GENOMICS

Division of Microbiology and Infectious Diseases

The Division of Microbiology and Infectious Diseases supports a substantial program in genomics research, including sequencing of human pathogens and invertebrate vectors of diseases, applying genomic and proteomic technologies to the study of microorganisms and infectious diseases, supporting genomic databases, and providing high-quality genomic technological resources to the scientific community.

Genome Sequencing

A genome is an organism's complete set of genes, encoded as a specific sequence of paired DNA bases. Recent advances in molecular biology have given researchers powerful methods that can quickly and accurately determine the complete DNA sequence of the whole genome of virtually any organism, including disease-causing microorganisms and the insect and other invertebrate vectors that can transmit them.

Whole-genome sequencing is an enormously powerful tool for understanding and defeating infectious diseases. For example, scientists can compare and contrast genomes to identify genes that are unique to a particular microbe. They can then target these genes with specific drugs, incorporate the products of these genes into experimental vaccines, and develop more sensitive diagnostic tests. Moreover, sequence information can reveal small genetic variations between different strains of a given pathogen. Researchers can use these subtle differences to determine which genes affect a pathogen's virulence, which genes are involved in the development of antibiotic resistance, and how a virulent or resistant strain spreads within a susceptible population; better understanding of these phenomena will help to improve disease diagnosis and patient care. Finally, understanding how microbial genes interact with one another and the human host during infection will lead to new strategies for drug therapies and vaccine development.

To capitalize on the tremendous potential of genome sequencing, NIAID has invested heavily in projects to sequence the genomes of medically important microbes. Sequencing technology has advanced to the point where NIAID working alone can fund the determination of a bacterial species; however, NIAID collaborates with other funding agencies to sequence the larger genomes of protozoan and fungal pathogens. In total, NIAID has supported 57 genome-sequencing projects for 45 bacteria, 3 fungi, 8 parasitic protozoa, and 1 invertebrate vector of disease. The bacterial species include those that cause anthrax, plague, tuberculosis, gonorrhea, chlamydia, cholera, strep throat, scarlet fever, and food-borne diseases. DNA sequencing projects have also been completed for protozoan parasites *Cryptosporidium parvum, Entamoeba histolytica, Toxoplasma gondii,* and *Trypanosoma brucei* and the fungi *Aspergillus fumigatus* and *Cryptococcus neoformans.* NIAID's data release policies ensure that both raw genome sequence data and associated annotations are available to scientists around the world through publicly accessible databases. A list of NIAID-supported large-scale pathogen genome-sequencing projects is provided on page 89.

Study of the genomics of malaria has been particularly successful; for the first time, researchers have in hand the complete genetic sequences of the infectious organism, its natural host, and the insect that transmits it. In 2002, the International Malaria Genome Sequencing Consortium—funded in part by NIAID—published the genome sequence of *Plasmodium falciparum,* the parasite that causes the most severe form of malaria. NIAID also supported the rapid sequencing of the genome of *Anopheles gambiae,* the mosquito that transmits the malaria parasite to humans. Researchers therefore now have the genome sequences of all three organisms involved in malaria—the mosquito vector, the

malaria parasite, and the human host. This has provided scientists with a unique opportunity to unravel the complex interactions between these three species on a molecular level. Indeed, NIAID-supported scientists already have taken advantage of this valuable genomic information to gain new insights into the molecular mechanisms involved in insecticide resistance and to identify genes and gene products that are promising targets for new drug therapies.

The national biodefense effort has benefited substantially from genomic research as well, and NIAID has made a significant investment in sequencing microorganisms with the highest priority as agents of bioterrorism. For example, NIAID collaborated with the Office of Naval Research and the Department of Energy to sequence the genome of the Ames strain of *Bacillus anthracis,* the bacterium that causes anthrax. Other organisms important to biodefense that NIAID has helped to sequence include *Brucella suis; Burkholderia mallei; Clostridium perfringens; Coxiella burnetii;* and *Rickettsia typhi; Staphylococcus aureus; Yersinia pestis; Mycobacterium tuberculosis;* food-borne bacterial pathogens including *Escherichia coli, Vibrio cholerae, Shigella,* and *Salmonella;* and parasitic protozoa including *Cryptosporidium parvum, Giardia lamblia, Entamoeba histolytica,* and *Toxoplasma gondii.*

