Journal of Human Genetics (2013) 58, 47–50

www.nature.com/jhg

SHORT COMMUNICATION

Use of alternative promoters may hide genetic effects on phenotypic traits

Edward A Ruiz-Narváez[1,2]

Genome-wide association studies have identified a multitude of single-nucleotide polymorphisms (SNPs) associated with a wide spectrum of human phenotypic traits. However, the SNPs identified so far do not explain much of the expected genetic variation and they are poor predictors of the occurrence of disease. I recently advanced the hypothesis that there is person-to-person variation in the use of alternative regulatory elements (for example, gene promoters) and this new source of variation may explain in part the low genetic variation accounted for known genetic variants. In the present report a simple mathematical model is developed to explore the biological consequences of the proposed hypothesis. The model predicts that in presence of person-to-person variation in the use of alternative promoters the observable effects of genetic variants located inside promoters will be smaller than their actual effects. As a consequence, genetic variation because of those observed polymorphisms will be reduced. The present report suggests new paths of research to elucidate the genetic basis of human complex traits.
Journal of Human Genetics (2013) **58,** 47–50; doi:10.1038/jhg.2012.115; published online 4 October 2012

Keywords: alternative promoters; complex diseases; genetic variation; GWAS; missing heritability

INTRODUCTION

A persistent observation in the current field of human genetics is that single-nucleotide polymorphisms (SNPs) identified through genome-wide association studies only explain a small fraction of the genetic variation of complex human traits;[1–3] the so-called missing heritability problem. Several non-mutually exclusive hypotheses such as rare alleles with large effects, and gene–gene and gene–environment interactions (reviewed by Manolio *et al.*,[4] and Gibson[5]) have been proposed to explain this lack of explained genetic variation, and they may all account in part for the missing heritability. However, most of these hypotheses assume that the genome is read in the same way across all individuals and therefore, the same SNP has exactly the same functionality from a person to another person. This assumption may be not true for the use alternative regulatory elements (for example, gene promoters).[6] For example, in a gene with multiple promoters some persons would use preferentially a particular promoter, and other subjects would tend to use another promoter of the same gene. Under this scenario, a particular SNP would have functional significance only among those individuals who use the promoter inside which the SNP is located. The observed effect of that SNP would be attenuated relative to its actual effect on a phenotypic trait of interest.

In the present report, I explore some of the quantitative consequences of the hypothesis of alternative use of regulatory elements.

MATERIALS AND METHODS

The present model is based on the recently proposed hypothesis on the existence of inter-individual variation in the use of alternative promoters.[6] Let us assume a gene X that controls a continuous phenotypic trait Y. Expression of the gene X is under the control of M alternative promoters, and different SNPs may be present inside each promoter (Figure 1). The model assumes the existence of person-to-person variation in the use of the alternative promoters may be through the action of epigenetic marks (for example, DNA methylation). Although in the present work the model is restricted to a gene with only two promoters, P1 and P2, and two SNPs: G1 (inside P1) and G2 (inside P2), the results can be easily generalized to a gene with more than two promoters and more than one SNP inside each promoter. The SNP G1 has two alleles, A1 with frequency equal to p_1 and A2 with frequency equal to p_2. The allele A1 increases by a units the value of the phenotypic trait Y compared with the allele A2. The SNP G2 has two alleles, B1 with frequency equal to q_1 and B2 with frequency equal to q_2. The allele B1 increases by b units the value of the phenotypic trait Y compared with the allele B2. Each allele will affect the phenotypic trait Y only when its corresponding promoter is being used (that is, the allele A1 increases the value of Y only when the promoter P1 is used, and the allele B1 increases the value of Y only when the promoter P2 is used). The promoter P1 is used in a proportion f_1 of the chromosomes in the population, and the promoter P2 is used in a proportion f_2 of the chromosomes in the population. Chromosomes that use the promoter P1 have an increase of e units of the phenotypic trait Y compared with chromosomes that use the promoter P2. Hardy–Weinberg equilibrium is assumed for each SNP. As the goal of the present analysis is to show how genetic variability may be

[1]Slone Epidemiology Center at Boston University, Boston, MA, USA and [2]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
Correspondence: Dr EA Ruiz-Narváez, Slone Epidemiology Center at Boston University, 1010 Commonwealth Avenue, Boston, MA 02215, USA.
E-mail: eruiznar@bu.edu
Received 25 March 2012; revised 2 September 2012; accepted 6 September 2012; published online 4 October 2012

hidden because of the use of alternative promoters, in the following results it is assumed that we do not observe which one of the alternative promoters is being used. We only observe the genotypes in G1 and G2 as well as the individuals' values of the phenotypic trait *Y*.

