

Bioinformatics Integration Support Contract II

**IMPORT REQUIREMENTS
AND DESIGN REVIEW MEETING**

May 10, 2005

IMPORT

Immunology Database and Analysis Portal

Developed Under Contract Number: HHSN266200400076C

ADB Contract Number: N01-A1-40076

Delivered: June 8, 2004

Project Sponsor:

NORTHROP GRUMMAN
Information Technology

National Institutes of Health (NIH)
National Institute of Allergy and Infectious Diseases (NIAID)
Division of Microbiology and Infectious Diseases (DMID)

Prepared by:

Federal Enterprise Solutions
Health Solutions
2101 Gaither Rd, Suite 600
Rockville, Maryland 20850
(301) 527-6600
Fax: (301) 527-6401
kevin.biersack@ngc.com

Contents

1.0 Introduction 1

 1.1 Background..... 1

 1.2 Presentation References 1

2.0 Overview of Requirements 1

 2.1 Overview of BISC and the Immunology Database and Analysis Portal (ImmPort)..... 1

 2.2 Summary of the System Requirements Assessment 2

 2.2.1 HLA Program..... 2

 2.2.2 Population Genetics Program..... 2

 2.3 Research Projects 3

 2.3.1 Prioritization of Experimental Platforms 3

 2.3.2 Patient Phenotype Data 4

 2.3.3 Sample Data 5

 2.4 Issues to Resolve..... 5

 2.5 Public Reference Data..... 7

 2.5.1 Database Query 7

 2.6 Ontology 8

 2.7 Semantic Mapping 8

3.0 Key Design Features of the ImmPort System 8

4.0 Storyboards 9

APPENDIX A Participant List 12

1.0 INTRODUCTION

1.1 BACKGROUND

The Division of Allergy, Immunology, and Transplantation (DAIT) at the National Institute of Allergy and Infectious Diseases (NIAID) convened a meeting on May 10, 2005, for members of the Population Genetics Analysis Program: Immunity to Vaccines/Infections to review the requirements and design activities taking place under the Bioinformatics Integration Support Contract (BISC). The Northrop Grumman IT Team (consisting of Northrop Grumman IT, University of Texas Southwestern Medical Center, Kevric, and Unicorn) gave presentations on its requirements assessment, initial design, and screen mockups. Participants were encouraged to provide feedback on each of the presentations. This summary provides key points of the presentations and discussions.

Twenty-three participants attended the meeting, not including all members of the Northrop Grumman IT Team. For a meeting agenda, refer to PopGen Data Advisory Board Meeting: Immunology Database and Analysis Portal (ImmPort), Slide 2. A participant list is included as an appendix.

1.2 PRESENTATION REFERENCES

- PopGen Data Advisory Board Meeting: Immunology Database and Analysis Portal (ImmPort), Dr. Scheuermann, University of Texas Southwestern Medical Center, May 10, 2005
- Semantic Technologies in ImmPort, Dr. Roth, Unicorn, May 10, 2005
- Key Design Features of the ImmPort System, Dr. Klem, Northrop Grumman IT, May 10, 2005

2.0 OVERVIEW OF REQUIREMENTS

Dr. Scheuermann, University of Texas Southwestern Medical Center

2.1 OVERVIEW OF BISC AND THE IMMUNOLOGY DATABASE AND ANALYSIS PORTAL (IMMPORT)

After briefly reviewing the agenda and goals of this meeting, Dr. Scheuermann provided an overview of BISC. This contract provides advanced computer support for the collection, integration and analysis of immunology research data from diverse sources and their long-term storage in sustainable databases. Users of the system that is currently in development under BISC will include scientists associated with NIAID/DAIT extramural projects who conduct basic scientific research of genetic correlates of immune disease and clinical trials to evaluate the safety, toxicity, efficacy and underlying mechanisms of immune disease therapies.

