New models of collaboration in genome-wide association studies: the Genetic Association Information Network

The GAIN Collaborative Research Group

The Genetic Association Information Network (GAIN) is a public-private partnership established to investigate the genetic basis of common diseases through a series of collaborative genome-wide association studies. GAIN has used new approaches for project selection, data deposition and distribution, collaborative analysis, publication and protection from premature intellectual property claims. These demonstrate a new commitment to shared scientific knowledge that should facilitate rapid advances in understanding the genetics of complex diseases.

Genome-wide association (GWA) studies of large numbers of individuals genotyped for hundreds of thousands of common genetic variants have now convincingly been shown to be effective in identifying genes related to health and disease[1–10]. The growing understanding of genome variation provided by the International HapMap Consortium[11] and continued major advances in genotyping technology[12] have together made it possible to conduct high-throughput, cost-effective GWA studies in large numbers of individuals with detailed information on phenotypic traits and environmental exposures. The resulting data will be used to identify genetic variants potentially related to health and disease, to assess the prevalence of these variants in large and diverse samples and to examine possible modifiers of gene-disease relationships.

Many GWA studies thus far have focused on a single phenotype, such as diabetes or breast cancer[4–9], or on closely related phenotypes, such as ulcerative colitis and Crohn's disease[13], within a single study. However, the pressing need for replication of initial associations[14] and the opportunities for developing common methods across GWA studies have led to the formation of networks of collaborative GWA studies involving different study samples and multiple phenotypes. The Wellcome Trust Case Control Consortium (WTCCC) is one such network; its pioneering effort on seven complex diseases and common controls has proven the power and potential of this approach[10]. GAIN is another, currently involving six different studies with case-control or family trio designs. Such efforts to develop robust, common approaches to study selection, genotyping, quality control, data analysis and data sharing, as well as cross-study analyses of common phenotypes, common controls and genotyping artifacts, are topics of considerable current interest.

GAIN is a public-private partnership between the Foundation for the National Institutes of Health (FNIH), the US National Institutes of Health (NIH) and partners in the academic and private sectors. The FNIH was established by the US Congress to support the mission of NIH; it works to advance scientific research by linking the generosity of private-sector donors and partners to programs that support the NIH mission. GAIN involves four private-sector partners at present, including the founding lead partner, Pfizer, as well as Affymetrix, Perlegen Sciences and Abbott, and one academic-sector partner, the Eli and Edythe L. Broad Institute of MIT and Harvard. Commitments from these partners have supported the initial development of GAIN and the genotyping and data distribution for up to 18,000 samples. Subsequent commitments will be sought to support future GWA studies by GAIN and to extend its infrastructure for broader use.

The design and implementation of GAIN has been directed by a series of guiding principles (**Box 1**) that will be adhered to throughout the life of the project. GAIN will release data as broadly and rapidly as possible, with equal opportunity for access by all users who agree to protect the confidentiality of study participants and to respect the intellectual investment of the investigators contributing data and samples to GAIN. Here we describe the selection and characteristics of the first six GAIN studies, as well as the design, policies, protections and implementation of GAIN as a whole.

Selection and characteristics of initial GAIN studies

In early 2006, the FNIH solicited DNA samples from existing studies worldwide for genotyping by GAIN in a GWA study (http://www.fnih.org/GAIN/Instructions_for_Applicants.shtml; **Fig. 1**). Applications underwent a multistage review process similar to that conducted by the NIH, including an administrative review for adherence to the application format and submission requirements, a scientific peer

A full list of participating authors and research groups is given at the end of this paper.
e-mail: manolio@nih.gov

Published online 29 August 2007; doi:10.1038/ng2127

review assessing the likelihood of identifying important genotype-phenotype associations from the study, and a programmatic and technical review for quality and ease of use of the proposed data set, limitations on research use or data sharing, public health significance and diversity of represented populations.

Applicants submitted an online application describing the study design and population, the disease proposed for study, the available phenotypic and environmental exposure data, the proposed analytic strategies and follow-up studies and the willingness of contributing investigators and their institutions to abide by GAIN policies. In addition, to verify data availability and assess data quality, all applicants submitted electronic files of de-identified, individual-level participant data and accompanying documentation before their applications were referred for peer review.

Informaticians at the National Center for Biotechnology Information (NCBI) assessed the completeness and accuracy of data transfer, the extent of missing and out-of-range values, the quality of documentation, the extent of human curation needed to integrate the documentation with the phenotype data and the clarity of case and control definitions. This initial assessment was provided to peer reviewers and GAIN project staff, and NCBI personnel were available as needed during the peer review discussions.