Table 1 shows both the unobserved types of chromosomes according to the unobserved promoter use and observed genotypes in the G1 and G2 SNPs (upper half of the table, chromosome types H1 through H8). Note that additive effects are measured relative to the chromosome H8 (P2A2B2) that by definition has a value equal to zero for the phenotypic trait *Y*. Observed chromosomes based on only the genotypes in the G1 and G2 SNPs are shown in the bottom half of Table 1 (chromosome types J1 through J4). Chromosome frequencies are shown for four different models of linkage disequilibrium (LD) between the G1 and G2 SNPs. The most general scenario (model 1) makes no assumption about any particular value of LD (given by the D coefficient or covariance between the G1 and G2 SNPs). Models 2, 3 and 4 are particular cases of model 1. Model 2 assumes linkage equilibrium between the G1 and G2 SNPs. Scenarios portrayed by models 3 and 4 refer to complete LD between the G1 and G2 SNPs. In model 3, the A1 and B1 (and A2 and B2) alleles are always present together (that is, only unobserved chromosomes P1A1B1, P1A2B2, P2A1B1 and P2A2B2 exist in the population). An opposite pattern of complete LD is shown in model 4, where the A1 and B2 (and A2 and B1) alleles are always transmitted together (that is, only unobserved chromosomes P1A1B2, P1A2B1, P2A1B2 and P2A2B1 are present in the population). It must be noticed that observed chromosomes are obtained from the unobserved chromosomes after collapsing over promoters P1 and P2. For example, the observed J1 chromosome (A1B1) is a mixture of the unobserved H1 (P1A1B1) and H5 (P2A1B1) chromosomes. Phenotypic value of the J1 chromosome is the average of the phenotypic values of the H1 and H5 chromosomes weighted by the f_1 and f_2 proportions, respectively. Frequency of the J1 chromosome is just the sum of the frequencies of the H1 and H5 chromosomes. The rest of the observed chromosomes can be obtained in a similar way: J2 = H2 + H6, J3 = H3 + H7 and J4 = H4 + H8.

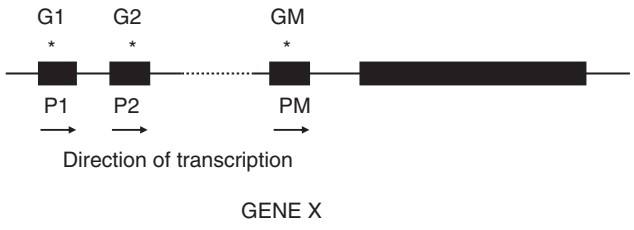

Figure 1 Gene with alternative promoters. A gene *X* is transcribed from *M* alternative promoters, P1, P2, …, PM. It is proposed the existence of person-to-person variation in which of the promoters is used. Each promoter contains different single-nucleotide polymorphisms (SNPs). A polymorphism G1 is located inside the P1 promoter, and a different polymorphism G2 is located inside the P2 promoter.

Summary statistics

The mean chromosome value of the phenotypic trait *Y* is equal to

$$E(Y_C) = \sum_i \text{phenotype}(i) \times \text{frequency}(i) = f_1(\text{ap}_1 + e) + f_2 \text{bq}_1 \quad (1)$$

where phenotype*(i)* and frequency*(i)* are the phenotype value and frequency of the i-th chromosome (either observed or unobserved).

Variance of the chromosome phenotype values would be equal to

$$V(Y_C) = \sum_i \text{frequency}(i) \times (\text{phenotype}(i) - E(Y_C))^2 \quad (2)$$

It is noteworthy that the mean of the chromosome phenotype values does not depend on LD and is the same for the actual (unobserved) and observed chromosomes. However, as it will be shown below, the actual variance due to unobserved chromosomes (that is, total variance) will be always greater or equal than the variance due to observed chromosomes. In other words the variance due to measurable genetic variation (that is, G1 and G2 SNPs) will fail to explain 100% of the actual variance due to the totality of unobserved chromosomes in the population.

RESULTS

Let us define K^2 as the ratio of the variance due to observed chromosomes to the total variance due to unobserved chromosomes. Figure 2 shows K^2 under three different particular scenarios: (1) no LD between the G1 and G2 SNPs, (2) positive LD between the G1 and G2 SNPs (that is, A1 and B1 alleles tend to be transmitted together) and (3) negative LD between the G1 and G2 SNPs (that is, A1 and B2 alleles tend to be transmitted together).

