Molecular Biology and Evolution

Multiple Functional Variants in *cis* Modulate *PDYN* Expression

Courtney C. Babbitt,*[,1,2] Jesse S. Silverman,[2] Ralph Haygood,[2] Jennifer M. Reininga,[1] Matthew V. Rockman,[3] and Gregory A. Wray[1,2]

[1]Institute for Genome Sciences & Policy, Duke University

[2]Department of Biology, Duke University

[3]Department of Biology and Center for Genomics and Systems Biology, New York University

***Corresponding author:** E-mail: courtney.babbitt@duke.edu.

Associate editor: Jonathan Pritchard

Abstract

Understanding genetic variation and its functional consequences within *cis*-regulatory regions remains an important challenge in human genetics and evolution. Here, we present a fine-scale functional analysis of segregating variation within the *cis*-regulatory region of *prodynorphin*, a gene that encodes an endogenous opioid precursor with roles in cognition and disease. In order to characterize the functional consequences of segregating variation in *cis* in a region under balancing selection in different human populations, we examined associations between specific polymorphisms and gene expression *in vivo* and *in vitro*. We identified five polymorphisms within the 5′ flanking region that affect transcript abundance: a 68-bp repeat recognized in prior studies, as well as two microsatellites and two single nucleotide polymorphisms not previously implicated as functional variants. The impact of these variants on transcription differs by brain region, sex, and cell type, implying interactions between *cis* genotype and the differentiated state of cells. The effects of individual variants on expression level are not additive in some combinations, implying epistatic interactions between nearby variants. These data reveal an unexpectedly complex relationship between segregating genetic variation and its expression-trait consequences and highlights the importance of close functional scrutiny of natural genetic variation within even relatively well-studied *cis*-regulatory regions.

Key words: cis-regulation, functional variation, human gene expression, polymorphism.

Research article

Introduction

Many functional polymorphisms within *cis*-regulatory regions influence transcription in humans (reviewed in Rockman and Wray 2002; Knight 2004; Pastinen and Hudson 2004). Some of these variants have clear trait, fitness, and disease consequences (Tournamille et al. 1995; Hamblin and Di Rienzo 2000; Bamshad et al. 2002; Enattah et al. 2002; Horan et al. 2003; Boodhoo et al. 2004; Hacking et al. 2004; Rockman et al. 2005; De Gobbi et al. 2006; Tishkoff et al. 2007; Luca et al. 2009). Genome-wide screens for polymorphisms that influence transcription indicate that segregating genetic variation influences the transcription of a large proportion of genes (Lo et al. 2003; Schadt et al. 2003; Pastinen and Hudson 2004; Stranger et al. 2007). With very few exceptions (Horan et al. 2003; Tao et al. 2006), however, little is known about the functional consequences of multiple sequence variants on the transcription of a single gene.

Here, we present an analysis of the impact of multiple segregating polymorphisms on the expression of *PDYN* (*prodynorphin*). The peptide encoded by *PDYN* can be processed into five distinct neuropeptides, which in experimental models play roles in nociception, dysphoria, dyskinesia, motor control, memory acquisition, learning, stress-induced analgesia, and modulation of reward in addiction (Kieffer and Gaveriaux-Ruff 2002; Corbett et al. 2006; Marinova 2006). Previous studies have implicated a 68-bp repeat located 1,250 bp upstream of the transcription start site in differential *PDYN* expression in humans (Zimprich et al. 2000; Nikoshkov et al. 2008), although another study did not detect this effect (Cirulli and Goldstein 2007). Segregating haplotypes in humans contain one to four 68-bp repeats, with two or three repeats the most prevalent in surveyed populations (Zimprich et al. 2000; Rockman et al. 2005). The 68-bp repeat has been the subject of several functional analyses and genetic association studies. One- and two-repeat haplotypes have lower inducibility *in vitro* than three- and four-repeat haplotypes, and that same pattern has been seen in associations between repeat number and expression level *in vivo* (Nikoshkov et al. 2008). This *cis*-regulatory variant has been investigated extensively in association studies for a spectrum of disease phenotypes, including epilepsy, schizophrenia, and drug and alcohol abuse (Chen et al. 2002; Stogmann et al. 2002; Ventriglia et al. 2002; Fagergren et al. 2003; Nomura et al. 2006; Xuei et al. 2006).

The evolutionary history of the regulatory region of *PDYN* is also intriguing as it shows a pattern of positive selection in the human lineage as compared with other nonhuman primates (Rockman et al. 2005) and shows evidence of positive selection favoring different *PDYN* regulatory alleles in different human populations (Rockman

Table 1. The *cis*-Regulatory Polymorphisms of *PDYN* Analyzed in This Study.[a]

Identification	Polymorphism	*PDYN* Position	Chromosome Position
SNP(−2746)	SNP C/T	−2746	1925448
MSAT(−2745)	Microsat 18, 19, 21, 23	−2745	1925447–1925413
SNP(−2357)	SNP C/A	−2357	1925059
SNP(−2343)	SNP A/T	−2343	1925045
SNP(−2116)	SNP G/C	−2116	1924818
SNP(−2081)^^	SNP T/C	−2081	1924783
SNP(−1754)^^	SNP C/T	−1754	1924456
SNP(−1647)	SNP C/T	−1647	1924349
68-bp Repeat 1	Repeat 1	−1454	1924156–1924089
68-bp Repeat 2	Repeat 2	−1386	1924088–1924021
68-bp Repeat 3	Repeat 3	−1318	1924020–1923953
68-bp Repeat 4	Repeat 4	Not annotated	Not present
SNP(−1148)^^	SNP T/C	−1148	1923850
SNP(−977)	SNP T/G	−977	1923679
MSAT(−362)	Microsat 7, 9	−362	1923064–1923047
SNP(−156)	SNP G/A	−156	1922858
TSS		1	1922702

[a] For each polymorphism, the following information is listed: the identifier, character states of the polymorphism, the distance 5′ from the transcriptional start site, and the position on chromosomal locations are based on human genome NCBI 36.2. Positions that are not annotated are external to the three copies of the 68-bp repeat in the references sequence. Known from a single haplotype (Rockman et al. 2005) but not used in this study.

et al. 2005). *PDYN* falls within a region of elevated F_{ST} between human populations, suggesting that the *cis*-regulatory region has been subject to different selection pressures among human populations. Selection driving differentiation between populations should also decrease variation within a given population. To address this, Rockman et al. (2005) also examined whether the microsatellite nearest the *PDYN* promoter 68-bp element (discussed below as MSAT(−2745)) exhibits the predicted signatures of selection. They found a significant reduction in heterozygosity and repeat-number variance in a number of populations. Biologically, regulation of *PDYN* may be a target for balancing selection due to its physiological role in modulating responses to several psychoactive substances and its known associations with protection against cocaine dependence or abuse (Chen et al. 2002) and with neurological disease states (Stogmann et al. 2002; Ventriglia et al. 2002). The signatures of selection suggest that variation in the repeat region has an impact on fitness. The question then is whether other variation in this regulatory region is important for function.

An ~3-kb region encompassing the 5′ untranslated region (UTR) and 5′ flanking sequence is sufficient to drive inducible expression of *PDYN* in neuroblastoma cells *in vitro* (Rockman et al. 2005). Although the 68-bp repeat is the most conspicuous component of genetic variation within this region, several other closely linked polymorphisms are also common within human populations (table 1). Because studies testing the effects of the 68-bp repeat on gene expression in a hybrid murine cell line (Zimprich et al. 2000) and in human brain regions (Cirulli and Goldstein 2007; Nikoshkov et al. 2008) have reached differing conclusions, we hypothesized that some of these other polymorphisms influence *PDYN* transcription, with potential consequences for traits associated with this gene.

In order to explore the functional consequences of a wider range of variation segregating within the 3-kb *cis*-regulatory region of *PDYN*, we assayed transcription using two approaches: *in vivo* measurements of allelic imbalance in four brain regions from 23 individuals using pyrosequencing and *in vitro* measurements of luciferase reporter expression driven by 19 natural haplotypes in two neuroblastoma cell lines. In order to assess the impact of specific variants on expression, we partitioned genetic effects using both standard multiple regression and a regression-tree approach. These statistical analyses allowed us to explore the functional consequences of two kinds of interactions that are rarely studied in regulatory sequences: interactions among multiple *cis* variants and interactions between *cis* variants and the differentiated state of cells. Our results indicate that variants within this 3-kb region are involved in both kinds of interactions, resulting in a complex relationship between segregating genetic variation within a population and gene expression traits. We also see some evidence of sex-specific effects on regulatory variants. These complex dynamics will likely differ between populations with additional segregating variants, highlighting the need to examine natural variation when dissecting regulatory function. These findings, together with those from a handful of other studies (Horan et al. 2003; Tao et al. 2006; Ayroles et al. 2009; Warner et al. 2009), suggest that *cis*-by-*cis*, *cis*-by-sex, and *cis*-by-cell type interactions may be more common within regulatory regions than is generally appreciated.

Materials and Methods

Genotyping and Allele-Specific *In Vivo* Assays

Brain tissue was obtained through the Kathleen Price Bryan Alzheimer's Disease Brain Bank at Duke University. Postmortem tissue samples were from the same brain regions as those described in Cirulli and Goldstein (2007). Although all of the samples included in this study were derived from individuals self-identified as being of European descent, it

Table 2. Multiple Regressions for *PDYN In Vivo* Expression.