Scientific peer review was conducted by an independent committee of leading scientific experts from government, industry and academia. Applications were evaluated on several criteria, including significance and complexity of the trait, the need for a GWA study, the appropriateness of the study design and population(s), the quality and completeness of phenotype and exposure measures to be provided to GAIN, the strength of the evidence for a genetic component for the trait, the anticipated size of a genetic effect and power to detect it, the advantages in terms of strategies for data management and data analysis and the advantages for replication studies and follow-up studies to identify the causative genetic variant(s). Applicants were provided with a brief summary of reviewers' comments after the review.

Based on initial peer review results, the GAIN Steering Committee identified a subset of applications to be considered for participation in GAIN. These were referred to a technical advisory group (TAG) composed of technical experts from academia, government and industry with expertise in genetics, epidemiology, bioethics, biostatistics and GWA studies. The TAG conducted an in-depth examination of the highest-priority applications, worked with the applicants to optimize the suitability of their proposed design for GAIN and provided recommendations to the Steering Committee to support the final selection process. Unlike the typical NIH peer review process, the TAG could recommend that FNIH negotiate inclusion of samples and data in addition to, or instead of, those proposed in the initial application as well as recommending other design modifications to maximize the scientific potential of successful applications.

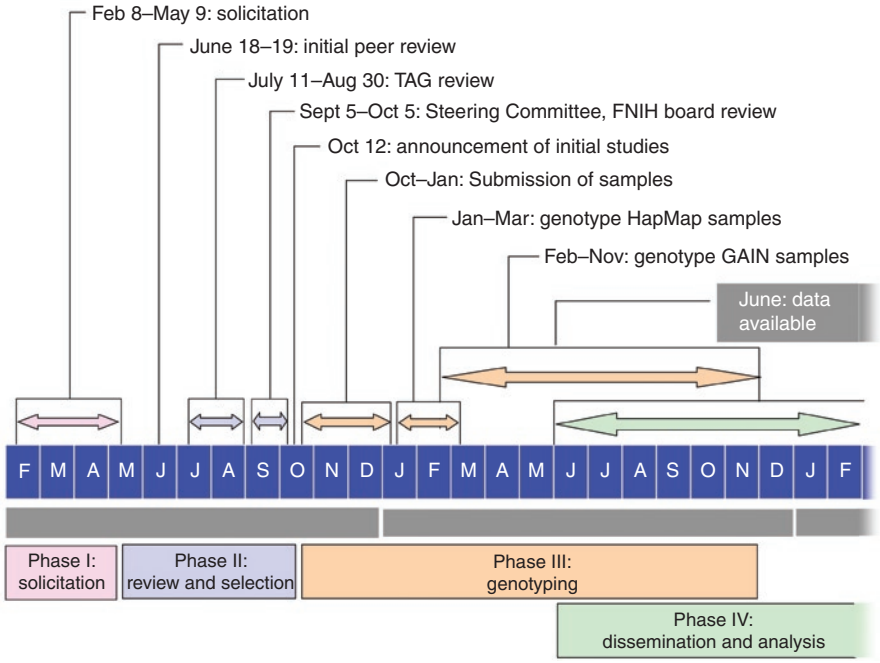

Figure 1 Timeline for GAIN solicitation, genotyping and release of data from initial 18,000 samples.

The TAG examined four areas in depth: (i) subject ascertainment, (ii) DNA specimen quality, (iii) power and analysis and (iv) data sharing and consent. Sample ascertainment criteria included sources, representativeness and comparability of cases and controls; reliability and comprehensiveness of trait definitions and phenotypic and exposure measures; diversity of ancestry, geographical region of origin and gender of study subjects; and potential overlap with ongoing studies. DNA specimen criteria included adequacy and consistency of DNA amount and concentration and specimen quality, as assessed by platform-specific SNP assay pass rates for roughly 3,200 SNPs in randomly selected samples submitted by each applicant. Power and analysis criteria included appropriateness of assumptions, impact of recommended reductions or increases in sample size, appropriateness of proposed analyses and availability of suitable follow-up samples. Data sharing and consent criteria included restrictions on the research use of data and samples (such as limitation to a single disease or to noncommercial investigators), adequacy of the procedure for obtaining consent for GWA genotyping and data sharing and provisions for the withdrawal of samples at a subject's request.

