Published in final edited form as:

Neurogenetics. 2014 March ; 15(1): 65–75. doi:10.1007/s10048-013-0388-3.

Genetic determinants of neuroglobin transcription

R Wang¹, E Halper-Stromberg², M Szymanski-Pierce², SS Bassett¹, and D Avramopoulos^{1,2}

¹Department of Psychiatry, Johns Hopkins University, 733 North Broadway, MRB-507, Baltimore, MD, 21205

²McKusick Nathans Institute of Genetic Medicine Johns Hopkins University, 733 North Broadway, MRB-507, Baltimore, MD, 21205

Abstract

Neuroglobin (NGB) is a neuron specific vertebrate globin shown to protect against hypoxia, ischemia, oxidative stress and the toxic effects of Amyloid-beta. Following on our and others' results highlighting the importance of *NGB* expression in disease we searched for genetic determinants of its expression. We found that a microRNA expressed with the *NGB* transcript shows significant target enrichments in the angiogenesis pathway and the Alzheimer disease / presenilin pathway. Using reporter constructs we identified potential promoter/enhancer elements between the transcription start site and 1142 bp upstream. Using 184 post mortem temporal lobe samples we replicated the reported negative effect of age, and after genotyping tagging SNPs we found one (rs981471) showing a significant correlation with the gene's expression and another (rs8014408) showing an interaction with age, the rare C allele being correlated with higher expression and faster decline. The two SNPs are towards the 3' end of *NGB* within the same LD block, 52 Kb apart and modestly correlated ($r^2 = 0.5$). Next generation sequencing of the same 184 temporal lobe samples and 79 confirmed AD patients across the entire gene region (including >12 Kb on the 3' and 5' flank) revealed limited coding variation, suggesting purifying selection of *NGB*, but did not identify regulatory or disease associated rare variants. A dinucleotide repeat in intron 1 with extensive evidence of functionality showed interesting but inconclusive results, as it was not amenable to further molecular analysis.

Keywords

Neuroglobin; genetics; gene expression; promoter; transcription; polymorphism; Alzheimer's disease; stroke

INTRODUCTION

Neuroglobin is a relatively new addition to the globin family that has received significant research interest. It is one of the four members of the family of heme binding metalloproteins, together with hemoglobin, myoglobin and cytoglobin. Hemoglobin is the best-known member and it is the main carrier of oxygen from the lungs to the periphery with near exclusive occupancy of red blood cells and the ability to reversibly bind oxygen. Myoglobin is also well studied and was the first protein to have its structure revealed by crystallography in 1957 (http://www.nsf.gov/news/news_summ.jsp?cntn_id=100689) [1]

Correspondence to: Dimitrios Avramopoulos, MD, PhD, McKusick Nathans Institute of Genetic Medicine and Department of Psychiatry, Johns Hopkins University, school of Medicine, 733 North Broadway, MRB-507, Baltimore, MD, 21205, Tel. 410 955-8323, Fax 410 955-7397, adimitr1@jhmi.edu.

The authors have no conflicts of interest to report.

while the gene was cloned in 1982 [2]. It is present mainly in skeletal muscle fibers and cardiomyocytes, and facilitates oxygen transport but also modulates the homeostasis of nitric oxide [3]. Neuroglobin was discovered by Burmester et al [4] in 2000 followed closely by the identification of cytoglobin by the same group [5]. While cytoglobin seems to be expressed in most human tissue types, neuroglobin has a more specific expression pattern, highest in the brain and retina, organs with exceptionally high oxygen demands. It is still unclear whether oxygen supply is a significant function of NGB. It maintains the ability to reversibly bind O₂ despite the relatively low sequence similarity with the other family members [4, 6], yet whether its biochemical properties support a role in O₂ storage and supply remains controversial. Trent et al [6] studied the kinetics of NGB and found that under known physiological conditions its affinity is too high to bind and release oxygen as needed for a function in transport. They proposed other functions that have been described for globins binding oxygen with high affinity, such as sensing of CO, NO, or O₂ or scavenging NO or O₂.

More consistent than the results supporting a role in oxygen supply are those that show neuroglobin to promote neuron survival. *NGB* overexpression in human neuronal cell cultures subjected to hypoxia results in enhanced cell survival as well as induction of *NGB* mRNA [7–10] although the latter has not been shown consistently [11, 12]. In vivo experiments in rats have shown Neuroglobin to protect against experimental stroke [13], a result also observed in mice [14], while NGB immunoreactivity has been found increased in the peri-infarct region in human stroke patients [15]. *NGB* overexpression has also been shown to protect against oxidative stress [16, 17] and to inhibit the intrinsic pathway of apoptosis by interacting with cytochrome c and preventing the activation of pro-caspase 9 [18]. The possible role of NGB in preventing apoptosis has been recently reviewed by Brittain et al [19]. At the cellular level NGB is localized in the cytosol near the mitochondria [20, 21], which is interesting given the roles of the mitochondria in oxidative phosphorylation and apoptosis.

NGB has been directly connected to Alzheimer's disease (AD). Li et al have shown it to protect against amyloid beta toxicity [22] and double-transgenic mice overexpressing NGB and mutant APP had a significantly attenuated AD phenotype. This was further supported by our data showing a genetic association and higher levels of *NGB* transcript in humans with AD [23] and data showing increased NGB protein in AD brains [24], suggesting that the gene might be induced by the disease and NGB levels might be a significant determinant of the risk to develop AD. In addition to this, our and others' genome wide data support overlap between genes changing expression with age and genes that change expression in AD [25, 26] suggesting that change with age might be in part responsible for the increased AD risk with age.

Neuroglobin is a highly phylogenetically conserved protein with 94% amino acid identity between human and mouse, and greater than 50% identity with Zebrafish (Figure 1). The *NGB* gene (Figure 2) is a small gene located on chromosome 14q24.3 and transcribed from the reverse strand. It spans 5.8 Kb and includes four exons with no known alternative splice forms. NGB shares less than 21% amino acid identity with myoglobins and less than 25% with hemoglobins [4]. The combination of divergence from the family and strict purifying selection pressure suggests that neuroglobin might have evolved to fill an important biological niche that no longer tolerates significant variation.

