Genetic variation and gene expression across multiple tissues and developmental stages in a non-human primate

Anna J. Jasinska[1,2], Ivette Zelaya[3], Susan K. Service[1], Christine Peterson[4], Rita M. Cantor[1,5], Oi-Wa Choi[1], Joseph DeYoung[1], Eleazar Eskin[5,6], Lynn A. Fairbanks[1], Scott Fears[1], Allison E. Furterer[7], Yu S. Huang[1,8], Vasily Ramensky[1], Christopher A. Schmitt[1,9], Hannes Svardal[10], Matthew J. Jorgensen[11], Jay R. Kaplan[11], Diego Villar[12], Bronwen L. Aken[13], Paul Flicek[13], Rishi Nag[13], Emily S. Wong[13], John Blangero[14], Thomas D. Dyer[14], Marina Bogomolov[15], Yoav Benjamini[16], George M. Weinstock[17], Ken Dewar[18], Chiara Sabatti[19], Richard K. Wilson[20,21], J. David Jentsch[22,23,24], Wesley Warren[20], Giovanni Coppola[1,25], Roger P Woods[23,25], Nelson B Freimer[1,5*]

[1]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA;
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[3]Interdepartmental Program in Bioinformatics, University of California Los Angeles, Los Angeles CA, USA
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston TX, USA
[5]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA,USA
[6]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA
[7]Interdepartmental Graduate Program in Neuroscience, University of California Los Angeles, Los Angeles CA, USA
[8]Current address: State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China
[9]Current address: Department of Anthropology, Boston University, Boston, MA, USA
[10]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[11]Department of Pathology, Wake Forest School of Medicine, Winston-Salem, NC, USA
[12]University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, UK
[13]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[14]South Texas Diabetes and Obesity Institute, UTHSCSA/UTRGV, Brownsville, TX, USA
[15]Faculty of Industrial Engineering and Management, Technion, Haifa, Israel
[16]Department of Statistics and Operation Research, Tel Aviv University, Tel Aviv, Israel

[17]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[18]Department of Human Genetics, McGill University, Montreal, Quebec, Canada
[19]Departments of Biomedical Data Science and Statistics, Stanford University, Stanford, California, USA
[20]The McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[21]Current Address: Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA
[22]Department of Psychology, University of California, Los Angeles, Los Angeles, California, USA
[23]Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[24]Current Address: Department of Psychology, Binghamton University, Binghamton, NY, USA
[25]Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles CA, USA
* to whom correspondence should be addressed at: nfreimer@mednet.ucla.edu

By analyzing multi-tissue gene expression and genome-wide genetic variation data in samples from a vervet monkey pedigree, we generated a transcriptome resource and produced the first catalogue of expression quantitative trait loci (eQTLs) in a non-human primate model. This catalogue contains more genome-wide significant eQTLs, per sample, than comparable human resources, and reveals sex and age-related expression patterns. Findings include a master regulatory locus that likely plays a role in immune function, and a locus regulating hippocampal long non-coding RNAs (lncRNAs) whose expression correlates with hippocampal volume. This resource will facilitate genetic investigation of quantitative traits, including brain and behavioral phenotypes relevant to neuropsychiatric disorders.

Efforts to understand how genetic variation contributes to common diseases and quantitative traits increasingly focus on the regulation of gene expression. This focus reflects the finding that most genetic loci demonstrating significant association to such phenotypes in genome wide association studies (GWAS) lie in non-coding portions of the genome[1], and are enriched for eQTLs; SNPs that regulate transcript levels, primarily those of nearby genes[2]. This observation has led to the idea that detailed genome wide eQTL catalogs may provide signposts for the specific variants responsible for GWAS signals[3].

The majority of known human eQTLs have been identified through gene expression studies in lymphocytes or lymphoblastoid cell lines obtained from adults[4]. As normal development and function in complex organisms depends on tightly regulated gene expression at specific developmental stages in specific cell types, most existing datasets describing human transcriptome characterization likely miss the data most relevant to understanding disease[5]. This lack is particularly striking for brain and behavior disorders, given the inaccessibility of the most relevant tissues in living individuals and the enormous modifications that occur in these tissues across development[6].

To remedy the lack of human data connecting genotypic variation and multi-tissue transcriptome variation, the United States National Institutes of Health launched the Genotype Tissue Expression (GTEx) project, using samples obtained from several hundred post-mortem donors[7]. GTEx has already provided an eQTL catalog, from multiple tissues, that is the most extensive such resource available[7]. However several limitations of GTEx, inherent to human research, motivate the generation and investigation of equivalent resources from model organisms. The advantages of model systems for genetic investigation of the regulation of gene expression include: (1) the feasibility of controlling for inter-individual heterogeneity in environmental exposures and of minimizing the interval between death and tissue preservation; (2) the practicability of obtaining sizable numbers of multi-tissue samples across a full range of developmental stages; and (3) the opportunity to systematically assess phenotypes of interest in the individuals from whom tissue samples are obtained. Non-human primate (NHP) species offer an additional advantage[8]; their brain circuitry, behavior, immune systems, and metabolism more closely resemble those of humans than do those of rodents and other model systems[9].

We report here, in a sample of Caribbean vervets (*Chlorocebus aethiops sabaeus*) from the Vervet Research Colony (VRC) extended pedigree, the first NHP resource combining genome-wide genotypes, multi-tissue expression data across post-natal development, and quantitative phenotypes relevant to human brain and behavior disorders. This Old World monkey population has expanded dramatically from a founding bottleneck occurring with the introduction of West African vervets to three Caribbean islands in the 17th Century[9]; it has experienced a drastic reduction in genetic variation and, like recently expanded human population isolates, displays an enrichment for potentially deleterious alleles at loci throughout the genome (Ramensky, unpublished data). Through necropsies performed under uniform

conditions, we obtained tissue samples from a series of vervets in which key environmental exposures, such as diet, had been carefully controlled. Using these resources we have delineated cross-tissue expression profiles across multiple developmental stages from birth to adulthood. We identified numerous local and distant eQTLs in each tissue, including a master regulatory locus that, via *IFIT1B*, a gene with a hypothesized role in immunie function, modulates expression in blood cells of multiple genes on several chromosomes. Additionally, we demonstrated the relevance of vervet tissue-specific eQTLs to higher-order traits, using hippocampus-specific local eQTLs to identify a set of lncRNAs as candidate regulators of hippocampal volume, a phenotype related to neuropsychiatric disorders[10].

Results

We investigated two datasets. Dataset 1, as described previously[11], consists of gene expression levels in 347 vervets obtained by hybridizing whole blood-derived RNA to microarrays designed for human sequences (Illumina HumanRef-8 v2). After filtering out probe sequences that were not represented in the vervet genome[12] or that contained common vervet SNPs[13], we estimated expression levels at 6,018 probes (Supplementary Data 1, Supplementary Table 1). Dataset 2 consists of RNA sequencing (RNA-Seq) reads from six tissues collected under identical conditions from each of 58 sequenced VRC monkeys (representing 10 developmental stages, from birth through adulthood, see Methods). Four of these tissues (caudate nucleus, hippocampus, pituitary, and adrenal) are components of a neuroendocrine circuit that plays a prominent role in brain and behavior[14,15] but is underrepresented in human studies, because of the inaccessibility of the constituent tissues. The other two tissues (cultured skin fibroblasts and whole blood) are readily accessible in humans, and thus have been more widely used in functional genomics studies. Before analyzing Dataset 2, we minimized spurious signals by excluding genes expressed in fewer than 10% of individuals or at a level lower than one read per tissue. Table 1 presents the number of genes analyzed for each of the six tissues. While most genes were expressed in multiple tissues, 147 genes demonstrated strong expression in only a single tissue (Supplementary Table 2). A principal components analysis (PCA) of Dataset 2 indicates that most tissues cluster separately from each other (Supplementary Fig. 1).

Multi-tissue expression data: variation by age and sex

The availability, in Dataset 2, of multiple samples from both sexes at each age/developmental time point enabled us to examine developmental trajectories and sex differences in gene expression for each tissue. To do so we conducted PCA on the expression of the 1,000 most variable genes, separately by tissue (Fig. 1). Among the six tissues, the pattern in the caudate nucleus displays the clearest association with development; PC1 distinguishes the monkeys in a nearly linear manner, with increasing age. All tissues except fibroblast show a sharp demarcation in expression pattern between males and females; this differentiation is observed on

PC1 for hippocampus and pituitary, on PC2 for caudate and blood, and on PC3 for adrenal.

