Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease

eGTEx Project*

Genetic variants have been associated with myriad molecular phenotypes that provide new insight into the range of mechanisms underlying genetic traits and diseases. Identifying any particular genetic variant's cascade of effects, from molecule to individual, requires assaying multiple layers of molecular complexity. We introduce the Enhancing GTEx (eGTEx) project that extends the GTEx project to combine gene expression with additional intermediate molecular measurements on the same tissues to provide a resource for studying how genetic differences cascade through molecular phenotypes to impact human health.

Introduction

Identifying the molecular and cellular basis of human genetic disease provides new opportunities for disease prevention and treatment. Genome-wide association studies (GWAS) have already yielded thousands of genetic associations, localizing regions of the genome that confer disease risk. However, within disease-associated regions, the causal variants and the mechanism of action often remain poorly understood. To address this challenge, the Genotype-Tissue Expression (GTEx) project has generated a systematic, multitissue reference for identifying genetic variants associated with changes in gene expression (expression quantitative trait loci, eQTLs). This resource supports research into potential mechanisms of action for disease-associated variants[1,2]. Beyond gene expression, a rapidly increasing array of molecular and sequencing-based assays is identifying genetic variants associated with many intermediate molecular phenotypes. Integration of multiple layers of molecular measurements will clarify the causes and consequences of changes in gene expression as well as identify new mechanisms underlying human disease.

Among the multiple molecular phenotypes recently used in QTL-based analyses are measurements of histone modification, chromatin accessibility, allele-specific expression (ASE), alternative splicing, DNA methylation, and protein expression. QTL-based analyses of histone modifications and chromatin accessibility provide insight into variants that influence transcription factor binding and nucleosome positioning[3,4]. Allele-specific expression QTLs (ase-QTLs) have used allelic ratios as a quantitative phenotype to increase the power of eQTL analyses[5–7] and can aid the localization of causal variants[8]. Alternative splicing QTLs (sQTLs) have been a major focus of multiple eQTL studies using RNA-seq and have been implicated as important contributors to human disease[9]. Methylation QTLs (meQTLs) have uncovered complex relationships between genetic, DNA methylation, and expression variation[10–12]. Ongoing and rapid advances in high-throughput protein quantification have enabled the detection of protein QTLs (pQTLs), which identify variants influencing both transcriptional and post-transcriptional mechanisms[13,14].

Beyond studies of QTL types in isolation, integrative or multi-omics analyses offer to elucidate the cascade of molecular effects of disease variants. For example, intersection of DNase I sensitivity QTLs (dsQTLs) and eQTLs established that over half of eQTLs are also associated with changes in chromatin accessibility[3]. Intersection of meQTLs and eQTLs identified variants with complex causal relationships depending on CpG and genic contexts[15,16]. Intersection of eQTLs and pQTLs exposed buffered effects, protein-specific effects, and overlap with disease-associated variants[13,14,17,18]. Recent integration of eight cellular phenotypes across the regulatory cascade from transcription factor binding to protein expression demonstrated essential contributions of splicing to disease-associated variation[9]. Such increased accessibility to multi-omics data offers new opportunities to develop and test holistic or 'systems genomics' approaches. These approaches offer to provide new opportunities for predictive modeling and enrich understanding of the multitude of effects for disease-associated variants and their interplay across diverse omics layers[19].

A major challenge to integration of multi-omics data in the study of human disease is that many multi-omics analyses have been conducted in cell lines instead of primary tissues. While GTEx has demonstrated the value of data from multiple tissues in identifying

*A list of members and affiliations appears at the end of the paper. Correspondence should be addressed to S.B.M. (smontgom@stanford.edu), K.G.A. (kardlie@broadinstitute.org), M.K. (manoli@mit.edu), M.P.S. (mpsnyder@stanford.edu) or B.E.S. (bstranger@medicine.bsd.uchicago.edu).

Published online 11 October 2017; doi:10.1038/ng.3969

tissue-specific mechanisms for disease-associated variants, there remains a need to obtain multi-omics reference data to study the effects of genetic variation across multiple tissues and multiple layers of molecular complexity. In addition to complementing studies of complex genetic diseases, expanding multitissue molecular data from 'normal' individuals can enhance cancer studies[20,21] (which currently comprise 28% of all requests for GTEx data use), by distinguishing cancer-specific alterations and elucidating the tissue specificity of certain cancers and their mutations[22,23].

Here we introduce the US National Institutes of Health (NIH) Common Fund's Enhancing GTEx (eGTEx) project, which seeks to complement the gene expression phenotypes determined in the GTEx project with intermediate phenotypes across the same tissues and individuals (**Fig. 1**). These additional data types will provide a more complete reference of how genetic differences cascade through molecular and cellular phenotypes to impact organismal phenotypes. To achieve this goal, eGTEx is applying diverse molecular assays to the GTEx sample collection, including DNase I hypersensitivity, ChIP–seq, DNA and RNA methylation, ASE, protein expression, somatic mutation, and telomere length assays. Together, the eGTEx reference aims to enable high-resolution identification of the mechanistic impacts of genetic variants and their role in human diseases, and it will serve as an enabling resource that will facilitate novel integrative and holistic computational methods development and biological insights.

