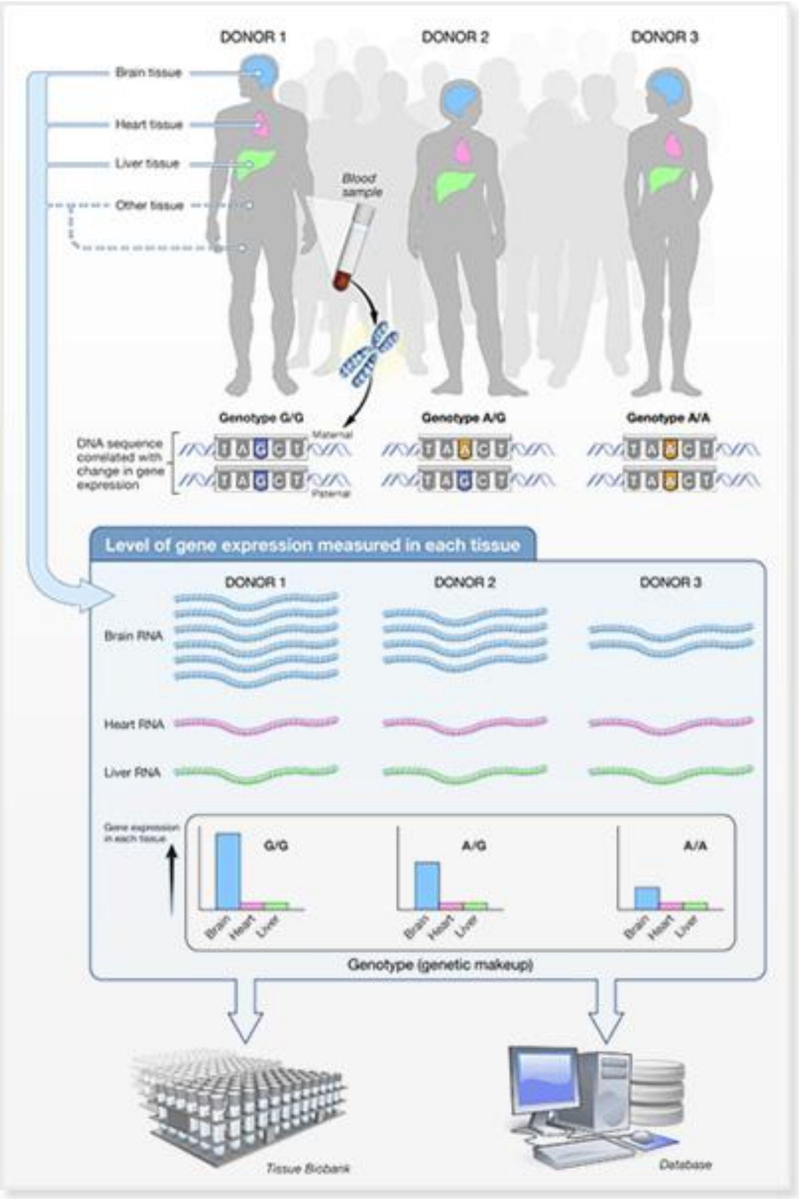
Supplementary Figures, Note, and Table

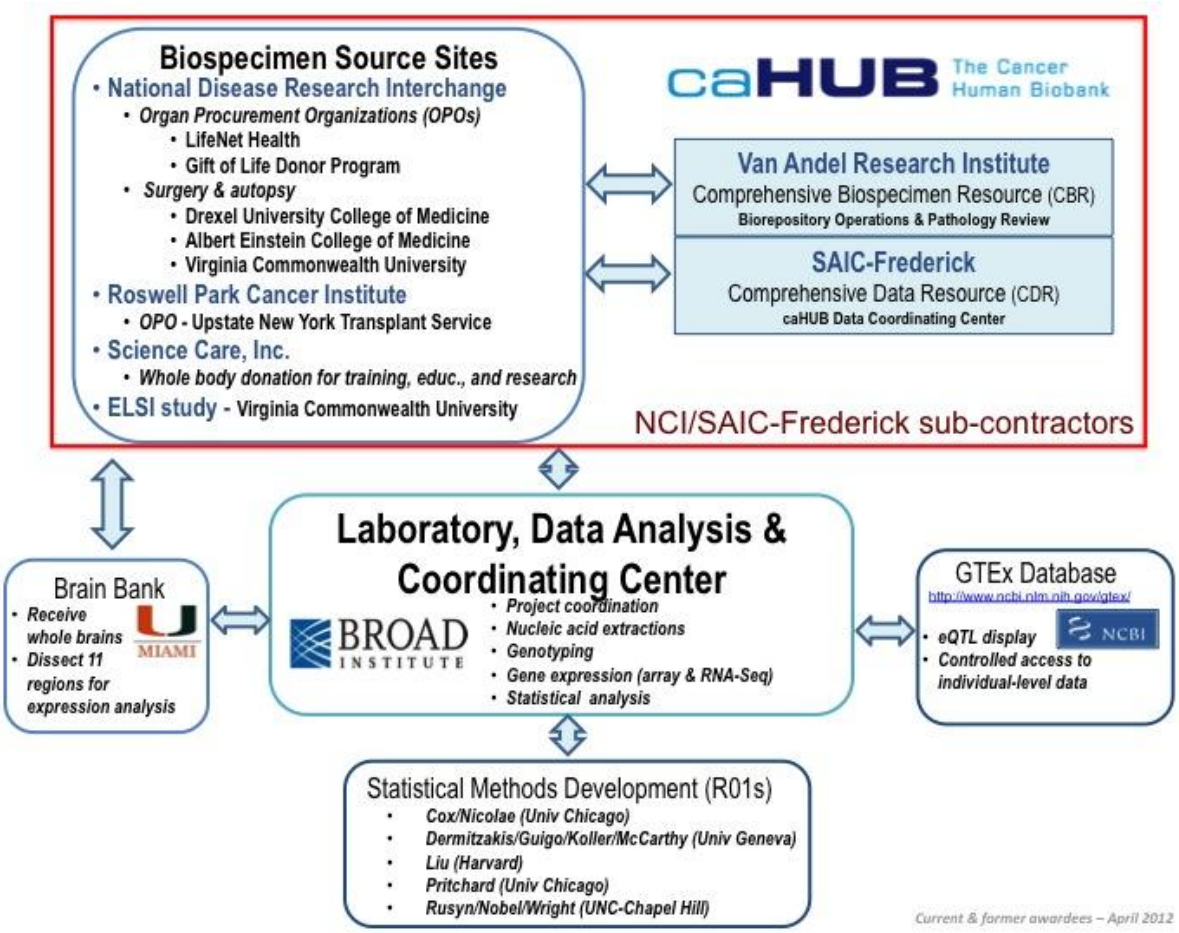
The Genotype-Tissue Expression (GTEx) project

The GTEx Consortium

Supplementary Figure 1: Design of the Genotype-Tissue Expression (GTEx) Project. Correlations between genotype and tissue-specific gene expression levels will help identify regions of the genome that influence whether and how much a gene is expressed. GTEx will help researchers to understand inherited susceptibility to disease and will be a resource database and tissue bank for many studies in the future.