As an initial exploration of the biology underlying the tissue-related expression patterns in the brain-neuroendocrine circuit examined here, we identified in each tissue, the genes in the top and bottom 10% of the distribution of PC loadings on PCs 1, 2, or 3 (200 genes total per tissue, per PC), in relation to age (caudate) or sex (caudate, hippocampus, pituitary, and adrenal) (Supplementary Tables 3, 4). The list of genes with age-related expression patterns in the caudate includes several genes that are both essential for nervous system development and implicated in the causation of human disorders. Figure 2 shows expression patterns by age-point of some notable examples; *ASPM*, which regulates neurogenesis in the cerebral cortex and, when mutated, results in primary microcephaly[16]; *NDRG1*, which (i) stimulates cellular differentiation and proliferation, (ii) is commonly involved in somatic rearrangements leading to medulloblastoma (the most prevalent pediatric brain tumor[17]), and (iii) when mutated, causes a progressive peripheral neuropathy, Charcot-Marie-Tooth (CMT) disease Type 4d; *GJB1*, which encodes a gap-junction protein mutated in CMT Type X[18]; and *HSPB8*, which encodes a heat-shock protein mutated in two peripheral neuropathies, axonal CMT Type 2L and distal hereditary motor neuropathy type IIA[19,20]. Several genes on this list contribute to postnatal myelination of the central nervous system[16,18,20,21], suggesting the possibility that the caudate age-related expression pattern at least partially reflects this process.

Among genes contributing to the sex-related expression patterns in specific tissues, perhaps the most striking examples are genes encoding the receptors for two structurally similar neuropeptides, the oxytocin receptor *OXTR* (in caudate) and the vasopressin receptor *AVPR1A* (in hippocampus). These two genes function in a sex specific manner mediated by the sex-steroids estrogen (for *OXTR*) and androgen (for *AVPR1A*)[22]. The distribution and function of these genes also differ dramatically between mammalian species, including among some that are closely related. For example, a polymorphism in the promoter of *Avpr1a* has been associated, in prairie voles, with inter-species differences in sex-related social behaviors, such as pair bonding[23]. In monogamous prairie vole males, additional polymorphisms in this region result in inter-individual variation in expression of *Avpr1a* in hippocampus and other tissues comprising a memory circuit, and these expression differences are related to sex-specific spatial behaviors[24]. Another notable example from the hippocampus gene list is *PDYN*, encoding the opioid peptide dynorphin, which modulates hippocampal synaptic plasticity[25] in a sex-specific manner, mediated by estrogen[26].

The lists of genes with sex-related expression patterns in pituitary and adrenal overlap substantially (40 of the top 200 genes in common) and include several molecules with functions in reproduction or in biological processes with a marked sex bias. Tissue-specific examples include, for pituitary, *TAC1*, encoding substance P, which regulates puberty onset and fertility[27] and *PTGER2* which plays a role in

ovulation and fertilization[28], and, for adrenal, *PRL*, which stimulates lactation in new mothers, plays a role in social behaviors, and is regulated by adrenal steroids[29].

Identification of eQTLs

We previously reported the first NHP genome-wide, high-resolution genetic variant set[13], obtained by whole genome sequencing (WGS) of 721 monkeys from the VRC pedigree, aligning sequencing reads to the vervet reference genome[12], calling variants, and identifying 497,163 WGS-based SNPs that tag common variation in most linkage disequilibrium (LD) blocks genome-wide. Using these SNPs we conducted separate GWAS of Datasets 1 and 2 to identify local (probes/genes < 1 Mb from an associated SNP) and distant (all other probe/gene-SNP associations) eQTLs in each dataset.

We used SOLAR[30] to estimate heritability of probe expression in Dataset 1, identifying significant heritability for 3,417 probes at a false discovery rate [FDR] threshold < 0.01 (Supplementary Data 1, 2). In a GWAS of each heritable probe, we identified 461 local and 215 distant eQTLs that were significant at Bonferroni-corrected thresholds of 4.8×10^{-8} for local and 1.5×10^{-11} for distant eQTLs (Table 1, Supplementary Data 3). Approximately 35% of probes with a significant eQTL (173/498) displayed at least one local *and* one distant significant association.

We conducted a GWAS for each gene-tissue combination in Dataset 2, using as phenotypes the full set of genes that passed expression level thresholds for that tissue. In this dataset we observed, for each of the four solid tissues, between 338-537 local eQTLs and 39-71 distant eQTLs, and for blood and fibroblasts, 70 and 196 local eQTLs and 4 and 23 distant eQTLs, respectively, all at Bonferroni corrected significance thresholds (7.6×10^{-10} [local] and 6.1×10^{-13} [distant]) (Table 1, Supplementary Data 4). The smaller number of eQTLs observed in blood likely reflects the high degree of inter-individual variability in the proportions of different cell types in this tissue compared to the other tissues analyzed (Supplementary Fig. 1); we have no obvious explanation for the relative paucity of eQTLs in fibroblasts, aside from the observation that fewer genes were analyzed in fibroblasts than in tissues with cellular heterogeneity.

The significant eQTLs in Table 1 exceeded highly conservative Bonferroni thresholds. We also applied FDR controlling procedures, to expand the list of local eQTLs for more exploratory investigations, and to make our results comparable to those of resources such as GTEx. Specifically, for the discovery of eGenes we controlled the FDR at 0.05 (see Methods), accounting for multiple testing using a hierarchical error controlling procedure developed for GWAS of multiple phenotypes[31], extended here to accommodate the analysis of multiple tissues. We observe that, in comparison with GTEx V6, despite having a smaller sample size we identify more local eQTLs for all tissues except for whole blood and adrenal (Table 2). We attribute the larger number of local eQTLs identified in the vervet sample,

relative to GTEx, to the more homogenous environment of colonied NHPs compared to humans, and to the more uniform process of collecting tissues in this study.

We considered the possibility that genotypic variation within the vervet pedigree could confound the effects of age in generating the strong loadings on PCs associated with age in caudate. Among the 200 genes with such strong loadings, 33 genes showed evidence of eQTLs, even when using the more liberal FDR controlling procedure. For these 33 genes, we modeled expression as a function of both age and genotype, using the most significant eQTLs, and found that genotype could not account for the association with age (data not shown).

Genomic Distribution of eQTLs

Reasoning that regulatory variants are most frequent in functional genomic regions, we analyzed the distribution of all eQTL, in both expression data sets, with respect to gene boundaries (transcription start site [TSS] and transcription end site [TES]), promoters, and enhancers. We categorized all SNPs along two binary dimensions: influence on gene expression (eQTL, yes/no, for any tissue and either dataset, at Bonferroni corrected significance thresholds) and location (in putative functional region, yes/no) and evaluated the significance of the odds ratio (OR) using Fisher's Exact Test. To account for LD between SNPs, we selected 22,892 genome-wide SNPs by LD pruning the entire set of 497,163 SNPs at $r^2<0.6$ in 14 distantly related individuals. This SNP set included 1,663 eQTL SNPs (1,380 that are local eQTLs, 239 that are distant eQTLs, and 44 that are both local and distant eQTLs).

Gene regions encompassing exons, introns and adjacent flanks show a clear enrichment for eQTLs (Supplementary Fig. 2), which is more significant for local than distant eQTL, likely due to the larger number of loci in the former category. As in other primates[32], vervet eQTLs are more frequent around gene TSS and TES (Supplementary Fig. 3). Conversely, intergenic regions show a significant deficit of eQTLs, both distant and local (Supplementary Fig. 2).

Chromatin immunoprecipitation with DNA sequencing (ChIP-seq) experiments using vervet liver samples[33] enabled us to classify genomic regions as active promoters (sites enriched in either trimethylated lysine 4 of histone H3 (H3K4me3) or in both H3K4me3 and acetylated lysine 27 on histone H3 (H3K27ac) marks) or active enhancers (enriched in H3K27ac only)[33]. Vervet promoter regions show stronger enrichment for eQTLs than either genic regions or enhancers (Supplementary Fig. 2). The combined set of all types of eQTLs (local and distant together), and local eQTLs alone showed the greatest enrichment in promoter regions dually-marked with both H3K27ac and H3K4me3, slightly lower enrichment in promoters marked with only H3K4me3 and a moderate enrichment in active enhancer regions (Supplementary Fig. 2). The depletion of eQTLs in intragenic regions and the eQTL enrichment in functional genomic regions is consistent with their presumed regulatory functions.

To evaluate whether vervet eQTLs, like human eQTLs, show enrichment at genome-wide significant human GWAS loci, we downloaded from www.ebi.ac.uk/gwas the coordinates of such loci, using LD to define a region of interest (ROI) around each associated index SNP. We then transposed these ROI coordinates to the vervet genome, and evaluated the association between local eQTLs and the vervet segments syntenic to the human GWAS ROI, using Fisher's Exact Test as described above. Vervet local eQTLs were enriched in these ROI: 13.2% of the vervet local eQTL SNPs lie in these ROI compared to 10.6% of the evaluated SNPs that are not vervet local eQTLs (OR=1.28, p=0.0027).

Validation of Distant eQTLs: a Master Regulatory Locus on Vervet Chromosome 9

Unlike most human expression data sets, our Dataset 1 is well-powered for discovery of distant eQTLs. Among the 215 such eQTLs that we identified at genome-wide significance thresholds, two loci stood out because they showed association to multiple unlinked genes. On vervet chromosome (CAE) 5, four SNPs in a 322 Kb region were each associated to five distant genes (Supplementary Table 5), but no local genes. On CAE 9, 76 SNPs across a ~500 Kb region displayed genome-wide significant local eQTL signals. Additionally, for each of these SNPs we identified multiple genome-wide significant distant eQTLs (ranging between five and 14 genes, on different vervet chromosomes, for each SNP, Fig. 3, Supplementary Table 5).

