Published in final edited form as:
Nat Genet. 2013 June ; 45(6): 580–585. doi:10.1038/ng.2653.

The Genotype-Tissue Expression (GTEx) project

The GTEx Consortium*

Abstract

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 are likely to contribute to many disease phenotypes. 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. 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.

URLs. GWAS, http://www.genome.gov/gwastudies; GTEx LDACC Data Portal, http://www.broadinstitute.org/gtex; caHUB, http://cahub.cancer.gov/; and http://biospecimens.cancer.gov/resources/sops/default.asp; dbGaP, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424; NIH Common Fund, http://commonfund.nih.gov/GTEx/ and http://genome.gov/gtex.; NCBI GTEx eQTL Browser, http://www.ncbi.nlm.nih.gov/gtex/test/GTEX2/gtex.cgi/; Request for Information, http://grants.nih.gov/grants/guide/notice-files/NOT-RM-12-028.html; seeQTL (http://www.bios.unc.edu/research/genomic_software/seeQTL/, and SCAN http://biospecimens.cancer.gov/resources/sops/default.asp; Community Resource Policy, http://gwas.nih.gov/03policy2.html. Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility, http://www.wellcome.ac.uk/About-us/Publications/Reports/Biomedical-science/WTD003208.htm; NIH Implementation Group Members, http://commonfund.nih.gov/GTEx/members.aspx.

AUTHOR CONTRIBUTIONS

Biospecimen and Data Collection, Processing, Quality Control, Storage, and Pathological Review
Cancer Human Biobank (caHUB) - Biospecimen Source Sites (BSS): ***National Disease Research Interchange:*** J.L., J.T., M.S., R.P., E.L., S.S; ***Gift of Life Donor Program:*** R.H.; ***LifeNet Health***: G.W.; ***Drexel University College of Medicine:*** F.G.; ***Albert Einstein Medical Center:*** N.Y.; ***Roswell Park Cancer Institute:*** B.F., M.M., E.K., B.G.; K.R.; ***Upstate New York Transplant Service:*** S.S., J.B.; ***Science Care, Inc.:*** H.M., J.S., J.F.; **caHUB - Ethical, Legal, and Social Implications (ELSI) Study:** ***Virginia Commonwealth University:*** L.S., H.T., M.M., L.B.; **caHUB - Comprehensive Biospecimen Resource (CBR):** ***Van Andel Institute:*** S.J., D.R., D.M., D.F., P.H., E.C., B.B., L.T., E.H., K.F.; **caHUB - Pathology Resource Center (PRC):** ***SAICF-Frederick, Inc.:*** L.S, J.R.; ***National Cancer Institute:*** P.B.; **caHUB - Comprehensive Data Resource (CDR):** ***SAICF-Frederick, Inc.:*** G.K., C.S., D.T., L.Q, K.G., S.N.; **caHUB - Operations Management:** ***SAICF-Frederick, Inc.:*** S.B., A.Z., A.S, R.B., K.R., K.V., D.B., M.C., N.D.-M., M.K., T.E., P.W.; ***Sapient Government Services:*** K.E.
Laboratory Analysis, Data Analysis, and Study Coordination: ***The Broad Institute of Harvard and MIT, Inc.:*** K.A., W.W., G.G., D.D., D.M., M.D.,A.T, T.Y., E.G., M.D, Y.M, G.G.
Brain Bank Operations: ***University of Miami School of Medicine:*** D.M., Y.M., M.B.
Statistical Methods Development and Data Analysis: ***Harvard University:*** J.L.; ***Mt Sinai School of Medicine:*** J.Z., Z.T.; ***University of Chicago:*** N.C., D.N., E.G., H.I., A.K., J.P., M.S., T.F., X.W.; ***University of Geneva:*** E.T.D., T.L.; ***Center for Genomic Regulaton:*** R.G., J.M., M.S.; ***Stanford University:*** D.K., A.B., S.M.; ***Oxford University:*** M.M., M.R., J.M.; ***University of North Carolina - Chapel Hill:*** I.R., A.N., F.W., A.S.
Database: ***National Center for Biomedical Information*** **(NCBI):** M.F., N.S., A.S., J.P.
Program Management: ***National Institutes of Health: Officer of the Director:*** J.M.A, E.L.W, L.K.D; ***National Human Genome Research Institute*** 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.; ***National Institute of Mental Health:*** T.R.I., S.E.K., A.R.L, P.K.B, T.L., Y.Y.; ***National Cancer Institute:*** C.C.C., J.B.V., S.S., N.C.L, J.D., H.F.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Note: The views presented in this article do not necessarily reflect those of the US National Institutes of Health.

