COMMENTARY

The Genotype-Tissue Expression (GTEx) project

The GTEx Consortium[*]

Genome-wide association studies have identified thousands of loci for common diseases, but, for the majority of these, the mechanisms underlying disease susceptibility remain unknown. Most associated variants are not correlated with protein-coding changes, suggesting that polymorphisms in regulatory regions probably contribute to many disease phenotypes. Here we describe the Genotype-Tissue Expression (GTEx) project, which will establish a resource database and associated tissue bank for the scientific community to study the relationship between genetic variation and gene expression in human tissues.

In the past decade, genome-wide association studies (GWAS) have documented a strong statistical association between common genetic variation at thousands of loci and more than 250 human traits[1]. Yet, the functional effects of most GWAS-implicated variants remain largely unexplained. The finding that nearly 90% of these sites occur outside of protein-coding sequences[2] suggests that many associated variants may instead have a role in gene regulation. The careful examination of gene expression and its relationship to genetic variation has thus become a critical next step in the elucidation of the genetic basis of common disease. Cell context is a key determinant of gene regulation; but, to date, the challenge of collecting large numbers of diverse tissues in humans has largely precluded such studies outside of a few easily sampled cell types.

Expression quantitative trait locus (eQTL) mapping offers a powerful approach to elucidate the genetic component underlying altered gene expression[3]. Studies primarily in blood, skin, liver, adipose and brain indicate that eQTLs are common in humans[4–6]. Genetic variation can also influence gene expression through alterations in splicing, noncoding RNA expression and RNA stability[7–9].

eQTLs regulating nearby or distant genes are commonly referred to as *cis* eQTLs and *trans* eQTLs, respectively[3]. Gene expression is differentially regulated across tissues, and many human transcripts are expressed in a limited set of cell types or during a limited developmental stage. Several studies have reported tissue-specific eQTLs[10,11], and combining eQTL studies with network analyses across multiple tissues has helped to define complex networks of gene interaction[12,13].

Complementing eQTL data with information on other molecular phenotypes, for example, from epigenomic assays[14], on the same tissues and linking to resources such as the Encyclopedia of DNA Elements (ENCODE)[15] will provide a powerful means of dissecting gene-regulatory and higher-order networks across multiple tissues. Analyzing multiple tissues will be important because evaluation of the functional consequences of a disease-associated SNP is ideally performed in a disease-relevant cell context. However, for most tissue types, human biospecimens are very difficult to obtain from living donors (for example, brain, heart and pancreas), and most eQTL studies so far have been performed with RNA isolated from immortalized lymphoblasts or lymphocytes[6] and a few additional readily sampled tissues.

To fully enable this critical next step in the study of the genetic basis of common disease, it will be of enormous value to have a resource of blood samples from individuals who have been comprehensively genotyped

(and eventually completely sequenced), with genotyping data linked to genome-wide gene expression patterns across a wide range of tissue types. Initially, this resource would enable the research community to perform a comprehensive search for eQTLs (both tissue-type specific and across tissue types) and establish their association with disease-associated variants from GWAS or sequencing studies. Eventually, as other molecular phenotypes are added, the relationship between genetic variation and gene expression could expand to include correlations with epigenetics and proteomics data as well as other molecular characteristics. Although such a catalog would have been unthinkable a few years ago, new genomic technologies are now making the problem approachable.

This convergence of unmet scientific need and new technologies prompted a US National Institutes of Health (NIH) workshop held in June 2008 to discuss the advisability and feasibility of a large-scale public resource for human genetic variation and gene expression across tissues. On the basis of the output from this workshop and ongoing consultation, the NIH developed the concept of the GTEx project (**Box 1**). Many of the specifics of the pilot project described here were contributed by funded investigators and were influenced by early, experimental biospecimen collections.