We evaluated Dataset 2 for replication of the CAE 5 and CAE 9 distant eQTLs, recognizing that the much smaller size of this dataset gave us limited power to achieve replication. Because the Dataset 2 data were obtained using a different platform from those in Dataset 1, and were from a mostly non-overlapping sample of monkeys (only 6 monkeys were represented in both datasets), we considered that such a replication would provide a validation of these eQTLs.

For the SNPs at the CAE 5 locus we did not observe any eQTLs in Dataset 2 (at a threshold of a marginal $p<0.05$). At the CAE 9 locus, we confirmed the distant eQTLs for six of the genes that showed such eQTLs in Dataset 1 (*RANBP10, LCMT1, ST7, TMEM57, YPEL4, NARF*), with at least one SNP demonstrating association at a marginal $p<0.05$ and in the same direction (Supplementary Table 5), with *ST7* continuing to show association at a Bonferroni level ($p<2.35 \times 10^{-5}$). This result suggests that the CAE 9 eQTL represents a master regulatory locus (MRL). This genomic segment contains a cluster of acid lipase genes and interferon-inducible genes, including *IFIT1B* (Interferon-Induced Protein With Tetratricopeptide Repeats 1B), a gene recently implicated in viral resistance in vervets, but not humans[34]. The same SNPs contributing to the MRL are also local eQTLs for *IFIT1B*, at genome-wide significant levels.

We conducted further analyses in Dataset 1 for a SNP (CAE9_82694171) that is both a significant distant eQTL for all 14 genes and a local eQTL for *IFIT1B*, at Bonferroni

corrected significance thresholds (Supplementary Table 6). This SNP accounts for 19-37% of the variance in expression level of the 14 genes not located on CAE 9. When we conditioned these analyses on expression of *IFIT1B*, the magnitude of these distant associations diminished substantially, with the percentage of variance accounted for by this SNP dropping to 10% or less for the 14 genes. Overall, these results support a scenario in which *IFIT1B*, under direct control of a local eQTL on CAE 9, influences expression of 14 other genes spread across the genome. As suggested by studies in human populations, such mediation by local eQTLs of distant eQTLs provides a further validation of the latter loci[35].

Identification of Hippocampus-Specific eQTLs in a Region Linked to Hippocampal Volume

As an initial investigation of the impact of vervet tissue-specific eQTLs on higher order traits we focused on a neuroanatomic trait, hippocampal volume. A previous MRI-based study in the VRC showed extremely high heritability (h^2 =0.95) for this phenotype[36]. We reasoned that genome wide QTL analysis of MRI-based hippocampal volume, in conjunction with eQTL analysis of hippocampal RNA, could reveal variants contributing to this trait. The strongest QTL signal for hippocampal volume (peak LOD score 3.42) occurred in an ~8.3 Mb segment of CAE 18. We identified in the center of this region, two hippocampus-specific local eQTLs, Bonferroni-significant at a genome-wide threshold, together with the genes that these eQTLs regulate (Fig. 4).

The genome-wide significant eQTL SNPs reside in, and regulate expression of, two lncRNAs located at a distance of 168 Kb from each other: *LOC103222765* (nine associated local eQTL SNPs) and *LOC103222769* (three associated local eQTL SNPs). An additional lncRNA gene, *LOC103222771*, situated two bp from *LOC103222769*, shows hippocampal specific association to six SNPs at a significance level ($p < 10^{-9}$) just above the genome-wide Bonferroni-corrected threshold. While all three genes display hippocampus-specific eQTLs, the genes themselves are expressed across all six tissues that we analyzed, and show no significant sex or age specific differences in expression patterns (data not shown). The incomplete database annotation of lncRNAs[37] limits comparative analyses of such genes among primates; a BLAST search found a homolog for *LOC103222765* in the white-tufted-ear marmoset and one for *LOC103222771*, in the crab-eating macaque. While *LOC103222765* overlaps a coding gene (*RAB31*), *LOC103222769* and *LOC103222771* do not overlap exons of any coding genes and therefore are more specifically classified as long intergenic non-coding RNA (lincRNA) genes, a class of genes known to function in cell differentiation and developmental regulation[38].

Given the physical proximity of these lncRNAs, we used multivariate conditional analyses to evaluate whether the regulation of these genes depends on a single or multiple independent eQTLs. For each lncRNA we designated a "lead SNP" (the SNP most significantly associated to its expression, Supplementary Table 7). For both *LOC103222769* and *LOC103222771*, modeling expression as a function of both lead

SNPs results in diminished significance levels for both SNPs (Supplementary Table 7), suggesting that one eQTL regulates both genes. Modeling *LOC103222765* expression as a function of its lead SNP and the lead SNP of the other two genes, the lead SNP for *LOC103222765* remains significant, while the other two lead SNPs are non-significant, confirming the "distinctness" of this signal (Supplementary Table 7). This analysis suggests two eQTLs in this region; one associated to *LOC103222765*, and the second associated to *LOC103222769* and *LOC103222771*.

We observed a positive correlation between hippocampal expression of *LOC103222765*, *LOC103222769* and *LOC103222771*, and hippocampal volume as assessed by MRI, in six monkeys for which both MRI and RNA-Seq data were available. To extend this observation, we assessed, using an independent platform, quantitative real-time PCR (qRT-PCR), *LOC103222765*, *LOC103222769* and *LOC103222771* hippocampal expression in these six monkeys and 10 additional monkeys for which both hippocampal RNA and MRI data were available. In this expanded sample set, we identified significant positive correlations (Fig. 5) between *LOC103222765, LOC103222769* and *LOC103222771* expression and hippocampal volume, suggesting that genetic variation regulating these lncRNAs has a strong impact on the MRI phenotype.

Discussion

The data presented here represent a first attempt to create an NHP resource for investigating the genetic contribution to inter-individual variation in gene expression across multiple tissues and across development. This vervet resource provides a complement to GTEx, which has become an essential tool for pinpointing the genes, and even the variants, underlying disease loci revealed by GWAS of human populations[39,40]. As with GTEx in humans, the sequenced multi-tissue vervet samples described here constitute a powerful asset, in a species closely related to humans, for investigating the genetic regulation of transcriptome variation.

Several features, however, differentiate the vervet resource from GTEx, reflecting aspects of the study design that are infeasible in human research. Notably, the age-based sampling design enabled us to delineate tissue-specific expression profiles in relation to developmental trajectories. Delineating these trajectories provides insights into biological processes that may be associated with the expression profiles of particular genes. For example, several genes that contribute to postnatal myelination of the central nervous system[16,18,20,21] contribute to the near linear age-related pattern observed in caudate, and suggest the possibility that the observed expression pattern at least partially reflects this process. So far we have only observed such a clear trajectory in this tissue. The tissues examined to date are, however, only a fraction of those available from the same set of monkeys. It will be possible to extend the investigations reported here to samples from an additional 60 brain regions and 20 peripheral tissues.

Three factors increased the relative signal-to-noise ratio of the vervet eQTL analyses: (i) the homogeneity of the vervet sample with respect to environmental exposures, such as diet; (ii) the greater control over necropsy conditions; and (iii) the restricted genetic background of the recently bottlenecked Caribbean vervet population. These factors enabled us to identify and validate a distant eQTL regulating expression of at least 14 distant genes. Furthermore, our data suggest that genetic variation regulating *IFIT1B*, one of a cluster of five *IFIT* genes, mediates this MRL.

The function of *IFIT1B* is poorly understood. It is a paralog of *IFIT1*, which is involved in innate antiviral immunity in mammals, broadly[41], and in regulation of gut microbiota in mouse[42]. Recent evidence indicates that in some mammalian species *IFIT1B* contributes to discrimination between "self versus non-self" transcripts based on the lack of 2' O-methylation on mRNA 5' caps in viruses, a so-called cap0 structure[34]. Notably, vervet *IFIT1B*, like that of mouse and gibbon, recognizes and inhibits replication of viruses with cap0-mRNAs, while human *IFIT1B* lacks this function[34]. It has been suggested that this functional divergence of *IFIT1B* antiviral activity reflects the divergence of the human lineage from that of other primates, in exposures and adaptations to particular sets of pathogens, including the arboviruses which are responsible for diseases such as encephalitis, dengue, and yellow fever.

Our results suggest that investigation of the genes regulated by *IFIT1B* in the vervet might help reveal mechanisms for its role in defense against viral pathogens. Recent evidence points to immune functions for the products of several of these genes. For example, *RANBP10*, a transcriptional coactivator, promotes viral gene expression and replication in HSV-1 infected cells[43]. *SUGT1*, a cell cycle regulator, is the homolog of *SGT1*, which plays an essential role in innate immunity in plants as well as mammals[44,45], while *TMEM57* has shown association in a human population, at a genome-wide significance threshold, to blood markers of inflammation[46].

