Human Lineage-Specific Transcriptional Regulation through GA-Binding Protein Transcription Factor Alpha (GABPa)

Alvaro Perdomo-Sabogal,[1,2] Katja Nowick,[*,1,2] Ilaria Piccini,[3,5] Ralf Sudbrak,[4,5] Hans Lehrach,[5] Marie-Laure Yaspo,[5] Hans-Jörg Warnatz,[†,5] and Robert Querfurth[*,†,5]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Leipzig, Germany

[2]Paul-Flechsig Institute for Brain Research, University of Leipzig, Leipzig, Germany

[3]Institute of Genetics of Heart Diseases (IfGH), Department of Cardiovascular Medicine, University Hospital Münster, 48149 Münster, Germany

[4]European Centre for Public Heath Genomics, UNU-MERIT, Unsiversity Maastricht,PO Box 616, 6200 MD Maastricht, The Netherlands

[5]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: nowick@bioinf.uni-leipzig.de, Robert.Querfurth@gmx.de.

Associate editor: Gregory Wray

Abstract

A substantial fraction of phenotypic differences between closely related species are likely caused by differences in gene regulation. While this has already been postulated over 30 years ago, only few examples of evolutionary changes in gene regulation have been verified. Here, we identified and investigated binding sites of the transcription factor GA-binding protein alpha (GABPa) aiming to discover *cis*-regulatory adaptations on the human lineage. By performing chromatin immunoprecipitation-sequencing experiments in a human cell line, we found 11,619 putative GABPa binding sites. Through sequence comparisons of the human GABPa binding regions with orthologous sequences from 34 mammals, we identified substitutions that have resulted in 224 putative human-specific GABPa binding sites. To experimentally assess the transcriptional impact of those substitutions, we selected four promoters for promoter-reporter gene assays using human and African green monkey cells. We compared the activities of wild-type promoters to mutated forms, where we have introduced one or more substitutions to mimic the ancestral state devoid of the GABPa consensus binding sequence. Similarly, we introduced the human-specific substitutions into chimpanzee and macaque promoter backgrounds. Our results demonstrate that the identified substitutions are functional, both in human and nonhuman promoters. In addition, we performed GABPa knock-down experiments and found 1,215 genes as strong candidates for primary targets. Further analyses of our data sets link *GABPa* to cognitive disorders, diabetes, KRAB zinc finger (KRAB-ZNF), and human-specific genes. Thus, we propose that differences in GABPa binding sites played important roles in the evolution of human-specific phenotypes.

Key words: **GABP, promoter assay, human-specific binding sites, human molecular evolution, KRAB zinc finger genes, ChIP-Seq, comparative genomics.**

Introduction

Regulation of gene expression is a major mechanism shaping phenotypic traits of organisms (Britten and Davidson 1971; King and Wilson 1975; Wray 2007). Gene expression is regulated to a large extent by transcription factors (TFs), proteins that bind to specific DNA sites of typically 5–15 bp, known as TF binding sites (TFBS) or *cis*-regulatory sites. Even though there is some sequence variation in the binding sites recognized by a certain TF, residing nucleotide substitutions can have a substantial impact on the affinity of the TF and hence on the levels at which the target genes are transcribed (Lin et al. 2007). Consequently, evolutionary changes in the sequence of a TFBS can change the expression of the target genes and lead to phenotypic differences and speciation.

Previous studies aiming at identifying species-specific *cis*-regulatory sites were driven by the interest in a particular gene (Huby et al. 2001; Rockman et al. 2005; Romanelli et al. 2009), or based on genome-wide bioinformatics approaches, for instance, the search for certain substitution patterns in multiple species alignments (Haygood et al. 2007; Taylor et al. 2008; Molineris et al. 2011). However, experimentally supported examples for species-specific *cis*-regulatory sites are still sparse. This is partially due to laborious experimental approaches required to confirm the functionality of a binding site. Some in vitro studies successfully demonstrated functionality of species-specific TFBSs using promoter reporter gene assays (Chabot et al. 2007), and a selected number of determined changes have also been tested *in vivo* in zebrafish and mice (Stedman et al. 2004; Rockman et al. 2005;

Article

Brinkman-Van der Linden et al. 2007; Prabhakar et al. 2008; Capra et al. 2013).

Here, we investigated binding sites of the GA-binding protein alpha subunit (GABPa) by using chromatin immunoprecipitation and comparative genomic approaches. *GABPa* belongs to the E-twenty-six (ETS) (E-26) family of DNA-binding factors and regulates a broad range of genes involved in cell cycle control, apoptosis, differentiation, hormonal regulation, and other critical cellular functions. GABPa is also recognized as an important transcriptional regulator in myeloid cell differentiation and neuromuscular function (Rosmarin et al. 2004; Yang et al. 2011). It further regulates the expression of genes implicated in mitochondrial functions (Yang et al. 2014) during early cleavage events of the embryo (Ristevski et al. 2004), and inactivation of GABPa induces early embryonic lethality in mice (Ristevski et al. 2004; Jaworski et al. 2007).

The GABPa binding motif has been confirmed by several studies (Batchelor et al. 1998; Lin et al. 2007). GABPa forms hetero-tetramers composed of two GABPa and two GABPb subunits to bind tandem repeats of the GGAA consensus motif (Batchelor et al. 1998). Repetitions of GABPa binding sites influence transcription levels in a synergistic manner (Yu et al. 1997). It is also common for GABPa to regulate bidirectional promoters (Collins et al. 2007). Importantly, the DNA-binding domain of the GABPa protein is entirely conserved in primates, mouse, dog, and cow, making it likely that GABPa binds to the same motif in all these species. Therefore, analyzing changes in DNA sequence resulting in gain or loss of GABPa binding sites will facilitate the identification of genes whose regulation has been altered by GABPa along the lineages leading to humans and hominids. In addition, given that GABPa preferentially binds in close proximity to the transcriptional start site (TSS) (Valouev et al. 2008), it is possible to evaluate potential binding site alterations straightforwardly by using promoter-reporter gene assays in human and nonhuman primate cells.

We performed chromatin immunoprecipitation-sequencing (ChIP-Seq) assays in human embryonic kidney (HEK293T) cells to identify TFBS for GABPa and independently supported the functionality of the binding sites by using two different RNA interference molecules to knockdown its expression. To search for GABPa consensus binding site emergence along the human lineage, we reconstructed the ancestral sequences of the binding sites within the ChIP-Seq peak regions based on multiple species alignments (fig. 1). We identified human and hominid-specific GABPa binding sites and exemplarily evaluated the functionality of four human wild-type (wt) promoters on gene expression by performing promoter reporter gene assays in HEK293T cells. By mutating the human/hominid-specific GABPa binding sites, we could also evaluate the activity of the ancestral states of these promoters lacking the GABPa consensus binding site. In addition, to test whether the human/hominid binding sites drive similar expression in a chimpanzee and rhesus macaque background, we also introduced the human/hominid-specific substitutions into the chimpanzee and rhesus macaque promoters. Lastly, we also introduced all promoter reporter gene constructs in COS-1 cells to test whether the activity of the binding sites would change within the background of a nonhuman primate cell line.

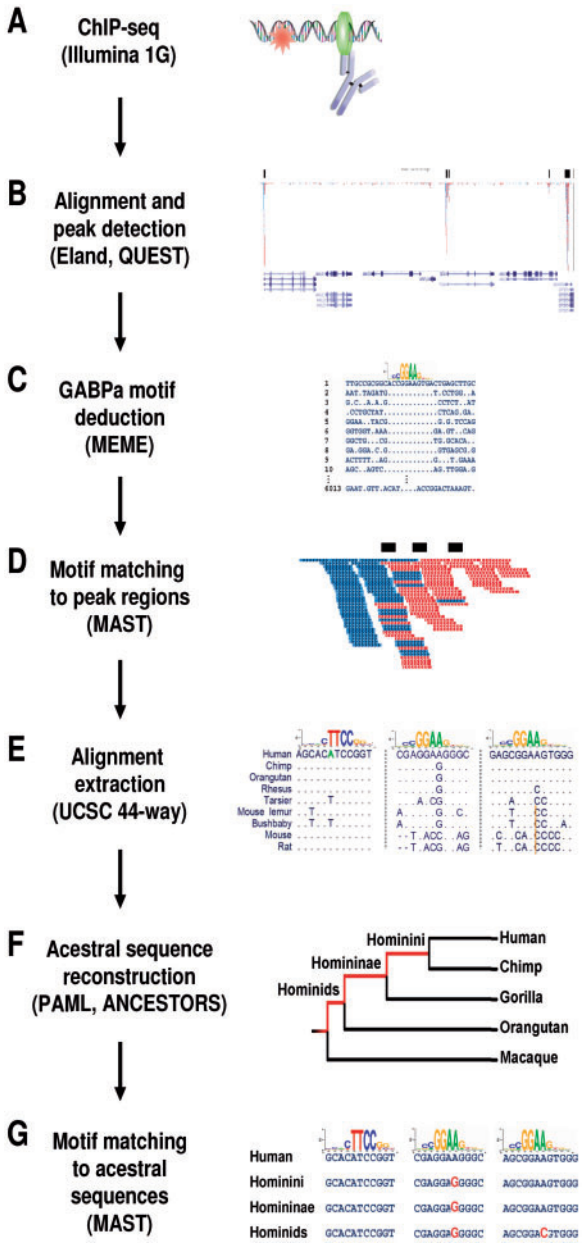

FIG. 1. Overview of the experiments and analyses performed in this study. (*A*) Schematic representation of ChIP-Seq experiments with a GABPa-specific antibody in HEK293T cells to identify binding sites for GABPa. (*B*) Peak detection using the peak-calling algorithm available in QuEST. (*C*) Search for the GABPa consensus binding sequence and deduction of a PWM using the motif discovery tool Multiple Em for Motif Elicitation (MEME). (*D*) Search for additional binding sites within the peak region using the motif alignment and scan tool MAST. (*E*) Extraction of multiple species alignments of the MAST-identified binding sites from the UCSC MultiZ 44 vertebrate alignments. (*F*) Ancestral sequence reconstruction along with the phylogeny of 34 mammalian species from the UCSC 44-vertebrate alignments using the tools PAML and ANCESTORS. (*G*) GABPa motif matching within reconstructed ancestral binding site sequences using MAST. The dots in panels *C* and *E* represent sequence conservation.