Only a few genetic studies have been reported for *NGB*. A study from our group showed a genetic association of *NGB* with AD [23] together with a reduction in expression with age and higher expression in AD which has now been replicated [24]. Shortly after Lin et al [27] reported variants in *NGB* associated with ischemic stroke, followed by another group

showing that the allele we reported as protective for AD also has a protective effect against traumatic brain injury [28]. In the present work we attempt an in depth search for genetic sequences and variants regulating *NGB* transcription. We employ bioinformatics, in vitro reporter assays, post mortem brain RNA measurements, quantitative trait mapping, and next generation sequencing on 263 individuals across the entire gene. Our results are an important starting point for future work on the regulation of *NGB*, which is likely to be of importance for brain health.

MATERIALS AND METHODS

Promoter constructs

We generated 9 constructs (shown in Figure 1) by introducing human sequences upstream of the *NGB* translation start site and as long as 2,056 bp, into the pGL3-Basic plasmid vector which is a firefly luciferase reporter vector (Promega). First human fragments were amplified using PfuTurbo® Polymerase (Stratagene) high fidelity Taq polymerase and restriction enzyme sites were added at both ends of the amplicons by inclusion in the primers. A' overhangs were then added by incubating the products for five minutes with Taq DNA polymerase (Invitrogen) and excess dATP. The amplicons were TA cloned into pCR®2.1 -TOPO vector (Invitrogen, Cat# K451020SC). The inserts were then excised and ligated into the pGL3-Basic plasmid by the specific restriction enzyme sites. We verified all constructs by Sanger sequencing using the commercial primer RVprimer3 (Promega) and pGLR 5'. The sequence coordinates of each of the constructs are reported on Table 1. To make the two constructs that include non-contiguous DNA sequences we used restriction enzyme digestion followed by ligation, selection of the appropriate size fragments of ligated products and verification by nucleotide sequencing.

Dual Luciferase reporter assays

Human Embryonic Kidney cells (HEK-293, ATCC® CRL-11268™), human neuroblastoma cells (SK-N-SH, ATCC no. HTB-11) and mouse brain neuroblastoma cells (Nuero 2A, ATCC® CCL-131™) were grown in ATCC suggested medium for each cell type (<http://www.atcc.org>) without antibiotics. Approximately 0.8×10^5 cells were plated 24 hours before transfection in order for cells to reach 90–95% confluence. The cells were co-transfected (Lipofectamine 2000, Invitrogen) using 0.8 µg of plasmid DNA and 0.08 µg phRL-SV40 control renilla plasmid using standard 24-well protocol. An empty pGL3-Basic vector and pGL3-control vector which contains a SV40 promoter are included in each experiment along with the construct under investigation.. Medium was replaced 5 hours post transfection. Dual luciferase assays (Promega) were performed in quadruplicate 24 hours post transfection following the manufacturers standard protocol (Tecan Genios Microplate Reader). Relative luciferase units (RLU) were calculated by dividing the firefly luciferase values by the Renilla control values for each transfection reaction.

Tissue samples

Tissue samples were acquired from the Harvard Brain Tissue Resource Center (HBTRC). They were dissected from flash frozen brain slices (at –80°C) from the superior temporal lobe (Brodmann area 22) of 184 neurologically normal subjects. They were 77.6% male, the average age at death was 62 ± 12.8 years and had an average time between death and tissue collection of 23 ± 5.3 hours. Total RNA integrity was confirmed by agarose gel electrophoresis. Genomic DNA from these samples was extracted from 10 mg of tissue using the Gentra Puregene Tissue Kit (Qiagen) following manufacturer's protocol. RNA was extracted from 30 mg of tissue using the RNeasy Lipid Tissue Mini Kit (Qiagen). Reverse transcription reactions on total RNA were performed using GeneAmp RNA PCR Kit (Applied Biosystems) and random hexamer primers following standard protocols. Real

time PCR experiments for *NGB* and normalization transcripts were performed in triplicate and on the same set of reverse-transcribed RNAs.

Patient samples

All AD patients (N=79) were unrelated, met NINCDS/ADRDA criteria for diagnosis of probable or possible AD or had a definite diagnosis after autopsy confirmation and were from the 3-site collection of the NIMH Genetics Initiative. Sample procurement, and data gathering for the NIMH samples are further described elsewhere [29, 30]. A subset of these (N=30) was collected by our site and after death had autopsies performed by our neuropathology department. Those were included in the sample because of their definite diagnosis. The remaining samples were selected to represent families that showed linkage to chromosome 14 in our original linkage report, which initiated our interest in *NGB* [31].

Real time PCR quantifications

Real time PCR reactions were performed in triplicate using the SYBR® Green qPCR Detection System (Invitrogen) as previously described [23]. Amplicon sequences were verified by Sanger sequencing. Gel electrophoresis and melting curves further confirmed that a single product was amplified and measured. Samples were run in triplicate and randomly distributed on two 384 well plates. Plate identity was accounted for in the analyses as described under "Statistical analysis". Real time measurements were made on an Applied Biosystems 7900HT sequence detection system and measurements were converted to relative quantities using six serial 2-fold dilutions in triplicate. Triplicates were examined for outlier measurements that were removed if present. Based on the literature on optimal normalization controls we chose three genes, *ACTB*, *POLR2F* and *MRIP* to control for variations in RNA input and reverse transcription efficiency [32, 33]. We performed real time experiments for all three as we have previously described [34], compared their inter-sample variability and their pair-wise correlations and used for normalization the mean of the least variable and most correlated pair, *ACTB* and *MRIP*. The normalized values were log base 2 transformed to achieve a normal distribution before proceeding to statistical analysis.

