Disease 2020

Discovering the genetic basis of human disease as a foundation for genomic medicine

May 10, 2013

The recommendations in this document represent the conclusions of a National Human Genome Research Institute (NHGRI) workshop held on January 30-31, 2013 to discuss a new strategic framework for the NHGRI Sequencing Program.

These conclusions were further refined over a period of several months based on questions from meeting participants, NHGRI leadership, external scientific advisors to the NHGRI Sequencing Program, and the National Council on Human Genome Research.

The conclusions are presented here as a set of recommendations to NHGRI from the invited meeting participants (see Appendix 1: workshop attendees).

A fundamental goal of human genetics is to discover the genes underlying human diseases, including the inherited genomic variations underlying simple Mendelian disorders and more complex oligogenic and polygenic common diseases and disease-related phenotypes; and the somatic genomic changes underlying cancer. In addition, the tools of human genomics can be used to gather information about the microbes that underlie infectious disease and co-inhabit our bodies. We also must learn how to use this information in clinical practice, to guide the development and application of prevention, diagnosis and therapy.

The past decade has seen (i) enormous advances in genomic technology, increasing sequencing power by nearly one-million-fold; (ii) development of new paradigms for gene discovery, including association studies to discover common and rare variant that predispose to disease, cancer genome studies and microbiome analysis; and (iii) initial steps to apply these findings in clinical settings, with the expectation that these will expand rapidly.

These advances have led to the identification of many genes and pathways that play a role in the basis of disease. However, we are still on the steep part of the learning curve: our knowledge remains far from complete.

It is now time to use the tools that have been developed to systematically define the genetic basis of human disease. The goal should be to create the knowledge needed to maximize the impact of genomic medicine. Achieving this goal will require dramatically accelerating the basic science research aimed at discovering and understanding the mechanisms of health and disease, to unleash efforts to develop better treatment and prevention strategies for disease.

We believe that laying the foundation for achieving this goal can be accomplished by the year 2020. **We refer to this project as Disease 2020**.

Disease 2020 will require a concerted effort, involving biomedical investigators and centers across the world -- with strong leadership among NIH Institutes, as well as partner agencies in other countries. Given its particular expertise and mission, NHGRI has a special responsibility to play in the leadership of Disease 2020, but significant progress can only be made if Disease 2020 is a collaboration with other NIH institutes and their scientific communities.

Disease 2020 is intended as a program for all Americans and all people. The incidence of human disease varies across subsets of the population, with specific diseases disproportionately affecting some groups more than others. These variations are not limited to self-reported race and skin color, but span the gamut of social and economic groups. Disease 2020 should also include all age groups from newborns to the elderly.

Disease 2020 will accelerate biomedical research in multiple ways.

First, Disease 2020 will reveal disease genes and pathways for many diseases and disease-related phenotypes.

Second, Disease 2020 will teach us a tremendous amount about the range of study designs needed to understand human disease. No single design is optimal. Rather a portfolio of complementary approaches will be required:

- Case-control designs have proven powerful for identifying genes related to clear disease endpoints.
- Clinic-based sample collections have advantages for studying rare events, such as rare diseases and responses to treatment, and for exploring the potential use of sequencing results at the point of care.
- Cohort studies with deep phenotyping of samples are a cost-effective way to identify genes affecting quantitative risk factors and epidemiological parameters, such as penetrance.
- Family-based studies are valuable for evaluating the impact of specific rare variants whose occurrence may be limited to a particular clan or lineage.

Third, Disease 2020 will propel our ability to interpret the entire human genome in health and disease. Current efforts are focused on analyzing the exome (or an "extended exome", including some non-coding conserved regions). While it is premature today to abandon exome-focused sequencing, an orderly transition to whole genome sequencing is inevitable. Today's focus on genes will expand to include genomic regions with currently unknown function, informed by emerging knowledge-bases related to the function of the non-protein encoding regions of the genome. Disease 2020 will thus help annotate the morbid anatomy of the non-protein-encoding portion of the human genome.

Fourth, Disease 2020 is only possible today because of past advances in genomic technologies and methods. The success of Disease 2020 will require continuing investment to propel such advances. For example, we need better methods for analyzing genomic sequence – including ways to improve the assembly of genome sequence, to achieve perfect identification of all types of alterations (from single base changes to small insertions/deletions to larger structural variants) and to annotate and interpret non-coding RNA and regulatory variants. We also need improved infrastructure for efficiently storing and sharing very large amounts of genomic and phenotypic information.

