Babbitt et al. BMC Genomics (2017) 18:435
DOI 10.1186/s12864-017-3831-2

BMC Genomics

RESEARCH ARTICLE Open Access

Gene expression and adaptive noncoding changes during human evolution

Courtney C. Babbitt[1,2,5*], Ralph Haygood[3], William J. Nielsen[1] and Gregory A. Wray[1,2,4]

Abstract

Background: Despite evidence for adaptive changes in both gene expression and non-protein-coding, putatively regulatory regions of the genome during human evolution, the relationship between gene expression and adaptive changes in *cis*-regulatory regions remains unclear.

Results: Here we present new measurements of gene expression in five tissues of humans and chimpanzees, and use them to assess this relationship. We then compare our results with previous studies of adaptive noncoding changes, analyzing correlations at the level of gene ontology groups, in order to gain statistical power to detect correlations.

Conclusions: Consistent with previous studies, we find little correlation between gene expression and adaptive noncoding changes at the level of individual genes; however, we do find significant correlations at the level of biological function ontology groups. The types of function include processes regulated by specific transcription factors, responses to genetic or chemical perturbations, and differentiation of cell types within the immune system. Among functional categories co-enriched with both differential expression and noncoding adaptation, prominent themes include cancer, particularly epithelial cancers, and neural development and function.

Keywords: Adaptation, Gene expression, Gene function, Gene regulation, Human evolution

Background

The evolutionary mechanisms responsible for divergence in gene expression between species are poorly understood. To begin with, expression level is for most genes a high-dimensional phenotype: almost without exception it differs among cell types, across the life cycle, and in response to numerous environmental factors [1, 2]. This makes it challenging to link positive selection on regulatory sequences to any particular aspect of a gene's expression. In addition, populations often harbor significant levels of genetic variation that influences gene expression [3], confounding attempts to distinguish between-species divergence from within-species variation. Finally, the full complement of *cis*-regulatory elements is rarely known, constraining attempts to carry out comprehensive scans for natural selection. Perhaps unsurprisingly, it has proven difficult to detect a clear relationship at a genomic scale between the distribution of positive selection on noncoding sequences and divergence in gene expression, particularly in multicellular organisms [4].

The question is whether this result is a true or false negative. One way to move beyond a quest for simple correlations is to carry out joint analysis of genes that contribute to related phenotypes. During human origins, for instance, many of the same functional categories of genes show enrichments for signatures of positive selection [5–8] and for changes in tissue-specific expression level [9–14], even though on a gene-by-gene basis no correlation is evident. This overlap in enrichments hints at cause-and-effect or common-effect relationships. For example, two genes whose products are part of the same biological process might both experience positive selection to change expression levels because they both alter the same quantitative organismal trait in the same direction. The clearest examples come from metabolic pathways [15–17], but in principle this relationship could apply to any set of genes that contribute to the same trait or set of traits. Working with functionally related sets of genes rather than single genes should, in

* Correspondence: cbabbitt@bio.umass.edu
[1]Department of Biology, Duke University, Durham, NC 27708, USA
[2]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA
Full list of author information is available at the end of the article

principle, improve our ability to detect relationships between signatures of positive selection and gene expression change.

To test this concept, we measured transcript abundance in five tissues in humans and chimpanzees, and then used the results to assess the relationship between gene expression and positive selection. We chose tissues informative to understanding the evolution of metabolism, namely adipose tissue, cerebellum, cortex, liver, and skeletal muscle. These tissues provide an opportunity to explore the "expensive tissue" hypothesis that a major shift in energy allocation among tissues occurred during human evolution in order to support the remarkable expansion of the metabolically expensive human brain [14, 18–21]. We analyzed our measurements of transcript abundance from all tissues together in a single statistical model so as to maximize our power to detect expression differences. We then looked at correlations between differential gene expression and the results from three noncoding DNA datasets. These included a combination of human accelerated regions (where there are significantly more changes on the human branch in a highly-conserved regions in other organisms, anywhere in the genome) [7, 8]; as well as rapidly evolving putative *cis*-regulatory regions upstream of coding sequences (as compared to changes in local introns) [6]. As expected, we find little correlation between gene expression changes and adaptive noncoding (putative *cis*-regulatory) changes at the level of individual genes; however, we did find appreciable correlations between the two for several informative kinds of biological functions. Our results demonstrate the utility of considering functional categories when studying the evolution of gene expression and provide novel insights into the genetic basis for human origins.

Results

Measurements of gene expression

We used RNA-Seq to measure expression of 14,341 genic regions in at least one of white adipose tissue, lateral cerebellum, frontal cortex, liver, and skeletal muscle sampled from four male humans and four male chimpanzees (see Additional file 1: Table S1 for details). Using a multi-factor Generalized Linear Model (GLM) to analyze our measurements of all tissues together [22], we identified 4828 genes whose expression differs between humans and chimpanzees (q-value <0.05, 5% FDR). This proportion, 34% (4828 of 14,341), is greater than in previous studies that analyzed different tissues in separate models [9, 13, 23, 24]. GO analyses of overall differential expression show significant expression differences in categories related to cell signaling, neural processes, ion transport and development (Fig. 1). In keeping with many previous studies [25–27], cluster analysis revealed a greater overall similarity in expression by tissue than by species; the tissue term in our GLM on average explained a greater fraction of differential expression (Kolmogorov-Smirnov test, D = 0.16069, p-value <2.2e-16) than did the species term (Fig. 1).

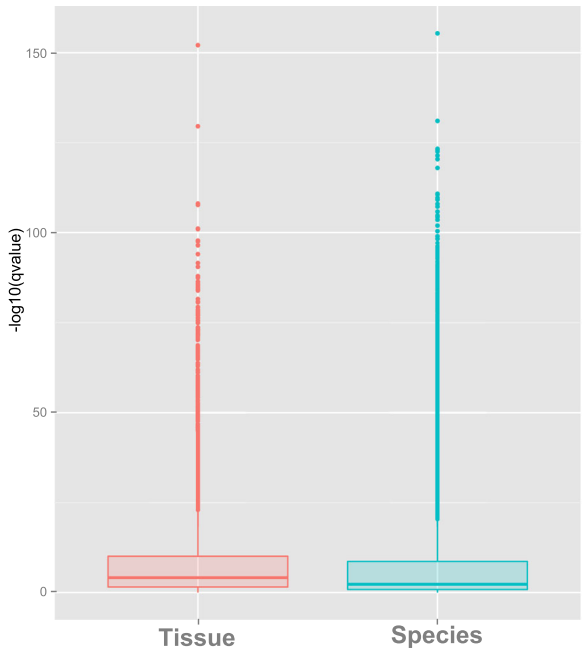

Fig. 1 Boxplot of the –log10(qvalue) distributions from the GLM examining the Tissue (left, orange) and Species (right, green) effects

