Workshop Report
The Future of DNA Sequencing at the National Human Genome Research Institute
March 23-24, 2009

What are the most important biomedical questions that can be addressed with large-scale sequence data? What are the most compelling sequence-based community resources that should be generated? What are the consequences of the rapid increase in sequencing capacity, and the rapid decrease in cost, afforded by the new technology platforms? In order to answer these questions, the National Human Genome Research Institute (NHGRI) convened a workshop to discuss the future of large-scale sequencing as one component[1] of the Institute's current two-year planning process for all of its scientific programs.

The need for this workshop was particularly underscored by the recent and ongoing rapid changes in sequencing technology, propelled by the "next generation" sequencing platforms. Introduced into production activities less than two years ago, the new sequencing platforms have already afforded an increase in throughput[2] of two orders of magnitude over the previous platforms, and this is likely to increase by nearly another order of magnitude in the next year or two. Furthermore, yet newer technologies are being developed and are expected to be available in the next three to five years. These rapid changes offer incredible new opportunities as well as major new challenges for the use of sequencing technology in general and to NHGRI's sequencing program specifically. As the technology continues to improve, new applications of genomic sequencing are constantly being developed, for example the sequencing of genomes from large numbers of individuals for disease and population studies, quantitative transcriptional analysis and epigenomics.

The 'disruptive' technological change has many other consequences. Most obviously, the ability to apply large-scale sequencing efficiently towards a larger number of problems will result in unprecedented demands on scientists' ability to find enough samples that are appropriate to addressing an expanded range of questions. To date, the most difficult problem has been obtaining samples for human disease or population studies that are properly consented for the work. One can also foresee

---

[1] Other components include workshops in specific topic areas, and solicitations of public responses to Web-based "white papers": see http://www.genome.gov/10001307. The outcome of the planning process will be an update of the Institute's previous five-year plan (*Nature*, Vol. 422, No. 6934, April 24, 2003, p. 835-847).

[2] The increase in sequencing capacity in just the NHGRI centers has been remarkable. In 2007, capacity was 0.15 Tb/year. In the next 12 months, we estimate that this will increase to over 30Tb/year, with another ~5-fold increase in the subsequent year. A note of caution is appropriate here---the data from the highest throughput new sequencing platforms consists so far of short reads—perhaps 35-75 basepairs as of this writing. This type of data is extremely useful for resequencing applications. However, higher coverage is required than was required for the older, long-read platforms, so the increase in raw data throughput, while phenomenal, must be interpreted with caution. In additional, short-read data are not yet useable in *de novo* assemblies for large, complex organisms. Another new platform, Roche/454, provided longer reads that are more useful for *de novo* applications, but data are more expensive to obtain than with the short read platforms.

increased demands of rigorous and systematic gathering and perhaps storage of samples from larger sets of organisms and of human tissues.

Another major consequence will be the enormous demands on informatics infrastructure and data analysis imposed by very large data sets, and the demands for new analytical tools that will enable the wider research community to use the data.

Finally, as technology improves, sequencing will become a significantly more dispersed activity with many research groups other than large sequencing centers able to produce more and more data. This change will create even more new challenges for sample collection, data analysis, data deposition, and informatics infrastructure.

This sociological change is an important consideration in the Institute's planning for the future of the sequencing program, and will have an effect on how NHGRI organizes its sequencing program, and the types of sequencing project that are no longer appropriate for support by NHGRI, but rather by another NIH institute or other funding body.

This report is intended to accurately reflect the major points that were raised at the workshop, but no programmatic decisions have yet been made on its basis. NHGRI intends to incorporate the discussions from the workshop with all of its other planning activities in making decisions about the future of the large-scale sequencing program, in the overall context of the many extramural programs that NHGRI funds, and in consultation with the National Council on Human Genome Research. It is especially important to emphasize that, while the discussions at the workshop raised many good ideas about new projects that will soon become possible with new sequencing platforms, it is not yet clear which among these (or others not discussed here) should go forward, nor how they should be funded.

It is also important to be clear that NHGRI has not been the only locus of funding for large-scale genomic sequencing activities. Part of the National Institutes of Health, the National Institute of Allergy and Infectious Diseases funds substantial sequencing capacity. The US Department of Energy, the Wellcome Trust Sanger Institute, and funding bodies in China, Japan, Germany, and elsewhere have funded large-scale sequencing programs. NHGRI has collaborated with all of these entities on major projects. Although the purpose of this workshop was to provide guidance to NHGRI, all mentions of the benefits from large-scale sequencing mentioned herein must be shared with our international colleagues. We anticipate that other large-scale sequencing programs will find some of the considerations raised at this workshop to be relevant to them.

The workshop agenda (see Appendix 1) was designed to raise a range of topics related to genome sequencing, and to elicit comments and advice from three distinct "breakout" sessions. The workshop concluded with a discussion and the development of a set of overall recommendations. This summary begins with those overall recommendations, either as explicitly stated in the final discussion or as raised

repeatedly during the workshop, followed by the recommendations from the breakout sessions.

General Recommendations

- NHGRI is uniquely positioned within the National Institutes of Health to undertake the development, assessment, and implementation of a broad range of projects involving large-scale application of genomic technologies.

- Computational biology methods, resources, and infrastructure have not kept pace with the increased rate of sequence output by the entire community, and the need to integrate sequence data with other biological and biomedical information is growing. Infrastructure, tools and expertise for computational biology will be needed by the entire scientific community as biology becomes more data-driven. NHGRI will need to play a role in filling this need, along with other funding agencies.

- NHGRI should maintain, with some modifications (see next point), a sequencing program that involves large-scale sequencing centers. There are many compelling projects that can only be properly done at a very large scale. Furthermore, the sequencing centers contribute much more than just data—they develop knowledge about how to approach and design many types of genome projects, drive understanding about what types of major biological and biomedical questions can be addressed by the application of sequence data, develop software tools and methods that can be dispersed to the community, implement rapidly improving new technology platforms so that they are "reduced to practice", set quality and other standards for sequencing and sequence data, and generally provide intellectual leadership that moves the field forward.

- NHGRI has an important (but by no means exclusive) role in encouraging the wider dispersion—enabled by ever cheaper "next-generation" genome sequencing---of the tools and knowledge to do genome projects across a range of topics, especially in medical sequencing. There are at least two ways that NHGRI could do this:

    i. by encouraging that the tools and knowledge generated in the large sequencing centers be made robust and more readily usable by smaller research laboratories.

    ii. by providing opportunities for smaller, more specialized groups to engage in "next-generation" sequencing projects that are not of an appropriate scale for the large centers. Many at the workshop thought that there were a number of valuable medical sequencing projects that could, and should, be done at a smaller scale than is appropriate for a project in a large

center. Such smaller sequencing efforts should not just be smaller versions of the large-scale centers. Rather, they should undertake well-defined, well-developed projects that address biological or biomedical problems in an integrated way, from sample acquisition through sequencing through analysis. NHGRI should solicit co-funding support from other funding sources for these specialized groups.