Table 1 List of actual and observed types of chromosomes with their respective phenotypic values and frequencies under linkage disequilibrium (LD) patterns

		Frequency under four different models of linkage disequilibrium (LD)[a]			
Type	*Phenotypic value*	*General (1)*	*No LD (2)* ($r^2=0$, $D'=0$)	*Complete positive LD (3)* ($r^2=1$, $D'=1$, $p_1=q_1$)	*Complete negative LD (4)* ($r^2=1$, $D'=-1$, $p_1=q_2$)
Actual (unobserved chromosomes)					
H1 = P1A1B1	$y_1 = a + e$	$z_1 = f_1(p_1q_1 + D)$	$z_1 = f_1p_1q_1$	$z_1 = f_1p_1$	$z_1 = 0$
H2 = P1A1B2	$y_2 = a + e$	$z_2 = f_1(p_1q_2 - D)$	$z_2 = f_1p_1q_2$	$z_2 = 0$	$z_2 = f_1p_1$
H3 = P1A2B1	$y_3 = e$	$z_3 = f_1(p_2q_1 - D)$	$z_3 = f_1p_2q_1$	$z_3 = 0$	$z_3 = f_1p_2$
H4 = P1A2B2	$y_4 = e$	$z_4 = f_1(p_2q_2 + D)$	$z_4 = f_1p_2q_2$	$z_4 = f_1p_2$	$z_4 = 0$
H5 = P2A1B1	$y_5 = b$	$z_5 = f_2(p_1q_1 + D)$	$z_5 = f_2p_1q_1$	$z_5 = f_2p_1$	$z_5 = 0$
H6 = P2A1B2	$y_6 = 0$	$z_6 = f_2(p_1q_2 - D)$	$z_6 = f_2p_1q_2$	$z_6 = 0$	$z_6 = f_2p_1$
H7 = P2A2B1	$y_7 = b$	$z_7 = f_2(p_2q_1 - D)$	$z_7 = f_2p_2q_1$	$z_7 = 0$	$z_7 = f_2p_2$
H8 = P2A2B2	$y_8 = 0$	$z_8 = f_2(p_2q_2 + D)$	$z_8 = f_2p_2q_2$	$z_8 = f_2p_2$	$z_8 = 0$
Observed chromosomes					
J1 = A1B1	$v_1 = f_1(a + e) + f_2b$	$w_1 = p_1q_1 + D$	$w_1 = p_1q_1$	$w_1 = p_1$	$w_1 = 0$
J2 = A1B2	$v_2 = f_1(a + e)$	$w_2 = p_1q_2 - D$	$w_2 = p_1q_2$	$w_2 = 0$	$w_2 = p_1$
J3 = A2B1	$v_3 = f_1e + f_2b$	$w_3 = p_2q_1 - D$	$w_3 = p_2q_1$	$w_3 = 0$	$w_3 = p_2$
J4 = A2B2	$v_4 = f_1e$	$w_4 = p_2q_2 + D$	$w_4 = p_2q_2$	$w_4 = p_2$	$w_4 = 0$

[a]D is the linkage disequilibrium coefficient that measures the deviation of the observed frequency of a haplotype from its expected frequency under linkage equilibrium.
D′ is defined as the LD coefficient normalized to the maximum D that is possible given the observed allele frequencies: $D' = D/D_{max}$ r^2 is the squared correlation coefficient between the G1 and G2 SNPs: $r^2 = D^2/(p_1p_2q_1q_2)$.

SCENARIO 1

Figure 2a shows K^2 under linkage equilibrium between the G1 and G2 SNPs (model 2 in Table 1) for different epigenetic effects and proportion of chromosomes using the promoter P1. The additive effects of the A1 and B1 alleles were assumed to be equal to 5 units of the continuous phenotypic trait ($a=b=5$ units). The epigenetic effect was allowed to take four different values: $e=0$, 5, 10 and 20 units of the phenotypic trait. It is clear that $K^2 \leqslant 1$, and the higher the epigenetic effect the lower the K^2 ratio (that is, the observed chromosomes explain less of the total variance due to the actual unobserved chromosomes). Even in absence of any epigenetic effect (that is, $e=0$ meaning that the P1 and P2 promoters have the same baseline level of the phenotypic trait Y) the observed chromosomes do not explain the totality of the variance due to unobserved chromosomes. The only instances when $K^2=1$ are when only one promoter is used in the population (that is, $f_1=1$, use of promoter P1 is fixed; or $f_1=0$, use of promoter P2 is fixed).