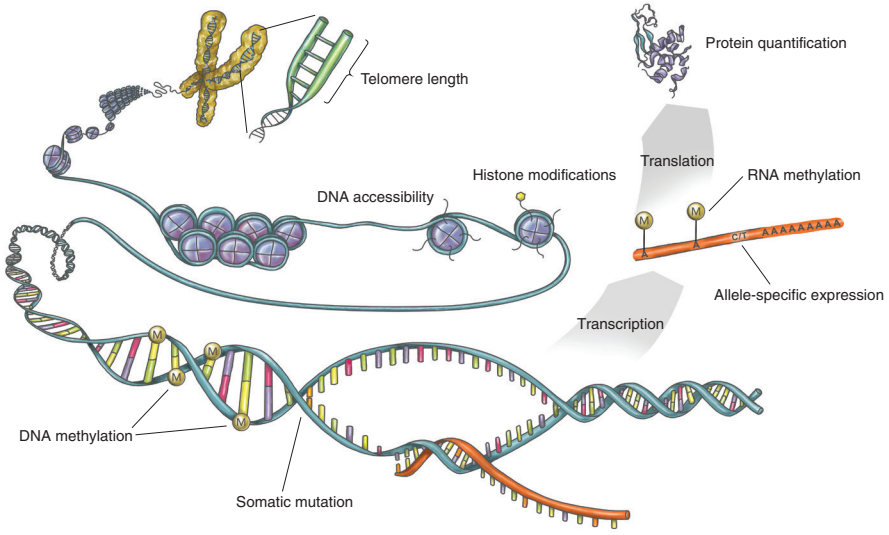

Figure 1 Quantifying layers of molecular and cellular phenotypes. The eGTEx project plans to study telomere length, DNA accessibility, histone modifications, DNA and RNA methylation, somatic mutation, allele-specific expression, and protein quantification across individuals and tissues.

The eGTEx project: study design and assays

The goal of the GTEx project is to establish a national multitissue cohort for molecular phenotypes. The current release of GTEx (database of Genotypes and Phenotypes (dbGaP) accession phs000424.v7.p2) provides 11,688 transcriptomes from 714 individuals and 53 tissues (median of 17 tissues per individual, 173 samples per tissue). The next release, v8, is expected to include 17,500 transcriptomes from ~850 individuals, and final data production for the project is targeted for late 2017. In addition to molecular data, GTEx includes pathology reports, histology images and reports, and donor characteristics, including ethnicity, age, and sex. Within GTEx, tissues are obtained from deceased donors with next-of-kin consent to the collection and banking of anonymized samples for scientific research[24]. Two existing strengths of the GTEx project are the large number of tissues collected from each donor, facilitating characterization of gene expression across a wide variety of tissues, and the relatively large size of the donor population, allowing one to evaluate the contribution of individual genetic variation. The first steps of assaying genetic variation and its impact on gene expression are the focus of two accompanying consortium papers[25,26]. However, fully understanding how a genetic variant regulates gene expression, such as through changes in DNA methylation or the binding affinity of a transcription factor, and subsequently connecting the downstream effects of differential gene expression through to protein abundance require additional molecular assays.

The goal of the eGTEx initiative is to enhance understanding of gene regulation by performing additional molecular analyses on the same tissues that underwent gene expression analysis. Because of the large size of the GTEx tissue collection (over 25,000 samples), the variable quality across the samples collected, and the relatively small aliquot remaining for each sample, the eGTEx initiative will analyze a subset of the entire collection. The study design for eGTEx activities was allocated across two 'dimensions' of analysis: phase I, involving a relatively small number of donors (~15) analyzed for a large number of different tissues (>20); and phase II, involving a relatively small number of tissues (4–6) analyzed in a larger number of donors (150–200). eGTEx has planned to use the same tissues from the same individuals for as many assays as possible. However, because available aliquots are limited, some assays require frozen tissue as input, and the throughput differs by assay, the extent of overlap and the number of phenotypes that will be generated from each individual sample will vary. The molecular phenotypes being studied are shown in **Table 1** and described in the following sections.

DNA accessibility

Systematic understanding of the impact of genetic variation on gene expression requires both comprehensive delineation of regulatory DNA and an understanding of the degree to which individual regulatory regions vary at the population level. DNA is tightly packaged into chromatin inside our cells, with 147-nucleotide segments of DNA wrapped around each histone octomer (themselves separated by ~50-nucleotide linkers). Displacement of nucleosomes through the binding of transcriptional regulators results in accessible regions of 'open chromatin', which can be mapped using endonucleases such as DNase I (refs. 27,28). Past work has shown that disease and trait associations are highly concentrated in accessible elements[29] and that allelic variation in DNA accessibility can precisely map the effects of sequence variation on transcription factor activity[3,30,31]. In eGTEx, we will examine DNA accessibility using both the DNase I hypersensitivity assay and the higher-resolution DNase I footprinting assay to map transcription factor occupancy within regulatory DNA at nucleotide-level resolution. The footprints revealed by DNA accessibility are highly unbiased and capture variation in diverse regulatory elements, including promoters, enhancers, silencers, insulators, and locus-control regions.

