive bacteria, including streptococcus, enterococcus, and lactococcus species. The emergence of the recent Shiga toxin–producing *E. coli* clone in Germany was probably the result of horizontal gene transfer, when a toxin-producing phage infected an enteroaggregative *E. coli* strain.[38]

Second, symbionts and avirulent relatives of pathogens often contain many of the same virulence-associated genes as do the microbes that typically cause disease.[39] The genes that we commonly associate with virulence may have been selected for the advantages they confer in promoting colonization of animal and plant hosts, in avoiding or surviving phagocytosis, and in enhancing competition against symbionts.[40-42] For example, the original role for bacterial toxins may have been to protect the bacterium against predation by protozoa and nematodes. The legionella protein IcmT facilitates the escape of the bacterium from human macrophages and also from the far more ancient predator, the free-living amoeba.[43] Virulence depends on the choreographed expression of particular combinations of genes at the right place and time in the right host. Commensals and other symbionts also serve as reservoirs of antibiotic resistance genes and genetic diversity.[44]

A third theme is the surprising diversity of genes associated with mechanisms of virulence. In a study of four closely related fungal species, all of which cause late blight disease but in different host plant species, investigators identified specific regions of the fungal genomes with evidence of accelerated rates of evolution, suggesting that these regions have been under strong positive selective pressure.[45] The genes in these regions produce effector molecules (see Glossary)