Design of the GTEx project
The GTEx project of the NIH Common Fund aims to establish a resource database and asso-

[*]*A full list of members and affiliations appears at the end of the paper. Correspondence should be addressed to G. Getz (gadgetz@broadinstitute.org), J.P.S. (struewij@mail.nih.gov) or W.W. (winckler@broadinstitute.org).*

BOX 1 GOALS OF THE GTEx PROJECT

- To create a data resource to enable the systematic study of genetic variation and the regulation of gene expression in multiple reference human tissues
- To provide the scientific community with a biospecimen resource including tissues, nucleic acids and cell lines upon which to determine other molecular phenotypes
- To support and disseminate the results of a study of the ethical, legal and social issues related to donor recruitment and consent
- To support the development of novel statistical methods for the analysis of human eQTLs, alone and in the context of other molecular phenotypes
- To make data available to the research community as rapidly as possible
- To support the dissemination of knowledge, standards and protocols related to biospecimen collection and analysis methods developed during the project

ciated tissue bank in which to study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues (**Supplementary Fig. 1**). The GTEx project began with a 2.5-year pilot phase to test the feasibility of establishing a rapid autopsy program that would yield high-quality nucleic acids and robust gene expression measurements. Having met milestones of donor enrollment, RNA quality and eQTL findings, the project is scaling up to include approximately 900 postmortem donors by the end of 2015. The power to detect eQTLs is dependent on multiple factors that are difficult to quantify precisely, but power estimates over a range of effect sizes and allele frequencies are described (**Fig. 1**).

GTEx donors are identified through low-PMI (post-mortem-interval) autopsy or organ and tissue transplantation settings. To compare the quality of results for tissues derived from autopsy and surgery, a small subset of tissue types routinely discarded during surgical amputation, such as skin, fat and muscle, are also collected. In addition, peripheral blood samples are collected and used as both a source of DNA for whole-genome SNP and copy number variant (CNV) genotyping and to establish lymphoblastoid cell lines. Skin samples are collected from the same region of the lower leg, both for measurement of gene expression and to establish fibroblast cultures. Quantification of gene expression is performed primarily through massively parallel sequencing of RNA, but some pilot-phase tissues were analyzed both by sequencing and by gene expression array to enable a comparison of the results with different technologies. eQTLs are identified and will be made accessible to the scientific community through the National Center for Biotechnology (NCBI) GTEx database and a GTEx data portal. In addition, GTEx raw data will be made available through the database of Genotypes and Phenotypes (dbGaP) on a periodic basis.

GTEx project structure during the pilot phase (**Supplementary Fig. 2**) included entities for biospecimen acquisition, processing, storage and verification; a study on ethical, legal and social issues (ELSI study); the Laboratory, Data Analysis and Coordinating Center (LDACC); the GTEx-eQTL browser; novel statistical methods development grants; and a brain bank. The scale-up is organized similarly to the pilot; the current structure of the project and information on funding opportunities are available from the NIH Common Fund website.

Biospecimen acquisition

These functions are designed and organized under the Cancer Human Biobank (caHUB) of the National Cancer Institute. caHUB has enrolled under contract several Biospecimen Source Sites (BSSs), a Comprehensive Biospecimen Resource (CBR), a Comprehensive Data Resource (CDR) and pathology and quality management teams to perform acquisition of biospecimens and associated data. Details on all standard operating procedures for donor enrollment and sample collection are available from the caHUB website.

Donors of either sex from any ancestry group are eligible if they are aged 21–70 and if biospecimen collection can start within 24 h of death. There are few medical exclusionary criteria: human immunodeficiency virus (HIV) infection or high-risk behaviors, viral hepatitis, metastatic cancer, chemotherapy or radiation therapy for any condition within the past 2 years, whole-blood transfusion in the past 48 h or body mass index of >35 or <18.5. Each BSS collects, where feasible, aliquots from many predesignated tissue sites and organs (**Supplementary Table 1**), including the brain of deceased donors who were not on a ventilator for the 24 h before death. Immediately after excision, most aliquots are stabilized in a solution containing alcohols (ethanol and methanol), acetic acid and a soluble organic compound that fixes primarily by protein precipitation (PAXgene Tissue, Qiagen) and shipped to the CBR. Only blood samples and

full-thickness skin biopsies are sent unfixed to the LDACC for cell line initiation. The majority of the brain and brainstem are also left unfixed and shipped overnight on wet ice to a brain bank. Further details of donor recruitment and sample collection, including standard operating procedures, are available through caHUB.