Histone modifications

Each histone protein in the chromatin fiber has a long amino acid tail that can be

Table 1 eGTEx study design

Molecular phenotype	Primary assay(s)	Targeted tissues (phase II)	Targeted sample number
DNA accessibility	DNase I hypersensitivity	Brain regions, heart, lung, muscle, esophagus, breast, prostate, skin	~1,135
Histone modifications	ChIP–seq	Brain regions, heart, lung, muscle	~600
DNA methylation	WGBS and capture bisulfite sequencing	Brain regions, heart, lung, muscle, thyroid	~2,000
Allele-specific expression	mmPCR–seq	All tissues	~2,000
Post-transcriptional RNA modifications	m^6A methylation capture sequencing	Brain regions, heart, lung, muscle	~300
Proteomic variation	MS, targeted arrays for transcription factors and cell signaling proteins	Brain, heart, lung, muscle, thyroid, colon, liver, prostate, pancreas, ovary, testis, breast	~1,000 (MS) ~2,500 (arrays)
Somatic variation	Deep exome sequencing, RNA-seq, SNP arrays	~20–25 tissues	~800
Telomere length	Luminex-based assay for telomere-repeat abundance	~20 tissues	~5,000

Molecular assays, targeted tissues, and sample number for eGTEx.

post-transcriptionally modified, thus serving both structural and informational roles[32]. An ever-growing multitude of histone modifications have been described, whose combinations mark diverse chromatin functions, including active enhancers, active promoters, poised promoters, repressed regions, heterochromatin, and transcribed functions[33]. Among the diverse chromatin states that have been broadly surveyed, enhancer regions marked by histone H3 lysine 27 acetylation (H3K27ac) were shown to be the most variable across tissues and cell types[33,34], the most variable across individuals[35], and the most highly enriched for disease-associated genetic variants[34]. Thus, to characterize the protein–DNA interactions through which gene regulation is mediated, eGTEx will perform ChIP–seq to profile the enhancer- and promoter-associated histone modification H3K27ac.

Because DNA accessibility and ChIP–seq assays require frozen material, they cannot be performed on as wide a spectrum of GTEx tissues as most other assays and instead will focus on the brain and tissues that were collected frozen for a subset of GTEx donors. Availability of DNase I hypersensitivity and H3K27ac data will aid in fine-mapping causal variants and improving localization and interpretation of tissue-specific and shared regulatory variants.

DNA methylation

DNA methylation of cytosine residues throughout the human genome is an important element of gene regulation and a key component of eGTEx. DNA methylation influences the binding of regulatory elements (such as transcription factors) to DNA[36] and is involved in imprinting[37,38], X-chromosome inactivation[39], and gene silencing[40]. eGTEx will use two complementary methods (whole-genome bisulfite sequencing (WGBS) and capture bisulfite sequencing) to characterize the DNA methylation landscapes of GTEx tissues, with a particular focus on distinct brain regions implicated in mental health. In addition, as disease-driven changes in cell type composition (specifically, neuron loss in many mental health disorders) have increasingly been appreciated, eGTEx will aim to account for this source of variation. To address this issue, eGTEx will perform fluorescence-activated nuclei sorting (FANS) to specifically isolate neuronal nuclei from GTEx brain tissues for WGBS analysis. These analyses will help identify variably methylated regions (VMRs), meQTLs, and regions of allele-specific methylation (ASM).

Allele-specific expression

Within individuals, ASE can validate *cis*-eQTLs or identify and characterize rare and private *cis*-regulatory effects[41]. However, the number of reads mapping to the coding heterozygous sites within a gene of interest limits the power to detect ASE. When using RNA-seq data, this is directly related to the gene's expression level, which varies from tissue to tissue. eGTEx will apply microfluidic multiplex PCR followed by deep sequencing (mmPCR–seq) to bypass this problem and provide high-depth ASE measurements[42]. This high-throughput, targeted approach decouples the power to detect ASE from the gene's expression level. Therefore, mmPCR–seq can provide ASE data from a wider range of tissues than RNA-seq. The assay works effectively with low-quality RNA or with low quantities of input, which makes it perfectly suited to process the limited quantities of RNA available from GTEx samples. Previously, mmPCR–seq has been applied to study the impact of imprinting and loss-of-function variants across multiple tissues[43–45]. As part of eGTEx, mmPCR–seq data will be generated for a few hundred genes across all available tissue samples from ~80 individuals. Data generated will validate and complement eQTLs identified by the GTEx project through joint analysis of ASE and total expression data[5–7]. In addition, these data will aid assessment of tissue specificity and detection of changes in the magnitude of effects across tissues[46].

Post-transcriptional RNA modifications

m^6A methylation recently emerged as an important post-transcriptional modification of RNA, affecting more than 7,000 protein-coding and noncoding genes and influencing protein translation, transcriptional gene regulation, RNA stability, alternative splicing, microRNA targeting, circadian rhythms, and overall gene function[47,48]. m^6A methylation is present in a wide variety of tissues and varies in abundance across tissues and across development. However, much remains unknown, including the variation in m^6A methylation across individuals in different tissues, the role that genetic variants have in guiding methylation changes, and the role of m^6A meQTLs in human disease association. Thus, a systematic survey of the inter-individual and inter-tissue variation in m^6A methylation within the GTEx cohort can have important implications for the study of human disease, by identifying tissue-specific m^6A methylation patterns, inter-individual differences in m^6A methylation, and individual genetic variants that act as m^6A meQTLs, all in the context of a deeply profiled cohort that benefits from genomic, transcriptomic, epigenomic, and proteomic measurements. To take advantage of this opportunity, eGTEx will carry out m^6A–seq experiments along two dimensions, the first exploring tissue diversity for a small set of individuals and the second exploring inter-individual variation for a small set of tissues. Along the tissue dimension, we will profile 20 tissues in 8 individuals, across a total of ~100 samples, as not all individuals have samples available or of sufficient quality for all tissues. Along the individual dimension, we will profile 4 tissues across 100 individuals, for a total of ~300 samples, once more on the basis of sample quality and availability in