The Northrop Grumman IT Team is developing the ImmPort system under BISC. ImmPort's main features are—

- A data warehouse
- Data marts
- Private project workspaces
- Ontology
- An analytical toolkit
- User support

For a description of the general principles guiding the development of ImmPort, refer to the PopGen Data Advisory Board Meeting presentation, Slide 6.

2.2 SUMMARY OF THE SYSTEM REQUIREMENTS ASSESSMENT

To define requirements for ImmPort, the Northrop Grumman IT Team focused its initial efforts on developing a functional understanding of experimental data and how scientists interpreted these data. Requirements gathering activities included conducting site visits to each of the Population Genetics projects and reviewing copies of their proposals to NIAID/DAIT. From these functional requirements, technical requirements (e.g., interfaces and approaches for data submission, query, and analysis) have been developed. For more information on the overall approach to the system requirements assessment, refer to the PopGen Data Advisory Board Meeting presentation, Slide 9.

At this point in time, NIAID/DAIT has determined that ImmPort will initially support the Human Leukocyte Antigen Region Genetics in Immune-Related Diseases Program (hereinafter referred to as the “HLA Program”) and the Population Genetics Analysis Program: Immunity to Vaccines/Infections (hereinafter referred to as “Population Genetics Program”). It is anticipated that ImmPort will incrementally expand its support to more programs/projects over the life of BISC. The initial release(s) of ImmPort will be designed to specifically support the needs of the HLA Program and Population Genetics Program users; however, the Northrop Grumman IT Team is approaching design and development activities in a manner that will facilitate implementation of a genericized design model to anticipate the diverse needs of what will be a much broader user community.

2.2.1 HLA Program

Dr. Scheuermann briefly reviewed general information on the HLA Program. This group will expand on the work of the International Histocompatibility Working Group. The awards have not yet been made, but Dr. Scheuermann noted that the HLA Request for Applications emphasizes state-of-the-art techniques to study the genetics of human Major Histocompatibility Complex (MHC)/disease and that it appears the HLA program will mainly focus on autoimmune diseases.

The National Center for Biotechnology Information (NCBI) maintains the dbMHC database, a publicly accessible platform for DNA and clinical data related to the human MHC. Dr. Scheuermann displayed dbMHC Web page screenshots. It is anticipated that ImmPort will directly link with the dbMHC and extract data for inclusion in ImmPort’s data warehouse. Although the dbMHC is of particular interest to the HLA Program, Population Genetics projects may also find the dbMHC useful, especially its data on anthropology/allele frequencies. The user can query dbMHC for HLA class I and class II allele and haplotype frequencies in specific populations and geographic areas. ImmPort will also become a data source for dbMHC, transferring relevant experimental data submitted to ImmPort.

The dbMHC’s query functionality enables users to compare results of previously processed data rather than having to recalculate data every time a query is submitted. Dr. Gulcher suggested that ImmPort should similarly enable the user to compare results of previously analyzed data (across data sources and studies) rather than actually processing (raw) data every time a query is run. Other participants expressed interest in being able to compare results within the Population Genetics Program. This would enable users to view suggestive associations and share them, which are generally not published.

2.2.2 Population Genetics Program

Dr. Scheuermann reviewed system-level requirements for the Population Genetics Program. The data housed for this program within ImmPort can be broadly categorized into administrative data and experimental data.

NIAID/DAIT and the Northrop Grumman IT Team have discussed how to administer ImmPort access for scientists associated with grants/contracts that have officially ended and asked for participants’ feedback on this matter. The following highlight some of the participants’ feedback:

- Dr. Cutter noted that much of the projects' key data may likely not be available until the fourth or fifth year of their contracts. That would be when the greatest opportunities for collaboration exist. It would be a shame to lose access just when ImmPort capabilities might be most useful.
- There may be interest in pooling a megacontrol study across groups after individual publications have been completed and/or original funding has ended. It would be disappointing to not have the opportunity to review metadata.
- Terminating a user's access after she has submitted data to the system would likely send a negative message to users.
- Programming for a user to lose system access on the basis of when funding ends appears to be a matter of conserving system resources. In reality, if users were concerned with losing access to the system (due to an impending end contract/grant date), those users would probably try to download the whole database.
- Comparing preprocessed data would present less of a strain than would reprocessing raw data.

