***Imm**unology Database and Analysis **Port**al*
ImmPort

Analysis of Shared Genotyping Data

27 June 2008 - WebEx

NORTHROP GRUMMAN Information Technology

SOUTHWESTERN

ImmPort Goals

- Develop an infrastructure for IT support of DAIT-funded investigators to:
  - Archive and manage experimental data
  - Analyze experimental data
  - Integrate with extensive knowledgebase
  - Analyze integrated data
  - ***Build upon data generated by previous studies***
- Develop data and vocabulary (ontologies) standards
- Support data analysis best practices

Initial work

- Phase I programs
  - Population Genetics Analysis Program: Immunity to Vaccines/Infections
  - Immune Function and Biodefense in Children, Elderly, and Immunocompromised Populations
  - HLA Region Genetics in Immune-mediated Diseases
  - Immune Tolerance Network
  - Atopic Dermatitis and Vaccinia Network (ADVN)
  - NIAID non-human primate colony
- Common theme
  - ***Genotype-phenotype associations***

Issue

- Sharing of genetic data
  - NIH data sharing compliance
  - Ethical considerations
  - Legal implications
  - IP implications
- Technical approaches

Bioethics Conference

- Data Sharing and the Bioethics of Collaborative Genetic Research Conference in JAN 2007
- Panel of bioethicists, scientists, legal experts and participant advocates
- Issues and recommendations
- Manuscript entitled “Ethical and Practical Issues Associated with Aggregating Databases” submitted to PLoSMedicine Policy Forum

IMMPORT

Panel

Name	Title	Institution	Expertise
Dr. David R. Karp	Associate Professor of Internal Medicine, Co-Investigator, BISC	UT Southwestern Medical Center	Co-Chair
Dr. Jeremy Sugarman	Professor of Bioethics and Medicine	Johns Hopkins University	Co-Chair
Dr. Gail Geller	Associate Professor of Pediatrics and Bioethics	Johns Hopkins University	Informed Consent, Human Genome Project
Dr. Jeffrey P. Kahn	Professor of Medicine and Director of the Center for Bioethics	University of Minnesota	Public Health Policy
Dr. Hank Greely	Professor of Law	Stanford Law School	Genetics and Law
Dr. Kathleen M. MacQueen	Senior Scientist	Family Health International	Protection of vulnerable populations ; global research ethics
Dr. Robert Cook-Deegan	Director, Center for Genome Ethics, Law, and Policy	Duke University	Human Genome Project and Public Policy
Dr. Bradley Malin	Assistant Professor of Computer Science	Vanderbilt University	Data Privacy
Dr. Daniel E. Ford	Vice-Dean for Clinical Investigation	Johns Hopkins University	Biospecimen usage
Dr. Joel Guthridge	Research Assistant Member	Oklahoma Medical Research Foundation	DAIT-supported ImmPort user
Dr. David Glass	Professor of Pediatrics	University of Cincinnati	DAIT-supported ImmPort user
Dr. Richard Kaslow	Professor of Epidemiology	UA Birmingham	DAIT-supported ImmPort user
Dr. Richard Scheuermann	Professor of Pathology, Principal Investigator, BISC	UT Southwestern Medical Center	Bioinformatics; Immunology
Shelly Carlin, JD, MPH		Dallas, TX	IRB member, subject advocate
Ms. Cheryl Kraft	Program Officer, BISC	NIH	BISC Program

IMMPORT

Specific Issues Discussed

- Change in informed consent standards
- Subjects may not have consented to participation in genetics research
- Subjects may not have consented to “secondary uses”
- IRB problems because of lack of consensus on data re-use
- Participant re-identification
- Lack of databases interoperability may lead to flawed analysis
- Compromise publication priority and intellectual property claims
- Participants and sponsors want to maximize discoveries

IMMPORT

Panel Recommendations

1. Determine whether initial consent and ethical approval will allow secondary research.
2. Ensure there are appropriate data security mechanisms and review bodies to protect privacy interests in aggregated databases.
3. Informed consent should take into account the potential incorporation of data into aggregated databases.
4. Address special challenges of using data obtained from existing databases.
5. Pursue efforts directed at standardization of data.
6. Establish data sharing rules, including attribution of contributions.
7. Adopt “best practices” to avoid identifiability of the data.