Because anthrax is a particularly dangerous bioterror agent, NIAID has expanded its sequencing efforts for *Bacillus anthracis* and has developed a comprehensive genomic analysis that includes sequencing of additional strains, clinical isolates, near neighbors, and related species. Under this expansion, sequencing projects were completed for six *Bacillus anthracis* strains and for two strains of the closely related bacterium *Bacillus cereus.* This effort has provided biomarkers to facilitate forensic strain identification; furthered the understanding of microbial pathogenesis; and facilitated the discovery of new targets for drugs, vaccines, and diagnostics to combat an anthrax attack.

Genomic Research

Obtaining the raw sequence of an organism's genome is only the first step in understanding it; annotating and organizing the sequence data are also required. Furthermore, the sequence data allow researchers to study an organism's proteome—the entire set of proteins that are encoded in the genome sequence. NIAID-supported investigators are applying such emerging genomic technologies to study microorganisms and infectious diseases. These studies include both basic research topics, such as the biology of a pathogen and the host's response to infection, and applied research such as development of medical diagnostics, drugs, and vaccines. Genomic technologies help scientists study infectious agents at the whole genome or proteome level. For example:

- Whole genome and proteome expression studies are being used to identify pathogen-specific genes and proteins involved in virulence, pathogenesis, and disease transmission.
- Proteomic technologies are being applied to both the pathogen and the host proteome to allow identification of candidate protein targets for the new vaccines, therapeutics, and diagnostics.
- Genomic technologies are providing platforms for examination of genetic variation within and between species, strains, and clinical isolates, as well as for study of host responses to infection, vaccines, and antibiotic drugs.

Genomic Resources, Reagents, and Technologies

NIAID facilitates distribution of genomic resources and technologies to the research community for functional genomic analysis of microbial pathogens and supports the development of bioinformatic and computational

tools that allow investigators to store and manipulate genomic and postgenomic data.

NIAID continues to support the Pathogen Functional Genomics Resource Center (PFGRC) at The Institute of Genomic Research (TIGR) in Rockville, Maryland. PFGRC was established in 2001 to distribute to the research community a wide range of genomic and related resources and technologies for the functional analysis of microbial pathogens and invertebrate vectors of infectious diseases. Considerable progress has been made toward this goal, including the generation and distribution to the research community of 19 organism-specific microarrays. This includes arrays for *Aspergillus fumigatus*, *Chlamydia*, human SARS chip, *Helicobacter pylori*, *Coronavirus* (animal and human), *Mycobacterium smegmatis*, *Mycobacterium tuberculosis*, *Neisseria gonorrhoeae*, *Plasmodium falciparium*, *Salmonella typhimurium*, *Staphylococcus aureus, Streptococcus agalactiae, Streptococcus pneumoniae,* and *Trypanosoma cruzi.*

The mission of PFGRC has been expanded to provide the research community with resources and reagents to conduct both basic and applied research on microorganisms with a high potential to be used as agents of bioterrorism, and has generated and distributed organism-specific microarrays for *Bacillus anthracis*, *Clostridium botulinum*, *Listeria monocytogenes*, *Vibrio cholerae*, and *Yersinia pestis*. In addition, microarray technology developed by Affymetrix, Inc., was added to PFGRC, and new genomic software tools have been developed for comparative genomics. PFGRC has provided severe acute respiratory syndrome (SARS) genomic resources to the broad scientific community to spur basic and applied research on SARS. PFGRC has also developed methods for generating organism-specific protein expression clones for human SARS coronavirus, *Bacillus anthracis*, *Vibrio cholerae,* and *Mycobacterium tuberculosis* and other pathogens. Further information is available at *www.niaid.nih.gov/dmid/genomes/pfgrc/default.htm*.