Genotyping

We genotyped 70 HapMap SNPs within and to the extent of linkage disequilibrium around *NGB*, (chr14:77,627,796-77,820,386 in hg19) chosen to be inter-correlated at $r^2 \geq 0.8$. We also genotyped 27 ancestry informative SNP markers (AIMs) [35] as the HBTRC did not retain ancestry information. All genotyping was done at the Johns Hopkins SNP center on a custom SNP panel using the Illumina Golden Gate platform. All SNP dbSNP IDs and Hardy Weinberg equilibrium test p-values are shown in Table 2, arranged by the reason they were genotyped (AIMs or *NGB*).

Statistical analyses

Statistical Analyses: SNP genotypes were coded in a quantitative manner. Principal component analyses of the AIM SNPs were performed in R using the "principal()" function in the "psych" package (Revelle, W. 2011, North-western University, R package version 1.0–95) replacing missing genotypes with the population's mean for the corresponding SNP. Statistical analyses of normalized log transformed expression data against genotypes were also performed in R using the "glm" function for fitting generalized linear models with formulas as described in the text, an identity link function and a Gaussian distribution. Age at death, sex, PMI, genotyping plate and the first principal component of the AIM SNPs were included in the model to correct for possible effects of these variables on the transcript

levels of the measurements. Interaction terms between genotype and age were also included in separate analysis.

Next generation sequencing

Targeted resequencing was performed at the Neuroglobin locus, including the gene itself and 12,540 upstream bases and 13,034 downstream bases (UCSC hg19, chr14:77,718,800-77,750,195), a total of 31,396 bp. First the samples were amplified by long range PCR in three amplicons each using primers Ap1-1F: 5'-GCCAACAACCTGTGTGTTCTAGC-3', Ap1-1R: GCATTTGATTGTGTCTCTTCACTC, Ap2-1F: 5'-TGAGCCAGCACAAGAGATG-3', Ap2-1R: 5'-TGAACCTCAATGACCAGCAC-3', Ap3-1F: 5'-TCTATCCCTGCTCCCTTCAA-3' and Ap3-1R: 5'-TGCTACAGCCCAGTTTTCCT-3'. We then used NEBNext DNA Sample Prep Reagent Set I (NEB, cat.# E6000L) to prepare DNA sample. The index primers and adaptor were from Illumina's Multiplexing Sample Preparation Oligonucleotide Kit (Illumina, P/N: 1005709) and the PCR cleaning kits from Qiagen (Qiagen, Cat. #28106, #28006). 100ng of Genomic DNA was amplified with TaKaRa LA Taq (TaKaRa, RR002M) at 68°C annealing temperature. The product was measured with Pico Green dsDNA Quantification Kit (Molecular Probes Cat. #P7589), then the three amplicons of each sample were mixed in equal parts (~3ug total DNA). The whole sample was fragmented to ~200bp size using Covaris M220 Focused-ultrasonicator™. The samples were then processed with Illumina standard protocol, which includes the steps of purification, repair ends, adenylate 3' end, adaptor ligation, gel purification and final enrichment of the fragments. 10nM of the final library of 12 multiplexed samples was sequenced in each lane on an Illumina GAI machine. Most were run with a single end protocol, generating reads of either 100 or 101 bases, with the exception of 24 control samples run in a paired end protocol generating 76 base reads. We compared these samples for any systematic differences in numbers of variants detected and allele calls from the remaining but we found no such difference.

Sequencing data analysis

The vast majority of the samples returned millions of alignable reads, providing coverage ranging from ~6K to ~43K per sample, with a median sample coverage of ~17K. We made high confidence variant calls for 184 control samples and 79 patient samples. For this we developed a pipeline using publicly available tools. We first removed reads with ambiguous bases from our fastq files, then aligned the remaining reads to the hg19 reference chromosome 14 using Bwa [36]. We used utilities offered by the GATK for rest of the steps [37]. These included, in the order they were applied, indel realignment, quality score recalibration, genotyping, and variant filtering [38]. For each of these steps we used all of our aligned reads, as our high coverage, predominantly single end protocol suggested that non-identical DNA fragments would produce multiple reads aligning to the same starting position and in the same orientation. Our final set of variants satisfied several filter criteria including a minimum variant quality of 30 (calculated by GATK; this is a phred scaled score), a minimum quality by depth of 5, a maximum strand bias of -0.1, a maximum homopolymer length of 5 bases, and a within sample maximum of 2 SNPs within any 10 base window.

RESULTS

Bioinformatic analysis

Figure 2 shows *NGB* in the UCSC genome browser with tracks relevant to regulation displayed. Note the region within intron 1 with a high concentration of evidence for regulatory sequence. This includes experimental histone acetylation information by chromatin immunoprecipitation and sequencing (ChIP-seq), sites of DNase hypersensitivity

indicating open chromatin and transcription factor binding sites as evidenced by ChIP-seq. There is a relative lack of such evidence elsewhere, although it must be stressed that no data from neuronal or similar cells are currently made available by ENCODE in this region with the exception of Retinoic acid differentiated SK-N-SH cells. Interestingly, the 3' untranslated region of *NGB* contains an experimentally validated micro RNA, hsa-mir-1260, shown exactly below the RefSeq gene track in Figure 2. We searched the miRNA target database, at www.targetscan.org, and found 193 genes with predicted conserved target sequences for this miRNA. When compared to all genes in the genome using the Panther bioinformatics tools at www.pantherdb.org, the mir-1260 targets show significant enrichments in multiple pathways shown in Table 3. It is interesting that among these pathways are the platelet derived growth factor (PDGF), the overlapping angiogenesis pathway, and the Alzheimer disease-presenilin pathway. This raises the possibility that the transcriptional level of *NGB* might have effects relevant to hypoxia and AD that are not attributable to the *NGB* protein but rather to the regulation of other genes through hsa-mir-1260.