sufficient quantity. We will use a combination of two m^{6}A-profiling technologies: 'location analysis', by methylated RNA immunoprecipitation followed by sequencing (MeRIP–seq), which uses an additional RNA fractionation step before m^{6}A profiling, thus enabling the locations in the transcript where m^{6}A methylation occurs to be pinpointed, and 'level analysis', by m^{6}A- level and isoform characterization sequencing (m6A-LAIC-seq), which does not include a fractionation step and provides a more quantitative readout of overall m^{6}A methylation levels for each transcript. We will seek to profile these data matrices sufficiently densely to enable imputation across technology platforms, across tissues, and across individuals in the context of the larger GTEx and eGTEx data matrices.

Protein abundance

Proteins provide a critical link that allows the gap between RNA expression and phenotype to be filled. On one hand, protein levels can be considered as biomarkers for phenotypes; at the same time, they are heritable molecular phenotypes whose genetic basis can be linked to genotype or RNA expression. Preliminary studies have indicated that there is an imperfect correspondence between eQTLs and pQTLs: in particular, not all pQTLs coincide with eQTLs[13,14,17,18]. Measurement of protein expression in the GTEx cohort will allow validation of transcriptional regulation while simultaneously characterizing novel post-transcriptional regulation, as well as the relationship between transcriptional and post-transcriptional regulatory mechanisms. eGTEx will characterize protein expression using two complementary strategies: isobaric tandem mass tag (TMT)-based quantitative mass spectrometry (MS)[13] and high-throughput, robust microwestern technology[49] applying a custom panel set of antibodies targeting ~400 transcription factors and cell signaling proteins. Preliminary studies have shown that by using a customized sample preparation protocol that maximizes the protein yield of PAXgene preserved tissue samples and running each tissue sample in duplicate we can identify ~7,500 proteins per sample using MS. The microwestern assay has been optimized to use small input quantities and will allow quantification across the full spectrum of abundance levels (albeit on a subset of all proteins). Importantly, the microwestern assay allows for quantification of low-abundance proteins that are not typically captured using MS, thus rendering the two approaches highly complementary. The data generated will support detection of pQTLs from approximately 200 individuals per tissue, will enhance studies of tissue specificity and network relationships, and will facilitate multi-omics integrative analyses.

Somatic mutation

It is generally assumed that the trillions of cells in a human body have identical DNA sequences, but in reality we are a mosaic of genomes. In addition to expressing many known epigenetic differences, we are made up of subpopulations of cells with genetic differences, such as point mutations, structural changes, and differences in telomere length. The level of this mosaicism is largely unknown, but recent studies suggest that the extent of somatic variability in humans is considerable and contributes to disease phenotypes[50]. However, there have been few systematic and comprehensive studies of human somatic variability, particularly outside of readily obtainable tissues like blood. This gap in knowledge is a significant impediment to many ongoing and future studies of human phenotypic variation and disease susceptibility, such as the interpretation of somatic variability in cancer genome sequencing projects. The eGTEx project will systematically study somatic variability by performing several new assays in addition to leveraging the existing RNA sequencing data. Somatic point mutations will be evaluated by deep exome sequencing using NimbleGen's SeqCap EZ Human Exome Library v2.0 and sequencing to an average depth of 150× on an Illumina HiSeq 2000. Somatic structural variability will be analyzed using the deep exome sequencing data and SNP arrays for copy number variation (CNV) and copy-neutral loss of heterozygosity (LOH) events.

Telomere length

Telomere length has a central role in maintaining cellular proliferative potential and genome stability, and telomere shortening over the life course may be a critical mechanism underlying many age-related health conditions, including cardiovascular disease[51]. In contrast, longer telomeres may increase risk for some types of cancer[52]. Most epidemiological studies of the association between telomere length and disease risk are difficult to interpret, in part because it is not clear how well telomere length in peripheral blood cells reflects telomere length in disease-relevant tissues. We are addressing this gap in eGTEx by measuring average telomere length (i.e., telomere-repeat abundance) across many tissue types using a high-throughput, Luminex-based method[53–55]. We will characterize the relationships among telomere length measures taken from different tissues and determine whether tissue-specific telomere length is associated with the frequency of somatic events in the same tissue (i.e., CNV and LOH events, detected using exome sequencing and SNP array data). In addition, we will determine whether SNPs known to affect leukocyte telomere length (based on prior GWAS) also influence telomere length in other tissues. Knowledge of tissue-specific effects of SNPs on telomere length will enable Mendelian randomization studies to estimate the effects of telomere length on disease in a tissue-specific fashion. The results from this eGTEx project will provide a foundation for interpreting epidemiological findings and guide the design of future studies of the effect of telomere length on human health.