- Genome projects funded by NHGRI will increasingly relate to topics of great interest to the categorical NIH institutes. NHGRI will need additional staff resources to effectively administer the increasing number of genome projects that are, at the same time, growing in complexity due to their size, and are dependent on collaborations for the collection, characterization, and analysis of disease samples. Collaborations with other disease-specific NIH institutes may be useful in leveraging staffing.

- The new sequencing platforms excel in data production, but so far do not produce a "finished" genome, particularly for *de novo* assemblies. NHGRI should address the technical problem of how to finish genomes using the new technologies.

Recommendations from the Breakout Session Discussions

Topic 1: Strategic planning for selecting projects; sample coordination; ELSI and consent

How do we encourage the best range of new genome projects? What are the major issues were likely to be that would help or hinder such projects, and what are the consequences for these issues when new technologies permit genome projects to be undertaken more widely?

Overall

- NHGRI should continue to undertake "flagship" projects that produce important community resources with a significant useful lifetime.

- From their inception, projects should be designed to achieve full integration across target selection, sample acquisition, sequencing, analysis and long-term maintenance of the data.

- The lack of availability of high quality samples has already been a challenge for some large sequencing projects. NHGRI should encourage the creation of sample repositories to ensure continuous availability of high quality samples. Genome projects will increasingly involve samples whose collection was

funded by the categorical NIH institutes, so this will only work well in collaborations between NHGRI and other institutes.

- NHGRI should encourage community education about genome sequencing projects, including establishment of sample and data quality standards for genome projects, appropriate consent, and basic knowledge and tools for doing genome projects.

- As the number and diversity of genome projects increases, NHGRI should undertake to account for them. That is, NHGRI should attempt to track what large genome projects are being done worldwide.

Medical Sequencing

- For large sequencing projects, NHGRI should work with other ICs that have domain knowledge to identify projects, ideally to be cooperatively funded and staffed.

- One way to increase the diversity and creativity of new projects is to establish dedicated 'smaller' medical sequencing centers. There are a number of possible medical sequencing projects that are of a scale more appropriate to a smaller center.

- Ensure that these 'smaller' centers have close ties with larger centers for technology transfer and support

- NHGRI should encourage a clear articulation of the required quality of samples to clinical communities.

Organismal Sequencing

- The current NHGRI "working group" mechanism should be continued for organismal sequencing and pathogens and vectors projects in order to ensure that the best community projects are identified[3]. The working groups need better outreach to the community, though this should not be the rate-limiting step.

- Organismal projects should be planned as a "package" ready for use by the biological community. Packages would include, for example, the reference genome sequence, variation information, some cDNA/RNA sequencing information, and automated gene annotation. As capabilities improve, the "package" should evolve. This will improve the usability of the product by the

[3] A list of organismal genome projects currently underway, funded by NHGRI, is available at http://www.genome.gov/10002154.

community, and it will also help set a standard for an integrated genome project.

- For organismal genomes, the reference sequences should be upgraded from "draft" status to high quality assemblies where justified by the scientific need.

Topic 2: Genome sequencing

This discussion covers the range of opportunities for compelling sequencing projects that could be done in the next five years, given the likely trajectory of the new sequencing capabilities, and considering the ability of sequence data to be used in addressing important biological and biomedical questions. It also covers how sequencing activities might be optimally organized at NHGRI to get these done.

Scientific opportunities for sequencing

Human Genetics

- Provide a catalog of variation, to a minor allele frequency of less than 1% as an enduring resource for human population and disease genetics. The 1000 Genomes[4] project is already beginning this, but an expansion of this effort can be justified as giving access to rarer variants and to more populations. An expanded effort should include phenotyped samples.

- Characterize the sequence variation underlying most common diseases[5]:

  i. Whole genome sequencing to complement (or possibly supersede) genome-wide association studies for major common diseases. To attain sufficient power, it is estimated that each study would require sequencing of ~10,000 individuals (cases/controls). With steady increases in sequencing capacity, it may be feasible to undertake 20 studies in the next five years.

  ii. Whole genome sequencing in medically characterized populations in longitudinal studies. This would also bring in the dimension of understanding of e.g., treatment and environmental response. Populations that could be considered include:
    - Prospective (for example [the National Children's Study)
    - An already well-characterized population that continues to be in follow-up (for example Framingham)
    - Discrete populations (for example Iceland, Utah CEPH)

[4] http://www.1000genomes.org/page.php
[5] NHGRI has begun several projects to sequence candidate regions and under association peaks for four diseases. See URL http://www.genome.gov/20019648. In addition, NHGRI is collaborating with several other NIH institutes on projects of this type, related to specific diseases.

- Find modifiers of highly penetrant disease variants, for example in cystic fibrosis or sickle cell disease. This would, among other things, identify constituents of pathways.

- Take a systematic approach to identifying the causal variants underlying all Mendelian diseases. There are ~7000 Mendelian traits, and for most of these the responsible gene/variant is not known. Regional, whole exome, or whole genome approaches may be appropriate, depending on the degree of genetic characterization and the cost of genome sequencing.

Functional Genomics

- Implement deep transcriptome sequencing for purposes of genome annotation and understanding function, especially in humans and important model organisms .

    i. The current Genotype-Tissue Expression (GTEx[6]) Project pilot will sequence transcriptomes of 50 human tissues from each of 160 donors. One of the purposes of the pilot is to determine whether there is scientific justification for scaling to 1000 donors.
    ii. Projects involving pediatric or fetal samples could improve our understanding of human development.

- Develop of single cell methods to provide greatly improved precision across all aspects of functional genomics.
- Use the new sequencing technologies to do epigenetic analysis at genomic scale.

Cancer

- Cancer genome analysis by large-scale sequencing is an area where the new methods can make an immediate and substantial impact and where NHGRI and others should be supporting increased efforts.
- Sequence the full genomes of all tumors paired with constitutional DNA[7].
- Include transcriptome/epigenome analysis using large-scale sequencing approaches in tumor characterization.
- Analyze heritable cancers using the same approach discussed above for other human genetic disease.

The Human Microbiome

---

[6] GTEx (http://nihroadmap.nih.gov/GTEx/index.asp) is an effort funded by the NIH Roadmap program.
[7] Two projects, The Cancer Genome Atlas (http://cancergenome.nih.gov/), a collaboration with the National Cancer Institute, and the Tumor Sequencing Program (URL http://www.genome.gov/19517442) demonstrate that this is a productive approach.

