CALL FOR PAPERS | Functional Analysis of Sequence Variation

Identifying functional single nucleotide polymorphisms in the human CArGome

Craig C. Benson,1 Qian Zhou,2 Xiaochun Long,2 and Joseph M. Miano2

1Combined Internal Medicine-Pediatrics Residency Program, University of Rochester Medical Center; and 2Aab Cardiovascular Research Institute, University of Rochester School of Medicine and Dentistry, Rochester, New York

Submitted 21 June 2011; accepted in final form 14 July 2011

Benson CC, Zhou Q, Long X, Miano JM. Identifying functional single nucleotide polymorphisms in the human CArGome. Physiol Genomics 43: 1038–1048, 2011. First published July 19, 2011; doi:10.1152/physiolgenomics.00098.2011.—Regulatory SNPs (rSNPs) reside primarily within the nonprotein coding genome and are thought to disturb normal patterns of gene expression by altering DNA binding of transcription factors. Nevertheless, despite the explosive rise in SNP association studies, there is little information as to the function of rSNPs in human disease. Serum response factor (SRF) is a widely expressed DNA-binding transcription factor that has variable affinity to at least 1,216 permutations of a 10 bp transcription factor binding site (TFBS) known as the CArG box. We developed a robust in silico bioinformatics screening method to evaluate sequences around RefSeq genes for conserved CArG boxes. Utilizing a predetermined phastCons threshold score, we identified 8,252 strand-specific CArGs within an 8 kb window around the transcription start site of 5,213 genes, including all previously defined SRF target genes. We then interrogated this CArG dataset for the presence of previously annotated common polymorphisms. We found a total of 118 unique CArG boxes harboring a SNP within the 10 bp CArG sequence and 1,130 CArG boxes with SNPs located just outside the CArG element. Gel shift and luciferase reporter assays validated SRF binding and functional activity of several new CArG boxes. Importantly, SNPs within or just outside the CArG box often resulted in altered SRF binding and activity. Collectively, these findings demonstrate a powerful approach to computationally define rSNPs in the human CArGome and provide a foundation for similar analyses of other TFBS. Such information may find utility in genetic association studies of human disease where little insight is known regarding the functionality of rSNPs.

transcription factor binding site; bioinformatics; serum response factor; CArG box

SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) are the most common variant in the human genome. While the function of protein-coding SNPs can be easily deciphered and tested, it is far more difficult to ascertain SNP function in the vast sequence landscape comprising the nonprotein coding genome. Even as the number of SNP association studies continues to soar, there has been less effort to elucidate the function of nonprotein coding SNPs associated with human disease though it is recognized that such variants likely play a significant role in the development of complex disease traits (19, 28). Recent estimates suggest that ~88% of the SNPs identified as associated with a trait or disease are located in nonprotein coding regions (27). Regulatory SNPs (rSNPs) are defined as variant sequences located in or near transcription factor binding sites (TFBS) that affect gene expression by altering DNA binding properties of transcription factors (39). rSNPs may also have other effects such as altering the specificity of microRNAs and their binding sites or disrupting recombination, replication, or structural organization of the genome (47, 76). Despite the rise in rSNPs associated with human disease derived from genome-wide association studies (GWAS), we know very little about how each of these variants function on a molecular or cellular level (31). However, variations in the binding sites of RNA polymerase and nuclear factor-kB, frequently due to SNPs, were found to differ between humans and correlated with altered gene expression (36). Additionally, the recent example of a noncoding SNP that created a C/EBP TFBS correlating with increases in SORTI gene expression and changes in serum lipoprotein levels strongly supports the hypothesis that rSNPs contribute to disease (55). Thus, elucidating the functional significance of such disease-associated rSNPs has emerged as a critical research aim (5, 67).