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 effect of most GWAS-implicated variants remains largely unexplained. The finding that nearly 90% of these sites occur outside of protein coding sequences[2] suggests that many associated variants may instead play a role in gene regulation.

Expression quantitative trait loci (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, non-coding RNAs, 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 interactions[12,13].

Complementing eQTL data with other molecular phenotypes, such as epigenomic assays[14], on the same tissues, and linking to resources such as ENCODE[15], will provide a powerful means of dissecting gene regulatory and higher-order networks across multiple tissues. This will be important because evaluation of the functional consequences of a disease-associated SNP would ideally be assayed in a disease-relevant cell context. However, for most tissue types, human biospecimens are very difficult to obtain from living donors (i.e. brain, heart, pancreas, etc.), and most eQTL studies to date 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) and linked to genome-wide gene expression patterns across a wide range of tissue types. As a first step, 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 can expand to include correlations with epigenetics, proteomics, etc. While 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 technology prompted an 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. Based on the output of this workshop and ongoing consultation, NIH Program staff developed the concept of the Genotype-Tissue Expression (GTEx) project (See 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 associated tissue bank in which to study the relationship between genetic variation and gene expression, and other molecular phenotypes, in multiple reference tissues (Supplementary Figure 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 post-mortem donors by the end of 2015. The power to detect eQTLs is dependent on multiple factors that are difficult to quantify precisely, but values over a range of effect sizes and allele frequencies are described in Figure 1.

GTEx donors are identified through low-post-mortem-interval (PMI) autopsy or organ and tissue transplant settings. To compare the quality of results from autopsy vs. surgically derived tissue, a small subset of tissue types routinely discarded during a surgical amputation, such as skin, fat, and muscle, are also being collected. In addition, peripheral blood samples are collected and used as both a source of DNA for whole-genome single-nucleotide polymorphism (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 derived primarily from massively parallel sequencing of RNA, but some pilot-phase tissues were analyzed both by sequencing and by gene expression array to enable a technology comparison. eQTLs will be identified and made accessible to the scientific community through the National Center for Biotechnology (NCBI) GTEx database and a GTEx data portal. In addition, the GTEx raw data will be made available through the database of Genotypes and Phenotypes (dbGaP) on a periodic basis.

The GTEx structure during the pilot phase is depicted in Supplementary Figure 2, and includes entities for biospecimen acquisition, processing, storage, and verification; a study on ethical, legal, and social issues (ELSI); 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 and information about funding opportunities are available on the NIH Common Fund web site.

Biospecimen acquisition

These functions are designed and organized under the National Cancer Institute’s (NCI) cancer Human Biobank (caHUB). caHUB has enrolled under contract several Biospecimen Source Sites (BSS), a Comprehensive Biospecimen Resource (CBR), a Comprehensive Data Resource (CDR), and Pathology and Quality Management teams to perform acquisition of biospecimens and associated data. All Standard Operating Procedures for donor enrollment and sample collection are available on the caHUB web site.

Donors of any racial and ethnic group and sex who are age 21–70 in whom biospecimen collection can start within 24 hours of death are eligible. 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 past 48 hours, or body mass index ≥ 35 or ≤18.5. Each BSS collects, where feasible, aliquots from many pre-designated tissue sites and organs (Supplementary Table 1), including the brain of deceased donors who were not on a ventilator for the 24 hours prior to death. Immediately after excision, most of the 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 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 accessing stored tissue samples are currently being developed. The CBR's histologic sections are viewed as digitally scanned images, which allow precise annotations to be made to indicate where downstream studies, e.g., tissue microarrays and laser capture microextraction, on selected portions of a specimen can focus (e.g., lymphoid nodules in ileal mucosa or the squamous epithelium of the esophageal mucosa).