2.3 RESEARCH PROJECTS

Research projects are distinct entities from contracts/grants in the system. Association with a contract/grant is a condition for NIAID/DAIT-funded research scientists to gain and retain access to the system. Research projects, on the other hand, provide an area for submitting/storing experimental results in the system. Users identified as principal investigators on NIAID/DAIT-funded contracts/grants can establish a research project in ImmPort and can share their prepublication data with other ImmPort users by assigning them access to their research projects.

2.3.1 Prioritization of Experimental Platforms

Dr. Scheuermann presented the experiment types used by Population Genetics projects, their corresponding experimental platforms, and a preliminary prioritization ranking given to each platform. The experimental platforms ranked as "1" are prioritized for inclusion in the ImmPort version 1.0 release in October 2005 or soon thereafter. The timeframe for items ranked as "2" and "3" have not been determined. The following are highlights of the discussion on experiment types and experimental platforms:

- Part of the reason that microarray data is a priority is because there is a fair amount of precedence on how to support microarray data.
- "Pyrosequencing" should be changed to "sequencing."
- The first priority in supporting any platform is to be able to capture data and parse them into ImmPort, i.e., data submission and storage. Subsequent support capabilities will include making the appropriate analytical tools available for that data.
- Is the initial data set coming into the system raw data or minimally processed data? Dr. Scheuermann suggested that if the projects are satisfied with the approaches that the other projects use to genotype, then ImmPort should use minimally processed data. There will be no need to provide the raw data for others to process.
- Nearly all Population Genetics projects are performing some form of genotyping. Results for genotyping are fairly easy to parse.

Dr. Scheuermann explained that the flow cytometry (FACS) experimental platform is relatively unique to the immunology community and hoped that ImmPort's support of this platform would provide added value to the Population Genetics projects. However, participants commented that parsing results for the FACS experimental platform would likely not be easy. The result of a FACS experiment is a rendered image. Recreating a tool in ImmPort to re-render an image would not be useful. For storing FACS data in ImmPort, one may want to consider the first result set as that file which does not require proprietary

software. A participant suggested looking into FlowJo software for FACS, which uses open source tools generating non-proprietary formats. The suggestion was also made to have two versions of each FACS results file: one for those who have the proprietary software to process the file, the other to capture small images for those who do not have the proprietary software (which they can use to pinpoint those images that are of interest and would like to investigate more).

Dr. Wilson suggested speaking with FACS experts before continuing planning efforts on how to support the FACS experimental platform.

2.3.2 Patient Phenotype Data

Dr. Scheuermann presented phenotype data that may be captured in ImmPort. A participant commented that the residential location data may be too specific. He would likely be able to identify some of his study's participants based on two or three of the data items that are listed. It was also noted that the 'age' data field needs to be more vague. To illustrate, deCODE is required to use 5-year intervals when designating someone's age. Several participants expressed interest in having a bulk upload capability for these data.

The Northrop Grumman IT Team will be in charge of semantically mapping each Population Genetics project's data to ImmPort. Semantic mapping will enable individual projects to continue using their current data items/codes, while enabling their data to be combined with data from other sources for user-driven operations (e.g., queries, analyses) and integration in ImmPort. A participant expressed that the individual Population Genetics projects may need to drive the semantic mapping process, but Dr. Scheuermann assured him that the team will perform the bulk of the semantic mapping, while working with the individual Population Genetics projects to refine and validate their mapping schemas.