- Analyze microbiomes of many more normal subjects than is now being considered for the Human Microbiome Project (HMP)[8], in order to obtain a fuller appreciation for the range of microbial communities in the human, and how their composition relates to environmental and other factors.
- Include the sequencing of host genomes.
- Integrate microbiome information with other projects, for example 1000 Genomes or GTEx.
- Analyze the microbiome of model organisms, for example to enable experimental analysis.
- Use sequencing to attain a more fundamental understanding of microbe biology, for example microbial communities, gene transfer, and other fundamental biology.

Genome Evolution and Model Organisms

- NHGRI has so far chosen new organismal genome projects[4] with several goals in mind. These remain important aims:

    i. Inform human biology, for example by using comparative analysis to annotate the human genome, or to identify regions of the genome under recent selection.
    ii. Understand the evolution of every basepair in the human (mammalian) lineage. Among other things, this will provide insight into the functional or other constraints on the human genome.
    iii. Understand the mechanisms and forces, such as adaptation, selection, duplication, deletion, birth and death of genes, horizontal gene transfer, etc. that shape the human genome, and genomes in general. Also, understand the genomic features that accompany the evolution of major novel anatomic or physiological features which are of fundamental biological and biomedical interest, for example an adaptive immune system or a myelinated nervous system.
    iv. Enable biomedical research, for example by adding value to important experimental or disease model systems, and furthering study of pathogens and vectors of human disease.

- However, with regard to organismal sequencing, NHGRI is at the point where it should assess what has been done (considering organisms, spacing/number of examples/quality of sequence), and what more it should do. How far has it gotten in attaining these goals?

[8] The HMP (http://nihroadmap.nih.gov/hmp/) is a pilot project funded by the NIH Roadmap program.

- It is clearly still important for understanding human biology to have genome sequences from multiple non-human primates at high quality with reasonable phylogenetic coverage. Among other things, primate sequence will facilitate insight into recent human evolution.

- Population genetics of model organisms is an important, relatively unexplored area. NHGRI does fund some population genomic sequencing on fly and yeast models, and others fund similar studies on mice. But much more would be important, and other organisms may provide unique value (e.g. stickleback).

Possible organization of sequencing activities

The Genome sequencing breakout group discussed the merits of research groups operating on different scales to address the scientific opportunities.

- Large-scale centers will continue to be important because many of the most compelling opportunities—especially those involving the sequencing of many human genomes--- must be done at large scale. In addition, large-scale centers make a number of critical contributions beyond data production of data[9]. For example, they implement new sequencing technologies and, in doing so, develop a deep understanding of the capabilities and opportunities for improvement of the technologies than either the manufacturers or smaller laboratories would do. Large-scale centers also disseminate technology and develop bioinformatic tools. They are also a source of innovation and intellectual leadership about the design and many pragmatic aspects of genome projects.

- Medium-scale centers will be enabled by the new platforms, and are likely to have unique strengths. For example, they could more easily[10] be designed to undertake vertically integrated projects, incorporating geneticists, data producers, and those with analysis expertise. They could concentrate on one or a small set of related disease areas, and thus bring all the required materials and expertise to bear on a biological problem. They could provide an opportunity to identify and carry out novel and creative kinds of sequencing

[9] Large-scale centers have dedicated technology implementation budgets that allow them to explore e.g., increasing machine performance, reagent use reduction, sample preparation, characteristics/quality of the data, use and assembly of the data, and integration and optimization of "production related" informatics. This has led repeatedly to significantly improved cost efficiencies that have been disseminated to the wider community, knowledge about data quality and its effect on the results of genome projects, productive discussions with the instrument makers regarding improved quality of data and reagents, opportunities for supply-chain improvement, and multiple other contributions. Large-scale centers also have dedicated budgets for outreach and collaboration with communities that use genomic data they produce.

[10] Medium scale efforts, if designed appropriately, could have specific advantages over large centers in key areas. Large-scale efforts are usually evaluated substantially on the basis of their efficiency in very high-throughput production. In addition, to justify their high overall level of funding, they undertake very many large projects. Although the projects that they undertake do require close collaboration with the community, the organization and incentives are more aligned with high-throughput production.

projects that are of high value, but may not be suitable for very large-scale work. They could also provide a different kind of opportunity for new trainees who would then go on to incorporate genomics into biomedical research. Medium-scale centers would produce enough data, and would be funded at a sufficient level, that the requirements for data deposition based on community resource projects should be the same as for large-scale centers.

- R01 research that involves significant genome sequencing will be enabled by the new platforms. Such groups would be well suited, for example, to explore new areas of genome science. These may or may not be appropriate for NHGRI funding. However, they may be appropriate to the extent that they are providing insight into novel, larger aspects of genome science, rather than, for example, sequencing the genome of a new species or of an individual person.

Initiatives

The Genome sequencing breakout group discussed several examples of specific initiatives that could be undertaken by NHGRI.

- A solicitation for "medium-scale centers" to focus on Mendelian diseases.

- A technology development initiative to develop methods to obtain genomic data from single cells or small collections of cells.

- An initiative to develop methods to work on "difficult" regions of the genome using next generation technology.

- An initiative to develop *de novo* sequencing and assembly using-next generation technology.

- Development of a publically accessible progress report to provide information on an on-going basis about the status of current projects (i.e. the sample collection stage, the sequencing stage, and analysis stage), with clear articulation of anticipated timelines[11].

- A specific initiative to integrate genomics into medicine.

Topic 3: Downstream issues: Informatics and analysis

This discussion covered a range of issues related to what happens "downstream" of data production, including informatics and analysis. Points were raised regarding the proper balance of ensuring privacy while maximizing the ability to do analyses for medical sequencing data, chronically underfunded informatics infrastructure. We can anticipate

[11] The list of ongoing organismal projects is similar to this (http://www.genome.gov/10002154), though it needs to be updated.

that several considerations about costs and funding are going to change as sequencing becomes dispersed. For example, there will be an increased need for development of informatics tools and resources to handle the data, yet these will require ongoing maintenance placing new demands on limited funds. More generally, as the costs of data production come down, the relative costs of analysis and informatics increase. The discussion also emphasized that the entire field of biology is still adapting to using and publishing papers on large data sets, and that this could require changes in the way that papers are published, what data they contain, and how those data are referenced, etc.

Data security and privacy

- This issue is larger than an informatics problem; the size of the sequence data sets that can now be generated easily inherently creates problems because such data sets are individually unique and could be used to identify individual research participants. Developing effective approaches that minimize the possibility of harm to individuals while at the same time facilitating progress in medical research will require ongoing public discussion and analysis, and consensus-building with research and disease communities. Such discussions have already started, have been productive, but have not solved the problem so need to continue.

- Many genomic studies are using samples with consents that were written before the possibility of genomic work was contemplated. Usually, these do not provide for widespread data release. These "legacy" consents must be respected.

- There is a partial technical solution, which is to provide controlled data access only to researchers who have obtained prior approval based on criteria consistent with the consents[12].