SCENARIO 2

Figure 2b shows K^2 under positive LD between the G1 and G2 SNPs (that is, the A1 and B1 alleles tend to be transmitted together in the same chromosome) for different r^2 values (squared correlation between the G1 and G2 SNPs) and proportion of chromosomes using the promoter P1. The additive effects of the A1 and B1 alleles as well as the epigenetic effect were kept constant and equal to 5 units of the phenotypic trait ($a=b=e=5$ units). For this scenario $K^2 \leqslant 1$ too, and it is noteworthy that the stronger the LD between both SNPs (that is, the higher r^2) the more the observed chromosomes would explain the total variance. When $r^2=1.0$ (complete positive LD as shown in model 3 of Table 1) reduction of K^2 is attenuated in comparison with the case of linkage equilibrium ($r^2=0.0$).

SCENARIO 3

Figure 2c shows K^2 under negative LD between the G1 and G2 SNPs (that is, the A1 and B2 alleles tend to be transmitted together in the same chromosome) for different r^2 values and proportion of chromosomes using the promoter P1. The additive effects of the A1 and B1 alleles as well as the epigenetic effect were kept constant and equal to 5 units of the phenotypic trait ($a=b=e=5$ units). Similar to the previous two scenarios we have that $K^2 \leqslant 1$ however, in presence of negative LD the higher r^2 the lower the variance that is explained by the observed chromosomes. Maximum reduction of that K^2 is observed when $r^2=1.0$ (complete negative LD as shown in model 4 of Table 1). Only the haplotypes A1B2 and A2B1 are observed in the presence of complete negative LD, and as Figure 2c shows the variance due to the observed chromosomes can completely disappear ($K^2=0$). A simple calculation shows that K^2 vanishes when the proportion of chromosomes using the P1 promoter is equal to $f_1=b/(a+b)$. K^2 will disappear at $f_1=0.5$ when both A1 and B1 alleles have the same additive effect ($a=b$); at $f_1<0.5$ when the A1 allele has a higher additive effect than allele B1 ($a>b$); and at $f_1>0.5$ when the A1 allele has a lower additive effect that allele B1 ($a<b$).

