Chapter 25

Genomic and Postgenomic Research

The word *genomics* was first coined by T. Roderick from the Jackson Laboratories in 1986 as the name for the new field of science focused on the analysis and comparison of complete genome sequences of organisms and related high-throughput technologies.

A sequence provides the most fundamental information about an organism, and the genes and the regulatory sites encoded in the sequence will reveal the complete profile of the organism and information about its evolution *(1)*.

The development of genomics should be considered one of the most dramatic results of the advances in biomedical sciences in the 20th century.

25.1 Computational Methods for Genome Analysis

Two basic computational methods are used for genome analysis: gene finding and whole genome comparison *(2)*.

Gene Finding. Using a computational method that can scan the genome and analyze the statistical features of the sequence is a fast and remarkably accurate way to find the genes in the genome of prokaryotic organisms (bacteria, archaea, viruses) compared with the still difficult problem of finding genes in higher eukaryotes. By using modern bioinformatics software, finding the genes in a bacterial genome will result in a highly accurate, rich set of annotations that provide the basis for further research into the functions of those genes.

The absence of introns—those portions of the DNA that lie between two exons and are transcribed into a RNA but will not appear in that RNA after maturation and therefore are not expressed (as proteins) in the protein synthesis—will remove one of the major barriers to computational analysis of the genome sequence, allowing gene finding to identify more than 99% of the genes of most genomes without any human intervention. Next, these gene predictions can be further refined by searching for nearby regulatory sites such as the ribosome-binding sites, as well as by aligning protein sequences to other species. These steps can be automated using freely available software and databases *(2)*.

Gene finding in single-cell eukaryotes is of intermediate difficulty, with some organisms, such as *Trypanosoma brucei*, having so few introns that a bacterial gene finder is sufficient to find their genes. Other eukaryote organisms (e.g., *Plasmodium falciparum*) have numerous introns and would require the use of special-purpose gene finder, such as GlimmerM *(3, 4)*.

Whole Genome Comparison. This computational method refers to the problem of aligning the entire deoxyribonucleic acid (DNA) sequence of one organism to that of another, with the goal of detecting all similarities as well as rearrangements, insertions, deletions, and polymorphisms *(2)*. With the increasing availability of complete genome sequences from multiple, closely related species, such comparisons are providing a powerful tool for genomic analysis. Using *suffix trees*—data structures that contains all of the subsequences from a particular sequence and can be built and searched in linear time—this computational task can be accomplished in minimal time and space. Because the suffix tree algorithm is both time and space efficient, it is able to align large eukaryotic chromosomes with only slightly greater requirements than those for bacterial genomes *(2)*.

Bacterial Genome Annotation. The major goal of the bacterial genome annotation is to identify the functions of all genes in a genome as accurately and consistently as possible by using initially automated annotation methods for preliminary assignment of functions to genes, followed by a second stage of manual curation by teams of scientists.

25.2 Genomes of Pathogenic Enterobacteria

The family Enterobacteriaceae encompasses a diverse group of bacteria including many of the most important human pathogens (*Salmonella, Yersinia, Klebsiella, Shigella*), as well as one of the most enduring laboratory research organisms, the nonpathogenic *Escherichia coli* K12. Many of

V. St. Georgiev, *National Institute of Allergy and Infectious Diseases, NIH: Impact on Global Health*, vol. 2,
DOI 10.1007/978-1-60327-297-1_25, © Humana Press, a part of Springer Science+Business Media, LLC 2009

these pathogens have been subject to genome sequencing or are under study. Genome comparisons among these organisms have revealed the presence of a core set of genes and functions along a generally collinear genomic backbone. However, there are also many regions and points of difference, such as large insertions and deletions (including pathogenicity islands), integrated bacteriophages, small insertions and deletions, point mutations, and chromosomal rearrangements *(5)*.

25.2.1 Escherichia coli K12

The first genome sequence of *Escherichia coli* K12 (reference strain MG1655) was completed and published in 1997 *(6)*. Later, the genome sequence of two other genotypes of *E. coli*, the enterohemorrhagic *E. coli* O157:H7 (EHEC; strains EDL933 and RIMD 0509952-Sakai) *(7, 8)* and the uropathogenic *E. coli* (UPEC; strain CFT073) *(9)*, were sequenced and the information published. Currently, it is accepted that shigellae are part of the *E. coli* species complex, and information on the genome of *Shigella flexneri* strain 2a has been published *(10)*.

A comparison of all three pathogenic *E. coli* with the archetypal nonpathogenic *E. coli* K12 revealed that the genomes were essentially collinear, displaying both conservation in sequence and gene order *(5)*. The genes that were predicted to be encoded within the conserved sequence displayed more than 95% sequence identity and have been termed the *core genes*. Similar observations were made for the *Shigella flexneri* genome, which also shares 3.9 Mb of common sequence with *E. coli (10)*.

A comparison of the three *E. coli* genomes revealed that genes shared by all genomes amounted to 2,996 *(9)* from a total of 4,288, and about 5,400 and 5,500 predicted protein-coding sequences for *E. coli* K12, EHEC, and UPEC, respectively *(5)*. The region encoding these core genes is known as the *backbone sequence.*

It was also apparent from these comparisons that interdispersed throughout this backbone sequence were large regions unique to the different genotypes. Moreover, several studies had shown that some of these unique loci were present in clinical disease–causing isolates but were apparently absent from their comparatively benign relatives *(11)*. One such well-characterized region is the *locus of enterocyte effacement (LEE)* in the enteropathogenic *E. coli* (EPEC). Thus, an EPEC infection results in effacement of the intestinal microvilli and the intimate adherence of bacterial cells to enterocytes. Furthermore, EPEC also subverts the structural integrity of the cell and forces the polymerization of actin, which accumulates below the adhered EPEC cells, forming cup-like pedestals *(12)*. This is called an *attachment and effacing (AE)* lesion. Subsequently, LEE was found in all bacteria known to be able to elicit an AE lesion *(5)*.

25.2.1.1 Pathogenicity Islands

The presence of many regions in the backbone sequence similar to LEE have been characterized in both Gram-negative and Gram-positive bacteria *(13)*. This led to the concept of *pathogenicity islands (PAIs)* and the formulation of a definition to describe their features *(5)*.

Typically, PAIs are inserted adjacent to stable RNA genes and have an atypical G+C content. In addition to virulence-related functions, the pathogenicity islands often carry genes encoding transposase or integrase-like proteins and are unstable and self-mobilizable *(13, 14)*. It was also noted that PAIs possess a high proportion of gene fragments or disrupted genes when compared with the backbone regions *(15)*.

It is generally accepted that the pathogenic *E. coli* genotypes have evolved from a much smaller nonpathogenic relative by the acquisition of foreign DNA. This laterally acquired DNA has been attributed with conferring on the different genotypes the ability to colonize alternative niches in the host and the ability to cause a range of different disease outcomes *(5)*.

Although sharing some of the features of PAIs and considered to be parts of the PAIs, some genomic loci are unlikely to impinge on pathogenicity. To take account of this, the concept of PAIs has been extended to include islands or strain-specific loops, which represent discrete genetic loci that are lineage-specific but are as yet not known to be involved in virulence *(7, 8)*.

