Cis-regulatory mutations in human disease

Douglas J. Epstein

Advance Access publication date 29 July 2009

Abstract

Cis-acting regulatory sequences are required for the proper temporal and spatial control of gene expression. Variation in gene expression is highly heritable and a significant determinant of human disease susceptibility. The diversity of human genetic diseases attributed, in whole or in part, to mutations in non-coding regulatory sequences is on the rise. Improvements in genome-wide methods of associating genetic variation with human disease and predicting DNA with cis-regulatory potential are two of the major reasons for these recent advances. This review will highlight select examples from the literature that have successfully integrated genetic and genomic approaches to uncover the molecular basis by which cis-regulatory mutations alter gene expression and contribute to human disease. The fine mapping of disease-causing variants has led to the discovery of novel cis-acting regulatory elements that, in some instances, are located as far away as 1.5 Mb from the target gene. In other cases, the prior knowledge of the regulatory landscape surrounding the gene of interest aided in the selection of enhancers for mutation screening. The success of these studies should provide a framework for following up on the large number of genome-wide association studies that have identified common variants in non-coding regions of the genome that associate with increased risk of human diseases including, diabetes, autism, Crohn's, colorectal cancer, and asthma, to name a few.

Keywords: cis regulation; transcription; gene expression; human disease

INTRODUCTION

The field of gene regulation is currently undergoing a renaissance. With the successful annotation of most of the protein-coding portion of the human genome [1], the focus of much research has shifted toward deciphering the regulatory logic governing the temporal, spatial and quantitative aspects of gene expression that is embedded in the remaining 98% of DNA that does not encode for protein [2]. A flurry of papers stemming, in large part, from two broad areas of investigation has recently made a significant impact on the field of gene regulation. The first revolves around the genetic basis of human disease. Fueled by the power of linkage and genome-wide association studies, an ever-expanding list of human diseases has been associated with single nucleotide polymorphisms (SNPs) residing in noncoding regions of the genome [3]. These disease-associated SNPs are thought to directly control some aspect of target gene expression, or are linked to other DNA variants that possess regulatory activity. In a small but growing number of cases, the regulatory SNPs identified in human genetic studies have led to the identification of disease susceptibility loci and have served as useful entry points for unraveling the complexities of the gene regulatory landscape (Table 1) [3]. The second line of investigation that has revitalized gene expression research relates to the development of functional genomic approaches to screen noncoding DNA for regulatory potential. Genome-wide surveys of sequence conservation [4–6], histone modifications [7–8], DNAse I hypersensitivity [9] and DNA structure [10], have all significantly improved the detection of functional cis-acting regulatory sequences. This review will highlight recent examples from the literature that...
have successfully integrated genetic and genomic approaches to uncover the molecular basis by which *cis*-regulatory mutations alter gene expression and contribute to human disease.

**CIS-REGULATION OF GENE EXPRESSION**

The transcription of RNA polymerase II (Pol II)-dependent genes is mediated by the recruitment of general and sequence-specific transcription factors to *cis*-acting regulatory sequences including core and proximal promoter elements that reside within 1 kb of the transcription start site (TSS), as well as enhancers, repressors, insulators and locus control regions that can act over considerable distance (Figure 1) [11]. Productive transcription is also reliant on chromatin structure and modification states [7]. *Cis*-acting regulatory mutations generally disrupt some facet of the transcriptional activation process [11].

It is important to distinguish this class of mutations from those that interfere with target gene expression through other means, including mRNA splicing, stabilization, degradation, poly-adenylation, etc.

**CIS-REGULATORY MUTATIONS AND HUMAN DISEASE: SOME BASIC PRINCIPLES**

The notion that mutations in *cis*-acting regulatory sequences are a significant cause of human disease is not new. According to April 2009 statistics compiled by the Human Gene Mutation Database [12], 1459 regulatory mutations have been identified in over 700 genes that cause human-inherited disorders. Between 1% and 2% of disease causing point mutations are in noncoding regions of the genome. The majority of these regulatory mutations is located in proximal and distal promoter elements that map within 1 kb of the TSS.