Given the extensive time and cost of wet-lab rSNP evaluation, numerous in silico methods of whole genome screening have been developed to predict TFBSs that incorporate multispecies alignments and phylogenetic foot-printing (24) and prioritize high-probability candidate rSNPs (63). Examples include the RAVEN system to explore in silico TFBS variation based upon position weight matrices (PWM) (2) and a novel Bayesian-based method to construct phylogenetic trees from 18 mammalian genomes, from which >28,000 TFBS that contain potential rSNPs were discovered (86). Other more recent PWM-based methods attempted to predict alterations in protein-DNA binding affinities induced by sequence variations and potential rSNPs (44, 45). While promising from a computational perspective, these approaches are limited primarily from a high false discovery rate of TFBS, thereby decreasing the likelihood of elucidating functionally relevant rSNPs. To address these challenges, translational methods that combine biological data with in silico methods are increasingly being used. Notably, Torkamani and Schork (78) utilized the functional genomics dataset from the ENCODE Pilot Project to develop a predictive model of rSNPs that achieved 80% sensitivity and 99% specificity. Unfortunately, this approach was limited since it used only 1% of the human genome for functional significance (17). Other translational bioinformatics approaches using chromatin immunoprecipitation (ChIP)-on-chip data have been successful at identifying rSNPs but require...
a priori knowledge of DNA-protein binding patterns (1). Computational models based on high-throughput sequencing methods of ChIP-seq have advanced the discovery of biologically active TFBS but are contextually biased by the cell type/condition and are limited to only those transcription factors with specific antibodies (3, 40, 84). Thus, while these biological assays have advanced our understanding of the genome, there remains a need for less biased genome-wide discovery of potentially functional rSNPs identified by in silico techniques.

Development of a robust genome-wide in silico approach to the identification of TFBS and potential rSNPs based on biological data requires a well-characterized TF. For example, the TFBS for CREB is well defined, and a recent genome-wide survey identified >750,000 CREB sites in the genome (89). Another TF with a well-defined TFBS is serum response factor (SRF), which controls disparate programs of gene expression related to growth and muscle differentiation (32, 51). SRF binds a TFBS known as the CArG box, a 10 bp sequence generally found within 4,000 bp of a targeted gene’s transcription start site (TSS) (75). Previous studies have used various screening methods to identify SRF-target genes with CArG boxes that are conserved in sequence and space between human and mouse genomes (60, 69, 75, 88). The aggregate total number of CArG elements within the genome, known as the CArGome (75), control hundreds of target genes (53). Thus, SRF’s extensive library of validated and hypothetical CArG targets and critical biological importance (50) render it an ideal TF to study in silico rSNP identification methods. Additionally, identifying potential rSNPs in or around CArG boxes may expand our understanding of SRF target gene function, specifically in the context of disease.

To date, there has been no systematic attempt to identify rSNPs across the human CArGome. Binding interactions between SRF and CArG elements have been well characterized (42). Previous studies have shown that mutations in the 10 bp CArG sequence are known to alter the binding properties of SRF (72). Additionally, Han et al. (23) discovered a 12 bp insertion mutation located ~20 bp from a nearby CArG box that resulted in elevated SRF binding and increased gene expression of smooth muscle myosin light chain kinase in hypertensive rats. These results support the notion that genetic mutations in or near the CArG box alter gene expression and contribute to disease. Given our understanding of the CArG sequence and the increasing number of SRF targets in the genome, the CArG box is an ideal TFBS to use as a model for developing screening tools for functional rSNP identification. Herein, we describe an in silico screening approach to rapidly pinpoint functional rSNPs in the human CArGome. Validation studies suggest that most rSNPs in the CArGome have functional consequences for SRF binding and activity, thus providing a rich resource for future genetic association studies as well as novel insights into SRF function.

MATERIALS AND METHODS

Computer interrogation of CArGome for rSNP discovery. Perl scripts (available on request) were created to implement the process of identifying SNPs in the CArGome as depicted in Fig. 1. The Ensembl database (release 59) (18) was queried for all protein-coding Human Genome Organization (HUGO) genes in the GRCh37/hg19 genome assembly from chromosomes 1 to 22, X, and Y. For each HUGO gene, the longest RefSeq transcript was used as the canonical transcript (Supplemental Table S1).1 For each canonical transcript, the promoter sequence comprising 4,000 bp upstream and 4,000 bp downstream from the TSS was extracted. Each promoter sequence was scanned for CArG elements using a Perl script that matched the pattern CCwGG (consensus CArG box), allowing for 1 bp deviation (CArG-like box) as reviewed previously (53).

Using a set of 34 previously validated CArG boxes from mouse, we generated a conservation threshold score by first extracting the CArG sequences from the National Center for Biotechnology Information 37/mm9 mouse assembly from the Ensembl database (Supplemental Table S2). We initially evaluated CArG sequences for conservation with the GERP score (33). However, we found that a majority of the CArG sequences had no GERP score available. In contrast, all 34 CArG sequences used to develop the conservation threshold exhibited a phastCons score (11). Using the 46-way vertebrate phastCons dataset from the UCSC Genome Browser (71), we calculated the phastCons score for each CArG box by averaging the 10 individual nucleotide phastCons scores in the given CArG box. The cumulative average phastCons score for these 34 individual mouse CArG boxes was calculated to be 0.842 and subsequently used as the conservation cutoff score for the human RefSeq CArG scan (RS-CArGome).