Disease 2020 will require studying very large numbers of human samples with accompanying high quality phenotype information. Yet, it is tractable, in significant part because of the ability to leverage previous investments in large case-control consortia, cohort studies and clinic-based sample collections. Even in the relatively simpler case of Mendelian diseases, large numbers of patients and families have proven advantageous for identifying novel disease genes and studying the extent of phenotypic variation for individuals carrying particular gene variants. In the case of diseases with more complex genetics, much larger sample sets are required.

The NHGRI’s Sequencing Network (http://www.genome.gov/10001691) will have a special role to play in Disease 2020.

The Large-Scale Sequencing and Analysis Centers (LSAC), because of their size and experiences in large-scale coordinated efforts, are well suited to large-scale data generation, to lead the implementation of new more-efficient laboratory methods, to lead the transition to whole genome sequencing and, importantly, to develop new analysis strategies that will be essential to the success of the program. (The LSAC’s are expected to devote 80% of their effort to the Disease 2020 program, with the remainder going to exploratory projects in other directions and the development of new technological capabilities.)

The other components of the NHGRI Sequencing Network also have critical roles to play. These include the Centers for Mendelian Genomics (CMG), which focus on uncovering Mendelian disease variants; the Clinical Sequencing Exploratory Research Centers, which bring together clinicians, bioinformaticians, and ethicists to understand and address the challenges of utilizing genomic sequence data in the clinic; and Genome Sequencing Informatics Tools, which aim to broaden access by the scientific community to sequence analysis. In these areas, these program components will lead efforts, with the LSAC contributing as appropriate.

In the sections below, we expand on four domains of Disease 2020: common disease, Mendelian disease, cancer and microbiome. We close with a section on “Genomics in the Health Care Setting”. The section delineations are largely for organizational purposes. We recognize that there are important areas of overlap and complementarity across the sections. For example, Mendelian subtypes of common diseases have proven particularly informative. Similarly, cancer is influenced by both inherited genomic variation as well as somatic alterations.

Table of Contents:

1. Inherited basis of disease:
Identify the key genetic factors contributing to 100 important common diseases

Discovering the genes underlying common diseases is critical to biomedical progress because it (i) reveals the biological mechanism of diseases in a systematic and unbiased manner, (ii) generates a molecular taxonomy for human diseases, providing markers that can be used for prediction and prognosis, and (iii) it identifies pathways containing novel targets of treatment or prevention.

The goal for Disease 2020 should be to identify the key genes harboring inherited genetic variation that affects disease risk for 100 important common diseases. Some of the most prominent examples include the leading causes of morbidity and mortality, including cardiovascular disease, cancer, diabetes, psychiatric diseases and neurodegenerative diseases. The 100 common diseases to be studied should be chosen based on a combination of overall medical impact (including such factors as prevalence, morbidity and mortality) and the likelihood of identifying important genetic contributions. Diseases to be studied may thus include diseases that affect millions of people, as well as devastating diseases that affect smaller numbers of people.

While Disease 2020 should focus primarily on disease, it should also examine disease-related phenotypes (such quantitative risk factors such as body-mass index), responses to therapy (such as pharmacogenomics); important endophenotypes (such as the blood metabolome), which may have a simpler genetic basis; and phenotypes related to drug response, including severe adverse events observed in clinical trials and clinical use.

Geneticists have made tremendous progress in discovering inherited variants underlying common diseases by using common variant association studies using genotyping methods to systematically study genetic variants with frequencies exceeding 5-10%. Such studies have already discovered >2000 loci associated with >200 diseases and traits, with progress continuing at a rapid pace (http://www.genome.gov/gwastudies/). For example, they have discovered >100 loci associated with inflammatory bowel disease, > 100 loci associated with lipid levels and > 90 loci associated with schizophrenia; these discoveries have revealed important new biology and pathways of disease etiology. [Note: Association studies involving common variants are typically called *genome-wide association studies* or *GWAS*. However, rare-variants studies also involve association analysis across the genome. We therefore refer below to common variant association studies (CVAS) and rare variant association studies (RVAS). As currently practiced, CVAS captures variation across the entire genome, whereas RVAS is focused on exomes and thus misses non-coding variation.]