The most common weaknesses identified by the TAG review related to (i) the adequacy of consents and approvals for broad data sharing and (ii) the completeness and reliability of the phenotypic and exposure data proposed for inclusion in GAIN. Consent documents in several applications were problematic because of specific language restricting data use to a single investigator or site or excluding commercial users, or because of failure to document discussion of genetic research, data sharing, options for withdrawal or potential risks to participants. A smaller number of applications could not be included because of lack of availability of critical phenotypic or exposure data on controls, cases or both.

The Steering Committee, guided by the TAG's recommendations, considered criteria such as the relative public health impact of the diseases studied, overall diversity of ancestry and geographical region of origin of samples and availability of other existing or pending GWA scans in allocating the 18,000 available genotyping 'slots' for the first round of GAIN genotyping. The six studies recommended to, and approved by, the FNIH Board of Directors included four studies of mental health disorders (attention deficit hyperactivity disorder (ADHD), bipolar I disorder, major depressive disorder (MDD) and schizophrenia) and one study each of diabetic nephropathy and psoriasis (**Table 1**). The prominent representation of mental health studies reflects the large

Table 1 Investigators, conditions and samples included in initial GAIN genotyping

PI	Institution	dbGaP accession number	Condition	Key secondary phenotypes	Countries of origin of participants	Number of samples	Detectable genotype relative risk at DAF 20% and 50%[a]	Anticipated completion of genotyping
S.F.[21]	State University of New York Upstate Medical University	phs000016	ADHD	Quantitative measures of conduct disorder, sleep problems, autism spectrum, emotional lability	UK, Republic of Ireland, Germany, Belgium, Spain, Switzerland, Netherlands, Israel	959 offspring; 1,918 parents	1.59, 1.49	June 15, 2007
J.W.[22]	Joslin Diabetes Center	phs000018	Diabetic nephropathy	Serum creatinine and cystatin C	USA	905 cases; 890 controls[b]	1.61, 1.51	September 1, 2007
P.S.	University of North Carolina Chapel Hill	phs000020	Major depressive disorder	Personality traits of neuroticism and extroversion	Netherlands	1,860 cases; 1,860 controls	1.40, 1.33	September 15, 2007
G.A.	University of Michigan	phs000019	Psoriasis	Age at onset, type (plaque, guttate, inverse, pustular, erythrodermic), percentage body surface area affected, location, severity grading, nail and joint involvement	USA	1,449 cases; 1,453 controls	1.46, 1.38	October 1, 2007
P.G.[23]	Evanston Northwestern Healthcare	phs000021	Schizophrenia (European Americans)	Dimensional ratings of psychosis and mood disorder	USA and Australia	1,440 cases; 1,469 controls[c]	1.46, 1.38	December 1, 2007
			Schizophrenia (African Americans)			1,280 cases; 1,000 controls[c]	1.52, 1.43	
J.K.[24]	University of California San Diego	phs000017	Bipolar I disorder (European Americans)	Complete DIGS interview; various temperament, personality, circadian rhythm and trauma questionnaires	USA	1,158 cases; 1,158 controls[c]	1.53, 1.44	December 1, 2007
			Bipolar I disorder (African Americans)			500 cases; 500 controls[c]	1.88, 1.74	

PI, principal investigator. S.F., Stephen Faraone; J.W., James Warram; P.S., Patrick Sullivan (see http://www.tweelingenregister.org and http://www.nesda.nl); G.A., Gonçalo Abecasis (see http://www.sph.umich.edu/csg/abecasis/CASP/); P.G., Pablo Gejman; J.K., John Kelsoe. [a]Multiplicative genotype relative risks detectable with 80% power assuming 1% disease prevalence, 20% or 50% disease allele frequency (DAF) and type I error $\leq 10^{-7}$ (http://www.sph.umich.edu/csg/abecasis/CaTS/). Note that assuming a prevalence of 1% leads to slightly conservative power calculations when the true disease prevalence is higher, as risk alleles for high prevalence conditions can be detected with higher power for any given relative risk. [b]Support for genotyping and data deposition of the diabetic nephropathy study was shared with the National Institute of Diabetes and Digestive and Kidney Diseases; 453 cases and 445 controls will be genotyped by GAIN. [c]Controls for the bipolar disorder study are a subset of those for the schizophrenia study.

number and high quality of such applications submitted to GAIN, arising in part from the extensive foundation for genetic studies laid by the National Institute of Mental Health, which supported three of the six studies selected for GAIN. Inclusion of related disorders such as these provides a valuable opportunity for cross-study collaborations, such as analysis of mood disorders or psychosis, that are facilitated by the interactive nature of GAIN. Sharing of control subjects across the bipolar disorder and schizophrenia studies permits an assessment of this approach (which was used effectively by the WTCCC[10]), and when controls have consented to broader uses of their data, as the WTCCC controls and some GAIN controls have done, this provides a valuable resource for other types of medical and genomic research.