Results

Identification of GABPa Binding Sites by Chromatin Immunoprecipitation

To identify binding sites for GABPa, we performed ChIP-Seq experiments with a GABPa-specific antibody in HEK293T cells (fig. 1A). This antibody has been validated for ChIP-Seq of GABPa, resulting in highly specific GABPa peak regions (Valouev et al. 2008). Our peak calling from the ChIP-Seq reads resulted in a list of 6,208 GABPa binding peaks (see Materials and Methods for details).

In order to derive a GABPa consensus binding motif from the ChIP-Seq peak regions, we used 200 bp DNA sequences equally surrounding the center of the 6,208 peaks as input for the de novo motif discovery algorithm MEME (fig. 1B and supplementary table S1, Supplementary Material online) (Bailey et al. 2009). This allowed us to build a consensus binding sequence and a position-specific weight matrix (PWM) for GABPa of 11 bp in length (figs. 1C and 2). The PWM-contributing sites were in 93% located close to the peak centers (fig. 3A), indicating proper peak-calling from ChIP-Seq reads and further demonstrating the validity of our ChIP-Seq data. The identified PWM is very similar to the GABPa PWMs available in the JASPAR and TRANSFAC TFBS databases, and it is almost identical to the one found by Valouev et al. (2008) in Jurkat cells (fig. 2). Taken together, the binding sites and motif identified here seem to be reliable.

Under default parameters, the MEME algorithm assumes that each peak contains zero or one sequence motif. This assumption is advantageous to find nonrepetitive motif elements. However, as more than one motif is likely present in each peak region (Yu et al. 1997; fig. 3B), we searched for additional binding sites. To this end, we used the motif alignment and scan tool MAST (Bailey et al. 2009; fig. 1D). The MAST analysis revealed 11,619 PWM hits in 5,797 peak regions of 200 bp length, with the majority of peaks containing two binding sites, closely followed by peaks with single sites (fig. 3C and supplementary table S2, Supplementary Material online).

Out of the 6,208 genomic peaks for GABPa, 4,277 (69%) peaks were located within 300 bp up- and downstream of the TSSs of 11,848 UCSC transcripts corresponding to 3,994 putative target genes (Entrez IDs). When we extended the window to ± 5 kb centered to the TSSs, the number of peaks mapping to transcripts increased to 5,321 (86%), corresponding to 15,046 transcripts from 5,218 putative targets (supplementary table S3, Supplementary Material online). Further extending the window to ± 10 kb, the numbers increased to 5,465 peaks (88%) mapping to 18,730 transcripts and corresponding to 5,784 putative genes. The majority of peaks reside close to the TSS (fig. 3D). We also detected that GABPa binding sites seem to experience depletion between −150 and −100 bp distance from the closest TSS (fig. 3E). For downstream analyses we used mappings within ±5 kb centered on UCSC-annotated TSSs.

To further validate these putative GABPa binding regions, we performed gene reporter assays and GABPa knock-down experiments in the same cell lines (see below). In addition, we also generated a replicate ChIP-Seq data set and compared it to the peaks we obtained from the first ChIP-seq experiment. We found that 91% of the regions in the initial experiment (5,647 peaks) were also present in the regions from the replicate experiment (supplementary fig. S1, Supplementary Material online). According to the recommendation of the ENCODE consortium for ChIP-Seq experiments, we also calculated the Irreproducible Discovery Rate (IDR) (Landt et al. 2012) for the two replicates and found that 3,677 peaks (∼60%) overlapped at IDR less than 0.05 (supplementary fig. S2, Supplementary Material online), demonstrating reasonable consistency among the two replicates. Furthermore, we also analyzed agreement of our ChIP-Seq data with the ChIP-Seq data obtained by ENCODE in five different cell lines. We found a representative overlap between our data and these data sets, with approximately 60% of our peaks overlapping with at least one of the other data sets (supplementary fig. S3, Supplementary Material online).

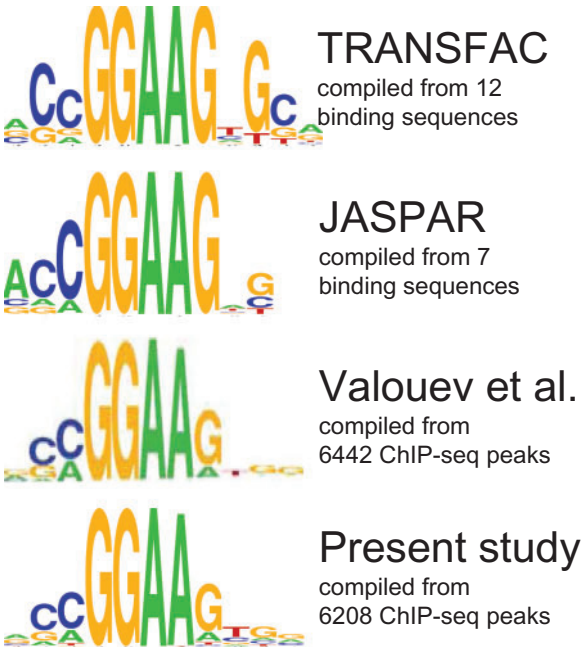

FIG. 2. Comparison of GABPa motifs from different studies and databases. Sequence logos represent the different PWMs.

Identification of Newly Evolved GABPa Binding Sites

In order to identify human-specific GABPa binding sites, we used all 11,619 GABPa binding sites to extract multiple sequence alignments from the UCSC MultiZ 44 vertebrate alignments. We obtained 11,008 alignments (fig. 1E), while for the remaining 611 binding site regions there was either no alignment available or the aligned regions were not contiguous. In addition, we used the eight available nonhuman primate genomes (Chimpanzee, Gorilla, Orangutan, Macaque, Marmoset, Tarsier, Mouse Lemur, and Bushbaby), which allowed us to identify those binding sites that are specific to the Human, Hominini (Human and Chimpanzee), Homininae (Homini and Gorilla), and Hominidae (Homininae and