Yet, the discoveries so far appear to explain only a minority of the “heritability” of these diseases. The remaining heritability is likely to be due in large part to contributions from additional variants – ranging from common to extremely rare in frequency.

With massively parallel sequencing, it is now possible to directly assay genetic variation at all frequencies. Importantly, new methods are needed to analyze and interpret rare genetic variants. (While there are many rare variants in the population, they must be aggregated together for study. This can be challenging, because one must ensure that rare causal variants are not swamped by a larger number of rare neutral variants.)

Moreover, the discoveries have been largely focused on European populations. The spectrum of genetic variation differs across populations, owing to both drift and selection. It will be important to expand studies to include a wider range of ethnic diversity, and to develop analytic methods to extract information from a range of populations.

Current evidence suggests that different diseases have different genetic architectures – including (i) the number of loci involved (genetic heterogeneity) and (ii) the balance between common and rare variants (frequency spectrum), which likely reflects the strength of selection. For example:

- Autism and schizophrenia appear to involve high locus heterogeneity and many extremely rare genetic variants, including a significant proportion of *de novo* mutations found in children that were not present in their parents.
- Type 2 diabetes, hyperlipidemias, hypertension, early myocardial infarction and inflammatory bowel disease appear to involve important roles for both common and rare genetic variants.
- Celiac disease appears to be largely attributable to common polymorphic variation.

We need to develop reliable methodologies to elucidate the genetic basis of disease across the full range of genetic architectures, and we need to validate them by applying them to a set of representative diseases at appropriate scale.

Common variants. Common-variant association studies (CVAS) have already localized thousands of disease-associated genetic variants to defined regions, but the causative genes and specific DNA variants responsible have been identified only in a small percentage of cases.

It will be critical to identify the causal genes in regions already implicated by CVAS -- especially where there is no obvious candidate gene. This will require extensive sequencing of large numbers of samples with replication in still larger samples. For this purpose, we may need to develop cost-effective methods for targeted sequencing, suitable for use on tens of thousands of samples.

In addition, CVAS need to be expanded to larger and more diverse populations and to the study of lower-frequency common variants (0.5 - 5%).

Rare variants. Systematic rare-variant association studies (RVAS) need to be undertaken. There are currently no published studies that have successfully performed exome-wide searches (let alone genome-wide searches) to definitively identify specific genes harboring rare variants associated with common complex disease. Discoveries to date have been based on candidate-gene studies.

Early hopes that it might be possible to identify accumulation of rare variants in small samples have given way to the recognition that very large samples will be required – with discovery samples likely exceeding 5,000 cases and with replication samples at a scale of perhaps 10-fold higher.

The fundamental methodologies for identifying genes based on rare variants remains in flux. Many methods have been proposed for analyzing coding regions based on their aggregate burden of rare variants, but they have yet to be extensively tested. No methodologies currently exist for analyzing the remaining ~99% of the genome, because it is unclear how to aggregate variants into sets that can be tested. Successful large-scale studies will be needed to solidify such methodology so that it can be systematically applied to hundreds of diseases.

More complex genetic mechanisms. The role of genetic interactions in human disease is often debated, especially in the context of “missing” heritability. Studies of gene x gene interaction will require even larger sample sizes than for discovery of loci. Studies of gene x environment interaction will require additional progress in measuring relevant environmental exposures.

Recontact and Expanded Characterization. Once genes have been identified as associated with a disease, it will be necessary to perform in-depth phenotyping of individuals carrying mutations in specific genes to understand their physiological consequences at a deeper level than can be done in initial characterization of a case-control or cohort study. This will require the ability and resources to recontact study participants (subject to their prior consent), and to carry out deeper phenotypic characterization, such as metabolic studies or functional imaging.

The ability to recontact study participants offers the ability to undertake a different kind of genetic study: genotype-to-phenotype (G2P) studies, in which very large numbers of participants are sequenced to select individuals for study based on their genotype. Individuals carrying clear loss-of-function mutations (in a heterozygous or homozygous state) can then be carefully phenotyped to provide insight into the null phenotype of uncharacterized genes. The optimal design for G2P cohorts remains to be worked out. Studies in a variety of populations could be valuable, including populations with high degrees of consanguinity, to enrich for the occurrence of rare variants in the homozygous state.