- In going forward, NHGRI and others should work towards encouraging consents that allow full unrestricted public access. However, the risks inherent in this approach are still not well understood and further analysis is needed.

- An interesting suggestion was made for a novel approach involving the identification of a community of informatics experts who are pre-approved ("bonded and licensed") to work on controlled access medical sequencing datasets to make them both acceptable from the perspective of individual safety and useful for the general research community .

Hardware infrastructure

[12] dbGaP is an example of a "controlled access" public database of genomic datasets . See http://www.ncbi.nlm.nih.gov/gap

- The current hardware infrastructure is insufficient for managing and analyzing the amounts of genome sequence data that the next-generation sequencers will produce. Meeting the hardware need is an "unfunded mandate" because the data are not useful without it, yet it is seldom fully planned for at all levels (by funding agencies, institutions, and often by investigators).

- As genome sequencing becomes more dispersed, the creation of data centers at the department or institute level can be contemplated, although individual labs will continue to need some funding for this purpose. Data centers only work well where a sufficient concentration of users exists.

- Cloud computing is an attractive alternative to centrally organized informatics hardware centers, and should be explored and developed. It addresses several issues in that it avoids having to build hardware capacity in multiple locations and may help serve more isolated PIs. Many companies offer this service, but they should not yet be absolutely relied upon.

<u>Software infrastructure</u>

- Informatics tool development remains important, and NHGRI could be even more stringent about making sure that applications for tool development projects propose new tools that are likely to be improvements over the state-of-the art for community use, rather than ones proposed based on local needs, abilities and interests.

- Tool maintenance ("hardening") needs more support. Funding mechanisms for this exist, and should be continued. It will be an ongoing challenge to identify which among many should receive ongoing support for maintenance.

- In publications on this topic, better "materials and methods" sections and computer parse-able results will go a long way toward addressing issues of accessibility, reproducibility, integration, and investigator education.

<u>Biology using computation</u>

- The demand is very high for people who can test biological hypotheses using computational approaches on large data sets; this includes people working as independent PI's, or in collaboration with biologists or genome centers.

- NHGRI should continue to support the entire field at all levels. NHGRI's R01 portfolio represents an opportunity for individual laboratory projects using big genomic data sets.

Appendix 1: Workshop Agenda

The Future of DNA Sequencing at the National Human Genome Research Institute
March 23-24, 2009

Purpose of the Meeting

The NHGRI is undertaking a two year scientific planning process to update the Institute's scientific vision that was elaborated over five years ago (*Nature*, Vol. 422, No. 6934, April 24, 2003, p. 835-847). Large-scale genomic sequencing has traditionally been the centerpiece as well as the largest of the Institute's extramural programs. This meeting is designed, as part of the overall NHGRI planning process, to consider questions of high importance for the future of its large-scale sequencing program within the overall Institute's mission: What are the most important biomedical questions that can be addressed with large-scale sequence data? What are the most compelling sequence-based community resources that should be generated? What are the consequences of the rapid increase in sequencing capacity, and the rapid decrease in cost, afforded by the new technology platforms?

Rapid changes in sequencing technology offer new opportunities as well as new challenges for NHGRI's sequencing program. As the technology improves, new applications of genome sequencing become possible, including sequencing of multiple human genomes for disease and population studies, and metagenomics. The new technologies also enable other applications, including quantitative transcriptional sequencing and epigenomics.

There are other consequences of the rapid technological change. Most obviously, the ability to apply large scale sequencing efficiently towards a larger number of problems results in significant new demands on the capacity to find samples appropriate to addressing an expanded range (and volume) of questions. To date, the most difficult problem has been obtaining samples for human disease or population studies that are properly consented for the work. One can also foresee increased demands of rigorous and systematic gathering and perhaps storage of samples for larger sets of organisms and for human tissues.

Another consequence will be the demands on informatics infrastructure and data analysis imposed by very large data sets, and the demands for new tools to enable the wider community to use the data.

Finally, there is a sociological consequence: as sequencing technology improves sequencing will become a significantly more dispersed activity. This change will create new challenges for sample collection, data analysis, and informatics infrastructure. This sociological change is an important consideration in the Institute's planning for the future of the sequencing program.

Within this context, the agenda for this workshop is designed to provide the Institute with guidance about how it should continue support for large-scale sequencing over the next five year period.

Specific Questions to Animate the Discussion

What kinds of projects are important for the NHGRI sequencing program to do?
What will be the "mission" and unique role of the NHGRI program?

- What is the next generation of important sequence-based community resource projects? What kind of biomedical and biological questions should motivate the program?
- One important goal of the next iteration of the large-scale sequencing program should be to make genome projects as easy to do as possible. How should the Institute accomplish that?
- How should the "upstream" issues be addressed in centralized and decentralized sequencing program scenarios?
- How should the community address the informatics needs for dealing with sequence information? What are the considerations for both large-scale, more centralized, and small-scale, more decentralized futures of the program? What is NHGRI's unique role in this?

Balance of centralization and decentralization within a future sequencing program:

- What types of activities would best be undertaken by larger centers with very large-scale sequencing capabilities? What magnitude of scale is needed to carry out these activities? What role should development and dissemination of technologies and computational tools have in larger centers?
- What types of activities would best be undertaken in smaller scale, less centralized centers? What steps could NHGRI take to facilitate wider dissemination of sequencing expertise?
- What are the implications of the balance for:
  -- the biomedical enterprise in general;
  -- the needed infrastructure (especially informatics and analysis downstream; and technology and samples upstream);
  -- technology development? How should those scenarios evolve over time, as technology changes? (2, 5, 10 years?)

Workshop on the Future of the Large-scale Sequencing Program

Purpose of the Meeting

The NHGRI is undertaking a two year scientific planning process to update the Institute's scientific vision that was elaborated over five years ago (*Nature*, Vol. 422, No. 6934, April 24, 2003, p. 835-847). Large-scale genomic sequencing has traditionally been the centerpiece as well as the largest of the Institute's extramural programs. This meeting is designed, as part of the overall NHGRI planning process, to consider questions of high importance for the future of its large-scale sequencing program within the overall Institute's mission: What are the most important biomedical questions that can be addressed with large-scale sequence data? What are the most compelling sequence-based community resources that should be generated? What are the consequences of the rapid increase in sequencing capacity, and the rapid decrease in cost, afforded by the new technology platforms?

Rapid changes in sequencing technology offer new opportunities as well as new challenges for NHGRI's sequencing program. As the technology improves, new applications of genome sequencing become possible, including sequencing of multiple human genomes for disease and population studies, and metagenomics. The new technologies also enable other applications, including quantitative transcriptional sequencing and epigenomics.