Promoter constructs

The position of the fragments that we inserted in constructs and explored via dual luciferase assays are shown on Figure 2 (see also Table 1). All fragments were chosen to start at the translation start site and extend up to ~2Kb 5'. They included the contiguous sequence with two exceptions that were constructed to clarify observed enhancer effects. The normalized relative levels of Luciferase activity in transfected HEK293, SK-N-SH and Neuro-2a cells are shown in Figure 3. Constructs as short as 201 bp showed activity in driving luciferase activity significantly above the pGL3-Basic vector which lacks a promoter (all p-values were below 0.0001) except for SK-N-SH cells where the change was marginal (p=0.07). Addition of more promoter sequence up to 379 bp consistently increased luciferase activity in HEK 293 and SK-N-SH cells while this continued up to 1124 bp for Neuro 2A cells. Larger constructs showed a reduction in luciferase luminescence across cell lines. In order to explore whether the reduction in larger constructs was due to insert size difference or repressor effects in the added sequence we generated two more constructs. The first, 0–385&1180–1568, was intended to test whether the 1180–1568 fragment would reduce the expression of the 0–385 fragment despite the total size remaining below 1 Kb. Similarly the second, 0–462&1681–2056 was intended to reveal potential repressor effects of the 1681–2056 fragment. The results are shown in Figure 4. Note that for technical reasons including the availability of restriction enzyme sites to allow us to generate these constructs, the fragment sizes do not exactly correspond to those used in the first seven constructs. The addition of the 1180–1568 fragment to the first 385 bp not only did not reduce the expression compared to the 376 bp construct, but rather increased it, making it unlikely that it includes sequences with suppressor function. The addition of the 1681–2056 fragment to the first 462 bp caused a small decrease in luciferase compared to the 1–478 fragment alone, far smaller than that seen in Figure 1 between the 1586 and 2056 bp fragments. Overall, our results suggest that cis regulation at the 5' end of the *NGB* translation start site is mostly driven by elements within the first 1124 bp of sequence, while the reduction observed in larger constructs is most likely due to the increased insert size.

Correlations of *NGB* expression with common genetic variation

Real time PCR quantification of the *NGB* transcript gave consistent amplification from cDNA with little variation within triplicates. It was clear however that the *NGB* transcript is at low abundance in the adult human temporal lobe, as the detection threshold (Ct) in our qPCR analysis was reached at cycle 26.5 compared to cycle 19 for the *ACTB* gene at the same cDNA input. This is an estimated 180-fold difference depending on amplification efficiency. The relatively low abundance is in agreement with the results of Sun et al [24].

Analysis of the ancestry informative SNP markers [35] by principal component analysis showed that most of our brain tissue samples shared common ancestry, presumably European, with two exceptions that were outliers for the first principal component and were removed from the dataset. In total we had 176 samples with both genotype and expression data for analysis. The total number of markers within the *NGB* LD block was 69. The average missing data rate was 0.2% and the maximum was 1.2%. All markers were in Hardy Weinberg equilibrium (Table 2). Due to the significant LD between markers, we first calculated the correlation matrix between markers using Haploview and then the number of independent comparisons by counting each marker once minus the maximum observed r^2 before removing it from the set. For example marker pairs at $r^2 = 0.3$ were counted as 1.7 tests. This correction is somewhat conservative as it does not account for multi-marker correlations. However, it is a great improvement to Bonferroni correction, estimating that our 69 markers accounted for 35.4 tests, setting a study wide significance threshold at 0.0014.

There were no significant correlations between transcript levels and PMI, sex, PCR plate or the first principal component of the genotypes (presumed to represent small remaining ethnic background differences). There was a significant effect of age on the levels of *NGB* transcript, with an approximate 1% decrease per year ($p=0.001$), which replicates our previous report [23].

Three of the tested SNPs showed nominally significant associations with transcript levels, rs8014408, rs1861420 and rs981471. The three were in LD with each other and one of them, rs981471, reached study wide significance (multiple test corrected study-wide $p < 0.048$). This SNP is ~100 Kb away from *NGB* and another gene (*TMEM63C*) separates them. However it is within the *NGB* LD block (chr14:77,627,796-77,820,386 in hg19) and therefore can be a proxy for variation closer to the gene. Alternately, long range regulation is also a possibility.

Because there is a correlation between *NGB* expression and age, suggesting that age related regulatory elements might be present, we proceeded to test these three SNPs for interactions between genotype and age on their effects on expression. One of the SNPs that previously showed only nominally significant association with *NGB* expression ($p=0.047$), rs8014408, now showed a strong interaction with age (interaction term $p=0.002$) and when the interaction term was included in the regression the genotype effect on expression was study wide significant (multiple test corrected $p = 0.028$). The rare allele carriers showed higher expression, which declined faster with age. This SNP is located 51 Kb 3' of *NGB* within the *TMEM63C* gene.

Sequencing for rare regulatory variants

Our sequencing of 184 controls with expression data and 79 cases revealed 275 high confidence variants in and around *NGB* including 256 snps and 19 indels. Of these, 98 had not been previously reported in dbSNP build 135 [39], including 82 SNPs and 16 indels. Most of the unknown single nucleotide variants (62 of 82 SNPs) were only seen once (1 in 526 chromosomes) while the most frequent was seen 9 times (allele frequency 1.7%). We chose 24 of these for verification via restriction enzyme digestion (see Table 4), biased towards those that could be genotyped by this method and those seen less frequently (20/24 were seen only once) as they are more likely to be false positives. We were able to verify all 25 of the variants, suggesting that our high read depth data had very high specificity. Rare alleles were distributed across all samples, with the exception of three cases that were clear outliers, carrying more than 10 standard deviations more rare alleles than the average. As we have no ethnicity information on the cases, we suspected that these represent members of a different population than the rest of the sample and they were removed.