In FY 2003, NIAID awarded a contract to TIGR to support a Microbial Genome Sequencing Center to allow for rapid and cost-efficient production of high-quality, microbial genome sequences; in early FY 2004, NIAID awarded a contract to Massachusetts Institute of Technology to support a similar sequencing center. Genomes to be sequenced include microorganisms considered agents of bioterrorism (NIAID Category A, B, and C agents), microorganisms responsible for emerging and re-emerging infectious diseases, related pathogens, clinical isolates, and invertebrate vectors of infectious diseases. These sequencing centers have the capacity to respond to national needs and government priorities for genome sequencing, filling in sequence gaps and thus providing genome sequencing data for multiple uses, including forensic strain identification and identification of targets for drugs, vaccines, and diagnostics. In FY 2004, NIAID supported new genome sequencing projects for additional strains of *Burkholderia mallei*, *Burkholderia pseudomallei*, coronaviruses, pathogenic *E.coli*, *Francisella tularensis*, *Influenza*, *Mycobacterium tuberculosis*, *Shigella*, *Vibrio cholerae*, and *Yersinia pestis*. In addition, NIAID approved sequencing projects for invertebrate vectors of infectious diseases, *Aedes aegypti*, *Culex pipiens,* and *Ixodes scapularis.* Further information can be found at *www.niaid.nih.gov/dmid/genomes/mscs*.

The Malaria Research and Reference Reagent Resource (MR4) Center *(www.malaria.mr4.org)* continued to provide expanded access to quality-controlled reagents for the international malaria research community in 2004.

Bioinformatics and Databases

NIAID has awarded several contracts to establish Bioinformatics Resource Centers. These centers will develop, populate, and maintain comprehensive relational databases to collect, store, display, annotate, query, and analyze genomic, structural, and related data for emerging and re-emerging pathogens, including those

important for biodefense. The centers will also develop and provide software tools to assist in data analysis. Eight centers were funded in FY 2004. The databases these centers maintain are a valuable genomic resource, providing the scientific community with easy access to large amounts of genomic and related data and bioinformatics tools for data analysis. Further information is available at *http://www.niaid.nih.gov/dmid/genomes/brc/default.htm*.

Genomics and Proteomics

NIAID awarded contracts for Biodefense Proteomics Research Programs: Identifying Targets for Therapeutic Interventions Using Proteomic Technology. The goals of this program are to develop and improve proteomic technologies, and to apply these technologies to pathogen and host cell proteomes for the discovery and identification of novel targets for the next generation of drugs, vaccines, diagnostics, and immunotherapeutics against microorganisms considered agents of bioterrorism. Seven centers have been funded to date; they focus on a range of NIAID category A–C biodefense agents. Further information is available at *www.niaid.nih.gov/dmid/genomes/prc/default.htm*.

NIAID continues to collaborate with the National Institute of General Medical Sciences (NIGMS) on the NIGMS Protein Structure Initiative, which supports research centers for the development of high-throughput methods and structural determination of proteins; further information can be found at *www.nigms.nih.gov/psi*. One project supports the determination and analysis of structures of more than 400 functionally relevant *Mycobacterium tuberculosis* proteins, and another project focuses on determining the protein structures from pathogenic protozoa. Structural and functional information on many proteins from this pathogen is now available in Web-based databases for access by the scientific community at *www.sgpp.org* and *www.doe-mbi.ucla.edu/TB*.

Division of Allergy, Immunology, and Transplantation

The Division of Allergy, Immunology, and Transplantation (DAIT) also supports genomics research. The human immune system is composed of complex networks of interacting cells, each programmed by precisely scripted genes. Underlying each immune response to a disease is a multistep pathway of interacting molecules influenced by an individual's unique genomic characteristics. The immune system plays a critical role in diseases such as rheumatoid arthritis, hay fever, contact dermatitis, insulin-dependent or type 1 diabetes, systemic lupus erythematosus, and graft rejection of transplanted solid organs, tissues, and cells. Each of these diseases has an underlying genetic component.