Brain Region	R^2	Coefficient for SNP(−156)[a]	Standard Error of Coefficient	*P* Value[b]
A. Multiple regressions of allelic imbalance on sequence features				
Frontal cortex	0.88	−0.38	0.1	0.0019*
Occipital cortex	0.87	−0.55	0.13	0.001*
Cerebellum	0.72	−0.32	0.16	0.06
Temporal cortex	0.86	−0.51	0.13	0.0015*
	Mean	Standard Error of Mean	*P* Value	
B. Expression from haplotypes containing A versus G at SNP(−156)[c]				
Frontal cortex	−0.4	0.07	0.0002*	
Occipital cortex	−0.59	0.07	<0.0001*	
Cerebellum	−0.29	0.11	0.02	
Temporal cortex	−0.52	0.06	<0.0001*	

[a] Partial regression coefficients are shown for SNP(−156) only, as no others were statistically significant.
[b] An asterisk indicates a *P* value < 0.01 from a two-tailed Wilcoxon ranked-sum test.
[c] Allelic imbalance specifically in relation to SNP(−156).

is important to note that the specific individuals included in the *in vivo* and *in vitro* analyses comprise distinct groups of individuals. The individuals included comprised 14 females and 10 males (supplementary table S3, Supplementary Material online). Genotypes of the regulatory single nucleotide polymorphisms (SNPs) in the individuals included in the *in vivo* assays were obtained through polymerase chain reaction (PCR) amplification using the high-fidelity polymerase Phusion (Finnzymes, Espoo, Finland) and then were directly bidirectionally sequenced (supplementary table S3, Supplementary Material online). The microsatellite genotypes were obtained by sizing PCR-amplified DNA fragments with an incorporated 5′ fluorescent-labeled primer on an ABI 3700 automated capillary sequencer (Applied Biosystems, Foster City, CA). Marker genotypes were assigned using the program Genotyper (Applied Biosystems). Pyrosequencing primers were designed around one reporter SNP (rs910080T/C) to analyze allele-specific expression. Allele-specific expression was performed one to three times, with the majority of samples being represented by 2. Each run consisted of three replicates of each cDNA sample. The raw data obtained from pyrosequencing and real-time PCR were controlled for quality before being analyzed as in Cirulli and Goldstein (2007). If the peak heights for both alleles of the SNP were less than 30, the data were not used. In some cases, peak heights slightly lower than 30 were allowed as long as the alternate allele measured higher. In this instance, samples with lower intensity peak heights were carefully inspected by hand for consistency with other replicate results. Allelic ratios were determined by dividing the score of one allele of the SNP by the score of the other allele. The average allelic ratio of the gDNA for a sample was then determined. To normalize the data, cDNA allelic ratios were divided by the sample's average gDNA ratio for each run. Standard deviations between sample replicates greater than 0.5 were discarded. The average of all normalized cDNA ratios from all replicates were averaged for one ratio per tissue for each sample.

For each tissue (frontal cortex, occipital cortex, cerebellum, and temporal cortex), both a regression tree analysis and a multiple regression analysis of the magnitude of allelic imbalance were performed. Specifically, the base-2 logarithm of the average (over technical replicates), normalized (relative to gDNA) ratio of the level of the lower-expressed reporter allele to the level of the higher-expressed reporter allele was analyzed. Thus, all the ratios were less than one (hence all the logarithms were negative): The smaller the ratio (and the logarithm), the greater the imbalance. The regression tree analyses were carried out in R (R Development Core Team 2005), and the multiple regression analyses were done in Excel. The regression tree analysis is described in detail below. Genotypes are listed in supplementary table S3, Supplementary Material online, and were coded as 0 for homozygosity and 1 for heterozygosity, and the intercept was set to 0 (no heterozygosity should imply no allelic imbalance).

Haplotype Phasing and the Effect of SNP(−156)

We were able to infer phase between SNP(−156) and the reporter SNP rs910080 by analyzing the genotypes for 55 individuals in the Duke Brain Bank using the program PHASE 2.1 (Stephens et al. 2001; Stephens and Scheet 2005). All phase inferred haplotypes had posterior probabilities >0.96. Knowing which allele at SNP(−156), A or G, was physically linked to each allele, C or T, at the reporter SNP in each sampled individual heterozygous for SNP(−156) enabled us to ascertain whether higher expression was associated with A or G at SNP(−156). For these individuals, the measured reporter T-to-C ratio was taken to be the A-to-G or G-to-A ratio, according to the inferred haplotype of the individual. The *P* values in table 2 are two-tailed *P* values for Wilcoxon's ranked-sum test, performed in JMP (SAS Institute Inc.)

Cloning and Sequencing

Constructs were made from DNA received from anonymous genomic DNA from Austrian populations (Stogmann et al. 2002). Constructs were made from a subset of the haplotypes used in the analysis of Rockman et al. (2005) (supplementary table S2, Supplementary Material online). These particular haplotypes were selected to maximize the

number of *cis*-regulatory variants assayed from the larger population of haplotypes in Rockman et al. (2005). The 3-kb *PDYN* *cis*-regulatory haplotypes were isolated through PCR amplification using the high-fidelity polymerase Phusion (Finnzymes, Espoo, Finland). Individual PCR-amplification products were cloned into pGL3-basic luciferase reporter vector using an *Acc*65I restriction site on the 5′ end and an *Nhe*I site incorporated into the primer on the 3′ end. Constructs were then prepared using the Wizard midi-prep kit (Promega, Madison, WI) and sequenced against references from Rockman et al. (2005).

PDYN Reverse Transcription PCR

To verify that SH-SY5Y and IMR-32 cells naturally express *PDYN*, fragments of cDNA were reverse transcribed and then PCR amplified with primers jumping between exons 3 and 4. RNA was extracted from SH-SY5Y and IMR-32 cells with an Aurum total RNA extraction kit (Bio-Rad, Hercules, CA). The cDNA single-strand synthesis used a high capacity cDNA archive kit (Applied Biosystems). PCRs were then completed with the absolute qPCR SYBR kit (Abgene, Epsom, United Kingdom). In both cell types, the expected amplicon was observed at 94 bp (supplementary fig. S1, Supplementary Material online). These results confirm that these cells are appropriate models for *PDYN* expression and that the expression readings observed were not an artifact of experimental conditions.

Transfection, Cell Culture, and Expression Measurement

SH-SY5Y cells were cultured in a 1:1 mixture of Ham's F12K and MEME (ATCC, Manassas, VA) with 1 mM sodium pyruvate and 0.1 mM nonessential amino acids (Gibco BRL, Gaithersburg, MD), supplemented with 10% fetal bovine serum (HyClone, Logan, UT). IMR-32 cells were cultured in MEME (Sigma–Aldrich Corp St Louis, MO) with 1 mM sodium pyruvate and 0.1 mM nonessential amino acids (Gibco BRL, supplemented with 10% FBS (ATCC). Both cell lines were acquired from ATCC and maintained at 37 °C and 5% CO_2. Transfections were performed in 24-well plates. SH-SY5Y and IMR-32 cells were seeded at 8×10^5 cells/ml at a volume of 500 μl. Cells were transfected 24 h after seeding using Lipofectamine 2000 (Invitrogen, Carlsbad, CA). The transfection mixture used was 2 μl Lipofectamine 2000, 100 μl OPTI-MEM, 700-ng reporter construct, and 200 ng of Renilla-TK as a coreporter. The Renilla construct served as an internal control for well-to-well variation. For the control wells, 427 ng of empty pGL3basic was added to the transfection mixture, which is the molar equivalent of the other constructs. Forty-two hours after transfection, cells were lysed with 100 μl of passive lysis buffer (Promega) for 20 min. Lysates were read in the automated 96-well Veritas Luminomiter (Turner Biosystems, Sunnyvale, CA) with autoinjection following procedures obtained from Promega, using a 2-s delay and 10-s read time. To control for variation, we completed all experiments on 3 separate days and on each day transfected six separate wells for each construct.

In Vitro Expression Analysis

The complex and stochastic nature of gene expression makes it difficult to fully control all factors that affect transcription even *in vitro*. In order to separate biological signal from experimental noise, we fitted a mixed-model analysis of variance (ANOVA) to our measurements. Specifically, for each cell type and day, we first computed the ratio between the luciferase reporter and Renilla coreporter expression for each well. We then subtracted the arithmetic mean over six replicates of measured expression from a promoterless pGL3 vector, from each value of measured expression from a vector containing a *PDYN* *cis*-regulatory haplotype, to remove background levels of expression from the analysis. We then computed the base-2 logarithm of the difference. For each cell type, we then used restricted maximum likelihood to fit the "normalization model"

$$y_{ijk} = \mu + H_i + D_j + (HD)_{ij} + \varepsilon_{ijk}, \qquad (1)$$

where y_{ijk} is the logarithmically transformed, background-subtracted value of measured expression for haplotype i (1–19), day j (1–3), and well k (1–6), μ is the overall mean, H_i is the fixed main effect of haplotype i, D_j is the random main effect of day j, $(HD)_{ij}$ is the random interaction effect of haplotype i and day j, and ϵ_{ijk} is the residual. This was done in R (R Development Core Team 2005), and our code is available on request. This model, which resembles models commonly used in the analysis of microarray data (Wolfinger et al. 2001), accounts for not only purely technical, well-to-well variation but also systematic effects of day arising from, for example, day-to-day variation in the cell cultures. R^2 was calculated for fixed and random effects estimated from three separate expression data sets for *PDYN* expression. The R^2 for fixed and random effects, respectively, was 0.32 and 0.21 in the IMR-32 cell line and 0.25 and 0.43 in the SH-SH5Y cell lines. The total R^2 was 0.58 in the IMR-32 cell line and 0.68 in the SH-SY5Y cell line. The fits indicate that random effects are substantial, accounting for 21% and 43% of expression variation in the two cell types. For each cell type, we used the Tukey–Kramer procedure to assess the significance of each pairwise difference between estimated H_i's. The estimated H_i's constitute our best measures of the typical expression from the haplotypes (supplementary table S1, Supplementary Material online). The H_i's were also used as input to the regression-tree analyses described in the next section.