Pilot projects will be needed to explore how best to recontact and phenotype targeted study participants – considering the scientific, logistic and ELSI issues.

Plan for Common Diseases

As a foundation for the larger Disease 2020 project, the Large-Scale Sequencing Centers will initially focus on a limited number (5-10) of diseases with the aim of achieving informative well-powered studies providing clear answers. Diseases will be chosen based on the following criteria: (i) overall medical impact (including such factors as prevalence, morbidity and mortality), (ii) likelihood of success, based on expected genetic architecture; (iii) availability of large numbers of appropriately consented high-quality samples and existing high-quality deep phenotypic data; (iv) engagement of effective communities of disease researchers; and (v) the offering of important analytical and/or technical challenges.

These projects must be large, well-powered, full-scale demonstration projects designed to explore the power of the approach and to provide critical information that will help guide the selection and design of future projects.

The initial projects should be important diseases where it is possible to obtain >10,000 clearly affected cases, such as: diabetes, early-onset myocardial infarction, inflammatory bowel disease, several kinds of cancers, schizophrenia, autism, bipolar affective disorder, Alzheimer’s disease, neurodegenerative disease, chronic kidney disease, and specific birth defects. It is also important to include projects to identify quantitative measures (such as metabolic or blood parameters or protection against diseases) in existing large cohorts. The samples should ideally come from multiple ethnic groups.

Projects will aim to identify genomic regions and variants underlying disease using an appropriate mix of studies including:

(1.1) large-scale exome sequencing projects to identify rare variants, as well as the development and testing of approaches to analyze whole-genome sequence,

(1.2) expanded discovery of common variants, through exome chips or low-pass sequencing,

(1.3) targeted sequencing of high priority (based on GWAS or *a priori* biologic information) genomic regions for follow-up in large sample sizes of appropriately phenotyped individuals.

(1.4) pilot studies in recall study participants in a genotype-specific manner for more detailed phenotypic characterization.

The projects will aim to produce both biological discoveries and methodological advances.

2. Inherited basis of disease: Identify the genes underlying essentially all Mendelian diseases

Disease 2020 should aim to identify the genes underlying essentially all Mendelian diseases, and develop procedures to readily recognize the presence of disease-causing mutations in these genes in clinical settings.

Discovering the genes underlying Mendelian diseases will provide critical information to (i) aid in treating patients with these diseases, (ii) shed light on related diseases and biological mechanisms in the general population, and (iii) may provide insights into the genetic basis of common disorders.

Geneticists have already made remarkable progress in revealing the genes underlying >3000 monogenic disorders, leading to treatments or therapeutic hypotheses for some disorders and shed extraordinary light on biology more generally. Yet, the catalog of known Mendelian disorders contains >1500 additional diseases. In addition, many more conditions caused by null mutations are likely currently unrecognized. It is time to complete this critical catalog for medicine by systematic sequencing of patients with Mendelian disorders and by using appropriate population samples.

Clear Mendelian conditions. For some well-defined Mendelian conditions in which the full set of causative genes has not been identified, there exist collections (of varying size) of patients and/or families. In some cases, the identification of the disease gene(s) will be straightforward. In other cases, the challenges include recognizing mutations in non-coding regions and overcoming substantial genetic heterogeneity.

The NHGRI’s Centers for Mendelian Genomics (CMG’s) are ideally suited for these studies, which require close clinical focus on the phenotype to be successful. This work will be the primary challenge for these Centers.

Uncertain modes of inheritance. For some phenotypes, the nature of inheritance is unclear -- although there is reason to suspect a role for some highly penetrant alleles (which may be *de novo* events in some cases). Examples include certain congenital neurodevelopmental disorders, certain blood disorders, spina bifada and cleft-lip and palate. The understanding of such phenotypes falls in a gray zone between common disease and Mendelian disease.

If the phenotype is due to highly penetrant alleles in a handful of genes, it may be possible to identify genes from analyses of modest sizes. In other cases, larger sample sizes and new analytical methods may be needed. Identifying the genes underlying recessive phenotypes may benefit from large-scale ascertainment in populations with higher rates of consanguinity.