For each CArG sequence found in the human scan of HUGO genes, a 10 bp average phastCons score was calculated and compared with the conservation cutoff described above. Those human CArG boxes meeting the cutoff score were classified as conserved and tabulated with respect to its nearest RefSeq neighbor. A set of 207 previously validated mammalian SRF-target genes was also cross-referenced with our tabular CArG list (Supplemental Table S3).

For each conserved CArG element identified, its absolute chromosomal coordinates in the human genome were used to query dbSNP 131 stored within the Ensembl Variation database to identify genotyped SNPs residing within the 10 bp conserved CArG or within an arbitrarily set 35 bp flanking sequence of the conserved CArG box (70). This distance surrounding the CArG box was selected as many CArG elements function in conjunction with neighboring cis-acting elements, such as ETS binding sites that may be located several helical turns of DNA away from the CArG box (4). Finally, the promoter sequences for each human gene that had conserved CArG boxes and locally identified SNPs were mapped on the human genome, along with the conserved CArG and SNPs, and visualized using the UCSC Genome Browser (38). A Perl script was written to calculate CArG distances relative to the CArG boxes along with their distribution among conserved regions. CArG nucleotide sequence analysis was conducted with WebLogo (13). Evaluation of possible CArG-related gene clusters was performed based on CArG box genomic coordinates to identify CArG boxes located in the vicinity of multiple genes.

Functional analysis of CArG boxes and SRF target genes. Genomic Regions Enrichment of Annotations Tool (GREAT) analysis was conducted to evaluate CArG box enrichment based on the biological process Gene Ontology (GO) term (49). All conserved CArG sequences identified in the hg19 assembly were utilized for GREAT analysis using version 1.6. GREAT analysis was performed against a whole genome background and using the “basal plus extension” association rule, in which a proximal gene association region is within 5 kb upstream and 1 kb downstream from the TSS and the distal gene association region is up to 1,000 kb.

Analysis for overrepresented TFBS in our set of 5,213 suspected SRF-target genes was conducted using the oPOSSUM’s Human Single Site analysis, version 2.0 (29). HUGO gene names were submitted to the oPOSSUM web server (http://www.cisreg.ca/cgi-bin/oPOSSUM/oopossum) for analysis with the default analysis settings, including top 10% of conserved regions, TFBS matrix match score of 80%, and a sequence length of 5 kb up and downstream. A total

---

1 The online version of this article contains supplemental material.
of 1,370 genes were excluded from analysis by oPOSSUM due to lack of strict one-to-one ortholog gene assignment between mouse and human.

**Electrophoretic mobility shift assay.** Eight of the conserved CArG sequences with SNPs were selected for validation with electrophoretic mobility shift assay (EMSA) and named according to their nearest proximity to a RefSeq gene (Table 1). We chose these eight CArG boxes to study based on their novelty as potential SRF targets and/or their consensus CArG sequence, rendering them most susceptible to SRF binding. Seven of these potential SRF target genes had an SNP located within the CArG box (ADRB2, KCNA3, CDH3, RALYL, PLAGL2, KLF6, GRK6), and for one of the target genes the SNP was located outside the CArG box.

