MULTI-IC SYMPOSIUM ON APPLICATION OF GENOMIC TECHNOLOGIES TO POPULATION-BASED STUDIES

June 5-6, 2006 *(Final Version, 7/3/06)*

Executive Summary

The advent of cost-effective, high-throughput methods for characterizing common human genetic variation (primarily whole genome association genotyping), and for identifying rare variants contributing to complex traits (primarily sequencing), now permits the assessment on a population level of disease risk associated with common genetic variants. Because disease-related variants represent another type of risk factor, proven epidemiologic approaches for identifying and characterizing potential risk factors can and should be applied to determining their health impact.

The addition of genomic markers to existing observational studies and clinical trials, particularly studies with widespread applicability to the general population, provides superb opportunities to assess the prevalence, magnitude, consistency and modifiers of risk-related genetic variants on a population basis. Many, if not most, NIH Institutes and Centers (ICs) have made long-term, substantial investments in population-based studies that are beginning to be used for genomic research, but the addition of a genetic component raises the following questions:

- Genomic technologies: Which technologies to use, applied to whom, and in what sequence? How can one keep pace with, or even anticipate, the rapid evolution of these technologies to ensure that the most current and reliable approaches are used?
- Bioinformatics/Statistical Analysis: How to manage the mass of data produced by whole genome association and sequencing technologies? How can access be facilitated for outside groups? What approaches are needed to promote comparability across studies with different genetic, environmental, and phenotypic measures?
- Participant protections: How to ensure compliance with existing consents and human subjects approvals, and facilitate adequate consent for future studies? How can participant confidentiality be protected and risks of participation in such studies reduced? How can this be done to maintain commitments to confidentiality and restrictions on data use contained in previous informed consent documents?
- Leveraging resources: What population studies, phenotypes, and biospecimens are available? Can phenotypic or environmental measures be added to existing studies or are new studies needed? How can one extend the use of large-scale studies to meet the needs of multiple ICs?
- Power: Are the available samples sufficient to establish statistical significance of findings of public health importance? If not, are there other compelling reasons to conduct the study?
- Prioritization: What is the appropriate succession of genetic study designs for a given disease? How many studies are needed for a disease? Which studies can address not only etiology but also outcomes, pharmacogenetics, and gene-gene or gene-environment interactions? How should one select and prioritize initial whole genome studies, replication studies, and follow-up sequencing or functional studies?

- Funding: How to fund and coordinate addition of genomic technologies to ongoing studies? What issues related to data access and intellectual property rights need to be addressed in existing vs *de novo* studies?

These questions are not unique to any single IC, but many ICs are currently working in relative isolation to address them. Sharing of experience across ICs, now that the majority are engaged in these issues in some way, is likely to provide a broad array of options that can be tailored to meet an individual IC's needs. For these reasons, a one-and-a-half-day symposium was organized by a multi-IC planning group (see roster) with the following goals:

1. To identify common, critical issues that have been encountered in applying genomic technologies to population studies at NIH and creative approaches to solving them;

2. To develop approaches for prioritizing and conducting population studies using genomic technologies for use by individual ICs as desired; and

3. To identify new tools for genomics, categorization of phenotypes, and database standardization required for genome-wide association and sequence-based studies.

This document summarizes the recommendations and action items of the symposium attendees, all of whom were NIH staff. Presenters' slides and background materials are available on the symposium website (http://www.genome.gov/Pages/Extranets/PopulationGenomicsTraining/ accessed by NIH username and password). Next steps should include: consultation with the scientific community on symposium recommendations, within individual ICs or jointly; establishment of shared information resources such as consent models, technology evaluations, and consortium agreements for IC use; development of new technologies needed to advance population-based genomic research; and, potentially, development of multi-IC programs for adding genomic technologies to existing studies.

Recommendations

Symposium participants provided recommendations regarding resources and issues to consider in soliciting, funding, and implementing the application of genomic technologies to population studies; and regarding new genomic tools needed for population studies and approaches for facilitating their use. Recommendations were intended to reflect shared experience and to be advisory rather than prescriptive. Participants recognized that the recommendations will need updating as these technologies and the science they enable continue to evolve.