The clinical data collected for each GTEx donor belongs to one of two categories: donor-level data or sample level data. Donor level data encompasses all clinical measures of the donor, which includes 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) or next of kin (post-mortem) 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 the PAXgene fixative. Both donor and sample level data are checked for quality and completeness before being released.

Brain Bank

Aliquots from a single region of the cortex and cerebellum are sampled and preserved in PAXgene Tissue at the BSS, while 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 at the Brain Bank, frozen samples from additional anatomical regions of the brain are analyzed at the LDACC and the remaining brain banked for future uses.

Laboratory, Data Analysis, and Coordinating Center (LDACC)

The LDACC performs nucleic acid extractions and quality assessment, DNA genotyping, and RNA expression analysis. The LDACC integrates results of 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 Human Omni5M-Quad BeadChip to collect whole genome SNP and CNV information from each donor's blood sample (or an alternate tissue if blood is unavailable). This assay contains over 4 million probes, with robust coverage of both SNPs and CNVs. DNA is also characterized using the Illumina HumanExome BeadChip to obtain high quality SNP calls within coding regions.

A portion of each tissue is processed for RNA and DNA extraction, quantification, and quality assessment. Extraction protocols that preserve both messenger RNA and microRNA are being used, and are available on the data portal. For measurement of gene expression, 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), during the pilot, to establish comparability of these methods using post-mortem tissue. The 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 moderate- and some low-expressed transcripts, but will have limited ability to accurately quantify rare transcripts and splice isoforms. It should provide gene expression measurements equal to or more accurate than expression arrays and with a higher dynamic range (i.e. coefficient of variation <0.1 for at least 12,000 genes; Supplementary Figure 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 be powered to detect ASE in the top tertile of expressed genes (Supplementary Figure 4). 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. Since 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 Browse

eQTLs are available and queryable 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 and methods 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 RFA form an analysis consortium that will provide innovative approaches to analyses of GTEx data and other similar datasets. Investigators also collaborate with the LDACC to perform data quality assessment/quality control before release into dbGaP. The initial GTEx Consortium publications, anticipated in 2013, will include genome-wide analysis of *cis*- and *trans*-eQTLs, allele-specific expression, splicing quantitative trait loci, and a comparison of array and RNA-Seq based gene expression results.

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 (genomic vs. single gene or small gene networks, proteomic vs. single proteins or small networks of proteins, etc.) 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, while those that take full advantage of the unique aspects of GTEx, such as the multiple tissues from each donor and the gene expression information, are particularly sought. All data resulting from 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 design of the full GTEx project, we built a framework to evaluate the statistical power to detect eQTLs. The statistical power depends on various parameters, some known more accurately than others. These parameters include the number of donors, the eQTL effect size, the noise, as well as the significance threshold, which is based on the number of hypotheses tested. Assuming we are testing *cis*-eQTLs between each of the 20,000 genes and 10 non-redundant *cis*-SNPs (on average) in vicinity (±100 kb) of each gene, the overall number of hypotheses is 200,000. Therefore, using a Bonferroni correction, we set the significance threshold, α, to be 0.05/200,000. For a *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 log-normally distributed with a log standard deviation of 0.13 within each genotype class. 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 (denoted by Δ). Figure 1a shows the statistical power of a *cis*-eQTL analysis, and Figure 1b a *trans*-eQTL analysis, using an ANOVA statistical test as a function the number of subjects and the minor allele frequency (MAF), and assumes Δ=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, since 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%. The statistical power may be higher using methods that leverage the fact that multiple tissues are collected and analyzed for each donor. Since the underlying parameters are merely rough estimates, we repeated the power analysis with different values (10 to 20 SNPs, 20,000–100,000 transcripts) and show that 80% power is achieved for MAF between 3 and 4% for *cis*-eQTLs. For *trans*-eQTLs, this range in transcripts results powered MAFs between 4 and 5% (Supplementary Figure 5).