Changes in expression divergence and tissue specificity

We then examined the trends of genes that have diverged in expression levels between humans and chimpanzees in individual tissues. We calculated a specificity score for differential expression (human vs. chimpanzee) in a given tissue relative to the rest of our gene expression data [as in 6, 28]. This score is not based on the magnitude of expression levels, but rather on the distribution of expression over tissues. A gene that is mainly differentially expressed in one tissue (even if at low numbers) between humans and chimpanzees will have a specificity score closer to one, and genes not showing changes in expression will have scores closer to 0. We found a strong correlation of tissue specificity and changing gene expression between humans and chimpanzees in all five tissues (adipose: rho = 0.962,liver: rho = 0.828, cerebellum: rho = 0.708, rho = cerebral cortex: 0.833, rho = skeletal muscle: 0.757; all 5 p-values <2.2e-16) (example tissue shown in Fig. 2). These results suggest that the more tissue specific a gene's expression, the more likely its expression is to change over evolutionary time, even during the short divergence time between the human and chimpanzee lineages. In contrast, the categories of genes we found enriched for similar expression levels

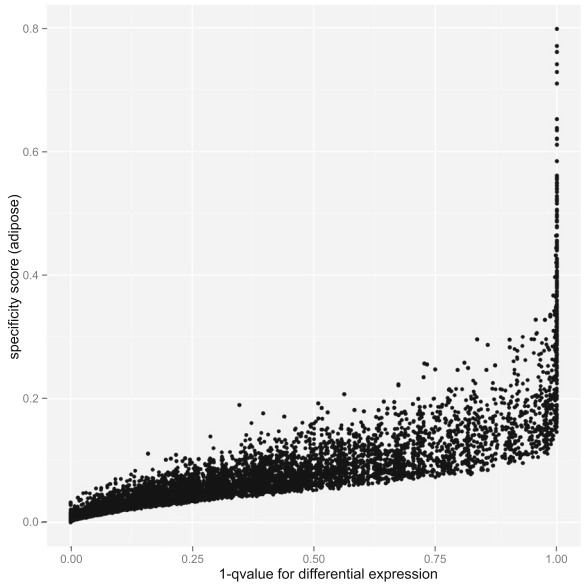

Fig. 2 Correlation between tissue specificity and gene expression divergence between humans and chimpanzees. There is a significant correlation between higher specificity to a tissue and expression divergence (adipose tissue as an example here)

across the tissues we measured here are mainly comprised of housekeeping processes, such as "transcription regulation" and "DNA binding". This is consistent with the idea that tissue-specific selection pressures are more easily accommodated by changes in cis-regulation than by changes in protein structure, because the latter are more likely to have deleterious pleiotropic effects [2, 29].

Functions of differentially expressed genes across humans and non-human primates

Considering just the large (>50 genes) Gene Ontology categories (Fig. 3) we see themes of metabolism, signaling, development, and nervous system as differentially expressed between humans and chimpanzee (Fig. 3). These results are intriguing, but only represent biological processes functioning in normal tissue.

To further understand the implications of these expression differences, we analyzed their distribution over multiple gene sets using the Molecular Signatures Database (MSigDB) [30, 31]; see "Methods" for details). MSigDB consists of 10,348 sets grouped into 20 collections representing broad aspects of gene function and organization. The collections form a shallow, partially hierarchical, tree structure. One of the challenges in interrogating these large gene ontology sets is the redundancy in, and heterogeneous size of, these gene set ontologies [32]. For example, the C3:All collection is the union of the C3:MIR and C3:TFT collections with no deeper branching. Therefore, we considered only 16 leaf (of the tree structure) collections (e.g., C3:MIR and C3:TFT but not C3:All). In order to emphasize broad trends among changes in gene expression between humans and chimpanzees, we restricted attention to sets containing at least 50 genes whose expression we measured.

Table 1 lists the 20 (of 5247) MSigDB gene sets most enriched with genes scoring high for overall differential expression between humans and chimpanzees (see Additional file 2: Tables S2a–S2f for full results, both overall and per tissue). One of the most unexpected and striking patterns is the clustering of statistically differentially expressed genes at specific cytogenetic bands (Fig. 4, Table 1). Some of these regions have interesting evolutionary histories in primates, and have important links to human disease. There is a common inversion in chromosome 17q11.31 that segregates within human populations [33] and is correlated with a number of neurological conditions, such as Parkinson's [34] and intellectual disabilities [35] and intercranial volume [36]. This region is also bounded by a segmental duplication that seems to have been under positive selection in the primate lineage [37], and contains genes such as MAPT and BRCA1. The 17q11 region is associated with a number of immune conditions, such as tuberculosis and leprosy in human populations [38], and has an expanded repertoire of cytokines [39], while the 12q13 region is associated with Type I diabetes [40]. Additionally, the 17q11 and 12q13 cytogenetic bands have gene clusters that have expanded during primate evolution [41]. There is previous evidence of positive selection in regions with duplications, both in coding [42] and in regulatory sequence [37], which we discuss below. Additionally, Table 1 shows that differentially expressed genes between humans and chimpanzee across all five tissues are enriched for processes related to neural signaling [43, 44] and pathways in cancer [45–47], based on the data from those studies.

We then went on to examine whether the differentially expressed genes between humans and chimpanzees have roles in human disease or dysregulation. The KEGG pathways that show enrichments are clearly related to two general groups of processes: neuronal function and cancer (Table 1, Additional file 3: Table S3). There are a number of categories related to "neuronal signaling." For processes related to cancer, the signal appears to be coming from differential gene expression in processes related to cell signaling and adhesion (e.g. "Focal adhesion", "Gap Junction", and "ECM Receptor Interaction"). This is suggestive that healthy differential gene expression between humans and chimpanzees overlaps with gene also involved in some late-onset disease susceptibilities.

Genic correlations between differential expression and noncoding adaptation

Four thousand five hundred eighty-one genes whose expression we measured were also analyzed in our