Dr. Scheuermann displayed a slide with a more generic framework for capturing phenotype data. Several participants expressed concern about the impracticalities, potential noncompliance with HIPAA, and the burden of capturing the data presented in the slide. However, capturing phenotype data are optional. Furthermore, the ontological and semantic mapping components of ImmPort (discussed at a later session) address much of the participants' concern over difficulties resulting from different data models, data codes, and units of measure utilized by each project. The following are highlights of some of the participants' feedback:

- Most of the phenotype data will not be readily accessible to the Population Genetics projects. For example, patient records are not accessible.
- The reliability of sources of phenotype data needs to be considered when deciding what information to capture.
- There are impracticalities involved with history by medical record, and it may not be useful for a backend analysis.
- Collecting these data will take too much time and effort. A participant wants to simplify and standardize the data choices (e.g., 'yes' and 'no'). Participants advise to strive for simple derivative diagnoses/data.
- Capture high-level data. It is impractical to try to capture very detailed phenotype data.
- Trying to capture these data out of what the participants' sites have already collected may be problematic.
- A participant questioned the choice of a relational database. Sometimes it's better to use patient-oriented records. Laboratory tests are not all the same.
- The Population Genetics projects have a lot of nonoverlapping phenotypes.
- A few participants expressed interest in having the capability to match results across phenotypes sooner rather than later.

- The Population Genetics projects will have access to a data dictionary explaining the data items to be submitted.

Dr. Klem emphasized that submission of phenotype data is optional. Although ImmPort will not require phenotype data, it will be able to capture any phenotype data that may be submitted.

2.3.3 Sample Data

Biological samples can undergo several intermediary processes before they are ready for use in experiments. Dr. Scheuermann presented a simplified approach for capturing the preparation process in which the sample used in an experiment is referred to as a 'sample' and the immediate source from which the sample is processed is referred to as a 'sample source.' Using this approach, a DNA sample may have peripheral blood mononuclear cells (PBMC) as the sample source. Using 'sample' and 'sample source' avoids having to define 'primary' and 'secondary' samples. Yet, this approach provides flexibility to move up and down the hierarchy of processes that may be involved in preparing a sample.

The following are highlights of the discussion:

- ImmPort may want to consider using a research proposal process to coordinate ImmPort users' data analysis efforts to avoid overlap. However, having a few groups perform overlapping analyses may be useful.
- A participant suggested having downloadable APIs for their own machines.
- Providing a capability to download data from ImmPort increases the risk of unauthorized use by non-NIAID/DAIT-funded persons. Ms. Kraft commented that users will need to agree to standardized usage agreements before they access ImmPort.
- Some participants wondered whether their study consent forms adequately informed their participants of the possibility that their (de-identified) data would be deposited in a database that is accessible to other users beyond the immediate Population Genetics project. Dr. Nabavi noted that this scenario is under review, but there is no indication at this time for concern that deposition in ImmPort falls beyond the scope of a participant's consent.
- Measures such as censoring certain information or implementing pooling algorithms should be taken to help retain the confidentiality of study participants' records.
- Dr. Scheuermann clarified that the Population Genetics projects will decide for themselves which data they will submit to ImmPort.
- A participant advised that the Team should develop a data architecture that will be useful to the widest group of users possible. Capturing data granularity is not necessarily of interest to many of the Population Genetics projects.
- As consumers of information, scientists generally approach data with a level of skepticism. They generally question the validity of data. At the same time, a participant noted that submitting these data could be misleading, because there may be some users who assume that all the data in ImmPort will have been validated already. For these reasons, the Northrop Grumman IT Team may find that collecting such information may not be as useful to ImmPort users as it had thought originally.
- The Northrop Grumman IT Team will preliminarily identify core data elements (a subset of data that are considered important to capture) and circulate them to the Population Genetics projects.

2.4 ISSUES TO RESOLVE

Dr. Scheuermann asked participants for feedback on the following issues:

- What is the definition of primary data?
- What is the nature of processed data?

- How much primary data should be archived?
- How much processed data should be retained?
- Should experimental data be submitted using a standardized file format and corresponding header file, a manual user interface for the metadata, or use both?

Feedback included the following key points:

Primary data/processed data

- Minimally processed data should be the first level of data to be brought into ImmPort (i.e., do not capture raw, preprocessed data).
- ImmPort can capture some sort of quality score at different data levels.
- Have a decentralized quality control model in which the individual users apply their own criteria to determine which data meet quality control standards for submission and which data should not be submitted to ImmPort.
- Questionable data can be handled in one of two ways: Submit them to ImmPort with annotation or exclude them from the submission.
- Annotation can be used to explain quality control steps/algorithms.
- ImmPort should anticipate the possibility that an investigator may change her opinion on the quality of genotype data she had previously submitted.
- Using the private project workspace enables a user to run operations on his prepublication experimental data using ImmPort resources (e.g., analysis tools, access to reference data, etc.) and facilitates sharing and/or collaboration with other users.

How much data to archive/store

- One approach is to not make any qualitative judgment on what should be submitted and let the user make the decision on how much to submit. (Of course, data will need to meet minimum technical standards.)
- Storage of processed data results is most important.
- Some participants expressed interest in being able to match/compare data analysis results among different populations and across different experimental platforms.

Batch submission versus a manual submission interface

- Participants agreed that having a batch data loading capability is more desirable than a wizard-driven data load interface.
- The wizard-driven interface may be more suitable when ImmPort becomes more established (and data requirements are stable) and when some scientists begin to use ImmPort as their primary data repository.
- A participant suggested building in SQL capabilities during the data load process. He opined that XML is more complicated.

Other comments include the following:

- A participant advised against making study participant unique identifiers (used at the individual Population Genetics project sites) available in ImmPort's public data warehouse. There should be rules addressing how to handle unique identifiers.
- Query results should not return individual-level records. A user querying the public data warehouse should see results from groups of people (e.g., 5-person groups).

- Experimental data made available in the public data warehouse will be associated with the appropriate PI/research study.

Updating the data warehouse

Some disadvantages of operating a real-time database were discussed. A participant opined that to keep track of different changes to a real-time database is an “accounting” problem that will need extra effort from both ends (the Northrop Grumman IT Team and projects submitting data to ImmPort). A lot more programming is needed to update data in the public data warehouse in real-time than if the team were to take the approach of completely replacing the database on a periodic basis (e.g., monthly, quarterly). In this scenario, projects would send new data submissions (including data promoted from private project workspaces to the public data warehouse) and updates for inclusion in the next release. The participants generally agreed that this would be a better approach than having a real-time database. (Previous data versions in the public data warehouse will be archived, not deleted.)

2.5 PUBLIC REFERENCE DATA

Dr. Scheuermann explained that in considering which publicly available reference data to make available in ImmPort, the following questions were asked:

- Which data would be most useful?
- How to link data together
- How to visualize linked data

Data marts will provide customized views of data subsets and supersets and can present the same data in different ways (e.g., genotypic view, phenotypic view, experiment-focused view).

Linking genes to pathway data will enable users to link genes to larger biological processes.

ImmPort will have its own human-specific list of immunology-related genes. At this time, the draft list consists of approximately 1800 genes selected on the basis of those genes under study by the Population Genetics projects and LocusLink database searches using Gene Ontology immunology terms. For a description of how the list was compiled, refer to the PopGen Data Advisory Board Meeting presentation, Slide 39. For a visual representation of the different information that will be linked to any given gene, refer to Slide 43.

Dr. Scheuermann asked participants whether using HapMap data to recommend tagSNPs for the ImmPort genes would be useful to them. He also asked participants for feedback on LD-select. For the specific slide, a participant noted that the region shown has very short LD blocks. However, in general, such analysis would be useful as long as there is good coverage. ImmPort should flag the user if an LD block is not well covered. Another participant suggested starting with LD blocks with the most information to see if this yields useful results to the Population Genetics projects. Participants agreed that this would be a good project for interested individuals from different sites to work together and give input.

Dr. Scheuermann walked through examples of gene expression data that can be used for cross-experiment comparisons using expression rank and linking to metabolic pathway information based on a specific gene (See the BioCyc Database Collection at www.biocyc.org for an example of a pathway/genome database).