The multiple regression analyses of the in vitro data were done in Excel. SNP(−2746) and SNP(−2343) are in complete linkage disequilibrium (LD) among our haplotypes and hence were treated as one sequence feature.

Regression-Tree Analyses of *In Vitro* and *In Vivo* Data

We also carried out regression-tree analyses on both the *in vivo* and *in vitro* data. The first phase of regression tree building finds a tree accounting for as much variation among fiducial expression levels as possible; this tree has a leaf node for each haplotype. Such a tree is generally overfitted, in that the relevance of a predictor (here a polymorphism) to the expression variation seen in our experimental data is not necessarily representative of its relevance to expression in the later splits of the regression tree. To correct

for this, we crossvalidated the trees by using subsets of each data set and evaluated their performance on the remainder. Such crossvalidation, repeated for many subsets of the measured cases, is the standard approach to pruning a regression tree so as to avoid overfitting (Breiman et al. 1984). The second phase of tree building prunes the tree to the size whose mean-squared prediction error over the crossvalidation subsets is minimal. For example, for the *in vitro* analysis, we had 19 unique haplotypes. For this data set, we used 19-fold leave-one-out crossvalidation, which we favor over the widely used 10-fold crossvalidation. Higher-fold crossvalidation is more accurate but more computationally expensive (Breiman et al. 1984, p. 78); however, our data set is sufficiently small that the computational burden was not problematic. Crossvalidation involved each of the 19 subsets of our haplotypes consisting of all but one haplotype. For each subset and for each size from 1 to 18 leaf nodes, the regression tree of that size was computed and used to predict expression from the excluded haplotype, and the squared prediction error was tabulated. The tree for all 19 haplotypes was pruned to the size whose mean-squared prediction error over the subsets was minimal. Although pruning criteria favoring simpler trees are often used in regression tree analyses, the minimum-error criterion is appropriate for exploratory and descriptive studies such as the present one. (Among the motivations for alternative pruning criteria are fluctuations arising in 10-fold and other nonexhaustive crossvalidation schemes, which do not arise in our scheme.) This same pruning procedure was also used on the *in vivo* data. We performed all these computations using the R system for statistical analyses (R Development Core Team 2005). Our R code is available on request.

Binding-Site Analysis

The haplotype sequences were submitted to AliBaba 2.1 (Grabe 2002) and P-Match (Chekmenev et al. 2005). Additionally, Motif Locator (Thijs et al. 2002) was used for SNP(−156), as the other programs had no predictions for this motif. In all programs, the default parameters were used. Transcription factors that were predicted to bind directly to the polymorphism in the *PDYN* *cis*-regulatory region were then recorded. Transcription factor–binding sites were classified as predicting the transcription factor would invariantly bind to a motif for both all polymorphism states, bind only to ancestral polymorphism states, or bind only at the derived polymorphic state, with ancestral states classified relative to the Great Apes.

Results

Identifying Additional Sequence Features Underlying Variation in *PDYN* Expression

To begin investigating the functional impact of *cis*-regulatory haplotype variation on *PDYN* expression, we measured luciferase expression driven by 19 naturally occurring *PDYN* haplotypes of European origin (Rockman et al. 2005) in two independently derived neuroblastoma cell lines, SH-SY5Y and IMR-32. Both of these cell lines endogenously express *PDYN* constitutively (supplementary fig. S1, Supplementary Material online) and are therefore appropriate for functional tests of its regulatory region. Additionally, in a previous study, we found functional variation in a regulatory region that differed between these cell lines (Warner et al. 2009), suggesting differences in *trans* factors that influence gene expression. The haplotypes we examined comprise a subset of the common variation in haplotypes described earlier (Rockman et al. 2005). Although previous studies have emphasized the impact of the 68-bp repeat on transcription (Zimprich et al. 2000; Nikoshkov et al. 2008), we found no consistent relationship between repeat number and expression in either cell line (fig. 1). However, there are clearly repeatable and statistically significant differences in expression between constructs. These results suggest that 68-bp repeat number is not the only sequence feature that influences *PDYN* expression. We therefore extended our analysis to examine the functional consequences of the other polymorphic sites on expression level, using both *in vivo* and *in vitro* functional assays.

Variants Correlated with Expression Changes *In Vivo*

In vivo assays of gene expression in humans cannot easily control for genetic background and physiological status but nonetheless provide the most biologically relevant measures of expression difference. In order to measure the functional effects of polymorphisms within the 3-kb 5′ flanking region of *PDYN* on *in vivo* expression, we measured allele-specific transcript abundance in four brain regions. Allele-specific measurements of transcript abundance can be used to assay the effects of *cis*-acting variation (Yan et al. 2002; Pastinen and Hudson 2004; Wittkopp et al. 2004). These assays utilize a polymorphism in a transcript (the reporter SNP) as a marker for each of the allele states (the putative functional variants). If variants near the reporter SNP affect expression differently on the two chromosomes, there will be an allelic imbalance, or departure from 1:1 ratio, in the abundance of transcripts bearing the two alleles of the reporter SNP. Because both regulatory haplotypes are exposed to the same complement of *trans*-acting factors, allelic imbalance is most plausibly the result of variation in *cis* (Knight 2004). Variants within 5′ flanking regions, such as the regions examined in this study, are unlikely to influence posttranscriptional events such as alternative splicing or message stability; thus, an association between a 5′ variant and allelic imbalance indicates that it, or a variant in LD with it, influences transcription.

We used pyrosequencing to measure allelic imbalance of *PDYN* transcripts within four brain regions: frontal cortex, temporal cortex, occipital cortex, and cerebellum. These same regions were previously assayed from some of the same individuals and tested for imbalance associated with the 68-bp repeat alone (Cirulli and Goldstein 2007). Here, we analyzed an expanded sample set, both in terms of the number of individuals (23 individuals, all of whom were heterozygous at the reporter SNP) and of possible associated polymorphisms.

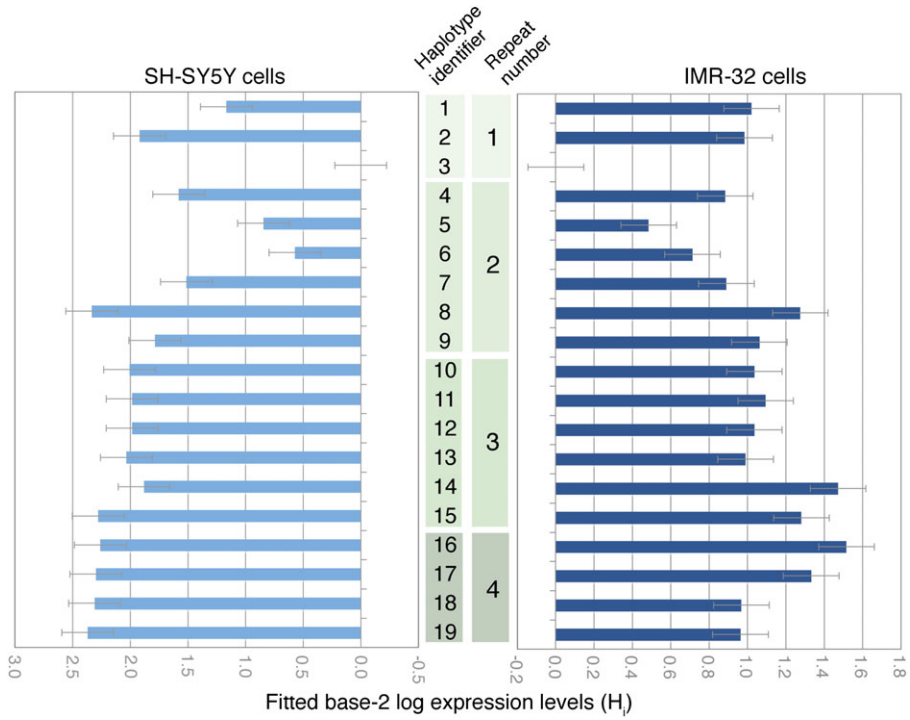

FIG. 1. Fitted expression levels driven by *PDYN cis*-regulatory haplotypes in two neuroblastoma cell lines. For ease of comparison, construct expression has been normalized so that the lowest expressed construct is equal to zero. All other constructs examined are then some value greater than zero. Constructs are organized vertically by repeat number. Error bars represent standard errors of fitted construct effects. The haplotype identifiers are listed, see supplementary table S2, Supplementary Material online, for haplotype-sequence states.

Using multiple regression analyses, we found that heterozygosity for one sequence feature, the A/G SNP at position −156 relative to the transcription start site, is clearly and consistently associated with allelic imbalance (table 2). The association is strongest for occipital cortex, and temporal cortex, weaker for frontal cortex, and weakest for cerebellum (table 2*A* and *B*). Interestingly, 68-bp repeat number was not associated with allelic imbalance in any of these four brain regions, suggesting that this variant might not be functionally important in the regions we examined.