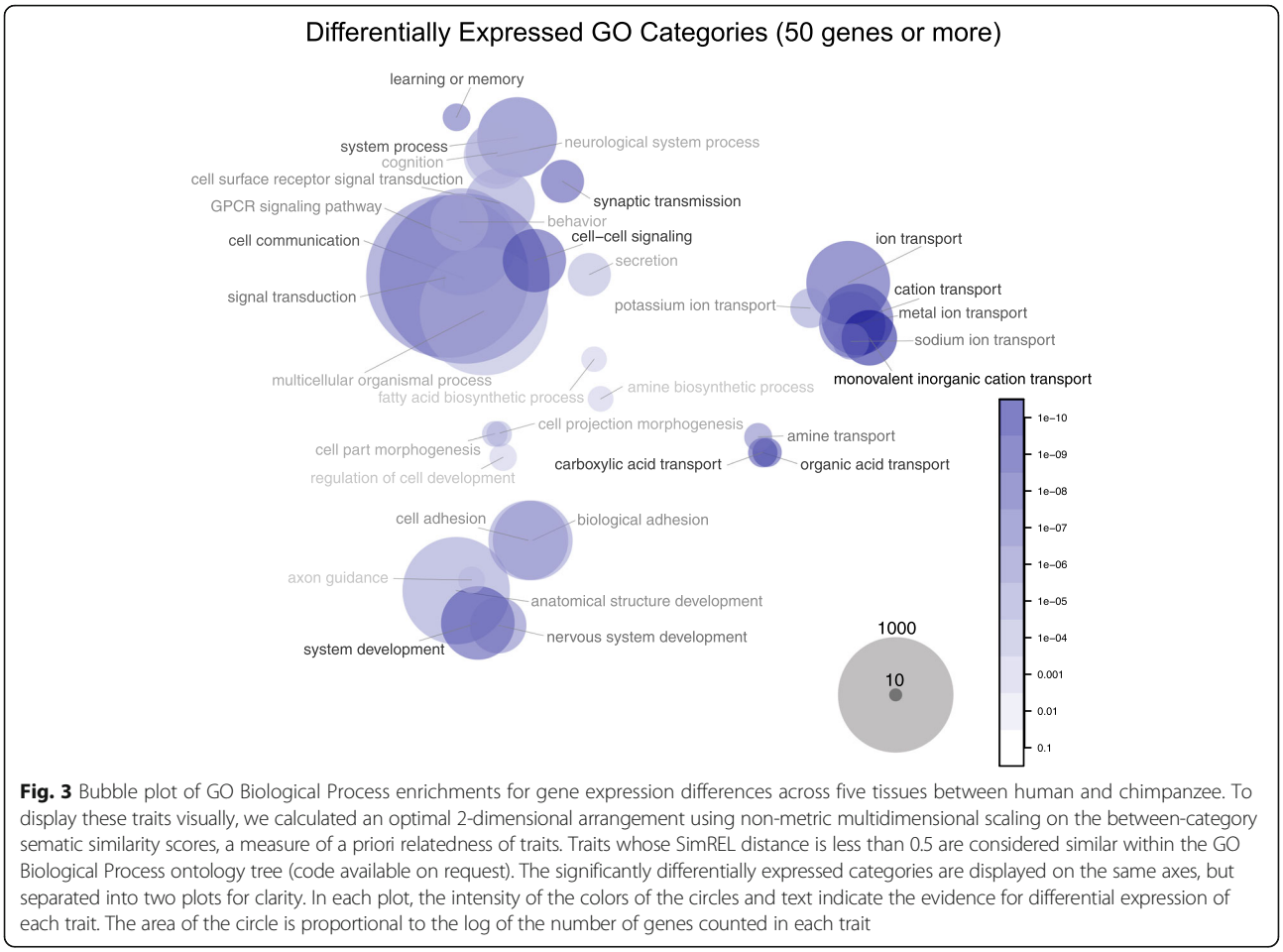

Fig. 3 Bubble plot of GO Biological Process enrichments for gene expression differences across five tissues between human and chimpanzee. To display these traits visually, we calculated an optimal 2-dimensional arrangement using non-metric multidimensional scaling on the between-category sematic similarity scores, a measure of a priori relatedness of traits. Traits whose SimREL distance is less than 0.5 are considered similar within the GO Biological Process ontology tree (code available on request). The significantly differentially expressed categories are displayed on the same axes, but separated into two plots for clarity. In each plot, the intensity of the colors of the circles and text indicate the evidence for differential expression of each trait. The area of the circle is proportional to the log of the number of genes counted in each trait

previous study of noncoding (putative *cis*-regulatory) regions immediately upstream from transcription start sites [6]. Consistent with that study, the correlation between our current scores for overall differential expression between humans and chimpanzees (q-value <0.05) and our previous scores for adaptive sequence changes during human evolution is negligible (rank correlation $r_r = 0.0074$, one-tailed $p = 0.32$ via permutation test with 10,000 permutations). The tissue with the highest correlation between scores for per-tissue differential expression and scores for adaptive sequence changes is cerebral cortex, but the correlation is weak ($r_r = 0.031$, $p = 0.032$). Similar analyses using scores for adaptive sequence changes from two other studies of noncoding regions [7, 8] yield similar results (see Additional file 4: Table S4, with total, and by tissue, tabs for full results). Thus, per gene, the relationship between gene expression and adaptive noncoding changes is apparently weak.

Beyond the possibility of noise or errors in the scores, this finding is unsurprising for several reasons explained above. We measured expression in only a few tissues from healthy adults, giving us a limited view of this high-dimensional expression phenotype. For these and other reasons, we had not expected to see a strong relationship between gene expression and adaptive noncoding changes at the level of individual genes.

There are correlations between differential expression and adaptation at the level of ontology categories

However, we hypothesized that there might be a stronger relationship at the level of biological function than at the level of individual genes. Conceivably, both genes whose expression differs between humans and chimpanzees and genes near noncoding regions whose sequences changed adaptively during human evolution might tend to affect the same biological functions. Of course, biological function is an extremely broad notion. As the diversity of MSigDB collections attests, genes can be categorized in many different ways, from their molecular characteristics through their contributions to normal development and physiology to their involvement in pathologies such as cancer. Accordingly, our question amounts to whether there is some type of biological function with respect to which there is a stronger relationship between gene

Table 1 Functions of overall differentially expressed genes

Set	Collection	# genes	r_rb	SE(r_rb)
chr17q11	C1:All	66	0.88	0.021
chr12q13	C1:All	152	0.84	0.026
NIKOLSKY BREAST CANCER 17Q11 Q21 AMPLICON	C2:CGP	80	0.84	0.033
chr17q21	C1:All	164	0.69	0.033
chr4q21	C1:All	61	0.62	0.066
chr1p13	C1:All	79	0.59	0.06
GNF2 DNM1	C4:CGN	63	0.41	0.046
RICKMAN HEAD AND NECK CANCER A	C2:CGP	77	0.34	0.06
LEIN NEURON MARKERS	C2:CGP	50	0.34	0.052
CAHOY NEURONAL	C6:All	84	0.33	0.053
GCM MAP1B	C4:CGN	53	0.33	0.076
ANASTASSIOU CANCER MESENCHYMAL TRANSITION SIGNATURE	C2:CGP	53	0.33	0.069
SABATES COLORECTAL ADENOMA UP	C2:CGP	86	0.33	0.049
GNF2 CCNA2	C4:CGN	52	0.32	0.065
KRAS.KIDNEY UP.V1 UP	C6:All	113	0.32	0.047
VOLTAGE GATED CATION CHANNEL ACTIVITY	C5:MF	57	0.31	0.061
NAKAYAMA SOFT TISSUE TUMORS PCA2 UP	C2:CGP	71	0.31	0.053
GCM MAPK10	C4:CGN	71	0.31	0.059
CERVERA SDHB TARGETS 1 UP	C2:CGP	95	0.3	0.052
VOLTAGE GATED CHANNEL ACTIVITY	C5:MF	63	0.3	0.058

The 20 (of 5247) MSigDB gene sets most enriched with genes scoring high for overall differential expression between humans and chimpanzees are listed. Collection is the MSigDB collection containing the set (one of 16). genes is the number of genes whose expression we measured in the set (at least 50). r_{rb} is the rank-biserial correlation between scores for differential expression and membership in the set; sets are ordered by decreasing r_{rb}. SE(r_{rb}) is the standard error of r_{rb} via bootstrapping with 10,000 replicates

expression and adaptive noncoding changes. We addressed this question using the commonly accessed MSigDB gene ontology collections [30, 31].