25.2.2 Salmonella Pathogenicity Islands

Currently, there are more than 2,300 *Salmonella* serovars in two species, *S. enterica* and *S. bongori*. All salmonellae are closely related, sharing a median DNA identity for the reciprocal best match of between 85% and 95% *(16, 17)*. Despite their homogeneity, there are still significant differences in the pathogenesis and host range of the different *Salmonella* serovars. Thus, whereas *S. enterica* subspecies *enterica* serovar Typhi (*S. typhi*) is only pathogenic to humans causing severe typhoid fever, *S. typhimurium* causes gastroenteritis in humans but also a systemic infection in mice and has a broad host range *(16)*.

Like *E. coli*, the salmonellae are also known to possess PAIs, known as *Salmonella* pathogenicity islands (SPIs). It is thought that SPIs have been acquired laterally. For example, the gene products encoded by SPI-1 *(18, 19)* and SPI-2

(20,21) have been shown to play important roles in the different stages of the infection process. Both of these islands possess type III secretion systems and their associated secreted protein effectors. SPI-1 is known to confer on all salmonellae the ability to invade epithelial cells. SPI-2 is important in various aspects of the systemic infection, allowing *Salmonella* to spread from the intestinal tissue into the blood and eventually to infect, and survive within, the macrophages of the liver and spleen *(22)*.

SPI-3, like LEE and PAI-1 of UPEC, is inserted alongside the *selC* tRNA gene and carries the gene *mgtC*, which is required for the intramacrophage survival and growth in the low-magnesium environment thought to be encountered in the phagosome *(23)*.

Other *Salmonella* SPIs encode type III–secreted effector proteins, chaperone-usher fimbrial operons, Vi antigen biosynthetic gene, a type IVB pilus operon, and many other determinants associated with the salmonellae enteropathogenicity *(15)*.

25.2.3 Yersinia High-Pathogenicity Island

Although the mobile nature of PAIs is frequently discussed in the literature, there is little direct experimental evidence to support these observations. One possible explanation for this may be that on integration, the mobility genes of the PAIs subsequently become degraded, thereby fixing their position *(5)*. Certainly, there is evidence to support this hypothesis, as many proposed PAIs carry integrase or transposase pseudogenes or remnants. One excellent example of this is the high-pathogenicity island (HPI) first characterized in *Yersinia (24)*.

The *Yersinia* HPIs can be split into two lineages based on the integrity of the phage integrase gene (*int*) carried in the island: (i) *Y. enterocolitica* biotype 1B and (ii) *Y. pestis* and *Y. pseudotuberculosis*. The *Y. enterocolitica* HPI *int* gene carries a point mutation, whereas the analogous gene is intact in the *Y. pestis* and *Y. pseudotuberculosis* HPIs.

The *Yersinia* HPI is a 35- to 43-kb island that possesses genes for the production and uptake of the siderophore yersiniabactin, as well as genes, such as *int*, thought to be involved in the mobility of the island. HPI-like elements are widely distributed in enterobacteria, including *E. coli*, *Klebsiella*, *Enterobacter*, and *Citrobacter* spp., and like many prophages, these HPIs are found adjacent to *asn*-tRNA genes *(8)*. tRNA genes are common sites for bacteriophage integration into the genome *(25)*.

Integration at these sites typically involves site-specific recombination between short stretches of identical DNA located on the phage (*attP*) and at the integration site on the bacterial genomes (*attB*). The tRNA genes represent common sites for the integration of many other PAIs and bacteriophages, with the *secC* tRNA locus being the most heavily used integration site in the enterics *(9)*.

25.2.4 Bacteriophages (Prophages)

Integrated bacteriophages, also known as prophages, are also commonly found in bacterial genomes *(5)*. For example, in the S loops of the *E. coli* O157:H7 strain EDL933 (EHEC) unique regions, nearly 50% were phage related. In addition to the 18 prophage sequences detected in the genome of EHEC strain Sakai *(8)*, the genomes of *E. coli* K12, UPEC, and *S. flexneri* have all been shown to carry multiple prophage or prophage-like elements *(6, 7, 9, 10)*. Moreover, comparison of the genome sequences of EHEC O157:H7 strain EDL933 and strain Sakai revealed marked variations in the complement and integration sites of the prophages, as did internal regions within highly related phages *(8, 26)*.

In addition to genes essential for their own replication, phages often carry genes that, for example, prevent superinfection by other bacteriophages, such as *old* and *tin (27, 28)*. However, other genes carried in prophages appear to be of nonphage origin and can encode determinants that enhance the virulence of the bacterial host by a process known as *lysogenic conversion (29)*.

In addition to the presence of the LEE PAI and the ability to elicit AE lesion, another defining characteristics of the enterohemorrhagic *E. coli* (EHEC) is the production of Shiga toxins (Stx). The Shiga toxins represent a family of potent cytotoxins that, on entry into the eukaryotic cell, will act as glycosylases by cleaving the 28S ribosomal RNA (rRNA) thereby inactivating the ribosome and consequently preventing the protein synthesis *(30)*.

Other enteric pathogens such as *S. typhi*, *S. typhimurium*, and *Y. pestis* are also known to possess significant numbers of prophages *(15, 16, 31)*. Thus, the principal virulence determinants of the salmonellae are the type III secretion systems, carried by SPI-1 and SPI-2, and their associated protein effectors *(32, 33)*. A significant number of these type III secreted effector proteins are present in the genomes of prophages and have a dramatic influence on the ability of their bacterial hosts to cause disease *(5)*.

25.2.5 Other Characteristic Features of the Enterobacterial Genomes

Small Insertions and Deletions. Even though the large PAIs play a major role in defining the phenotypes of different strains of the enteric bacteria, there are many other differences resulting from small insertions and deletions, which must be taken into account when considering the overall genomic picture of Enterobacteriaceae *(5)*.

Thus, the comparisons between *E. coli* K12 and *E. coli* O157:H7 and between *S. typhi* and *S. typhimurium* have indicated the existence of many small differences that exist aside from the large pathogenicity islands. For example, the number of separate insertion and deletion events has shown that there are 145 events of 10 genes or fewer compared with 12 events of 20 genes or more for the *S. typhi* and *S. typhimurium* comparison. Furthermore, comparison between *S. typhi* and *E. coli* revealed 504 events of 10 genes or fewer compared with just 25 events of 20 genes or more. Even taking into account that the larger islands contain many more genes per insertion or deletion event, it becomes clear that nearly equivalent numbers of species-specific genes are attributable to insertion or deletion events involving 10 genes or fewer as are due to events involving 20 genes or more. These data should lend credence to the assertion that the acquisition and exchange of *small islands* is important in defining the overall phenotype of the organism *(5)*. In the majority of cases studied to date, there is no evidence to suggest the presence of genes that may allow these small islands to be self-mobile. It is far more likely that small islands of this type are exchanged between members of a species and constitute part of the species gene pool. Once acquired by one member of the species, they can be easily exchanged by generalized transduction mechanisms, followed by homologous recombination between the near identical flanking genes to allow integration into the chromosome *(5)*.

This sort of mechanism of genetic exchange would also make possible nonorthologous gene replacement, involving the exchange of related genes at identical regions in the backbone. A specific example to illustrate such a possibility is the observed *capsular switching* of *Neisseria meningitides (34)* and *Streptococcus pneumoniae (35, 36)* for which different sets of genes responsible for the biosynthesis of different capsular polysaccharides are found at identical regions in the chromosome and flanked by conserved genes. The implied mechanism for capsular switching involves replacement of the polysaccharide-specific gene sites by homologous recombination between the chromosome and exogenous DNA in the flanking genes *(5)*.