We identified 3 exonic variants, including two non-synonymous variants. Each are known, rare, and seen in both cases and controls. Setting minor allele frequency thresholds between 0.5 and 10% we looked for enrichment of rare alleles in cases or controls. We found that at thresholds between 1% and 6% there was a significant enrichment of rare alleles in controls (best $p=0.0004$ for a MAF threshold of 0.2%). This however was driven by a long 8-SNP, 18 Kb haplotype shared by 9 controls and no AD case. One of those SNPs was a known variant, rs72681205 located in the 3' UTR of *NGB*, after mir-1260. Given the number of variants we examined overall, the presence of all 9 alleles only in controls is not statistically significant. We concluded that there was no statistically significant difference in the number of rare alleles carried by cases or controls. Common SNPs were tagged by SNPs in our common variant genotyping array. However, none of them tagged well the two common variants with significant effects on *NGB* expression, rs981471 and rs8014408. In the same manner we explored correlations between *NGB* expression and number of rare alleles for multiple thresholds, but we found no significant correlation. Only 3 variants were found within the 1 kb upstream of *NGB* that we showed above to drive luciferase activity, one of which was novel. They were all rare variants, with a total of 8 alternative alleles, 5 in the controls and 3 in the cases. There was no difference in expression for those carrying the rare alleles.

A CT repeat in intron 1 of *NGB* attracted our attention because of evidence from public data that it binds multiple transcription factors and that it resides at a site of Histone 3 Lysine 27 Acetylation, a mark of active enhancers [40], the only such region seen in Figure 2. We were not able however to reliably analyze this region. It consists of a stretch of 130 C and T nucleotides followed by a run of 29 C nucleotides in the reference genome. Our sequencing results clearly indicated that it is polymorphic with apparently different common patterns between cases and controls but due to the noise and the existence of reads of many more than two different lengths, no specific alleles could be unequivocally called for any individual and this visual impression of a difference could not be confirmed. Our efforts to genotype the variant as a repeat length polymorphism were also not successful due the amplification of products of multiple lengths, a phenomenon we attribute to PCR artifacts due to the length, the high GC content and the highly repetitive nature of the region. Therefore, we report our observation on this region not as a conclusive result but as a likely target for follow up by alternative methods.

DISCUSSION

We have performed an extensive analysis of the genetically determined regulation of the *NGB* gene. Using a series of in vitro reporter assays in three different cell lines we have found that sequences up to 376 bp upstream from the translation start site and perhaps as much as ~1Kb based on the results from Neuro-2A cells can contribute to the induction of luciferase expression. This is highly consistent with data from Zhang et al [41] who identified the transcription start site (TSS) 306 bp upstream from the initiation methionine codon (Translation Start Site) and used reporter assays in two additional cell lines to show that the proximal promoter lies between bp -164 and +306 relative to the TSS. This almost exactly corresponds to our 1-478 construct, which is among our constructs with the highest activity. By replicating these results in three additional cell types in our study, there is now good determination of the proximal promoter of *NGB*, which will be useful for future studies of the gene's regulation.

Our main interest was in identifying genetic variation that influences *NGB* expression. The results of our sequencing, however, did not support the existence of variants that alter the regulatory role of this region. We found no common variants within 1 Kb of the initiating methionine, and the few rare variants did not show any enrichment in cases or controls or

any significant effect on the transcript levels. It is notable however that sequencing this non-coding region in 263 individuals only identified 3 rare variants, and it could suggest conservation due to functional importance.

In toto, we sequenced 31.4 kb that included the *NGB* gene together with more than 12 Kb of 3' and 5' sequences in 184 controls with gene expression data and 79 AD cases. We analyzed the data for enrichment of rare variants in the group of AD cases as well as for correlations between *NGB* transcript levels and the presence of rare alleles without significant results. This was also true when the region under study was restricted to the 1 Kb proximal to the transcription start site. This negative result excludes the presence of variants in the immediate vicinity of the gene that have a strong influence on its expression.

Our search for common variation affecting *NGB* transcription was more fruitful. We identified two common genetic variants whose genotypes show correlation with *NGB* mRNA levels in post mortem samples without pathology (controls) and modify the effects of ageing on the gene's transcript levels. This interaction with age is of particular interest given our identification of an effect of age on *NGB* [23] and our previous work showing strong overlap with genes changing transcription in AD affected brains [26]. The present study replicates the effect of age on *NGB* that we observed in those studies. Note that although *NGB* did not make the list of genes reported in our genome wide exploration [26] at the false discovery rate of 0.05, it did show a reduction with age with a p-value of 0.02 in that smaller sample. Here the reduction we observe with age is ~ 1% per year, which is about half in magnitude than our previous results [23]. This could be the result of the "winner's curse" affecting all replication studies, the result of different age distributions of the samples, or both. The reduction of *NGB* expression with age has been also shown in rodents [42]. The contrast between the lack of positive results from sequencing regions close to *NGB* and the findings from common variants that are further and within another neighboring gene (*TMEM63C*) suggests that the effects we see are happening over a long distance. Although the associated SNPs are within the extent of LD around the *NGB* gene, our sequencing did not identify any more proximal SNPs that could explain the effect. *NGB* is located in a gene-dense region, the distance with its neighbors being only 6 and 4 Kb away. The associated SNPs, rs8014408 and rs981471, are 51 and 102 Kb away from the gene respectively. Enhancers have been observed in distances that are an order of magnitude further therefore it is likely that these SNPs tag variation much further from the gene than we sequenced. Clearly, future follow up attempts need to extend the search area, a feat complicated by the presence of other genes.

Data from the ENCODE project as seen in the image from the UCSC genome browser in Figure 2 show very little evidence for regions of regulatory potential, especially for the presence of enhancers as predicted by H3K4 mono and tri- methylation and H3K27 acetylation [40]. The notable exception is a region in intron 1 between 77,735,926 and 77,736,479 bp in hg19. This region contains a dinucleotide repeat of 56 CT followed closely by a run of 37 Cytosines with 4 interspersed Thymidines. Our sequencing data suggested that this area is highly polymorphic with different patterns of reads observed in each individual, and these patterns appeared different between AD cases and controls and high or low *NGB* transcript samples. Unfortunately, the sequence context makes it impossible to accurately align short reads within this region and to get reliable allele calls. Within each individual's data we found reads of multiple lengths, which could be due to experimental artifacts. Attempts to type this region by PCR amplification also resulted in multiple alleles. It is not unlikely that this repeat is subject to frequent somatic mutation as we have previously described [43], nevertheless and despite our best efforts we were unable to reliably call genotypes and test whether it is relevant to AD, the gene's regulation or the effects we detected from the distant SNPs. We feel that it is important to be aware of the

existence of this potential polymorphic regulatory region when further exploring the regulation of *NGB*.