Collaboration between the LSAC's and the CMG's will be needed to identify the most productive cases to study and to develop effective study designs.

G2P studies in recalled study participants. As described in the section on common disease above, genotype-to-phenotype studies in recalled study participants can potentially be used to identify the phenotypes of individuals who are homozygous or heterozygous null for specific genes. This information can contribute importantly to understanding the full range of Mendelian disease genes.

Here too, collaboration between the LSAC's and the CMG's will be important.

Plan for Mendelian Diseases

As a foundation for the larger Disease 2020 project:

(2.1) Clear Mendelian traits without extreme genetic heterogeneity. The CMG's will take the lead in sequencing and analyzing families and singleton patients, for all such traits.

(2.2) Traits with uncertain inheritance or clear evidence of extreme genetic heterogeneity. Projects will be undertaken on 5-10 such traits, across the CMG's and LSAC's.

(2.2) Genotype-to-phenotype studies. Projects will be taken to characterize the ability to identify the phenotype associated with specific loss-of-function mutations in populations with the potential for recall and with diverse genetic structures. Pilot projects will be undertaken to explore the feasibility of obtaining and phenotyping recalled study participants. The work will be undertaken collaboratively among CMG's, LSAC's and NHGRI staff.

3. Mutational basis of disease:
Identify genes that drive cancer initiation, progression and treatment response in all significant cancer types.

Discovering the genes that underlie cancer is critical to biomedical progress because it will define (i) the appropriate targets and pathways for cancer therapy in each patient, (ii) the mechanisms of acquired resistance (iii) prognostic and predictive markers, (iv) mechanisms underlying metastases; (v) appropriate cell- and animal models for disease studies, and (vi) homogeneous subsets of patients for clinical trials, (vii) high-risk population for targeted cancer screening and prevention.

The goal for Disease 2020 should be to identify the genes that drive cancer initiation, progression and treatment response in all significant cancer types.

Genetic studies have been central to propelling progress in cancer, beginning with early discoveries in the 1980s of oncogenes and tumor suppressor genes that defined pathways for growth factor signaling.

More recently, systematic genomic studies have been substantially broadened our understanding of the genomic basic of cancer. They have expanded the number of genes in classical growth-related pathways, and begun to characterize their frequency in different cancer types. They have also pointed to many new classes of cancer-related genes, including those encoding proteins affecting lineage specification; epigenomic regulation; RNA splicing, protein homeostasis, and other cellular processes. Genomic studies have identified important new therapeutic targets, some of which have already led to approved drugs.

Many of these studies have been undertaken under the auspices of The Cancer Genome Atlas (a long term collaboration between NHGRI and the National Cancer Institute; http://cancergenome.nih.gov/) and of the International Cancer Genome Consortium.

While these discoveries represent tremendous progress, they fall far short of the knowledge base that is need to fully understand cancer biology and guide therapy for cancer patients.

We need to understand the full set and frequency of genomic alterations that play a role in every significant cancer type. Studies to date have been limited in that:

- they have analyzed only modest number of samples from a limited subset of cancer types, and have in no case achieved saturation (which will require at least 2000 samples to overcome the noise from background mutation rates);
- they have not obtained and systematically integrated comprehensive information from somatic alterations in the cancer genome (including point mutations, amplifications,

deletions and rearrangements), transcriptome and methylome, as well as germline variation that may predispose to cancer and affect treatment response;

- they have largely focused on large primary tumors with high proportions of neoplastic cells (in order to have sufficient material to distribute through the TCGA network)
- they have not yet taken advantage of the 10,000s of samples available only as formalin-fixed paraffin-embedded (FFPE) samples, which represent a wider swath of tumors and which may have richer clinical information about treatment response and patient outcome;
- they have not systematically characterized the genomic changes seen in metastases or in response to treatment;
- they have not systematically explored tumor heterogeneity and clonal evolution;
- they have not been systematically characterized important naturally occurring and engineered animal models of cancer (including mouse and dog) and cell-based models of cancer, for which an understanding of the relationship to human cancers at a mechanistic level could provide a powerful tool for therapeutic studies;

In short, there is a tremendous amount of information still to be gained from genomic characterization of cancer. Given the importance of cancer and the demonstrated power of genomic analysis in this disease, our goal should be to obtain a complete genomic picture of this disease.