Data release, 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, following NIH guidelines. It is recognized that quantifying the risk of identifying a donor based on genetic and other information lies on a continuum and is a complex issue dependent on many factors, such as other sources of data and evolving analytical methods[16,17]. Sharing of any information unique to an individual carries a small but difficult to define risk of identifiability, but this must be balanced with the benefits of data sharing to advance science.

Some data from the GTEx project is openly available, meaning that it can be accessed directly through the Internet. However, in order to reduce risks of sharing potentially identifying data, some data elements are available to the scientific community only through a controlled-access system, NCBI’s dbGaP. Standard Operating Procedures (SOPs), data collection instruments, histopathological interpretations, molecular data that does not provide direct genetic variation information (e.g., expression arrays, summary sequence-based gene expression estimates stripped of variant information, eQTL results), laboratory processing variables (e.g., cDNA library preparation methods), and a very limited set of medical and socio-demographic variables (e.g., sex, age at death in intervals) will be openly available. The LDACC will host an open access data portal while specimen acquisition SOPs and associated data collection instruments will be available through caHUB. Other medical and social history information, molecular results that contain direct genetic variation information (e.g., SNP genotyping files, RNA-Seq reads) and summary results that allow accurate inference of allele frequencies[18] will be available only through controlled access. Direct HIPAA identifiers, including dates that include month and day, will not be available through either open or controlled access.

Implementation of these data release policies and processes are 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, then data is 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 comprises genotype data from 190 individuals from which 1814 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 further releasing data two to four times per year until the project is completed. We expect total enrollment to over 400 by 2013, over 700 in 2014, and 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 pre-publication data release relies on the scientific community to respect the data producer’s interest to publish a full analysis of their data first. While 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 in the “Statistical Analysis and Development of Methods” section.

Ethical, legal and social implications (ELSI)

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 deceased donor specimens has direct implications for donor patient families. In recognition of this, sites were required to obtain written or recorded verbal authorization from next-of-kin for deceased donor participation in GTEx, typically through an addendum or modification to an existing authorization form for research donation of tissues and organs. It included statements common in consent forms such as the intention to perform genetic analyses, establish cell lines, and to share data with the scientific community. Work underway is more closely identifying familial concerns and may result in a modification to authorization. Living surgery donors participate only after full, written informed consent is obtained.

In addition, an ELSI study of the consent/authorization process is being carried out at one BSS, to assess both the effectiveness of the consent/authorization process in informing participants of the risks and benefits of the study and its potential psychosocial impact on donors and/or their families. The ELSI study is fully integrated with the 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 to unravel 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge and thank the donors and their families for making organ and tissue donations, both for transplantation and for the GTEx research study.

The authors acknowledge the following funding sources: Contracts X10S170, X10S171, and X10172, SAIC-Frederick, Inc., Frederick, MD, USA; US National Cancer Institute, and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the National Disease Research Interchange, Philadelphia, PA, USA, the Roswell Park Cancer Institute, Buffalo, NY, USA and Science Care, Inc., Phoenix, AZ, USA; Contract HHSN268201000029C, National Heart, Lung, and Blood Institute and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the Broad Institute of Harvard and MIT, Inc. (Wendy Winckler, contact PI), Cambridge, MA, USA; R01 DA006227-17, National Institute of Drug Abuse, National Institute of Mental Health, and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA to the University of Miami School of Medicine (Deborah Mash, PI), Miami, FL, USA; Contract 10ST1035, SAIC-Frederick, Inc., Frederick, MD, USA; National Cancer Institute, and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to Van Andel Institute, Inc., Grand Rapids, MI, USA; Prime contract HHSN261200800001E, National Cancer Institute and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to SAIC-Frederick, Inc., Frederick, MD, USA; R01 MH090941, National Institute of Mental Health and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the University of Geneva (Emmanouil Dermitzakis, contact PI), Geneva, Switzerland; R01 MH090951, National Institute of Mental Health and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the University of Chicago (Jonathan Pritchard, PI), Chicago, IL, USA; R01 MH090937, National Institute of Mental Health, National Human Genome Research Institute, National Heart Lung and Blood Institute and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the University of Chicago (Nancy Cox, contact PI), Chicago, IL, USA; R01 MH090936, National Institute of Mental Health and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to the University of North Carolina at Chapel Hill (Ivan Rusyn, contact PI), Chapel Hill, NC, USA; R01 MH090948, National Institute of Mental Health, National Human Genome Research Institute and NIH Common Fund, National Institutes of Health, Bethesda, MD, USA to Harvard University (Jun Liu, contact PI), Cambridge, MA, USA. This research was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