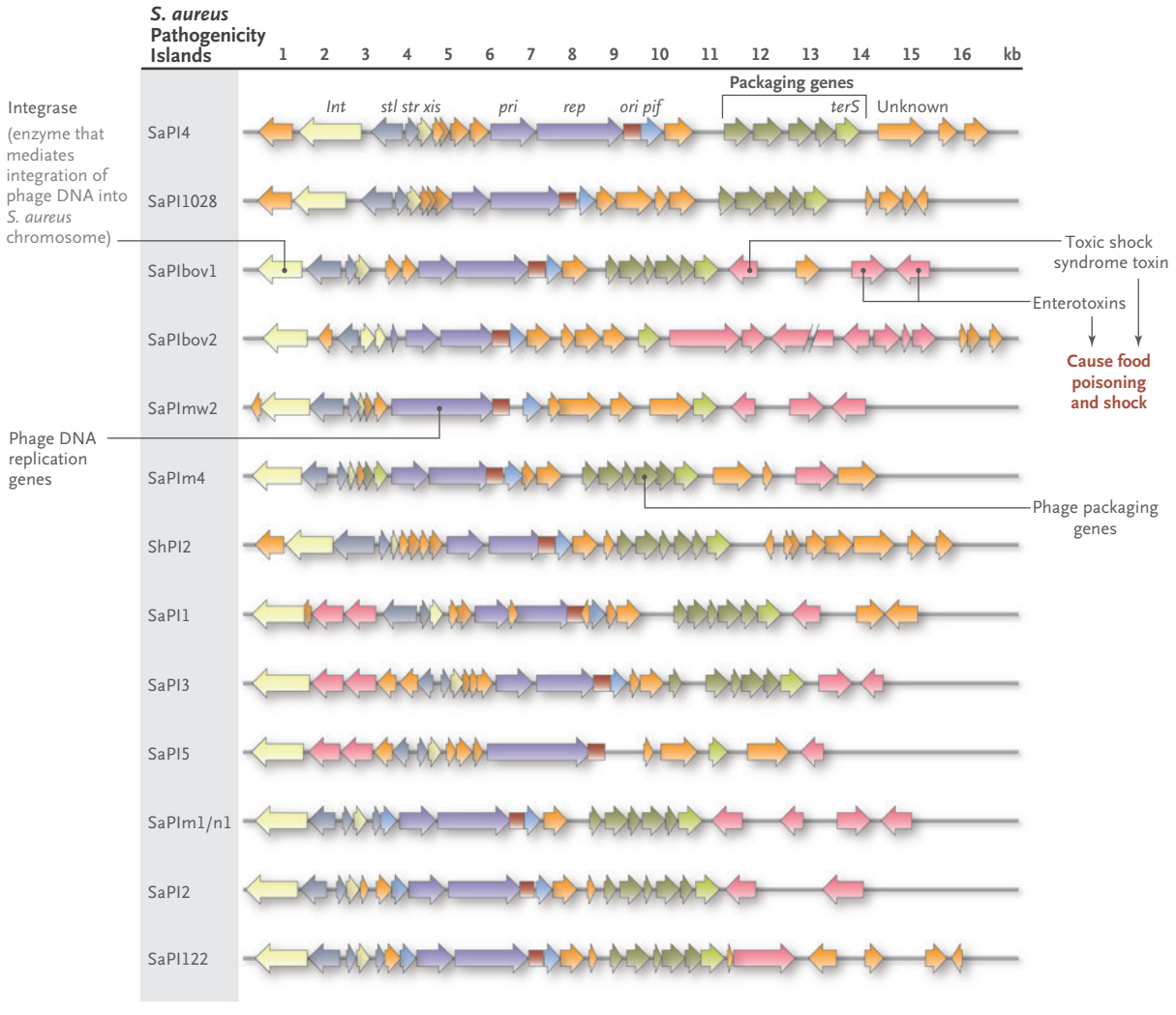

Figure 2. Unexpected Diversity of Virulence Factors for *Staphylococcus aureus*, as Shown by Comparative Genome Analysis.

Shown are phage-related, virulence-associated genomic islands from different strains of *S. aureus* (listed to the left of the genomes), with their genes colored according to functional categories. The islands are located at specific phage integration sites in the *S. aureus* genome and are responsible for the production of toxins by these strains. The numbers at the top are distance markers along the genomic islands. Data are adapted from Novick et al.[37]

that interact with host plant proteins and elicit host cell death. One of these fungal pathogens, the agent responsible for the 19th-century Irish potato famine, expresses 196 related effectors of unexpected complexity and diversity.[46]

A fourth theme is genome reduction and pseudogene formation (see Glossary), especially in pathogens with a relatively specialized lifestyle and with restricted numbers and types of habitats, niches, and hosts.[20] This is illustrated by an unusual multidrug resistant strain of *Salmonella enterica* serovar Typhimurium that emerged in sub-Saharan Africa in the early 1990s to become the most common cause of invasive bacterial disease in some regions of that continent. Bacteremia and meningitis are common features of this disease, as they are for typhoid and paratyphoid fevers. The genome sequence of this strain reveals a large number of partially degraded and deleted genes, many of which are also degraded or deleted in the genomes of salmonella serovars Typhi and Paratyphi A.[47]

THE HUMAN MICROBIOME AND METAGENOMICS

An interactive graphic showing microbial genomics and tool development is available at NEJM.org

The role of the human indigenous microbiota in human health and disease has received a great deal of attention in the past 5 years.[7,8,13,48] Surveys of bacterial phylogenetic diversity that are based on comparative analyses of ribosomal RNA gene sequences recovered directly from clinical specimens have confirmed habitat- and individual-specific patterns in healthy persons.[49,50] Yet, core features of the indigenous microbial communities are conserved in healthy persons.[51]

A metagenomic analysis (see Glossary) of fecal samples from 124 healthy European subjects identified an average of 536,112 unique genes in each of these samples, 99.1% of which were bacterial and 0.8% of which were archaeal — and a total of 3.3 million unique genes overall, or 150 times the number of genes in the human genome.[9] Approximately 38% of an individual's fecal gene pool is shared by at least half of all other individuals. The shared gene products are predicted to mediate degradation of complex sugars, such as pectin and sorbitol, and of glycans harvested from the host diet or intestinal lining, as well as fermentation of mannose, fructose, cellulose, and sucrose (to short-chain fatty acids) and vitamin biosynthesis. These conserved genes constitute an accessory human genome that facilitates dietary energy harvest and nutrition. Alterations in the human microbiome are associated with a number of diseases in which no single organism seems to explain either the presence or the absence of disease. For these diseases (of which Crohn's disease is a leading example), the concept of community as pathogen has been proposed.[52] Elucidation of the role played by altered microbial communities in such conditions and the associated mechanisms are likely to emerge from the application of genomic approaches during the next decade.