There are other consequences of the rapid technological change. Most obviously, the ability to apply large scale sequencing efficiently towards a larger number of problems results in significant new demands on the capacity to find samples appropriate to addressing an expanded range (and volume) of questions. To date, the most difficult problem has been obtaining samples for human disease or population studies that are properly consented for the work. One can also foresee increased demands of rigorous and systematic gathering and perhaps storage of samples for larger sets of organisms and for human tissues.

Another consequence will be the demands on informatics infrastructure and data analysis imposed by very large data sets, and the demands for new tools to enable the wider community to use the data.

Finally, there is a sociological consequence: as sequencing technology improves sequencing will become a significantly more dispersed activity. This change will create new challenges for sample collection, data analysis, and informatics infrastructure. This sociological change is an important consideration in the Institute's planning for the future of the sequencing program.

Within this context, the agenda for this workshop is designed to provide the Institute with guidance about how it should continue support for large-scale sequencing over the next five year period.

Specific Questions to Animate the Discussion

What kinds of projects are important for the NHGRI sequencing program to do?
What will be the "mission" and unique role of the NHGRI program?

- What are the next generation of important sequence-based community resource projects? What kind of biomedical and biological questions should motivate the program?
- One important goal of the next iteration of the large-scale sequencing program should be to make genome projects as easy to do as possible. How should the Institute accomplish that?
- How should the "upstream" issues be addressed in centralized and decentralized sequencing program scenarios?
- How should the informatics needs for dealing with sequence information be addressed? What are the considerations for both large-scale, more centralized, and small-scale, more decentralized futures of the program? What is NHGRI's unique role in this?

Balance of centralization and decentralization within a future sequencing program:

- What types of activities would best be undertaken by larger centers with very large-scale sequencing capabilities? What magnitude of scale is needed to carry out these activities? What role should development and dissemination of technologies and computational tools have in larger centers?
- What types of activities would best be undertaken in smaller scale, less centralized centers? What steps could NHGRI take to facilitate wider dissemination of sequencing expertise?
- What are the implications of the balance for:
  -- the biomedical enterprise in general;
  -- the needed infrastructure (especially informatics and analysis downstream; and technology and samples upstream);
  -- technology development? how should those scenarios evolve over time, as technology changes? (2, 5, 10 years?)

Workshop on the Future of the Large-scale Sequencing Program

March 23-24, 2009
Mariott Bethesda North Hotel and Conference Center

Rick Lifton and Barbara Wold, Co-chairs

Monday, March 23, 2009

12:30 pm **Welcome** — Alan Guttmacher

12:40 pm **Introduction, scope, and deliverables** — Rick Lifton, Barbara Wold, Adam Felsenfeld

- The changing NHGRI sequencing program: projects and technical capabilities.
- Scope of meeting
- Meeting charge

1:00 pm **"Biological Opportunities for the Future of DNA Sequencing"**

The talks in this session are intended to encourage discussion about the most significant and important biomedical work that requires sequencing, and to touch on some of the organizational and logistical issues that they raise (20 minutes presentation; 10 minutes Q&A)

Speakers have been asked to address, as appropriate:

- The ideas and their significance
- How the community will use and benefit from the data
- The approximate capacity (type and amount) needed and cost, and the time horizon for the work being feasible. Is this best done in a centralized, or decentralized way?
- The demands on samples
- The demands on current technologies—do we need next-next gen for these?
- The demands for informatics and analysis
- The best general organization for the project, e.g., is it better centralized, or better as "investigator initiated" research integrated with biology?

1:00 pm	Scientific opportunities in a world of cheap high-quality sequence	Eric Lander
1:30 pm	The most important questions for building a scientific foundation for medical sequencing	David Altshuler
2:00 pm	The five most compelling large-scale medical sequencing projects	Rick Lifton
2:30 pm	**Break**	
2:50 pm	Infectious diseases-pathogens, vaccines, vectors, host factors	Jane Carlton
3:20 pm	Organismal biology: comparative genomics, genome evolution, model organisms	Bob Waterston
3:50 pm	What type of genome (large-scale?) sequencing do we need to enable personal genomics?	Francis Collins
4:20 pm	How should new technologies transform the culture of genome sequencing?	Ron Davis
4:50 pm	**Break**	
5:10 pm	**General discussion**	Rick Lifton, Barbara Wold, NHGRI staff

- What other large, compelling projects are appropriate?
- What not-so-large sequencing activities are still important to be done as part of the NHGRI program?

It is important to remember the following points during the discussion:

- NHGRI will not be able to do everything. Sequencing should be increasingly done as part of regular biology. How do we convince the community?
- By being very ambitious, we risk creating a gap of highly important but "smaller scale" work. How do we avoid that gap?
- Upstream and downstream issues (samples and computation) are myriad
- How do we get the proper community input on appropriate targets?

6:00 pm **Plenary talk: Computational challenges** Sean Eddy

- NHGRI predicts that genome sequence output from just the current international large-scale public genome centers could exceed 40 Tb/year in the next two years with no increase in funding. The future volume and utility of genome sequence information raises multiple challenges for computational infrastructure for biology. Many of the challenges are not specific to sequence data, and apply to other data types (imaging, medical records, integration of multiple data types). Without addressing these very significant challenges, the practical utility of generating very cheap sequence data will not be fully realized.
    - Demands on storage of information. How will public research keep up? Will we have to choose what data to keep? How?
    - Demands on moving information over the internet. The 1000 Genomes project alone produces data in sufficient amounts that it can overwhelm available internet bandwidth. How will NIH provide for its scientific community?
    - Computational analysis of the data. What infrastructure will be needed? This includes computational resources (databases, user centers, distributed computing) and also robust computational tools.
    - Quality standards for data. As increasing amounts of data are produced by more and more laboratories, what should be done to help insure that large datasets are of high enough (or well-enough defined) quality to be interoperable?
    - Integration with other data types. In the near future, sequence and medical records data will be integrated for both clinical and research applications. What are the computational infrastructure needs? What are the needs for a system of reliable and secure access to such data?