Genomic research supported by DAIT is yielding insights into the functional and structural dimensions of immune system regulation, hypersensitivity, and inflammation in diseases such as asthma, the dysregulation of immune responses that results in autoimmune disease, and basic mechanisms of immune tolerance and graft rejection. This research is important in the following areas:

- **Asthma and allergic diseases.** DAIT-supported research on the genetics of asthma, hypersensitivity, inflammation, and T cell mediation enables us to understand the mechanisms underlying these immune responses, resulting in improved diagnostic, prevention, and treatment strategies. Through genomic research, DAIT-supported investigators discovered that interleukin-4 (IL-4), a cytokine that is produced by helper T cells and mast cells, stimulates antibody production by B cells in a series of reactions involving several genes. Further studies on IL-4 might provide a marker for measuring asthma risk and severity.
- **Autoimmune diseases.** DAIT supports research on type 1 diabetes and other

autoimmune diseases that involve more than a single gene. Recent developments in genomics such as high-resolution DNA analysis and bioinformatics tools are making it possible to understand the underlying genetic causes of these complex diseases. For example, one approach compares the genes of individuals who have an autoimmune disease with those of healthy individuals to identify genetic and genomic differences that might be the underlying cause of disease. Between 10 and 20 distinct loci on the human genome may be responsible for susceptibility to type 1 diabetes. This knowledge will increase our ability to predict, diagnose, and treat this disease.

- **Transplantation.** DAIT-supported research on the genetics of graft rejection and immune tolerance is breaking new ground in the transplantation of solid organs, tissues, and cells for the prevention and treatment of disease. Genomic research funded by DAIT has identified surrogate markers of graft rejection in kidney transplant recipients. This research holds promise for the development of a noninvasive predictor of graft rejection based on gene expression analysis in urinary cells of transplant recipients.

- **Basic immunology research.** Basic research in immunology furthers our understanding of the properties, interactions, and functions of the cells of the immune system and the genetic aspects of immune system regulation and provides information about essential structural immunobiology. Recent breakthroughs in the basic science of immunogenetics inform clinical immunology, which may lead to the development of new immune-based therapies. Examples of basic immunology research supported by DAIT include the following:

  – Use of large-scale gene- and protein-expression analysis tools to describe pathways of cellular activation;

  – Discovery of anti-inflammatory and immunosuppressive agents using DNA-based screening methods; and

  – Analysis of genomic databases of T cell receptors and immunoglobulin gene sequences to link structural, functional, and clinical information.

Multicenter Research Programs

DAIT supports several multicenter research programs that include significant genomic efforts aimed at understanding the underlying mechanisms of immune-mediated diseases.

Immune Tolerance Network (ITN). The ITN is an international consortium of more than 80 investigators in the United States, Canada, Europe, and Australia dedicated to the clinical evaluation of novel, tolerance-inducing therapies in autoimmune diseases; asthma and allergic diseases; and rejection of transplanted organs, tissues, and cells. The goal of these therapies is to re-educate the immune system to eliminate harmful immune responses while preserving protective immunity against infectious agents. To understand the underlying mechanisms of action of the candidate therapies and to monitor tolerance, ITN has established state-of-the-art core laboratory facilities to conduct integrated mechanistic studies and to develop and evaluate markers and assays to measure the induction, maintenance, and loss of tolerance in humans. These core facilities include microarray analyses of gene expression, bioinformatics approaches to develop analytic tools for clinical and scientific data sets from the ITN-sponsored trials, enzyme-linked immunospot analyses of protein expression, and cellular assays for T cell reactivity. ITN has completed 2 clinical trials; 23 trials are ongoing or in development. ITN is co-sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases and the Juvenile Diabetes Research Foundation International. More information on the ITN is available at *www.immunetolerance.org*.

Autoimmunity Centers of Excellence (ACEs). ACEs support collaborative basic and clinical research on autoimmune diseases, including single-site and multisite pilot clinical trials of promising immunomodulatory therapies. ACEs presently are enrolling participants in several clinical trials, including a trial of anti-CD20 in SLE and a trial of anti-C5 in lupus nephritis.

International Histocompatibility Working Group (IHWG). IHWG is a network of more than 200 laboratories in more than 70 countries that applies new molecular techniques to population-based studies of histocompatibility genes. Histocompatibility genes allow the immune system to respond to specific pathogens, but these genes also play a role in the unwanted immune responses that occur in graft rejection and autoimmune diseases. Recent advances in genomics will facilitate the work of the human leukocyte antigen class II genes and related polymorphisms and their role in immunity, disease susceptibility, and graft rejection. Genomic techniques developed by IHWG investigators and others have shown a greater diversity among histocompatibility genes than was previously detected by conventional serologic methods. This work will bridge the gap between serologic and genomic definitions of these genes.