Just as the human genetics field is increasingly employing GTEx data to refine the mapping information obtained by GWAS[5], we used the hippocampal eQTLs discovered in the vervet to identify a set of lncRNAs as candidate causal genes for a higher order phenotype, hippocampal volume. The exceptional genetic and environmental homogeneity of the relatively small vervet study sample likely facilitated these findings, and supports the extension of multi-tissue vervet eQTL studies as a strategy for identifying loci with a large impact on higher-order phenotypes, generally. While expanding expression resources in other NHP species will create additional opportunities to identify eQTLs that are informative for various biomedical investigations[8,47], the Caribbean vervet is unique among NHPs in having abundant natural populations available for such investigations, with an essentially identical genetic background to the samples studied here[9,11]. For example, the lead SNPs for the eQTLs contributing to hippocampal volume in the VRC each occur at a relatively high frequency in these island populations (Supplementary Information). We therefore anticipate that most, if not all, of the

findings presented here can be followed up through well-powered association studies in these populations.

Methods

Study Sample

The vervet monkeys used in this study are part of the Vervet Research Colony (VRC), established by UCLA during the 1970's and 1980's from 57 founder animals captured from wild populations in St. Kitts and Nevis[9]. In 2008 the VRC was moved to Wake Forest School of Medicine; the MRI phenotypes included in this study were collected when the colony was in California (see Supplementary Information for more details).

Gene Expression Phenotypes

Two data sets of gene expression measurements were collected. Dataset 1 consisted of microarray (Illumina HumanRef-8 v2) assays of RNA obtained from whole blood in 347 vervets, while Dataset 2 consisted of RNA-Seq data from 58 animals, with six tissues (whole blood, cultured fibroblast, caudate, hippocampus, adrenal, and pituitary) assayed in each animal.

Dataset 1: Microarrays From Whole Blood

The microarray data set has been described in Jasinska et al.[11] and is available at NCBI at the BioProject PRJNA115831. Details on RNA extraction, cDNA synthesis, and initial data processing are presented in Supplementary Information. To obtain a set of probes usable in vervet from the Illumina HumanRef-8 v2 microarray (originally developed for assaying gene expression in humans), we used the vervet reference sequence to select probes that contain no vervet indels and demonstrate $\leq$ five mismatches, with a maximum of one mismatch in the 16 nt central portion of the probe. To prevent bias in the measurement of expression due to SNP interference with hybridization, we excluded probes targeting sequences with common SNPs identified in the VRC pedigree. A total of 11,001 probes passed these filters (Supplementary Table 1). Illumina provides a "detection p-value" for each subject and probe; $p<0.05$ indicates significant detection of a given probe in a specific individual. We retained for analysis 6,018 probes that were detected with detection p-values of $p<0.05$ in at least 5% of monkeys, and tested for association 3,417 probes that were significantly heritable.

Dataset 2: RNA-Seq Data from Six Tissues

Tissues harvested during experimental necropsies were obtained from 60 monkeys representing 10 developmental stages, ranging from neonates (7 days), through infants (90 days and one year), young juveniles (1.25, 1.5, 1.75, 2 years old), subadults (2.5, 3 years old) to adults (4+ years old), with six monkeys (3 male and 3 female) from each developmental time point. Two monkeys (a 1.75 year old female

and a 7 day old male) for which we did not have WGS data were excluded from the eQTL study. Altogether, in the eQTL study we included 11 vervets below one year old, 23 vervets between one to two years old, and 24 vervets between two and four years old, 29 males and 29 females. Details regarding tissue collection and RNA collection procedures are in Supplementary Information.

We conducted RNA-Seq in six tissues: two brain tissues (caudate and hippocampus), two neuroendocrine tissues (adrenal and pituitary) and two peripheral tissues serving as a source of biomarkers (blood and fibroblasts). From purified RNA, we created two types of cDNA libraries, poly-A RNA (blood, fibroblasts, adrenal and pituitary) and total RNA (blood, caudate, hippocampus) cDNA libraries. Details on library preparation are in Supplementary Information. The RNA-Seq read data were made available through NCBI as BioProject PRJNA219198.

RNA-Seq reads were aligned to the vervet genomic assembly Chlorocebus_sabeus 1.1 http://www.ncbi.nlm.nih.gov/assembly/GCF_000409795.2 by the ultrafast STAR aligner[48] using our standardized pipeline. STAR was run using default parameters, which allow a maximum of ten mismatches. Gene expression was measured as total read counts per gene. For paired end experiments, total fragments are considered. Fragment counts that aligned to known exonic regions based on the NCBI *Chlorocebus sabaeus* Annotation Release 100 were quantified using the HTSeq package[49]. The counts for all 33,994 genes were then combined, and lowly expressed genes, defined as genes with a mean of < 1 across all samples, as well as genes detected in fewer than 10% of individuals were filtered out. Finally, quantile normalization was applied to the remaining genes to obtain normalized gene counts.

Quantitative real-time PCR (qRT-PCR) hippocampal expression data were generated by the following methods. 900ng of cDNA was generated from hippocampal mRNA using the SuperScript® III First-Strand Synthesis System (Life Technologies). Quantitative real-time PCR was performed in two steps using the SensiFAST™ SYBR® No-ROX Kit (Bioline) and the Roche LightCycler® 480 platform, with three technical replicates per assay, per animal. Specific primers (see Supplementary information for sequences) were designed using Primer3 software, and relative transcript abundance was calculated by the delta-delta Ct method, where results were first normalized to a housekeeping gene (glyceraldehyde 3-phosphate dehydrogenase; *GAPDH*).

Hippocampal Volume Phenotype

Estimates of hippocampal volume were measured in 347 vervets >2 years of age using MRI. Details of the image acquisition and processing protocol were described previously[36] and are outlined in Supplementary Information. Prior to genetic analysis, hippocampal volume was log transformed, regressed on sex and age using SOLAR[30], and residuals used as the final phenotype.

Genotype Data

Genotype data were generated through whole genome sequencing of 725 members of the VRC[13]. Genotypes from 721 VRC vervets that passed all QC procedures can be directly queried via the EVA at EBI (www.ebi.ac.uk/eva) using the PRJEB7923 accession number. Two genotype data sets were used in the current study[13]: (1) The Association Mapping SNP Set consists of 497,163 SNPs on the 29 vervet autosomes. In this set of ~500K SNPs, there were an average of 198 SNPs per Mb of vervet sequence, and the largest gap size between adjacent SNPs was 5 Kb. (2) The Linkage Mapping SNP Set consists of 147,967 markers on the 29 vervet autosomes. In this set of ~148K SNPs, there were an average of 58.2 SNPs per Mb of vervet sequence, and the average gap size between adjacent SNPs was 17.5 Kb.

The software package Loki[50], which implements Markov Chain Monte Carlo methods, was used to estimate the multipoint identical by decent (MIBD) allele-sharing among all vervet family members from the genotype data. As long stretches of IBD were evident among these very closely related animals, a reduced marker density was sufficient to evaluate MIBD at 1cM intervals; we used a 9,752 subset of the 148K SNP data set. The correspondence between physical and genetic positions in the vervet was facilitated by a vervet linkage map[51], constructed using a set of 360 STR markers. Both the physical and genetic position of these markers was known, and genetic locations of SNPs were found by interpolation.

Statistical Analysis

Principal Components Analysis

In Dataset 2, the top 1,000 most variable genes were selected for each tissue, and PCA applied to log2-transformed counts per million, using the singular value decomposition and the prcomp function in R (https://www.R-project.org, version 3.2.3). Expression was mean-centered prior to analysis.

Mapping of Gene Expression and Hippocampal Volume Phenotypes

We expected greater power for association analyses of gene expression traits compared to more complex phenotypes. Therefore we applied genome wide association analyses to these traits. For the higher-order phenotype examined (hippocampal volume) we anticipated having power only to detect loci with a much stronger effect, and therefore utilized linkage analysis for this trait.

Heritability and Multipoint Linkage Analysis We estimated familial aggregation (heritability) of traits using SOLAR, which implements a variance components method to estimate the proportion of phenotypic variance due to additive genetic factors (narrow sense heritability). This model partitions total variability into polygenic and environmental components. The environmental component is unique to individuals while the polygenic component is shared between individuals as a function of their pedigree kinship. If the variance in phenotype Y due to the polygenic component is designated as σ_g^2 and the environmental component as σ_e^2, then in this model $Var(Y) = \sigma_g^2 + \sigma_e^2$, and the covariance between phenotype values

of individuals *i* and *j* is $Cov(Y_i,Y_j)=2\ \varphi_{ij}\ \sigma_g^2$, where φ_{ij} is the kinship between individuals *i* and *j*.

Whole genome multipoint linkage analysis of hippocampal volume was also implemented in SOLAR, which uses a variance components approach to partition the genetic covariance between relatives for each trait into locus-specific heritability (h2q) and residual genetic heritability (h2r). Linkage analysis was performed at 1cM intervals using the likelihood ratio statistic.