Pathology review and clinical annotation

At the CBR, an aliquot from each sampled tissue is paraffin embedded, sectioned and stained for histological analysis. A dedicated team of pathologists reviews slides from all tissue specimens to verify the organ source and to characterize both general pathological characteristics, such as autolysis, as well as organ-specific pathological states and inflammation. Of course, not all organs will be entirely normal, but donor eligibility is broad and is not restricted to specific diseases or conditions, and it is expected that many organs will be free of major disease processes. An aliquot of each tissue, fixed and stabilized in PAXgene Tissue solution but without paraffin embedding, is sent to the LDACC for molecular analysis. Policies and systems for
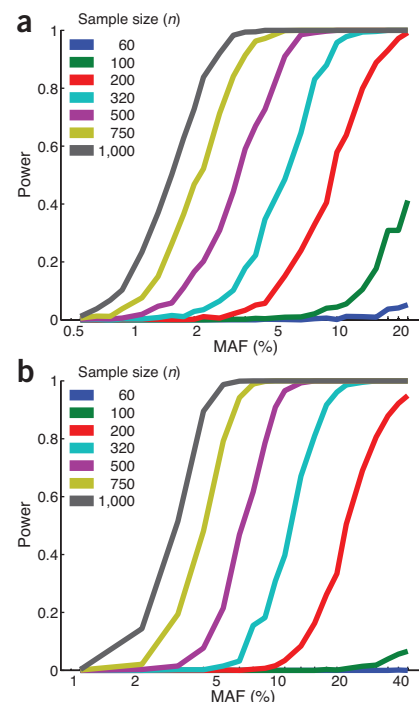
Figure 1 Effect of sample size and MAF on power to detect eQTLs. (**a**) Power for *cis*-eQTL analysis in which we assume $\alpha = 0.05/200{,}000$, reflecting Bonferroni correction for 200,000 hypotheses based on 20,000 genes and an average of 10 non-redundant SNPs in the region ±100 kb of each gene. (**b**) Power for *trans*-eQTL analysis in which we test 20,000 genes against 5 million SNPs in a total of 1×10^{11} tests with $\alpha = 5 \times 10^{-13}$.

accessing stored tissue samples are currently being developed. The CBR's histological sections are viewed as digitally scanned images, allowing precise annotations to be made to indicate where downstream studies, for example, tissue microarray and laser-capture microextraction, on selected portions of a specimen can focus (for example, lymphoid nodules in the ileal mucosa or squamous epithelium in the esophageal mucosa).

The clinical data collected for each GTEx donor belong to one of two categories: donor-level data or sample-level data. Donor-level data encompass all clinical measures of the donor, which include basic demographics, medication use, medical history, laboratory test results and the circumstances surrounding the donor's death. These data are collected from the donor (surgical biospecimens) or next of kin (post-mortem biospecimens) and verified against the donor's medical record, when readily available. Summary frequency distributions for clinical variables are available in dbGaP. Sample-level data are attributes belonging to each sample collected and include the tissue type, ischemic time, comments from the prosector and pathology reviewer, and process metadata such as batch ID and the amount of time spent in PAXgene Tissue fixative. Both donor- and sample-level data are examined for quality and completeness before being released.

Brain bank
Aliquots from single regions of the cortex and cerebellum are sampled and preserved in PAXgene Tissue at the BSS, and the remaining whole brain, with attached brain stem and cervical spinal cord, when possible, is shipped on wet ice to an NIH-funded brain bank. After sectioning of the brain at the brain bank, frozen samples from additional anatomical regions of the brain are analyzed at the LDACC, and the remaining brain is banked for future uses.

LDACC
The LDACC performs nucleic acid extractions and quality assessment, DNA genotyping and RNA expression analysis. The LDACC integrates results from the molecular analysis with phenotype data, performs eQTL analysis, deposits data into dbGaP and provides a portal for open-access data, standard operating procedures for sample processing and data generation, and results.

DNA is genotyped using the Illumina HumanOmni5M-Quad BeadChip to collect whole-genome SNP and CNV information from each donor's blood sample (or an alternate tissue, if blood is unavailable). The Illumina assay contains over 4 million probes,

with robust coverage of both SNPs and CNVs. DNA is also characterized using the Illumina Infinium HumanExome BeadChip to obtain high-quality SNP calls in coding regions.