For each MSigDB gene set delineated above and for each previous study of noncoding regions mentioned above [6–8], we computed a rank-biserial correlation (see Materials and Methods) of the kind presented in Table 1 for enrichment with genes scoring high for adaptive sequence changes in putatively regulatory regions of the human genome. As in our previous meta-analysis of such studies [5], we combined these rank-biserial correlations across the studies to obtain one rank-biserial correlation for each set, restricting attention to sets represented in at least two of the studies, with correlations not significantly heterogeneous across these studies, and containing at least 50 genes analyzed

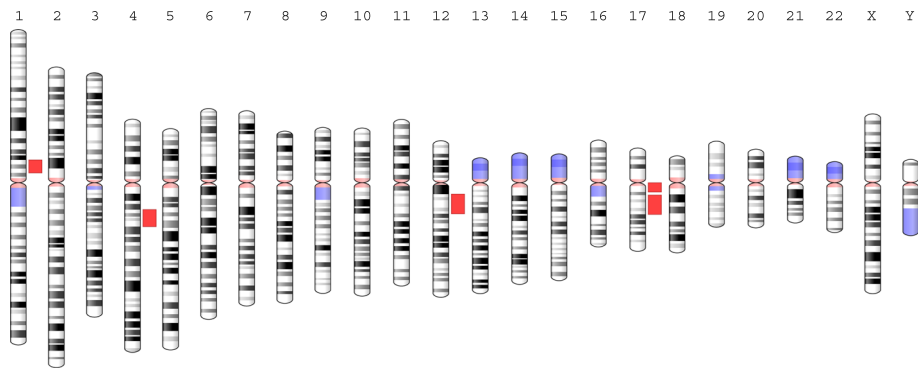

Fig. 4 Ideogram of clustering of significant differentially expressed genes in humans compared to chimpanzees across tissues (red bars highlight the chromosomal enrichment regions in Table 1)

on the average over these studies (see "Methods" for details). For each MSigDB collection delineated above, we then computed the correlation over gene sets in the collection between enrichment rank-biserial correlations for differential expression, either overall or per-tissue, and enrichment rank-biserial correlations for noncoding adaptation. Interpreting these correlations is complicated by overlap among sets within a collection—a typical gene is a member of several sets—which tends to inflate the correlations. Thus, for each correlation, we corrected by computing a *p*-value using a form of permutation testing that accounts for overlap among sets.

Table 2 lists the results for overall differential expression (see Additional file 5: Tables S5b–S5f for per-tissue results). The most significant correlations represent several types of function with respect to which there is appreciable congruity between gene expression and adaptive noncoding changes. The C3:TFT collection contains sets of genes with binding sites for a transcription factor according to the TRANSFAC database. The sets most co-enriched with both overall differential expression and noncoding adaptation include targets of *FOXM1*, *POU3F1* and *POU3F2*, *GATA1*, and *NKX2–5*, which are regulators of the cell cycle, neuron development, erythrocyte development, and cardiomyocyte development, respectively [48–51]. The C2:CGP collection contains sets of genes whose regulation is altered in response to a genetic or chemical perturbation. The most co-enriched sets include MOHANKUMAR_TLX1_TARGETS_DN [52], genes down-regulated by TLX1 in a breast cancer cell line; POOLA_INVASIVE_BREAST_CANCER_UP [53], genes up-regulated in tissues from patients with vs. without breast cancer. Beyond these three collections, the most co-enriched sets include C4:CGN:MORF_THPO and MORF_EPHA7 [54], genes in the expression neighborhoods of the cancer-associated genes THPO and EPHA7, respectively; and C5:BP:SYNAPTIC_TRANSMISSION and NERVOUS_SYSTEM_DEVELOPMENT, genes involved in these biological processes according to the GO database [55]. There are prominent themes here, notably cancer and neural development and function.

Discussion

The correlations that we detected between positive selection and changes in gene expression are significant, but not strong. There are several plausible reasons for this, beyond whatever noise exists in the expression measurements and selection scores. It is possible, for instance, that some expression differentiation is neutral or even deleterious, and that some of the adaptation happened along the chimpanzee rather than human lineage. It is also important to bear in mind that both datasets contain only a subset of all possible genes and tissue types over the four individuals. Additionally, several recent studies have shown the importance of looking over multiple tissue types to gain a clearer understanding of important *cis*-regulatory effects within human populations, where eQTLs differ among or across environments [56–58]. Including more expression data would also likely improve the correlation between expression differences

Table 2 Functional correlations between overall differential expression and noncoding adaptation

Collection #	sets	r_r	p(r_r)	q(r_r)
C3:TFT (transcription factor targets)	472	0.35	< 0.0001	< 0.0005
C2:CGP (chemical and genetic perturbations)	645	0.28	< 0.0001	< 0.0005
C7:All (immunologic signatures)	1723	0.14	< 0.0001	< 0.0005
C4:CGN (cancer gene neighborhoods)	88	0.5	0.012	0.045
C6:All (oncogenic signatures)	109	0.22	0.019	0.057
C4:CM (cancer modules)	87	0.29	0.036	0.09
C5:CC (GO cellular components)	37	0.33	0.14	0.27
C3:MIR (microRNA targets)	109	0.11	0.14	0.27
C1:All (positional gene sets)	9	0.17	0.34	0.53
H:All (hallmark gene sets)	25	0.089	0.35	0.53
C2:CP:Reactome (Reactome gene sets)	39	0.058	0.4	0.55
C2:CP:KEGG (KEGG gene sets)	11	−0.17	0.7	0.88
C2:CP:Other (other canonical pathways)	6	−0.49	0.8	0.93
C5:BP (GO biological processes)	123	−0.28	0.95	0.96
C5:MF (GO molecular functions)	44	−0.41	0.96	0.96

All MSigDB collections we considered are listed except for C2:CP:BioCarta, in which no gene set contained at least 50 genes analyzed in a previous study of adaptive noncoding changes. # sets is the number of gene sets we considered in the collection. r_r is the rank correlation between rank-biserial correlations for enrichment with differential expression (each of which is a rank-biserial correlation as in Table 1) and rank-biserial correlations for enrichment with noncoding adaptation (each of which is likewise a rank-biserial correlation). $p(r_r)$ is the one-tailed *p*-value of r_r via permutation test with 10,000 permutations; collections are ordered by increasing $p(r_r)$, which reflects r_r, # sets, and patterns of overlap among gene sets in the collection

and the scans for selection in our study as well as in other, more distal putative enhancer regions [7, 8].