We were able to determine haplotype phasing between SNP(−156) and the reporter SNP (rs910080T/C) using PHASE 2.1 (Stephens et al. 2001; Stephens and Scheet 2005). In order to determine which allelic state at SNP(−156) is associated with higher or lower expression, we first asked whether there was a trend in the ratio of expression in SNP(−156) heterozygotes. Only SNP(−156) was tested, because it was implicated by the multiple regression analysis and inferring phasing of more distal sequence features and the reporter SNP would be less certain. Expression is significantly lower from haplotypes with A at SNP(−156) in three of the brain regions surveyed (table 2*B*). The remaining brain region, cerebellum, shows a similar, but a less robust trend. Clearly, the presence of an A at SNP(−156) is associated with lower expression *in vivo*. A limitation of *in vivo* assays is that we cannot experimentally verify whether this change in expression is caused by SNP(−156) itself or a polymorphism in LD with it. However, it seems unlikely that the other known variants in the region (table 1) are causal, given their failure to exhibit strong associations with allelic imbalance in the multiple regression analyses (table 2) as well as in a regression-tree analysis (discussed below).

We also tested whether any of the polymorphisms showed sex-specific functional effects. For the female individuals, SNP(−156) drops out of the analyses, because all the female samples are A/G at this SNP. The unsigned linear regression analyses for frontal cortex show significant associations with the 68-bp repeat (0.014) and MSAT(−2745) ($P = 0.018$). The 68-bp repeat is also significant in the cerebellum ($P = 0.025$). We also found that there is an SNP within repeat 1 that shows up as significant in the frontal and occipital cortex, as well as the cerebellum ($P = 0.023$, 0.007, and 0.040, respectively) in females. For the male individuals, the data sets are almost too small to analyze, but the linear regressions now show no significant features, including SNP(−156). These data are suggestive of sex-specific interactions in *PDYN* regulation, but as the sample sizes are very small, these results should be interpreted cautiously.

Variants Correlated with Expression Changes *In Vitro*

In vitro assays provide a valuable complement to *in vivo* measurements, because different haplotypes can be tested in the same genetic background and under nearly uniform physiological conditions. Although a transformed cell line growing in culture is somewhat less natural than *in vivo* samples, *in vitro* assays provide substantially greater power

detect genetic effects by controlling for two of the most likely confounding effects. We therefore returned to the 19 natural haplotypes of the *PDYN cis*-regulatory region described earlier and sought to accurately identify specific sequence features that influence expression level.

Our analytical strategy was as follows (see Materials and Methods for details): First, we fitted a mixed-model ANOVA (the normalization model) to our measurements for each cell type, in order to obtain a fiducial expression level for each haplotype in that cell type (supplementary table S1, Supplementary Material online). Next, for each cell type, we performed a multiple regression of the fiducial expression level on haplotype sequence features, coding SNPs as binary (0 vs. 1) variables. The only variant we did not consider is a SNP that lies within the 68-bp repeat and varies within each of them, because the number of permutations of SNP by repeat number is simply too large relative to the number of haplotypes tested.

The multiple regression analysis for expression in SH-SY5Y cells is statistically significant overall (F-statistic $P = 0.0066$), indicating an effect of genotype on expression level (fig. 3*A*). Five sequence features are statistically significant individually (*t*-statistic *P* from 0.002 to 0.019) (table 3). For expression in IMR-32 cells, the regression is marginally significant overall (F-statistic $P = 0.053$), and two sequence features are marginally significant individually (*t*-statistic $P = 0.074$ and 0.097) (table 3). Both of the features weakly implicated in IMR-32 cells are strongly implicated in SH-SY5Y cells. Although copy number of the 68-bp repeat is implicated in SH-SY5Y cells, it is only one of several features implicated and does not have the strongest support. Thus, these analyses suggest that *PDYN* regulation is similar but not identical in these two cell types and that the 68-bp repeat is a relevant but not a dominant factor.

Table 3. Multiple Regression of Haplotype Sequence Features on Fiducial Expression *In Vitro*.

Sequence Feature	Coefficient	Standard of Coefficient	*P* Value[a]
SH-SY5Y			
SNP 2746/SNP −2343	−1.85	0.49	0.004**
MSAT −2745	0.32	0.19	0.13
SNP −2357	0.31	0.3	0.32
SNP −2116	−2.36	0.82	0.018**
SNP −2081	0.25	0.3	0.42
68-bp Repeat count	0.22	0.07	0.016**
SNP −977	−0.04	0.34	0.92
MSAT −362	−1.1	0.25	0.002**
SNP −156	1.8	0.49	0.005**
IMR-32			
SNP −2746/SNP −2343	−0.7	0.44	0.14
MSAT −2745	0.04	0.17	0.83
SNP -2357	−0.11	0.27	0.7
SNP −2116	−0.44	0.73	0.56
SNP −2081	−0.05	0.27	0.87
68-bp Repeat count	0.1	0.07	0.15
SNP −977	0	0.31	1
MSAT −362	−0.45	0.22	0.073*
SNP −156	0.81	0.44	0.097*

[a] One asterisk indicates a *P* value $0.05 > P < 0.01$ and two indicate $P < 0.05$.

Regression-Tree Analyses of the *In Vivo* Assays

Although multiple regression analysis is the conventional statistical approach to testing for an association between a specific genetic variant and expression level, this approach is not suited to a situation where multiple variants interact nonadditively. Such interactions are likely to be common within *cis*-regulatory regions, given that many transcription factors bind cooperatively to DNA and to each other (reviewed in Lemon and Tjian 2000). Although interaction terms can be added to regression models, the number of potential pairwise interactions is typically large; in our case, this number exceeds the number of haplotypes or individuals we had available for analysis. A second limitation of regression analysis is that it assumes additive dependence on sequence features. However, there is no particular reason to expect additive effects, or even monotonic effects when more than two alleles are present, as in microsatellites.

Therefore, we carried out regression-tree analyses as a complement to multiple regression analyses. We constructed a regression tree of fiducial expression level on haplotype sequence features for each cell type or brain region. The goal of a regression-tree analysis is to identify a branching sequence of if–then conditions, represented by nodes in a tree, that accurately predict a response variable using observed values of predictor variables, where the predictors may be subject to complicated interdependencies (Breiman et al. 1984). Unlike multiple regression analyses, regression-tree analyses do not assess statistical significance of predictors, nor of the splits they govern. Instead, the emphasis is on predictive accuracy, as demonstrated by cross-validation. Attractive qualities of regression-tree analysis include its modest assumptions (e.g., it assumes nothing about the probability distribution of the response given the predictors) and interpretive simplicity. Here, the response is expression in a given cell type driven by a haplotype (after accounting for random variation via the normalization model), and the predictors are sequence features of the haplotype.

We began by revisiting the allelic imbalance results, which assay *in vivo* expression (fig. 2). For a given cell type, the sequence feature that is the strongest predictor of expression is the first split in the tree, and subsequent splits are the next strongest, and so on. For each split in figure 2, the subtree to the right (red) represents haplotypes yielding higher expression, whereas the subtree to the left (blue) represents haplotypes yielding lower expression. The following information is included in each box: the name of a sequence feature, the state of the sequence feature, the average deviation (on a base-2 logarithmic scale) of expression for haplotypes having the given state of the sequence feature from average expression for all haplotypes, and the number of haplotypes having the given state of the sequence feature. Each split is annotated with the percentage of expression variation (sum of squared deviations from mean) it accounts for. As in the multiple regression analyses, SNP(−156) was the feature most strongly implicated

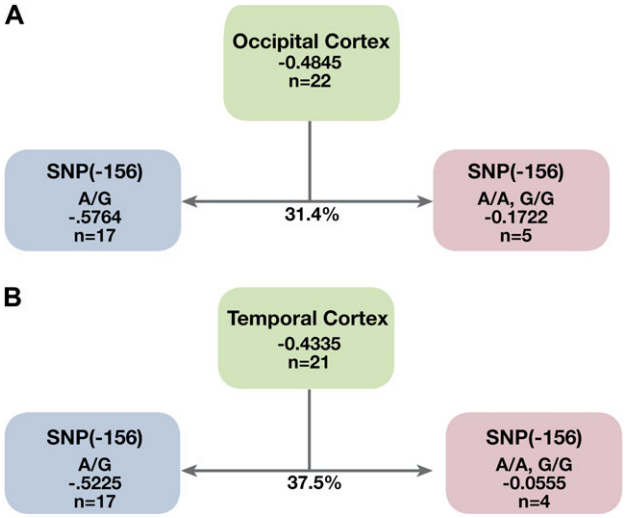

FIG. 2. Regression tree analyses of *PDYN* expression *in vivo*. (*A*,*B*) Regression trees for *PDYN in vivo* expression. (*A*) Regression tree for allelic imbalance in occipital cortex. (*B*) Regression tree for allelic imbalance in temporal cortex.

as affecting expression level (fig. 2*A* and *B*). The same two brain regions, occipital cortex and temporal cortex, showed the strongest support for affects of SNP(−156) on allelic imbalance in both multiple regression and regression tree analyses (compare table 2 and fig. 2). For the sex-specific analysis, SNP(−156) remains significant in males, but all females are heterozygous at this SNP, so it cannot be analyzed. For the females, the only nontrivial tree in seen in occipital cortex for the SNP in repeat one. The 68-bp repeat was not associated with allelic imbalance in any of the four brain regions. Thus, SNP(−156) stands out as the variant most strongly associated with expression level by two different methods of analysis and in the same two brain regions.