A portion of each tissue is processed for RNA and DNA extraction, quantification and quality assessment. Extraction protocols that preserve both mRNA and microRNA are being used and are available from the data portal. For measurement of gene expression during the pilot phase, the LDACC analyzed approximately 1,000 samples using both microarrays (Affymetrix Human Gene 1.1 ST Array) and next-generation RNA sequencing (Illumina HiSeq 2000) to establish the comparability of these methods using post-mortem tissue. RNA sequencing (RNA-seq) uses a 76-base, paired-end Illumina TruSeq RNA protocol, averaging ~50 million aligned reads per sample. This read depth was selected to maximize sequencing value with the budget available and should make it possible to accurately measure moderately expressed transcripts, as well as some with low-level expression, but will have limited ability to accurately quantify rare transcripts and splice isoforms. It should provide gene expression measurements that have equivalent or better accuracy than those obtained with expression arrays and should include a higher dynamic range (with coefficient of variation < 0.1 for at least 12,000 genes; **Supplementary Fig. 3**). RNA-seq allows one to evaluate allele-specific expression in heterozygous individuals, improving the power to identify *cis* regulatory variants. With the target depth of 50 million aligned reads, we expect to have sufficient power to detect allele-specific expression in the top tertile of expressed genes (**Supplementary Fig. 4** and **Supplementary Note**). As the cost of RNA-seq drops, greater read depth will be possible, but, with current resources, the strategy is to maximize the number of samples analyzed.

The fresh-blood and full-thickness skin samples are used to establish Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines and primary fibroblast cell lines. Because many existing human eQTL studies have used EBV-immortalized cell lines, having these lines in addition to all the other peripheral tissues will allow researchers to evaluate the limitations of using only a lymphoblastoid cell line.

GTEx-eQTL browser
eQTLs are available and can be queried in browsers hosted both at the LDACC GTEx portal and at NCBI, who will verify the eQTL results provided by the project and both display them and make them available to other genome browsers and the scientific community.

Statistical analysis development
To promote the analysis of eQTL results across a wide range of human tissues, the NIH funded five centers to develop improved methods for statistical analysis. Investigators funded through this request for applications (RFA) form an analysis consortium that will provide innovative approaches for the analyses of GTEx data and other similar data sets. Investigators also collaborate with the LDACC to perform data quality assessment and quality control before release of the data into dbGaP. The initial GTEx Consortium publications, anticipated in 2013, will include genome-wide analysis of *cis* and *trans* eQTLs, allele-specific expression and splicing quantitative trait loci and a comparison of gene expression results obtained by array and RNA-seq.

Sample access and molecular analyses
The NIH is interested in making maximal use of this unique biospecimen resource, rich with clinical and genomic information. An access system, including mechanisms for requesting samples, is under development. Except for the fibroblast and lymphoblastoid cell lines, biospecimens are of limited quantity and are non-renewable. Potential uses that are comprehensive (for example, genomic versus single gene or small gene network and proteomic versus single protein or small protein network) and complementary to existing gene expression and variation data are preferred. Scientific questions that are equally well addressed using other sample sets will probably not be suitable, whereas those that take full advantage of the unique aspects of GTEx data, such as the multiple tissues from each donor and the gene expression information, are particularly sought. All data resulting from the analysis of GTEx samples must be made widely available to the scientific community. In addition to scientific review, all proposals to use GTEx samples would also go through a Biospecimen Access Committee (currently being formed).

Power analysis
To set expectations and guide the design of the full GTEx project, we built a framework to evaluate the statistical power to detect eQTLs. Statistical power depends on various parameters, some known more accurately than others. These parameters include the number of donors, the eQTL effect size and the presence of noise, as well as the significance threshold selected, which is chosen on the basis of the number of hypotheses tested. Assuming we are testing the *cis* eQTL effects of the 10 nonredundant SNPs (on average) in the vicinity (±100 kb of the start site) of each of 20,000 genes, the overall number of hypotheses is