PATHOGEN DISCOVERY AND DIAGNOSTICS

Genomic approaches have introduced a new era in the discovery and detection of microbial pathogens. The robustness, reliability, and portability of molecular sequence-based data for phylogenetic assessments and for characterization of previously unrecognized pathogens, coupled with technology developments, recommend genomic approaches for both research and routine clinical applications[53-63] (Table 1 and Fig. 3; interactive graphic, available with the full text of this article at NEJM.org). Broad-range molecular methods for microbial discovery were introduced two decades ago.[54,64] Approaches for targeting differentially abundant or phylogenetically informative molecules have now been joined by less efficient but more powerful methods for broad sequence surveys of clinical and environmental samples with the use of high-density DNA microarrays[55,65] and shotgun sequencing[56,66] (see Glossary). The advantages of DNA microarrays include the simultaneous detection of diverse sequences with widely varying relative abundance and recovery of captured sequences of interest directly from the microarray. A panviral DNA microarray with oligonucleotides designed from all known viral genera was used to characterize the novel causative agent of the severe acute respiratory syndrome (SARS)[55] and has been used to detect viruses in nasopharyngeal aspirates from children with a variety of acute respiratory syndromes.[65] The disadvantages of DNA microarrays include their insensitivity to rare microbial sequences in the presence of highly abundant host sequences (i.e., those obtained from host tissues) and their reliance on previous knowledge of microbial sequence diversity for oligonucleotide design.

High-throughput shotgun sequencing offers important new opportunities for the detection and discovery of microbial pathogens. This approach has revealed both previously known viruses (e.g., rotavirus, adenovirus, calicivirus, and astrovirus) and unknown viruses (e.g., novel types of picobirnavirus, enterovirus, TT virus, and norovirus) in fecal samples from children with unexplained acute diarrhea[66] and a novel Old World arenavirus that caused fatal disease in three recipients of organs from a single donor.[56] Dramatic advances in sequencing technology highlight the need to understand the diversity of microbial sequences in healthy subjects and to develop better methods for distinguishing rare, genuine microbial sequences from sequencing errors.

Sequence-based characterization of pathogens enables the design and development of sensitive and specific diagnostic assays and, in some cases, methods for cultivation of the pathogen. Characterization of the 16S ribosomal RNA gene from the agent of Whipple's disease, *T. whipplei*, led to a molecular diagnostic assay for this disease agent.[67]

Table 1. Examples of the Use of Microbial Genomics to Enhance the Management of Infectious Diseases.*

Application and Disease or Pathogen	Approach	Reference
Epidemiology		
Methicillin-resistant *Staphylococcus aureus*	Whole-genome sequencing	Harris et al.[33]
Tuberculosis	Whole-genome sequencing	Gardy et al.[53]
Cholera	Whole-genome sequencing	Chin et al.[12]
Influenza	Whole-genome sequencing	Smith et al.[35]
Pathogen discovery		
Bartonella henselae	Broad-range PCR	Relman et al.[54]
Severe acute respiratory syndrome coronavirus	Viral microarray	Wang et al.[55]
Novel arenavirus	Deep sequencing	Palacios et al.[56]
Diagnostic testing		
Tuberculosis	Rapid PCR	Boehme et al.[57]
HIV and HIV resistance	RT-PCR	Li et al.[58] and Panel on Antiretroviral Guidelines for Adults and Adolescents[59]
H1N1 influenza A	RT-PCR	Wang et al.[60]
Whipple's disease	PCR	Fenollar et al.[61]
Therapeutic use		
Schistosoma mansoni	High-throughput screening	Sayed et al.[62]
Preventive use		
Meningococcus B vaccine	Reverse vaccinology	Giuliani et al.[63]

* HIV denotes human immunodeficiency virus, PCR polymerase chain reaction, and RT reverse transcriptase.