7:00 pm **Adjourn**

Tuesday, March 24, 2009

9:00 am	**Continental Breakfast**	
9:30 am	**Charge to break-out groups**	

10:00 am	**Breakout sessions**
	• Upstream issues: strategic planning for selecting projects; sample coordination; ELSI and consent (Bill Gelbart/Adam Felsenfeld)
	  o What long-term issues are raised for centralized and de-centralized futures?
	  o How can NHGRI be most effective, and what should it encourage, in each scenario?
	• Downstream issues: Informatics and analysis-same questions as above (Sean Eddy/Peter Good)
	• What are the most compelling sequencing projects of the next 5 years? Which are appropriate to a centralized model, and which de-centralized? Is size the only criterion? How should NHGRI best encourage each? (David Valle/Jane Peterson)	

11:30 pm	**Break; lunch; breakout chairs prepare reports**	
1:00 pm	**Breakout reports (15 minutes each), discussion, and synthesis** • What are the most compelling projects that have been identified? Are there others? • What will they require? What should be NHGRI's role/priorities in encouraging them? • What is NHGRI's role in encouraging the upstream (samples) and downstream (analysis and informatics) aspects? • What are the consequences for informatics infrastructure and data deposition? For technology development and cost reduction? For innovation? For other aspects of the sequencing enterprise?	Bill Gelbart, Sean Eddy, David Valle, NHGRI staff

2:00 pm	**Final discussion and recommendations**	Rick Lifton, Barbara Wold

- What should be NHGRI's priorities for the sequencing program? How should NHGRI encourage them?
  - What should the program structure look like? What is the right balance of "large scale sequencing", "small scale sequencing", informatics, analysis, genotyping, etc. that should occur within the program?
  - In a structure with a few, large-scale centers, what is the right general approach to selecting new targets that balances the need to identify the most important questions, innovation, and community input?
  - In a structure that includes a number of smaller groups with sequencing capabilities, how should NHGRI encourage this in the most productive way possible, given that we can't fund everything?
- What overall investment is appropriate, and how (roughly) should it be allocated among project priorities, and between sequencing and other related (samples, analysis, informatics) activities?

3:00 pm **Adjourn**

Appendix 2

National Human Genome Research Institute (NHGRI)
National Institutes of Health

Workshop on the Future of the Large-Scale Sequencing Program

March 23-24, 2009

Bethesda, MD

PARTICIPANT LIST

Goncalo Abecasis, D.Phil.
Associate Professor of Biostatistics
University of Michigan
goncalo@umich.edu

David Altshuler, M.D., Ph.D.
Professor of Genetics and Medicine
Harvard Medical School
Director
Program in Medical and Population Genetics
Broad Institute of MIT and Harvard
Massachusetts General Hospital
altshuler@molbio.mgh.harvard.edu

Deborah Applebaum-Bowden, Ph.D.
Program Director
National Heart, Lung, and Blood Institute, NIH
applebad@nhlbi.nih.gov

Anna D. Barker, Ph.D.
Deputy Director
National Cancer Institute, NIH
barkera@mail.nih.gov

Jim Battey, M.D., Ph.D.
Director
National Institute on Deafness and Other
Communication Disorders, NIH
batteyj@nidcd.nih.gov

Tsegahiwot T. Belachew
Scientific Program Analyst
National Human Genome Research Institute, NIH
belachewt@mail.nih.gov

Leslie Biesecker, M.D.
Chief
Genetic Disease Research Branch
National Human Genome Research Institute, NIH
leslieb@helix.nih.gov

Bruce Birren, Ph.D.
Co-Director
Genome Sequencing and Analysis Program
Broad Institute of MIT and Harvard
bwb@broad.mit.edu

Michael Boehnke, Ph.D.
Professor
University of Michigan
boehnke@umich.edu

Eric Boerwinkle, Ph.D.
Professor and Director
School of Public Health
The University of Texas Health Science Center
at Houston
eric.boerwinkle@uth.tmc.edu

Vivien Bonazzi, Ph.D.
Program Director
Computational Biology and Bioinformatics
National Human Genome Research Institute, NIH
bonazziv@mail.nih.gov

David Botstein, Ph.D.
Anthony B. Evnin Professor of Genomics
Director
Lewis-Sigler Institute
Princeton University
botstein@princeton.edu

Joy Boyer
Senior Program Analyst
Ethical, Legal, and Social Implications Program
National Human Genome Research Institute, NIH
boyerj@mail.nih.gov

Lisa D. Brooks, Ph.D.
Program Director
Genetic Variation
National Human Genome Research Institute, NIH
lisa.brooks@nih.gov

Jane M. Carlton, Ph.D.
Associate Professor
New York University
jane.carlton@nyumc.edu

Stephen J. Chanock, M.D.
Chief
Laboratory of Translation Genomics
National Cancer Institute, NIH
chanocks@mail.nih.gov

Mark Chee, Ph.D.
President and Chief Executive Officer
Prognosys Biosciences, Inc.
mchee@prognosysbio.com

Mildred Cho, Ph.D.
Associate Professor
Stanford University
micho@stanford.edu

Deanna M. Church, Ph.D.
National Center for Biotechnology Information
National Library of Medicine, NIH
church@ncbi.nlm.nih.gov

George Church, Ph.D.
Professor
Harvard University
gmc@mit.edu

Francis Collins, M.D., Ph.D.
Special Volunteer
National Human Genome Research Institute, NIH
4rancis@mial.nih.gov

Carolyn Compton, M.D., Ph.D.
Director, Office of Biorepositories and Biospecimen Research
National Cancer Institute, NIH
comptcar@mail.nih.gov

Ronald W. Davis, Ph.D.
Professor
Department of Biochemistry and Genetics
Director
Stanford Genome Technology Center
Stanford University
dbowe@stanford.edu

Camilla Day, Ph.D.
Scientific Review Officer and Executive Director, CIDR Board of Governors
Center for Inherited Disease Research (CIDR)
National Human Genome Research Institute, NIH
dayc@mail.nih.gov

Daniel Drell, Ph.D.
Biologist
U.S. Department of Energy
daniel.drell@science.doe.gov

Sean R. Eddy, Ph.D.
Group Leader
Howard Hughes Medical Institute
eddys@janelia.hhmi.org

Evan Eichler, Ph.D.
Professor
University of Washington
eee@gs.washington.edu

Adam Felsenfeld, Ph.D.
Program Director
National Human Genome Research Institute, NIH
felsenfa@mail.nih.gov

Paul Flicek, D.Sc.
Team Leader
European Bioinformatics Institute
flicek@ebi.ac.uk

Kelly A. Frazer, Ph.D.
Director of Genomics
The Scripps Research Institute
frazer.kelly@scrippshealth.org

Nelson B. Freimer, M.D.
Professor
University of California, Los Angeles
nfreimer@mednet.ucla.edu

Stephen H. Friend, M.D., Ph.D.
Senior Vice President and Franchise Head
Merck & Company, Inc.
stephen_friend@merck.com

Stacey Gabriel, Ph.D.
Director
Genetic Analysis Platform
Broad Institute of MIT and Harvard
stacey@broad.mit.edu

Weiniu Gan, Ph.D.
Program Director
National Heart, Lung, and Blood Institute, NIH
ganw2@nhlbi.nih.gov

Susan Garges, Ph.D.
Program Director
Human Microbiome Project
National Human Genome Research Institute, NIH
sgarges@mail.nih.gov

William M. Gelbart, Ph.D.
Professor
Harvard University
gelbart@morgan.harvard.edu

Daniela S. Gerhard, Ph.D.
Director
Office of Cancer Genomics
National Cancer Institute, NIH
gerhardd@mail.nih.gov