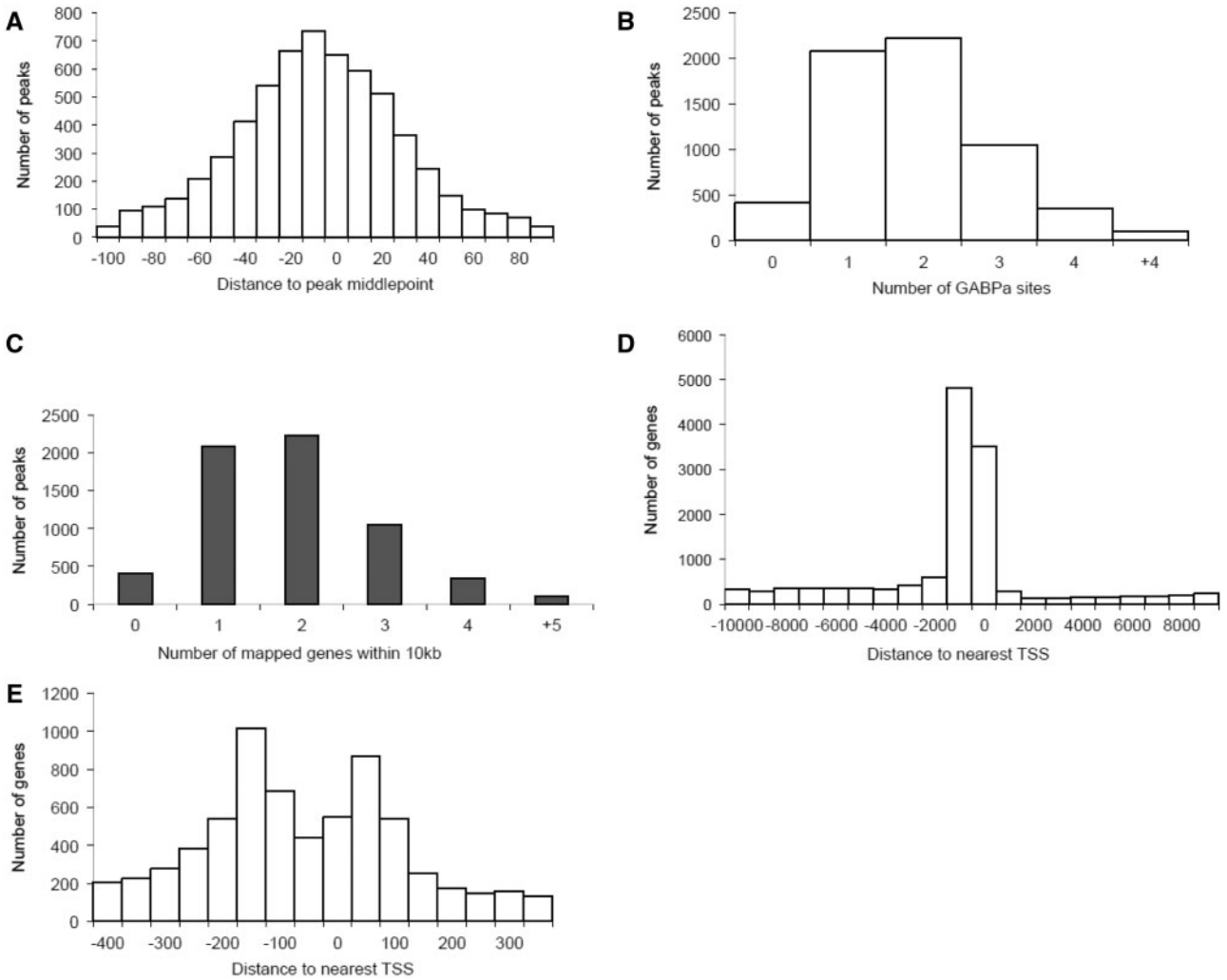

FIG. 3. GABPa peaks map close to gene starts and harbor GABPa binding sites residing closely to the peak centers. (*A*) Distance of the sites contributing to the MEME motif (6,031 of 6,208 in total) to the ChIP peak centers. (*B*) Number of GABPa motifs found per each ChIP-Seq peak. (*C*) Distribution of motif occurrences within 200 bp surrounding the ChIP peak centers. (*D*) Distance of peak calls to the nearest TSSs of UCSC genes within 10 kb, centered on the TSS. (*E*) ± 400 bp zoom of the peak calls to the nearest TSS. The *x* axis shows the distance to the nearest TSS in base pairs. Negative values represent upstream, positive values downstream regions.

Orangutan) lineages. We could reconstruct a hominid ancestral sequence for all except for 65 binding sites which miss aligned sequences (supplementary table S4, Supplementary Material online). We then searched within these ancestral sequences for the presence of GABPa motifs (supplementary table S5, Supplementary Material online).

We discovered 224 GABPa binding sites that are specific to humans. These binding sites correspond to 219 ChIP-Seq peaks mapping within ±5 kb centered on TSSs of 217 genes (supplementary table S6, Supplementary Material online). Three of these genes are human-specific (*CEP170*, *RPL41*, and *GUSBP4*), and three more are Hominidae-specific (*STAG3L4*, *USP6*, and *ZNF383*) (Zhang et al. 2010) (supplementary table S7, Supplementary Material online). Two peaks with human-specific binding sites did not map to known genes. Manual inspection revealed that one of these peaks is located 317 nt upstream of the transfer RNA Phe (anticodon GAA) gene (uc021qjx.1), whereas the other one is 3,810 nt upstream of a human cDNA (uc021suf.1). Among the 217 promoters that gained human-specific GABPa binding sites, we detected 12 KRAB-ZNF TFs. Our subsequent test revealed that KRAB-ZNFs are indeed significantly overrepresented among genes with human-specific GABPa binding sites in their promoters (P value = 0.01116, Fisher's exact test). For the ancestral sequences of Hominini, Homininae, and Hominids, we identified 57, 244, and 310 lineage-specific binding sites (supplementary table S8, Supplementary Material online) mapping to 44, 240, and 326 genes, respectively (supplementary table S9, Supplementary Material online). Binding site appearances for all ancestral branches leading to human are shown in supplementary figure S4, Supplementary Material online.

To gain insights into the functions of those genes belonging to the 217 promoters that gained human-specific GABPa binding sites, we further performed enrichment tests based on Gene Ontology (GO) annotations. We found several enriched GO terms, for example, terms associated with heart development, RNA processing of tRNAs and mRNAs,

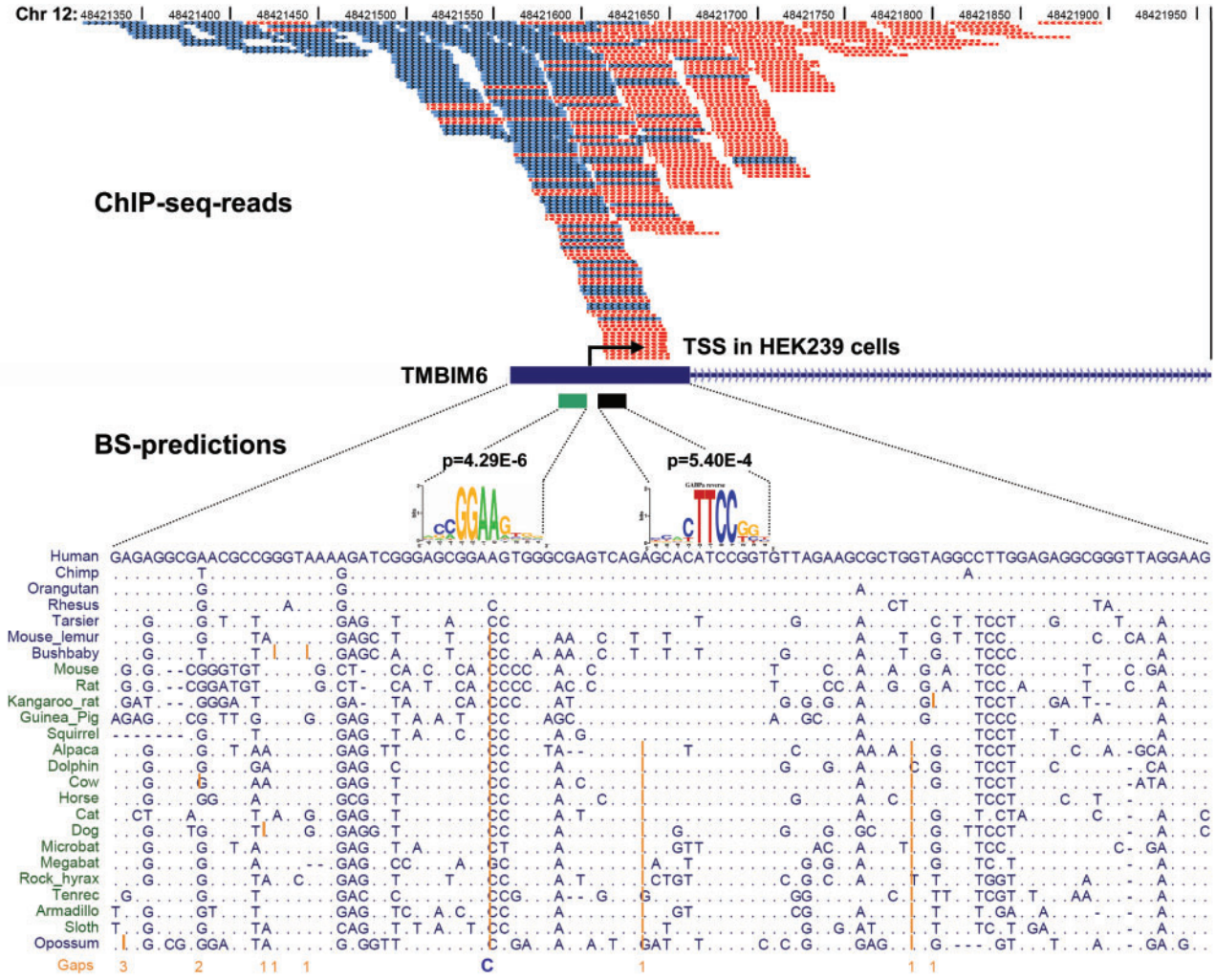

Fig. 4. Genomic view of GABPa ChIP-Seq reads spanning the *TMBIM6* promoter including GABPa binding site predictions and multiple species alignment of the first exon. ChIP-Seq reads are colored in blue (forward reads) and red (reverse reads). The first exon (5′-UTR) is shown as blue bar with a black arrow indicating the TSS in HEK293T cells as determined by RNA-Seq. GABPa binding site predictions are shown as green and black boxes. Within the blowup in the lower part, including the UCSC multiple species alignment of exon 1, binding sites are shown as sequence logos of the GABPa PWM aligned to their matching positions. Within the alignment, dots indicate identity to the human reference sequence, whereas orange vertical bars indicate bases that are not depicted. Orange numbers below represent the sum of bases not depicted. The blue (*C*) illustrates the presence of a single cytosine in all nonhaplorhini at the indicated site.

mammary morphogenesis and development, lipid biosynthesis and metabolism, signaling pathways involved in ventral spinal cord interneuron and spinal cord motor neuron cell fate specification, and dorsal/ventral neural tube patterning (supplementary table S10, Supplementary Material online).

Functional Analysis of Newly Evolved GABPa Binding Sites Using Reporter Gene Assays

From the set of promoters with human-specific GABPa binding sites, we selected four promoters for further experimental validation: the promoters of *ZNF197*, *ANTXR1*, and *TMBIM6* and the bidirectional promoter of *ZNF398*/*ZNF425*. *ZNF197*, *ZNF398*, and *ZNF425* were chosen as representatives of the KRAB-ZNF family. The promoter of *ZNF197* has two binding sites; one is conserved among mammals, whereas the other one is specific to humans. The anthrax toxin receptor-1 gene (*ANTXR1*) harbors three GABPa binding sites in the human promoter, but only two in chimpanzee and rhesus. In addition, this gene is highly expressed in HEK293T cells. Even though we were particularly interested in human-specific binding site gains, we included the *TMBIM6* promoter, which harbors a hominid-specific binding site, because of its strong ChIP-Seq peak and expression levels of the corresponding gene (Sultan et al. 2008). Interestingly, MAST analysis predicted another GABPa binding site next to the hominid-specific binding site for *TMBIM6*, which is deeply conserved. However, this variation was found in only 0.94% of all 11,008 binding sites and it does not match the GABPa core consensus sequence "GGAA" (see fig. 4). Lastly, the promoter of the KRAB-ZNF genes *ZNF398*/*ZNF425* was chosen as a representative of a bi-directional promoter having TSSs located approximately 130 bp apart. It harbors two overlapping GABPa binding sites caused by two human-specific single nucleotide changes. This promoter was cloned in both

directions to account for bidirectional transcription. A genomic view of ChIP-Seq peaks, cloned fragments, sequence differences to the human reference sequence, and binding site predictions can be found in supplementary figure S5, Supplementary Material online.