*The GTEx Consortium Collaborators

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[13], Kristin Ardlie[14], Wendy Winckler[14], Gad Getz[14,15], David DeLuca[14], Daniel MacArthur[14,15], Manolis Kellis[14,16], Alexander Thomson[14], Taylor Young[14], Ellen Gelfand[14], Molly Donovan[14], Yan Meng[14], George Grant[14], Deborah Mash[17], Yvonne Marcus[17], Margaret Basile[17], Jun Liu[18], Jun Zhu[19], Zhidong Tu[19], Nancy Cox[20], Dan Nicolae[20], Eric Gamazon[20], Haky Im[20], Anuar Konkashbaev[20], Jonathan Pritchard[20,21], Matthew Stevens[20], Timothèe Flutre[20], Xiaoquan Wen[20], Emmanouil T. Dermitzakis[22], Tuuli Lappalainen[22], Roderic Guigo[23], Jean Monlong[23], Michael Sammeth[23], Daphne Koller[24], Alexis Battle[24], Sara Mostafavi[24], Mark McCarthy[25], Manual Rivas[25], Julian Maller[25], Ivan Rusyn[26], Andrew Nobel[26], Fred Wright[26], Andrey Shabalin[26], Mike Feolo[27], Nataliya Sharopova[27], Anne Sturcke[27], Justin Paschal[27], James M Anderson[28], Elizabeth L Wilder[28], Leslie K Derr[28], Eric D Green[29], Jeffery P Struewing[29], Gary Temple[29], Simona Volpi[29], Joy T Boyer[29], Elizabeth J Thomson[29], Mark S Guyer[29], Cathy Ng[29], Assya Abdallah[29], Deborah Colantuoni[29], Thomas R Insel[30], Susan E Koester[30], A Roger Little[30], Patrick K Bender[30], Thomas Lehner[30], Yin Yao[30], Carolyn C Compton[12], Jimmie B Vaught[12], Sherilyn Sawyer[12], Nicole C Lockhart[12], Joanne Demchok[12], Helen M Moore[12].

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

References

1. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–888. [PubMed: 18988837]

2. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]

3. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008; 24:408–415. [PubMed: 18597885]

4. Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–428. [PubMed: 18344981]

5. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]

6. Stranger BE, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. [PubMed: 17873874]

7. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

8. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 2010; 6:e1001236. [PubMed: 21151575]

9. Borel C, et al. Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. Genome Res. 2011; 21:68–73. [PubMed: 21147911]

10. Petretto E, et al. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. Plos Computational Biology. 2010; 6

11. Grundberg E, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature Genetics. 2012

12. Zhong H, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet. 2010; 6:e1000932. [PubMed: 20463879]

13. Zhao E, et al. Obesity and genetics regulate microRNAs in islets, liver, and adipose of diabetic mice. Mamm Genome. 2009; 20:476–485. [PubMed: 19727952]

14. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology. 2010; 28:1045–1048.

15. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

16. Craig DW, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet. 2011; 12:730–736. [PubMed: 21921928]

17. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nature Genetics. 2012; 44:603–608. [PubMed: 22484626]

18. Jacobs KB, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet. 2009; 41:1253–1257. [PubMed: 19801980]

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 tissue, nucleic acid, 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 expression Quantitative Trait Loci (eQTL) alone and in the context of other molecular phenotypes.
- To make the 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.