Small-scale sequencing efforts can make important contributions to this goal, especially to explore specific hypotheses and to characterize unusual patients.

However, the goal of obtaining a comprehensive picture of cancer will require

(i) large scale;
(ii) major technology development (including efficient analysis of FFPE samples and single-cells); and
(iii) development of new analytical methods and tools (such as for accurate determine of mutations and rearrangements, interpretation of heterogeneity, pattern recognition, and determination of statistical significance.) The LSACs thus have a crucial role to play.

The Disease 2020 plan for cancer has been developed with input from The National Cancer Institute (NCI), and will be further developed in full partnership between NCI and NHGRI.

Plan for Cancer

As a foundation for the larger Disease 2020 project, the NHGRI's Large-Scale Sequencing Centers should focus on critical challenges that require large scale; major technology development; and/or major new analytical methods and tools.

The NHGRI Sequencing Program, through the LSAC's, will continue its substantial efforts in cancer sequencing. The largest part of this effort has been through the TCGA. NHGRI intends to continue this productive collaboration in the next few years and, beyond that, to work closely with NCI in identifying future priorities and projects in cancer sequencing to look for opportunities for productive collaboration.

Cancer projects will be chosen based on the following criteria: (i) high medical importance; (ii) likelihood of success; (iii) availability of excellent samples; and (iv) important analytical and/or technical challenges.

Projects will aim to characterize the genomic basis of at least 50 cancer types, by using an appropriate mix of studies:

(3.1) Complete the catalog for primary tumors. Driving characterization toward saturation for targets mutated in at least 1% of tumors of a given type, through increased sample number and use of archived FFPE samples.

(3.2) Progression. Characterize the genomic changes that drive cancer progression and heterogeneity.

(3.3) Resistance and Response to Treatment. Characterize the genomic changes that explain primary and secondary resistance to treatment, or remarkable clinical response.

(3.4) Germline Variation. Discover the germline variants affecting cancer susceptibility, as well as those affecting drug response and pharmacogenomics. (These projects concern inherited variation, and thus formally fit under Section 1 above.)

(3.5) Models. Characterize animal models (mouse and dog) and cell models to understand their relationship with human disease at a mechanistic level, to provide powerful tools for pre-clinical studies.

4. Microbial basis of disease: Identify microbes and their communities that cause and correlate with disease.

There is a growing appreciation of the role of the commensal microbiome in human health and disease.

The effects of the microbiome may be due to a single microbial pathogens; to changes in the balance of beneficial and detrimental organisms; to interaction with host immune, inflammatory, and other pathways by microbes; or to other mechanisms. For example: (i) acne appears to involve specific sub-species of the ubiquitous species *Propionibacterium acnes*, discovered by analyzing large numbers of both full-length 16S rRNA sequences and whole genomes of P. acnes strains; (ii) Antibiotics can reduce beneficial commensal microbes, allowing rare, resistant pathogens such as *Clostridium difficile* to overgrow and cause severe diarrhea; and viruses are widespread in healthy and diseased subjects, but their role is unclear. Even where not causal, an altered microbiome composition resulting from underlying disease may provide an early warning sign.

Gaining a full understanding of the microbial contributions to disease will advance biomedical progress by (i) revealing underlying mechanisms of disease in both acute and chronic disease, (ii) providing new markers for diagnostics and prognosis, and (iii) providing targets for potential therapeutics, including non-antibiotics.

High-throughput methods are now making it possible to recognize the multitudes of taxa present in a microbiome, to recognize the related species within each taxon, and to recognize polymorphic variation within species. Each level of description has an important role to play in the study of disease. Key correlates of disease may involve a species within a taxon. And, polymorphisms within species are an essential tool for public health studies of outbreaks, infection control in hospitals, and treatment of individual infection.

Current studies of the microbiome are limited by numbers and demographics of subjects, by breadth of host anatomical sites examined, by sparse longitudinal data, and by limited information about environmental effects (such as diet).

Disease 2020 should aim (i) to identify and characterize the genomes of all key microbes (prokaryotic, eukaryotic, and viral) in the human body and (ii) it should undertake rigorous and carefully designed pilot projects to explore association between the microbiome and common human diseases.