Regression-Tree Analyses of the *In Vitro* Assays

Next, we analyzed reporter gene results, which assay effects on expression *in vitro*. As expected, more variants are implicated by the *in vitro* than the *in vivo* measurements (compare fig. 2 and fig. 3). The regression trees for the SH-SY5Y cell line (fig. 3*A*) indicate that the microsatellite 2,745-bp upstream of the transcription start site is the most important factor in explaining differences in expression between haplotypes. The presence of 18, 19, or 21 copies of the dinucleotide $(GT)_n$ at MSAT(−2745) within a given haplotype leads to higher expression, whereas with 23 repeat expression was lower. For those haplotypes that had 21 or fewer repeats of the microsatellite, the presence of three or four copies of the 68-bp repeat was associated with higher expression than one or two copies, consistent with results from previous studies (Zimprich et al. 2000; Nikoshkov et al. 2008). For those haplotypes containing one or two copies of the 68-bp repeat, a second microsatellite, MSAT(−362), was correlated with expression differences. Again, fewer repeats was correlated with higher expression. The fourth variant implicated as affecting expression is SNP(−2746), which is physically adjacent to the most predictive feature, MSAT(−2745).

The regression trees for the IMR-32 cell type (fig. 3*B*) also showed that the most important predictor of differences in expression between haplotypes is the number of repeats in MSAT(−2745). The 68-bp repeat number is again the second-best predictor of expression, with those haplotypes containing three or four copies showing higher expression than those with one or two copies (fig. 3*B*). Next, MSAT(−2745) is further associated with expression differences; haplotypes with 21 repeats have higher expression than haplotypes with 18 or 19 repeats. Finally, the 68-bp repeat appears again on the tree, with the two-repeat haplotype showing higher expression than the one-repeat haplotype, although this result is based on only two haplotypes.

In these trees, splits nearer the top tend to account for larger percentages of the expression variation among haplotypes. Although regression-tree analysis does not assign *P* values to predictors (as multiple regression analysis does), the percentage of variation accounted for by all splits involving a predictor quantifies the importance of the predictor. Thus, our *in vitro* data strongly implicate MSAT(−2745) and the 68-bp repeat, which, respectively, are involved in the first- and second-level splits in every tree and account for 62.9% and 18.7% of the expression variation on the average in these analyses.

In SH-SY5Y cells, all but one of the sequence features implicated by regression tree analysis are also implicated by multiple regression analysis. The exception, MSAT(−2745), is instructive. Ignoring other features, the relationship between expression level and repeat count at MSAT(−2745) is not monotonic: Haplotypes containing 19 or 21 repeats are associated with higher expression than those containing 18 or 23 repeats. In general, such a relationship with one predictor can arise through correlations among predictors, but this does not appear to be the case here. MSAT(−2745) is strongly correlated (magnitude of correlation coefficient greater than 0.5, when SNPs are coded as binary variables) with two other features, SNP(−2746)/SNP(−2343) (these two SNPs are in complete LD among our haplotypes, so we analyze them as one feature) and SNP(−2116). Among haplotypes with C/A at SNP(−2746)/SNP(−2343), the relationship with MSAT(−2745) remains nonmonotonic, and although no equally informative restriction with respect to SNP(−2116) exists, haplotypes with G at SNP(−2116) include both the highest- and the lowest-expressing haplotypes in SH-SY5Y cells, with 21 and 23 repeats at MSAT(−2745), respectively. Because the statistical significance of SNP(−2116) in our multiple regression analysis arises from the latter two haplotypes and is not predictive in the regression trees, we have not included it in our set of implicated variants. Thus, although it is impossible to be certain without making and measuring haplotypes varying solely at MSAT(−2745), our data suggest that the relationship between expression and MSAT(−2745) may be genuinely nonmonotonic, at least in some genetic backgrounds. Previous studies have shown that polymorphisms in dinucleotide repeat number can

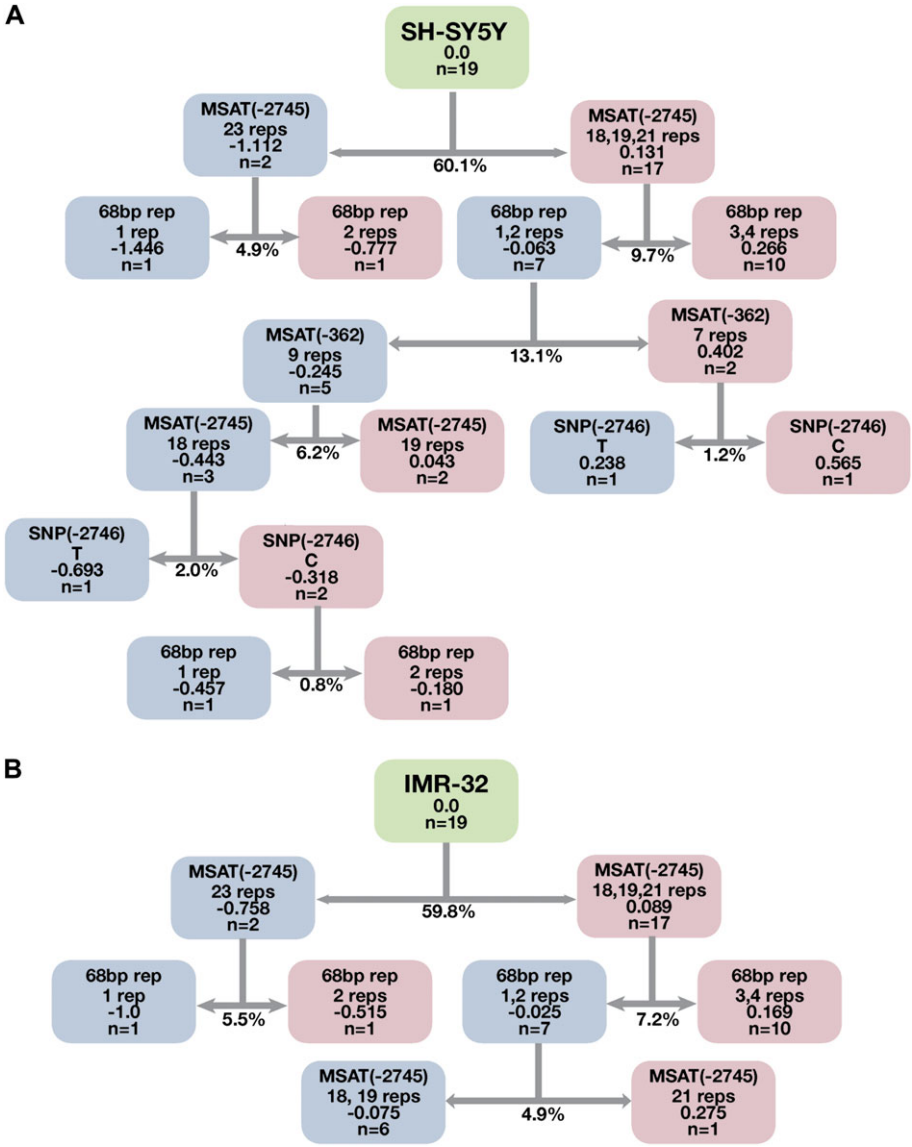

Fig. 3. Regression trees for *PDYN in vitro* expression. (*A*,*B*) Regression trees follow the same format as in figure 2. (*A*) Regression trees for expression in SH-SY5Y cells. (*B*) Regression trees for expression in IMR-32 cells.

have a large impact on expression (Rothenburg et al. 2001) and can effect expression in a nonmonotonic manner (Chen et al. 2007). Given the apparent complexity of the relationship, it is not surprising that multiple regression failed to recognize the importance of MSAT(−2745). The relationship appears to be similarly complex for IMR-32 cells, presumably contributing to the larger *P* values of multiple regression for this cell type. For both cell types, the regression trees account for more variance with fewer predictors than the multiple regressions (SH-SY5Y: 98% with four predictors vs. 86% with nine predictors; IMR-32: 77% with two predictors vs. 76% with nine predictors).

Binding-Site Analysis

In order to begin exploring the possible molecular bases for the polymorphisms that influence expression, we scanned each region containing one of the variants for putative transcription factor–binding sites using a variety of bioinformatic tools (Grabe 2002; Thijs et al. 2002; Chekmenev et al. 2005). Eleven of the 13 known variants contained at least one possible transcription factor–binding site. Additionally, nine of the 11 variants that had at least one implicated binding site also showed that different transcription factors were expected to bind the ancestral and derived variants. However, the bioinformatic predictions varied substantially depending on the program and parameters used.

MSAT(−2745), the variant associated with the largest effects *in vitro*, and MSAT(−362) do not have any predicted transcription factor–binding sites that vary between repeat state. Within the 68-bp repeat, three potential interactions were found between potential binding proteins. In addition, one of these interactions was lost when the ancestral repeat internal SNP is replaced with the derived SNP. Previous work by Zimprich et al. (2000) confirmed

a biochemical interaction between the transcription factor AP-1 and the 68-bp repeat region, which is one of the putative factors that was predicted bioinformatically. Within the 68-bp repeat, the transcription factors ER and SP1 are also both predicted to bind. These proteins are known to interact with each other, enhancing SP1 binding (Sun et al. 1998). In addition, ER is known to interact with GATA-1, which was found to have a putative binding site only in repeats containing the ancestral SNP. GATA-1 is also known to interact with SP1 (Fischer et al. 1993) and GATA-1 binding is predicted to be lost in repeats with the derived mutation. The same interaction between GATA-1 and SP1 may also occur between SNP(−2116) and SNP(−2081), where GATA-1 is only implicated as being able to bind with the derived mutation at SNP(−2116). AliBaba and P-Match do not predict binding sites at SNP(−156), but other prediction algorithms, Motif Locator and Motif Scanner (Thijs et al. 2002) predict both an AP-1 and MEIS-1 site for SNP(−156)A and not for SNP(−156)G. Note, however, that binding-site identification based on sequence alone generally produces substantial rates of false positives and false negatives (Tompa et al. 2005), so these results must be interpreted cautiously.