**Table 1. Functionally validated SNPs in the human RS-CArGome**

<table>
<thead>
<tr>
<th>HUGO Gene</th>
<th>CArG Sequence</th>
<th>Position to TSS</th>
<th>Coordinates</th>
<th>Strand</th>
<th>phastCons</th>
<th>rsSNP</th>
<th>Position in CArG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADRB2</td>
<td>CCA(A/G)TTTGG</td>
<td>−2458</td>
<td>chr5:148203698-148203707</td>
<td>−1</td>
<td>0.941732</td>
<td>rs35118767</td>
<td>4</td>
</tr>
<tr>
<td>KCNA3</td>
<td>CCATTTTGG(G/A)</td>
<td>−2423</td>
<td>chr1:111220069-111220078</td>
<td>1</td>
<td>0.99685</td>
<td>rs61805969</td>
<td>10</td>
</tr>
<tr>
<td>RALYL</td>
<td>CCATAAAAG(G/T)</td>
<td>2056</td>
<td>chr8:18620218-18620227</td>
<td>1</td>
<td>0.929134</td>
<td>rs62530172</td>
<td>10</td>
</tr>
<tr>
<td>KLF6</td>
<td>(C/A)CTTATTGG</td>
<td>2124</td>
<td>chr10:3825334-3825343</td>
<td>−1</td>
<td>0.994488</td>
<td>rs10795076</td>
<td>1</td>
</tr>
<tr>
<td>ABHD5</td>
<td>(C/A)TTTGGG(A/G)</td>
<td>−2405</td>
<td>chr3:43729957-43729966</td>
<td>1</td>
<td>0.948034</td>
<td>rs34104729</td>
<td>12</td>
</tr>
<tr>
<td>GRK6</td>
<td>CGCTAA(T/C)TTG</td>
<td>−282</td>
<td>chr5:176853405-176853414</td>
<td>1</td>
<td>0.998425</td>
<td>rs3763079</td>
<td>7</td>
</tr>
<tr>
<td>CDH3</td>
<td>CCTTTAG(A/G)GG</td>
<td>2637</td>
<td>chr16:68681376-68681385</td>
<td>1</td>
<td>0.992913</td>
<td>rs16958232</td>
<td>8</td>
</tr>
<tr>
<td>PLAGL2</td>
<td>CGAG(A/G)TAAGG</td>
<td>464</td>
<td>chr20:37095121-37095130</td>
<td>−1</td>
<td>0.999213</td>
<td>rs13037347</td>
<td>5</td>
</tr>
</tbody>
</table>

Each row represents a HUGO gene with a single nucleotide polymorphism (SNP) (dbSNP 131) mapped within or near the gene’s conserved CArG box. HUGO Gene names in boldface contain true CArG sequences, while lightfaced gene names harbor CArG-like sequences with 1 bp substitution (see Discussion for differences between consensus CArG and CArG-like sequences). Wild-type and SNP alleles are shown in parentheses for SNPs within or near the CArG box; the first nucleotide in each parenthesis represents the wild-type allele and the second nucleotide designates the SNP allele. Position to transcription start site (TSS) is the number of bp from the start of the CArG box to the TSS of the possible serum response factor-target gene (negative is upstream; positive is downstream). Conserved CArG Coordinates and Strand are based on Ensembl (Release 59), using the H. S. Feb 2009 (HG 19) Assembly. phastCons score is an average of the 10 individual phastCons scores for each nucleotide in the CArG box (see Methods). The reference SNP (rs) number for each CArG-SNP was obtained from dbSNP 131. Position in CArG is the location of the SNP in relation to position 1 of the CArG box. *SNP that lies outside the CArG box.
located 2 bp downstream from the CArG box (ABHD5). Double-stranded oligonucleotides (Integrative DNA Technologies, Coralville, IA) corresponding to each CArG element were heated to 65°C for 10 min and then slowly annealed. EMSA was performed with in vitro translated SRF as described previously using a well-defined SRF binding CArG box to the CNNI gene as a positive control (52). Wild-type (WT) oligonucleotide, WT sequence with antibody to SRF, and SNP-modified oligonucleotide samples for each CArG box were studied. Samples were run on a 5% nondenaturing polyacrylamide gel, vacuum dried, and exposed for varying lengths of time to x-ray film.

Luciferase assays. We designed primers (Supplemental Table D4) flanking each of the eight CArG elements shown in Table 1 and PCR cloned intervening regions from total genomic DNA obtained from primary-derived human coronary artery smooth muscle cells. The PCR products were sequence verified and then cloned into the pGL3 basic vector for luciferase assay. Site-directed mutagenesis (Quik Change, Stratagene) was done with primers that contained the SNP defined from our genomic scan. Cells (CO87 and rat pulmonary artery smooth muscle cells) were transfected with either the WT or SNP sequence in the absence or presence of SRF-VP16 (43) and luciferase activity measured 24 h later. Data were normalized to a control Renilla reporter that was cotransfected in all samples. Results are reported as the mean (± SD) of four replicates. One-way ANOVA with Tukey’s t-test was done to compute any statistical differences between WT, WT+SRF, SNP, and SNP+SRF for each gene, but limited reporting to evaluations between WT and SNP reporters with SRF. A probability value of <0.05 was considered statistically significant.