Rigorous Methodology: Distinguishing cause and effect. The greatest challenge for microbiome studies is to develop and deploy experimental designs from which it will be possible to draw reasonable inferences about the direction of causality. Unlike the situation for inherited genetic variation or somatic variation in cancer, it is likely that many of the microbiome changes observed in disease are the result -- rather than the cause -- of a disease state. Longitudinal studies will be essential for creating plausible hypotheses. In the longer term, interventional studies (in which the microbiome is perturbed in controlled ways) will then be essential for testing hypotheses.

The importance of rigorous methodology cannot be overstated. There are currently no published studies that have demonstrated a definitive association between a microbial community structure and disease. Reports to date have largely been based on limited numbers of subjects, that produced hypotheses but not conclusions. Moreover, early suggestions that it is possible to identify disease-related changes at higher community taxonomic levels (for example, the ratio of the phyla Bacteroidetes to Firmicutes being diagnostic of obesity) in small sample sets have not proven reproducible. The current view is that the situation may be complex, with many taxa involved, including viruses and eukaryotes.

Expanded genomic analysis. Creating systematic catalogs will require deeper sampling. We need to develop more sensitive and reliable methodologies for targeted sequencing (eg 16S rRNA) and for comprehensive sequencing of microbial genomes and transcriptomes. Identifying strain-specific genes will likely require large-scale sequencing of thousands of samples of each species (for example, there are over 90 serotypes of Streptococcus pneumoniae and many distinct disease-causing strains of E. coli). To do this, we will need to develop cost-effective methods for whole-genome sequencing, assembly, and analysis suitable for use on up to 100,000s of samples of both pathogenic and commensal microbes. Efficient single-cell sequencing will likely be an essential tool. Importantly, these projects will all require improved computational tools.

Current consortia. Initial consortium-scale microbiome projects – such as the Human Microbiome Project in the USA and the MetaHIT project of the European Union — have been recently completed. These studies involve hundreds of subjects, sampled on several occasions at multiple body sites. Various NIH ICs are actively engaged in this work. NHGRI should coordinate closely with the other ICs in designing the next round of microbiome studies.

Plan for microbes

As a foundation for the larger Disease 2020 project, the Large-Scale Sequencing and Analysis Centers will focus on pilot efforts in a small number (no more than 3) diseases with the aim of achieving large enough studies to provide clear direction for future study design.

Diseases will be chosen based on the following criteria: (i) high medical importance; (ii) high likelihood of involving a microbial component; (iii) likelihood of success, based on

expected microbial role; (iv) availability of large numbers of excellent samples with longitudinal sampling and well annotated metadata; (v) engagement of effective communities of disease researchers; and (vi) important analytical and/or technical challenges. Where feasible, these projects could be coordinated with projects aimed at studying the inherited basis of these diseases. However, to the extent that new sample sets would need to be identified, phenotyped, and followed over time, this element may only be feasible in the longer-term, and would depend critically on collaboration with other funding bodies.

For these diseases, projects will aim to identify microbial correlates underlying disease, by using an appropriate mix of studies including:

(4.1) large-scale targeted sequencing to identify bacteria and eukaryotes.

(4.2) large-scale shotgun sequencing to identify bacteria, eukaryotes, and viruses and define strain-specific markers.

(4.3) large-scale shotgun sequencing of cDNA to identify gene expression patterns.

(4.4) expanded discovery of rare microbes, and validation of biomarkers, through custom capture chips.

(4.5) targeted sequencing of organism-specific genes or regions for follow up in expanded cohorts (at least 1000s of samples).

(1.6) whole genome sequencing of cultured, pure microbes where possible.

The projects will aim to produce both biological discoveries and methodological advances in data production and analysis.

5. Genomics in clinical and health-care settings

We are still far from understanding the genomic basis of most disease. Accordingly, NHGRI's primary focus must remain on discovering the genes underlying disease.

Nonetheless, it is important to begin pilot studies to explore the use of genomic information in clinical and health-care settings – including in analysis and interpretation; understanding by patients and physicians; ethical and legal issues; and laboratory standardization, turn-around time and cost. This will require demonstration projects involving both healthy and diseased individuals focused on diagnostics, pharmacogenetics, and primary prevention.