Point Mutations and Pseudogenes. One of the most surprising observations to come from enterobacterial genome research has been the discovery of a large number of *pseudogenes.* The pseudogenes appeared to be untranslatable due to the presence of stop codons, frameshifts, internal deletions, or insertion sequence (IS) element insertions. The presence of pseudogenes seems to run contrary to the general assumption that the bacterial genome is a highly "streamlined" system that does not carry "junk DNA" *(5)*.

For example, *Salmonella typhi*, the etiologic agent of typhoid fever, is host restricted and appears only capable of infecting a human host, whereas *S. typhimurium*, which causes a milder disease in humans, has a much broader host range. Upon analysis, the genome of *S. typhi* contained more than 200 pseudogenes *(15)*, whereas it was predicted that the number of pseudogenes in the genome of *S. typhimurium* would be around 39 *(16)*. From this observation, it becomes clear that the pseudogenes in *S. typhi* were not randomly spread throughout its genome—in fact, they were overrepresented in genes that were unique to *S. typhi* when compared with *E. coli*, and many of the pseudogenes in *S. typhi* have intact counterparts in *S. typhimurium* that have been shown to be involved in aspects of virulence and host interaction. Given this distribution of pseudogenes, it has been suggested that the host specificity of *S. typhi* may be the result of the loss of its ability to interact with a broader range of hosts caused by functional inactivation of the necessary genes *(15)*. In contrast with other microorganisms containing multiple pseudogenes, such as *Mycobacterium leprae (37)*, most of the pseudogenes in *S. typhi* were caused by a single mutation, suggesting that they have been inactivated relatively recently.

Taken together, these observations suggest an evolutionary scenario in which the recent ancestor of *S. typhi* had changed its niche in a human host, evolving from an ancestor (similar to *S. typhimurium*) limited to localized infection and invasion around the gut epithelium into one capable of invading the deeper tissues of the human hosts *(5)*.

A similar evolutionary scenario has been suggested for another recently evolved enteric pathogen, *Yersinia pestis.* This bacterium has also recently changed from a gut bacterium (*Y. pseudotuberculosis*), transmitted via the fecal-oral route, to an organism capable of using a flea vector for systemic infection *(38, 39)*. Again, this change in niche was accompanied by pseudogene formation, and genes involved in virulence and host interaction are overrepresented in the set of genes inactivated *(31)*.

Yet another example of such an evolutional scenario is *Shigella flexneri* 2a, a member of the species *E. coli* (which is predicted to have more than 250 pseudogenes), and is again restricted to the human body *(10)*.

All of these organisms demonstrate that the enterobacterial evolution has been a process that has involved both gene loss and gene gain, and that the remnants of the genes lost in the evolutionary process can be readily detected *(5)*.

25.3 Bacterial Proteomes as Complements of Genomes

The focus in the postgenomic era is on functional genomics, in which proteomics plays an essential role. The living cell is a dynamic and complex system that cannot be predicted from the genome sequence. Whereas genomes will disclose important information on the biological importance of the organism, it is still static and will not reveal information on the expression of a particular gene or of posttranslational modifications or on how a protein is regulated in a specific biological situation *(40)*.

Thus, whereas the complete genome sequence provides the basis for experimental identification of expressed proteins at the cellular level, very little has been accomplished to identify all expressed and potentially modified proteins.

Direct investigation of the total content of proteins in a cell is the task of proteomics. Proteomics is defined as the complete set of posttranslationally modified and processed proteins in a well-defined biological environment under specific circumstances, such as growth conditions and time of investigation *(40, 41)*.

Proteomics can be studied by following two separate steps: separation of the proteins in a sample, followed by identification of the proteins. The common methodology used for separating proteins is two-dimensional polyacrylamide gel electrophoresis (2D PAGE). The principal method for large-scale identification is mass spectroscopy (MS), but other identification methods, such as *N*-terminal sequencing, immunoblotting, overexpression, spot colocalization, and gene knockouts, can also be used.

25.3.1 Two-Dimensional Polyacrylamide Gel Electrophoresis

Because of its high-resolution power, 2D PAGE is currently the best methodology to achieve global visualization of the proteins of a microorganism. In the first dimension, isoelectric focusing is carried out to separate the proteins in a pH gradient according to their isoelectric point (p*I*). In the second dimension, the proteins are separated according to their molecular weight by SDS-PAGE (sodium dodecyl sulfate–PAGE). The resulting gel image presents itself as a pattern of spots in which p*I* and the relative molecular weight (M_r) can be recognized as in a coordinate system *(40)*. A critical step during the 2D PAGE procedure is the sample preparation, as there is no single method that can be universally applied because different reagents are superior with respect to different samples. To this end, chaotropes such as urea, which act by changing the parameters of the solvent, are used in most 2D PAGE procedures.

Major problems to overcome in 2D PAGE sample preparation arise because of limited entry into the gel of high-molecular-weight proteins and the presence of highly hydrophobic and/or basic proteins *(42, 43)*.

For protein separation, the protein mixture is loaded onto an acrylamide gel strip in which a pH gradient is established. When a high voltage is applied over the strip, the proteins will focus at the pH at which they carry zero net charge. The pH gradient is established during the focusing using either carrier ampholytes in a slab gel *(44)* or a precast polyacrylamide gel with an immobilized pH gradient (IPG) *(45)*. The latter method is advantageous because of improved reproducibility. Samples can be applied to IPG dry strips preferably by rehydration. Rehydration of dried IPGs under application of a low voltage (10 to 50 V) has significantly improved the recovery especially of high-molecular-weight proteins.

25.3.2 Mass Spectrometry

Mass spectrometry is the method of choice for identifying proteins in proteomics. The proteins are converted into gas phase ions that can be measured with an accuracy better than 50 ppm *(40)*. Two widely used techniques for ionization are matrix-assisted laser desorption ionization (MALDI) *(46)* and electrospray ionization *(47)*. MALDI is usually coupled with a TOF (time of flight) device for measuring the masses. The ionized peptides are then accelerated by the application of accelerated field and the TOF until they reach a detector to calculate their mass/charge ratio *(40)*.

In electrospray ionization, the peptides are sprayed into the spectrometer *(47)*. Ionization is achieved when the charged droplets evaporate. An alternative procedure for measuring masses is the ion trap *(48)*, which selects ions with certain mass/charge ratios by keeping them in sinusoidal motion between two electrodes.

25.4 NIAID Research Programs in Genomic Research

In 1995, the first microbe sequencing project, *Haemophilus influenzae* (a bacterium causing upper respiratory infection), was completed with a speed that stunned scientists (http://www3.niaid.nih.gov/research/topics/pathogen/Introduction.htm). Encouraged by the success of that initial effort, researchers have continued to sequence an astonishing array of other medically important microorganisms. To this end, NIAID has made significant investments in large-scale sequencing projects, including projects to sequence the complete genomes of many pathogens, such as the bacteria that cause tuberculosis, gonorrhea, chlamydia, and cholera, as well as organisms that are considered agents of bioterrorism. In addition, NIAID is collaborating with other funding agencies to sequence larger genomes of protozoan pathogens such as the organism causing malaria.

The availability of microbial and human DNA sequences opens up new opportunities and allows scientists to perform functional analyses of genes and proteins in whole genomes and cells, as well as the host's immune response and an individual's genetic susceptibility to pathogens. When scientists identify microbial genes that play a role in disease, drugs can be designed to block the activities controlled by those genes. Because most genes contain the instructions for making proteins, drugs can be designed to inhibit specific proteins or to use those proteins as candidates for vaccine testing. Genetic variations can also be used to study the spread of a virulent or drug-resistant form of a pathogen.