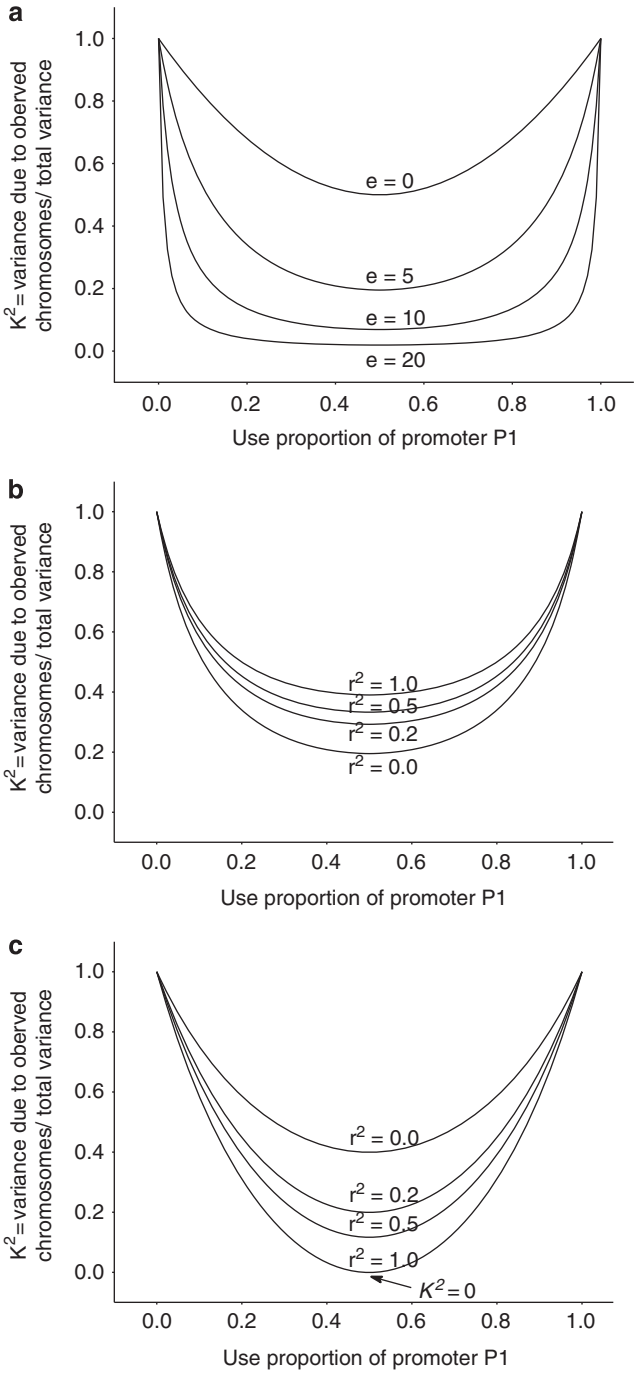

Figure 2 Proportion of total variance that is explained by observed chromosomes. In presence of person-to-person variation in the use of alternative promoters the variance due to observed chromosomes is always lower that the total variance of the genetic system ($K^2<1$). Only when use of one promoter is fixed in the population ($f_1=0$ or $f_1=1$) the observed chromosomes would explain 100% of the total genetic variance ($K^2=1$). K^2 variation is under three possible scenarios of linkage disequilibrium (LD) between the G1 and G2 SNPs: (**a**) linkage equilibrium, (**b**) positive LD, and (**c**) negative LD. The additive effects of the A1 and B1 alleles were assumed to be equal to 5 units of the phenotypic trait. The epigenetic effect was allowed to take four different values in (**a**) (e = 0, 5, 10 and 20 units), and kept constant in (**b**, **c**) (e = 5 units).

DISCUSSION

The current model offers a potential mechanism to explain in part why genetic variants discovered so far do not explain much of the expected genetic variability. Although part of the unexplained variability may be due to rare genetic polymorphisms still to be found,[4] the model predicts that person-to-person variation in the use of alternative promoters would reduce the observed genetic variance of a genetic system. Thus, even a complete knowledge of all the genetic variants involved in a particular phenotypic trait would be no enough to explain the whole genetic variance of the trait.

Three major factors explain the reduction of the genetic variance according to the model discussed in the present work. First, the observed additive effects of the SNPs inside each of the alternative promoters are attenuated in comparison with their actual effects. For example, because the allele A1 of the G1 SNP exerts its effect only when the promoter P1 is being used, its observed additive effect would be reduced by a factor equal to f_1 relative to its actual effect. The same situation applies for the B1 allele of the G2 SNP whose observed additive effect would be attenuated by a factor equal to f_2. Second, because the use of alternative promoters is not being measured (for example, in current genetic epidemiology studies such scenario is not even considered as a possibility) the dimensionality of the observed data would be always lower than the actual dimensionality of the population data. The number of observed chromosomes will be less than the number of actual chromosomes in the population. Third, different promoters may have different baseline levels of the phenotypic trait under study further reducing the proportion of the actual variance that is due to measured genetic polymorphisms.

Recent published evidence supports the proposed hypothesis of person-to-person variation in the use of alternative promoters. Turner *et al.*[7] reported the presence of high inter-individual variability in the methylation patterns of alternative promoters of the glucocorticoid receptor (*NR3C1*) gene in 26 healthy subjects, suggesting person-to-person variation in epigenetic regulatory mechanisms. A small study that measured promoter activity of the aromatase (*CYP19A1*) gene in skin fibroblasts from four normal volunteers found that one subject showed increased activity of the promoters I.3 and II in response to cyclic adenosine monophosphate, in contrast to the other three subjects who expressed the cyclic adenosine monophosphate-unresponsive promoter I.4.[8] In non-malignant lung tissue from 15 patients with non-small-cell lung cancer, two cases used mostly promoters I.3 and II of the *CYP19A1* gene and the rest of patients used the promoter I.4.[9] It is noteworthy that may even exist ethnic differences in the use of alternative promoters. A recent study in 101 women with uterine leiomyoma (31 African American, 34 white American and 36 Japanese women) reported that leiomyoma tissue from African American women expressed the promoter II in higher proportion compared with Japanese women.[10] At last, the *CD36* gene showed inter-individual variability in the use of four out of five alternative promoters in cultured monocytes from 10 subjects.[11]

The present results, published evidence about variability in the use of alternative promoters, and the fact that more than half of human genes have alternative promoters,[12] with a mean of 3.1 promoters per gene[13] stress the need to carry out extensive studies in human populations to determine and quantify inter-individual variation in the use of alternative promoters. To date there are few approaches to assess the use of alternative promoters in a genome-wide scale. Singer *et al.*[14] developed a promoter tiling array that can identify about 35 000 alternative promoters from almost 7000 human genes, and Jacox *et al.*[15] described a computational approach to determine alternative promoter usage in nearly 1500 genes using the Affymetrix Exon 1.0 array (Affymetrix, Santa Clara, CA, USA). Although those microarrays only interrogate a subset of genes in the genome (that is, those genes with known alternative promoters) they would provide enough data to test the proposed hypothesis in a genome-wide scale. A comprehensive assessment should ideally measure person-to-person variation across different types of tissue.