Subsequent determination of its complete genome sequence[23,68] provided additional potential target sequences and the basis for a more sensitive diagnostic test.[61] It also provided insight into the metabolic defects of this bacterium, such that cell-free growth medium could be designed to include missing, needed growth factors.[69]

THERAPEUTICS AND DRUG DISCOVERY

Genome sequences provide the blueprint for essential microbial and viral components, the disruption of which can lead to growth inhibition and death. These same sequences can sometimes indicate resistance of the microbe or virus to a particular drug. Although drug susceptibility and resistance are often governed by multiple genetic components, some drug-resistance traits are encoded by single genes and can therefore be easily predicted by detecting or sequencing such genes. Examples include rifampin resistance in *M. tuberculosis*, methicillin resistance in *S. aureus*, trimethoprim–sulfamethoxazole resistance in *T. whipplei*,[70] and resistance to some antiretroviral drugs in HIV. Genome sequences have also provided new targets and leads for the development of new antimicrobials.

The standard of care for the management of HIV infection now includes targeted drug selection with the use of a profile for HIV-drug susceptibility that is derived from the sequence of the infecting HIV species.[59] Testing for genotypic resistance is recommended for patients with HIV infection when they enter care and when there is a suboptimal reduction in viral load while they are receiving first- or second-line antiretroviral regimens. Clinically important resistance mutations occur in HIV genes encoding the reverse-transcriptase, protease, envelope, and integrase proteins. Interpretation of these mutant genotypes is facilitated by several databases, including those maintained by the International Antiviral Society–USA[71] and a research group at Stanford University.[72] Genotypic analysis is cheaper and faster than phenotypic analysis for HIV-drug resistance and is often more sensitive for detecting resistant strains within mixtures of drug-susceptible viruses.[73]