RESULTS

In silico CArG sequence conservation screening. We previously reported on the initial definition of the mammalian CArGome through a computational biology approach that analyzed only a small fraction of RefSeq genes (75). Here, we used the latest sequence assemblies to interrogate an 8 kb window of genomic sequence centered at the TSS in each of 18,925 protein-coding human genes for the presence of CArG boxes. We refer to this collection of CArG boxes as the RefSeq CArGome (RS-CArGome). This analysis revealed a total of 142,597 CArG boxes over a total genomic interval of 151.4 Mb. This number of CArG boxes agrees well with the theoretical number of CArG boxes found in our genomic scan, suggesting there are at least 1,216 permutations of CArG in the human genome (53), a number we factored in our Perl script (see MATERIALS AND METHODS).

The RS-CArGome encompasses 5,213 unique protein-coding genes (27.5% of the initial 18,925 protein-coding genes searched). Of the 209 previously validated mammalian SRF-target genes, all were identified with the initial CArG matching algorithm, with 72% meeting the phastCons conservation threshold value of 0.842. As an independent test for CArG element enrichment, we also evaluated our set of 5,213 suspected SRF-target genes using the oPOSSUM algorithm (29). This analysis confirmed that the majority of sequences common to our set of 5,213 genes were indeed CArG-like sequences (Supplemental Table S7). Other binding sites found (e.g., TATA and homeodomain) reflect the AT-rich central domain of CArG boxes.

We further defined the genomic region where the 8,252 conserved CArG boxes were located and calculated the distance of each CArG box from the TSS (Fig. 3). Given that some conserved CArG elements were discovered in the promoter region of multiple genes (see below), we assumed the gene closest to the CArG box to be the priority SRF target gene and thus determined the genomic region and calculated the distance based on this CArG-SRF target gene pairing (Supplemental Table S5). Over half of the conserved CArG boxes were located downstream from the TSS as indicated by the skewed histogram (Fig. 3A). The two regions with the highest number of conserved CArG elements were 500 bp immediately up- and downstream from the TSS, which is consistent with data showing a direct interaction between SRF and a subunit of the RNA polymerase II holoenzyme (34). While conserved CArG boxes were most commonly found in the immediate 5′ promoter region, a surprisingly high percentage (30%) was...
found within exons (Fig. 3B). We speculate that regulatory element conservation will coincide with protein coding information in many of these cases.

We next used the GREAT algorithm (49) to evaluate the functional significance of conserved CArG regions adjacent to putative SRF-target genes. GREAT analysis revealed enrichment for GO terms of biological processes related to cellular differentiation, the cytoskeleton, nervous system development, muscle cell development, and contraction, as well as tissue morphogenesis (Fig. 4). These classifications are consistent with known roles for SRF in normal and pathological processes (50).

Identification of CArG boxes near histone gene clusters. Bidirectional promoters (79) and co-regulated gene expression patterns (85) are prevalent in the human genome; however, there has been no known SRF-binding CArG elements identified to date that control divergently transcribed genes or gene clusters. To explore the possibility that a given CArG box may be active for multiple and/or divergently transcribed SRF-target genes, we searched for conserved CArG boxes that are located in genomic regions corresponding to divergent promoters. We identified a total of 646 CArG boxes that were located in divergently arrayed promoter regions of two or more genes (Supplemental Table S8).

Fig. 3. Conserved CArG distance to transcription start site (TSS) and genomic region location. A: histogram depicting the frequency of conserved CArGs (y-axis) classified into bins based on the distance of the CArG box to the TSS (in kb, x-axis) of the closest RefSeq gene (represented by the dashed vertical line). B: distribution of conserved CArG elements around associated RefSeq genes.

Fig. 4. Genomic Regions Enrichment of Annotations Tool (GREAT) enrichment analysis of theoretical SRF-target genes based on biological process Gene Ontology (GO) terms. Enrichment analysis was performed using the GREAT algorithm (http://great.stanford.edu/) for each of the 8,252 conserved CArG sequences identified with the in silico method. The GREAT algorithm associates each human gene (n = 17,578) with a regulatory domain in the human genome (hg19 assembly) and calculates the total fraction of the genome annotated with GO terms (e.g., myofibril assembly). The submitted sequences that fall in each annotated GO term region are counted as “hits.” A binomial test compares the expected number of hits in a genome region with the observed number of hits. Listed in the figure are the most significantly enriched biological process GO terms (out of 7,170) for the conserved CArG sequences. Expected and observed counts for each GO term are listed with binomial test P value.
erved CARG boxes adjacent to 26 histone genes on both chromosome 1 and 6. For example, we identified a cluster of three conserved CARG boxes between the divergently transcribed histone genes HIST2H2BE and HIST2H2AC separated by ~300 bp. Despite the close proximity of these three CARG elements over this short genomic interval, we were unable to validate them as SRF dependent using luciferase reporter assay or expression analysis of the histone genes following SRF knockdown (data not shown). This suggests either we have not found a permissive context for these CARG boxes to exhibit function or they represent nonfunctional CARG boxes.