200,000. Therefore, using Bonferroni correction, we set the significance threshold α to 0.05/200,000. For *trans*-eQTL analysis, a conservative estimate of α is ~5×10^{-13} (20,000 transcripts tested against 5 million loci). We model the expression data as having a log normal distribution with a log standard deviation of 0.13 within each genotype class (AA, AB, BB). This level of noise is based on estimates from initial GTEx data. The effect size depends both on the minor allele frequency of the SNP (MAF) and the actual log expression change between genotype classes (Δ). **Figure 1a** shows the statistical power of *cis*-eQTL analysis, and **Figure 1b** shows *trans*-eQTL analysis, with each analysis using an ANOVA statistical test as a function of the number of subjects and MAF and assuming $\Delta = 0.13$ (equivalent to detecting a log expression change similar to the standard deviation within a single genotype class). A final GTEx resource of 900 or more donors would realistically yield ~750 samples of any given tissue, as not all organs are available for collection from each donor. At an effective sample size of 750, we would have 80% power to detect *cis* eQTLs with MAF as low as 2% and *trans* eQTLs with MAF as low as 4%. Statistical power may be higher using methods that leverage the fact that multiple tissues are collected and analyzed for each donor. Because the underlying parameters were merely rough estimates, we repeated power analysis with different values (10–20 SNPs and 20,000–100,000 transcripts) and showed that 80% power is achieved for MAFs between 3 and 4% for *cis* eQTLs. For *trans* eQTLs, this range in transcript numbers gives sufficient power with MAFs between 4 and 5% (**Supplementary Fig. 5**).

Data access and publication policy
GTEx is designated by NIH as a community resource and, as such, aims to share as much of the data (some of which will be unique and identifiable) as rapidly as possible, according to NIH guidelines. It is recognized that quantifying the risk of identifying a donor on the basis of genetic and other information lies on a continuum and is a complex issue dependent on many factors, such as the availability other sources of data and the evolution of analytical methods[16,17]. Sharing of any information unique to an individual carries a small but difficult-to-define risk of allowing identification of the donor, but this risk must be balanced with the benefits of data sharing to the advancement of science.

Some data from the GTEx project is openly available, meaning that it can be accessed directly through the Internet. However, to reduce risks of sharing potentially identifying data, some data elements are available to the scientific community only through a controlled-access system, dbGaP. Standard operating procedures, details of data collection instruments, histopathological interpretations, molecular data that do not provide direct genetic variation information (for example, data from expression arrays, summary sequence-based gene expression estimates stripped of variant information and eQTL results), laboratory processing variables (for example, cDNA library preparation methods) and a very limited set of medical and sociodemographic variables (for example, sex and age at death in intervals) will be openly available. The LDACC will host an open-access data portal, and specimen acquisition standard operating procedures and information on associated data collection instruments will be available through caHUB. Medical and other epidemiological information, molecular results that contain direct genetic variation information (for example, SNP genotyping files and RNA-seq reads) and summary results that allow accurate inference of allele frequencies[18] will be available only through controlled access. Direct HIPAA (Health Insurance Portability and Accountability Act) identifiers, such as dates that include the month and day, will not be available through either open or controlled access.

Implementation of these data release policies and processes is a topic of ongoing discussion and may need to be modified as risks of identifiability are better quantified for various data types and as the size of the study increases. After initial processing of raw data (such as sequence reads and genotyping files), basic data quality checks are completed by the LDACC and statistical methods investigators, and data are then released immediately through dbGaP. The first dbGaP data release, consisting of data from 62 individuals, occurred in June 2012. For the pilot phase of the project, which concluded in January 2013, the data set comprised genotype data from 190 individuals from whom 1,814 total tissues (from 47 separate tissue sites) were profiled by RNA-seq to a median depth of 80 million aligned reads. These data are in the process of being released to dbGaP, and we anticipate releasing data two to four times per year until the project is completed. We expect total enrollment to increase to over 400 by 2013, to over 700 in 2014 and to approximately 900 by the end of 2015.

The GTEx project falls under the Ft. Lauderdale meeting principles of rapid, pre-publication data release. These principles involve publication of a manuscript near the outset to describe the scope and vision of the project and plans to make data available. The continued success of rapid prepublication data release relies on the scientific community to respect the data producer's interest to publish a full analysis of the data first. Although others are free to analyze GTEx data immediately upon release, the GTEx Consortium envisions publication of both a comprehensive description of the sample acquisition and processing system and a series of genome-wide analyses of genetic variation and gene expression, as described for statistical analysis and the development of methods.