Richard A. Gibbs, Ph.D.
Director
Human Genome Sequencing Center
Baylor College of Medicine
agibbs@bcm.edu

Geoffrey S. Ginsburg, M.D., Ph.D.
Director
Center for Genome Medicine
Duke University
geoffrey.ginsburg@duke.edu

Maria Y. Giovanni, Ph.D.
Assistant Director for Microbial Genomics and Advanced Technologies
National Institute of Allergy and Infectious Diseases, NIH
mgiovanni@niaid.nih.gov

Peter Good, Ph.D.
Program Director
National Human Genome Research Institute, NIH
goodp@mail.nih.gov

Bettie Graham, Ph.D.
Associate Director
Division of Extramural Research
National Human Genome Research Institute, NIH
grahamb@mail.nih.gov

Eric Green, M.D., Ph.D.
Scientific Director
National Human Genome Research Institute, NIH
egreen@nhgri.nih.gov

Alan E. Guttmacher, M.D.
Acting Director
National Human Genome Research Institute, NIH
guttmach@mail.nih.gov

Mark Guyer, Ph.D.
Director
Division of Extramural Research
National Human Genome Research Institute, NIH
guyerm@exchange.nih.gov

John Hogenesch, Ph.D.
Associate Professor
Department of Pharmacology
Associate Director
Penn Genome Frontiers Institute
University of Pennsylvania
hogenesc@mail.med.upenn.edu

Peter W.H. Holland, Ph.D., D.Sc.
Professor
University of Oxford
peter.holland@zoo.ox.ac.uk

Bruce Howard, M.D.
Head
Section on Human Genetics
National Institute of Child Health and Human Development, NIH
howard@helix.nih.gov

Timothy Hubbard, Ph.D.
Head
Department of Informatics
Wellcome Trust Sanger Institute
th@sanger.ac.uk

Michael W. Hunkapiller, Ph.D.
General Partner
Alloy Ventures
mwhunk@alloyventures.com

Thomas Insel, M.D.
Director
National Institute of Mental Health, NIH
insel@mail.nih.gov

David Jaffe, Ph.D.
Director of Computational Research and Development
Broad Institute of MIT and Harvard
jaffe@broad.mit.edu

Christopher Juenger, M.S.
Scientific Program Analyst
National Human Genome Research Institute, NIH
juengerc@mail.nih.gov

Karen Kennedy, Ph.D.
Head of Strategic Relations
Wellcome Trust Sanger Institute
kk6@sanger.ac.uk

Jim Kent, Ph.D.
Research Scientist
University of California, Santa Cruz
kent@soe.ucsc.edu

Jean-Pierre Kocher, Ph.D.
Chair
Division of Biomedical Statistics and Informatics
Assistant Professor of Medical Informatics
Mayo Clinic
kocher.jeanpierre@mayo.edu

Leonid Kruglyak
Professor of Genomics and Evolutionary Biology
Princeton University
leonid@genomics.princeton.edu

Eric S. Lander, Ph.D.
Director
Broad Institute of MIT and Harvard
lander@broad.mit.edu

Charles Lee, Ph.D.
Associate Professor
Harvard Medical School
Brigham and Women's Hospital
clee@rics.bwh.harvard.edu

Thomas Lehner, Ph.D., M.P.H.
Chief
Genomics Research Branch
National Institute of Mental Health, NIH
tlehner@mail.nih.gov

Ann Lichens-Park, Ph.D.
National Program Leader
U.S. Department of Agriculture
apark@csrees.usda.gov

Richard P. Lifton, M.D., Ph.D.
Sterling Professor and Chairman
Department of Genetics
Yale University School of Medicine
richard.lifton@yale.edu

Kerstin Lindblad-Toh, Ph.D.
Director
Vertebrate Sequencing and Analysis
Broad Institute of MIT and Harvard
kersli@broad.mit.edu

David J. Lipman, M.D.
Director
National Center for Biotechnology Information
National Library of Medicine, NIH
d12a@nih.gov

Teri Manolio, M.D., Ph.D.
Director
Office of Population Genomics
National Human Genome Research Institute, NIH
manolio@nih.gov

Elaine Mardis, Ph.D.
Co-Director
The Genome Center
Washington University Medical School
emardis@wustl.edu

Elliott H. Margulies, Ph.D.
Head
Genome Informatics Section
Investigator
Genome Technology Branch
National Human Genome Research Institute, NIH
elliott@nhgri.nih.gov

Jean E. McEwen, Ph.D., J.D.
Program Director
Ethical, Legal, and Social Implications
National Human Genome Research Institute, NIH
jm522n@nih.gov

Pamela McInnes, D.D.S., M.Sc.
Director
Division of Extramural Research
National Institute of Dental and Craniofacial Research, NIH
pmcinnes@nidcr.nih.gov

Keith McKenney, Ph.D.
Scientific Review Officer
National Human Genome Research Institute, NIH
mckenneyk@mail.nih.gov

Catherine McKeon, Ph.D.
Senior Advisor for Genetic Research
National Institute of Diabetes and Digestive and Kidney Diseases, NIH
mckeonc@mail.nih.gov

John D. McPherson, Ph.D.
Director and Senior Principal Investigator
Department of Cancer Genomics
Ontario Institute for Cancer Research
john.mcpherson@oicr.on.ca

Alan M. Michelson, M.D., Ph.D.
Associate Director for Basic Research
Office of the Director
National Heart, Lung, and Blood Institute, NIH
michelsonam@nhlbi.nih.gov

Patrice M. Milos, Ph.D.
Vice President and Chief Scientific Officer
Helicos BioSciences Corporation
pmilos@helicosbio.com

Anika Mirick
Scientific Program Analyst
National Human Genome Research Institute, NIH
miricka@mail.nih.gov

Michael J. Morgan, Ph.D.
Chief Scientific Officer
Genome Canada
mmorgan@genomecanada.ca

Jim Mullikin, Ph.D.
Associate Investigator
Genome Technology Branch
National Human Genome Research Institute, NIH
mullikin@mail.nih.gov

Donna Muzny, M.Sc.
Director of Operations
Human Genome Sequencing Center
Baylor College of Medicine
donnam@bcm.edu

Richard M. Myers, Ph.D.
Director
HudsonAlpha Institute for Biotechnology
rmyers@hudsonalpha.org

Deborah Nickerson, Ph.D.
Professor
University of Washington
debnick@u.washington.edu

John E. Niederhuber, M.D.
Director
National Cancer Institute, NIH
niederhj@mail.nih.gov

Chad Nusbaum, Ph.D.
Co-Director
Genome Sequencing and Analysis Program
Broad Institute of MIT and Harvard
chad@broad.mit.edu