Association Analysis Association between specific SNPs and gene expression phenotypes was evaluated using EMMAX[52]. EMMAX employs a linear mixed model approach, where SNP genotype is a fixed effect, and correlation of phenotype values among individuals is accounted for using an identity by state (IBS) approximation to kinship. Association analyses used the full set of 497,163 SNP markers, and included age and sex as covariates (both Dataset 1 and 2; where in Dataset 2 age, in days, corresponds to developmental stage), as well as batch (Dataset 1 only). It is common to try to account for unmeasured factors influencing global gene expression by including probabilistic estimation of expression residuals (PEER) factors as covariates[53]. We considered the controlled nature of the study environment and experimental design to preclude the need for this adjustment.

Multiple Testing Considerations in eQTL

We used a Bonferroni correction to account for multiple testing across genes, SNPs, and tissues as our primary error-controlling strategy for the identification of eQTL. Thresholds for Dataset 2 were more stringent, as more genes were tested than in Dataset 1 (~25K vs. ~3K) and multiple tissues were analyzed in Dataset 2. Dataset 1 was analyzed association to 3,417 heritable probes. The local eQTL significance threshold (4.8×10^{-8}) was corrected for the testing of SNPs within 1 Mb of 3,417 probes, and the distant eQTL significance threshold (1.5×10^{-11}) accounted for genome-wide testing of 3,417 probes. Dataset 2 significance thresholds were constructed in a similar fashion, but also accounted for testing of 163,770 gene-tissue combinations (the number of genes tested per tissue is in Table 1). The RNA-Seq local eQTL threshold was 7.6×10^{-10}, and the distant eQTL threshold was 6.1×10^{-13}.

To identify multi-tissue eGenes and the tissues in which they are active, and the associated SNPs in each of these tissues, we used a hierarchical approach based on Peterson et al. (2016)[31] which groups the hypotheses into a tree with three levels: genes in level 1, tissues in level 2, and SNPs in level 3. Testing proceeds sequentially starting from the top of the tree in a manner that accounts for each previous selection step. This method allows control of the FDR of local eGenes (defined as those genes whose expression is regulated in at least one tissue by some genetic variants located within 1 Mb of the gene) and of the expected average false discovery proportion of the tissues in which we claim this regulation is present across the discovered eGenes. P-values are defined by building up from the bottom

of the tree. Specifically, to obtain a p-value for the null hypothesis of no local regulation for a given gene in a given tissue (corresponding to a hypothesis in level 2 of the tree), we applied Simes' combination rule[54] to the p-values obtained via EMMAX for the hypotheses of no association between the expression of the gene in the tissue and each of the SNPs in the local neighborhood (corresponding to the hypotheses in level 3 of the tree). To obtain a p-value for the null hypothesis of no local regulation for a given gene in any of the tissues under study (corresponding to a hypothesis in level 1 of the tree), we applied Simes' combination rule to the gene x tissues p-values just described. We then tested the global null hypotheses of no local regulation in any tissue for all the genes in our study, applying the Benjamini Hochberg procedure[55] to control the FDR at the 0.05 level. For those genes for which we were able to reject the null hypotheses of no local regulation, we examined the tissue-specific p-values, applying the Benjamini Bogomolov procedure that allows the identification of significant findings controlling for the initial selection[56]. Finally, the individual SNPs responsible for regulation of the gene in each tissue were identified, again using a selection-adjusted threshold. An R package implementing this procedure is available at http://www.bioinformatics.org/treeqtl/.[57]

We compared the number of eGenes identified in each tissue using the above procedure with the results of GTEx (Analysis Release V6; dbGaP Accession phs000424.v6.p1), as presented on the GTEx portal web site. GTEx used a permutation strategy (described in the Analysis Methods section of the web portal) to identify eGenes.

Association between eQTLs and genomic features

We estimated the association of eQTLs to various genomic features (TSS/TES, functional genomic regions, human GWAS loci) by categorizing SNPs in two binary dimensions (eQTL and location in or near a specific genomic feature) and estimating the odds ratio from this contingency table. The significance of the association was evaluated using Fisher's Exact Test. To account for LD between SNPs (which violates the assumption of independence of SNP counts), we selected 22,892 genome-wide SNPs based on LD pruning the entire set of 497,163 SNPs at $r^2<0.6$ in 14 unrelated individuals. This SNP set included 1,663 eQTL SNPs (1,380 local eQTLs, 239 distant eQTLs, and 44 SNPs are both local and distant eQTLs).

References

1. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
2. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
3. Albert, F.W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**, 197-212 (2015).
4. Gilad, Y., Rifkin, S.A. & Pritchard, J.K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* **24**, 408-15 (2008).
5. Gibson, G., Powell, J.E. & Marigorta, U.M. Expression quantitative trait locus analysis for translational medicine. *Genome Med* **7**, 60 (2015).
6. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-9 (2011).
7. Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-5 (2015).
8. Rogers, J. & Gibbs, R.A. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat Rev Genet* **15**, 347-59 (2014).
9. Jasinska, A.J. *et al.* Systems biology of the vervet monkey. *ILAR J* **54**, 122-43 (2013).
10. Stein, J.L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet* **44**, 552-61 (2012).
11. Jasinska, A.J. *et al.* Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Hum Mol Genet* **18**, 4415-27 (2009).
12. Warren, W.C. *et al.* The genome of the vervet (Chlorocebus aethiops sabaeus). *Genome Res* **25**, 1921-33 (2015).
13. Huang, Y.S. *et al.* Sequencing strategies and characterization of 721 vervet monkey genomes for future genetic analyses of medically relevant traits. *BMC Biol* **13**, 41 (2015).
14. Arnett, M.G., Muglia, L.M., Laryea, G. & Muglia, L.J. Genetic Approaches to Hypothalamic-Pituitary-Adrenal Axis Regulation. *Neuropsychopharmacology* **41**, 245-60 (2016).
15. McEwen, B.S., Gray, J.D. & Nasca, C. 60 YEARS OF NEUROENDOCRINOLOGY: Redefining neuroendocrinology: stress, sex and cognitive and emotional regulation. *J Endocrinol* **226**, T67-83 (2015).
16. Bond, J. *et al.* ASPM is a major determinant of cerebral cortical size. *Nat Genet* **32**, 316-20 (2002).
17. Northcott, P.A. *et al.* Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49-56 (2012).
18. Bergoffen, J. *et al.* Connexin mutations in X-linked Charcot-Marie-Tooth disease. *Science* **262**, 2039-42 (1993).
19. Irobi, J. *et al.* Hot-spot residue in small heat-shock protein 22 causes distal motor neuropathy. *Nat Genet* **36**, 597-601 (2004).

20. Tang, B.S. *et al.* Small heat-shock protein 22 mutated in autosomal dominant Charcot-Marie-Tooth disease type 2L. *Hum Genet* **116**, 222-4 (2005).
21. Sargiannidou, I. *et al.* Connexin32 mutations cause loss of function in Schwann cells and oligodendrocytes leading to PNS and CNS myelination defects. *J Neurosci* **29**, 4736-49 (2009).
22. Insel, T.R. The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. *Neuron* **65**, 768-79 (2010).
23. Young, L.J. & Hammock, E.A. On switches and knobs, microsatellites and monogamy. *Trends Genet* **23**, 209-12 (2007).
24. Okhovat, M., Berrio, A., Wallace, G., Ophir, A.G. & Phelps, S.M. Sexual fidelity trade-offs promote regulatory variation in the prairie vole brain. *Science* **350**, 1371-4 (2015).
25. Weisskopf, M.G., Zalutsky, R.A. & Nicoll, R.A. The opioid peptide dynorphin mediates heterosynaptic depression of hippocampal mossy fibre synapses and modulates long-term potentiation. *Nature* **362**, 423-7 (1993).
26. Harte-Hargrove, L.C., Varga-Wesson, A., Duffy, A.M., Milner, T.A. & Scharfman, H.E. Opioid receptor-dependent sex differences in synaptic plasticity in the hippocampal mossy fiber pathway of the adult rat. *J Neurosci* **35**, 1723-38 (2015).
27. Simavli, S. *et al.* Substance p regulates puberty onset and fertility in the female mouse. *Endocrinology* **156**, 2313-22 (2015).
28. Tamba, S. *et al.* Timely interaction between prostaglandin and chemokine signaling is a prerequisite for successful fertilization. *Proc Natl Acad Sci U S A* **105**, 14539-44 (2008).
29. Egli, M., Leeners, B. & Kruger, T.H. Prolactin secretion patterns: basic mechanisms and clinical implications for reproduction. *Reproduction* **140**, 643-54 (2010).
30. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**, 1198-211 (1998).
31. Peterson, C.B., Bogomolov, M., Benjamini, Y. & Sabatti, C. Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. *Genet Epidemiol* **40**, 45-56 (2016).
32. Tung, J., Zhou, X., Alberts, S.C., Stephens, M. & Gilad, Y. The genetic architecture of gene expression levels in wild baboons. *Elife* **4**(2015).
33. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-66 (2015).
34. Daugherty, M.D., Schaller, A.M., Geballe, A.P. & Malik, H.S. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. *Elife* **5**(2016).
35. Pierce, B.L. *et al.* Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet* **10**, e1004818 (2014).
36. Fears, S.C. *et al.* Identifying heritable brain phenotypes in an extended pedigree of vervet monkeys. *J Neurosci* **29**, 2867-75 (2009).
37. Mattick, J.S. & Rinn, J.L. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* **22**, 5-7 (2015).

38. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26-46 (2013).
39. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
40. Wang, J. *et al.* Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *Am J Hum Genet* **98**, 697-708 (2016).
41. Pichlmair, A. *et al.* IFIT1 is an antiviral protein that recognizes 5'-triphosphate RNA. *Nat Immunol* **12**, 624-30 (2011).
42. Brodziak, F., Meharg, C., Blaut, M. & Loh, G. Differences in mucosal gene expression in the colon of two inbred mouse strains after colonization with commensal gut bacteria. *PLoS One* **8**, e72317 (2013).
43. Sato, Y. *et al.* Cellular Transcriptional Coactivator RanBP10 and Herpes Simplex Virus 1 ICP0 Interact and Synergistically Promote Viral Gene Expression and Replication. *J Virol* **90**, 3173-86 (2016).
44. Azevedo, C. *et al.* The RAR1 interactor SGT1, an essential component of R gene-triggered disease resistance. *Science* **295**, 2073-6 (2002).
45. Mayor, A., Martinon, F., De Smedt, T., Petrilli, V. & Tschopp, J. A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. *Nat Immunol* **8**, 497-503 (2007).
46. Naitza, S. *et al.* A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS Genet* **8**, e1002480 (2012).
47. Bakken, T.E. *et al.* A comprehensive transcriptional map of primate brain development. *Nature* **535**, 367-75 (2016).
48. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
49. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-9 (2015).
50. Heath, S.C., Snow, G.L., Thompson, E.A., Tseng, C. & Wijsman, E.M. MCMC segregation and linkage analysis. *Genet Epidemiol* **14**, 1011-6 (1997).
51. Jasinska, A.J. *et al.* A genetic linkage map of the vervet monkey (Chlorocebus aethiops sabaeus). *Mamm Genome* **18**, 347-60 (2007).
52. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
53. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-7 (2012).
54. Simes, R.J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751-754 (1986).
55. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
56. Benjamini, Y. & Bogomolov, M. Selective inference on multiple families of hypotheses. *J. R. Stat. Soc. B* **76**, 297-318 (2014).

57. Peterson, C.B., Bogomolov, M., Benjamini, Y. & Sabatti, C. TreeQTL: hierarchical error control for eQTL findings. *Bioinformatics* **32**, 2556-8 (2016).

Figure Legends

Figure 1. PCA of 1,000 genes, for each tissue, with the most variable expression levels. Numbers in the labels for x and y axes indicate the proportion of total variance accounted for by that PC.

Figure 2. Boxplot of log counts per million (CPM) expression in Caudate vs. timepoint, for four genes with a strong relationship between expression pattern and age.

Figure 3. Master regulatory locus on vervet chromosome CAE 9. Upper panel: Ensemble view of the CAE 9 region. Lower panel: The minimum –log10(p-value) for each SNP in association analyses vs. expression of microarray probes on different chromosomes. The symbols are color-coded to represent the number of probes significantly associated to each SNP: 1-2 probes (black), 3-4 probes (yellow), 5-6 probes (blue), 7-10 probes (green), 11-14 probes (red).

Figure 4. Hippocampal volume QTL and local hippocampal volume eQTL in RNA-Seq analysis. Top panel: purple dotted line is the multipoint LOD score for hippocampal volume. Circles represent evident for association of SNPs to expression of *LOC103222765* (red), *LOC103222769* (blue) and *LOC103222771* (green). Solid circles indicate genome-wide significant associations. The region between the black vertical lines is blown up in the middle and bottom panels. The horizontal dotted line represents the genome-wide significant threshold for local eQTL. Middle panel: SNPs with –log10(p-value)>8 for association to expression in hippocampus, color codes are as in the top panel. Bottom panel: Genes sited between 68.7 and 69 Mb (the eQTL region). Color codes are as in the top panel.

Figure 5. Correlation of hippocampal volume (MRI) with hippocampal expression of *LOC103222765* (left), *LOC103222769* (middle) and *LOC103222771* (right). The expression data are from qRT-PCR. Hippocampal volume measurements are residuals from a regression on covariates of age and sex.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements. Thanks to Stephanie Groman for assistance with tissue resources. Thanks to Tara Chavanne, Kelsey Finnie, Margaret Long, Jean Gardin and Dianna Swaim for technical assistance with necropsies and tissue collections. This work was supported by the following grants, all from the U.S. National Institutes of Health: U54HG00307907 (to RKW); P40RR019963/OD010965 (to JRK); R01RR016300/OD010980 (to NBF); R37MH060233 (to Daniel Geschwind);

UL1DE019580 (to Robert Bilder); PL1NS062410 (to Christopher Evans); RL1MH083270 (to JDJ); P30NS062691 (to NBF and GC); R01MH101782 (to CS and EE). RN, BLA, PF acknowledge support from Wellcome Trust grant numbers WT095908 and WT098051 and the European Molecular Biology Laboratory. ESW was supported by an EMBO Advanced Fellowship (aALTF1672-2014).

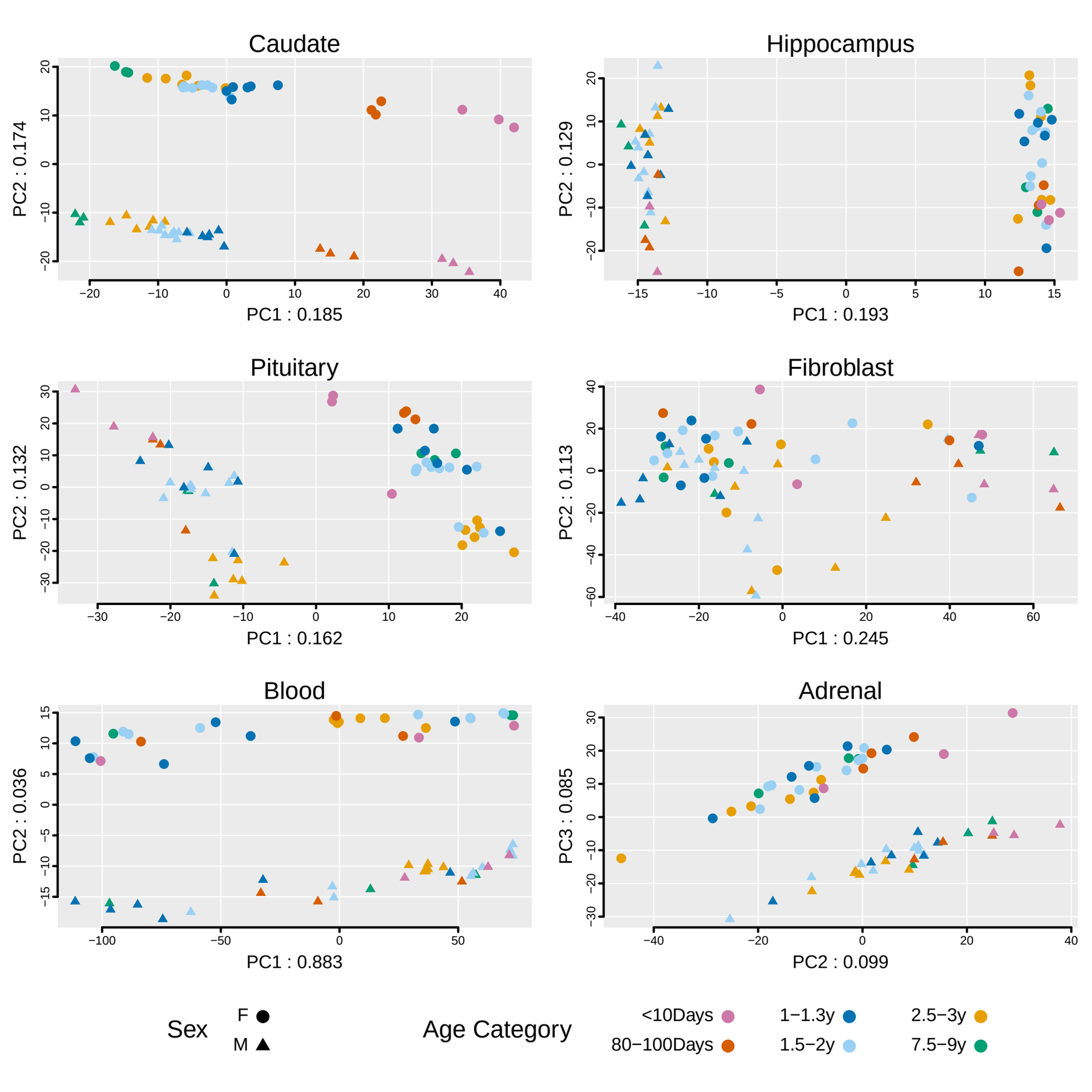
Author contributions: AJJ, JRK, GMW, KD, RKW, JDJ, WW, RPW, and NBF designed the study. AJJ, IZ, OWC, JD, LAF, SF, AEF, YSH, VR, CAS, JDJ, GC, and RPW produced the data. AJJ, IZ, SKS, CP, RMC, EE, LAF, SF, YSH, VR, CAS, HS, DV, BLA, PF, RN, ESW, JB, TDD, MB, YB, CS, and GC analyzed the data. OWC, JD, and MJJ managed data and samples. AJJ, SKS, and NBF wrote the paper. All authors reviewed the final draft.