Additional interesting data shown here includes the *in silico* analysis of the microRNA expressed from the 3' untranslated of the *NGB* transcript. This experimentally validated miRNA shows a very interesting pattern of target enrichments in the angiogenesis pathway and the Alzheimer disease-presenilin pathway. This result, together with the pre-existing data on links between *NGB* expression, resistance to hypoxia and AD, suggests an additional mechanism that should be taken into account when studying these effects.

In summary, through bioinformatic analysis, in vitro assays, expression analysis, common variant genotyping and extensive sequencing, we have provided a systematic assessment of genetic regulation of *NGB*. We have found no variation influencing proximal regulatory elements but likely remote effects that appear to be age dependent. Given the robust data of change in *NGB* expression with age this can be of particular interest. Additional interesting points include the potential functions of the *NGB* co-expressed miRNA and the presence of a polymorphic and likely functional CT-repeat in intron 1. Together with the work of others on this gene whose transcript levels appear to be important for health and disease we hope this work will accelerate the progress in understanding its involvement and its translational potential.

Acknowledgments

We thank Dr. Francine Benes of the Harvard Brain Tissue resource center for providing brain tissue. We thank Dr Andrew McCallion and Dr David Valle for conversations, and expert advice in many aspects of the project. This work was supported by NIA grants to DA and SSB (RO1AG022099 and RO1AG021804) and an award from the Neurosciences Education and Research Foundation to DA.

REFERENCES

1. Theorell H, de Duve C. Crystalline human myoglobin from heart-muscle and urine. Arch Biochem. 1947; 12:113–124. [PubMed: 20277228]
2. Wood D, Blanchetot A, Jeffreys AJ. Molecular cloning of seal myoglobin mRNA. Nucleic Acids Res. 1982; 10:7133–7144. [PubMed: 6185919]
3. Kanatous SB, Mammen PP. Regulation of myoglobin expression. J Exp Biol. 2010; 213:2741–2747. [PubMed: 20675543]
4. Burmester T, Weich B, Reinhardt S, Hankeln T. A vertebrate globin expressed in the brain. Nature. 2000; 407:520–523. [PubMed: 11029004]
5. Burmester T, Ebner B, Weich B, Hankeln T. Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. Mol Biol Evol. 2002; 19:416–421. [PubMed: 11919282]
6. Trent JT 3rd, Watts RA, Hargrove MS. Human neuroglobin, a hexacoordinate hemoglobin that reversibly binds oxygen. J Biol Chem. 2001; 276:30106–30110. [PubMed: 11429401]
7. Duong TT, Witting PK, Antao ST, Parry SN, Kennerson M, Lai B, Vogt S, Lay PA, Harris HH. Multiple protective activities of neuroglobin in cultured neuronal cells exposed to hypoxia re-oxygenation injury. J Neurochem. 2009; 108:1143–1154. [PubMed: 19154338]
8. Shao G, Gong KR, Li J, Xu XJ, Gao CY, Zeng XZ, Lu GW, Huo X. Antihypoxic effects of neuroglobin in hypoxia-preconditioned mice and SH-SY5Y cells. Neurosignals. 2009; 17:196–202. [PubMed: 19325253]
9. Schmidt-Kastner R, Haberkamp M, Schmitz C, Hankeln T, Burmester T. Neuroglobin mRNA expression after transient global brain ischemia and prolonged hypoxia in cell culture. Brain Res. 2006; 1103:173–180. [PubMed: 16796995]
10. Sun Y, Jin K, Mao XO, Zhu Y, Greenberg DA. Neuroglobin is up-regulated by and protects neurons from hypoxic-ischemic injury. Proc Natl Acad Sci U S A. 2001; 98:15306–15311. [PubMed: 11742077]