Multiple Autoimmune Disease Genetics Consortium (MADGC). MADGC is a repository of genetic and clinical data and specimens from families in which two or more individuals are affected by two or more distinct autoimmune diseases. This resource provides materials to promote research aimed at discovering the human immune response genes involved in autoimmunity. More information can be found at *www.madgc.org*.

North American Rheumatoid Arthritis Consortium (NARAC). NARAC is a collaborative registry and repository of information on families with rheumatoid arthritis. The NARAC database contains information on 902 families, encompassing 1,522 patient visits. Of the 902 families, data for more than half have been validated, including 600 affected sibling pairs. The family registry and the repository samples should facilitate the characterization of the genes underlying susceptibility to rheumatoid arthritis and are available to all investigators. This registry is cosponsored by the National Institute of Arthritis and Musculoskeletal and Skin Diseases and the Arthritis Foundation. More information can be found at *www.naracdata.org*.

Primary Immunodeficiency Diseases Registry and Consortium. In FY 2003, the Primary Immunodeficiency Diseases Consortium was established with support from NIAID and the National Institute of Child Health and Human Development. The Consortium (1) provides leadership and mentoring; facilitates collaborations; enhances coordination of research efforts; and solicits, reviews, recommends, and makes awards for pilot or small research projects; (2) maintains a primary immunodeficiency diseases registry, which provides data to the research community about the clinical characteristics and prevalence of these diseases; and (3) develops a repository of specimens from subjects with primary immunodeficiency diseases. Additional information on Consortium activities is available at *www.usidnet.org*.

The following is a list of NIAID-supported large-scale pathogen genome-sequencing projects active in fiscal year 2004:

ORGANISM	DISEASE
Aedes aegypti	invertebrate vector for yellow fever
Anopheles gambiae	malaria
Aspergillus fumigatus	aspergillosis
Bacillus anthracis	anthrax
Bacillus cereus	food poisoning
Brugia malayi	elephantiasis
Burkholderia mallei	glanders
Burkholderia pseudomallei	melioidosis
Burkholderia thailandensis	non-virulent strain
Clostridium perfringens	gas gangrene
Coccidioides immitis	respiratory infections; coccidioidomycosis
Cryptococcus neoformans	cryptococcosis
Cryptosporidium parvum	food-borne and water-borne diseases, gastritis
Culex pipens	invertebrate vector for West Nile virus
Escherichia coli K1 RS218	meningitis
Ehrlichia spp.	ehrlichiosis
Entamoeba histolytica	dysentery
Francisella tularensis	tularemia
Giardia lamblia	giardiasis
Histoplasma capsulatum	histoplasmosis
Influenza viruses	influenza
Ixodes scapularis	invertebrate vector for Lyme Disease
Legionella pneumophila	Legionnaire's disease
Leishmania major	cutaneous leishmaniasis
Mycobacterium tuberculosis	tuberculosis
Mycobacterium smegmatis	non-virulent strain
Nematode species	helminthiases
Plasmodium vivax	malaria
Pneumocystis carinii	pneumonia, opportunistic disease
Rickettsia rickettsii	Rocky Mountain spotted fever
Rickettsia typhi	typhus
Salmonella typhi	typhoid fever
Schistosoma mansoni	dermatitis, Katayama fever, liver inflammation, fibrosis
Streptococcus agalactiae	Group B Streptococcus
Stretococcus pneumoniae	pneumonia, meningitis
Toxoplasma gondii	toxoplasmosis, congenital, and ocular infections, opportunistic disease
Trichomonas vaginalis	vaginitis
Trypanosoma brucei	trypanosomiasis
Trypanosoma cruzi	Chagas disease
Vibrio cholerae	cholera
Wolbachia	endosymbiont of filarial nematodes and insect vectors
Yersinia pestis	plague