Integrative multi-omics analyses

Integrative analyses that combine both GTEx and eGTEx data will complement the collection of molecular phenotype data (**Box 1**). These efforts will aim to (i) determine the extent to which different molecular phenotypes vary across tissues and elucidate which factors mediate the levels of each phenotype. For example, it may be found that the variation in mRNA abundance for a particular gene largely occurs between individuals while that in protein abundance occurs between tissues, suggesting the existence of tissue-specific regulatory steps that occur at post-transcriptional levels. Integrative analyses will also aim to (ii) establish the similarity of covariation and coexpression networks across tissues and (iii) identify which loss-of-function mutations are expressed at the RNA and protein levels and to what degree this expression varies across tissues. Recent evidence suggests that a subset of loss-of-function effects are compensated for at the protein level, highlighting the utility of multi-omics data in personal genome interpretation[56].

Studies using eGTEx data will integrate and enhance diverse external projects and resources. As an example, characterization of tissue-specific methylation and expression patterns will benefit from the expanding catalog of chromatin modifications (H3K4me1, H3K27ac, H3K9me3, etc.), DNase I hypersensitivity sites, and binding sites for transcription factors and other regulatory DNA-binding factors from the Encyclopedia of DNA Elements (ENCODE) and NIH Roadmap Epigenomics resources. Roadmap Epigenomics in particular has an extensive data set including many of the same brain regions being profiled by eGTEx and will allow identification of brain-specific regulatory regions. To better enable these types of integrated studies, eGTEx and ENCODE have developed the ENTEX collaborative project to focus on deeply profiling a small subset of directly overlapping tissues using shared technologies.

Box 1 Examples of integrative analyses across tissues using eGTEx data

- Determine the relative variability of molecular phenotypes.
- Compare covariation networks across molecular phenotypes.
- Determine whether genes/proteins with loss-of-function variants are expressed. For example, examine regulatory features, mRNA levels, and protein levels.
- Map QTLs for each molecular phenotype to determine where most functional genetic variation resides.
- Construct integrative regulatory networks using systems genomics approaches.
- Connect regions of allele-specific chromatin accessibility, allele-specific methylation, and allele-specific gene expression.
- Perform integrated analysis of patterns of X-chromosome inactivation.
- Quantify tissue-specific levels of somatic mutations and their relationship to heterogeneity in gene expression levels.
- Associate levels of methylation and expression at telomere maintenance genes (for example, *TERC*, *TERT*, *DKC1*) with telomere length measurements.
- Examine multi-omics enrichments of trait-associated variation.
- Support holistic predictive modeling across molecular phenotypes.

Integrative eGTEx analyses will continue to focus on methodologies that enhance GWAS interpretation. Such methods will take advantage of recent efforts to impute molecular phenotypes[57], test colocalization of trait and molecular association signals[58,59], and jointly model multi-omics data[60–62]. The specific focus on brain tissues in GTEx and eGTEx is particularly useful for contextualizing GWAS of neurological and neuropsychiatric traits. Combinatorial analyses of GWAS associations and diverse molecular factors can reveal important disease-associated SNPs that may have fallen below standard association thresholds or help identify the likely driver SNP from a group of associated and tightly linked variants. Increasingly, as whole-genome-sequenced cohorts become available, integration with eGTEx is expected to enhance new methods for predicting and identifying the tissue context of diverse classes of genetic variation.

Data release and community impact

The GTEx and eGTEx projects are community resources committed to rapid and complete data release. Thus far, wet lab analysis of germline genetic variation and RNA-seq-based mRNA expression measurements for the main GTEx project have been performed at the Broad Institute, serving as the Laboratory, Data Analysis, and Coordinating Center (LDACC) for the project. The entire consortium has contributed to all aspects of the analysis pipeline, but the generation of the vast majority of the data at a single laboratory has facilitated the integration and coordinated release of these data. eGTEx experiments, on the other hand, will be generating a heterogeneous collection of data in more than seven laboratories, making data integration more challenging. Some of the data generated (for example, protein expression) will have little to no privacy concerns and can be made available without access restrictions. Other molecular phenotypes, such as DNA methylation, will be similar to RNA-seq expression, encompassing both raw sequence data that will be deposited into the controlled-access database dbGaP and processed data and results that can mostly be made available without access controls through the GTEx portal (see URLs). The LDACC, which is continuing to release the primary GTEx data, will also serve as the coordination center for release of the controlled-access data for the eGTEx assays. This will ensure that data IDs and metadata are harmonized across the entire project and will enable eGTEx data to be included with the major GTEx releases. Additionally, to the extent possible, open-access eGTEx data will be integrated and made available through the GTEx portal alongside gene expression and eQTL results.

Given the heterogeneous nature of the data, the differing capacities of the participating laboratories, and differences in how samples will need to be batched across the various assays for quality control purposes, we expect components of eGTEx data to be released intermittently over the next several years, with this publication serving as a guide to the overall eGTEx effort. The phase I data, representing analysis of a wide range of tissues from a relatively small number of donors, will be generated first by most groups and thus should be released before phase II data, representing a smaller number of tissues analyzed over a larger number of individuals. We plan initial data deposition in late 2017 to early 2018, with no publication embargo after data release.