11. Hundahl C, Stoltenberg M, Fago A, Weber RE, Dewilde S, Fordel E, Danscher G. Effects of short-term hypoxia on neuroglobin levels and localization in mouse brain tissues. *Neuropathol Appl Neurobiol*. 2005; 31:610–617. [PubMed: 16281909]
12. Fordel E, Geuens E, Dewilde S, De Coen W, Moens L. Hypoxia/ischemia and the regulation of neuroglobin and cytoglobin expression. *IUBMB Life*. 2004; 56:681–687. [PubMed: 15804832]
13. Sun Y, Jin K, Peel A, Mao XO, Xie L, Greenberg DA. Neuroglobin protects the brain from experimental stroke in vivo. *Proc Natl Acad Sci U S A*. 2003; 100:3497–3500. [PubMed: 12621155]
14. Khan AA, Wang Y, Sun Y, Mao XO, Xie L, Miles E, Graboski J, Chen S, Ellerby LM, Jin K, Greenberg DA. Neuroglobin-overexpressing transgenic mice are resistant to cerebral and myocardial ischemia. *Proc Natl Acad Sci U S A*. 2006; 103:17944–17948. [PubMed: 17098866]
15. Jin K, Mao Y, Mao X, Xie L, Greenberg DA. Neuroglobin expression in ischemic stroke. *Stroke*. 2010; 41:557–559. [PubMed: 20075359]
16. Li RC, Morris MW, Lee SK, Pouranfar F, Wang Y, Gozal D. Neuroglobin protects PC12 cells against oxidative stress. *Brain Res*. 2008; 1190:159–166. [PubMed: 18083144]
17. Antao ST, Duong TT, Aran R, Witting PK. Neuroglobin overexpression in cultured human neuronal cells protects against hydrogen peroxide insult via activating phosphoinositide 3 kinase and opening the mitochondrial K(ATP) channel. *Antioxid Redox Signal*. 2010; 13:769–781. [PubMed: 20367258]
18. Raychaudhuri S, Skommer J, Henty K, Birch N, Brittain T. Neuroglobin protects nerve cells from apoptosis by inhibiting the intrinsic pathway of cell death. *Apoptosis*. 2010; 15:401–411. [PubMed: 20091232]
19. Brittain T, Skommer J, Henty K, Birch N, Raychaudhuri S. A role for human neuroglobin in apoptosis. *IUBMB Life*. 2010; 62:878–885. [PubMed: 21190290]
20. Schmidt M, Giessel A, Laufs T, Hankeln T, Wolfrum U, Burmester T. How does the eye breathe? Evidence for neuroglobin-mediated oxygen supply in the mammalian retina. *J Biol Chem*. 2003; 278:1932–1935. [PubMed: 12409290]
21. Geuens E, Brouns I, Flamez D, Dewilde S, Timmermans JP, Moens L. A globin in the nucleus! *J Biol Chem*. 2003; 278:30417–30420. [PubMed: 12796507]
22. Li RC, Pouranfar F, Lee SK, Morris MW, Wang Y, Gozal D. Neuroglobin protects PC12 cells against beta-amyloid-induced cell injury. *Neurobiol Aging*. 2008; 29:1815–1822. [PubMed: 17560688]
23. Szymanski M, Wang R, Fallin MD, Bassett SS, Avramopoulos D. Neuroglobin and Alzheimer's dementia: genetic association and gene expression changes. *Neurobiol Aging*. 2010; 31:1835–1842. [PubMed: 19010568]
24. Sun F, Mao X, Xie L, Greenberg DA, Jin K. Neuroglobin protein is upregulated in Alzheimer's disease. *J Alzheimers Dis*. 2013; 36:659–663. [PubMed: 23648513]
25. Miller JA, Oldham MC, Geschwind DH. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J Neurosci*. 2008; 28:1410–1420. [PubMed: 18256261]
26. Avramopoulos D, Szymanski M, Wang R, Bassett S. Gene expression reveals overlap between normal aging and Alzheimer's disease genes. *Neurobiol Aging*. 2011; 32:2319.e27–2319.e34. [PubMed: 20570407]
27. Lin Y, Fang L, Xue XH, Murong SX, Wang N, Wu ZY. Association between Ngb polymorphisms and ischemic stroke in the Southern Chinese Han population. *BMC Med Genet*. 2008; 9:110–2350-9-110.
28. Chuang PY, Conley YP, Poloyac SM, Okonkwo DO, Ren D, Sherwood PR, Hravnak M, Alexander SA. Neuroglobin genetic polymorphisms and their relationship to functional outcomes after traumatic brain injury. *J Neurotrauma*. 2010; 27:999–1006. [PubMed: 20345238]
29. Blacker D, Haines JL, Rodes L, Terwedow H, Go RC, Harrell LE, Perry RT, Bassett SS, Chase G, Meyers D, Albert MS, Tanzi R. ApoE-4 and age at onset of Alzheimer's disease: the NIMH genetics initiative. *Neurology*. 1997; 48:139–147. [PubMed: 9008509]
30. Bassett SS, Avramopoulos D, Fallin D. Evidence for parent of origin effect in late-onset Alzheimer disease. *Am J Med Genet*. 2002; 114:679–686. [PubMed: 12210287]

31. Avramopoulos D, Fallin MD, Bassett SS. Linkage to chromosome 14q in Alzheimer's disease (AD) patients without psychotic symptoms. *Am J Med Genet B Neuropsychiatr Genet.* 2005; 132B:9–13. [PubMed: 15389761]
32. Hoerndli FJ, Toigo M, Schild A, Gotz J, Day PJ. Reference genes identified in SH-SY5Y cells using custom-made gene arrays with validation by quantitative polymerase chain reaction. *Anal Biochem.* 2004; 335:30–41. [PubMed: 15519568]
33. Johansson S, Fuchs A, Okvist A, Karimi M, Harper C, Garrick T, Sheedy D, Hurd Y, Bakalkin G, Ekstrom TJ. Validation of endogenous controls for quantitative gene expression analysis: application on brain cortices of human chronic alcoholics. *Brain Res.* 2007; 1132:20–28. [PubMed: 17188656]
34. Szymanski M, Wang R, Bassett SS, Avramopoulos D. Alzheimer's risk variants in the clusterin gene are associated with alternative splicing. *Transl Psychiatry.* 2011; 1 10.1038/tp.2011.17.
35. Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A. SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet.* 2007; 1:273–280. [PubMed: 19083773]
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
38. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011; 43:491–498. [PubMed: 21478889]
39. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29:308–311. [PubMed: 11125122]
40. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A.* 2010; 107:21931–21936. [PubMed: 21106759]
41. Zhang W, Tian Z, Sha S, Cheng LY, Philipsen S, Tan-Un KC. Functional and sequence analysis of human neuroglobin gene promoter region. *Biochim Biophys Acta.* 2011; 1809:236–244. [PubMed: 21362510]
42. Sun Y, Jin K, Mao XO, Xie L, Peel A, Childs JT, Logvinova A, Wang X, Greenberg DA. Effect of aging on neuroglobin expression in rodent brain. *Neurobiol Aging.* 2005; 26:275–278. [PubMed: 15582755]
43. Talbot CC Jr, Avramopoulos D, Gerken S, Chakravarti A, Armour JA, Matsunami N, White R, Antonarakis SE. The tetranucleotide repeat polymorphism D21S1245 demonstrates hypermutability in germline and somatic cells. *Hum Mol Genet.* 1995; 4:1193–1199. [PubMed: 8528208]

Figure 1.
Amino acid sequence of the human, mouse chicken and zebrafish comparison shows very strong conservation. The sequence is shaded from black to white according to the number of species carrying the same amino acid at that position.

Figure 2. Screen shot from the UCSC genome browser including the track of genes from RefSeq showing the location and exons of *NGB*, a microRNA track, and ENCODE project tracks showing Histone acetylation, DNase hypersensitivity and Transcription factor binding sites.