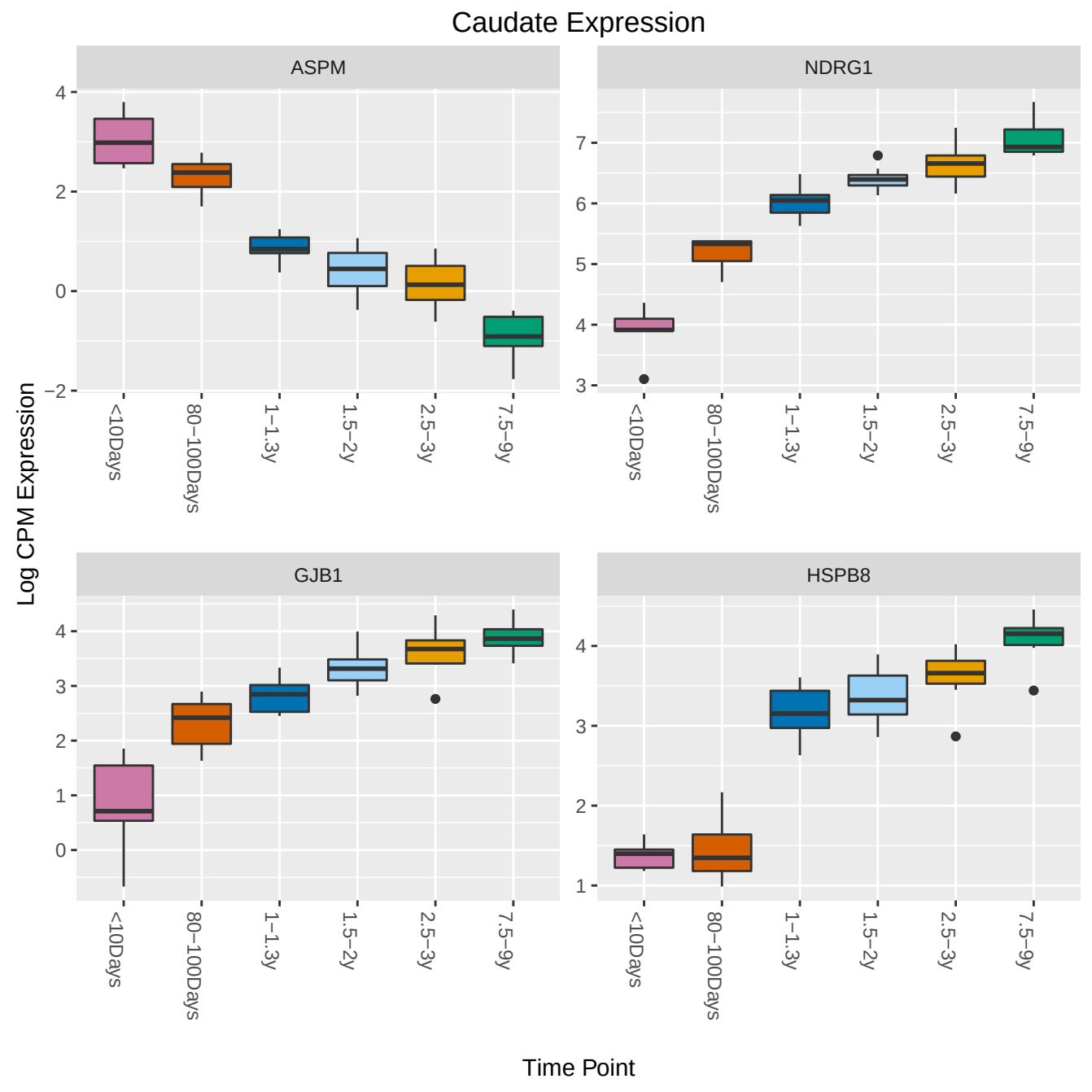
Author Information:

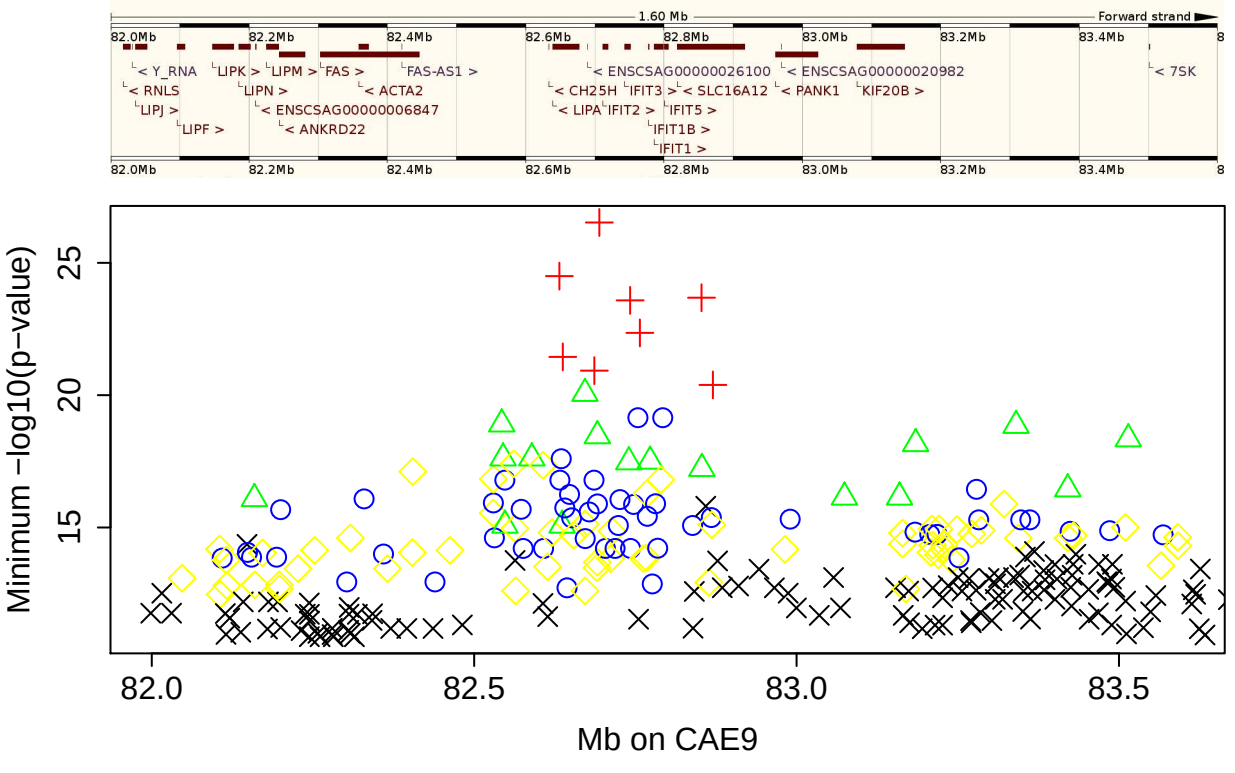
The microarray data set is available at NCBI at the BioProject PRJNA115831. The RNA-Seq read data were made available through NCBI as BioProject PRJNA219198. Genotypes from 721 VRC vervets that passed all QC procedures are publically available at EMBL-EBI (https://www.ebi.ac.uk/ena/data/view/ERP008917); these data can be directly queried via the EVA at EBI (www.ebi.ac.uk/eva).