Highest priority recommendations are shown below as near-term administrative and scientific action items, followed by more intermediate priority goals, each with suggested leadership. Other recommendations, many of which are currently being addressed, will be moved forward as leadership for them is identified. Sentiment appeared strong for convening a second multi-IC symposium in 9-12 months, to share accrued experience in applying genomics to population studies and to assess progress in the action items identified at the 2006 symposium. The areas of statistical analysis and computational biology, which were not addressed in depth here, may deserve particular emphasis at that time.

Near-Term Administrative Action Items:	Leadership
1. Key elements of consent for genome-wide association studies (GWAS) should be collected, updated frequently, and made available to ICs and possibly to the outside community. A repository of model consent forms could be developed.	Symposium Panel 3, in collaboration with Nabel GWAS Data Sharing Committee
2. Examples or collections of successful consortium agreements and genotyping quality control standards would be helpful.	Nabel Committee, with NIH/OD
3. Existing efforts should be coordinated, and new efforts initiated as needed, to develop common data elements for key phenotypes and environmental exposures for use in GWAS.	GEI can serve as a pilot, with GAIN, NCBI, caBIG, NHLBI
Near-Term Scientific Action Items:	**Leadership**
4. Agreed-upon standards for quality of genotyping and sequencing data should be disseminated.	NHGRI
5. Rigorous algorithms should be developed to define approaches to follow-up GWA signals with sequencing: in which samples, over what interval, and what fraction of the interval (exons, promoters, conserved sequences, etc).	GEI
6. Standards for defining validity and replication of GWA findings should be developed.	NHGRI and NCI
Intermediate Priority Goals:	**Leadership**
7. Efforts should be made to identify and prioritize high-impact exposures, such as those that are readily modifiable or that have substantial relevance to many diseases and traits. The long-term goal of these efforts should be to develop standardized tools for definition, collection, and analysis.	NIEHS, with GEI
8. GWA applications should be evaluated in review for plans to promote data accessibility. Review of GWAS may need to be multi-tiered, to ensure adequate evaluation of phenotype and study design (standardization, bias) as well as genomic issues such as genotyping technology and genetic effect.	Nabel Committee, with CSR, NHGRI, NIDCD
9. Accessible sources of data structures and formats for GWAS should be provided, to reduce reinventing the wheel and improve ability to compare and pool studies in the future.	NCBI, with GAIN, NHLBI, NEI
10. The benefits and risks of electronically tracking the research use of GWA data should be explored; consideration should be given to asking that GWA study name be used in abstracts of publications.	Nabel Committee, with NCBI

Other Recommendations:

11. The database of uncommon SNPs should be expanded. **(ongoing, NCBI and NHGRI)**

12. A template consent form that is widely, though not necessarily universally, acceptable may be useful if made available to IC staff.

13. The value of performing genome wide SNP genotyping on existing cell lines and making these data widely available should be explored.

14. A set of frequently asked questions for genetics and genetic epidemiology may be useful.

15. Investigators should be encouraged to deposit GWA data from well-characterized control samples in the NCBI database, though biases in participant selection and validity, and poor comparability of phenotypic measures, may limit the utility of such controls for comparison to cases drawn from other sources.

16. Trans-NIH policies, or guidelines if policies are unnecessary or premature, are needed for consent, data release, intellectual property, and publication. **(ongoing, Nabel Committee)**

17. Increased dialogue and engagement between IRBs and NIH is needed regarding the acceptability of broad consent, the inability to identify individual genetic variants to be studied, the need for data sharing, etc.; approaches could include FAQs, presentations at meetings such as PRIM&R, newsletters, publication in journal "IRB."

18. A central IRB should be considered for GWA studies. An important charge will be to address potential conflicts in previously signed consent forms for pre-existing studies with evolving societal and scientific concerns, and to determine when exemptions or waivers could be granted or re-consenting of individual participants may be needed.

19. Consideration should be given to future development of guidelines for distribution of biospecimens, including DNA, blood/serum, or tissue, from GWAS.