Defining functional CARG-SNPs. A total of 115 unique, genotyped SNPs were identified across the entire RS-CARGome, representing 1.4% of all conserved CARG boxes (Supplemental Table S9). In four cases (TSPAN7, MDK, CHRM4, and TN), the SNP resulted in a tolerable central A-T substitution, suggesting functional SRF-CARG binding would be preserved. Approximately 14% of the conserved CARG elements (1,130) were found to have 1,232 unique SNPs located 35 bp up- or downstream from the CARG box (Supplemental Table S10). For the cumulative 1,347 unique SNPs identified, there are 475 transversions, 784 transitions, 35 deletions, 40 insertions, and 3 mixed polymorphisms. The transversion-to-transition ratio of 1:1.65 is a slightly higher ratio of transversions than typically seen in the genome at large, but consistent with other analysis of rSNPs (21).

We selected eight novel CARG sequences from diverse GO Term Functional classifications exhibiting SNPs for further experimentation (Table 1). We have begun to catalogue these CARG-SNPs in a database using the UCSC genome browser (Fig. 5). Gel shift assays showed that each WT CARG sequence displayed variable binding to in vitro translated human SRF as evidenced by a supershift of the CARG-SRF nucleoprotein complex upon addition of an antibody to SRF (Fig. 6). We next tested the same sequences with the SNP introduced and found in seven of eight cases, the CARG-SNP mutation resulted in reduced SRF binding (Fig. 6). These results suggest that the majority of SNPs within CARG boxes will exhibit attenuated SRF binding. The CARG box associated with ABHD5, which had an SNP 2 bp outside the CARG box, also showed decreased binding. Interestingly, one of the CARG-SNPs (associated with CDH3) exhibited greater SRF binding than that observed with the WT CARG sequence (Fig. 6). This result was unexpected since the CARG-SNP in CDH3 caused a second nucleotide deviation from the consensus CARG and would therefore be less likely to support SRF binding based on the known binding behavior of SRF to CARG (53). This finding underscores unanticipated complexity of SRF binding to CARG sequences (see discussion).

To further evaluate the functional activity of the eight WT CARG sequences and the consequence of each CARG-SNP therein, we cloned a portion of regulatory sequences encompassing each CARG box (Supplemental Table S4) from human genomic DNA and cloned the sequences into a luciferase reporter plasmid. We then performed site-directed mutagenesis to create each CARG-SNP and transfected two cell types with either the WT CARG or each respective CARG-SNP for SRF-dependent luciferase activity. In two independent studies, we noted that WT CARG sequences that bound SRF tended to exhibit enhanced luciferase activity upon cotransfection with SRF-VP16 (see MATERIALS AND METHODS). Introduction of the SNP impaired basal luciferase activity in four of the eight
CArG boxes studied and reduced SRF-dependent transactivation, consistent with reduced SRF binding (Fig. 7). This overall trend of decreased promoter activity mirrors the EMSA binding data. However, given that SRF interacts with >60 cofactors (53), it is unclear if the specific luciferase assay utilized in this study is the appropriate context to evaluate all CArG-SNPs found by our methods. Also, consistent with the EMSA results, the promoter activity for CDH3 was significantly increased with the introduction of the SNP. Taken together, these results have greatly expanded the human CArGome and demonstrate the functional activity of rSNPs that fall within or adjacent to CArG boxes.

DISCUSSION

Efforts to elucidate the impact of rSNPs on human disease have garnered significant attention since completion of the sequencing phase of the human genome project. Much of the work in this area has concentrated on GWAS. With the advent of SNP chips, the number of GWAS has drastically accelerated in recent years (25). While estimates vary widely, some indicate the number of SNPs associated with human disease to be as high as 52,000 (33). Unfortunately, we are left with little understanding of biological mechanisms or function of these 52,000 SNPs, which remains one of the major criticisms of GWAS (15, 48). While identification of genetic risk has notable merit (46), understanding the functional significance of SNPs will lead to a deeper understanding of the cellular and molecular mechanisms of disease and facilitate the development of more clinically-relevant preventive and therapeutic interventions (27).