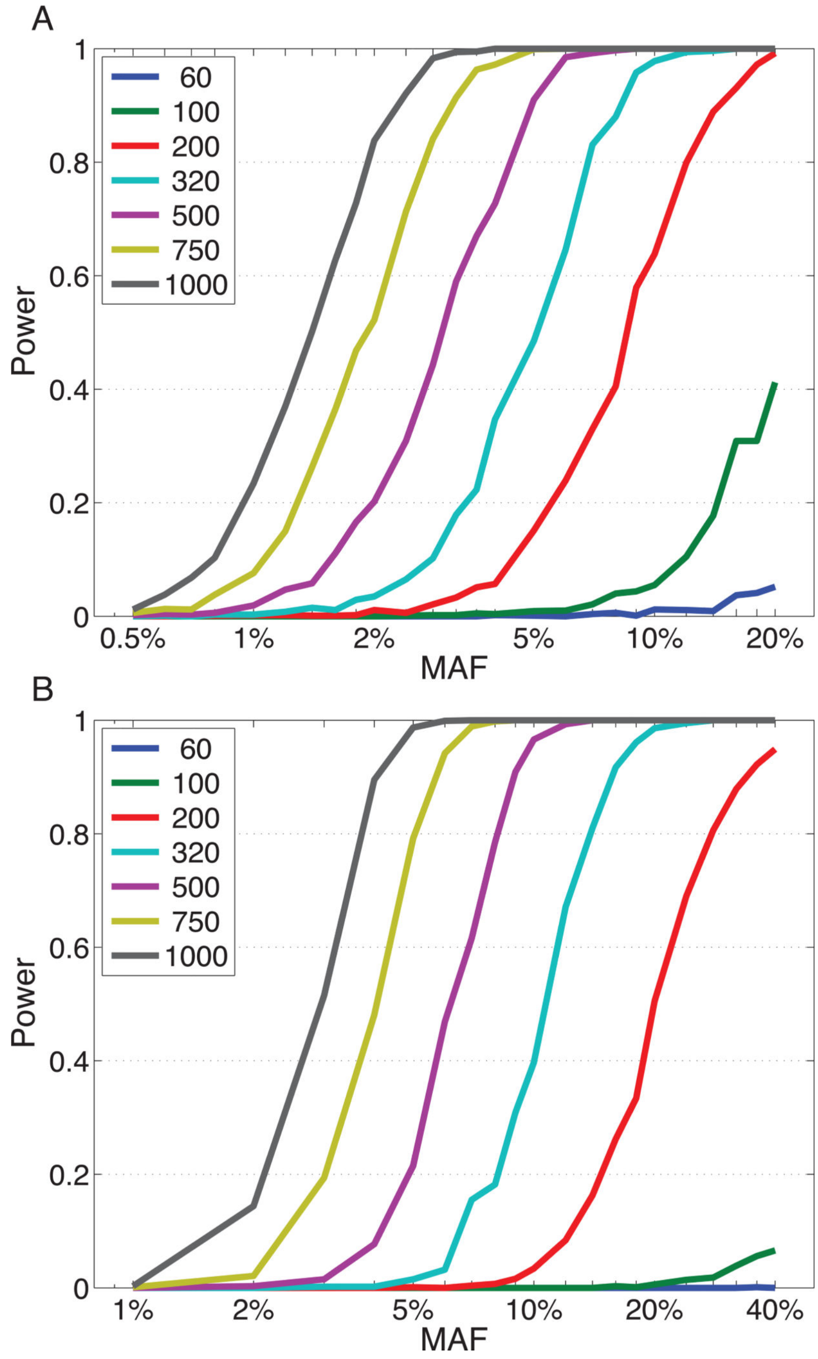

Figure 1.

Effect of sample size and MAF on power to detect eQTLs. (a) Power for *cis*-eQTL analysis in which we assume α = 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 $\alpha = 5 \times 10^{-13}$.