Discussion

Multiple Segregating Variants in *cis* Influence *PDYN* Transcription

We used both *in vivo* and *in vitro* assays to search for functional variants within a 3-kb region 5′ of *PDYN* that affect its expression. The results from both multiple regression and regression-tree analyses implicate five segregating variants in transcriptional regulation: one previously examined in several association studies (Chen et al. 2002; Stogmann et al. 2002; Ventriglia et al. 2002; Ray et al. 2005; Nomura et al. 2006; Nikoshkov et al. 2008) and molecular analyses (Zimprich et al. 2000), a second in association studies (Geijer et al. 1997; Xuei et al. 2006; Yuferov et al. 2009), and three additional variants not previously recognized as functional in their impact on gene expression (table 1). Yuferov et al. (2009) recently determined that three SNPs in perfect LD within the 3′ UTR also affect expression. Thus, at least six common functional variants have been identified that influence *PDYN* expression to date.

The complex expression profile of *PDYN* makes understanding the consequences of genetic variation within its *cis*-regulatory region challenging. *PDYN* is expressed in many areas of the brain (Telkov et al. 1998; Hurd 2002; Nikoshkov et al. 2005), as well as in other tissues, including the spinal cord (Ji et al. 2002), immune cells (Sun et al. 2006), and both testes and ovaries (Civelli et al. 1985; Kaynard et al. 1992). Genes with complex expression profiles may be more likely to harbor multiple functional genetic variants than those with simple expression profiles for two reasons: First, their *cis*-regulatory regions are likely to be more extensive and therefore provide a larger mutational target. Second, their protein products are more likely to be involved in multiple biological processes, increasing the likelihood of balancing selection among diverse functional demands. *PDYN* meets both criteria, with a moderately extensive *cis*-regulatory region (Carrion et al. 1999; Zimprich et al. 2000). The six segregating variants that affect *PDYN* expression (table 1) may therefore be atypical of human genes as a whole but not particularly unusual for genes with complex expression profiles.

There is also evidence of balancing selection among human populations (Rockman et al. 2005). Balancing selection can maintain advantageous genetic diversity in populations and, therefore, maintain phenotypic variation. Selection apparently drove independent increases in the frequency of the two- and three-repeat alleles in different populations (Rockman et al. 2005). There is also evidence that MSAT(−2745) is under positive selection within populations, as shown by reduced variation at this microsatellite within certain populations. This reduction of variance was originally interpreted as being due to linkage with the 68-bp repeat (Rockman et al. 2005) but could also be due to the independent role of MSAT(−2745) now that it is known to have a distinct functional consequence (this study).

Most genes with clear signatures of balancing selection in humans are associated with immune function or protection (e.g., Takahata et al. 1992; Peiper et al. 1995; Tournamille et al. 1995; Verrelli et al. 2002; Tung et al. 2009). *PDYN* provides an interesting example of balancing selection on the regulation of a gene primarily involved in behavioral phenotypes. *PDYN* products influence a number of physiological phenotypes, such as reward, mood regulation, stress response, and motor function (Drolet et al. 2001; Hauser et al. 2005). It is possible that the signatures of recent selection at this locus have been driven by differences in the use of environmental opioids or different triggers of endogenous opioids. Other neuropeptides have been experimentally shown to alter behaviors in other species (Lim et al. 2004; Saetre et al. 2004). Understanding the links between the polymorphisms in different populations affecting *PDYN* expression and the fitness consequences of that change poses a substantial, but important, challenge.

A prediction based on the results of this study is that there are most likely additional functional polymorphisms segregating in other populations and that between populations there may be differences in the relative importance of a polymorphism on expression and on fitness. For example, SNP(−156)A, here driving lower expression levels, is maintained at a much higher allele frequency in the CEPH (Utah residents with ancestry from northern and western Europe) population than the other HapMap populations (The International HapMap Consortium 2007). One practical impact is that future association studies of *PDYN* expression with disease phenotypes might benefit from deep sequencing in this region and in multiple populations. These data would allow for a more detailed understanding of the important regulatory variants beyond the 68-bp repeat in population(s) of interest.

For a variety of reasons, the existence of additional functional variants cannot be ruled out. We measured allelic imbalance from just four regions of the adult central

nervous system and examined the haplotypes in two cell lines, a small subset of the full range of locations where *PDYN* is expressed. In addition, we sampled a small fraction of natural genetic variation: 23 individuals for the *in vivo* assays and 19 haplotypes for the *in vitro* assays. Furthermore, we only assayed for effects on transcript abundance, but additional variants could influence *PDYN* expression through utilization of alternate transcription start sites or alternative splicing downstream of the canonical transcription start site (Telkov et al. 1998; Nikoshkov et al. 2005). (The reporter SNP used in the allelic imbalance assays in this study is present in all known *PDYN* alternative transcripts and would therefore not identify variants influencing these other regulatory processes.)

The number of cases where multiple functional variants in *cis* have been shown to influence the expression of a given gene is small but growing. Examples include *HG1* (Horan et al. 2003), *KRT1* (Tao et al. 2006), *LCT* (Enattah et al. 2002; Tishkoff et al. 2007; Enattah et al. 2008), *TH* (Warner et al. 2009), and *PDYN* (Zimprich et al. 2000). It is worth noting, however, that relatively few studies have explicitly searched for multiple variants that influence the expression of a particular gene in humans. More cases need to be examined in detail in order to understand what fraction of human genes harbor multiple regulatory polymorphisms effecting transcription and how these regions are being shaped by selection within populations.

Cell Type- and Sex-Specific Influences on the Expression Phenotypes of Functional Variants

The spatial heterogeneity of *PDYN* expression within the brain in particular (Telkov et al. 1998; Hurd 2002; Nikoshkov et al. 2005) raises the possibility that functional variants could affect restricted regions of its overall expression profile, through interactions with *trans*-acting factors specific to subsets of cell types. Because transcriptional activators often differ among tissues, this may be a relatively common situation. Evidence for interactions with *trans*-acting factors come from two kinds of comparisons: First, the rank order of haplotypes differs among cell lines in the *in vitro* assays (fig. 1), as does the order of appearance of variants in the regression tree (fig. 3*A* and *B*). Two of the variants, MSAT(−362) and SNP(−2746), are implicated in SH-SY5Y but not IMR-32 cells (fig. 3*A*). These differences are likely driven by *trans* effects in the differentiated state of the neuroblastoma cells and demonstrate the utility of examining regulatory regions of genes with pleiotropic effects in multiple contexts.

Second, although *PDYN* is endogenously expressed in all the brain regions that we assayed, expression-trait consequences sometimes differed. For instance, SNP(−156) explains expression differences *in vivo* to different degrees among different brain regions (table 2 and fig. 2). Because these assays control for genetic background, these differences are driven by the differentiated state of the cells included in the tissue samples, with SNP(−156), or a SNP in LD with it, playing a more or less important role, depending on the context. Brain region–specific consequences of variants affecting *PDYN* expression have previously been noted for the 68-bp repeat (Nikoshkov et al. 2008) and SNPs in the 3′ UTR (Yuferov et al. 2009). Although we only found a statistically significant correlation for the 68-bp variant *in vivo* in a subset of samples and regions, we examined different regions of the brain than Nikoshkov et al. (2008) examined, and that study included relatively few female samples. Together, the results of these three studies suggest that genetic differences in *cis* affect interactions with *trans*-acting factors to alter *PDYN* expression and that these interactions are mediated through at least three different variants in *cis*, namely, the 68-bp repeat (Nikoshkov et al. 2008), SNP(−156) (this study), and one or more of three SNPs in the 3′ UTR that are in perfect LD (Yuferov et al. 2009). The only variant we found to affect expression both *in vivo* and *in vitro* in all of the samples was SNP(−156) (figs. 2 and 3). The fact that the other four variants only showed measurable effects on expression in the cell culture assays may reflect biological reality or the greater sensitivity of *in vitro* assays to detect subtle functional consequences.

As *in vivo* tissue samples are a complex mix of neuronal and glial cell types and that cell types express distinct suites of transcription factors, differences in expression consequences among brain regions and cell types are perhaps unsurprising. These results suggest that cellular environment influences the trait consequences of most of the known variants that influence *PDYN* expression.

Our results are also suggestive of sex-specific interactions with the regulatory variants of *PDYN*. There are now a number of studies from model systems (Bhasin et al. 2008; e.g., Ayroles et al. 2009; and reviewed in Williams and Carroll 2009) demonstrating significant effects of gender on gene expression. There is also evidence of global sex-specific changes in primate brain gene expression (Reinius et al. 2008) and in disease susceptibilities in humans (reviewed in Ober et al. 2008). We see effects of the 68-bp repeat, an SNP within the first repeat, and MSAT(−2745) in the female individuals. The SNP within the repeat may be another potential functional variant, although with our small sample size, we cannot distinguish its effects from the effect of repeat number. That we see significant interactions with the 68-bp repeat and MSAT(−2745) in the *in vivo* assays in female individuals may mean that these variants play as or more important a functional role as SNP(−156) in *PDYN* expression in female frontal cortex and that those roles may change between brain regions (here comparing between frontal cortex and cerebellum). Our sample sizes for these analyses are very small, but these results are reasonably suggestive of sex-specific effects of individual variants. Neuropsychological association studies of *PDYN* may benefit from both combined and sex-specific association analyses to understand variants involved in disease susceptibilities or progression.