In silico models of TFBS prediction are lauded for cellular and environmental independence and perform adequately in lower organisms, but their success as predictors of functional rSNPs in more complex organisms has been limited (14). This weakness may stem in part from incomplete PWMs. For instance, the PWM stored in the Jaspar database (64) for CArG boxes (MA0083.1 accessed at http://jaspar.genereg.net on 13 April 2011) is based on one study (62) and does not represent what is currently known for this well-characterized TFBS (53). Furthermore, while identifying conserved nonprotein coding regions via multispecies alignments and phylogenetic footprinting has further improved in silico rSNP identification models (30, 68, 77, 83), many of these comparative approaches require strict multispecies alignment of entire promoter regions and, as a result, have relatively poor sensitivity and specificity of TFBS prediction in vivo.

We have created the largest dataset of putative SRF-CArG sequences known to date. Our sequence-based approach was intentionally broad, so as to capture all CArG elements that adhere to our extensive knowledge of its well-validated sequence motif. Discovery of transcription factor target genes by array-based methods would likely capture only a subset of all target genes. Here, we utilize computational independence to avoid the limitations of experimental context introduced by
The advantage of our method for identifying rSNPs in the RS-CArGome is based upon the integration of in silico and biological techniques. First, we employed a pure sequence-based algorithm based on the known CArG sequences that should bind SRF to minimize TFBS prediction errors. Second, we used a conserved, multispecies approach by utilizing a phastCons score, which is based on whole genome alignments from 46 vertebrate species. While whole genome alignment algorithms have acknowledged weaknesses (9), the advantage of using phastCons is that a conservation probability score is determined for each individual base in the aligned genomes and can be scaled to include only a specific sequence of interest. In our case, only the 10 bp CArG box adjacent to RefSeq genes required conservation, as opposed to previous approaches that require entire promoter sequence alignment based on fewer species (82). Third, we screened the human RS-CArGome to identify rSNPs within or adjacent to the CArG box to test whether such rSNPs induced alterations in SRF-CArG binding and function. Importantly, we demonstrate decreased SRF-CArG binding in seven of eight selected CArG boxes harboring an internal or external rSNP. We further show decreased promoter activity in four of these CArG-SNPs via luciferase assay. By definition, an SNP substitution in a conserved CArG will result in a CArG-like box, while an SNP in a CArG-like sequence should result in a nonconforming SRF binding site. This variation in the CArG box is significant because SRF binds to conserved CArG elements more tightly than to CArG-like boxes (26, 51). Consistent with this notion, validation EMSA gels for ADRB2, KCNA3, RALYL, and KLF6, carrying an SNP located within a consensus CArG box, show diminished SRF binding. Additionally, the decreased binding pattern displayed in ABHD5 suggests that SNPs in close proximity to the CArG box may also impact SRF-dependent transcription by potentially disrupting SRF’s interaction with neighboring transcription factors. Our in silico sequence and conservation-based approach significantly expands the human CArGome, and the resulting dataset represents one of the largest collections of computationally predicted TFBS reported to date with insights into the functional consequences of CArG-SNPs. Interestingly, we found 14 CArG-SNPs that fall within haplotype blocks linked to various human phenotypes, including Type 2 diabetes, coronary artery disease, and ischemic stroke (Supplemental Table S11). Additionally we conducted linkage disequilibrium (LD) analysis with our CArG-SNPs and discovered a number of known GWAS SNPs in LD with CArG-SNPs (Supplemental Table S12). None of the CArG-SNPs reside within the 10 bp CArG element. Whether the CArG-SNPs alter SRF binding and/or contribute to phenotypic etiology is unknown.

Our present study likely underestimates the number of biologically active human CArG boxes. While a conservation-based approach has merits, it may have resulted in an increased number of false negative CArG boxes by eliminating nonconserved CArG boxes. Additionally, the discovery of a functionally active CArG box >10 bp (37, 41, 54) and the use of immunoprecipitation-based assays to identify several putative SRF target genes containing a CArG box deviating from the 1 bp substitution rule (88) suggest there are >1,216 permutations of the CArG box than we searched for here. In this context, we show that an atypical CArG box with two substitutions associated with the CDH3 gene supports greater SRF binding compared with the same CArG element with just 1 bp substitution. Finally, although SRF has known microRNA targets (56, 90) and SNPs have been shown to alter microRNA function (74), we did not include microRNA targets in this analysis. However, a recent microRNA-seq study revealed a large number of CArG-containing microRNAs in gastrointestinal smooth muscle with many shown to be regulated by SRF (58).