The present model can be easily extended to include cases of genes with more than two promoters and more than one SNP in each of the promoters. In a gene with multiple promoters, the observed additive effect of a particular SNP would be reduced by a factor equal to the proportion of chromosomes in the population using the promoter in which the SNP is located. The model may also be used for other types of alternative regulatory elements such as multiple enhancers affecting gene expression; the so-called shadow enhancers.[16–18] A limitation of the presented model is that depends on the knowledge about alternative promoters or regulatory elements in general. More experimental work such as chromatin immunoprecipitation-chip assays validated with transgenic models is needed to identify new regulatory elements.

In summary, the present report shows that in presence of inter-individual variation in the use of alternative promoters the observable effects of genetic variants will be lower than their actual effects. The proposed model may explain in part why genome-wide association studies-identified variants are in most part poor predictors of human complex traits. Future studies are needed to determine and quantify the person-to-person variability in the use of alternative promoters as well as to identify new regulatory elements in the human genome.

ACKNOWLEDGEMENTS

This work was supported by grant 11SDG7390014 of the American Heart Association, and by grants R01CA058420 and R01CA098663 from the National Cancer Institute, Division of Cancer Control and Population Science (http://www.cancercontrol.cancer.gov). The content is solely the responsibility of the author and does not necessarily represent the official views of the American Heart Association, the National Cancer Institute or the National Institutes of Health.

1 Hofker, M. & Wijmenga, C. A supersized list of obesity genes. *Nat. Genet.* **41,** 139–140 (2009).
2 Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40,** 575–583 (2008).
3 Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P. *et al.* Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40,** 609–615 (2008).
4 Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
5 Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
6 Ruiz-Narvaez, E. A. What is a functional locus? Understanding the genetic basis of complex phenotypic traits. *Med. Hypotheses* **76,** 638–642 (2011).
7 Turner, J. D., Pelascini, L. P., Macedo, J. A. & Muller, C. P. Highly individual methylation patterns of alternative glucocorticoid receptor promoters suggest individualized epigenetic regulatory mechanisms. *Nucleic Acids Res.* **36,** 7207–7218 (2008).
8 Demura, M. & Bulun, S. E. CpG dinucleotide methylation of the CYP19 I.3/II promoter modulates cAMP-stimulated aromatase activity. *Mol. Cell Endocrinol.* **283,** 127–132 (2008).
9 Demura, M., Demura, Y., Ameshima, S., Ishizaki, T., Sasaki, M., Miyamori, I. *et al.* Changes in aromatase (CYP19) gene promoter usage in non-small cell lung cancer. *Lung Cancer* **73,** 289–293 (2011).
10 Ishikawa, H., Reierstad, S., Demura, M., Rademaker, A. W., Kasai, T., Inoue, M. *et al.* High aromatase expression in uterine leiomyoma tissues of African-American women. *J. Clin. Endocrinol. Metab.* **94,** 1752–1756 (2009).
11 Andersen, M., Lenhard, B., Whatling, C., Eriksson, P. & Odeberg, J. Alternative promoter usage of the membrane glycoprotein CD36. *BMC Mol. Biol.* **7,** 8 (2006).
12 Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A. *et al.* Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes. *Genome Res.* **17,** 1005–1014 (2007).
13 Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16,** 55–65 (2006).
14 Singer, G. A., Wu, J., Yan, P., Plass, C., Huang, T. H. & Davuluri, R. V. Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics* **9,** 349 (2008).
15 Jacox, E., Gotea, V., Ovcharenko, I. & Elnitski, L. Tissue-specific and ubiquitous expression patterns from alternative promoters of human genes. *PLoS One* **5,** e12274 (2010).
16 Hong, J. W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* **321,** 1314 (2008).
17 Wray, G. A. & Babbitt, C. C. Genetics. Enhancing gene regulation. *Science* **321,** 1300–1301 (2008).
18 Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F. & Stern, D. L. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466,** 490–493 (2010).