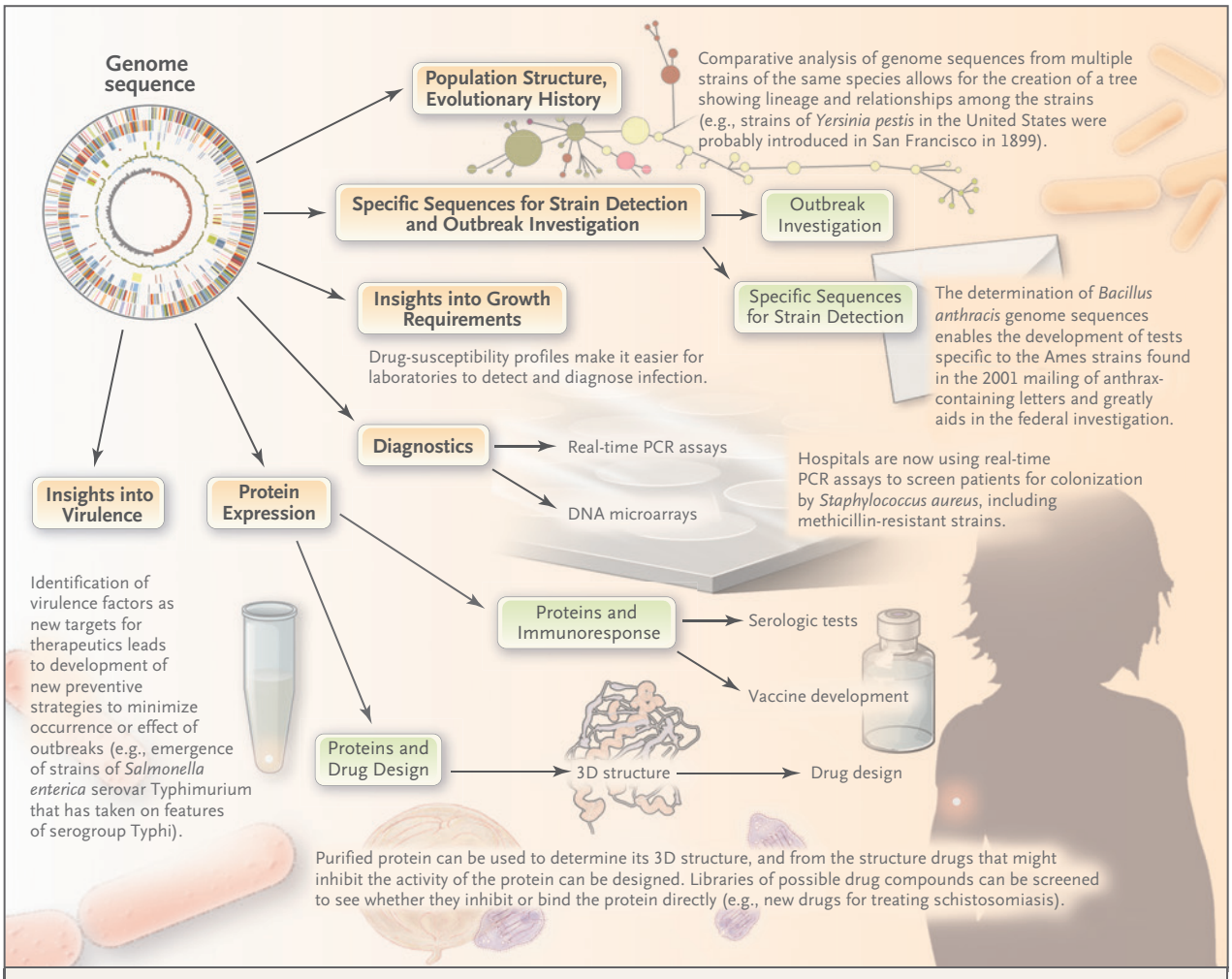

Figure 3. Microbial Genomics and Tool Development.

A genome sequence facilitates the development of a variety of tools and approaches for understanding, manipulating, and mitigating the overall effect of a microbe. The sequence provides insight into the population structure and evolutionary history of a microbe for epidemiologic investigation, information with which to develop new diagnostic tests and cultivation methods, new targets of drug development, and antigens for vaccine development.

However, commercial assays of both types do not routinely detect resistant viruses when they are less than 10 to 20% of the overall circulating virus population. With newer sequencing techniques, less abundant strains are easier to detect and characterize. Although the clinical relevance of rare resistant variants is not fully understood, the pretreatment detection of such variants has been shown to have clinical value.[58] Traditional phenotypic testing (measuring the ability of the virus to replicate in the presence of the antiviral drug) is still recommended for patients in whom viruses are suspected of having complex drug-resistance mutation patterns.

Schistosomiasis is a chronic and debilitating disease that affects approximately 210 million people in 76 countries around the globe and results in some 280,000 deaths per year in sub-Saharan Africa alone. Praziquantel has been the drug of choice for the treatment of schistosomiasis but is in danger of losing efficacy because of parasite resistance. *Schistosoma mansoni* is one of three helminths for which there is now a draft genome sequence available to the public.[74] Besides enabling the study of gene and protein expression,[75] the nuclear genome of *S. mansoni* and its approximately 11,800 putative genes point to critical compounds and processes on which the worm depends to survive in its host. These compounds and processes reveal potential new drug

targets, one of which is a redox enzyme, thioredoxin–glutathione reductase.[74] Quantitative high-throughput screening of small-molecule libraries for compounds with activity against the *S. mansoni* thioredoxin–glutathione reductase has already identified some candidate drugs.[62]

Microbes produce a wealth of druglike molecules, the vast majority of which remain uncharacterized.[76,77] Because many of these molecules are not expressed under typical laboratory conditions, they often escape detection when laboratory culture filtrates are screened for druglike properties. Some of these molecules can now be identified by recognizing the relevant genes in the parent organism's genome with the use of computational tools and detecting the molecules with mass spectroscopy techniques.[78,79] Derivative compounds can be designed and tested.

VACCINES

In the same way that genome sequences reveal drug-resistance profiles, vulnerabilities, and synthetic capabilities of microbes and viruses, these sequences also provide clues about antigenic repertoire. This information can be exploited for vaccine design and other immunoprophylactic interventions. Genome-based antigen discovery has also been undertaken for more complex pathogens. One approach, known as reverse vaccinology, involves cloning and expressing all proteins that are predicted (from the organism's complete genome) to be secreted or surface-associated, starting with the complete genome sequence (Fig. 3).[80] After immunizing mice with each of the proteins, each of the corresponding antiserum samples is tested for its ability to neutralize or kill the original target organism. On the basis of this approach, a small group of proteins from group B meningococcus,[81] a pathogen that has so far eluded vaccine development, has shown promise as a candidate multivalent subunit vaccine. A similar approach has been taken with group B streptococcus[82] and extraintestinal pathogenic *E. coli*.[83] Protective antigens that are discovered through these sorts of methods may have been previously ignored because they are not immunogenic during natural infections.

FUTURE DIRECTIONS

Without question, the techniques for microbial and viral genome sequencing are becoming increasingly rapid and less expensive. Genome sequencing of a microbe or virus will soon be easier than characterization of its growth-based behavior in the laboratory. In the next 3 to 5 years, direct shotgun sequencing of the DNA and RNA in a clinical sample may become a routine matter. What is less clear is how clinically relevant information will be most effectively extracted from the ensuing massive amounts of data. In the near term, genomic and metagenomic analyses of microbes are most likely to be useful in areas such as the cataloguing and understanding of microbial and viral diversity in the human body, the identification of molecular determinants of virulence and symbiosis, and real-time tracking of particular strains of pathogens. Such analyses will also provide a deeper understanding of how pathogens spread and cause disease and will identify new targets for therapies and antigens for vaccines. Thoughtfully designed clinical and epidemiologic studies will be required to see the full realization of these benefits.

Disclosure forms provided by the author are available with the full text of this article at NEJM.org.

REFERENCES

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496-512.
2. National Center for Biotechnology Information. Genome. (http://www.ncbi.nlm.nih.gov/genome?db=genome.)
3. Squires B, Macken C, Garcia-Sastre A, et al. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res 2008;36:D497-D503.
4. Los Alamos National Laboratory. Los Alamos HIV sequence database, 2011. (http://www.hiv.lanl.gov.)
5. Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. Science 2006;312:1355-9.
6. Marcy Y, Ouverney C, Bik EM, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A 2007;104:11889-94.
7. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature 2007;449:804-10.
8. Peterson J, Garges S, Giovanni M, et al. The NIH Human Microbiome Project. Genome Res 2009;19:2317-23.
9. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59-65.
10. Tettelin H, Feldblyum T. Bacterial genome sequencing. Methods Mol Biol 2009;551:231-47.
11. Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol 2010;13:625-31.
12. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med 2011;364:33-42.
13. Dethlefsen L, McFall-Ngai M, Relman DA. An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 2007;449:811-8.

14. Welch RA, Burland V, Plunkett G III, et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A 2002;99:17020-4.
15. Rasko DA, Rosovitz MJ, Myers GS, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 2008;190:6881-93.
16. Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A 2005;102:13950-5. [Erratum, Proc Natl Acad Sci U S A 2005;102:16530.]
17. Doolittle RF. Biodiversity: microbial genomes multiply. Nature 2002;416:697-700.
18. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304-51. [Erratum, Science 2001;292:1838.]
19. Fraser-Liggett CM. Insights on biology and evolution from microbial genome sequencing. Genome Res 2005;15:1603-10.
20. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature 2007;449:835-42.
21. Cummings CA, Brinig MM, Lepp PW, van de Pas S, Relman DA. Bordetella species are distinguished by patterns of substantial gene loss and host adaptation. J Bacteriol 2004;186:1484-92.
22. Preston A, Parkhill J, Maskell DJ. The bordetellae: lessons from genomics. Nat Rev Microbiol 2004;2:379-90.
23. Bentley SD, Maiwald M, Murphy LD, et al. Sequencing and analysis of the genome of the Whipple's disease bacterium Tropheryma whipplei. Lancet 2003;361:637-44.
24. Parkhill J, Wren BW, Thomson NR, et al. Genome sequence of Yersinia pestis, the causative agent of plague. Nature 2001;413:523-7.
25. Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. Nature 2001;409:1007-11.
26. Gogarten JP, Doolittle WF, Lawrence JG. Prokaryotic evolution in light of gene transfer. Mol Biol Evol 2002;19:2226-38.
27. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 2009;33:376-93.
28. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science 2010;327:167-70.
29. Baker S, Hanage WP, Holt KE. Navigating the future of bacterial molecular epidemiology. Curr Opin Microbiol 2010;13:640-5.
30. Morelli G, Song Y, Mazzoni CJ, et al. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 2010;42:1140-3.
31. Falush D, Wirth T, Linz B, et al. Traces of human migrations in Helicobacter pylori populations. Science 2003;299:1582-5.
32. Moodley Y, Linz B, Yamaoka Y, et al. The peopling of the Pacific from a bacterial perspective. Science 2009;323:527-30.
33. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 2010;327:469-74.
34. Russell CA, Jones TC, Barr IG, et al. The global circulation of seasonal influenza A (H3N2) viruses. Science 2008;320:340-6.
35. Smith GJ, Vijaykrishna D, Bahl J, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 2009;459:1122-5.
36. McHardy AC, Adams B. The role of genomics in tracking the evolution of influenza A virus. PLoS Pathog 2009;5(10):e1000566.
37. Novick RP, Christie GE, Penades JR. The phage-related chromosomal islands of Gram-positive bacteria. Nat Rev Microbiol 2010;8:541-51.
38. Scheutz F, Møller Nielsen E, Frimodt-Møller J, et al. Characteristics of the enteroaggregative Shiga toxin/verotoxin-producing Escherichia coli O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011. Euro Surveill 2011;16:pii=19889.
39. Chen PE, Cook C, Stewart AC, et al. Genomic characterization of the Yersinia genus. Genome Biol 2010;11:R1.
40. Levin BR. The evolution and maintenance of virulence in microparasites. Emerg Infect Dis 1996;2:93-102.
41. Le Gall T, Clermont O, Gouriou S, et al. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. Mol Biol Evol 2007;24:2373-84.
42. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. Pathogenicity-associated islands in extraintestinal pathogenic Escherichia coli are fitness elements involved in intestinal colonization. J Bacteriol 2010;192:4885-93.
43. Molmeret M, Alli OA, Zink S, Flieger A, Cianciotto NP, Kwaik YA. icmT Is essential for pore formation-mediated egress of Legionella pneumophila from mammalian and protozoan cells. Infect Immun 2002;70:69-78.
44. Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 2009;325:1128-31.
45. Raffaele S, Farrer RA, Cano LM, et al. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 2010;330:1540-3.
46. Haas BJ, Kamoun S, Zody MC, et al. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 2009;461:393-8.
47. Kingsley RA, Msefula CL, Thomson NR, et al. Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome Res 2009;19:2279-87.
48. Relman DA, Falkow S. The meaning and impact of the human genome sequence for microbiology. Trends Microbiol 2001;9:206-8.
49. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. Science 2005;308:1635-8.
50. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science 2009;326:1694-7.
51. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature 2009;457:480-4.
52. Peterson DA, Frank DN, Pace NR, Gordon JI. Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases. Cell Host Microbe 2008;3:417-27.
53. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 2011;364:730-9. [Erratum, N Engl J Med 2011;364:2174.]
54. Relman DA, Loutit JS, Schmidt TM, Falkow S, Tompkins LS. The agent of bacillary angiomatosis — an approach to the identification of uncultured pathogens. N Engl J Med 1990;323:1573-80.
55. Wang D, Urisman A, Liu YT, et al. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 2003;1(2):E2.
56. Palacios G, Druce J, Du L, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 2008;358:991-8. [Erratum, N Engl J Med 2008;358:1204.]
57. Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 2010;363:1005-15.
58. Li JZ, Paredes R, Ribaudo HJ, et al. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. JAMA 2011;305:1327-35.
59. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Washington, DC: Department of Health and Human Services, 2009:1-161.
60. Wang R, Sheng ZM, Taubenberger JK. Detection of novel (swine origin) H1N1 influenza A virus by quantitative real-time reverse transcription-PCR. J Clin Microbiol 2009;47:2675-7. [Erratum, J Clin Microbiol 2009;47:3403.]
61. Fenollar F, Laouira S, Lepidi H, Rolain JM, Raoult D. Value of Tropheryma whipplei quantitative polymerase chain reaction assay for the diagnosis of Whipple disease: usefulness of saliva and stool specimens for first-line screening. Clin Infect Dis 2008;47:659-67.