25.4.1 Genomic Initiatives

NIAID has launched initiatives to provide comprehensive genomic, proteomic, and bioinformatic resources. These resources, listed below, are available to scientists conducting basic and applied research on a broad array of pathogenic microorganisms (http://www3.niaid.nih.gov/research/topics/pathogen/initiatives.htm):

- *NIAID's Microbial Sequencing Centers (NSCs).* The NIAID's Microbial Sequencing Centers are state-of-the-art high-throughput DNA sequencing centers that can sequence genomes of microbes and invertebrate vectors of infectious diseases. Genomes that can be sequenced include microorganisms considered agents of bioterrorism and those responsible for emerging and re-emerging infectious diseases.
- *NIAID's Pathogen Functional Genomics Resource Center (PFGRC).* NIAID's Pathogen Functional Genomics Resource Center is a centralized facility that provides scientists with the resources and reagents necessary to conduct functional genomics research on human pathogens and invertebrate vectors at no cost. The PFGRC provides scientists with genomic resources and reagents such as microarrays, protein expression clones, genotyping, and bioinformatics services. The PFGRC supports the training of scientists in the latest techniques in functional genomics and emerging genomic technologies.
- *NIAID's Proteomics Centers.* The primary goal of these centers is to characterize the pathogen and/or host cell proteome by identifying proteins associated with the biology of the microorganisms, mechanisms of microbial pathogenesis, innate and adaptive immune responses to infectious agents, and/or non–immune-mediated host responses that contribute to microbial pathogenesis. It is anticipated that the research programs will discover targets for potential candidates for the next generation of vaccines, therapeutics, and diagnostics. This will be accomplished by using existing proteomics technologies, augmenting existing technologies, and creating novel proteomics approaches as well as performing early-stage validation of these targets.
- *Administrative Resource for Biodefense Proteomic Centers (ARBPCs).* The ARBPCs consolidate data generated by each proteomics research center and make it available to the scientific community through a publicly accessible Web site. This database (www.proteomicsresource.org) serves as a central information source for reagents and validated protein targets and has recently been populated with the first data released.
- *NIAID's Bioinformatics Resource Centers.* The NIAID's Bioinformatics Resource Centers will design, develop, maintain, and continuously update multiorganism databases, especially those related to biodefense. Organisms of particular interest are the NIAID Category A to C priority pathogens and those causing emerging and re-emerging diseases. The ultimate goal is to establish databases that will allow scientists to access a large amount of genomic and related data. This will facilitate the identification of potential targets for the development of vaccines, therapeutics, and diagnostics. Each contract will include establishing and maintaining an analysis resource that will serve as a companion to the databases to provide, develop, and enhance standard and advanced analytical tools to help researchers access and analyze data.
- A joint collaboration between NIAID and the National Institute of General Medical Sciences (NIGMS) is providing research funding for the *Protein Structure Initiative (PSI) Centers*:

  - ○ *TB Structural Genomics Consortium.* A collaboration of scientists in six countries formed to determine and analyze the structures of about 400 proteins from *Mycobacterium tuberculosis*. The group seeks to optimize the technical and management aspects of high-throughput structure determination and will develop a database of structures and functions. NIAID, which is co-funding this project with NIGMS, anticipates that this information will also lead to the design of new and improved drugs and vaccines for tuberculosis.
  - ○ *Structural Genomics of Pathogenic Protozoa Consortium.* This consortium is aiming to develop new ways to solve protein structures from organisms known as protozoans, many species of which cause

deadly diseases such as sleeping sickness, malaria, and Chagas' disease.

25.4.2 Recent Programmatic Accomplishments

25.4.2.1 Genome Sequencing

The National Institute of Allergy and Infectious Diseases is providing support to the *Microbial Genome Sequencing Centers* (MSCs) at the J. Craig Venter Institute [formerly, the Institute for Genomic Research (TIGR)], the Broad Institute at the Massachusetts Institute of Technology (MIT), and Harvard University for a rapid and cost-efficient production of high-quality, microbial genome sequences and primary annotations. NIAID's MSCs (http://www.niaid.nih.gov/dmid/genomes/mscs/) are responding to the scientific community and national and federal agencies' priorities for genome sequencing, filling in sequence gaps, and therefore providing genome sequencing data for multiple uses including understanding the biology of microorganisms, forensic strain identification, and identifying targets for drugs, vaccines, and diagnostics. In addition, the NIAID's MSCs have developed Web sites that provide descriptive information about the sequencing projects and their progress (http://www.broad.mit.edu/seq/msc/and http://msc.tigr.org/status.shtml).

Genomes to be sequenced include microorganisms considered to be potential agents of bioterrorism (NIAID Category A, B, and C), related organisms, clinical isolates, closely related species, and invertebrate vectors of infectious diseases and microorganisms responsible for emerging and re-emerging infectious diseases.

In addition, in response to a recommendation from a 2002 NIAID-sponsored Blue Ribbon Panel on Bioterrorism and its Implication for Biomedical Research to support genomic sequencing of microorganisms considered agents of bioterrorism and related organisms, the MSCs will address the institute's need for additional sequencing of such microorganisms and invertebrate vectors of disease and/or those that are responsible for emerging and re-emerging diseases (http://www.niaid.nih.gov/dmid/genomes/mscs/overview.htm). The panel's recommendation included careful selection of species, strains, and clinical isolates to generate genomic data for different uses such as identification of strains and targets for diagnostics, vaccines, antimicrobials, and other drug developments.

The MSCs have the capacity to rapidly and cost-effectively sequence genomic DNA and provide preliminary identification of open reading frames and annotation of gene function for a wide variety of microorganisms, including viruses, bacteria, protozoa, parasites, and fungi. Sequencing projects will be considered for both complete, finished genome sequencing and other levels of sequence coverage. The choice and justification of complete versus draft sequence is likely to depend on the nature and scope of the proposed project.

Large-scale prepublication information on genome sequences is a unique research resource for the scientific community, and rapid and unrestricted sharing of microbial genome sequence data is essential for advancing research on infectious agents responsible for human disease. Therefore, it is anticipated that prepublication data on genome sequences produced at the NIAID Microbial Sequencing Centers will be made freely and publicly available via an appropriate publicly searchable database as rapidly as possible.

Completed Genome Sequencing Projects in 2006

NIAID-supported investigators have completed 131 genome sequencing projects for 105 bacteria, 8 fungi, 15 parasitic protozoa, 2 invertebrate vectors of infectious diseases, and one plant (http://www.niaid.nih.gov/dmid/genomes/mscs/req_process.htm). In addition, NIAID completed the sequence for 1,467 influenza genomes. In 2006, genome sequencing projects were completed for 22 pathogens as described in Section 23.16.2. Genome sequencing data is publicly available through Web sites such as GenBank, and data for the influenza genome sequences have been published in 2006.

Furthermore, through the NIAID's Microbial Sequencing Centers, the NIAID has funded the sequence, assembly, and annotation of three invertebrate vectors of infectious diseases. In 2006, the final sequence, assembly, and the annotation of *Aedes aegyptii* were released, as well as the preliminary sequence and assembly of the genomes for *Ixodes scapularis* and *Culex pipiens*; the final results for *I. scapularis* and *C. pipiens* will be released in 2007.