ImmPort

Re-identification

- By design, the research facilitated by combination and re-analysis of datasets within de-identified databases do not constitute human subjects research under US regulations. This may lead to a false sense of security as complex phenotypes and extensive genotypes may be used to re-identify individuals.
- This needs to be prevented using technologies and policies that address the appropriate uses of those data.
  - Only data in de-identified, aggregated form should be accessible to all users.
  - Access to record level data must be limited either by keeping all data within the confines of the secure database and providing users with sophisticated analytical capacity, or by requiring legal agreements to prevent attempted re-identification of subjects.
  - Downloading and re-distribution of data should be carefully controlled.
  - If allowed at all, extraction of record level data should only be for specified, pre-approved purposes.

Technical approach

- Data can be stored in encrypted form in the database server. A researcher would only be able to query a processing engine that would interact with the database and report the results. Analysis of record level information would still be possible, but that analysis would be done by the database host, and not by downloading the data to client computers.

Related Projects

- dbGAP
- GAIN
- GWAS

dbGAP

dbGaP Overview

The **d**ata**b**ase of **G**enotype **a**nd **P**henotype (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.

The data in dbGaP will be pre-competitive, and will not be protected by intellectual property patents. Investigators who agree to the terms of dbGaP data use may not restrict other investigators' use of primary dbGaP data by filing intellectual property patents on it. However, the use of primary data from dbGaP to develop commercial products and tests to meet public health needs is encouraged.

ImmPort

dbGAP

Data Content and Organization

Open-Access Data

Open-access data can be browsed online or downloaded from dbGaP without prior permission or authorization. These data will include, but may not be limited to, the following:

dbGaP Data Type	Where to Find It
Studies	'Study' column when browsing studies
	Result of a search under the tab 'Studies'
	Part of the breadcrumb path of a variable or document
Study Documents	Link from 'Browse Studies'
	Link under 'Associated Documents' on study report
	Result of a search under the tab 'Study Documents'
Phenotypic Variables	Link under 'Browse Studies'
	Link under 'Associated Variables' on study report
	Result of a search under the tab 'Variables'
Genotype-Phenotype Analyses	Link under 'Associated Analyses' on variable report
	Link under 'Associated Analyses' on study report

Please note that this is a general description of what is available to open- access users. Data available to open-access users may vary between studies and may also differ from what is described here without notice. You can find more details regarding data access policies for specific studies on the individual study report pages.

Controlled-Access Data

Controlled-access data can only be obtained if a user has been authorized by the appropriate Data Access Committee (DAC). Information on requesting controlled data access, is available below. Data available to authorized investigators may include the following:

- de-identified phenotypes and genotypes for individual study subjects
- pedigrees
- pre-computed univariate associations between genotype and phenotype (if not made available on the public site)

Since data access policies are determined on a per-study basis, data available to users with controlled access authorization may vary between studies and may also change from what is described here without notice. You can find more details regarding data access policies for a specific study on the study report page along with a link to the appropriate authorizing body.