Orthologous promoters were cloned from human, chimpanzee, and rhesus macaque genomic DNA. For each promoter, two fragments were cloned, one representing the wt and the other one a mutated (mut) form. For human mutated forms, the binding sites were modified by one or two single nucleotide mutations to mimic the ancestral state incompatible with GABPa binding. Inversely, original sequences for chimpanzee and macaque were altered to generate the human-specific GABPa binding sites. All wt and mutated promoters were cloned into a modified *firefly* luciferase reporter gene vector (pGL3, see Materials and Methods) and verified by whole-insert sequencing. Reporter gene expression was measured in human HEK293T cells and COS-1 cells derived from African green monkey and normalized to a cotransformed plasmid stably expressing *Renilla* luciferase.

For all selected promoters we detected a measurable effect of the sequence differences in GABPa binding sites on the promoter activity (see figs. 5 and 6). The human *ZNF197* wt promoter that harbors one conserved and one human-specific GABPa binding site showed significantly higher activity than the chimpanzee and rhesus macaque wt promoters (figs. 5A and 6). The introduction of a single nucleotide mutation to create the human-specific binding site within the sequence background of the chimpanzee and rhesus macaque promoter sequences resulted in a significant increase of the promoter activity in both cell lines, lifting reporter activities almost to the level of the human wt sequence. However, the activity of the human promoter did not change upon the binding site disruption to the ancestral state, thus indicating that also other sequences adjacent to the human GABPa binding site might contribute to the overall promoter activity.

The human *ANTXR1* promoter with two conserved and one human-specific binding site showed significantly higher activity than the chimpanzee and rhesus macaque wt orthologs in at least one of the two cell lines (figs. 5*B* and 6). Introduction of the human binding site into chimpanzee and macaque promoters raised activity levels significantly in three of the four cases, namely for chimpanzee in both cell lines and for macaque in HEK293T cells. Single nucleotide mutations of the human-specific binding site for creating the ancestral state caused significantly decreased reporter activity in COS-1 cells and also decreased activity in HEK293T cells.

The promoter of *TMBIM6* contains a hominid-specific GABPa binding site (see fig. 4). As expected, the activities of the human and chimpanzee wt promoter were higher than the activity of the rhesus macaque wt promoter in both cell lines. Surprisingly, the human wt promoter activity was found to be significantly higher than that of the chimpanzee wt promoter (figs. 5C and 6). This suggests an influence of sequence differences between humans and chimpanzees on the promoter activity. The introduction of the hominid-specific site into the macaque promoter resulted in a highly significant change in activity, incrementing the intensities above chimpanzee wt activity. Disruption of the hominid-specific binding site in human decreased the activity slightly below chimpanzee wt activity, whereas disruption of the chimpanzee binding site decreased the activity below macaque wt activity.

The bidirectional promoter of *ZNF398*/*ZNF425* showed significant differences in the activities of the human, chimpanzee, and rhesus macaque wt promoters in the direction of *ZNF398* in COS-1 cells and *ZNF425* in HEK293T cells (figs. 5B, 5E and 6). The introduction of the human binding site into chimpanzee and rhesus macaque promoters resulted in a significant increase in activity in both cell lines. The reversion of the human binding site to the chimpanzee sequence caused more than a 2-fold reduction in promoter activity in both cell lines.

In summary, the introduction of human GABPa binding sites into chimpanzee or rhesus macaque promoters resulted in significant increase in reporter gene expression in 17 of 18 cases in both cell lines. In contrast, the disruption of GABPa binding sites in human and chimpanzee promoters led to a significant decrease of the reporter gene activity in 9 out of 12 cases. In none of the cases we observed conflicting effects, since binding site introduction did never lead to significant activity decrease, and binding site disruption did never result in any significant activity increase.

Identification of Differential Gene Expression after GABPa Knock-Down

To obtain an independent confirmation of the functionality of all GABPa binding sites determined here by ChIP-Seq, we carried out two knock-down experiments of GABPa in HEK293T cells. Knock-down was performed through transfection of cells with two independent types of silencing molecules (Qiagen SI00423311 and Invitrogen HSS103907), followed by genome-wide expression profiling at two different time points (RNA extraction at 24 and 72 h after transfection) (see Materials and Methods). We only considered genes that were significantly differentially expressed (DE) with both siRNA molecules. Among 14,873 expressed genes, we found 1,156 (24 h after transfection) and 3,238 (72 h) to be DE. For the 72 h experiment, more DE genes than expected by chance harbored GABPa ChIP-Seq peaks in their regulatory sequence ($P < 0.001$) (see Materials and Methods). In total, out of 4,531 genes that are expressed at the 72 h time point and have at least one GABPa binding site, about a quarter (1,215) showed significant changes in expression after GABPa knock-down. These genes are strong candidate target genes of GABPa (supplementary table S11, Supplementary Material online).

Analyses of DE Genes

Next, we used GO enrichment analyses to determine the terms associated with those 1,215 strong candidate target genes of GABPa (see Materials and Methods). A number of functional categories were significantly enriched; including several GO terms associated with RNA processing, ribosome

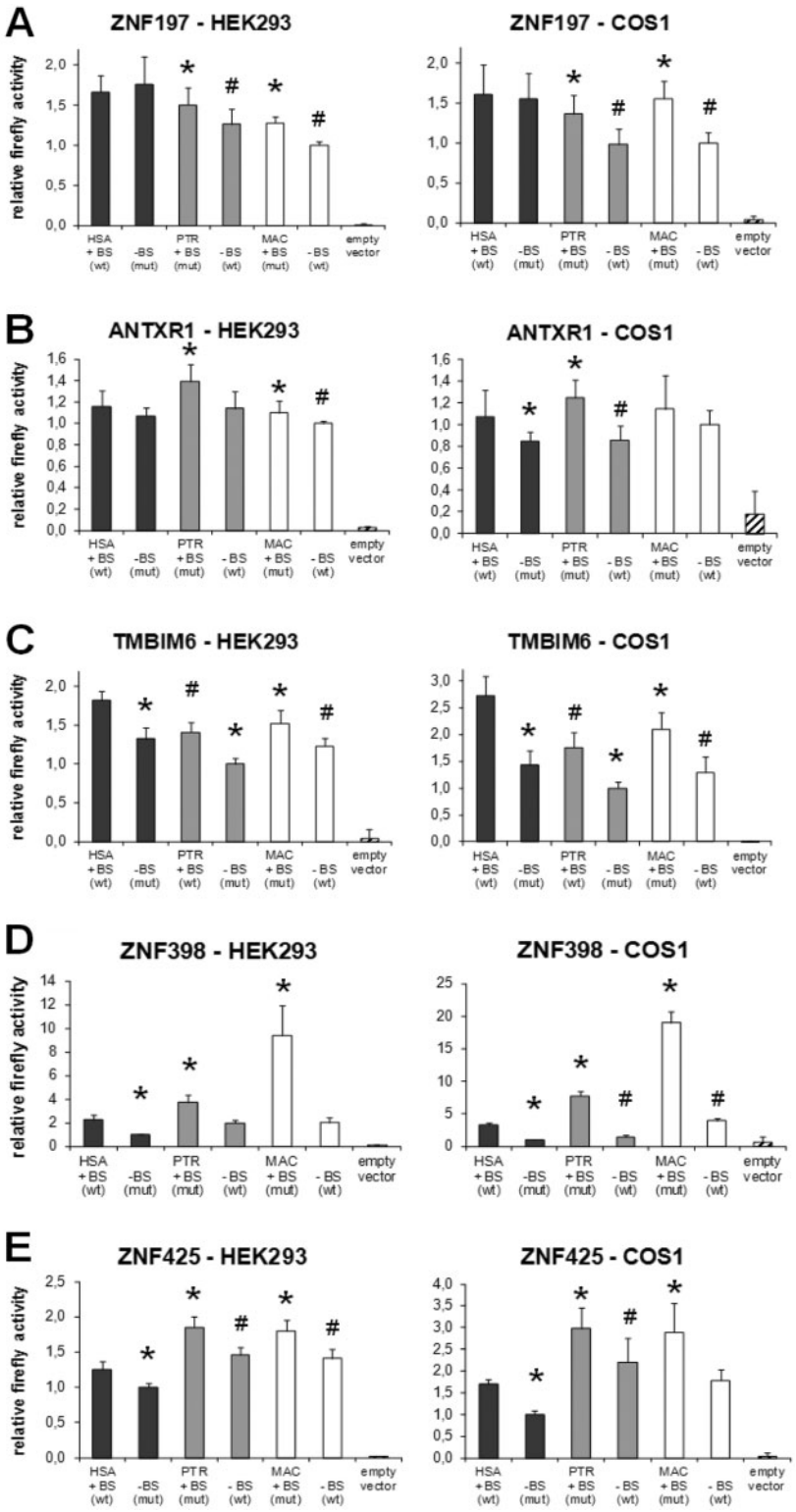

Fig. 5. Normalized *firefly* luciferase activities of human, chimpanzee, and rhesus macaque wt and mutated (mut) promoters. Bars represent average *firefly* to *Renilla* ratio in black for human (HSA), in gray for chimpanzee (PTR) and in white for macaque (MAC) promoters. For each species, the left bar refers to the wt and the right bar to the mutated promoter. (+binding site) or (−binding site) indicate presence or absence of GABPa binding sites in wt promoters and indicate introduction or disruption of sites in mutated promoters. For each promoter, measured activities were normalized to the construct with the lowest promoter activity level in HEK293T cells (set to one; supplementary table S16, Supplementary Material online). Standard errors were calculated from at least six replicates. (*) indicates significant differences between wt and mutated promoter activities according to a one-tailed Welch's *t*-test, while (#) indicates significant difference of wt chimpanzee or macaque promoters compared with human wt activity according to a two-tailed Welch's *t*-test (supplementary tables S17–S21, Supplementary Material online). The raw data for all constructs and the statistical significance are available in supplementary material (supplementary tables S17–S21, Supplementary Material online).

biogenesis, regulation of lipid metabolism, protein ubiquitination, neuron projection development and innate, and adaptive immune response, among others (supplementary table S12, Supplementary Material online).