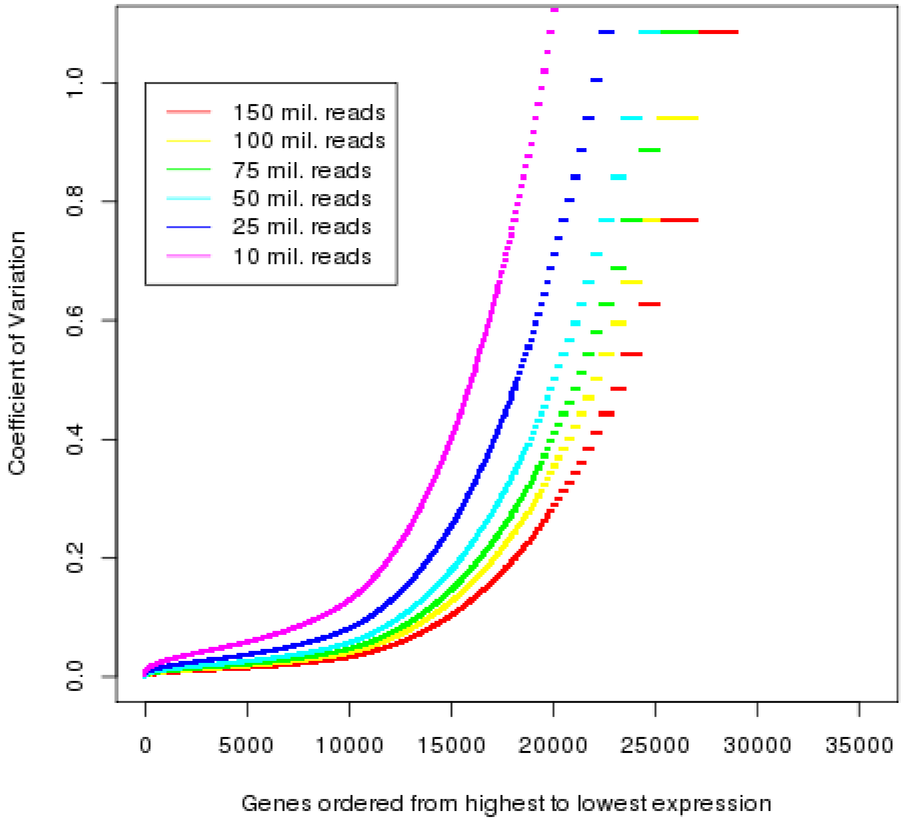


Supplementary Figure 2: The GTEx Project Consortium Organization.