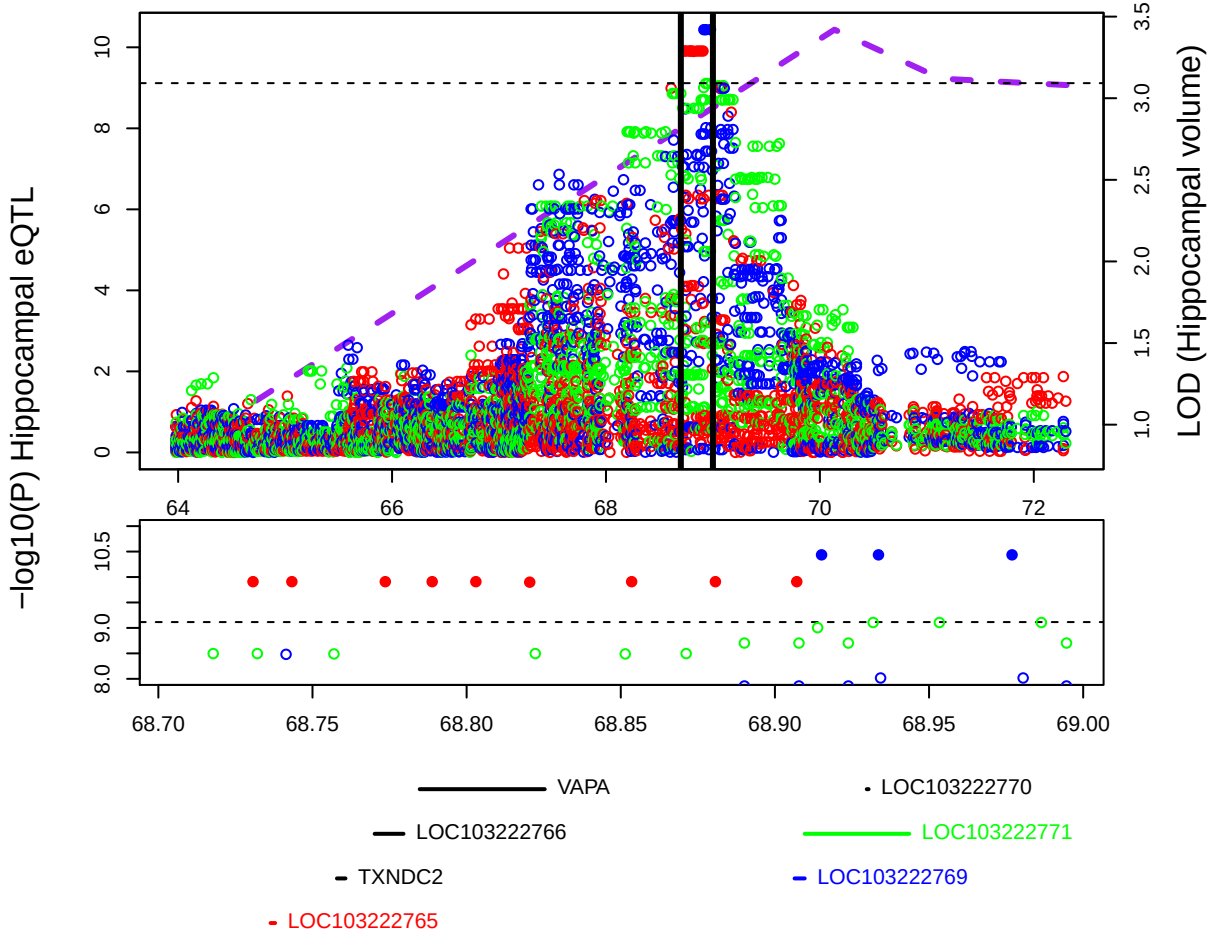
The authors declare that they have no competing financial interests

Correspondence and requests for materials should be addressed to
nfreimer@mednet.ucla.edu


Caudate
PC2 : 0.174
PC1 : 0.185
Hippocampus
PC2 : 0.129
PC1 : 0.193
Pituitary
PC2 : 0.132
PC1 : 0.162
Fibroblast
PC2 : 0.113
PC1 : 0.245
Blood
PC2 : 0.036
PC1 : 0.883
Adrenal
PC3 : 0.085
PC2 : 0.099
Sex
F
M
Age Category
<10Days
80−100Days
1−1.3y
1.5−2y
2.5−3y
7.5−9y


Caudate Expression
ASPM
NDRG1
GJB1
HSPB8
Log CPM Expression
Time Point
<10Days
80–100Days
1–1.3y
1.5–2y
2.5–3y
7.5–9y


1.60 Mb
Forward strand
82.0Mb
82.2Mb
82.4Mb
82.6Mb
82.8Mb
83.0Mb
83.2Mb
83.4Mb
< Y_RNA
LIPK >
LIPM >
FAS >
FAS-AS1 >
< RNLS
LIPN >
< ACTA2
LIPJ >
< ENSCSAG00000006847
LIPF >
< ANKRD22
< ENSCSAG00000026100
< ENSCSAG00000020982
< 7SK
< CH25H
IFIT3 >
< SLC16A12
< PANK1
KIF20B >
< LIPA
IFIT2 >
IFIT5 >
IFIT1B >
IFIT1 >
Minimum −log10(p−value)
15
20
25
82.0
82.5
83.0
83.5
Mb on CAE9


−log10(P) Hippocampal eQTL
LOD (Hippocampal volume)
0
2
4
6
8
10
1.0
1.5
2.0
2.5
3.0
3.5
64
66
68
70
72
8.0
9.0
10.5
68.70
68.75
68.80
68.85
68.90
68.95
69.00
VAPA
LOC103222766
TXNDC2
LOC103222765
LOC103222770
LOC103222771
LOC103222769

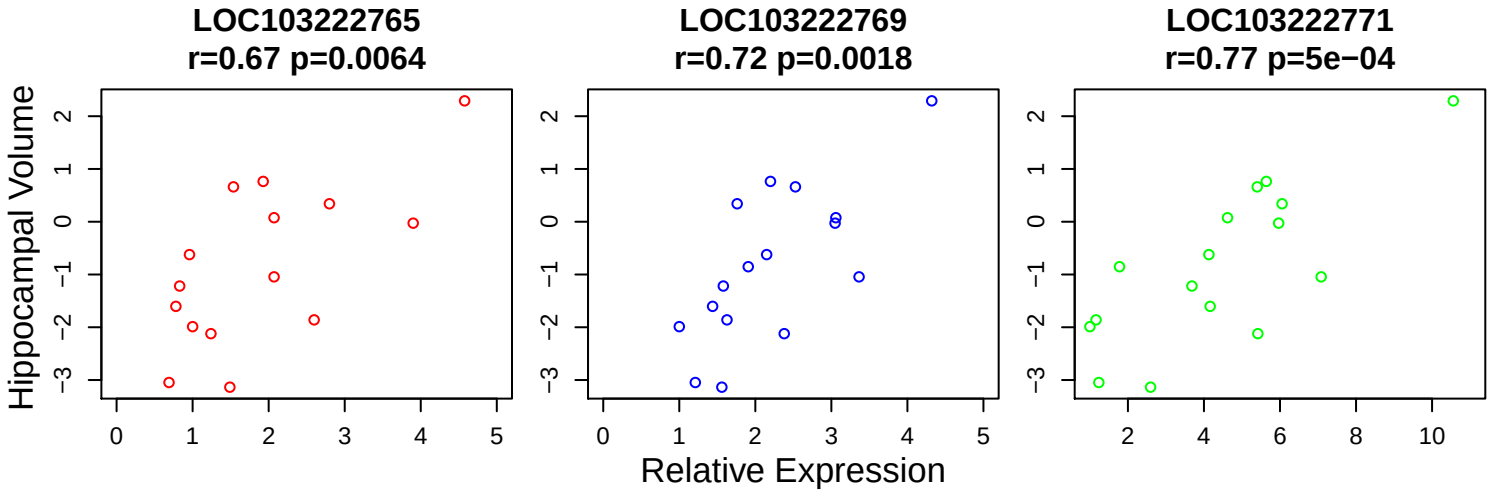

LOC103222765
r=0.67 p=0.0064
LOC103222769
r=0.72 p=0.0018
LOC103222771
r=0.77 p=5e−04
Hippocampal Volume
Relative Expression

Table 1. Gene expression data sets. The number of probes/genes with at least one significant local and distant eQTL (at Bonferroni corrected thresholds) are presented.

Tissue	Probes/genes analyzed[a]	Local eQTL[b]	Distant eQTL[c]
Dataset 1: Microarray			
Blood	3,417	461	215
Dataset 2: RNA-seq			
Adrenal Cortex	25,187	506	60
Pituitary Gland	27,236	537	71
Caudate	28,282	386	39
Hippocampus	26,957	338	47
Fibroblast	22,328	196	23
Blood	33,780	70	4

[a]microarray dataset (Dataset 1) with an initial set of 22,184 probes on Illumina HumanRef-8 v2 (6,018 probes passed filters described in Supplementary Table 1, 3,417 were heritable); RNA-seq (Dataset 2) with an initial set of 33,994 genes annotated in vervet

[b]Bonferroni threshold for Dataset 1: 4.8×10^{-8}; Bonferroni threshold for Dataset 2: 7.6×10^{-10}

[c]Bonferroni threshold for Dataset 1:1.5×10^{-11}; Bonferroni threshold for Dataset 2: 6.1×10^{-13}

Table 2. Comparison of the number of local eGenes by tissue in our study and GTEx

Tissue	GTEx number of individuals	GTEx number of eGenes	vervet number of individuals	vervet number of eGenes
Adrenal Cortex	126	3259	58	2759
Caudate	100	2447	58	2863
Hippocampus	81	1134	58	2350
Pituitary Gland	87	2160	58	3169
Blood	338	6784	58	609