Genome Sequencing Projects in Progress

In 2006, NIAID supported nearly 40 large-scale genome sequencing projects for additional strains of viruses, bacteria, fungi, parasites, viruses, and invertebrate vectors. New projects included additional strains of *Borrelia, Clostridium, Escherichia coli, Salmonella, Streptococcus pneumonia, Ureaplasma, Coccidioides, Penicillium marneffei, Talaromyces stipitatus, Lacazia loboi, Histoplasma capsulatum, Blastomyces dermatitidis, Cryptosporidium muris,* and dengue viruses, as well as additional sequencing and annotation of *Aedes aegyptii*.

25.4.2.2 Influenza Genome Sequencing Project

In 2004, NIAID launched the *Influenza Genome Sequencing Project (IGSP)* (http://www.niaid.nih.gov/dmid/genomes/mscs/influenza.htm), which has provided the scientific community with complete genome sequence data for thousands of human and animal influenza viruses. The influenza sequence data has been rapidly placed in the public domain, through GenBank, an international searchable database, and the NIAID-funded Bioinformatics Resource Center with accompanying data analysis tools. All of the information will enable scientists to further study how influenza viruses evolve, spread, and cause disease and may ultimately lead to improved methods of treatment and prevention. This sequence information is now providing a larger and more representative sample of influenza than was previously publicly available. The Influenza Genome Sequencing Project has the capacity to sequence more than 200 genomes per month and is a collaborative effort among NIAID (including the NIAID's Division of Intramural Research), the National Center for Biotechnology Information at the National Library of Medicine, NIH (NCBI/NLM/NIH), the J. Craig Venter Institute, the Wadsworth Center at the New York State Department of Health, St. Jude Children's Research Hospital in Memphis, Ohio State University, the Canterbury Health Laboratories (New Zealand), Los Alamos National Laboratories, OIE/FAO International Reference Laboratory, Baylor College of Medicine, and others. As of September 2006, 1,467 complete genome sequences for influenza viruses were released to GenBank, including the H1N1, H1N2, and H3N2 viruses from human clinical isolates collected globally from 1931 to 2006, as well as other isolates from different hosts including birds, horses, swine, and ducks. Additional information can be found at http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/shipment.cgi

25.4.2.3 Functional Genomics

NIAID is continuing its support for the *Pathogen Functional Genomics Resource Center (PFGRC)* (http://www.niaid.nih.gov/dmid/genomes/pfgrc/default.htm) at The Institute for Genomic Research (TIGR) (currently part of the J. Craig Venter Institute). The PFGRC was established in 2001 to provide and distribute to the broader research community a wide range of genomic resources, reagents, data, and technologies for the functional analysis of microbial pathogens and invertebrate vectors of infectious diseases. In addition, the PFGRC was expanded to provide the research community with the resources and reagents needed to conduct both basic and applied research on microorganisms responsible for emerging and re-emerging infectious diseases and those considered agents of bioterrorism.

Bioinformatic Software Tools and Web Site Enhancements

One of the priorities for the PFGRC has been to provide the scientific community with access to the reagents and genomic and proteomic data that the PFGRC generated. A new software tool, called *SNP filtering tool*, was developed for Affymetrix resequencing arrays to analyze the single nucleotide polymorphism (SNP) data. Enhancements have been made to other tools for microarray data analysis, including a tool for analyzing slide images. A new layout for the TIGR-PFGRC Web site (http://pfgrc.tigr.org/) has been developed and launched and has the potential to be more user-friendly for the scientific community to access the PFGRC research and development projects, poster presentations, publications, reagents, and their descriptions and data.

DNA Microarrays

The number of organism-specific microarrays produced and distributed to the scientific community increased to 28 in 2006 and now includes arrays for viruses, bacteria, fungi, and parasites. The available organism-specific arrays include *Aspergillus fumigatus*, *Aspergillus nidulans, Candida albicans, Chlamydia,* coronaviruses (animal and human), human SARS chip, *Helicobacter pylori, Mycobacterium smegmatis*, *Neisseria gonorrhoeae*, *Plasmodium falciparum*, *Plasmodium vivax*, *Pseudomonas aeruginosa, Staphylococcus aureus, Streptococcus agalactiae, Streptococcus pneumoniae, Trypanosoma brucei, and Trypanosoma cruzi*. In addition, organism-specific microarrays were produced and distributed for organisms considered agents of bioterrorism, including *Bacillus anthracis*, *Burkholderia*, *Clostridium botulinum*, *Francisella tularensis, Giardia lamblia, Listeria monocytogenes, Mycobacterium tuberculosis, Rickettsia prowazekii, Salmonella typhimurium*, *Vibrio cholerae*, and *Yersinia pestis*.

PFGRC has continued to collaborate with the National Institute of Dental and Craniofacial Research (NIDCR/NIH) in producing and distributing five organism-specific microarrays, including arrays for *Actinobacillus actinomycetemcomitans, Fusobacterium nucleatum*, *Porphyromonas gingivalis, Streptococcus mutans*, and *Treponema denticola*.

Protein Expression Clones

PFGRC has also developed the methods and pipeline for generating organism-specific clones for protein expression.

Seven complete clone sets are now available for human severe acute respiratory syndrome coronavirus (SARS-CoV), *Bacillus anthracis*, *Yersinia pestis*, *Francisella tularensis*, *Streptococcus pneumoniae, Staphylococcus aureus*, and *Mycobacterium tuberculosis.* In addition, individual custom clone sets are available for more than 20 organisms upon request.

Comparative Genomics

Comparative genomics analysis using the available *Bacillus anthracis* sequence data and the discovery of the SNPs were used to develop a new bacterial typing system for screening anthrax strains. This system allowed NIAID-funded scientists to define detailed phylogenetic lineages of *Bacillus anthracis* and to identify three major lineages (A, B, C) with the ancestral root located between the A+B and C branches. In addition, a genotyping Genechip, which has been developed and validated for *Bacillus anthracis*, will be used to genotype about 300 different strains of *Bacillus anthracis*.

PFGRC has developed additional comparative genomic platforms for both facilitating the resequencing a bacterial genome on a chip to identify sequence variation among strains and to discover novel genes. A pilot project has been completed with *Streptococcus pneumoniae* for sequencing different strains using resequencing chip technology. In collaboration with the Department of Homeland Security (DHS), a resequencing chip has been developed and is now being used to screen a number of *Francisella tularensis* strains to identify SNPs and genetic polymorphisms. Sixteen *Francisella tularensis* strains are being genotyped by using the newly developed resequencing chip. Additional collaboration with DHS led to the development of a gene discovery platform aimed at discovering novel genes among different strains of *Yersinia pestis*. To this end, nine strains are being analyzed using this platform to discover novel gene sets.

Proteomics

PFGRC is developing proteomics technologies for protein arrays and comparative profiling of microbial proteins. A protein expression platform is under development, and a pilot comparative protein profiling project using *Staphylococcus aureus* has already been completed and published. A protein profiling project using *Yersinia pestis* to compare proteomes in different strains is now under way, complementing ongoing proteomics projects supported by NIAID; numerous proteins are currently being identified that are differently abundant during different growth conditions.

A new project was added in 2006 for comparative profiling of proteins on the proteomes of *E. coli* and *Shigella dysenteriae* to provide the scientific community with reference data on differential protein expression in animal models versus cultured systems infected with the pathogen.