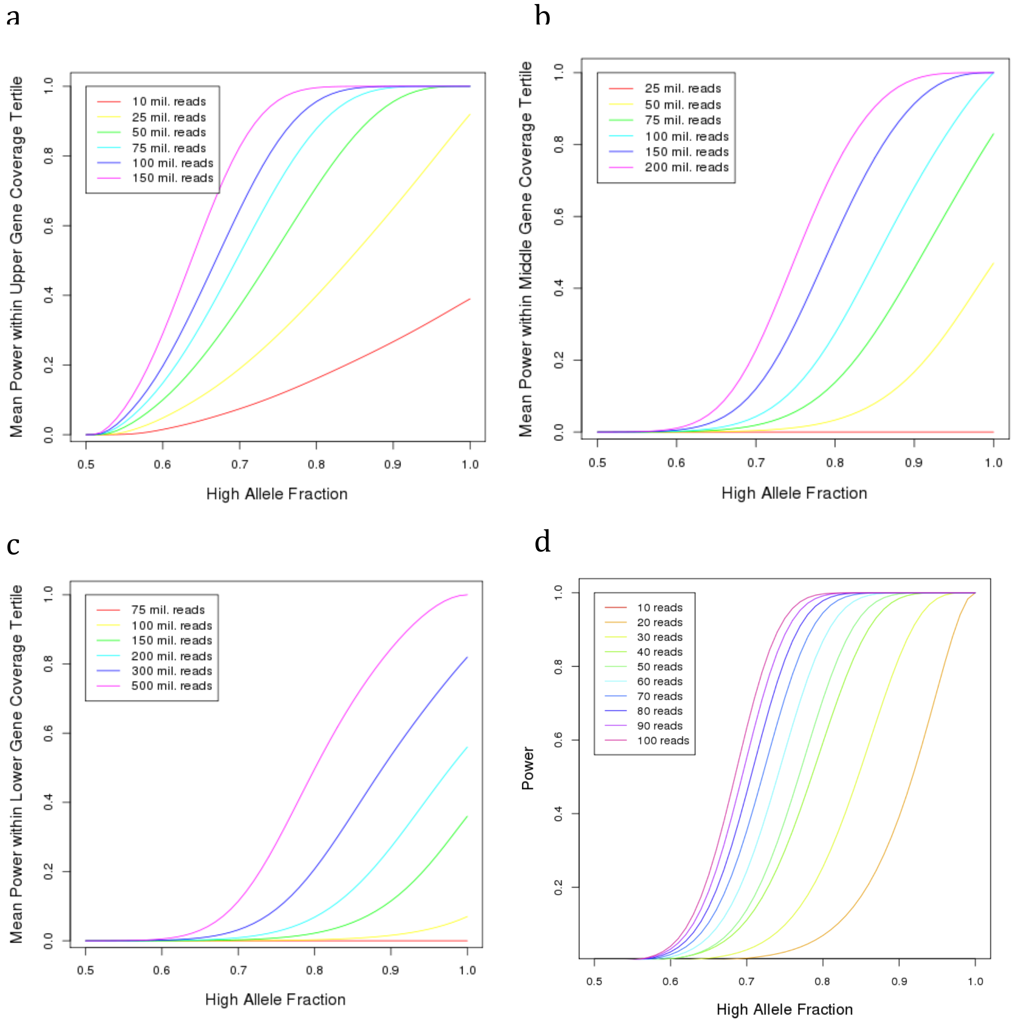

Biospecimen Source Sites
• National Disease Research Interchange
• Organ Procurement Organizations (OPOs)
• LifeNet Health
• Gift of Life Donor Program
• Surgery & autopsy
• Drexel University College of Medicine
• Albert Einstein College of Medicine
• Virginia Commonwealth University
• Roswell Park Cancer Institute
• OPO - Upstate New York Transplant Service
• Science Care, Inc.
• Whole body donation for training, educ., and research
• ELSI study - Virginia Commonwealth University
caHUB The Cancer Human Biobank
Van Andel Research Institute
Comprehensive Biospecimen Resource (CBR)
Biorepository Operations & Pathology Review
SAIC-Frederick
Comprehensive Data Resource (CDR)
caHUB Data Coordinating Center
NCI/SAIC-Frederick sub-contractors
Brain Bank
• Receive whole brains
• Dissect 11 regions for expression analysis
MIAMI
Laboratory, Data Analysis & Coordinating Center
BROAD INSTITUTE
• Project coordination
• Nucleic acid extractions
• Genotyping
• Gene expression (array & RNA-Seq)
• Statistical analysis
GTEx Database
http://www.ncbi.nlm.nih.gov/gtex/
• eQTL display
• Controlled access to individual-level data
NCBI
Statistical Methods Development (R01s)
• Cox/Nicolae (Univ Chicago)
• Dermitzakis/Guigo/Koller/McCarthy (Univ Geneva)
• Liu (Harvard)
• Pritchard (Univ Chicago)
• Rusyn/Nobel/Wright (UNC-Chapel Hill)
Current & former awardees – April 2012

Supplementary Figure 3 – Sampling variation of gene expression.

Variation in observed expression levels due to the sampling process of RNA-seq. The coefficient of variation (CV) modeled using the observed read distribution in a deep RNA-seq run (176 Million reads) and scaled down to yields between 10 and 150 million reads. The CV (=standard deviation/mean) is estimated as $\sqrt{l}/l$ following a Poisson distribution. The GTEx target of 50 million mapped reads provides a CV of less than 0.1 for >12,000 genes.