2.5.1 Database Query

Dr. Scheuermann showed the Plasmodium Resource Database, which has a robust set of canned queries that ImmPort may use as a model for some of its query choices.

ImmPort will support a variety of complex database queries without requiring that the user know the database structure or SQL.

2.6 ONTOLOGY

Dr. Fan, Kevric

Dr. Fan demonstrated the ImmPort ontology using the Protégé ontology editor. The ImmPort ontology currently has over 50,000 entries and may leverage other existing ontologies (e.g., FMA, proteomics ontology, etc.). Several participants expressed interest in gaining access to the ontology soon. The ontology may be made available to the core group of Population Genetics projects that agree to beta test ImmPort in late August 2005. The ontology will be released with ImmPort version 1.0.

The ImmPort ontology is currently in an Oracle database, but will be viewable via ImmPort through a browser. The ImmPort ontology can also be released in a general OWL file format.

The Northrop Grumman IT Team will work with the Population Genetics projects to validate the ImmPort ontology.

The ontology will not initially be used in the batch data submission process but will be used to support data query features.

ImmPort should provide the user with filter options to make using the ontology easier. A participant asked whether the user will take the time to view the ontology. That is for the user to decide.

Dr. Kammer noted that the ontology may provide some guidance on determining the scope of the experimental metadata items (e.g., phenotype data, biological sample data, etc.) that individual projects may want to track.

2.7 SEMANTIC MAPPING

Dr. Roth, Unicorn

Dr. Roth provided a brief description of semantic mapping and the role it will play in ImmPort. The team will handle most of the burden of mapping activities between ImmPort and the individual projects; however, the team will need to work with the projects to validate the mapping schemas. Although the mapping illustration shown to participants was simplistic, semantics mapping can represent very complex relationships. The ontology and the semantic mapping will evolve as ImmPort expands.

It is not yet determined whether Unicorn's semantic mapping tool will be part of ImmPort version 1.0.

3.0 KEY DESIGN FEATURES OF THE IMMPORT SYSTEM

Dr. Klem, Northrop Grumman IT

Dr. Klem's presentation provided a brief overview of major design areas of ImmPort, including—

- Private project workspaces
- Collaborative workspaces
- Data submission features
- Query options
- Analysis pipelines

- Journaling
- Public experimental data
- Public reference data
- Ontology

Dr. Klem reviewed some key science-based concepts/terms applied in ImmPort, including experiments, experiment groups, research projects, primary data sets, and secondary data sets. An understanding of these terms is necessary to be able to organize and retrieve experimental data in ImmPort. The Northrop Grumman IT team will provide a written description of these terms and other terms to NIAID/DAIT and the Population Genetics projects in June 2005. The team would like to obtain the projects' feedback/critique during the next data advisory board meeting.

The team will map to the projects' current file formats for the projects to be able to load data.

Mr. Desborough suggested making the ontology available during data submission via the manual load interface for assistance.

A suggestion is made to focus on SNP and expression data.

For more information on ImmPort's design, refer to the Key Design Features of the ImmPort System presentation.

4.0 STORYBOARDS

Dr. Scheuermann, University of Texas Southwestern Medical Center, and Mr. Chiu, Northrop Grumman IT

Dr. Scheuermann and Mr. Chiu presented storyboard designs to obtain initial feedback from the participants on design details, business rules, and underlying requirements. The following section groups the participants' feedback according to the discussion topics prompted by the storyboard screens.

Querying the private project workspace

- Querying the private project workspace in effect runs queries on the metadata of the experiments and returns data sets.
- The ability to group complex queries using parentheses is needed.
- Permit a user to construct queries using SQL.
- Permit a user to modify a previously constructed query.
- Permit a user to change logical parameters.
- When one enters the query page, enable the user to perform a query in both the private project workspace and the public data warehouse.