25.4.2.4 Population Genetics Analysis Program: Immunity to Vaccines/Infections

In 2006, NIAID continued to support the *Population Genetics Analysis Program: Immunity to Vaccines/Infections*. A joint project between NIAID’s Division of Allergy, Immunity, and Transplantation (DAIT) and the Division of Microbiology and Infectious Diseases (DMID), this program is aimed to identify associations between specific genetic variations or polymorphisms in immune response genes and the susceptibility to infection or response to vaccination, with a focus on one or more NIAID Category A to C pathogens and influenza.

NIAID awarded six centers to study the genetic basis for the variable human response to immunization (smallpox, typhoid fever, cholera, and anthrax) and susceptibility to disease (tuberculosis, influenza, encapsulated bacterial diseases, and West Nile virus infection). The centers are comparing genetic variance in specific immune response genes as well as more generally associated genetic variance across the whole genome in affected and nonaffected individuals. The physiologic differences associated with these genome variations will also be studied. In 2006, these centers focused on recruiting the samples needed for genotyping. For example, more than 1,100 smallpox-vaccinated individuals and controls were recruited and blood and peripheral blood mononuclear cell (PBMC) samples were obtained for whole genome association studies, which were conducted in 2007.

In another example, one of the centers used genome-wide linkage approaches to map, isolate, and validate human host genes that confer susceptibility to influenza infection. Nearly 1,000 individuals with susceptibility to influenza and 2,000 control individuals were recruited using an Iceland genealogy database. By late 2006, the center had recruited more than 600 individuals and had genotyped more than 500 in this subproject of the study.

25.4.2.5 Microbial Bioinformatics

During 2006, NIAID continued its support of the eight *Bioinformatics Resource Centers (BRCs)* (http://www.niaid.nih.gov/dmid/genomes/brc/default.htm) with the goal of providing the scientific community with a publicly accessible resource that allows easy access to genomic and related data for the NIAID Category A to C priority pathogens, invertebrate vectors of infectious diseases, and pathogens causing emerging and re-emerging infectious diseases. The

BRCs are supported by multidisciplinary teams of scientists to develop new and improved computational tools and interfaces that can facilitate the analysis and interpretation of the genomic-related data by the scientific community. In 2006, each publicly accessible BRC Web site continued to be developed, the user interfaces were improved, and a variety of genomics data types were integrated, including gene expression and proteomics information, host/pathogen interactions, and signaling/metabolic pathways data. A public portal of information, data, and open-source software tools generated by all the BRCs is available at http://www.brc-central.org/. In 2006, many genomes of microbial species were sequenced by the NIAID's Microbial Sequencing Centers as well as by other national and international sequencing efforts, and the BRCs provided either long-term maintenance of the genome sequence data and annotation or the initial annotation for a number of particular microbial genomes. For example, NIAID's BRC VectorBase collaborated with NIAID's MSCs to annotate the genome of *Aedes aegyptii* with the scientific community and will continue the curation of this genome.

25.4.2.6 Microbial Proteomics

In 2006, NIAID continued to support contracts for seven *Biodefense Proteomics Research Centers (BPRCs)* to characterize the proteome of NIAID Category A to C bioweapon agents and to develop and enhance innovative proteomic technologies and apply them to the understanding of the pathogen and/or host cell proteome (http://www.niaid.nih.gov/dmid/genomes/prc/default.htm). These centers conducted a range of proteomics studies, including six Category A pathogens, six Category B pathogens, and one Category C emerging disease organism. Data, reagents, and protocols developed in the research centers are released to the NIAID-funded Administrative Resource for Biodefense Proteomics Research Centers (www.proteomicsresource.org) Web site within 2 months of validation. The Administrative Resource Web site was created to integrate the diverse data generated by the BPRCs. In 2005, more than 700 potential targets for vaccines, therapeutics, and diagnostics were generated. Examples of progress include:

(i) The elucidation of five SARS-CoV open reading frame (ORF) structures.
(ii) Cloning for expression studies of 99% of the ORFs for *V. cholerae*.
(iii) Development of multiple protocols for extracellular components and membrane subfractionation prior to mass spectroscopy.
(iv) Accurate time and mass tag databases have been populated for *Salmonella typhimurium.*

In 2006, more than 2,400 potential new pathogen targets for vaccines, therapeutics, and diagnostics were identified, and more than 5,700 new corresponding host targets were generated. In addition:

(i) Two more SARS-CoV structures were solved.
(ii) Ninety-six percent of the ORFs for *B. anthracis* were cloned with 47% sequence validated.
(iii) A custom *B. anthracis* Affymetrix GeneChip was developed.
(iv) Fifty-three polyclonal sera generated against novel *Toxoplasma gondii* and *Cryptosporidium parvum* proteins were characterized, and accurate time and mass tag databases were populated for *Salmonella typhi*, monkeypox, and vaccinia virus.

25.4.2.7 Transgenomic Activities

- NIAID staff are participating in two related NIH-wide genomic initiatives that focus on examining and identifying genetic variations across the human genome (genes) that may be linked or influence susceptibility or risk to a common human disease, such as asthma, autoimmunity, cancer, eye diseases, mental illness, and infectious diseases, or response to treatment as a vaccine. The approach is to conduct genome-wide association studies in which a dense set of SNPs across the human genome is genotyped in a large defined group of controls and diseases samples to identify genetic variations that may contribute to or have a role in the disease, with the hope of identifying an association between a genetic variant in a gene or group of genes and the disease.
  - *GAIN (Genome Association Identification Network)*. GAIN is a public-private partnership alliance with Pfizer Corp., Affymetrix, and NIH and managed by the NIH Foundation to bring new scientific and financial resources to NIH for genome-wide association studies (www.fnih.org/GAIN/GAIN_home.shtml). Initially, Pfizer Corp. has committed US$20 million for management and genotyping capacity for five common diseases in partnership with Perlegen Sciences. Investigators were invited initially to submit applications to have genotyping performed on existing DNA samples from patients with specific diseases and control individuals in case control studies. The GAIN initiative proposes to raise additional private funds for genotyping of more common diseases.
  - *GWAS (Genome-Wide Association Studies)*. GWAS is a trans-NIH committee that is focused on developing an NIH-wide policy for sharing data obtained in NIH-supported or conducted Genome-Wide Association Studies. The policy is to focus on data sharing

procedures, data access principles, intellectual policy, and issues related to protection of research participants. In 2006, the proposed policy has been shared with the scientific community for public comment as a NIH Guide Request for Information (RFI).

- NIAID has continued to participate in a coordinated federal effort in biodefense genomics and is a major participant in the *National Inter-Agency Genomics Sciences Coordinating Committee (NIGSCC)*, which includes many federal agencies. This committee was formed in 2002 to address the most serious gaps in the comprehensive genomic analysis of microorganisms considered agents of bioterrorism. A comprehensive list of microorganisms considered agents of bioterrorism was developed that identifies species, strains, and clinical and environmental isolates that have been sequenced, that are currently being sequenced, and that should be sequenced.

In 2003, the committee focused on Category A agents and provided the CDC with new technological approaches for sequencing additional smallpox viral strains. Affymetrix-based microarray technology for genome sequencing was established, as well as additional bioinformatics expertise for analyzing the genomic sequencing data. In 2004, as a result of this continuing coordination of federal agencies in genome sequencing efforts for biodefense, NIAID developed a formal interagency agreement with the Department of Homeland Security (DHS) to perform comparative genomics analysis to characterize biothreat agents at the genetic level and to examine polymorphisms for identifying genetic variations and relatedness within and between species.