Figure 3.
Levels of Luciferase activity in three cell types transfected with empty vector, vector with a CMV promoter and the 7 different basic constructs

Figure 4.
Levels of Luciferase activity in Neuro-2A cell transfected with empty vector, vector with a CMV promoter and the 2 constructs designed to assess repressor effects of sequences beyond 1 Kb from the transcription start site.

Table 1

Genome coordinates of inserts in Luciferase expression constructs

construct	chromosome	start bp-hg19	stop bp-hg19
1-201	14	77,737,282	77,737,481
1-306	14	77,737,282	77,737,587
1-376	14	77,737,282	77,737,657
1-478	14	77,737,282	77,737,759
1-1124	14	77,737,282	77,738,405
182-1124	14	77,737,460	77,738,406
1-1568	14	77,737,282	77,738,849
0-385&1180-1568	14	77,737,281 77,738,461	77,737,666 77,738,849
1-2056	14	77,737,282	77,739,337
1-462&1681-2056	14	77,737,281 77,738,962	77,737,743 77,739,337

Table 2

SNPs included in the common variant genotyping. Ancestry informative markers and NGB SNPs are shown separately, together with the p-values for Hardy Weinberg equilibrium.

AIM SNP ID#	HW-p	NGB SNP ID#	HW-p	NGB SNP ID#	HW-p
rs10141763	0.555	rs10111183	0.130	rs3783988	0.120
rs1024116	0.685	rs10134823	0.488	rs3813539	0.451
rs12913832	0.939	rs1046428	0.64	rs388764	0.814
rs1321333	0.944	rs11627257	0.105	rs391966	0.386
rs1335873	0.193	rs11627633	0.287	rs409109	0.493
rs1498444	0.940	rs11844745	0.598	rs4243654	0.788
rs1886510	0.427	rs11847091	0.478	rs428868	0.014
rs1978806	0.847	rs11850308	0.024	rs438931	0.744
rs2026721	0.513	rs12433986	0.010	rs4396457	0.703
rs2040411	0.974	rs1467011	0.543	rs4467006	0.072
rs2065160	0.540	rs17105685	0.849	rs450587	0.053
rs2065982	0.513	rs17105732	0.338	rs4540995	0.019
rs2303798	0.000	rs17750078	0.611	rs4899644	0.689
rs2304925	0.742	rs17750404	0.367	rs4899650	0.340
rs239031	0.925	rs17812088	0.188	rs4899651	0.598
rs2572307	0.945	rs1861420	0.964	rs4903549	0.048
rs2814778	0.848	rs1861888	0.168	rs4903560	0.120
rs3785181	0.148	rs2017792	0.241	rs7159558	0.202
rs5997008	0.000	rs2020996	0.950	rs733416	0.675
rs722098	0.052	rs2058916	0.399	rs735477	0.038
rs727811	0.817	rs2098380	0.729	rs747273	0.215
rs730570	0.701	rs2110908	0.279	rs8004558	0.828
rs773658	0.850	rs2110912	0.470	rs8014408	0.082
rs7897550	0.636	rs2270422	0.960	rs8015231	0.019
rs881929	0.828	rs2287381	0.317	rs8016187	0.091
rs896788	0.060	rs2287383	0.123	rs8017067	0.359
rs917118	0.636	rs2287386	0.196	rs8021076	0.014
		rs2287396	0.544	rs8177539	0.895

AIM SNP ID#	HW-p	NGB SNP ID#	HW-p	NGB SNP ID#	HW-p
		rs2287397	0.966	rs8177544	0.938
		rs2287398	0.562	rs8177565	0.962
		rs2540896	0.907	rs8177569	0.000
		rs2655997	0.239	rs888059	0.517
		rs369202	0.322	rs888060	0.756
		rs376645	0.048	rs905734	0.010
		rs3783986	0.968	rs981471	0.373

Table 3

Pathway enrichments among the genes that are predicted to be regulated by the microRNA mir-1260 in the 3' untranslated region of *NGB*. Gene predictions were performed at the target database targets can at www.targetscan.org and enrichment tests with corrected significance levels were calculated using the Panther bioinformatics tools at www.pantherdb.org

	N genes observed	N genes expected	Corrected p
Axon guidance mediated by semaphorins	4	0.41	0.0008
PDGF signaling pathway	7	1.52	0.0009
Angiogenesis	7	1.82	0.0026
Insulin/IGF pathway - mitogen activated protein kinase kinase/MAP kinase cascade	3	0.33	0.0048
Ras Pathway	4	0.75	0.0073
Wnt signaling pathway	8	3.02	0.0118
Interleukin signaling pathway	5	1.54	0.0198
Alzheimer disease - presenilin pathway	4	1.16	0.0303

Table 4

Location, carrier sample, enzyme used for validation and alternative allele for rare detected non-annotated SNPs

POSITION	Sample ID	Enzyme	REF/ALT
77719185	AN01685	BsaI	C/T
77719860	AN13112	BseRI	C/G
77722416	AN00317;AN08807	BsmAI	C/G
77724555	NC01683	BstXI	C/A
77725622	AN12089	EaRI	C/T
77728028	AN04568	AciI	G/A
77728526	AN01787	MnII	C/A
77729553	AN10642	BseRI	G/A
77729744	AN09545	BseRI	G/A
77730033	AN14867;NC00204	MnII	A/C
77732190	AN16146	AciI	C/T
77732243	AN13565;AN01107;NC01544	AciI	C/T
77734751	AN16568	EaRI	G/A
77741509	NC00642	PstI	C/G
77741948	NC00642	BseRI	G/C
77743135	AN16017	NlaIII	C/T
77743510	AN07239	NlaIII	G/A
77745127	NC00123	AciI	C/G
77745254	AN05666	AciI	G/A
77746528	AN12916	NlaIII	C/T
77746666	NC03416	BsmAI	C/T
77747302	AN06708	AciI	C/T
77748738	NC02240	AciI	G/A
77749306	AN06004;AN12916	AciI	G/C