In addition, we also analyzed 2,023 genes found to be DE after GABPa knock-down (72 h time point, see Materials and Methods) that did not presented GABPa binding sites in their promoter regions (supplementary table S11, Supplementary Material online). We found very similar GO groups enriched, suggesting that among the DE genes are not only primary targets of GABPa, but also a big group of secondary targets involved in similar processes. For instance, for both groups (DE genes with ChIP-Seq peaks and DE genes without ChIP-Seq peaks for GABPa) we found significant enrichments for biological processes associated with lipid and fat biosynthesis and metabolism such as lipoprotein transport, glycosphingolipid metabolic process, lipid biosynthetic process, fatty acid transmembrane transport, and regulation of fatty acids, mitochondrial biogenesis, among others (supplementary table S13, Supplementary Material online). In addition, we also found in all GO enrichment analyses an overrepresentation of terms related to tissue development, such as regulation of endothelial tube morphogenesis and endothelial cell proliferation, and epithelial cell differentiation involved in mammary gland alveolus development.

Analysis of DE Genes with Human-Specific GABPa Binding Sites

Of the 217 genes with human-specific binding sites, we found 52 that showed significant changes in gene expression 72 h after GABPa knock-down with both siRNA molecules (supplementary table S7, Supplementary Material online). We found an enrichment of genes with human-specific binding sites among the DE genes (Fisher's Exact test, P value $= 5.79^{-10}$). Four (*CEP170*, *RPL41*, *GUSBP4*, and *STAG3L4*) out of the six human and great ape specific genes having human-specific GABPa binding sites showed significant changes in gene expression after GABPa knock-down at least with one siRNA molecule (supplementary table S7, Supplementary Material online). This indicates that many newly evolved binding sites are functional.

Since GABPa has been associated with several human medical conditions (Bauer-Mehren et al. 2010), we explored if some of these 52 genes have also been associated with the very same altered phenotypes. We used the gene-disease association database "DisGeNET" (Bauer-Mehren et al. 2010, 2011), finding information for 17 genes, many of them have been associated with diabetes (*PCMT1*, *UGGT2*), Parkinson disease (*RPL6*, *TDP2*, *HSPA8*), and breast cancer (*ACOT13*, *ANTXR1*, *BAG4*, *EMG1*, *HSPA8*, *NEK7*, *YPEL3*, *ZNF398*), among other diseases (supplementary table S14, Supplementary Material online).

Discussion

By using ChIP-Seq and knock-down experiments, we found that GABPa regulates a substantial fraction of human genes. In total, we identified binding sites of GABPa in the vicinity of 5,321 genes. Out of them, 217 genes have gained human-specific GABPa binding sites. We found that GABPa binds to almost one-third (31%) of the promoters of genes expressed in HEK293T cells. Analyses of differential expression after GABPa knock-down allowed us to identify 1,215 genes with at least one GABPa binding site and significant changes in gene expression. This constitutes a set of strong candidates for being primary targets of GABPa. We also found that around 2,000 genes are likely to be secondary targets of GABPa. In general, our results indicated that GABPa is mainly involved in the regulation, direct or indirect, of genes that are important for neurological processes, lipid biosynthesis and metabolism, endothelial and epithelial cell morphogenesis, and mammary morphogenesis and development.

Newly Evolved GABPa Binding Sites Are Functional

To find binding sites that are specific to humans or hominids, we reconstructed the ancestral sequences for the 11,008 human GABPa binding sites in HEK293T cells based on the UCSC 44-vertebrate whole-genome alignments. This approach relies on the accuracy of the UCSC alignments. UCSC multiZ alignments of human–chimpanzee and human–macaque have been estimated to be problematic (while not necessarily wrong) for 0.004% and 0.02% of the aligned nucleotides, respectively (Prakash and Tompa 2007). Theoretically, this would imply that for the 11,008 human–macaque alignments corresponding each to 11bp of a GABPa binding site, a fraction of 24 nt was problematically aligned. However, this fraction is likely even smaller as most problematic alignments have been found in intronic and intergenic regions (Prakash and Tompa 2007). The majority of the GABPa binding sites reside in proximal promoter regions where mammalian genomic sequences are particularly conserved (Mahony et al. 2007), allowing for accurate overall alignments. We thus expect that our results were not significantly influenced by erroneous alignments.

To test whether the identified binding sites are functional, we selected four of them for experimental validation, three of them carrying human-specific, and one carrying a hominoid-specific GABPa binding site. We carried out dual luciferase reporter assays of human, chimpanzee, and macaque promoters in human HEK293T and African green monkey-derived COS-1 cells. The relative reporter activity was similar in both cell lines, thus experimentally indicating independence from the species-specific cellular background. Importantly, this also indicates that the activity of human-specific binding sites can be demonstrated even within the genomic/proteomic background of a catharine monkey cell line. The insertion of the human GABPa binding site into chimpanzee or rhesus macaque promoters resulted in significant increase in reporter gene expression in the majority of the cases. Conversely, disruption of GABPa binding sites in human and chimpanzee promoters led to a significant decrease of the reporter gene activity.

Mutation analyses for the promoters of *ZNF197* and *ANTXR1* showed that an insertion of a human-specific GABPa binding site into chimpanzee and macaque promoters resulted in a significant and consistent increase in reporter activity. In contrast, we did not observe a decrease in activity when disrupting the newly evolved binding sites in the human promoters of these two genes. In general, this

Promoter (length)	# of BS per peak	Species	Wild type (wt)	Mutated (mut)	Log2 ratios (mut/wt) HEK293	COS-1
ZNF398	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -1.16 ***	↓ -1.69 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.92 ***	↑ 2.45 ***
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 2.22 ***	↑ 2.28 ***
ZNF425	2	HSA	CTCGGAAGCG---GAAGCCG	CTCGGCAGCG---GAGGCCG	↓ -0.32 ***	↓ -0.76 ***
(329bp)	0	PTR	CTCGGcAGCG---GAcGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.34 ***	↑ 0.44 **
	0	MAC	CcCGGcAatGgctGggGCCG	CTCGGAAGCG---GAAGCCG	↑ 0.35 ***	↑ 0.69 ***
ZNF197	2	HSA	TGCCGGAAGGGC	TGCCGCAAGGGC	→ 0.08 ns	→ -0.04 ns
(430bp)	1	PTR	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.25 *	↑ 0.47 ***
	1	MAC	TGCCGcAAGGGC	TGCCGGAAGGGC	↑ 0.35 ***	↑ 0.64 ***
ANTXR1	3	HSA	GCGAGGAAGGGC	GCGAGGAGGGGC	↓ -0.11 ns	↓ -0.34 *
(633bp)	2	PTR	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.29 **	↑ 0.54 ***
	2	MAC	GCGAGGAgGGGC	GCGAGGAAGGGC	↑ 0.14 *	↑ 0.19 ns
TMBIM6	1 (2)	HSA	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.45 ***	↓ -0.93 ***
(576bp)	1 (2)	PTR	GAGCGGAAGTGG	GAGCGGACGTGG	↓ -0.49 ***	↓ -0.81 ***
	0 (1)	MAC	GAGCGGAcGTGG	GAGCGGAAGTGG	↑ 0.30 ***	↑ 0.70 ***

Fig. 6. The introduction and disruption of GABPa consensus binding sites significantly influence reporter gene activities. For each gene, the number of predicted binding sites within 200 bp surrounding the peak centers is indicated. Species are denoted by HSA, *Homo sapiens*; PTR, *Pan troglodytes* (chimpanzee); and MAC, *Macacca mulatta* (macaque). Sequences are shown for wt and mutated sites. Underlined bases indicate differences from the human wt sequence. Mutated bases are colored in green or red indicating generation or disruption of a GABPa binding site, respectively. Green arrows depict higher activity of mutated over wt promoter, red arrows indicate lower activity, and yellow arrows represent no change. Differences in mutated and wt promoter activities are given as log2 ratios of average luciferase to *Renilla* ratios. Significance levels, as determined by Welch's *t*-test for unequal variances, are indicated as (*) P value < 0.05, (**) P value < 0.01, (***) P value < 0.001, and (ns) not significant.

does not necessarily render the human-specific binding sites irrelevant; under different conditions these binding sites may still have a functional impact. Instead, our findings could indicate the presence of additional mutations that facilitate the binding of one or more factors to compensate for the activating property of the new GABPa binding sites. Indeed, both promoters (*ZNF197* and *ANTXR1*) harbor additional human-specific mutations within less than 100 bp to the newly evolved GABPa binding sites. This could illustrate a scenario where an increase in gene expression was beneficial at some time during evolution, with a subsequent period of decreased evolutionary pressure. In the case of *TMBIM6* the human wt promoter drives higher reporter activity compared to the chimpanzee wt promoter, even though both species share a GABPa consensus-binding site. However, the human promoter (including exon 1) harbors two additional single nucleotide mutations in very close proximity that might account for the observed difference (see fig. 4). Disruption of two overlapping GABPa binding sites in the bidirectional promoter of *ZNF398*/*ZNF425* resulted in greater than 2-fold activity reduction in the direction of *ZNF398* and greater than 1.2-fold reduction in the direction of *ZNF425*. To convert the rhesus macaque-binding site to the human binding site, it was necessary to introduce six mutations, accompanied by a 3 bp deletion. Interestingly, this strong intervention had only a moderate effect in the direction of *ZNF425*. Conversely, in the direction of *ZNF398*, we observed an almost 5-fold increase in activity. This might indicate context-specificity, in other words, which gene is affected depends on the tissue or cell type.