- NIAID continues to participate in the *Microbe Project Interagency Working Group (IWG)*, which has developed a coordinated, interagency, 5-year action plan on microbial genomics, including functional genomics and bioinformatics in 2001 (http://www.ostp.gov/html/microbial/start.htm). In 2003, the Microbe Project Interagency Working Group developed guidelines for sharing prepublication genomic sequencing data that serve as guiding principles, so that federal agencies have consistent policies for sharing sequencing data with the scientific community and can then implement their own detailed version of the data release plan. In 2004, the Microbe Project IWG supported a workshop on "An Experimental Approach to Genome Annotation," which was coordinated by the American Society for Microbiology, and discussed issues faced in annotating microbial genome sequences that have been completed or will be completed in the next few years. In 2005, the Microbe Project IWG developed a Strategic Plan and Implementation Steps as an updated action plan for coordinating microbial genomics among federal agencies, and the plan was finalized in 2006.
- NIAID continues to participate with other federal agencies in coordinating medical diagnostics for biodefense and influenza across the federal government and in facilitating the development of a set of contracts to support advanced development toward the approval of new or improved point-of-care diagnostic tests for the influenza virus and early manufacturing and commercialization.
- NIAID continues to participate in the NIH Roadmap Initiatives, including Lead Science Officers for one of the National Centers for Biomedical Computation and one of the National Technology Centers for Networks and Pathways. Seven biomedical computing centers are developing a universal computing infrastructure and creating innovative software programs and other tools that would enable the biomedical community to integrate, analyze, model, simulate, and share data on human health and disease. Five technology centers were created in 2004 and 2005 to cooperate in a U.S. national effort to develop new technologies for proteomics and the study of dynamic biological systems.

25.4.3 Resources for Researchers

25.4.3.1 NIAID-Supported Sequencing Centers

- Bioinformatics Resource Centers
- Microbial Sequencing Centers
- Pathogen Functional Genomics Resource Center
- Proteomics Research Centers

25.4.3.2 Genome Sequence Databases

- Administrative Resource for Biodefense Proteomic Centers
- ATCC Animal Virology Collection
- BRC (Bioinformatics Research Centers) Central
- Malaria Research and Reference Reagent Resource (MR4) Center

25.4.3.3 Sequencing Projects

- NIAID Influenza Genome Sequencing Project
- The Microbe Project: U.S. Federal Efforts in Microbial Research
- Network on Antimicrobial Resistance in *Staphylococcus aureus* (NARSA)

25.5 Recent Scientific Advances

- *Supramolecular Architecture of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV).* Coronaviruses derive their name from their protruding oligomers of the spike glycoprotein (S), which forms a coronal ridge around the virion. The understanding of the virion and its organization has previously been limited to x-ray crystallography of homogenous symmetric virions, whereas coronaviruses are neither homogenous nor symmetric. In this study, a novel methodology of single-particle image analysis was applied to selected coronavirus features to obtain a detailed model of the oligomeric state and spatial relationships among viral structural proteins. The two-dimensional structures of S, M, and N structural proteins of SARS-CoV and two other coronaviruses were determined and refined to a resolution of approximately 4 nm. These results demonstrated a higher level of supramolecular organization than was previously known for coronaviruses and provided the first detailed view of the coronavirus ultrastructure. Understanding the architecture of the virion is a necessary first step to defining the assembly pathway of SARS-CoV and may aid in developing new or improved therapeutics *(49)*.
- *Large-Scale Sequence Analysis of Avian Influenza Isolates.* Avian influenza is a significant global human health threat because of its potential to infect humans and result in a global influenza pandemic. However, very little sequence information for avian influenza virus (AIV) has been in the public domain. A more comprehensive collection of publicly available sequence data for AIV is necessary for research on influenza to understand how flu evolves, spreads, and causes disease, to shed light on the emergence of influenza epidemics and pandemics, and to uncover new targets for drugs, vaccines, and diagnostics. In this study, the investigators released genomic data from the first large-scale sequencing of AIV isolates, doubling the amount of AIV sequence data in the public domain. These sequence data include 2,196 AIV genes and 169 complete genomes from a diverse sample of birds. The preliminary analysis of these sequences, along with other AIV data from the public domain, revealed new information about AIV, including the identification of a genome sequence that may be a determinant of virulence. This study provides valuable sequencing data to the scientific community and demonstrates how informative large-scale sequence analysis can be in identifying potential markers of disease *(50)*.

First Large-Scale Sequencing and Analysis of Human Influenza Viruses Supported by the NIAID-Funded Influenza Genome Sequencing Project. The analysis of the first 209 full genome sequences from human influenza strains, deposited in GenBank through the NIAID Influenza Genome Sequencing Project, was published in 2006 *(51)*. Influenza isolates were chosen in a relatively unbiased manner, allowing a comprehensive look at the influenza virus population circulating within the same geographic region over several seasons, which provided a real picture of the dynamics of influenza virus mutation and evolution. Analysis demonstrated that the circulating strains of influenza included alternative minor lineages that could provide genetic variation for the dominant strain. This may allow a novel strain to emerge within a human host and would explain the unexpected emergence of the Fujian influenza strain in 2003–2004 that resulted in a vaccine mismatch. These findings demonstrate the usefulness of full genomic sequences for providing new information on influenza viruses and lend further support for the need for large-scale influenza sequencing and the availability of sequence data in the public domain. Within the influenza community, public availability of influenza sequence data and sharing of strains has been an important issue. The NIAID has been instrumental in promoting the sharing of influenza sequence information, notably by sequencing more than 1,400 complete influenza genome sequences and depositing the sequences in the public domain through GenBank as soon as sequencing has been completed.

References

1. Smith, H. O. (2004) History of microbial genomics. In: *Microbial Genomes* (Fraser, C. M., Read, T. D., and Nelson, K. E., eds.), Humana Press, Totowa, NJ, pp. 3–16.
2. Salzberg, S. L. and Delcher, A. L. (2004) Tools for gene finding and whole genome comparison. In: *Microbial Genomes* (Fraser, C. M., Read, T. D., and Nelson, K. E., eds.), Humana Press, Totowa, NJ, pp. 19–31.
3. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., and Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding, *Genomics*, **59**, 24–31.
4. Pertea, M. and Salzberg, S. L. (2002) Computational gene finding in plants, *Plant Mol. Biol.*, **48**, 39–48.
5. Parkhill, J. and Thomson, N. R. (2004) The genomes of pathogenic Enterobacteria. In: *Microbial Genomes* (Fraser, C. M., Read, T. D., and Nelson, K. E., eds.), Humana Press, Totowa, NJ, pp. 269–289.
6. Blattner, F. R., Plunkett, G., Bloch, C. A., et al. (1997) The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453–1474.
7. Perna, N. T., Plunkett, G., 3rd, Burland, V., et al. (2001) Genome sequence of enterohemorrhagic *Escherichia coli* O157:H7, *Nature*, **409**, 529–533.
8. Hayashi, T., Makino, K., Ohnishi, M., et al. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, *DNA Res.*, **8**, 11–22.
9. Welch, R. A., Burland, V., Plunkett, G. 3rd, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*, *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 17020–17024.