62. Sayed AA, Simeonov A, Thomas CJ, Inglese J, Austin CP, Williams DL. Identification of oxadiazoles as new drug leads for the control of schistosomiasis. Nat Med 2008;14:407-12.
63. Giuliani MM, Adu-Bobie J, Comanducci M, et al. A universal vaccine for serogroup B meningococcus. Proc Natl Acad Sci U S A 2006;103:10834-9.
64. Lipkin WI, Travis GH, Carbone KM, Wilson MC. Isolation and characterization of Borna disease agent cDNA clones. Proc Natl Acad Sci U S A 1990;87:4184-8.
65. Chiu CY, Urisman A, Greenhow TL, et al. Utility of DNA microarrays for detection of viruses in acute respiratory tract infections in children. J Pediatr 2008;153:76-83.
66. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D. Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 2008;4(2):e1000011.
67. Relman DA, Schmidt TM, MacDermott RP, Falkow S. Identification of the uncultured bacillus of Whipple's disease. N Engl J Med 1992;327:293-301.
68. Raoult D, Ogata H, Audic S, et al. Tropheryma whipplei Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res 2003;13:1800-9.
69. Renesto P, Crapoulet N, Ogata H, et al. Genome-based design of a cell-free culture medium for Tropheryma whipplei. Lancet 2003;362:447-9.
70. Bakkali N, Fenollar F, Biswas S, Rolain JM, Raoult D. Acquired resistance to trimethoprim-sulfamethoxazole during Whipple disease and expression of the causative target gene. J Infect Dis 2008;198:101-8.
71. International Antiviral Society–USA. HIV drug resistance mutations. San Francisco: IAS–USA, 2011. (http://www.iasusa.org/resistance_mutations/index.html.)
72. Stanford University HIV Drug Resistance Database home page. (http://hivdb.stanford.edu.)
73. Llibre JM, Schapiro JM, Clotet B. Clinical implications of genotypic resistance to the newer antiretroviral drugs in HIV-1-infected patients with virological failure. Clin Infect Dis 2010;50:872-81.
74. Berriman M, Haas BJ, LoVerde PT, et al. The genome of the blood fluke Schistosoma mansoni. Nature 2009;460:352-8.
75. Han ZG, Brindley PJ, Wang SY, Chen Z. Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. Annu Rev Genomics Hum Genet 2009;10:211-40.
76. Fischbach MA. Antibiotics from microbes: converging to kill. Curr Opin Microbiol 2009;12:520-7.
77. Zimmermann M, Fischbach MA. A family of pyrazinone natural products from a conserved nonribosomal peptide synthetase in Staphylococcus aureus. Chem Biol 2010;17:925-30.
78. Wieland Brown LC, Acker MG, Clardy J, Walsh CT, Fischbach MA. Thirteen post-translational modifications convert a 14-residue peptide into the antibiotic thiocillin. Proc Natl Acad Sci U S A 2009;106:2549-53.
79. Van Voorhis WC, Hol WG, Myler PJ, Stewart LJ. The role of medical structural genomics in discovering new drugs for infectious diseases. PLoS Comput Biol 2009;5(10):e1000530.
80. Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. Immunity 2010;33:530-41.
81. Pizza M, Scarlato V, Masignani V, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 2000;287:1816-20.
82. Tettelin H, Medini D, Donati C, Masignani V. Towards a universal group B Streptococcus vaccine using multistrain genome analysis. Expert Rev Vaccines 2006;5:687-94.
83. Moriel DG, Bertoldi I, Spagnuolo A, et al. Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic Escherichia coli. Proc Natl Acad Sci U S A 2010;107:9072-7.