Nevertheless, there are other studies that support the idea that one can detect the effects of adaptation on *cis*-regulation on human traits. A correlation between expression divergence in other tissues and mutations in short core promoter regions has been reported [12]. More recently, an analysis of positive selection across the human genome using human population genomic data (1000 Genomes Project [59]) found that signatures of adaptation are common in regulatory regions (as defined by open chromatin) [60]. These are signals of more recent positive selection than were detected in Haygood et al. [[6]], suggesting that these regulatory regions are continuing to adapt to novel challenges and environments, and, presumably, this correlates with changes in gene expression.

The enrichments for differential expression between humans and chimpanzees by cytogenetic band was somewhat unexpected. For example, the category C4:CGN (cancer gene neighborhoods, q-value = 0.045, Table 2), shows up as significant in our correlations between noncoding adaptation and differential expression. It is possible there are more global regulators of expression for these regions that differ between species. Compared to the genomic regions highlighted in [37, 42, 61], there does seem to be an overlap between these regions and regions where segmental duplications have occurred in the human genome. This could influence some of our signal of positive selection due to orthology issues; however, there is previous evidence of positive selection in regions with duplications, both in coding [42, 61] and in regulatory sequence [37].

We also found a strong correlation between tissue specificity and shifting gene expression between species. This makes sense in the light of pleiotropic concerns, where genes with a greater degree of tissue specificity can change expression pattern or levels over relatively short evolutionary time spans.

One important question is whether there are pleiotropic consequences of these adaptive changes in gene expression that have accrued during human evolution. We find that disease pathways involved in signaling and certain forms of cancer are enriched in the genes that show differential expression between humans and chimpanzees. There appears to be a substantial difference in the frequency of epithelial derived cancers, such as breast, ovary, and prostate between humans and non-human primates: incidence of cancer is significantly higher in humans compared to non-human primates in captivity [62–65], reviewed in [66] and human fibroblasts also show reduced apoptotic ability as compared to chimpanzees and other non-human primates [67]. Although much of this difference is likely driven by environmental and dietary factors, some of the difference may be due to genomic differences between the species. It will take functional studies to understand how these gene expression differences are driving the differences in organismal phenotypes between humans and our closest relatives, and to understand how much of the differences in gene expression we see are due to genetic as opposed to environmental factors.

Conclusions

We find that adaptation in regulatory regions may have driven many of these changes in gene expression, but that may have also had other attendant consequences related to human disease states. In looking at changes across tissues and evolutionary time, we see that shifting specificities across tissues show how gene expression can be the raw material for selection. These data present a window into how positive selection has worked to change gene expression between humans and chimpanzees at the level of larger groups of changing gene expression. These changes underlie the more obvious changes in phenotype, but may have the important side effect of also underling differential disease susceptibilities that may not have been visible to selection.

Methods

Sample preparation and sequencing

We obtained samples of white adipose tissue, lateral cerebellum, frontal cortex, liver, and skeletal muscle from four male humans and four male chimpanzees. The small sample size is due to the limited resources of chimpanzee post-mortem tissues, all from the same individual. Our experimental design was to match the number of individuals across the two species. All samples were obtained through opportunistic sampling of individuals that died of causes unrelated to this research. We obtained human samples from BioChain (Newark, CA) and the NICHD Brain and Tissue Bank for Developmental Disorders and non-human primate samples from the Southwest Foundation for Biomedical Research, and the New Iberia Research Center. We matched samples from each individual across tissues, so that all samples labeled, for example, "human 1" represent the same *trans* environment.

All brain samples from non-human primates were dissected by the same researcher in order to maintain consistency. We stored samples at −80 °C. For more information, see Additional file 1: Table S1.

We isolated total RNA from the samples using RNeasy kits (Qiagen, Valencia, CA), including a DNase I treatment step. Ten micrograms of RNA became the starting material for an RNA-Seq library per sample. We constructed libraries using SOLiD Total RNA-Seq kits (Ambion, Austin, TX), which yield strand-specific reads.

Libraries were sequenced at the Duke University Sequencing and Genomic Technologies Shared Resource (http://genome.duke.edu/cores-and-services/sequencing-and-genomic-technologies/). Sequencing yielded approximately 70 million 50–35 bp paired-end sequences per library.

Sequence mapping and count analysis

We mapped sequences into the human and chimpanzee genomes (hg19, panTro3) using Tophat version 1.4.1 [68]. We constructed orthologous gene models using the Primate Orthologous Exon Database, version 2 (http://precedings.nature.com/documents/7054/version/1). We filtered them using Ensembl (http://useast.ensembl.org/index.html), removing homology types one2many and many2many while retaining homology types one2one and apparentone2one (human-chimpanzee). We also removed ribosomal genes in the RPL, RPS, MRPL, and MRPS families, which include many paralogs with unclear homologies.

We derived counts per gene using HT-Seq [69]. We analyzed them using edgeR [70], defining a multi-factor Generalized Linear Model (GLM) with tissue, species, and species-within-tissue-interaction effects [22]. As a score for differential expression of a gene, we used the negative of the natural logarithm of the adjusted p-value for the appropriate contrast of the GLM. The positional information in Fig. 4 was plotted using the Genome Decoration Page at NCBI.

Tissue specificity analysis

We defined the specificity score of the gene for a tissue as the square of the cosine of the angle between the vector and the axis corresponding to the tissue [as in 6, 28]. This measure depends on the distribution of expression over tissues, but not the overall magnitude of expression. A highly specific gene to one tissue has specificity scores near 1 for this tissue and near 0 for others, whereas housekeeping genes, for example, might be highly expressed in all tissues, and have specificity values around 0.

Analyses of gene ontology groups and comparisons between studies of adaptation and gene expression

We used the Molecular Signatures Database (MSigDB) version 5.0 [30, 31]. More specifically, we used the 15 leaf collections in the tree of all 20 MSigDB collections, with one addition. One MSigDB collection, the C2:CP (canonical pathways) collection, is neither a leaf nor strictly an internal node of the branching pathways, because it includes not only the union of the C2:CP:BioCarta, C2:CP:KEGG, and C2:CP:Reactome collections but also 253 other gene sets. Therefore, we treated these 253 sets as an additional leaf collection, the C2:CP:Other collection.

As a rank-biserial correlation for enrichment of a gene set with genes scoring high for differential expression, either overall or per-tissue, we used the rank-biserial correlation, r_{rb}, over genes whose expression we measured between scores for differential expression and membership in the set. R_{rb} measures association between an ordinal variable and a dichotomous variable [71]. It is proportional to the standard (Pearson) correlation between the ranks of the ordinal variable and any two values, say, 0 and 1, for the dichotomous variable. It is a linear function of the Mann-Whitney U statistic, and a test of $r_{\mathrm{rb}} = 0$ is equivalent to a Mann-Whitney test, which was used to test for categorical enrichment in several previous studies of adaptive sequence changes [6, 28, 72, 73]. An important advantage of r_{rb} for our purposes is that there are well established procedures for combining it across studies [74]. We estimated the standard error of r_{rb}, $\mathrm{SE}(r_{\mathrm{rb}})$, as the standard deviation of r_{rb} over 10,000 bootstrap replicates.