Notably, in a previous study where GABPa binding sites were introduced into six promoters that are not regulated by GABPa in their wt form, only one promoter was activated (Collins et al. 2007). This may indicate that the introduction of GABPa binding sites per se is mostly insufficient to affect gene expression. However, we found that by inserting human-specific GABPa binding sites into chimpanzee and rhesus macaque promoter regions, the reporter gene activity increased consistently. Our results thus support the conclusion that binding sites need to be placed in the right genomic context to exert an influence on gene expression.

Our knock-down experiments further revealed that 1,215 of the genes with at least one GABPa binding site changed in expression after GABPa knock-down. Interestingly, genes with human-specific binding sites were enriched among the genes with differential expression after knock-down. This indicates that at least a representative number of the predicted newly evolved binding sites contribute to transcriptional regulation by GABPa in

humans, including *ANTXR1* and *ZNF398*, which we had also tested in our promoter assays.

Human-Specific GABPa Binding Sites Regulate Genes That Are Potentially Important for Human Evolution and Human Diseases

Among the genes with GABPa binding sites specific to the human lineage, we found a significant overrepresentation of KRAB-ZNF TFs genes. KRAB-ZNFs, a relatively young and fast evolving family of TF genes with several primate-specific family members (Hamilton et al. 2006; Huntley et al. 2006; Nowick et al. 2010), seem prone to acquire new GABPa binding sites. KRAB-ZNF genes play a role in a number of biological processes, including development and brain functions (Najmabadi et al. 2011; Zhang et al. 2011) and have been considered as excellent candidates for contributing to postzygotic isolation and other barriers that could lead to speciation processes (Nowick et al. 2013). Consequently, it would be interesting to further investigate how GABPa, presumably via regulating the expression of KRAB-ZNF genes expressed in brain cells, has contributed to the evolution of human-specific phenotypes.

Among the genes with human-specific GABPa binding sites, we also found enrichment for genes involved in RNA processing, especially of tRNAs, and signaling pathways involved in ventral spinal cord interneuron and spinal cord motor neuron cell fate specification and dorsal/ventral neural tube patterning (supplementary table S10, Supplementary Material online). For instance, the promoters of *DARS* and *PARS2*, genes that encode for aminoacyl tRNA synthetases (aspartyl- and prolyl-tRNAs) and catalyze the attachment of amino acids to their cognate tRNAs, carry human-specific GABPa binding sites. Both of these genes were DE with at least one siRNA molecule. Functional changes in these two aminoacyl tRNA synthetases have been related with neuronal disorders of the peripheral nervous system, amyotrophic lateral sclerosis, and ataxia (Park et al. 2008), brain stem and spinal cord leukoencephalopathy with cerebellar and dorsal column dysfunctions (Sissler et al. 2007; Taft et al. 2013), and Alpers syndrome (Sofou et al. 2015).

ALDOA, *HSPA8*, and *TP73* are three further examples of genes with human-specific GABPa binding sites that showed significant changes in gene expression after GABPa knock-down. They, as well as *TMBIM6*, which we have tested with our reporter assays, have been associated with cognitive diseases such as autism, Alzheimer's disease (Harris et al. 2007; Kumar et al. 2009), and Parkinson's disease (Naidoo 2009; Lauterbach 2013), control of neocortical regionalization and brain pathologies like loss of Cajal-Retzius cells (Meyer et al. 2004), abnormal accumulation of cerebrospinal fluid in the brain (Yang et al. 2000), and neuron apoptosis (Pozniak et al. 2002). In agreement with our results, it has been shown that expression changes in *GABPa* might have consequences for mitochondrial functions in patients with Alzheimer's disease (Sheng et al. 2012).

Genes that harbor human-specific GABPa binding sites and that significantly change their expression after GABPa knock-down have also been associated with the same medical conditions with which GABPa has been associated. For instance, protein-L-isoaspartate (D-aspartate) O-methyltransferase (*PCMT1*), UDP-glucose glycoprotein glucosyltransferase 2 (*UGGT2*), as well as *TMBIM6*, have been linked to diabetes types I and II. *PCMT1* encodes the Protein-L-isoaspartate (D-aspartate) O-methyltransferase, a repair enzyme for which reduced expression levels have been associated with delaying the appearance and decreasing the severity of diabetes type I (Wägner et al. 2007, 2008). Similarly, patients with diabetes type II showed significantly lower *UGGT2* expression (Marchetti et al. 2007). Noteworthy, we also found enrichments for processes associated with triglyceride biosynthesis and cholesterol transport, both being connected with the occurrence and prevalence of several human metabolic diseases, for example, abnormal blood sugar levels. Thus, we suggest that human-specific GABPa binding sites might be associated with human-specific changes in metabolic functions, potentially increasing the risk for developing diabetic phenotypes and other metabolic syndromes.

It was previously demonstrated that GABPa is associated with the transcriptional regulation of genes directly involved in controlling cell migration in breast epithelial cells (Odrowaz and Sharrocks 2012). Interestingly, among the genes with human-specific binding sites, we found enrichment of genes involved in mammary gland development (supplementary table S10, Supplementary Material online). For instance, *PHB2*, also known as Repressor of Estrogen Receptor Activity (*REA*), is important for mammary gland cell proliferation, breast alveolus development, and postnatal breast development (Mussi et al. 2006). PHB2 has a human-specific GABPa binding site 26 bp downstream the TSS and shows significant change in expression after GABP knock-down. Within this study, and given the uniqueness of the human breast, this particular human-specific GABPa binding site is the best candidate for a functional contribution to this human trait.

Conclusions

In summary, we identified human-specific TFBSs for GABPa using ChIP-Seq data together with comparative genomic analysis and experimentally demonstrated the functionality of selected newly evolved GABPa binding sites. Our results depict a scenario that may connect gene regulation by GABPa to human speciation and to shaping human-specific phenotypes, including brain, breast, and metabolic functions among others. Our work can also serve as an example for using a bioinformatics approach with the increasing number of publicly available ChIP-Seq data for TFs to study binding site evolution on a "TF-ome"-wide level.

Materials and Methods

Chromatin Immunoprecipitation-Sequencing

We performed ChIP-Seq according to a published protocol (Warnatz et al. 2011). Briefly, 5×10^7 HEK293T cells were cross-linked for 10 min at room temperature with 1%

formaldehyde, nuclei were prepared following the published protocol and chromatin was sheared to 100–500 bp size by 45 cycles of 30 s on/off at the highest amplitude using a Bioruptor water bath sonicator (Diagenode). Nuclear extracts were immunoprecipitated with 10 μg rabbit anti-GABP-α (H-180X, Santa Cruz Biotechnology sc-22810) and 70 μl Protein G-Dynabeads (Invitrogen). After washing of beads, protein-DNA complexes were eluted, crosslinks were reversed overnight, and DNA was purified according to the manufacturer's protocol. For sequencing library preparation, 2 ng ChIP DNA and 10 ng Input DNA were subjected to end-repair, addition of adenine bases and ligation of sequencing adapters, followed by DNA amplification through polymerase chain reaction (PCR) and subsequent gel purification for sequencing on an Illumina Genome Analyzer GAII according to the manufacturer's protocol for 36 bp reads. Reads were aligned to the human genomic sequence (hg18) using Eland, resulting in 6,955,499 GABPa ChIP reads with unique match to the genome (allowing up to two mismatches) and 2,948,346 corresponding reads from the input DNA. A replicate ChIP-seq experiment performed later for validation of the initial experiment resulted in 16,856,422 GABPa ChIP reads and 26,104,399 reads from the input DNA.