Links | Help | About Immport

IMMPORT

dbGAP

Browse dbGaP

TUTORIAL | ABOUT dbGAP

By Studies | By Diseases | Advanced Search

Study	Embargo Release	Details	Participants	Type of Study	Project
CIDR: Genome Wide Association Study in Familial Parkinson Disease (PD)	Feb 13, 2009	V D	1991	Case-control	
Framingham SHARe	Version 1: Oct 19, 2008. Version 2: Feb 01, 2009.	V D A	14489	Community-based, longitudinal, family-based cohort	
GAIN: Collaborative Association Study of Psoriasis	Aug 13, 2008	V D	2875	Case-control	GAIN
GAIN: Genotyping the 270 HapMap samples for GAIN by Broad			-	Parent-offspring trios	GAIN
GAIN: Genotyping the 270 HapMap samples for GAIN by Perlegen			-	Parent-offspring trios	GAIN
GAIN: International Multi-Center ADHD Genetics Project	Mar 26, 2008	V D	2835	Parent-offspring trios	GAIN
GAIN: Linking Genome-Wide Association Study of Schizophrenia	Version 1: Nov 07, 2008. Version 2: Dec 11, 2008.	V D	5066	Case-control	GAIN
GAIN: Major Depression: Stage 1 Genomewide Association in Population-Based Samples	Jul 15, 2008	V D	3741	Case-control	GAIN
GAIN: Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes	Jul 09, 2008	V D	1825	Case-control	GAIN
GAIN: Whole Genome Association Study of Bipolar Disorder	Version 1: Nov 25, 2008. Version 2: Dec 30, 2008.	V D	3261	Case-control	GAIN
GAW16 Framingham and Simulated Data	Oct 19, 2008	V D	7130	Longitudinal, population-based	
Genome-wide Association Study of Neuroblastoma		V D	1032	Case-control	
Genome-wide study in amyotrophic lateral sclerosis and controls: first stage analysis		V D	544	Case-control	
Ischemic Stroke Genetics Study (ISGS)		V D A	485	Case-control	
Mayo-Perlegen LEAPS (Linked Efforts to Accelerate Parkinson's Solutions) Collaboration	Mar 03, 2008	V D A	1550	Case-control	
NEI Age-Related Eye Disease Study (AREDS)	Jun 11, 2007	V D A	600	Case-control	
NINDS Parkinson's Disease	Oct 12, 2007	V D A	535	Case-control	
NINDS Parkinsonism Study	Oct 12, 2007	V D	1283	Case-set	
NINDS Repository Cerebrovascular Disease/Stroke Study		V D	870	Case-set	
NINDS Repository Motor Neuron Disease/ALS Study		V D	1790	Case-set	
NINDS Repository Neurologically Normal Control Collection	Oct 12, 2007	V D	2723	Control-set	
Study of Irish Amyotrophic Lateral Sclerosis (SIALS)			-	Case-control	
The Finland-United States Investigation of NIDDM Genetics (FUSION) study		V A	2335	Case-control	
Whole Genome Association Study of Systemic Lupus Erythematosus			4651	Case-control	

Genetic Association Information Network (GAIN)

Overview

The Genetic Association Information Network (GAIN) is a public-private partnership of the Foundation for the National Institutes of Health, Inc. (FNIH) and includes partnerships with the National Institutes of Health (NIH) and the private sector.

GAIN will support a series of Genome-Wide Association Studies (GWAS) designed to identify specific points of DNA variation associated with the occurrence of a particular common disease. Investigators from existing case-control or trio (parent-offspring) studies have been invited to submit samples and data on roughly 2,000 participants for assay of 300,000 to 500,000 single nucleotide polymorphisms designed to capture roughly 80 percent of the common variation in the human genome. Specific genes involved in disease processes can be identified once particular areas of the genome are associated with disease occurrence.

With the sequencing of the human genome, lower cost and more efficient technologies, and the exploration of common patterns of genetic variation through the International Haplotype Map (HapMap) project, we now have the techniques and tools necessary to uncover causal genes that contribute to disease susceptibility, and thus facilitate discovery of new molecular targets for prevention, diagnosis, and treatment.

Applications for access to GAIN genotyping were submitted on May 9, 2006 and projects selected for genotyping were announced at the American Society of Human Genetics annual meeting on October 10, 2006.

The GAIN Kick-off Meeting and Analysis Workshop was held November 29-30, 2006. Visit the Workshop Web site: GAIN Kick-Off Meeting and Analysis Workshop [capconcorp.com].

The next GAIN Analysis Workshop will be held in September or October of 2007, and will be announced on the GAIN Web site at Foundation for the National Institutes of Health: Genetic Association Information Network [fnih.gov].

GAIN Project Director for Foundation for the National Institutes of Health

David Wholley, M.A.
E-mail: wholleyd@mail.nih.gov

Links | Help | About Immport

GAIN

genome.gov
National Human Genome Research Institute
National Institutes of Health

Search

Home | About NHGRI | Newsroom | Staff

Research | Grants | Health | Policy & Ethics | Educational Resources | Careers & Training

Policy and Legislation Database
Search federal and state laws and policy.

Current Issues:

- **Genetic Discrimination**
  Protection for personal genetic information.
- **Genetic Testing**
  Issues, questions, legislation, policy.
- **Intellectual Property and Genomics**
  Gene patenting, guidelines, policy.
- **Coverage and Reimbursement of Genetic Tests**
  Who will pay for genetic testing?
- **Health Issues**
  Genetic testing, Personalized Medicine, Informed Consent.

Policy & Ethics

Critical issues and legislation surrounding genetic research

Feature

GINA: Keeping Pace with the Times

This week's *New England Journal of Medicine* features an article on the Genetic Information Nondiscrimination Act (GINA) co-authored by NHGRI's Francis Collins, M.D., Ph.D., and M.K. Holohan, J.D., and Kathy Hudson, Ph.D., director of the Johns Hopkins Genetics and Public Policy Center. The new federal law protects people from discrimination in employment and health insurance based on genetic information.

Read the article and its Quick Guide to GINA [content.nejm.org]

Learn more about the Genetic Information Nondiscrimination Act of 2008

Highlights

GWAS

Funding Opportunities
- Funding Opportunities (RFAs, PAs) & Notices
- Unsolicited Applications (Parent Announcements)
- Research Training & Career Development
- Small Business (SBIR/STTR)
- Contract Opportunities

NIH-Wide Initiatives
- New Investigators Program
- Multiple Principal Investigators
- Genome-Wide Association Studies (GWAS)
- NIH Roadmap for Medical Research
- NIH Blueprint for Neuroscience Research

Global OER Resources
- Glossary & Acronyms
- Frequently Used Links
- Frequent Questions

Genome-Wide Association Studies (GWAS)

The NIH is interested in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease. For the purposes of this policy, a genome-wide association study is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition. Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine. In addition, rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease. The purpose of this Website is to support the implementation of the GWAS Policy.

The NIH will continue to release additional guidance information on this site. Please e-mail GWAS@mail.nih.gov with any questions.

Policy Guidance

- Governance Structure - (07/27/2007) (PowerPoint - 224 KB)
- Frequently Asked Questions - (12/12/2007)
- Points to Consider for IRBs and Institutions - (11/29/2007) (PDF - 176 kb)
- General Brochure - (11/16/2007) (pdf - 607 kb)
- Data Sharing Plan - (11/27/2007) (pdf - 168 kb)
- End to End Diagram - (11/16/2007) (pdf - 176 kb)

IMMPORT

ImmPort Implementation Plan - Goals

1. Protect participant privacy
2. Prevent re-identification
3. Support analysis of aggregated data
   a. Public & public
   b. Public & private
   c. Private & private
4. Facilitate collaborative access to participant-level data

ImmPort Implementation Plan - Preconditions

The following data would need to be available in either the Semi-Public Workspace (SPW) or in the Private Project Workspace (PPW):

1. Genotype data
    a. Sequence feature ID (e.g. SNP ID; HLA-A beta strand 3)
    b. Variation type for each sister chromosome (e.g. A/G; Type 4/7)
2. Phenotype data
    a. Discrete variable
        i. Phenotypic characteristic (e.g. disease; blood glucose; gender)
        ii. Characteristic value (e.g. affected/unaffected; normal/abnormal; M/F)
    b. Continuous variable
        i. Phenotypic characteristic (e.g. ANA titer; IFN gamma [])
        ii. Characteristic value (e.g. 1/512; 23 nM)

ImmPort Implementation Plan - Results

The following results would be provided:

1. For discrete variables:
   a. n x n contingency table
   b. statistics derived from chi-squared and Fisher exact test analysis
2. For continuous variables:
   a. logistic regression

IMMPORT

ImmPort Implementation Plan - Steps

1. User selects relevant data set sources from SPW and/or PPW based on study metadata, e.g. studies in which participants with Type 1 diabetes are included and in which HLA-DQ1 genotype data is available.
2. User filters participant set based on selected phenotypic characteristics
3. User selects discrete vs. continuous variable options
4. User selects sequence features for analysis (All as one option). These will then correspond to the dependent variables in the analysis experiment.
5. User selects statistical test
   a. For discrete independent variables – chi-squared or Fisher's exact
   b. For continuous independent variables – logistic regression
6. User runs pipeline
7. System extracts relevant record-level data from database and aggregates accordingly.
8. System generates contingency table from aggregated data.
9. System test all values in contingency table:
   a. If all values are >5, system proceeds to Step #10.
   b. If any value is <5, system displays warning message that the number of subjects in one or more categories is too small to maintain participant anonymity and prompts the user to re-run the analysis using less stringent parameters.
10. System performs statistical analysis and displays summary results.