Ethical, legal and social issues
The GTEx project involves potentially sensitive recruitment, institutional review board (IRB) and consent issues, particularly for deceased donors and their families. The collection of biospecimens from deceased individuals is not legally classified as human subjects research under 45 CFR 46; nonetheless, the depth of the genetic information obtained from the specimens of deceased donors has direct implications for the families of the donors. In recognition of this understanding, sites were required to obtain written or recorded verbal authorization from next of kin for the participation of deceased donors in GTEx, typically through an addendum or modification to an existing authorization form for donation of tissues and organs for research. This authorization included statements common in consent forms, such as the intention to perform genetic analyses, establish cell lines and share data with the scientific community. Work under way is more closely identifying familial concerns and may result in modifications to authorization procedures. Living surgery donors participate only after full, written informed consent is obtained.

In addition, an ELSI study of the consent and authorization process is being carried out at one BSS to assess both the effectiveness of the process in informing participants of the risks and benefits of the study and its potential psychosocial effects on donors and/or their families. The ELSI study is fully integrated with biospecimen collection efforts and will be expanded during the scale-up of the GTEx program.

Conclusions
A large-scale GTEx resource will be a powerful tool in unraveling the complex patterns of genetic variation and gene regulation across diverse human tissue types. The GTEx project will aid in the interpretation of GWAS findings for translational research by providing data and resources on eQTLs in a wide range of tissues of relevance to many diseases. But the value of a large GTEx resource, especially one that includes other molecular phenotypes,

goes well beyond GWAS follow-up, by providing a deeper understanding of the functional elements of the genome and their underlying biological mechanisms.

URLs. Catalog of published GWAS, http://www.genome.gov/gwastudies; GTEx LDACC data portal, http://www.broadinstitute.org/gtex/; caHUB, http://cahub.cancer.gov/; caHUB standard operating procedures, http://biospecimens.cancer.gov/resources/sops/default.asp; GTEx project on dbGaP, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424; US GTEx project on NIH Common Fund, http://commonfund.nih.gov/GTEx/; GTEx on National Human Genome Research Institute, http://genome.gov/gtex/; NCBI GTEx eQTL Browser, http://www.ncbi.nlm.nih.gov/gtex/test/GTEX2/gtex.cgi/; Request for information for the GTEx project, http://grants.nih.gov/grants/guide/notice-files/NOT-RM-12-028.html; seeQTL, http://www.bios.unc.edu/research/genomic_software/seeQTL/; SCAN, http://www.scandb.org/newinterface/about.html; US NIH community resource policy for GWAS, http://gwas.nih.gov/03policy2.html; Sharing Data from Large-Scale Biological Research Projects: a System of Tripartite Responsibility (Wellcome Trust), http://www.wellcome.ac.uk/About-us/Publications/Reports/Biomedical-science/WTD003208.htm; US NIH GTEx working group members, http://commonfund.nih.gov/GTEx/members.aspx.

AUTHOR CONTRIBUTIONS
Biospecimen and data collection, processing, quality control, storage and pathological review. caHUB-BSS: J. Lonsdale, J.T., M. Salvatore, R.P., E.L., S. Shad, R.H., G.W., F.G., N.Y., B.F., M. Moser, E.K., B.G., K. Ramsey, S. Sullivan, J.B., H.M., J.S. and J.F. caHUB-ELSI study: L. Siminoff, H.T., M. Mosavel and L.B. caHUB-CBR: S.J., D.R., D. Maxim, D.F., P.H., E.C., B.B., L.T., E.H. and K.F. caHUB-PRC: L. Sobin, J.R. and P.B. caHUB-CDR: G.K., C.S., D.T., L.Q., K.G. and S.N. caHUB–Operations Management: S.B., A.Z., A. Smith, R.B., K. Robinson, K.V., D.B., M.C., N.D.-M., M. Kennedy, T.E., P.W. and K.E.

Laboratory analysis, data analysis and study coordination. K.A., W.W., G. Getz, D.D., D. MacArthur, M. Kellis, A.T., T.Y., E. Gelfand, M.D., Y. Meng and G. Grant. **Brain bank operations.** D. Mash, Y. Marcus and M.B. **Statistical methods development and data analysis.** J. Liu, J.Z., Z.T., E.T.D., T. Lappalainen, R.G., J. Monlong, M. Sammeth, D.K., A.B., S.M., M. McCarthy, M.R., J. Maller, I.R., A.N., F.W., A. Shabalin, N.J.C., D.L.N., E.R.G., H.K.I., A.K., J. Pritchard, M. Stevens, T.F. and X.W. **Database.** M.F., N.S., A. Sturcke and J. Paschal. **Program management.** J.M.A., E.L.W., L.K.D., E.D.G., J.P.S., G.T., S.V., J.T.B., E.J.T., M.S.G., C.N., A.A., D.C., T.R.I., S.E.K., A.R.L., P.K.B., T. Lehner, Y.Y., C.C.C., J.B.V., S. Sawyer, N.C.L., J.D. and H.F.M.

1. - Altshuler, D., Daly, M.J. & Lander, E.S. *Science* **322**, 881–888 (2008).
2. - Hindorff, L.A. *et al. Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
3. - Gilad, Y., Rifkin, S.A. & Pritchard, J.K. *Trends Genet.* **24**, 408–415 (2008).
4. Emilsson, V. *et al. Nature* **452**, 423–428 (2008).
5. Schadt, E.E. *et al. PLoS Biol.* **6**, e107 (2008).
6. - Stranger, B.E. *et al. Nat. Genet.* **39**, 1217–1224 (2007).
7. Pickrell, J.K. *et al. Nature* **464**, 768–772 (2010).
8. - Pickrell, J.K., Pai, A.A., Gilad, Y. & Pritchard, J.K. *PLoS Genet.* **6**, e1001236 (2010).
9. Borel, C. *et al. Genome Res.* **21**, 68–73 (2011).
10. Petretto, E. *et al. PLoS Comput. Biol.* **6**, e1000737 (2010).
11. Grundberg, E. *et al. Nat. Genet.* **44**, 1084–1089 (2012).
12. Zhong, H. *et al. PLoS Genet.* **6**, e1000932 (2010).
13. Zhao, E. *et al. Mamm. Genome* **20**, 476–485 (2009).
14. Bernstein, B.E. *et al. Nat. Biotechnol.* **28**, 1045–1048 (2010).
15. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
16. Craig, D.W. *et al. Nat. Rev. Genet.* **12**, 730–736 (2011).
17. Schadt, E.E., Woo, S. & Hao, K. *Nat. Genet.* **44**, 603–608 (2012).
18. Jacobs, K.B. *et al. Nat. Genet.* **41**, 1253–1257 (2009).

John Lonsdale[1], Jeffrey Thomas[1], Mike Salvatore[1], Rebecca Phillips[1], Edmund Lo[1], Saboor Shad[1], Richard Hasz[2], Gary Walters[3], Fernando Garcia[4], Nancy Young[5], Barbara Foster[6], Mike Moser[6], Ellen Karasik[6], Bryan Gillard[6], Kimberley Ramsey[6], Susan Sullivan[7], Jason Bridge[7], Harold Magazine[8], John Syron[8], Johnelle Fleming[8], Laura Siminoff[9], Heather Traino[9], Maghboeba Mosavel[9], Laura Barker[9], Scott Jewell[10], Dan Rohrer[10], Dan Maxim[10], Dana Filkins[10], Philip Harbach[10], Eddie Cortadillo[10], Bree Berghuis[10], Lisa Turner[10], Eric Hudson[10], Kristin Feenstra[10], Leslie Sobin[11], James Robb[11], Phillip Branton[12], Greg Korzeniewski[11], Charles Shive[11], David Tabor[11], Liqun Qi[11], Kevin Groch[11], Sreenath Nampally[11], Steve Buia[11], Angela Zimmerman[11], Anna Smith[11], Robin Burges[11], Karna Robinson[11], Kim Valentino[11], Deborah Bradbury[11], Mark Cosentino[11], Norma Diaz-Mayoral[11], Mary Kennedy[11], Theresa Engel[11], Penelope Williams[11], Kenyon Erickson[12], Kristin Ardlie[13], Wendy Winckler[13], Gad Getz[13,14], David DeLuca[13], Daniel MacArthur[13,14], Manolis Kellis[13,15], Alexander Thomson[13], Taylor Young[13], Ellen Gelfand[13], Molly Donovan[13], Yan Meng[13], George Grant[13], Deborah Mash[16], Yvonne Marcus[16], Margaret Basile[16], Jun Liu[17], Jun Zhu[18], Zhidong Tu[18], Nancy J Cox[19], Dan L Nicolae[19], Eric R Gamazon[19], Hae Kyung Im[19], Anuar Konkashbaev[19], Jonathan Pritchard[19,20], Matthew Stevens[19], Timothèe Flutre[19], Xiaoquan Wen[19], Emmanouil T Dermitzakis[21], Tuuli Lappalainen[21], Roderic Guigo[22], Jean Monlong[22], Michael Sammeth[22], Daphne Koller[23], Alexis Battle[23], Sara Mostafavi[23], Mark McCarthy[24], Manual Rivas[24], Julian Maller[24], Ivan Rusyn[25], Andrew Nobel[25], Fred Wright[25], Andrey Shabalin[25], Mike Feolo[26], Nataliya Sharopova[26], Anne Sturcke[26], Justin Paschal[26], James M Anderson[26], Elizabeth L Wilder[26], Leslie K Derr[27], Eric D Green[28], Jeffery P Struewing[28], Gary Temple[28], Simona Volpi[28], Joy T Boyer[28], Elizabeth J Thomson[28], Mark S Guyer[28], Cathy Ng[28], Assya Abdallah[28], Deborah Colantuoni[28], Thomas R Insel[29], Susan E Koester[29], A Roger Little[29], Patrick K Bender[29], Thomas Lehner[29], Yin Yao[29], Carolyn C Compton[30], Jimmie B Vaught[30], Sherilyn Sawyer[30], Nicole C Lockhart[30], Joanne Demchok[30] & Helen F Moore[30]

[1]National Disease Research Interchange, Philadelphia, Pennsylvania, USA. [2]Gift of Life Donor Program, Philadelphia, Pennsylvania, USA. [3]LifeNet Health, Virginia Beach, Virginia, USA. [4]Drexel University College of Medicine, Philadelphia, Pennsylvania, USA. [5]Albert Einstein Medical Center, Philadelphia, Pennsylvania, USA. [6]Roswell Park Cancer Institute, Buffalo, New York, USA. [7]Upstate New York Transplant Service, Buffalo, New York, USA. [8]Science Care, Inc., Phoenix, Arizona, USA. [9]Virginia Commonwealth University, Richmond, Virginia, USA. [10]Van Andel Institute, Grand Rapids, Michigan, USA. [11]SAIC-Frederick, Inc., Frederick, Maryland, USA. [12]Sapient Government Services, Arlington, Virginia, USA. [13]The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [14]Massachusetts General Hospital Cancer Center, Boston, Massachusetts, USA. [15]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [16]University of Miami School of Medicine, Miami, Florida, USA. [17]Harvard University, Boston, Massachusetts, USA. [18]Mount Sinai School of Medicine, New York, New York, USA. [19]University of Chicago, Chicago, Illinois, USA. [20]Howard Hughes Medical Institute, Chicago, Illinois, USA. [21]University of Geneva, Geneva, Switzerland. [22]Center for Genomic Regulation, Barcelona, Spain. [23]Stanford University, Palo Alto, California, USA. [24]Oxford University, Oxford, UK. [25]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [26]National Center for Biotechnology Information, National Library of Medicine, US National Institutes of Health, Bethesda, Maryland, USA. [27]Division of Program Coordination, Planning and Strategic Initiatives, Office of Strategic Coordination (Common Fund), Office of the Director, US National Institutes of Health, Bethesda, Maryland, USA. [28]National Human Genome Research Institute, Bethesda, Maryland, USA. [29]National Institute of Mental Health, Bethesda, Maryland, USA. [30]US National Cancer Institute, Bethesda, Maryland, USA.