Peak Calling, Gene Mapping, MEME, and MAST Analysis

ChIP-Seq reads were analyzed in three steps as published previously (Valouev et al. 2008). To find regions of high sequencing read density (peaks) within 6.96 million reads from GABPa ChIP-Seq, we used the peak-calling algorithm available in QuEST. In total, we found 6,208 genomic peaks for GABPa (fig. 1*B* and supplementary table S15, Supplementary Material online). From a replicate experiment with 16.86 million ChIP-Seq reads, we identified 8,311 genomic peaks of GABPa, which were used for calculation of the overlap of regions among replicates and for calculation of the IDR as described before (Landt et al. 2012). The 6,208 peaks from the first replicate were mapped to all UCSC known transcripts that start within ±5 kb of the peak (hg19). For annotation we used 82,961 UCSC known transcripts (Hsu et al. 2006) corresponding to 23,460 Entrez genes. UCSC transcript IDs were converted to Entrez gene IDs using UCSCs knownToLocusLink table. 1,116 peaks were mapped to UCSC transcripts that do not correspond to Entrez genes (supplementary table S3, Supplementary Material online). A majority of these correspond to partial novel nucleic acids, CDS, human cDNAs, among others. After extraction of peak-associated sequences comprising 200 bp surrounding each peak center via the UCSC table browser (Karolchik et al. 2004), we applied MEME (Bailey et al. 2009) to identify overrepresented motifs. 97% of the 6,208 peaks contributed to the motif identified by MEME (supplementary table S1, Supplementary Material online). Using default parameters, MEME assumes that each sequence contains zero or one motif. The derived PWM was then used to run the MEME tool MAST that reports the occurrences of PWM hits for each sequence in the input set at a particular stringency (set to $P = 0.001$) (fig. 1*G* and supplementary table S2, Supplementary Material online).

Multiple Species Alignment Extraction and Conversion

UCSC provides the 44 way-vertebrate alignments in a multiple alignment format (MAF) that consists of short blocks (1–200 bp) of multiple alignments that can be concatenated. We extracted the alignments corresponding to GABPa binding sites within the ChIP peak regions via UCSCs table browser (Karolchik et al. 2004) and converted the MAF-formatted alignments into the commonly used FASTA format, while excluding nonsyntenic blocks and species with missing sequence data (e.g., insertions not included in the MAF alignments; supplementary table S4, Supplementary Material online).

Ancestral Sequence Reconstruction

Ancestral sequences were calculated using ANCESTORS (Blanchette et al. 2004) obtained from http://ancestor.bioinfo.uqam.ca/programs/anc.tar. The program requires a multiple species alignment and a phylogenetic tree including branch lengths. Therefore, we reconstructed the ancestral sequences along with the phylogeny of 34 mammalian species from the UCSC 44-vertebrate alignments (fig. 1*F*) (Diallo et al. 2007). The approach implemented in ANCESTORS is suitable for reconstructing ancestral sequences, including the most likely scenario of insertions and deletions observed in alignments while retaining an extremely high degree of accuracy (Diallo et al. 2007). To calculate branch lengths, all alignments were concatenated and run through BASEML, a maximum likelihood-based program of the PAML package (Yang 2007). The nucleotide substitution model HKY was used in both programs. Phylogeny was taken from UCSC (phyloP44wayPlacMammal) available at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/vertebrate.mod.

Cloning and Plasmid Preparation

PCR primers were designed using the Primer3 online service and extended by 29 bp Gateway attB tails (Invitrogen) at the 5′-end of each primer. Touch-down PCR was performed as described previously (Ralser et al. 2006), except for the supplementation of each reaction with 0.001U *Pfu* polymerase. Mutations were introduced by primer-mediated mutagenesis. To facilitate cloning, the Gateway cloning cassette (Invitrogen) was amplified with the forward primer attP1 and the reverse primer attP2 and cloned into the pGL3 reporter vector (Promega). PCR products were purified and cloned upstream of the luciferase gene in the modified pGL3 vector using BP Clonase II Enzyme Mix (Invitrogen) following the manufacturer's instructions. Plasmids were transformed into the methylation-deficient Dam⁻ *Escherichia coli* strain GM2929. Inserts of positive clones were validated using capillary Sanger sequencing (Services in Molecular Biology, Berlin, Germany). DNA concentration was measured on a Nanodrop UV spectrophotometer

(NanoDrop Technologies) and standardized to 50 ng/μl for transfections.

Cell Culture, Transient Transfection, and Reporter Gene Activity Assays

HEK293T and COS-1 cells were cultivated in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 100 U/ml penicillin/G-streptomycin (Biochrom) and 10% heat-inactivated fetal bovine serum (Biochrom) at 37 °C and 5% CO_2. We seeded approximately 15,000 (HEK293T) and approximately 5,000 (COS-1) cells per well in clear-bottom 96-well plates (Costar). Twenty-four hours after seeding, we cotransfected 150 ng of experimental firefly luciferase plasmid together with 10 ng of *Renilla* luciferase control plasmid (pRL-TK, Promega) in five replicates using Lipofectamine 2000 following the manufacturer's recommendations. Cells were lyzed 24 h posttransfection. We measured firefly luciferase and *Renilla* luciferase activities using the Centro LB960 luminometer (Berthold) and the Dual Luciferase Kit (Promega). We followed the protocol suggested by the manufacturer with the exception of injecting 25 μl each of the firefly luciferase and *Renilla* luciferase substrate reagents. All measurements were performed at least in three technical and two biological replicates, including new dilution and concentration adjustments of reporter plasmids.

Inhibition of GABPa Expression by RNA Interference

RNA interference experiments for GABPa were performed using two independent types of silencing molecules against GABPa, namely one unmodified synthetic small interfering RNA (Qiagen SI00423311) and one chemically modified synthetic small interfering RNA (Invitrogen HSS103907). HEK293T cells were seeded in 12-well plates together with siRNA-HiPerFect complexes according to the HiPerFect fast forward protocol (Qiagen). For the mock transfections, cells were treated with HiPerFect reagent only. Knock-down transfections were performed in triplicates; mock transfections were performed in quadruplicates. Total RNA was extracted from cultured cells at 24 and 72 h posttransfection using the RNeasy mini kit (Qiagen) following the manufacturer's instructions. All RNA samples were DNase-treated, purified, quantified, and inspected for integrity. For hybridizations on microarrays, biotinylated cRNA was synthesized using the GeneChip expression 3′-amplification one-cycle target labeling and control reagents (Affymetrix). Following integrity control, the cRNA was hybridized to the Affymetrix GeneChip HG-U133Plus2. The arrays were washed, stained, and scanned following the recommended protocols from Affymetrix.

Gene expression was analyzed using the "affy" package from Bioconductor (Gautier et al. 2004). In particular, we determined Robust Multi-Array Robust Multi-array Average (RMA) normalized expression values for each probeset (Bolstad et al. 2003), filtered for probe sets that were not reliably detected in any sample (detection P value >0.05), and combined rma values of probesets referring to the same gene into one mean expression value as described before (Nowick et al. 2009). We determined genes that were DE between knock-down and mock transfected samples for each time point using the package "multtest" (Pollard et al. 2005) implemented in Bioconductor. We then overlapped the DE genes of each time point with the candidate genes from our ChIP-Seq experiments, obtaining 392 and 1,215 genes overlapping for 24 and 72 h experiments, respectively. Permutation tests indicated that the observed overlap is higher than expected by chance for the 72 h time point (P value $= 0.001$), conversely to the results obtained for 24 h treatment (P value $= 0.137$). This suggests that for the 72 h treatment we obtained more genes than we would expect by chance. Consequently, we chose the data of the 72 h treatment for further analyses.

GO Enrichment Analyses

In total, we performed three GO analyses using the hyper geometric test implemented in FUNC (Prüfer et al. 2007). In the first analysis, we tested for enrichment of GO groups in the 217 genes located ±5 kb around the human-specific GABPa TFBS compared to the genes with at least one ChIP-Seq peak in the promoter region (± 5) (supplementary tables S6 and S11, Supplementary Material online). In the second analysis, we assessed enrichment among the set of genes that were DE after GABPa knock-down with both siRNA molecules (Qiagen: SI00423311 and Invitrogen: HSS103907) and had at least one GABPa binding site in our ChIP-Seq data compared with the set of DE genes not having ChIP-Seq peaks in the promoter region (supplementary tables S11 and S12, Supplementary Material online). In the third test, we analyzed whether DE genes after GABPa knock-down and without a GABPa binding site in our ChIP-Seq data were enriched for particular GO groups (supplementary tables S11 and S13, Supplementary Material online). We followed the enrichment tests by refinement of the enriched GO groups as implemented in FUNC (Prüfer et al. 2007) applying P value cutoffs after refinement of $P < 0.05$ for the first and third tests test, and $P < 0.01$ for the second one.

Supplementary Material

Supplementary figures S1–S5 and tables S1–S21 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Departamento Administrativo de Ciencia, Tecnología e Innovacioín Colciencias from Colombia, call Francisco José de Caldas 497/2009 (A.P-S.), by the Volkswagen Foundation within the initiative "Evolutionary Biology" (K.N.), by the Max Planck Society, and by the European Union under its Sixth Framework Program (AnEUploidy [LSHG-CT-2006-037627] and APES [NEST-2004-Path-HUM-28594]). The authors thank Lutz Walter for providing primate DNA samples, Martin Lange for providing the cloning vector, and Tatiana Borodina for assistance in the preparation of ChIP-sequencing libraries. Special thanks to Paz Polak and Andrew Hufton for continuous discussion and support.

References

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 37:W202–W208.

Batchelor AH, Piper DE, de la Brousse FC, McKnight SL, Wolberger C. 1998. The structure of GABPalpha/beta: an ETS domain- ankyrin repeat heterodimer bound to DNA. *Science* 279:1037.

Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. 2011. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One* 6:e20284.

Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. 2010. DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics* 26:2924–2926.

Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*. 14:2412–2423.

Bolstad BM, Irizarry RA, Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.

Brinkman-Van der Linden ECM, Hurtado-Ziola N, Hayakawa T, Wiggleton L, Benirschke K, Varki A, Varki N. 2007. Human-specific expression of siglec-6 in the placenta. *Glycobiology* 17:922–931.

Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol*. 46:111–138.

Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci*. 368:20130025.

Chabot A, Shrit RA, Blekhman R, Gilad Y. 2007. Using reporter gene assays to identify cis regulatory differences between humans and chimpanzees. *Genetics* 176:2069–2076.

Collins PJ, Kobayashi Y, Nguyen L, Trinklein ND, Myers RM. 2007. The ets-related transcription factor GABP directs bidirectional transcription. *PLoS Genet*. 3:e208.

Diallo AB, Makarenkov V, Blanchette M. 2007. Exact and heuristic algorithms for the indel maximum likelihood problem. *J Comput Biol*. 14:446–461.

Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics* 20:307–315.

Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, Stubbs L. 2006. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res*. 16:584–594.

Harris SE, Fox H, Wright AF, Hayward C, Starr JM, Whalley LJ, Deary IJ. 2007. A genetic association analysis of cognitive ability and cognitive ageing using 325 markers for 109 genes associated with oxidative stress or cognition. *BMC Genet*. 8:43–43.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39:1140–1144.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* 22:1036–1046.

Huby T, Dachet C, Lawn RM, Wickings J, Chapman MJ, Thillet J. 2001. Functional analysis of the chimpanzee and human apo(a) promoter sequences: identification of sequence variations responsible for elevated transcriptional activity in chimpanzee. *J Biol Chem*. 276:22209–22214.

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res*. 16:669–677.

Jaworski A, Smith CL, Burden SJ. 2007. GA-binding protein is dispensable for neuromuscular synapse formation and synapse-specific gene expression. *Mol Cell Biol*. 27:5040–5046.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC table browser data retrieval tool. *Nucleic Acids Res*. 32:D493–D496.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.

Kumar RA, Karamohamed S, Sutcliffe JS, Cook EH, Geschwind DH, Dobyns WB, Scherer SW, Christian SL, Marshall CR, Badner JA, et al. 2009. Association and mutation analyses of 16p11.2 autism candidate genes. *PLoS One* 4(2): e4582.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 22:1813–1831.

Lauterbach EC. 2013. Psychotropics regulate Skp1a, Aldh1a1, and Hspa8 transcription —potential to delay Parkinson's disease. *Prog NeuroPsychopharmacol Biol Psychiatry*. 40:236–239.

Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, Myers RM, Weng Z. 2007. Transcription factor binding and modified histones in human bidirectional promoters. *Genome Res*. 17:818–827.

Mahony S, Corcoran DL, Feingold E, Benos PV. 2007. Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome. *Genome Biol*. 8:R84.

Marchetti P, Bugliani M, Lupi R, Marselli L, Masini M, Boggi U, Filipponi F, Weir GC, Eizirik DL, Cnop M. 2007. The endoplasmic reticulum in pancreatic beta cells of type 2 diabetes patients. *Diabetologia* 50:2486–2494.

Meyer G, Socorro AC, Garcia CGP, Millan LM, Walker N, Caput D. 2004. Developmental roles of p73 in cajal-retzius cells and cortical patterning. *J Neurosci*. 24:9878–9887.

Molineris I, Grassi E, Ala U, Di Cunto F, Provero P. 2011. Evolution of promoter affinity for transcription factors in the human lineage. *Mol Biol Evol*. 28:2173–2183.

Mussi P, Liao L, Park S-E, Ciana P, Maggi A, Katzenellenbogen BS, Xu J, O'Malley BW. 2006. Haploinsufficiency of the corepressor of estrogen receptor activity (REA) enhances estrogen receptor function in the mammary gland. *Proc Natl Acad Sci U S A*. 103:16716–16721.

Naidoo N. 2009. ER and aging—protein folding and the ER stress response. *Ageing Res Rev*. 8:150–159.

Najmabadi H, Jamali P, Zecha A, Mohseni M, Püttmann L, Vahid LN, Jensen C, Moheb LA, Bienek M, Larti F, et al. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 478:57–63.

Nowick K, Carneiro M, Faria R. 2013. A prominent role of KRAB-ZNF transcription factors in mammalian speciation? *Trends Genet* 29(3):130–139.

Nowick K, Gernat T, Almaas E, Stubbs L, Robinson GE. 2009. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A*. 106:22358–22363.

Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol*. 27:2606–2617.

Odrowaz Z, Sharrocks AD. 2012. The ETS transcription factors ELK1 and GABPA regulate different gene networks to control MCF10A breast epithelial cell migration. *PLoS One* 7:e49892.

Park SG, Schimmel P, Kim S. 2008. Aminoacyl tRNA synthetases and their connections to disease. *Proc Natl Acad Sci U S A*. 105:11043–11049.

Pollard KS, Dudoit S, van der Laan MJ. 2005. Multiple testing procedures: the multtest package and applications to genomics. In: R Gentleman, V Carey, W Huber, R Irizarry, S Dudoit, editors. *Bioinformatics and computational biology solutions using R and bioconductor*. New York: Springer. p. 249–271.

Pozniak CD, Barnabe-Heider F, Rymar VV, Lee AF, Sadikot AF, Miller FD. 2002. p73 is required for survival and maintenance of CNS neurons. *J Neurosci*. 22:9800.

Prabhakar S, Afzal V, Pennacchio LA, Rubin EM, Noonan JP, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, et al. 2008. Human-

specific gain of function in a developmental enhancer. *Science* 321:1346–1350.
Prakash A, Tompa M. 2007. Measuring the accuracy of genome-size multiple alignments. *Genome Biol*. 8:R124.
Prüfer K, Muetzel B, Do H-H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 8:41.
Ralser M, Querfurth R, Warnatz HJ, Lehrach H, Yaspo ML, Krobitsch S. 2006. An efficient and economic enhancer mix for PCR. *Biochem Biophys Res Commun*. 347:747–751.
Ristevski S, O'Leary DA, Thornell AP, Owen MJ, Kola I, Hertzog PJ. 2004. The ETS transcription factor GABPα is essential for early embryogenesis. *Mol Cell Biol*. 24:5844–5849.
Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol*. 3:e387.
Romanelli MG, Lorenzi P, Sangalli A, Diani E, Mottes M. 2009. Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5′ flanking region. *Genomics* 93:227–234.
Rosmarin AG, Resendes KK, Yang Z, McMillan JN, Fleming SL. 2004. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol Dis*. 32:143–154.
Sheng B, Wang X, Su B, Lee H, Casadesus G, Perry G, Zhu X. 2012. Impaired mitochondrial biogenesis contributes to mitochondrial dysfunction in Alzheimer's disease. *J Neurochem*. 120:419–429.
Sissler M, Andel RJV, Scheper GC, Krageloh-Mann I, Uziel G, Coster R, Muravina TI, Bugiani M, Pronk JC, Florentz C, et al. 2007. Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation. *Nat Genet*. 39:534–539.
Sofou K, Kollberg G, Holmström M, Dávila M, Darin N, Gustafsson CM, Holme E, Oldfors A, Tulinius M, Asin-Cayuela J. 2015. Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. *Mol Genet Genomic Med*. 3:59–68.
Stedman HH, Mitchell MA, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428:415–418.
Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321:956–960.
Taft RJ, Vanderver A, Leventer RJ, Damiani SA, Simons C, Grimmond SM, Miller D, Schmidt J, Lockhart PJ, Pope K, et al. 2013. Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. *Am J Hum Genet*. 92:774–780.
Taylor MS, Massingham T, Hayashizaki Y, Carninci P, Goldman N, Semple CAM. 2008. Rapidly evolving human promoter regions. *Nat Genet*. 40:1262–1263.
Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*. 5:829–834.
Wägner AM, Cloos P, Bergholdt R, Boissy P, Andersen TL, Henriksen DB, Christiansen C, Christgau S, Pociot F, Nerup J. 2007. Post-translational protein modifications in type 1 diabetes: a role for the repair enzyme protein-l-isoaspartate (d-aspartate) O-methyltransferase? *Diabetologia* 50:676–681.
Wägner AM, Cloos P, Bergholdt R, Eising S, Brorsson C, Stalhut M, Christgau S, Nerup J, Pociot F. 2008. Posttranslational protein modifications in type 1 diabetes–genetic studies with PCMT1, the repair enzyme protein isoaspartate methyltransferase (PIMT) encoding gene. *Rev Diabet Stud*. 5:225–231.
Warnatz H-J, Schmidt D, Manke T, Piccini I, Sultan M, Borodina T, Balzereit D, Wruck W, Soldatov A, Vingron M, et al. 2011. The BTB and CNC homology 1 (BACH1) target genes are involved in the oxidative stress response and in control of the cell cycle. *J Biol Chem*. 286:23521–23532.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*. 8:206–216.
Yang A, Walker N, Bronson R, Kaghad M, Oosterwegel M, Bonnin J, Vagner C, Bonnet H, Dikkes P, Sharpe A, et al. 2000. p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* 404:99–103.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591. 10.1093/molbev/msm088
Yang Z-F, Drumea K, Cormier J, Wang J, Zhu X, Rosmarin AG. 2011. GABP transcription factor is required for myeloid differentiation, in part, through its control of Gfi-1 expression. *Blood* 118:2243–2253.
Yu M, Yang XY, Schmidt T, Chinenov Y, Wang R, Martin ME. 1997. GA-binding protein-dependent transcription initiator elements. Effect of helical spacing between polyomavirus enhancer a factor 3(PEA3)/Ets-binding sites on initiator activity. *J Biol Chem*. 272:29060–29067.
Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol*. 9:e1001179.
Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*. 8:e1000494.