Functional Variants Interact and Can Be Suggestive of Nonadditive Interactions

With multiple variants affecting *PDYN* expression segregating in human populations, it becomes important to

understand whether their effects are additive or epistatic. Regression-Tree analysis can be particularly informative in understanding interactions between predictors, among which there may be both causal (epistasis) and statistical (LD) interdependencies. Our results provide several possible cases of epistatic interactions. For instance, 68-bp repeat number has different consequences for expression, depending on repeat number of MSAT(−2745) (fig. 3*A* and *B*). Similarly, the consequences of SNP(−2746) depend on 68-bp repeat number (fig. 3*A* and *B*). Tracing through the regression trees reveals more complex situations. For example, fewer repeats of MSAT(−2745) are predictive of higher expression on the first split in the IMR-32 tree (fig. 3*B*), with the 68-bp repeat as the next split and then another split on the state of MSAT(−2745). This back-and-forth appearance on the same regression tree may be indicative of interactions between different states of these sequence features.

We also found a case where a multiallelic variant has nonmonotonic effects on expression. Haplotypes containing 19 or 21 repeats of MSAT(−2745) are associated with higher expression than those containing 18 or 23 repeats when other features are not considered (see fig. 3*A* and Results for details). This effect could arise through interactions among sites or by repeat number affecting secondary structure (Iglesias et al. 2004).

Because so few genes are known to harbor multiple functional variants that influence transcription (see above), it is perhaps not surprising that there are even fewer cases where interactions among variants have been shown to be either additive or epistatic in nature. Measuring these interactions will generally require examining many haplotypes in order to determine whether effects on expression are truly causal or due to LD and whether real effects are additive or epistatic. *In vitro* assays may be particularly useful in this regard, as it is possible to engineer specific combinations of variants not found naturally due to LD and test their impact on expression level.

The 68-bp Repeat Influences *PDYN* Expression

The 68-bp repeat has been the focus of several prior studies (Zimprich et al. 2000) and therefore merits special mention. Our results suggest that the number of 68-bp repeats has a functional impact, although this variant was not a robust predictor of expression level, significant only in the two cell lines and in frontal cortex and cerebellum in the female *in vivo* analysis (table 3 and fig. 3). This finding falls somewhere between the results of Cirulli and Goldstein (2007), who found no association, and those of Zimprich et al. (2000) and Nikoshkov et al. (2008), who reported that a higher repeat number may drive higher expression *in vitro* and *in vivo*. These discrepancies could arise for a number of biological or technical reasons. Cirulli and Goldstein (2007) examined *in vivo* expression in fewer individuals than we examined. Nikoshkov et al. (2008) surveyed different brain regions than we examined; their results may indicate expression traits specific to particular brain regions rather than an inconsistency among studies. Finally, Zimprich et al. (2000) used a mouse/rat neuroblastoma/glioma hybrid cell line for transient transfection assays rather than human cell lines, and their expression constructs incorporated smaller portion of the flanking region that did not contain all of the functional variants examined here. This last case highlights the importance of examining regulatory regions in a species-specific context, as there are well-documented genome-wide changes in regulatory regions between human and mouse (Odom et al. 2007; Wilson et al. 2008).

Consistent with previous studies (Zimprich et al. 2000; Nikoshkov et al. 2008), our *in vitro* analyses found that haplotypes containing three and four 68-bp repeats associate with increased expression, whereas one and two repeats were associated with lower expression (fig. 3). An increase in expression with repeat number implies that the repeat contains binding sites for transcriptional activators. Within the repeat, there is an empirically documented binding site for the transcriptional activator AP-1 (Zimprich et al. 2000), whereas our bioinformatic screens (see Materials and Methods) identified multiple additional possible transcription factor–binding sites, almost all of which bind transcriptional activators. Although not empirically tested here, these results are compatible with the consistent finding that higher repeat number leads to higher expression (Zimprich et al. 2000; Nikoshkov et al. 2008). Collectively, our analysis and previous studies (Zimprich et al. 2000; Nikoshkov et al. 2008) provide evidence that variation in 68-bp repeat number is functionally significant. Evidence that recent natural selection has altered repeat-number allele frequencies among human populations (Rockman et al. 2005) further suggests that the resulting expression variation may have consequences for trait variation and fitness.

Conclusion

The impact of genetic variation on *PDYN* expression is unexpectedly complex. At least six different variants are now known to affect transcript abundance, all of which are segregating in human populations. Some of these variants show cell type–specific and sex–specific expression-trait consequences, implying interactions with *trans*-acting factors that are themselves differentially expressed among cell types. Segregating variants interact with each other to influence *PDYN* expression, in some cases nonadditively and, in the case of some multiallelic variants, nonmonotonically. The 68-bp repeat that has been the focus of several previous studies is not a consistent predictor of expression level across all expression environments and may be important in only a subset of the biological functions of *PDYN*; other polymorphisms, most notably MSAT(−2745) and SNP(−156), may be the primary drivers of expression variation in some tissues. For future studies, it should be clear that focusing exclusively on variation in the 68-bp repeat will provide too simplistic of a view of *PDYN cis*-regulatory function and trait associations and that other polymorphims may be playing important functional roles within a given population. Whether *PDYN* is an outlier or a fairly

typical example of the complex relationship between genetic variation and expression trait consequences can only be answered through detailed functional analyses of additional genes.

Supplementary Material

Supplementary tables S1–S3 and supplementary figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Funding for this project came from HOMINID grant NSF-BCS-08-27552 from the NSF and from the Institute for Genome Sciences & Policy of Duke University. R.H. was supported by an NSF Postdoctoral Fellowship in Biological Informatics. The authors would like to thank Jenny Tung and William Nielsen for assistance with pyrosequencing. Fritz Zimprich provided the genomic DNA samples used to create the expression constructs and the Kathleen Price Bryan Alzheimer's Disease Brain Bank at Duke University provided tissues samples used for the allelic imbalance measurements. We would also like to thank the members of the Wray laboratory and two anonymous reviewers for suggestions and comments.

References

Ayroles JF, Carbone MA, Stone EA, et al. (11 co-authors). 2009. Systems genetics of complex traits in *Drosophila melanogaster*. *Nat Genet*. 41:299–307.

Bamshad MJ, Mummidi S, Gonzalez E, et al. (11 co-authors). 2002. A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. *Proc Natl Acad Sci U S A*. 99:10539–10544.

Bhasin JM, Chakrabarti E, Peng DQ, Kulkarni A, Chen X, Smith JD. 2008. Sex specific gene regulation and expression QTLs in mouse macrophages from a strain intercross. *PLoS One*. 3:e1435.

Boodhoo A, Wong AM, Williamson D, Voon D, Lee S, Allcock RF, Price P. 2004. A promoter polymorphism in the central MHC gene, IKBL, influences the binding of transcription factors USF1 and E47 on disease-associated haplotypes. *Gene Expr*. 12:1–11.

Breiman L, Friedman J, Olshen R, Stone CJ. 1984. Classification and regression trees. New York: Chapman and Hall.

Carrion AM, Link WA, Ledo F, Mellstrom B, Naranjo JR. 1999. DREAM is a Ca^{2+}-regulated transcriptional repressor. *Nature* 398:80–84.

Chekmenev DS, Haid C, Kel AE. 2005. P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res*. 33:W432–W437.

Chen ACH, LaForge KS, Ho A, McHugh PF, Kellogg S, Bell K, Schluger RP, Leal SM, Kreek MJ. 2002. Potentially functional polymorphism in the promoter region of *prodynorphin* gene may be associated with protection against cocaine dependence or abuse. *Am J Med Genet*. 114:429–435.

Chen T-M, Kuo P-L, Hsa C-H, Tsai S-J, Chen M-J, Lin C-W, Sun HS. 2007. Microsatellite in the 3′ untranslated region of huma fibroblast growth factor 9 (FGF9) gene exhibits pleiotropic effect of modulating FGF9 protein expression. *Hum Mutat*. 28:98.

Cirulli ET, Goldstein DB. 2007. In vitro assays fail to predict in vivo effects of regulatory polymorphisms. *Hum Mol Genet*. 16:1931–1939.

Civelli O, Douglass J, Goldstein A, Herbert E. 1985. Sequence and expression of the rat *prodynorphin* gene. *Proc Natl Acad Sci U S A*. 82:4291–4295.

Corbett AD, Henderson G, McKnight AT, Paterson SJ. 2006. 75 years of opioid research: the exciting but vain quest for the Holy Grail. *Br J Pharmacol*. 147:S153–S162.

De Gobbi M, Viprakasit V, Hughes J, et al. (15 co-authors). 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312:1215–1217.

Drolet G, Dumont EC, Gosselin I, Kinkead R, Laforest S, Trottier JF. 2001. Role of endogenous opioid system in the regulation of the stress response. *Prog Neuropsychopharmacol Biol Psychiatry*. 25:729–741.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet*. 30:233–237.

Enattah NS, Jensen TG, Nielsen M, et al. (22 co-authors). 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet*. 82:57–72.