Searching project information

- Some participants expressed interest in enabling users to query for keywords describing other research projects in ImmPort. After reading the search results, users would then be able to contact the PIs of the research projects that are of interest to them.
- Ms. Kraft noted that to accommodate those users who may not want to be contacted by others, the ImmPort system could ask users as they register in the system whether they would like their contact information to be made available to other users.

- Make contract/grant information searchable since it is already available on the Computer Retrieval of Information on Scientific Projects (CRISP) database.

Data submission

- Dr. Baxter would like ImmPort to be able to handle large batch loading activities for experimental data in which metadata and results from many experiments are submitted together.
- A recommendation was made to look into Java applets to facilitate the FTP process.
- Several participants commented that the data submission function should be made available to non-PI users.

Upload wizard

- A participant noted that he liked that fields could be prepopulated with array design information.
- Under the sample tab, consider using the information entered for the first sample form to prepopulate the subsequent sample information that needs to be completed.

Downloading

- It is not yet determined whether downloading will be performed in real time; however, requests will be queued in real time. A notification mechanism will alert the user of an estimated time when a request will be completed.
- A user may want to retrieve a data set and then use a data analysis tool to perform operations on the data set. It would be inconvenient to have to wait a long time between selecting a data set and actually getting it into a tool to perform analysis.
- Consider providing more selectivity in the front end to enable a user to retrieve lower level data to improve retrieval time.

Copies of data sets

- It is not yet decided whether ImmPort will allow for more than one physical copy of the same data set or use virtual copies.
- Participants reasoned that one would normally want to modify data to make corrections, and any corrections should be made throughout all copies of that data set. However, changes to data sets for normalization purposes should be isolated to a specific data set copy.
- Having multiple physical copies may increase query optimization.
- A historical annotation tool is needed.

Other discussion

Dr. Gulcher wondered whether there are real world examples in which data sets on two different diseases from two different PIs would be merged at the raw data level. Would you combine two populations (e.g., diabetes and arthritis)? Are there phenotypes for which this would be done? Another participant noted that you would not so much query two groups with two diseases, but you would normally look for a gene and then maybe find out what populations it occurs in. You would generally compare results first, and then merge data afterwards (i.e., start with higher-level data first).

There is interest in being able to query for phenotypes that have been genotyped with SNPs near a particular gene. It would be useful to determine whether other projects have genotyped a gene and to be able to compare results. Let the gene direct a user to phenotypes that are not necessarily associated with the gene. You would not have to have a positive association. You just want to be able to determine

whether a phenotype has been genotyped. The intention is to be able to go from phenotype to genotype and the reverse.

Participants agreed to discuss during the next call whether projects were comfortable disclosing their lists of genotyped genes to other Population Genetics projects and other projects.

The dbMHC is a good example of a database that enables users to compare results from multiple projects. Although scientists may be studying different diseases, there may be common underlying genes (or vice versa).

APPENDIX A PARTICIPANT LIST

Dr. Susan Baxter, National Center for Genome Resources

Mr. Kevin Biersack, Northrop Grumman IT

Dr. Joseph Breen, NIAID

Dr. Gary Cutter, University of Alabama Birmingham

Dr. Valentina DiFrancesco, NIAID

Dr. Jeffrey Edberg, University of Alabama Birmingham

Dr. Vincent Ferretti, McGill University

Dr. Jeffrey Gulcher, deCODE Genetics

Dr. David Hall, Research Triangle Institute

Dr. David Kammer, University of Washington

Dr. Richard Kaslow, University of Alabama Birmingham

Dr. Perry Kirkham, NIAID

Dr. Edward Klem, Northrop Grumman IT

Ms. Cheryl Kraft, NIAID

Dr. Mark Loeb, McMaster University

Ms. Lara R. Miller, NIAID

Dr. Nasrin Nabavi, NIAID

Dr. Shane Pankratz, Mayo Clinic

Dr. Richard Scheuermann, University of Texas Southwestern Medical Center

Dr. Paul Targonski, Mayo Clinic

Dr. Christopher Wilson, University of Washington

Dr. Ashley Xia, NIAID

Dr. Alison Yao, NIAID