Similarly, as a rank-biserial correlation for enrichment of a gene set with genes scoring high for adaptive sequence changes according to any one of three previous studies of noncoding regions [6–8], we used the rank-biserial correlation between scores for adaptive sequence changes and membership in the set. As in standard fixed-effects meta-analyses [74], we combined r_{rb} across studies to obtain its weighted mean, $\mathrm{WM}(r_{\mathrm{rb}}) = \Sigma_i\ w_i\ (r_{\mathrm{rb}})_i$, where $w_i = (1/\mathrm{SE}(r_{\mathrm{rb}})_i)^2\ /\ \Sigma_j\ (1/\mathrm{SE}(r_{\mathrm{rb}})_j)^2$. Under the null hypothesis $(r_{\mathrm{rb}})_i = (r_{\mathrm{rb}})_j$ for every i and j from 1 to n, the heterogeneity statistic $Q = \Sigma_i\ w_i\ ((r_{\mathrm{rb}})_i - \mathrm{WM}(r_{\mathrm{rb}}))^2$ has approximately the chi-squared distribution with $n - 1$ degrees of freedom, where the set is represented in n studies. We restricted attention to sets with $n \geq 2$ and Q not significantly different from 0 ($\chi^2_{n-1}\ p > 0.05$). All code is available upon request.

Functional correlations between noncoding adaptation and gene expression

The data for noncoding adaptation (Additional file 4: Table S4) are from Haygood et al. [5]. That was a meta-anaylsis of adaptation in coding and noncoding DNA. The three noncoding DNA datasets included a combination of human accelerated regions (where there are significantly more changes on the human branch in a highly-conserved regions in other organisms, anywhere in the genome) [7, 8] and rapidly evolving putative *cis*-regulatory regions near coding sequence (as compared to local introns) [6]. In each survey, we mapped the regulatory regions to the nearest genes with UniProt [75] identifiers and then to MSigDB categories. For each category, we computed the rank-biserial correlation, rrb, between the score for positive selection and membership in the category.

What we term the functional correlation between differential expression, either overall or per-tissue, and noncoding adaptation for an MSigDB collection is the rank (Spearman) correlation, r_r, over gene sets in the collection; this is a correlation between enrichment rank-biserial correlations for differential expression (5% FDR q-value <0.05) and enrichment rank-biserial correlations for noncoding adaptation, restricting attention to sets containing at least 50 genes whose expression we measured. Additionally, noncoding categories have to be represented in at least two of the three previous studies of noncoding regions, with rank-biserial correlations for adaptation not significantly heterogeneous across these studies, and containing at least 50 genes analyzed on the average over these studies.

Gene sets overlap, so the enrichment rank-biserial correlations of different sets are not necessarily independent, which tends to inflate the correlations. Standard permutation testing, permuting rank-biserial correlations among sets within a collection, would be inappropriate. Therefore, instead, we permuted scores for differential expression and for adaptive sequence changes among genes, and for each permutation, we recomputed every enrichment rank-biserial correlation. These rank-biserial correlations represent a null hypothesis of no association between differential expression or adaptive sequence changes and membership in any gene set. Nonetheless, for any given permutation, they may exhibit correlations, not only by chance but also due to overlap among sets, which is preserved in their construction. For each collection, we estimated the one-tailed p-value of r_r, $p(r_r)$, as the fraction of 10,000 permutation replicates in which the replicate r_r for the collection was no less than the observed r_r for the collection or as <0.0001 if there were no such replicates. In Table 2 (and Additional file 5: Tables S5a–S5f), we list not only p-values but also q-values, which adjust for multiple comparisons [76], although we are performing correlations on already adjusted q-values.

Additional files

Additional file 1: Table S1. Details of individuals and samples. (XLSX 36 kb)

Additional file 2: Table S2. Functions of genes expressed differentially overall MSigDB gene sets. These sets are listed from most to least enriched with genes scoring high for differential expression between humans and chimpanzees overall. Collection is the MSigDB collection containing the set (one of 16). # genes is the number of genes whose expression we measured in the set (at least 50). rrb is the rank-biserial correlation between scores for differential expression and membership in the set; sets are ordered by decreasing rrb. SE(rrb) is the standard error of rrb via bootstrapping with 10,000 replicates. (XLSX 1260 kb)

Additional file 3: Table S3. A list of KEGG functions of overall differentially expressed genes. This list is restricted to KEGG pathways (i.e., MSigDB's C2:CP:KEGG collection) and restricted to sets containing at least 50 genes. (XLSX 13 kb)

Additional file 4: Table S4. Genic correlations between differential expression and noncoding adaptation. Correlations are over genes whose expression we measured that were also analyzed in the studies of noncoding regions by Haygood et al. [6], Pollard et al. [7], and Prabhakar et al. [8]. # genes is the number of genes with appropriate scores. rr is the rank correlation between scores for differential expression between humans and chimpanzees, either overall or per-tissue, and scores for adaptive sequence changes during human or chimpanzee evolution, according to a study of noncoding regions. p(rr) is the one-tailed p-value of rr via permutation test with 10,000 permutations. (XLSX 49 kb)

Additional file 5: Table S5. Functional correlations between differential expression overall and noncoding adaptation. All MSigDB collections we considered are listed except for C2:CP:BioCarta, in which no gene set contained at least 50 genes analyzed in a previous study of adaptive noncoding changes. # sets is the number of gene sets we considered in the collection. rr is the rank correlation between rank-biserial correlations for enrichment with differential expression and rank-biserial correlations for enrichment with noncoding adaptation. p(rr) is the one-tailed p-value of rr via permutation test with 10,000 permutations; collections are ordered by increasing p(rr), which reflects rr, # sets, and patterns of overlap among gene sets in the collection. (XLSX 63 kb)

Abbreviations

FDR: False discovery rate; GLM: Generalized linear model; MSigDB: Molecular signatures database

Acknowledgements

We thank the NICHD Brain and Tissue Bank for Developmental Disorders, the Southwest Foundation for Biomedical Research, the New England Primate Center, New Iberia Research Center and Oregon National Primate Research Center, and Julie Horvath at the North Carolina Museum of Natural Sciences. We thank Christine Wall and members of the Wray and Babbitt labs for comments and suggestions.

Author contributions

CCB and WN generated the gene expression datasets. RH performed the rank bi-serial analyses. CCB performed the rest of the analyses and CCB, RH and GAW wrote the manuscript. All authors have read and approved this manuscript for publication.

Funding

This project was funded by National Science Foundation grant NSF-BCS-08-27,552 (HOMINID). This sponsor had no role in the design of the study and collection, analysis, and interpretation of data or in writing this manuscript.