Two studies (on ADHD and MDD) involve European population samples, the schizophrenia study includes US and Australian population samples and the remaining three studies involve only US samples. Declining costs of genotyping and increasing recognition of the importance of large sample sizes led the Steering Committee to recommend increases in size above the originally anticipated 1,000 cases and 1,000 controls for the three studies with additional samples available (MDD, psoriasis and schizophrenia). In addition, emphasis on diversity led to the inclusion of African American samples originally proposed as replication samples in the bipolar disorder and schizophrenia studies. All studies had at least 80% power to detect genotype relative risks >1.6, assuming 1% disease prevalence, type I error $<10^{-7}$ and 20% disease allele frequency except for the smaller study of bipolar disorder in African Americans. Minimum detectable odds ratios are smaller, as odds ratios overestimate relative risks[15]. Follow-up analyses of candidate SNPs identified by the GAIN studies but attaining lower levels of significance will allow identification of associated SNPs with smaller effects. Finally, improvements in genotyping technology led the Steering Committee to recommend increasing the density of genetic markers to capture a greater proportion of genomic variation.

Genotyping and quality control

Perlegen Sciences and the Broad Institute's Genetic Analysis Platform are performing genotyping. Perlegen Sciences is using a proprietary, high-density oligonucleotide array-based platform with roughly 480,000 SNPs for the ADHD, MDD and psoriasis studies (http://www.perlegen.com/index.htm?science/HT_SNP_genotyping.html). The Broad Institute is using the Affymetrix SNP Array 5.0 platform with roughly 470,000 SNPs and 400,000 amplicons for copy number variants for the nephropathy study (http://www.affymetrix.com/support/technical/datasheets/genomewide_snp5_datasheet.pdf). For the schizophrenia and bipolar disorder studies, which include substantial numbers of African American subjects, the Broad Institute will use the SNP Array 6.0 platform with roughly 930,000 SNPs and 900,000 amplicons to ensure adequate genomic coverage in this population with lower levels of linkage disequilibrium genome-wide.

We assessed the quality of the genotype data produced by these three newly developed platforms by genotyping all 270 HapMap phase II samples on each platform. Comparison with the extensive genotyping already performed on these samples by the HapMap Project will also be used to facilitate cross-platform analyses.

Box 1 Guiding principles of GAIN

GAIN will use the most rigorous scientific approaches and maintain the highest ethical standards, as guided by the following principles:

- The greatest public benefit will be achieved if GAIN results are made immediately available for research use by any interested and qualified investigator or organization, within the limits of providing appropriate protection of research participants.
- Discovery of genetic variants related to health and disease and their translation into effective diagnostic, therapeutic and preventive strategies should be expedited.
- The best available human studies of diseases and traits, chosen to achieve programmatic balance among diseases, should be used for this discovery process.
- Findings supported or enabled by GAIN should be relevant and applicable to all population subgroups and segments of society.
- Investigators granted access to GAIN data should ensure confidentiality of study participants and follow any limitations specified by their informed consent.
- Intellectual contributions and efforts of investigators submitting samples should be appropriately recognized by any user of GAIN data, consistent with the principles that guide the use of other community resource projects within the genomics field.
- Access to GAIN data should be made available to GAIN partners and contributing investigators and other users at the same time and through the same access approval mechanisms.

Data from the Perlegen Sciences high-density array and the Affymetrix 5.0 and 6.0 SNP arrays are available at ftp://ftp.ncbi.nih.gov/dbgap/GAIN/genotypeQC/ and show excellent genomic coverage, call rates and concordance when compared with the HapMap Release 22 data for concordance and the phased HapMap Release 21 data for coverage (**Supplementary Table 1** online).

For each study, quality assessment samples will include a small number of reference HapMap samples, duplicates of study samples and samples from parents of study participants, where available, recognizing the limitations on the total number of samples to be genotyped. All data generated by the genotyping centers will be released (except for data from misidentified samples or those with possible pedigree errors), including the genotype calls and allelic intensity scores, quality scores to measure the confidence of genotype calls and image files to allow other users to apply alternative genotype calling algorithms. Using the genotype calls generated by Perlegen and the Broad Institute, NCBI will assess each genotyped SNP for call rate, minor allele frequency, Hardy-Weinberg equilibrium statistics, concordance with HapMap genotypes, concordance among internal duplicates and patterns of allelic segregation in the population and within trios. NCBI will also assess each genotyped sample for call rate, degree of heterozygosity, relatedness to other samples and sex misidentification. These statistics will be available to all users of the GAIN data and will also be used to generate a high-quality, filtered data set of samples and SNPs (**Box 2**).

Organization of the GAIN Collaborative Research Group

The GAIN organizational structure includes a Steering Committee, responsible for overall guidance of the project, and three key subcommittees (**Supplementary Note** online). The Principal Investigators' Group discusses study progress and policies, data access and proposed study-wide publications and presentations. The Genotyping Group has developed genotype data standards, oversees genotype data quality assessment and will ensure that data have been handled correctly for situations such as proper assignment of strands to alleles, loci on sex chromosomes and family samples. The Analysis Methodology Group compares analysis methods across the six GAIN studies, suggesting common approaches to analyses within each project as needed to enhance comparability, recommending analyses across multiple projects as appropriate and discussing challenges in analysis and interpretation.

Contributing investigators from the six selected studies came together with the Steering Committee, the FNIH/NIH project team, project partners, outside experts and other interested researchers for the first GAIN Analysis Workshop on November 29–30, 2006. This workshop examined new approaches to genotyping and analysis in GWA studies, including matters relating to study-specific and cross-study design and analysis, as well as GAIN policy matters. A second workshop will be held in Bethesda, Maryland in October 2007, and a third is planned for 2008. GAIN analysis workshops are open to the entire scientific community, and presentation materials are available through the GAIN website (http://www.fnih.org/GAIN2/analysis_workshops.shtml).

Access to GAIN data

GAIN data access policies are designed to facilitate the discovery of genetic variants related to health and disease in a manner that respects the privacy and informed consent of the research participants from whom the data were derived. The potential identifiability of genotype-phenotype data[16] (see http://www.genome.gov/19519198) and the applicability of human subjects regulations to coded data[17] (see http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm) were considered carefully in developing the GAIN policies. The extensive genotype and phenotype information included in GAIN raises important questions about possible risks to the confidentiality of individual participants in broad data-sharing models. Thus, GAIN policies were developed with deliberate attention to participant protections, both during data submission from the original studies and during data access and use by outside investigators.

A key aspect of the protections provided at the data submission step was removal of potentially identifying information before data submission, as described at http://www.fnih.org/GAIN/Project_Data_Sets.shtml. GAIN requirements for de-identification of submitted data sets were very similar to those described within the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[18]. NCBI staff reviewed the submitted data sets for removal of these identifiers; those with disallowed information were returned to applicants for redaction. In addition, applicants and their institutions described any limitations on use of data submitted to GAIN based on participant consent and institutional review board approval. They also confirmed the appropriateness of the data for inclusion and distribution through GAIN in the Applicant Policy Agreement (http://www.fnih.org/GAIN2/APPLICANT%20POLICY%20AGREEMENT2.pdf). The completeness of these statements was confirmed through the pre-review assessment of submitted data by the NCBI and through the TAG review of informed consent documents. As needed for clarification, or when requested by applicants, the GAIN project team explored and resolved matters relating to study-specific participant protections with applicants and

their institutions before studies were accepted for GAIN genotyping.

Substantial participant protections are also applied at the user level through a controlled access data request process managed by the GAIN Data Access Committee (DAC). The DAC is composed of senior NIH staff with expertise in the diseases under study and in genetics, epidemiology, bioethics and human subjects concerns. Researchers interested in obtaining controlled-access GAIN data submit a Data Access Request (DAR), cosigned by their institution, constituting their agreement to abide by the principles and practices detailed in the GAIN Data Use Certification (http://www.fnih.org/GAIN2/Data_Use_Certification.pdf, and **Box 3**). These conditions include keeping the data secure by implementing standard data security practices (http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf) and using data only for the approved research purposes; acknowledging GAIN policies on publications and intellectual property; and submitting periodic reports on data use to the DAC. Data users also agree not to distribute individual-level data in any form to any third parties (other than their own research staff who have agreed to the terms of the Data Use Certification) and not to attempt to identify individual study participants.

The GAIN DAC reviews each access request, checks for any federal sanctions on the requester and reviews the research use statement to ensure the proposed use is consistent with any data use limitations. GAIN project staff and the DAC also monitor use of GAIN data through periodic reports from approved users (including any violations, inadvertent or intentional, of GAIN policies), surveys of the literature and interactions with GAIN principal investigators, approved users and journal editors. Activities of the DAC and the general conduct of GAIN data distribution and use are reviewed by an independent Data Use Review Board of scientists, statisticians, ethicists and public representatives from outside NIH. This expert panel provides input on policy questions related to GAIN data use and human subjects protection, privacy concerns, intellectual property matters, publication timing and other topics as appropriate.

The genotype data generated for each study sample, plus de-identified, coded phenotype and exposure data, will be deposited in the controlled access section of the GAIN database within dbGaP (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gap). DbGaP is an NIH database for genotype-phenotype data sets developed and maintained by the NCBI. Each data set accepted for distribution through dbGaP is assigned an accession number (a permanent public unique identifier that is not reused) that can be used to reference a data set for bibliographic purposes.

DbGaP provides two levels of access—open and controlled—to allow broad release of non-sensitive summary data while providing oversight and investigator accountability for sensitive data sets involving individual-level genotype and phenotype data. Descriptive summaries of studies and measured variables, cluster plots for each SNP and original study documents will be indexed for searching by the general public through the open access portion, while access to individual-level data, including phenotypic data tables, genotype calls, gene probe intensity data, CEL files and pedigrees will require authorization through the controlled access process.

Simple, unadjusted genotype-phenotype association measures will be calculated and posted in the controlled access portion of the GAIN database to allow approved data users to check their initial analyses for consistency with these results. Posting these precomputed association data is also intended to discourage premature patent claims by placing the phenotype and genotype data and first-line analysis in the public domain. A variety of strategies for evaluating and correcting single-marker association distributions for genotype data quality, population substructure and cryptic relatedness are being explored and developed through intense collaborations among the Genotyping and Analysis Methodology Groups. The methods used to generate these results will be carefully described in distributed data sets and subsequent GAIN publications. GAIN investigators may choose to use different methods for genotype calling, imputation and association for their own study but have agreed to make their analyses available through the GAIN open or controlled access processes, as appropriate, after they publish a manuscript using the GAIN data.

Approved data users are expected to respect the exclusive rights of the GAIN contributing investigators (who designed the studies, recruited the participants and provided study samples and data) to publish their data within a reasonable timeframe. The names of approved users and the titles of their projects are listed on the open-access portion of the dbGaP website. GAIN project data sets will be deposited and made available to all approved users—including contributing investigators, outside users and GAIN partners—at the same time. However, for 9 months after the release of a specific GAIN project data set, only the contributing investigators have the right to submit abstracts and publications and to make presentations based on the data and samples they contributed to GAIN. During this period, approved users have access to the data set but agree not to submit for presentation or publication any results derived from it. At the end of the 9-month period for a given GAIN data set, which will be specified within the GAIN database, approved users may publicly discuss and submit papers for publication on GAIN data for any purpose consistent

Box 2 GAIN genotyping quality standards and procedures

- The Perlegen genotyping center will include one standard HapMap sample on half of the 96-well plates and one study sample duplicate (a different duplicate sample on each plate) on the other half of the plates, along with two parents of a study sample, when available, on each plate to make a trio for studies not based on trio samples
- The Broad Institute genotyping center will include one standard HapMap sample on each plate, along with one study sample duplicate (with a different duplicate sample on each plate) and two parents of a study sample.
- Samples with fewer than 90% of the SNPs called will be removed.
- At least 90% of the SNPs in a study will meet the following minimum data quality standards; actual data quality is expected to be much better:
    1. Hardy-Weinberg deviation P value > 0.00033 in any plate
    2. Call rate minimum = 90% and average ≥ 97%
    3. For HapMap quality assessment samples, average call rates for both heterozygotes and homozygotes ≥ 97%
    4. Concordance in duplicate samples of ≥99.5%
    5. Quality scores meet a pre-determined minimum level, to be decided for each study by the Genotyping Group
    6. Minor allele frequencies meet a pre-determined minimum level, to be decided for each study by the Genotyping Group

with the policies and practices of GAIN, the FNIH and the NIH. Investigators using GAIN data will acknowledge GAIN, the contributing investigators and the funding organization that supported the contributing study in any resulting oral or written presentation, disclosure or publication. Although not a requirement for data access, data users are encouraged to collaborate with the contributing investigators to maximize efficiency and scientific productivity.

A key component of GAIN infrastructure lies in its intellectual property policies, which are modeled on the recommendations cited in NIH's Best Practices for the Licensing of Genomic Inventions (http://ott.od.nih.gov/NewPages/LicGenInv.pdf, accessed May 7, 2007) and the NIH Research Tools Policy (http://ott.od.nih.gov/policy/rt_guide_final.html, accessed May 7, 2007). The GAIN policies promote broad freedom of operation for all users of GAIN data by rapidly placing data in the public domain and by encouraging the initial genotype-phenotype associations identified through GAIN to remain unencumbered by intellectual property claims. The filing of patent applications in a manner that might restrict use of GAIN data could substantially diminish the public benefit provided by these community resources. Approved users, including GAIN principal investigators and their affiliated organizations, will be asked to acknowledge the GAIN Intellectual Property Policy of not pursuing intellectual property protections that would prevent or block access to or use of GAIN data or conclusions drawn directly from these data. This approach is expected to discourage premature claims on pre-competitive information while promoting opportunities to develop intellectual property and file appropriate claims on downstream discoveries.

Box 3 Conditions of GAIN data access acknowledged by approved data users and their institutions

- Use of GAIN data will be limited for a given data set to the parameters described within the GAIN database for appropriate research use of the data.
- Uses of the data must be consistent with federal, state and local laws and any relevant institutional policies.
- Data will not be used to identify or contact individual participants from any GAIN study.
- Individual participant data will be kept secure and will not be distributed.
- Names of approved users and titles of their projects will be listed on the open access GAIN website.
- Approved users will submit regular reports on research progress, including any matters involving data security.
- GAIN-supported data and the conclusions drawn directly from them should remain freely available, without requirement for licensing or undue intellectual property encumbrances.
- No abstracts, presentations or publications based on GAIN data will be submitted or presented by anyone other than the original study investigators for 9 months after the GAIN project data set is released for research use.
- Presentations and publications after this 9-month period will acknowledge GAIN, the investigators who contributed the phenotype data and DNA samples from their original study and the primary funding organization that supported the contributing study.

Conclusion

Genome-wide association studies hold tremendous promise for unraveling the genetics of complex diseases, but the genomics community faces enormous challenges in analyzing data sets of billions of genotypes, distinguishing the small number of true positive associations and following their leads to causative variants and effective interventions. GWA studies, almost regardless of the trait(s) under study, present many common challenges in analysis and interpretation that are likely to have common solutions. These solutions, and the potential for combining phenotype and genotype data across studies to enhance statistical power, are best developed through collaborative approaches such as GAIN, in which significant data resources can be developed and made available for the collective benefit of the entire scientific community. The power of this approach was recently demonstrated in the WTCCC, where accuracy of genotype calling was substantially increased by combining results across studies[10] and where common loci were identified in diseases that appear to be dissimilar, such as type 1 diabetes and Crohn's disease[19,20].

Sharing these data outside the GAIN collaborative group, and encouraging the scientific community to use these data responsibly and participate actively in their analysis, should speed the identification of variants related to complex diseases and the development of effective new treatments. Extensive sharing should also facilitate replication of initial GWA findings and development of new analytical methods to maximize the knowledge obtainable through GWA studies. However, access to GAIN data carries significant responsibilities, particularly in protecting the confidentiality and respecting the informed consent of the study participants and also in ensuring that the use of these data is not restricted by premature claims on intellectual property. GAIN represents a new experiment in these policy areas as well as in the scientific pursuit of the genetics of complex diseases. The entire scientific community is invited to participate.

Accession codes. dbGaP accession numbers are as follows: for ADHD, phs000016; for diabetic nephropathy, phs000018; for major depressive disorder, phs000020; for psoriasis, phs000019; for schizophrenia, phs000021; and for bipolar I disorder, phs000017.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors express their appreciation to P. de Bakker for providing estimates of genomic coverage for the GAIN genotyping platforms. The Broad Institute Center for Genotyping and Analysis is supported by grant U54 RR020278-01 (S. Gabriel, principal investigator) from the National Center for Research Resources.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/

Published online at http://www.nature.com/naturegenetics

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
2. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
3. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
4. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
5. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
6. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
7. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
8. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).

9. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
10. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
11. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
12. Wang, W.Y.S., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
13. Duerr, R.H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
14. Chanock, S.J. *et al.* NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
15. Schlesselman, J.J. *Case-Control Studies: Design, Conduct, and Analysis* 33–34 (Oxford University Press, New York, 1982).
16. McGuire, A.L. & Gibbs, R.A. Genetics. No longer de-identified. *Science* **312**, 370–371 (2006).
17. Department of Health and Human Services. 45 CFR 46 Subpart A: Federal Policy for the Protection of Human Subjects (revised). *Fed. Regist.* **70**, 36325–36328 (2005).
18. Department of Health and Human Services. 45 CFR 164.514: Other requirements relating to uses and disclosures of protected health information. *Fed. Reg.* **65**, 82818–82819 and **67**, 53270 (2000).
19. Todd, J.A., *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).
20. Parkes, M., *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**, 830–832 (2007).
21. Kuntsi, J., Neale, B.M., Chen, W., Faraone, S.V. & Asherson, P. The IMAGE project: methodological issues for the molecular genetic analysis of ADHD. *Behav. Brain Funct.* **2**, 27 (2006).
22. Mueller, P.W. *et al.* Genetics of Kidneys in Diabetes (GoKinD) Study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *J. Am. Soc. Nephrol.* **17**, 1782–1790 (2006).
23. Suarez, B.K. *et al.* Genomewide linkage scan of 409 European-Ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. *Am. J. Hum. Genet.* **78**, 315–333 (2006).
24. Dick, D.M. *et al.* Genomewide linkage analyses of bipolar disorder: A new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative. *Am. J. Hum. Genet.* **73**, 107–114 (2003).

The complete list of authors (the GAIN Collaborative Research Group) is as follows: Teri A Manolio[1], Laura Lyman Rodriguez[1], Lisa Brooks[1], Gonçalo Abecasis[2], the Collaborative Association Study of Psoriasis, Dennis Ballinger[3], Mark Daly[4], Peter Donnelly[5], Stephen V Faraone[6], the International Multi-Center ADHD Genetics Project, Kelly Frazer[3,7], Stacey Gabriel[4], Pablo Gejman[8], the Molecular Genetics of Schizophrenia Collaboration, Alan Guttmacher[1], Emily L Harris[1], Thomas Insel[9], John R Kelsoe[10], the Bipolar Genome Study, Eric Lander[4], Norma McCowin[11], Matthew D Mailman[12,13], Elizabeth Nabel[14], James Ostell[13], Elizabeth Pugh[15], Stephen Sherry[13], Patrick F Sullivan[16], the Major Depression Stage 1 Genomewide Association in Population-Based Samples Study, John F Thompson[17], James Warram[18], the Genetics of Kidneys in Diabetes (GoKinD) Study, David Wholley[11], Patrice M Milos[19] & Francis S Collins[1]

[1]National Human Genome Research Institute, US National Institutes of Health (NIH), 31 Center Drive, Bethesda, Maryland 20892, USA. [2]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [3]Perlegen Sciences, 2021 Stierlin Court, Mountain View, California 94043, USA. [4]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [5]University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [6]State University of New York Upstate Medical University, 750 E. Adams Street, Syracuse, New York 13210, USA. [7]Scripps Research Institute, Genomic Medicine, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [8]Evanston Northwestern Healthcare and Northwestern University, 1001 University Place, Evanston, Illinois 60201, USA. [9]National Institute of Mental Health, NIH, 6001 Executive Drive, Bethesda, Maryland 20892, USA. [10]Departments of Psychiatry, University of California, San Diego and Veterans Administration San Diego Healthcare System, 9500 Gilman Drive, La Jolla, California 92093, USA. [11]Foundation for the NIH, 45 Center Drive, Bethesda, Maryland 20892, USA. [12]Genomic Medicine, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana 46285, USA. [13]National Library of Medicine, NIH, 45 Center Drive, Bethesda, Maryland 20892, USA. [14]National Heart, Lung, and Blood Institute, NIH, 31 Center Drive, Bethesda, Maryland 20892, USA. [15]Center for Inherited Disease Research, Johns Hopkins University, 333 Cassel Drive, Baltimore, Maryland 21224, USA. [16]University of North Carolina at Chapel Hill, 103 Mason Farm Road, Chapel Hill, North Carolina 27599, USA. [17]Pharmacogenomics, Pfizer Global Research and Development, Pfizer Inc., 1 Eastern Point Road, Groton, Connecticut 06340, USA. [18]Joslin Diabetes Center, One Joslin Place, Boston, Massachusetts 02215, USA. [19]Helicos BioSciences Corp., One Kendall Square, Cambridge, Massachusetts 02139, USA.