Fagergren P, Smith HR, Daunais JB, Nader MA, Porrino LF, Hurd YL. 2003. Temporal upregulation of *prodynorphin* mRNA in the primate striatum after cocaine self-administration. *Eur J Neurosci*. 17:2212–2218.

Fischer KD, Haese A, Nowock J. 1993. Cooperation of GATA-1 and Sp1 can result in synergistic transcriptional activation or interference. *J Biol Chem*. 268:23915–23923.

Geijer T, Jonsson E, Neiman J, Gyllander A, Sedvall G, Rydberg U, Terenius L. 1997. *Prodynorphin* allelic distribution in Scandinavian chronic alcoholism. *Clin Exp Res*. 21:1333–1336.

Grabe N. 2002. AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol*. 2:S1–S15.

Hacking D, Knight JC, Rockett K, Brown H, Frampton J, Kwiatkowski DP, Hull J, Udalova IA. 2004. Increased in vivo transcription of an IL-8 haplotype associated with respiratory syncytial virus disease-susceptibility. *Genes Immun*. 5:274–282.

Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet*. 66:1669–1679.

Hauser KF, Aldrich JV, Anderson KJ, et al. (13 co-authors). 2005. Pathobiology of dynorphins in trauma and disease. *Front Biosci*. 10:216–235.

Horan M, Millar DS, Hedderich J, Lewis G, Newsway V, Mo N, Fryklund L, Procter AM, Krawczak M, Cooper DN. 2003. Human growth hormone 1 (GH1) gene expression: complex haplotype dependent influence of polymorphic variation in the proximal promoter and locus control region. *Hum Mutat*. 21:408–423.

Hurd YL. 2002. Subjects with major depression or bipolar disorder show reduction of *prodynorphin* mRNA expression in discrete nuclei of the amygdaloid complex. *Mol Psychiatry*. 7:75–81.

Iglesias AR, Kindlund E, Tammi M, Wadelius C. 2004. Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene* 341:149–165.

Ji RR, Befort K, Brenner GJ, Woolf CJ. 2002. ERK MAP kinase activation in superficial spinal cord neurons induces *prodynorphin* and NK-1 upregulation and contributes to persistent inflammatory pain hypersensitivity. *J Neurosci*. 22:478–485.

Kaynard AH, McMurray CT, Douglass J, Curry TE Jr, Melner MH. 1992. Regulation of *prodynorphin* gene expression in the ovary: distal DNA regulatory elements confer gonadotropin regulation of promoter activity. *Mol Endocrinol*. 6:2244–2256.

Kieffer BL, Gaveriaux-Ruff C. 2002. Exploring the opioid system by gene knockout. *Prog Neurobiol*. 66:285–306.

Knight JR. 2004. Allele-specific gene expression uncovered. *Trends Genet*. 20:113–116.

Lemon B, Tjian R. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*. 14:2551–2569.

Lim MM, Wang Z, Olazábal DE, Ren X, Terwilliger EF, Young LJ. 2004. Enhanced partner preference in a promiscuous species

by manipulating the expression of a single gene. *Nature* 429: 754–757.
Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res*. 13:1855–1862.
Luca F, Kashyap S, Southard C, Zou M, Witonsky D, Di Rienzo A, Conzen SD. 2009. Adaptive variation regulates the expression of the human SGK1 gene in response to stress. *PLoS Genet*. 5:e1000489.
Marinova Z. 2006. Opioid and non-opioid activities of the dynorphins. The section of alcohol and drug dependence research. Stockholm, Sweden: Karolinska Institutet. p. 1–55.
Nikoshkov A, Hurd YL, Yakovleva T, Bazov I, Marinova Z, Cebers G, Pasikova N, Gharibyan A, Terenius l, Bakalkin G. 2005. *Prodynorphin* transcripts and proteins differentially expressed and regulated in the adult human brain. *Faseb J*. 19:1543.
Nikoshkov A, Drakenberg K, Wang X, Horvath MC, Keller E, Hurd YL. 2008. Opioid neuropeptide genotypes in relation to heroin abuse: dopamine tone contributes to reversed mesolimbic proenkephalin expression. *Proc Natl Acad Sci U S A*. 105:786–791.
Nomura A, Ujike H, Tanaka y, et al. (17 co-authors). 2006. Genetic variant of *prodynorphin* gene is risk factor for methamphetamine dependence. *Neurosci Lett*. 400:158–162.
Ober C, Loisel DA, Gilad Y. 2008. Sex-specific genetic architecture of human disease. *Nat Rev Genet*. 9:911–922.
Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*. 39:730–732.
Pastinen T, Hudson TJ. 2004. Cis-acting regulatory variation in the human genome. *Science* 306:647–650.
Peiper SC, Wang ZX, Neote K, et al. (11 co-authors). 1995. The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. *J Exp Med*. 181:1311–1317.
R Development Core Team. 2005. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Ray R, Doyle GA, Crowley JJ, Buono RJ, Oslin DW, Patkar AA, Mannelli P, DeMaria PAJ, O'Brien CP, Berrettini WH. 2005. A functional *prodynorphin* promoter polymorphism and opioid depenence. *Psychiatr Genet*. 15:295–298.
Reinius B, Saetre P, Leonard J, Blekhman R, Merino-Martinez R, Gilad Y, Jazin E. 2008. An evolutionarily conserved sexual signature in the primate brain. *PLoS Genet*. 4:e1000100.
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: 2208–2219.
Rockman MV, Wray GA. 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol*. 19:1991–2004.
Rothenburg S, Koch-Nolte F, Rich A, Haag F. 2001. A polymorphic dinucleotide repeat in the rat nucleoli gene forms Z-DNA and inhibits promoter activity. *Proc Natl Acad Sci U S A*. 98: 8985–8990.
Saetre P, Lindberg J, Leonard JA, Olsson K, Pettersson U, Ellegren H, Bergstrom TF, Vila C, Jazin E. 2004. From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res Mol Brain Res*. 126:198–206.
Schadt EE, Monks SA, Drake TA, et al. (14 co-authors). 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297–302.
Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*. 76:449–462.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 68:978–989.
Stogmann E, Zimprich A, Baumgartner C, Aull-Watschinger S, Hollt V, Zimprich F. 2002. A functional polymorphism in the *prodynorphin* gene promotor is associated with temporal lobe epilepsy. *Ann Neurol*. 51:260–263.
Stranger BE, Nica AC, Forrest MS, et al. (14 co-authors). 2007. Population genomics of human gene expression. *Nat Genet*. 39:1217–1224.
Sun BY, Tipton CM, Bidlack JA. 2006. The expression of *prodynorphin* gene is down-regulated by activation with lipopolysaccharide in U-937 macrophage cells. *J Neuroimmunol*. 174:52–62.
Sun G, Porter W, Safe S. 1998. Estrogen-induced retinoic acid receptor alpha 1 gene expression: role of estrogen receptor-Sp1 complex. *Mol Endocrinol*. 12:882–890.
Takahata N, Satta Y, Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130:925–938.
Tao H, Cox DR, Frazer KA. 2006. Allele-specific KRT1 expression is a complex trait. *PLoS Genet*. 2:848–858.
Telkov M, Geijer T, Terenius L. 1998. Human *prodynorphin* gene generates several tissue-specific transcripts. *Brain Res*. 804: 284–295.
The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 499:851–861.
Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouze P, De Moor B, Marchal K. 2002. INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif sampling. *Bioinformatics* 18:331–332.
Tishkoff SA, Reed FA, Ranciaro A, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 39:31–40.
Tompa M, Li N, Bailey TL, et al. (25 co-authors). 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*. 23:137–144.
Tournamille C, Colin Y, Cartron JP, Levankim C. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy negative individuals. *Nat Genet*. 10: 224–228.
Tung J, Primus A, Bouley A, Severson TF, Alberts SC, Wray GA. 2009. Evolution of a malaria resistance gene in wild primates. *Nature* 460:388–392.
Ventriglia M, Chiavetto LB, Bonvicini C, Tura GB, Bignotti S, Racagni G, Gennarelli M. 2002. Allelic variation in the human *prodynorphin* gene promoter and schizophrenia. *Neuropsychobiology*. 46:17–21.
Verrelli BC, McDonald JH, Argyropoulos G, Destrol-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. 2002. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am J Hum Genet*. 71:1112–1128.
Warner LR, Babbitt CC, Primus A, Severson TF, Haygood R, Wray GA. 2009. Functional consequences of genetic variation in primates on tyrosine hydroxylase (TH) expression in vitro. *Brain Res*. 1288:1–9.
Williams TM, Carroll SB. 2009. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nat Rev Genet*. 10:797–804.
Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavare S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* 322:434–438.
Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 430:85–88.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol*. 8:625–637.

Xuei X, Dick D, Flury-Wetherill L, et al. (16 co-authors). 2006. Association of the kappa-opioid system with alcohol dependence. *Mol Psychiatry*. 11:1016–1024.

Yan Y, Yuan W, Velculescu VE, Volgelstein B, Kinzler K. 2002. Allelic variation in human gene expression. *Science* 297.

Yuferov V, Ji F, Nielsen DA, Levran O, Ho A, Morgello S, Shi R, Ott J, Kreek MJ. 2009. A functional haplotype implicated in vulnerability to develop cocaine dependence is associated with reduced PDYN expression in human brain. *Neuropsychopharmacology* 34:1185–1197.

Zimprich A, Kraus J, Wöltje M, Mayer P, Rauch E, Höllt V. 2000. An allelic variation in the human *prodynorphin* gene promoter alters stimulus-induced expression. *J Neurochem*. 74: 472–477.