Our approach was limited to an 8 kb window of genomic sequence around HUGO protein-coding genes representing <5% of the human genome. If we consider the remaining nonredundant genomic sequence outside the window of analysis we interrogated and a phastCons threshold of 0.97099 (average of all 8,252 CArG boxes defined here, see Supplemental Table S5), we may infer a conservative estimate of some 50,000 CArG boxes outside the RS-CArGome. Assuming that a similar percentage of these CArG boxes will harbor rSNPs as those reported here (1.4%), we estimate there will be ~700 additional CArG-SNPs in the human genome. Moreover, the predominance of only a subset of CArG sequences (146/1,216) in >50% of the RS-CArGome indicates the existence of a selection bias for specific CArG sequences in this region of the genome. It will be interesting to determine whether this CArG sequence bias extends outside the RS-CArGome. Moreover, it may prove true that many nonprotein coding RNAs will be dependent on SRF for expression. In this context, it will be important to apply our computational method to the remainder of the genome and assess functionality using high-throughput technologies such as ChiP-seq. It must be emphasized however, that ChiP-seq alone, while powerful, is limited to the cell type and cell context under analysis and will always underestimate the totality of TFBS (59). For instance, SRF binding-site identification by ChIP-seq was used to find 2,429 combined density profile peaks in human Jurkat cells (80) and 1,262 peaks in a macrophage cell line (73). Furthermore, Cooper et al. (12) used ChiP-chip to study SRF binding in neuronal, smooth muscle, and Jurkat cells and identified 216 SRF binding sites. These numbers are lower than those generated through computational means (89). Nevertheless, in silico screening methods employed in isolation will have inherent limitations as well but will be enhanced in combination with wet-lab assays such as ChiP-seq to yield more robust results. In particular, incorporating the whole genome TFBS and chromatin state data from the ENCODE project will undoubtedly improve computational screening for functional elements in the genome and assist in identifying rSNPs linked to human disease (66).

Integrating remote CArG elements identified by in silico whole genome curation with ChiP-seq may reveal novel SRF functions. For example, studies in yeast suggest a role for nonpromoter CArG elements in DNA replication (6). Moreover, it is possible that CArG boxes associate with such genomic functions as recombination and structural integration of DNA with surrounding histones or nucleoskeletal proteins.
Interestingly, we identified multiple CArG boxes near histone gene clusters, but the set of three CArG elements within a ~300 bp region positioned between two divergently transcribed histone genes were not biologically active in our limited wet-lab analysis (data not shown). Moreover, expression of several zinc-finger transcription factors with multiple conserved CArG elements was found to be unaltered upon knockdown of SRF (data not shown). There are two possibilities for these surprising results. First, we may not have found the correct cell culture context for these CArG elements to display activity. Alternatively, the results may simply reflect the fact that these CArG boxes are inactive, which would imply the existence of false positives among our dataset of 8,252 conserved CArG boxes. While in silico approaches will likely have a higher false discovery rate, high throughput assays such as ChIP-seq may have a higher false negative rate stemming from the biological context employed in performing these assays.

Several individual rSNPs have been functionally validated using traditional wet-lab techniques, including luciferase reporter assays, EMSA, microarray, and ChIP studies (10, 16, 57, 65, 81). Furthermore, many screening, annotation, and prioritization resources exist to identify potentially functional SNPs as high probability for inclusion in GWAS, such as CASCAD (22), SNPnexus (7), SNPLogic (61), SNPinfo (87), and FitSNPs (8). The weakness of these tools is they are overly dependent on in silico algorithms, typically require an a priori set of genes, and lack wet-lab functional validation, especially with regard to non-protein-coding SNPs (35). The ORegAnno database stores experimentally identified DNA regulatory regions and regulatory variants (20). However, ORegAnno contains only 175 human rSNPs and appears to be outdated and not maintained for use as a practical resource for functional SNP evaluation. Thus, there remains a need for active central repositories of functionally validated rSNPs merged with disease-associated SNPs reported in GWAS studies. The results of this study represent a foundation for the development of such a resource.

ACKNOWLEDGMENTS

The authors thank Drs. Rob Fortuna and Brett Robbins from the University of Rochester Combined Internal Medicine-Pediatrics Residency Program for project support.

GRANTS

This work was supported by National Heart, Lung, and Blood Institute Grants HL-62572 and HL-091168 to J. M. Miano.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES