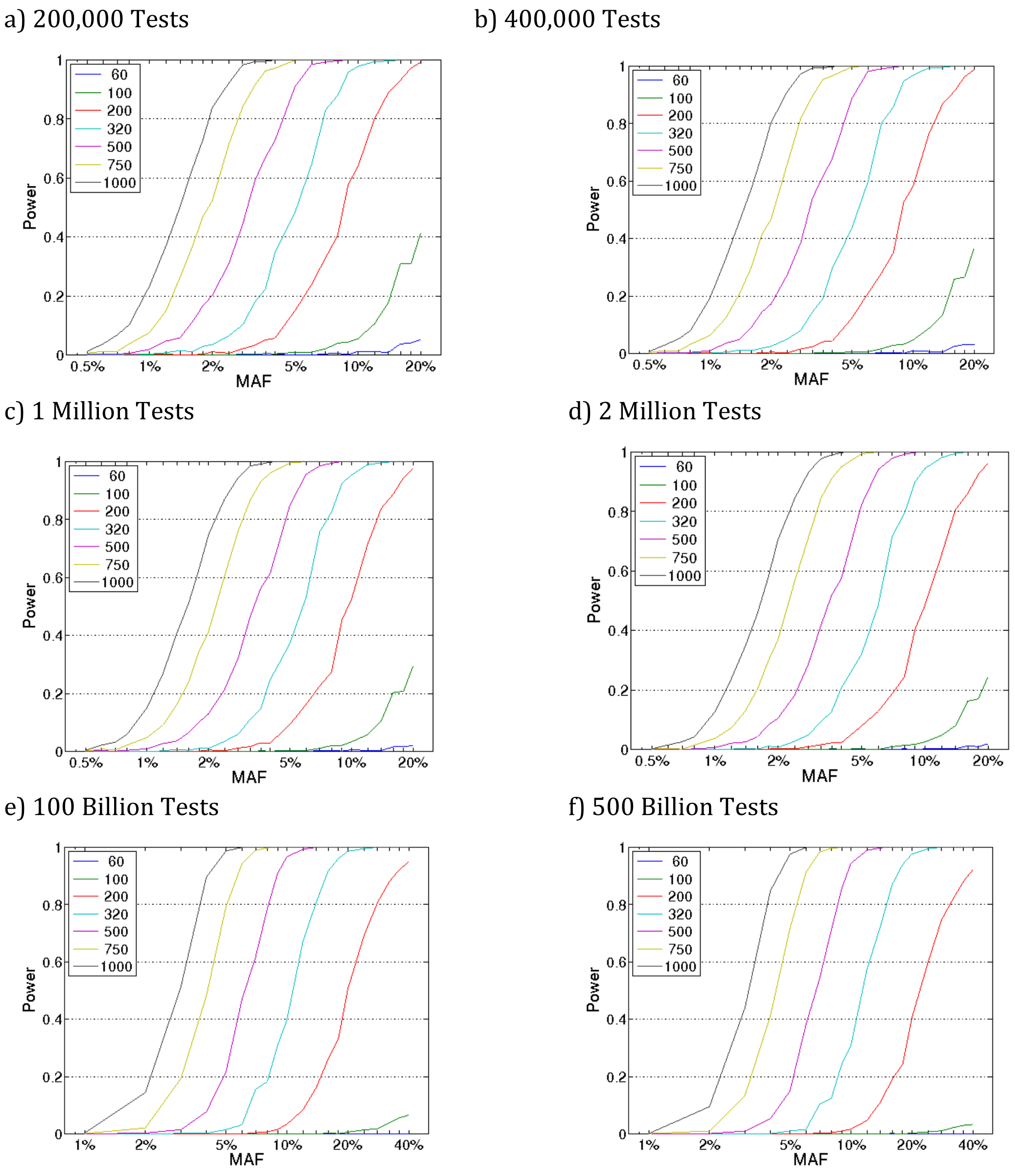


Supplementary Figure 4 - Power Analysis for Allele Specific Expression (ASE). Power analysis for allele specific expression, segregated by tertiles of sequencing coverage per gene (a: upper coverage tertile; b: middle coverage tertile; c: lower coverage tertile). The mean power for genes within a tertile is reported as a function of effect size (expressed as high allele fraction). A range of sequencing yields (from 10 million to 500 million reads) was assessed. 50 million reads are necessary to discover ASE in the upper tertile (a) of genes. For the lower tertile (c), as much as 500 million reads would be necessary to power ASE analysis. (d) Power analysis for allele specific expression based on coverage at heterozygous sites. Power is given as a function of effect size (x axis, expressed as high allele fraction) as well as the number of reads covering the putative ASE site (from 20 to 100).



Supplementary Figure 5 - Data for eQTL Power Analysis.

Increasing the number of *cis* eQTL tests performed has a modest effect on power. The power to detect eQTLs is given as a function of the minor allele frequency. The results for a range of sample sizes (60 to 1000) are given in each plot. Comparing 200,000 tests (a) with two million tests (d) reveals an increase in detectable MAF by 1%. In the case of 750 samples, increasing the tests from 200,000 to 2 million increases the detectable MAF from 3 to 4%. This relatively modest effect indicates that this power analysis is reasonably stable with regard to α. For *trans* eQTLs (e and f), the number of SNPs is constant at 5 million, but we can vary the number of transcripts from 20,000 to 100,000. Accordingly, increasing the number of tests from 1 billion to 5 billion results in a 1% shift in detectable MAF from 4 to 5%.


a) 200,000 Tests

b) 400,000 Tests

c) 1 Million Tests

d) 2 Million Tests

e) 100 Billion Tests

f) 500 Billion Tests

Supplementary Table 1: Complete list of tissues collected in at least one GTEx donor.
Adipose, Subcutaneous #
Adipose, Visceral (Omentum)
Adrenal Gland
Artery - Aorta (Ascending or thoracic)
Artery - Coronary
Artery - Tibial #
Bladder - Urinary *
Blood #
Brain, Cerebellar Hemisphere
Brain, Frontal Cortex
Brain, Hippocampus
Brain, Substantia nigra
Brain, Anterior cingulate cortex (BA24)
Brain, Amygdala
Brain, Caudate (basal ganglia)
Brain, Nucleus accumbens (basal ganglia)
Brain, Putamen (basal ganglia)
Brain, Hypothalamus
Brain, Spinal cord (cervical c-1)
Breast, Mammary Tissue
Cervix - Ectocervix *
Cervix - Endocervix *
Colon - Sigmoid
Colon - Transverse
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube *
Gastroesophageal junction
Heart - Right Atrium
Heart - Left Ventricle
Ilium - Terminal (Peyer's patch)
Kidney - Cortex
Kidney - Medulla *
Liver
Lung
Muscle, Skeletal #
Nerve - Tibial #
Ovary
Pancreas
Pituitary Gland
Prostate
Salivary Gland, Minor
Skin - Sun exposed (Leg) #
Skin - Not sun exposed (Suprapubic area)
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina

# These tissue samples are also being collected from surgical donors.
* Tissues collected only during the early pilot phase for which there will likely be fewer than 100 in total.

Supplementary Note - Power Analysis for Allele Specific Expression (ASE).

The power to detect ASE is primarily influenced by sequencing depth. The following power analysis is aimed at providing a sense of which ASE effect sizes will be detectable given the target sequencing depth of 50 million mapped reads in the GTEx project. We apply the binomial test as a means of detecting whether an observed allelic ratio in heterozygous site-spanning reads is significantly different from 0.5 (after correcting for multiple hypotheses). The coverage at heterozygous sites is not only dependent on the sequencing yield, but is also a function of the level of gene expression. We estimate the coverage distribution of expressed heterozygous sites using preliminary GTEx deep RNA-seq data.

One of the most interesting applications of GTEx data will be allele specific expression (ASE) analysis within eQTL genes. As such, we correct for 500 hypotheses in this analysis representing a rough estimate for the number of eQTL genes containing heterozygous sites within the coding region. Therefore we use as a significant threshold, $\alpha=0.05/500=10^{-4}$. We calculated the power within the lower, middle and upper tertiles of the observed coverage distribution. Within each tertile we report the mean power as a function of effect size. This analysis was repeated for a range of sequencing yields between 10 and 150 million reads (**Supplementary Figure 4a-c**, below), scaling the coverage distribution accordingly.

For a yield of 150 millions reads there were 14,933 genes with an average depth of at least 5 reads, or 3393 genes per tertile. For the upper tertile of genes, the targeted 50 million reads will provide sufficient power (>80%) to detect an expected allele fraction of 0.83, i.e. an odds-ratio of ~4.9. This power is achieved when having at least 40 reads covering the heterozygous site (**Supplementary Figure 4d**). Comparable power within the middle tertile would require 150 million reads. Based on current variations in sequencing yield, we expect several dozen GTEx samples (~3%) to achieve this depth providing extra power for the middle tertile. It appears that detecting ASE within the lower tertile would require over 500 million reads which could be achieved in a more targeted approach.