The Clinical Sequencing Exploratory Research (CSER) Centers were conceived to play a lead role in these studies, and will continue to play this role. Each CSER group is comprised of active clinicians, genome analysis researchers, decision-support tool developers, and ethics/psychosocial implications experts. The CSER consortium is well positioned to combine genome sequence analysis and ELSI research in an active clinical setting to provide guidance and evidence-supported best practices to the rapidly expanding clinical sequencing community.

The LSAC also have an important role to play in developing rigorous methods to scaling up genome sequencing and analysis, and in piloting a few larger-scale efforts in the next few years in order to better understand the longer-term opportunities for applications that require significant sequencing and analysis capacity.

Plan for Genomics in Clinical and Health-care Settings

As a foundation for the larger Disease 2020 project, the NHGRI Sequencing Network should pilot projects related to the components above. These projects should adapt as scientific knowledge is gained and the clinical use of genomics evolves rapidly.

Examples could include:

(5.1) Genomic screening in routine newborn screening programs. Initial studies could involve exome sequencing of newborns with common or unusual developmental defects. The critical challenge is to determine the success in interpreting the genomic information to (i) provide diagnoses for known conditions and (ii) discover new genes underlying conditions.

(5.2) Genome sequencing in cancer to guide choice of therapy and to study tumor response to treatment. Initial studies could involve (i) exome or genome sequencing performed in sufficiently rapid time frame to inform choice of treatment and (ii) exome or genome sequencing performed on longitudinal samples before and after treatment, to study primary and secondary resistance to both targeted and non-targeted treatments.

(5.3) Genomic approaches to infectious disease analysis in clinical settings. Studies could include hospital infection control (e.g. whole genome sequencing to track pathogen genotypes); nosocomial infection diagnostics (e.g. metagenomic sequencing of hospital acquired diarrhea, febrile transplant patients, etc. for bacteria and viruses); and anti-microbial treatment decisions by screening for resistance mutations/genes in viruses/bacteria with metagenomic sequencing. In each case, the goal would be to evaluate how genomic methods could provide more complete, more sensitive, more cost-effective or more timely analysis than currently used methods.

Appendix 1: NHGRI Sequencing Genomics Framework Meeting Attendees

Invited Participants

Michael Boehnke, Univ. of Michigan School of Public Health
Eric Boerwinkle, Univ. of Texas Health Science Center at Houston
Carlos Bustamante, Stanford School of Medicine
Stephen Chanock, National Cancer Institute, NIH
Nancy Cox, University of Chicago
Stacey Gabriel, Broad Institute
Levi Garraway, Dana-Farber Cancer Institute
Richard Gibbs, Baylor College of Medicine
David Goldstein, Duke University
Ramaswamy Govindan, Washington University at St. Louis
Eric Lander, Broad Institute
Rick Lifton, Yale University
Jim Mullikin, National Human Genome Research Institute, NIH
Deborah Nickerson, Washington University
Sharon Plon, Baylor College of Medicine
Julie Segre, National Human Genome Research Institute, NIH
Louis Staudt, National Cancer Institute, NIH
David Valle, Johns Hopkins Medical Institute
George Weinstock, Washington University
Rick Wilson, Washington University

Scientific Advisors to the NHGRI Genome Sequencing Program

Ewan Birney, European Bioinformatics Institute
Jeffrey Botkin, University of Utah
Rex Chisholm, Northwestern University
William Gelbart, Harvard University
Deirdre Meldrum, Arizona State University
Len Pennacchio, Lawrence Berkeley National Laboratory
Pamela Sankar, University of Pennsylvania

National Human Genome Research Center Extramural Staff

Alexi Archambault
Steve Benowitz
Vivien Bonazzi
Lisa Brooks
Deborah Colantuoni
Elise Feingold
Adam Felsenfeld
Eric Green
Mark Guyer
Lucia Hindorff
Nicole Lockhart
Teri Manolio
Terryn Marette
Jean McEwan
Jeannine Mjoseth
Brad Ozenberger
Jane Peterson
Lita Proctor
Laura Rodriguez
Jeff Schloss
Mike Smith
Heidi Sofia
Larry Thompson
Katya Vaydylevich
Lu Wang
Kris Wetterstrand