10. Jin, Q., Yuan, Z., Xu, J., et al. (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157, *Nucleic Acids Res.*, **30**, 4432–4441.
11. Knapp, S., Hacker, J., Jarchau, T., and Goebel, W. (1986) Large, unstable inserts in the chromosome affect virulence properties of uropathogenic *Escherichia coli* O6 strain 536, *J. Bacteriol.*, **168**, 22–30.
12. Levine, M. M. (1987) *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent, *J. Infect. Dis.*, **155**, 377–389.
13. Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution, *Mol. Microbiol.*, **23**, 1089–1097.
14. Blum, G., Ott, M., Lischewski, A., et al. (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen, *Infect. Immunol.*, **62**, 606–614.
15. Parkhill, J., Dougan, G., James, K. D., et al. (2001) Complete genome sequence of multiple drug resistant *Salmonella enterica* serovar Typhi CT18, *Nature*, **413**, 848–852.
16. McClelland, M., Sanderson, K. E., Spieth, J., et al. (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2, *Nature*, **413**, 852–856.
17. Reeves, P. and Stevenson, G. (1989) Cloning and nucleotide sequence of the *Salmonella typhimurium* LT2 gnd gene and its homology with the corresponding sequence of *Escherichia coli* K12, *Mol. Gen. Genet.*, **217**, 182–184.
18. Mills, D. M., Bajaj, V., and Lee, C. A. (1995) A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome, *Mol. Microbiol.* **15**, 749–759.
19. Galan, J. E. (1996) Molecular genetic bases of *Salmonella* entry into host cells, *Mol. Microbiol.*, **20**, 263–271.
20. Shea, J. E., Hensel, M., Gleeson, C., and Holden, D. W. (1996) Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium*, *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 2593–2597.
21. Ochman, H., Soncini, F. C., Solomon, F., and Groisman, E. A. (1996) Identification of a pathogenicity island required for *Salmonella* survival in host cells, *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 7800–7804.
22. Kingsley, R. A. and Baumler, A. J. (2002) Pathogenicity islands and host adaptation of *Salmonella* serovars, *Curr. Top. Microbiol. Immunol.*, **264**, 67–87.
23. Blanc-Potard, A. B. and Groisman, E. A. (1997) The *Salmonella selC* locus contains a pathogenicity island mediating intramacrophage survival, *EMBO J.*, **16**, 5376–5385.
24. Buchrieser, C., Prentice, M., and Carniel, E. (1998) The 102-kb unstable region of *Yersinia pestis* comprises a high-pathogenicity island linked to a pigmentation segment which undergoes internal rearrangement, *J. Bacteriol.*, **180**, 2321–2329.
25. Reiter, W. D., Palm, P., and Yeats, S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements, *Nucleic Acid Res.*, **17**, 1907–1914.
26. Makino, K., Yokoyama, K., Kubota, Y., et al. (1999) Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak, *Genes Genet. Syst.*, **74**, 227–239.
27. Mosig, G., Yu, S., Myung, H., et al. (1997) A novel mechanism of virus-virus interactions: bacteriophage P2 Tin protein inhibits phage T4 DNA synthesis by poisoning the T4 single-stranded DNA binding protein, go32, *Virology*, **230**, 72–81.
28. Myung, H. and Calendar, R. (1995) The *old* exonuclease of bacteriophage P2, *J. Bacteriol*., **177**, 497–501.
29. Davis, B. M. and Waldor, M. K. (2003) Filamentous phages linked to virulence of *Vibrio cholerae*, *Curr. Opin. Microbiol.*, **6**, 35–42.
30. Donohue-Rolfe, A., Acheson, D. W., and Keusch, G. T. (1999) Shiga toxin: purification, structure, and function, *Rev. Infect. Dis.*, **13**(Suppl. 4), S293–S297.
31. Parkhill, J., Wren, B. W., Thomson, N. R., et al. (2001) Genome sequence of Y*ersinia pestis*, the causative agent of plague, *Nature*, **413**, 523–527.
32. Hansen-Wester, I. and Hensel, M. (2001) *Salmonella* pathogenicity islands encoding type III secretion systems, *Microbes Infect.*, **3**, 549–559.
33. Lostroh, C. P. and Lee, C. A. (2001) The *Salmonella* pathogenicity island-1 type III secretion system, *Microbes Infect.*, **3**, 1281–1291.
34. Swartley, J. S., Marfin, A. A., Edupuganti, S., et al. (1997) Capsule switching of *Neisseria meningitides*, *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 271–276.
35. Dillard, J. P., Caimano, M., Kelly, T., and Yother, J. (1995) Capsules and cassettes: genetic organization of the capsule locus of *Streptococcus pneumoniae*, *Dev. Biol. Stand.*, **85**, 261–265.
36. Dillard, J. P. and Yother, J. (1994) Genetic and molecular characterization of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 3, *Mol. Microbiol.*, **12**, 959–972.
37. Cole, S. T., Eiglmeier, K., Parkhill, J., et al. (2001) Massive gene decay in the leprosy bacillus, *Nature*, **409**, 1007–1011.
38. Perry, R. D. and Fetherston, J. D. (1997) *Yersinia pestis* – etiologic agent of plague, *Clin. Microbiol. Rev*., **10**, 35–66.
39. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*, *Proc. Natl. Acad. Sci. U.S.A.*,**96**, 14043–14048.
40. Birkelund, S., Vandahl, B. B., Shaw, A. C., and Christiansen, G. (2004) Microbial proteomics. In: *Microbial Genomes* (Fraser, C. M., Read, T. D., and Nelson, K. E., eds.), Humana Press, Totowa, NJ, pp. 517–530.
41. Wilkins, M. P., Pasquali, C., Appel, R. D., et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis, *Biotechnology (NY)*, **14**, 61–65.
42. Santoni, V., Molloy, M., and Rabilloud, T. (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoresis*, **21**, 1054–1070.
43. Adessi, C., Miege, C., Albrieux, C., and Rabilloud, T. (1997) Two-dimensional electrophoresis of membrane proteins: a current challenge for immobilized pH gradients, *Electrophoresis*, **18**, 127–135.
44. Righetti, P. G. and Gianazza, E. (1980) New developments in isoelectric focusing, *J. Chromatogr.*, **184**, 415–456.
45. Bjellqvist, B., Ek, K., Righetti, P. G., et al. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications, *J. Biochem. Biophys. Methods*, **6**, 317–339.
46. Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons, *Analyt. Chem.*, **60**, 2299–2301.
47. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules, *Science*, **246**, 64–71.
48. Cooks, R. G., Glish, G. L., Kaiser, R. E., and McLuckey, S. A. (1991) Ion trap mass spectrometry, *Chem. Eng. News*, **69**, 26–41.
49. Neuman, B. W., Adair, B. D., Yoshioka, C., Quispe, J. D., Orca, G., Kuhn. P., Milligan, R. A., Yeager, M., and Buchmeier, M. J. (2006) Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy, *J. Virol.*, **80**(16),7918–7928.

50. Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., Xu, X., Wang, J., Ma, J., Fan, Y., Rakestraw, K. M., Webster, R. G., Hoffmann, E., Krauss, S., Zheng, J., Zhang, Z., and Naeve, C. W. (2006) Large-scale sequence analysis of avian influenza isolates, *Science*, **311,**1576–1580.

51. Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., Dernovoy, D., Tatusova, T., Bao, Y., St. George, K., Taylor, J., Lipman, D. J., Fraser, C.M., Taubenberger, J. K., and Salzberg, S. L. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution, *Nature*, **437**, 1162–1166.