20. Issues to consider in prioritizing GWAS may include:
    a. The scientific and public health rationale for the study design
    b. Evidence of heritability of the condition or trait
    b. Reasons to suspect finding a common allele that confers a significant risk
    c. Quality and extent of available phenotypic and exposure data
    d. Epidemiologic features of this trait that make it a promising candidate for study (e.g., environmental and behavioral risk factors, special clinical relevance, special population, public health impact)

21. Descriptions of currently funded case-control and cohort studies believed by NIH staff or investigators to be suitable for addition of genomic technologies, or already pursuing genomic research, could be added to databases such as the ClinicalTrials.gov website.

22. The feasibility of enhancing the search functions of ClinicalTrials.gov for GWA studies should be explored.

23. ClinicalTrials.gov should consider developing a parallel site for observational studies, as

relevance and user-friendliness of the current site for non-intervention studies are limited.

24. ICs should encourage addition of ancillary phenotypic and exposure measures to their existing studies if these would serve the needs of other ICs without interfering with the parent study.

25. Publications derived from existing study datasets should acknowledge the contribution of parent study investigators and credit the grants that supported the data collection.

26. Public concerns about research use of GWA genotype-phenotype data and whether the consent process accomplishes what it is intended to should be investigated. **(ongoing in part, NHGRI ELSI program)**

27. Consideration should be given to asking investigators to provide a template and documentation of the phenotypic and environmental data to be submitted to the GWA database at the time of application for NIH funding.

28. Consideration should be given to providing incentives for analysis of datasets incorporating genetic, exposure, and outcome data in large population studies, and for encouraging collaboration with population study investigator, to promote informed and productive use of these complex data sets. Support for collaborative efforts such as awarding small analysis grants, assisting outside investigators in applying for access, and inviting them to participate in cohort study functions have been very effective in bringing new investigators and disciplines into population-based studies.

29. A single standardized database for genotypes and phenotypes should be created and maintained by NIH through coordination of NCBI, caBIG, and similar efforts. **(ongoing)**

30. Limited subsets of phenotypic and exposure data that are amenable to common definition and standardized collection in GWAS should be identified in near future

31. Efficient methods for transmitting and handling terabytes of data are needed.

32. Databases should be tailored for intended users, anticipating who users are likely to be.

33. Web-based interfaces and tools are needed for rapidly visualizing associations in GWAS.

34. Automated data analysis tools should be developed to identify heterozygotes in DNA sequence traces more efficiently.

35. “Federation” of datasets should be considered for housing very large capacity, infrequently used data outside of central databases.

36. Comprehensive GWA panels are needed for different populations. **(ongoing in part, extension of HapMap)**

37. A “cosmopolitan” GWA panel (that will work in numerous or all populations) should be developed, either through shared resources or public availability of custom sets appropriate for admixed or under-represented populations.

38. Flexible and cost-effective technologies are needed for studies involving varying numbers

of SNPs per subject, ranging from genome-wide (~10^6 SNPs) through replication studies (~10^4 – 10^3 SNPs) through candidate SNP characterization (~ 10^1 SNPs). **(ongoing, NHGRI)**

39. Better methods should be developed for scoring structural variations. **(ongoing, NHGRI)**

40. Continued improvements are needed in sequencing technology, moving toward the $1,000 genome. **(ongoing, NHGRI)**

41. Effective methods should be developed for targeted resequencing of regions of 100kb – 1 Mb that show evidence of association to produce extended haplotypes.

42. Better methods for phenotyping (rigorous, standardized, inexpensive, non-invasive, limited burden, appropriate for controls) are needed, particularly for phenotypes relevant to a wide variety of diseases and disability.

43. Better methods for measuring environmental exposures should be developed. **(ongoing, GEI Exposure Biology component)**

44. Improved education of non-epidemiologists regarding the biases inherent in clinical case series and convenience controls is needed.

45. Methods for weighting SNPs in GWAS according to prior likelihood of association should be explored.

46. Better methods for optimizing efficient use of limited DNA in a series of initial GWAS and replication samples are needed.

47. Better methods for assessing gene-gene and gene-environment interactions should be developed. **(ongoing in part, NHLBI, NIGMS)**

Planning Group Roster

James Battey, NIDCD
Stephen Chanock, NCI
Katrina Gwinn-Hardy, NINDS
Teri Manolio, NHGRI
Rebekah Rasooly, NIDDK
Winifred Rossi, NIA
Gerald Sharp, NIAID