Impact of eGTEx on the research community and future directions

The tissues from GTEx donors are collected without focus on a disease state and are representative of a US-based human population[24]. With detailed molecular data being collected across diverse tissues by eGTEx, the resource provides a snapshot of normal (or non-diseased-state) genetic and genomic variation among individuals and across tissues. As such, the resource serves as a reference for disease-focused research, where investigators can compare eGTEx data with genomic data obtained from disease cohorts. Much in the way that eQTL studies are being used to generate hypotheses regarding causal genes and mechanisms underlying GWAS trait associations, these novel eGTEx genomic data are expected to be widely used to elucidate additional mechanisms contributing to diverse human disease.

The diversity of genetic and genomic data types being surveyed by eGTEx will facilitate statistical methods development in the area of data integration. Although analysis methods for specific pairs of data types (for example, genetic and transcriptomic, or eQTLs) have become relatively standardized by the community, methods development for holistic analysis of diverse data types remains an active research area. The eGTEx population-based genomics data characterizing multiple modalities of genome function from genetic to epigenetic to transcriptomic and proteomic variation will provide a rich primary-tissue-based resource for development of 'best practices' for integrative data analysis.

As new computational and experimental approaches continue to elucidate the function of the genome, eGTEx aims to provide a rich data source that enables the integration of multi-omics data in the interpretation of health and disease. These efforts will help pave the route toward an increased understanding of genome function, the elucidation of novel molecular therapeutics, and the integration of high-throughput molecular diagnostics in individualized patient care.

URLs. NIH Common Fund GTEx Project, http://commonfund.nih.gov/GTEx; GTEx Portal, http://www.gtexportal.org/.

ACKNOWLEDGMENTS

The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the US National Institutes of Health (NIH; see URLs). Additional funds were provided by the National Cancer Institute (NCI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Drug Abuse (NIDA), National Institute of Mental Health (NIMH), and National Institute of Neurological Disorders and Stroke (NINDS). Donors

were enrolled at Biospecimen Source Sites funded by Leidos Biomedical. Leidos subcontracts to the National Disease Research Interchange (10XS170) and the Roswell Park Cancer Institute (10XS171). The LDACC was funded through a contract (HHSN268201000029C) to the Broad Institute. Biorepository operations were funded through a Leidos subcontract to the Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami grant DA006227. E.K.T. is supported by a Hewlett-Packard Stanford Graduate Fellowship and a doctoral scholarship from the Natural Science and Engineering Council of Canada. NIH grant U01MH104393 supported A.P.F., K.D.H., L.F.R., and P.F.H. NIH grant U01HG007598 supported B.E.S. NIH grant U01HG007599 supported J.A.S. NIH grant U01HG007593 supported J.B.L. and S.B.M. NIH grant U01HG007591 supported J.M.A. NIH grant U01HG007610 supported M.K. NIH grant U01HG007601 supported B.L.P. NIH grant U01HL131042 supported M.P.S. and H.T.