Availability of data and materials

The code for analysis during the current study will be available from the corresponding author on request. The aligned bam files (to hg19 and panTro3) can be found at the Short Read Archive under SUB2571516 and SUB2716550 and under Bioproiect PRJNA387923 and PRJNA382384 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA382384 and https://www.ncbi.nlm.nih.gov/bioproject/PRJNA382384).

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable per Exemption 45 CFR 46.101(b)(4) of OHRP regulations (https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/index.html). The tissues used in this study were publicly available and de-identified, and obtained from the following sources: NICHD Brain and Tissue Bank for Developmental Disorders, and the Biochain Institute, Inc. (human samples), and the Southwest National Primate Research Center (chimpanzee).

Publisher's Note

Author details

[1]Department of Biology, Duke University, Durham, NC 27708, USA. [2]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA. [3]Ronin Institute, Montclair, NJ 07043, USA. [4]Department of Evolutionary Anthropology, Duke University, Durham, NC 27708, USA. [5]Present Address: Department of Biology, University of Massachusetts Amherst, Amherst, MA 01003, USA.

Received: 19 December 2016 Accepted: 31 May 2017
Published online: 05 June 2017

References

1. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 2003;20(9):1377–419.
2. Carroll SB. Evolution at two levels: on genes and form. PLoS Biol. 2005;3(7):1159–66.
3. Garfield D, Haygood R, Nielsen WJ, Wray GA. Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin *Strongylocentrotus purpuratus*. Evol Dev. 2012;14(2):152–67.
4. Shibata Y, Sheffield N, Fedrigo O, Babbitt CC, Wortham M, Tawari AK, London D, Song L, Lee B, Iyer VR, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012;8(6):e1002789.
5. Haygood R, Babbitt CC, Fedrigo O, Wray GA. Contrasts between adaptive coding and non-coding changes during human evolution. Proc Natl Acad Sci U S A. 2010;107(17):7853–7.
6. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet. 2007;39:1140–4.
7. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2006;2:1599–611.
8. Prabhakar S, Noonan JP, Paabo S, Rubin EM. Accelerated evolution of conserved noncoding sequences in humans. Science. 2006;314(5800):786.
9. Babbitt CC, Fedrigo O, Pfefferle AD, Horvath J, Furey TS, Wray GA. Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. Genome Biology and Evolution. 2010;2010:67–79.
10. Babbitt CC, Warner LR, Fedrigo O, Wall CE, Wray GA. Genomic signatures of diet-related shifts in primate evolution. Proc R Soc B. 2011;278:961–9.
11. Khaitovich P, Tang K, Franz H, Kelso J, Hellmann I, Enard W, Lachmann M, Paabo S. Positive selection on gene expression in the human brain. Curr Biol. 2006;16(10):R356–8.
12. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science. 2005;309(5742):1850–4.
13. Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet. 2008;4(11):e1000271.
14. Bozek K, Wei Y, Yan Z, Liu X, Xiong J, Sugimoto M, Tomita M, Paabo S, Pieszek R, Sherwood CC, et al. Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. PLoS Biol. 2014;12(5):e1001871.
15. Bullard JH, Mostovoy Y, Dudoit S, Brem RB. Polygenic and directional regulatory evolution across pathways in Saccharomyces. Proc Natl Acad Sci U S A. 2010;107(11):5058–63.
16. Sezgin E, Duvernell DD, Matzkin LM, Duan Y, Zhu CT, Verrelli BC, Eanes WF. Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in *Drosophila melanogaster*. Genetics. 2004;168(2):923–31.
17. Crawford DL, Powers DA. Molecular basis of evolutionary adaptation at the lactate dehydrogenase-B locus in the fish *Fundulus heteroclitus*. Proc Natl Acad Sci U S A. 1989;86(23):9365–9.
18. Aiello LC, Wheeler P. The expensive-tissue hypothesis - the brain and the digestive-system in human and primate evolution. Curr Anthropol. 1995;36(2):199–221.
19. Leonard WR, Robertson ML, Snodgrass JJ, Kuzawa CW. Metabolic correlates of hominid brain evolution. Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology. 2003;136(1):5–15.
20. Finch CE, Stanford CB. Meat-adaptive genes and the evolution of slower aging in humans. Q Rev Biol. 2004;79(1):3–50.
21. Ungar PS, Grine FE, Teaford MF. Diet in early homo: a review of the evidence and a new model of adaptive versatility. Annu Rev Anthropol. 2006;35:209–28.
22. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97.
23. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 2010;20(2):180–9.
24. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet. 2011;7(2):e1001316.
25. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660–5.
26. Gilad Y, Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Res. 2015;4:121.
27. Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol. 2015;16:287.
28. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. Patterns of positive selection in six mammalian genomes. PLos Genetics 2008, 4(8).
29. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:206–16.
30. Liberzon A. A description of the molecular signatures database (MSigDB) web site. Methods Mol Biol. 2014;1150:153–60.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
32. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
33. Boettger LM, Handsaker RE, Zody MC, McCarroll SA. Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet. 2012;44(8):881–5.
34. Wider C, Vilarino-Guell C, Jasinska-Myga B, Heckman MG, Soto-Ortolaza AI, Cobb SA, Aasly JO, Gibson JM, Lynch T, Uitti RJ, et al. Association of the MAPT locus with Parkinson's disease. Eur J Neurol. 2010;17(3):483–6.
35. Dubourg C, Sanlaville D, Doco-Fenzy M, Le Caignec C, Missirian C, Jaillard S, Schluth-Bolard C, Landais E, Boute O, Philip N, et al. Clinical and molecular characterization of 17q21.31 microdeletion syndrome in 14 French patients with mental retardation. Eur J Med Genet. 2011;54(2):144–51.
36. Ikram MA, Fornage M, Smith AV, Seshadri S, Schmidt R, Debette S, Vrooman HA, Sigurdsson S, Ropele S, Taal HR, et al. Common variants at 6q22 and 17q21 are associated with intracranial volume. Nat Genet. 2012;44(5):539–44.
37. Bekpen C, Tastekin I, Siswara P, Akdis CA, Eichler EE. Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region. Genome Res. 2012;22(6):1050–8.
38. Jamieson SE, Miller EN, Black GF, Peacock CS, Cordell HJ, Howson JM, Shaw MA, Burgner D, Xu W, Lins-Lainson Z, et al. Evidence for a cluster of genes on chromosome 17q11-q21 controlling susceptibility to tuberculosis and leprosy in Brazilians. Genes Immun. 2004;5(1):46–57.
39. Nomiyama H, Fukuda S, Iio M, Tanase S, Miura R, Yoshie O. Organization of the chemokine gene cluster on human chromosome 17q11.2 containing the genes for CC chemokine MPIF-1, HCC-2, HCC-1, LEC, and RANTES. J Interf Cytokine Res. 1999;19(3):227–34.
40. Keene KL, Quinlan AR, Hou X, Hall IM, Mychaleckyj JC, Onengut-Gumuscu S, Concannon P. Evidence for two independent associations with type 1 diabetes at the 12q13 locus. Genes Immun. 2012;13(1):66–70.
41. Zhang Y, Song G, Vinar T, Green ED, Siepel A, Miller W. Reconstructing the Evolutionary History of Complex Human Gene Clusters. In: Research in Computational Biology RECOMB 2008. Springer; 2008: 29–49.
42. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. Positive selection of a gene family during the emergence of humans and African apes. Nature. 2001;413(6855):514–9.
43. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, et al. A transcriptome database for

astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci. 2008;28(1):264–78.
44. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445(7124):168–76.
45. Rickman DS, Millon R, De Reynies A, Thomas E, Wasylyk C, Muller D, Abecassis J, Wasylyk B. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. Oncogene. 2008;27(51):6607–22.
46. Anastassiou D, Rumjantseva V, Cheng WY, Huang JZ, Canoll PD, Yamashiro DJ, Kandel JJ. Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. BMC Cancer. 2011;11:529-37.
47. Nakayama R, Nemoto T, Takahashi H, Ohta T, Kawai A, Seki K, Yoshida T, Toyama Y, Ichikawa H, Hasegawa T. Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. Mod Pathol. 2007;20(7):749–59.
48. Wierstra I. The transcription factor FOXM1 (Forkhead box M1): proliferation-specific expression, transcription factor function, target genes, mouse models, and normal biological roles. Adv Cancer Res. 2013;118:97–398.
49. Nelms BL, Labosky PA. In: Transcriptional Control of Neural Crest Development. San Rafael (CA): Morgan & Claypool Life Sciences; 2010.
50. Simon MC. Transcription factor GATA-1 and erythroid development. Proc Soc Exp Biol Med. 1993;202(2):115–21.
51. Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP, Maron BJ, Seidman CE, Seidman JG. Congenital heart disease caused by mutations in the transcription factor NKX2-5. Science. 1998;281(5373):108–11.
52. Mohankumar KM, Xu XQ, Zhu T, Kannan N, Miller LD, Liu ET, Gluckman PD, Sukumar S, Emerald BS, Lobie PE. HOXA1-stimulated oncogenicity is mediated by selective upregulation of components of the p44/42 MAP kinase pathway in human mammary carcinoma cells. Oncogene. 2007;26(27):3998–4008.
53. Poola I, DeWitty RL, Marshalleck JJ, Bhatnagar R, Abraham J, Leffall LD. Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis. Nat Med. 2005;11(5):481–3.
54. Brentani H, Caballero OL, Camargo AA, da Silva AM, da Silva WA Jr, Dias Neto E, Grivet M, Gruber A, Guimaraes PE, Hide W, Iseli C, Jongeneel CV, Kelso J, Nagai MA, Ojopi EP, Osorio EC, Reis EM, Riggins GJ, Simpson AJ, de Souza S, Stevenson BJ, Strausberg RL, Tajara EH, Verjovski-Almeida S, Acencio ML, Bengtson MH, Bettoni F, Bodmer WF, Briones MR, Camargo LP, Cavenee W, Cerutti JM, Coelho Andrade LE, Costa dos Santos PC, Ramos Costa MC, da Silva IT, Estécio MR, Sa Ferreira K, Furnari FB, Faria M Jr, Galante PA, Guimaraes GS, Holanda AJ, Kimura ET, Leerkes MR, Lu X, Maciel RM, Martins EA, Massirer KB, Melo AS, Mestriner CA, Miracca EC, Miranda LL, Nobrega FG, Oliveira PS, Paquola AC, Pandolfi JR, Campos Pardini MI, Passetti F, Quackenbush J, Schnabel B, Sogayar MC, Souza JE, Valentini SR, Zaiats AC, Amaral EJ, Arnaldi LA, de Araújo AG, de Bessa SA, Bicknell DC, Ribeiro de Camaro ME, Carraro DM, Carrer H, Carvalho AF, Colin C, Costa F, Curcio C, Guerreiro da Silva ID, Pereira da Silva N, Dellamano M, El-Dorry H, Espreafico EM, Scattone Ferreira AJ, Ayres Ferreira C, Fortes MA, Gama AH, Giannella-Neto D, Giannella ML, Giorgi RR, Goldman GH, Goldman MH, Hackel C, Ho PL, Kimura EM, Kowalski LP, Krieger JE, Leite LC, Lopes A, Luna AM, Mackay A, Mari SK, Marques AA, Martins WK, Montagnini A, Mourão Neto M, Nascimento AL, Neville AM, Nobrega MP, O'Hare MJ, Otsuka AY, Ruas de Melo AI, Paco-Larson ML, Guimarães Pereira G, Pereira da Silva N, Pesquero JB, Pessoa JG, Rahal P, Rainho CA, Rodrigues V, Rogatto SR, Romano CM, Romeiro JG, Rossi BM, Rusticci M, Guerra de Sá R, Sant' Anna SC, Sarmazo ML, Silva TC, Soares FA, Sonati Mde F, de Freitas Sousa J, Queiroz D, Valente V, Vettore AL, Villanova FE, Zago MA, Zalcberg H; Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium; Human Cancer Genome Project Sequencing Consortium. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A. 2003;100(23):13418-23.
55. Consortium TGO. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
56. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44(10):1084–9.
57. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325(5945):1246–50.
58. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011;7(2):e1002003.
59. Genomes Project C, Abecasis GR, Auton A, Brooks LD, MA DP, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
60. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24(6):885–95.
61. Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. Adaptive evolution of young gene duplicates in mammals. Genome Res. 2009;19(5):859–67.
62. Seibold HR, Wolf RH. Neoplasms and proliferative lesions in 1065 nonhuman primate necropsies. Lab Animal Science. 1973;23:533–9.
63. Beniashvili DS. An overview of the world literature on spontaneous tumors in nonhuman primates. J Med Primatol. 1989;18:423–37.
64. Scott GBD. Comparative primate pathology. New York, NY: Oxford University Press; 1992.
65. McClure HM. Tumors in nonhuman primates: observations during a six-year period in the Yerkes primate Center Colony. Am J Phys Anthropol. 1973;38(2):425–9.
66. Varki A. A chimpanzee genome project is a Biomedical imperative. Genome Res. 2000;10:1065–70.
67. Arora G, Mezencev R, McDonald JF. Human cells display reduced apoptotic function relative to chimpanzee cells. PLoS One. 2012;7(9):e46182.
68. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
69. Anders S, Pyl PT, Huber W. HTSeq-a python framework to work with high-throughput sequencing data. Bioinformatics. 2014.
70. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
71. Kraemer HC. Biserial Correlation. In: Encyclopaedia of Statistical Sciences. vol. Volume 1: Wiley: Hoboken; 1982:276–79.
72. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153–7.
73. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005;3:976–85.
74. Hedges LV. Fixed effects models. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 285–99.
75. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69.
76. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5.