Christopher J. O'Donnell, M.D., M.P.H.
Senior Advisor to the Director for Genome Research
Framingham Heart Study
odonnellc@nhlbi.nih.gov

Susan E. Old, Ph.D.
Senior Advisor for Translational Research
National Center for Research Resources, NIH
oldse@mail.nih.gov

Pilar N. Ossorio, Ph.D., J.D.
Professor
School of Law
University of Wisconsin-Madison
pnossorio@wisc.edu

James M. Ostell, Ph.D.
Chief
Information Engineering Branch
National Center for Biotechnology Information
National Library of Medicine, NIH
ostell@ncbi.nlm.nih.gov

Francis Ouellette
Associate Director, Informatics and Biocomputing
Principal Investigator
Ontario Institute for Cancer Research
francis@oicr.on.ca

Brad Ozenberger, Ph.D.
Program Director
National Human Genome Research Institute, NIH
bozenberger@mail.nih.gov

David C. Page, M.D.
Director
Whitehead Institute
Professor of Biology
Massachusetts Institute of Technology
Investigator
Howard Hughes Medical Institute
page_admin@wi.mit.edu

Abbas Parsian, Ph.D.
Program Director
National Institute of Alcohol Abuse and Alcoholism, NIH
parsiana@mail.nih.gov

Len Pennacchio, Ph.D.
Senior Staff Scientist
Lawrence Berkeley National Laboratory
lapennacchio@lbl.gov

Jane Peterson, Ph.D.
Associate Director
Division of Extramural Research
Program Director
Large-Scale Sequencing Program
National Human Genome Research Institute, NIH
jane_peterson@nih.gov

Jonathan Pollock, Ph.D.
Chief
Genetics and Molecular Neurobiology Research Branch
National Institute on Drug Abuse, NIH
jpollock@mail.nih.gov

Stephen M. Prescott, M.D.
President
Oklahoma Medical Research Foundation
steve-prescott@omrf.org

Harold Riethman, Ph.D.
Associate Professor
The Wistar Institute
riethman@wistar.org

Jeff Rogers, Ph.D.
Associate Professor
Baylor College of Medicine
jr13@bcm.tmc.edu

Daniel Rokhsar, Ph.D.
Program Head, Computational Genomics
Lawrence Berkeley National Laboratory
U.S. Department of Energy
dsrokhsar@lbl.gov

Carl Schaefer, Ph.D.
Director of Biomedical Informatics
Center for Biomedical Informatics and Information Technology
National Cancer Institute, NIH
schaefec@mail.nih.gov

Jeffery A. Schloss, Ph.D.
Program Director
Technology Development Coordination
National Human Genome Research Institute, NIH
schlossj@exchange.nih.gov

Stephan C. Schuster, Ph.D.
Professor
The Pennsylvania State University
scs@bx.psu.edu

Stephen Sherry, Ph.D.
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine, NIH
sherry@ncbi.nlm.nih.gov

Martin Shumway, M.Sc.
Staff Scientist
National Center for Biotechnology Information
National Library of Medicine, NIH
shmwaym@ncbi.nlm.nih.gov

Arend Sidow, Ph.D.
Stanford University
arend@stanford.edu

Adam Siepel, Ph.D.
Assistant Professor
Cornell University
acs4@cornell.edu

Jane Silverthorne, Ph.D.
Deputy Director
National Science Foundation
jsilvert@nsf.gov

Douglas R. Smith, Ph.D.
Director
Science and Technology
Agencourt Bioscience Corporation
dsmith@agencourt.com

Geoff Spencer
Public Affairs Specialist
National Human Genome Research Institute, NIH
spencerg@mail.nih.gov

John Stamatoyannopoulos, M.D.
Assistant Professor
University of Washington
jstam@u.washington.edu

Paul Sternberg, Ph.D.
Professor of Biology
California Institute of Technology
pws@caltech.edu

Marvin Stodolsky, Ph.D.
U.S. Department of Energy
marvin.stodolsky@science.doe.gov

Robert L. Strausberg, Ph.D.
Deputy Director
Department of Genome Medicine
J. Craig Venter Institute
rls@jcvi.org

Jeff Struewing, M.D., M.S.
Program Director
National Human Genome Research Institute, NIH
struewij@mail.nih.gov

Danilo Tagle, Ph.D.
Program Director
National Institute of Neurological Disorders and Stroke, NIH
tagled@mail.nih.gov

Gilles Thomas, Ph.D.
Professor
Centre Leon Berard
Fondation Synergie Lyon Cancer
gd.thomas@me.com

Larry Thompson, M.S., M.F.A.
Chief
Communications and Public Liaison Branch
National Human Genome Research Institute, NIH
thompsl@mail.nih.gov

David Valle, M.D.
Professor and Director
Institute of Genetic Medicine
Johns Hopkins University
dvalle@jhmi.edu

Joseph Vockley, Ph.D.
Scientific Projects Director
National Cancer Institute, NIH
vockleyj@mail.nih.gov

Lu Wang, Ph.D.
Program Director
National Human Genome Research Institute, NIH
wanglu@mail.nih.gov

Robert H. Waterston, M.D., Ph.D.
William H. Gates Endowed Chair in
Biomedical Sciences and Professor
University of Washington
waterston@gs.washington.edu

Richard M. Weinshilboum, M.D.
Professor
Mayo Clinic
weinshilboum.richard@mayo.edu

George Weinstock, Ph.D.
Professor of Genetics and Molecular Microbiology
Washington University
geowei@mac.com

Chris Wellington
Scientific Program Analyst
National Human Genome Research Institute, NIH
wellingtonc@mail.nih.gov

Kris Wetterstrand, M.S.
Senior Program Analyst
National Human Genome Research Institute, NIH
wettersk@mail.nih.gov

David A. Wheeler, Ph.D.
Director
Department of Bioinformatics and Cancer Genomics
Baylor College of Medicine
wheeler@bcm.tmc.edu

Owen White, Ph.D.
Director of Bioinformatics
Institute for Genome Sciences
University of Maryland, Baltimore
owhite@som.umaryland.edu

Alan Williamson, Ph.D.
arwill37@btinternet.com

Rick Wilson, Ph.D.
Director
The Genome Center
Washington University Medical School
rwilson@wustl.edu

Barbara J. Wold, Ph.D.
Bren Professor of Molecular Biology
California Institute of Technology
Director
Beckman Institute
woldb@caltech.edu

Liming Yang, Ph.D.
Health Scientist Administrator
Division of Biomedical Technology
National Center for Research Resources, NIH
lyang@mail.nih.gov

Yin Yao, Ph.D.
Chief
Genetic and Molecular Epidemiology of Mental Disorders Program
National Institute of Mental Health, NIH
kay1yao@mail.nih.gov

Julia Zhang
Scientific Program Analyst
National Human Genome Research Institute, NIH
zhangjul@mail.nih.gov