AUTHOR CONTRIBUTIONS

All authors contributed to study design. L.E.B., R.H., M.H., C.J., M.J., G.K., W.F.L., J.T.L., A.M., B. Mestichelli, K.M., B.R., M.S., S.S., J.A.T., G.W., M. Washington, J.W., J.B., B.A.F., B.M.G., E.K., R. Kumar, M.M., M.T. Moser, S.D.J., R.G.M., D.C.R., D.R.V., D.A.D., and D.C.M. were part of the biospecimen collection group. S.E.G., P.G., S.K., A.R.L., C.M., H.M.M., A.R., J.P.S., and S.V. were NIH program management. K.D.H., P.F.H., L.F.R., L.H., Y.L., B. Molinie, Y.P., N.R., L.W., N.V.W., M.C., E.T.G., Q.L., S. Linder, R.Z., K.S.S., E.K.T., L.S.C., K.D., J.A.D., F.J., M.G.K., L.J., S. Lin, M. Wang, R.J., X.L., J.C., D.B., M.D., J.H., E.H., A.J., R. Kaul, K.L., M.T. Maurano, J.N., F.J.N., R.S., M.S.F., C.L., M.O., A.S., F.W., J.M.A., A.P.F., J.B.L., B.L.P., J.A.S., H.T., K.G.A., M.K., M.P.S., S.B.M., and B.E.S. were part of the eGTEx project working group. The writing group included E.K.T., J.M.A., M.T. Maurano, H.T., M.S., S.V., R. Kaul, J.A.S., L.F.R., B.L.P., H.M.M., K.G.A., M.K., S.B.M., and B.E.S. and was led by K.G.A., M.K., M.P.S., S.B.M., and B.E.S.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Nicolae, D.L. *et al. PLoS Genet.* **6**, e1000888 (2010).
2. GTEx Consortium. *Science* **348**, 648–660 (2015).
3. Degner, J.F. *et al. Nature* **482**, 390–394 (2012).
4. McVicker, G. *et al. Science* **342**, 747–749 (2013).
5. Sun, W. *Biometrics* **68**, 1–11 (2012).
6. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. *Nat. Methods* **12**, 1061–1063 (2015).
7. Kumasaka, N., Knights, A.J. & Gaffney, D.J. *Nat. Genet.* **48**, 206–213 (2016).
8. Lappalainen, T. *et al. Nature* **501**, 506–511 (2013).
9. Li, Y.I. *et al. Science* **352**, 600–604 (2016).
10. Gibbs, J.R. *et al. PLoS Genet.* **6**, e1000952 (2010).
11. Bell, J.T. *et al. Genome Biol.* **12**, R10 (2011).
12. Gutierrez-Arcelus, M. *et al. eLife* **2**, e00523 (2013).
13. Wu, L. *et al. Nature* **499**, 79–82 (2013).
14. Hause, R.J. *et al. Am. J. Hum. Genet.* **95**, 194–208 (2014).
15. Banovich, N.E. *et al. PLoS Genet.* **10**, e1004663 (2014).
16. Gutierrez-Arcelus, M. *et al. PLoS Genet.* **11**, e1004958 (2015).
17. Battle, A. *et al. Science* **347**, 664–667 (2015).
18. Cenik, C. *et al. Genome Res.* **25**, 1610–1621 (2015).
19. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. & Kim, D. *Nat. Rev. Genet.* **16**, 85–97 (2015).
20. Vucic, E.A. *et al. Genome Res.* **22**, 188–195 (2012).
21. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. *Cell* **160**, 48–61 (2015).
22. Fernandez-Banet, J. *et al. Nat. Methods* **13**, 9–10 (2016).
23. Kosti, I., Jain, N., Aran, D., Butte, A.J. & Sirota, M. *Sci. Rep.* **6**, 24799 (2016).
24. Carithers, L.J. *et al. Biopreserv. Biobank.* **13**, 311–319 (2015).
25. GTEx Consortium. *Nature* http://dx.doi.org/10.1038/nature24277 (2017).
26. Li, X. *et al. Nature* http://dx.doi.org/10.1038/nature24267 (2017).
27. Weintraub, H. & Groudine, M. *Science* **193**, 848–856 (1976).
28. Wu, C., Wong, Y.C. & Elgin, S.C. *Cell* **16**, 807–814 (1979).
29. Maurano, M.T. *et al. Science* **337**, 1190–1195 (2012).
30. Maurano, M.T. *et al. Nat. Genet.* **47**, 1393–1401 (2015).
31. Neph, S. *et al. Nature* **489**, 83–90 (2012).
32. Bannister, A.J. & Kouzarides, T. *Cell Res.* **21**, 381–395 (2011).
33. Ernst, J. *et al. Nature* **473**, 43–49 (2011).
34. Roadmap Epigenomics Consortium. *Nature* **518**, 317–330 (2015).
35. Kasowski, M. *et al. Science* **342**, 750–752 (2013).
36. Maurano, M.T. *et al. Cell Rep.* **12**, 1184–1195 (2015).
37. Pervjakova, N. *et al. Epigenomics* **8**, 789–799 (2016).
38. Li, E., Beard, C. & Jaenisch, R. *Nature* **366**, 362–365 (1993).
39. Payer, B. & Lee, J.T. *Annu. Rev. Genet.* **42**, 733–772 (2008).
40. Curradi, M., Izzo, A., Badaracco, G. & Landsberger, N. *Mol. Cell. Biol.* **22**, 3157–3173 (2002).
41. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. *Genome Biol.* **16**, 195 (2015).
42. Zhang, R. *et al. Nat. Methods* **11**, 51–54 (2014).
43. Kukurba, K.R. *et al. PLoS Genet.* **10**, e1004304 (2014).
44. Rivas, M.A. *et al. Science* **348**, 666–669 (2015).
45. Baran, Y. *et al. Genome Res.* **25**, 927–936 (2015).
46. Pirinen, M. *et al. Bioinformatics* **31**, 2497–2504 (2015).
47. Dominissini, D. *et al. Nature* **485**, 201–206 (2012).
48. Meyer, K.D. *et al. Cell* **149**, 1635–1646 (2012).
49. Ciaccio, M.F., Wagner, J.P., Chuu, C.P., Lauffenburger, D.A. & Jones, R.B. *Nat. Methods* **7**, 148–155 (2010).
50. O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E. & Snyder, M.P. *Proc. Natl. Acad. Sci. USA* **109**, 18018–18023 (2012).
51. Haycock, P.C. *et al. Br. Med. J.* **349**, g4227 (2014).
52. Stone, R.C. *et al. PLoS Genet.* **12**, e1006144 (2016).
53. Kibriya, M.G., Jasmine, F., Roy, S., Ahsan, H. & Pierce, B. *Cancer Epidemiol. Biomarkers Prev.* **23**, 2667–2672 (2014).
54. Pierce, B.L. *et al. Int. J. Mol. Epidemiol. Genet.* **7**, 18–23 (2016).
55. Kibriya, M.G., Jasmine, F., Roy, S., Ahsan, H. & Pierce, B.L. *PLoS One* **11**, e0155548 (2016).
56. Jagannathan, S. & Bradley, R.K. *Genome Res.* **26**, 1639–1650 (2016).
57. Gamazon, E.R. *et al. Nat. Genet.* **47**, 1091–1098 (2015).
58. Nica, A.C. *et al. PLoS Genet.* **6**, e1000895 (2010).
59. Hormozdiari, F. *et al. Am. J. Hum. Genet.* **99**, 1245–1260 003 (2016).
60. Civelek, M. & Lusis, A.J. *Nat. Rev. Genet.* **15**, 34–48 (2014).
61. Parikshak, N.N., Gandal, M.J. & Geschwind, D.H. *Nat. Rev. Genet.* **16**, 441–458 (2015).
62. Zhu, J. *et al. PLoS Biol.* **10**, e1001301 (2012).

eGTEx Project:
Barbara E Stranger[1–3], Lori E Brigham[4], Richard Hasz[5], Marcus Hunter[6], Christopher Johns[7], Mark Johnson[8], Gene Kopen[9], William F Leinweber[9], John T Lonsdale[9], Alisa McDonald[9], Bernadette Mestichelli[9], Kevin Myer[6], Brian Roe[6], Michael Salvatore[9], Saboor Shad[9], Jeffrey A Thomas[9], Gary Walters[8], Michael Washington[8], Joseph Wheeler[7], Jason Bridge[10], Barbara A Foster[11], Bryan M Gillard[11], Ellen Karasik[11], Rachna Kumar[11], Mark Miklos[11], Michael T Moser[11], Scott D Jewell[12], Robert G Montroy[12], Daniel C Rohrer[12], Dana R Valley[12], David A Davis[13], Deborah C Mash[13], Sarah E Gould[14], Ping Guan[15], Susan Koester[16], A Roger Little[17], Casey Martin[14], Helen M Moore[15], Abhi Rao[15], Jeffery P Struewing[14], Simona Volpi[14], Kasper D Hansen[18–20], Peter F Hickey[20], Lindsay F Rizzardi[18], Lei Hou[21,22], Yaping Liu[21,22], Benoit Molinie[22], Yongjin Park[21,22], Nicola Rinaldi[21,22], Li Wang[22], Nicholas Van Wittenberghe[22], Melina Claussnitzer[22–24], Ellen T Gelfand[22], Qin Li[25], Sandra Linder[25,26], Rui Zhang[25], Kevin S Smith[26], Emily K Tsang[26,27], Lin S Chen[28], Kathryn Demanelis[28], Jennifer A Doherty[29], Farzana Jasmine[28], Muhammad G Kibriya[28], Lihua Jiang[25], Shin Lin[25,30], Meng Wang[25], Ruiqi Jian[25], Xiao Li[25], Joanne Chan[25], Daniel Bates[31], Morgan Diegel[31], Jessica Halow[31], Eric Haugen[31], Audra Johnson[31], Rajinder Kaul[31], Kristen Lee[31], Matthew T Maurano[32], Jemma Nelson[31], Fidencio J Neri[31], Richard Sandstrom[31], Marian S Fernando[1,2], Caroline Linke[1,2], Meritxell Oliva[1,2], Andrew Skol[1–3], Fan Wu[1,2], Joshua M Akey[33,34], Andrew P Feinberg[18,35–37], Jin Billy Li[25], Brandon L Pierce[28], John A Stamatoyannopoulos[31,38,39], Hua Tang[25], Kristin G Ardlie[22], Manolis Kellis[21,22], Michael P Snyder[25] & Stephen B Montgomery[25,26]

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, USA. [2]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois, USA. [3]Center for Data-Intensive Science, University of Chicago, Chicago, Illinois, USA. [4]Washington Regional Transplant Community, Annandale, Virginia, USA. [5]Gift of Life Donor Program, Philadelphia, Pennsylvania, USA. [6]LifeGift, Houston, Texas, USA. [7]Center for Organ Recovery and Education, Pittsburgh, Pennsylvania, USA. [8]LifeNet Health, Virginia Beach, Virginia, USA. [9]National Disease Research Interchange, Philadelphia, Pennsylvania, USA. [10]Unyts, Buffalo, New York, USA. [11]Pharmacology and Therapeutics, Roswell Park Cancer Institute, Buffalo, New York, USA. [12]Van Andel Research Institute, Grand Rapids, Michigan, USA. [13]Brain Endowment Bank, Miller School of Medicine, University of Miami, Miami, Florida, USA. [14]Division of Genomic Medicine, National Human Genome Research Institute, Rockville, Maryland, USA. [15]Biorepositories and Biospecimen Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland, USA. [16]Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, US National Institutes of Health, Bethesda, Maryland, USA. [17]National Institute on Drug Abuse, US National Institutes of Health, Bethesda, Maryland, USA. [18]Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [19]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [20]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [21]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [22]Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, USA. [23]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA. [24]Department of Natural Science, University of Hohenheim, Stuttgart, Germany. [25]Department of Genetics, Stanford University, Stanford, California, USA. [26]Department of Pathology, Stanford University, Stanford, California, USA. [27]Biomedical Informatics Program, Stanford University, Stanford, California, USA. [28]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA. [29]Huntsman Cancer Institute, Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA. [30]Division of Cardiology, University of Washington, Seattle, Washington, USA. [31]Altius Institute for Biomedical Sciences, Seattle, Washington, USA. [32]Institute for Systems Genetics, New York University Langone Medical Center, New York, New York, USA. [33]Lewis Sigler Institute, Princeton University, Princeton, New Jersey, USA. [34]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA. [35]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [36]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [37]Department of Mental Health, Johns Hopkins University School of Public Health, Baltimore, Maryland, USA. [38]Department of Medicine, University of Washington, Seattle, Washington, USA. [39]Department of Genome Sciences, University of Washington, Seattle, Washington, USA.

COMPETING FINANCIAL INTERESTS

M.P.S. is a cofounder of Personalis and Q bio and is on the scientific advisory boards of Personalis, Epinomics, and Genapsys.