*Cis*-regulatory mutations affect a broad range of morphological, physiological and neurological phenotypes (Table 1). Classic examples of diseases caused by regulatory mutations include beta-thalassemia, hemophilia and atherosclerosis [13]. Each of these simple Mendelian disorders is caused by the disruption of a single gene: beta-chain of hemoglobin (*HBB*); coagulation Factor IX and low-density lipoprotein receptor (*LDLR*), respectively. *Cis*-regulatory mutations in each of these genes have several features in common. Firstly, they are less frequent than coding mutations. Typically, the promoter region of a candidate disease gene is only screened for mutations after coding mutations have been ruled out. Secondly, *cis*-regulatory mutations result in a significant reduction in target gene transcription in the relevant cell type (erythrocytes/*HBB*,

### Table 1: Examples of regulatory mutations associated with human disease

<table>
<thead>
<tr>
<th>Gene</th>
<th>Disease</th>
<th>Location of rSNP</th>
<th>TF-binding site affected</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBB</td>
<td>β-thalassemia</td>
<td>Promoter</td>
<td>Several (TATA, CACCC, EKLF)</td>
<td>[13]</td>
</tr>
<tr>
<td>F9</td>
<td>Hemophilia B</td>
<td>Promoter</td>
<td>Several (HNF4, C/EBP)</td>
<td>[13]</td>
</tr>
<tr>
<td>LDLR</td>
<td>Familial hypercholesterolemia</td>
<td>Promoter</td>
<td>Several (SPI, SRE repeat)</td>
<td>[13]</td>
</tr>
<tr>
<td>CollA</td>
<td>Osteoporosis</td>
<td>Intron 1 (+2kb)</td>
<td>SPI (gain)</td>
<td>[4]</td>
</tr>
<tr>
<td>RET</td>
<td>Hirschprung</td>
<td>Introns (+9.7 kb)</td>
<td>Unknown</td>
<td>[23]</td>
</tr>
<tr>
<td>HBA</td>
<td>α-thalassemia</td>
<td>Upstream (−13 kb)</td>
<td>GATA1 (gain)</td>
<td>[25]</td>
</tr>
<tr>
<td>SHH</td>
<td>Preaxial polydactyly</td>
<td>Upstream (−1 Mb)</td>
<td>Unknown</td>
<td>[29]</td>
</tr>
<tr>
<td>SHH</td>
<td>Holoprosencephaly</td>
<td>Upstream (−470 kb)</td>
<td>Six3</td>
<td>[37]</td>
</tr>
<tr>
<td>SOX9</td>
<td>Pierre Robin Sequence</td>
<td>Upstream (−1.5 Mb)</td>
<td>Msx1</td>
<td>[41]</td>
</tr>
<tr>
<td>IRF6</td>
<td>Nonsyndromic cleft lip</td>
<td>Upstream (−14kb)</td>
<td>Ap2</td>
<td>[42]</td>
</tr>
</tbody>
</table>

Figure 1: *Cis*-regulatory sequences implicated in gene expression. Schematic showing the location of proximal and distal regulatory elements with respect to the transcriptional TSS of a given gene. Abbreviations: BE, boundary element (gray oval); CP, core promoter (orange oval); E, enhancer (green oval); IE, intronic enhancer (green oval); PP, proximal enhancer (blue oval); R, repressor (red oval). Exons are represented by black boxes.
liver/Factor IX, liver (primarily)/LDLR). Thirdly, each of the regulatory mutations alters the DNA sequence of a transcription factor-binding site that impairs the recruitment of a key transcription factor required for RNA polymerase II (pol II)-dependent synthesis of mRNA transcripts [13].

Mutations in cis-acting regulatory sequences can also predispose individuals to disease by increasing the amount of transcribed product. For example, individuals carrying a common polymorphism (G/T) in an Sp1-binding site in the first Col1A1 intron are at increased risk of osteoporosis, a bone fragility disorder resulting from a reduction in bone mass [14]. Electromobility shift assays (EMSAs) showed that the zinc finger containing transcription factor, Sp1, bound to the risk allele ‘T’ with increased affinity over the ‘G’ allele, causing a 3-fold increase in Col1A1 transcription [15]. Consequently, the osteoblasts of G/T individuals showed an altered ratio of type I collagen chains compared to G/G homozygotes, likely accounting for the decrease in bone mineral density.

**THE GENETICS OF GENE EXPRESSION**

The previous example demonstrates how a common regulatory SNP can predispose individuals to a complex genetic disease by modifying the expression level of a given gene. Several studies have recently expanded on this concept by showing that variation in gene expression is widespread in the human genome [16]. Humans are more polymorphic at functional regulatory sequences than they are in coding exons [17]. Interestingly, variation in gene expression is highly heritable and can be mapped in humans, and other organisms, as a quantitative trait [18–20]. These experiments used DNA microarrays to measure the abundance of thousands of mRNA transcripts expressed in cells derived from extensively genotyped pedigrees. Genome-wide linkage and association studies were then performed to map the genetic determinants underlying the gene expression differences between individuals. Whereas trans-acting determinants were generally more numerous, cis-acting signals showed a consistently stronger influence on gene expression [16]. Cis-acting determinants were enriched close to the TSS with only 5% mapping further than 20-kb upstream [21]. While there is clearly a bias in rSNPs at promoters, experimental evidence is accumulating which suggests that variation in the sequence of local and remote enhancers also plays an important role in the genetics of gene regulation and disease pathogenesis [22].

**A CANDIDATE GENE APPROACH TO IDENTIFYING REGULATORY MUTATIONS**

What do you do when mutations aren’t identified in a candidate gene despite the preponderance of genetic evidence supporting its association with a particular disease? More and more researchers are facing this predicament. This was the case for the Chakravarti lab in their effort to identify the genetic risk factors associated with Hirschprung disease (HSCR) [23]. HSCR is a congenital defect of the colon (aganglionic megacolon) resulting from the failure of neural crest-derived enteric neurons to populate and innervate the gut. HSCR is a relatively common disorder (1/5000 live births) with a complex pattern of inheritance. Rare coding mutations in the receptor tyrosine kinase RET, in combination with mutations in other genes, account for less than 30% of HSCR cases. This raises the obvious question: what is the cause of HSCR in the 70% of cases with no known genetic lesion? The possibilities include, additional rare mutations in many other genes, or common low-penetrance variants in a small number of genes. Initial support for the latter hypothesis came from linkage studies in an inbred Mennonite population, which identified a high frequency HSCR-associated RET haplotype on chromosome 10q that lacked coding mutations in RET [23]. Importantly, for the common variant hypothesis, the RET Mennonite haplotype is also present in the general population and is overtransmitted to offspring with HSCR. Using a combination of family-based association studies, resequencing and transmission disequilibrium tests (TDTs), the genomic interval harboring the disease-associated variant(s) was narrowed down to a 900 bp multispecies conserved sequence (MCS) in RET intron 1 [23]. Three nucleotide variants in complete linkage disequilibrium were segregating in the HSCR-associated MCS. Inheritance of these common non-coding variants increases the risk of HSCR by 20-fold compared to the rare coding mutations. The next step was to determine the significance of the HSCR-associated variants on RET expression. Quantifying differences in gene expression with
respect to genotype, while feasible in some readily accessible human cell types, remains a significant challenge during human embryonic development. It is therefore common practice to use in vitro assays, or experimental organisms as a proxy. To determine whether the MCS possessed cis-regulatory function, the authors performed luciferase reporter assays in Neuro-2a cells [23]. The wild-type MCS conferred robust luciferase expression, whereas transcriptional activity from the mutant MCS was significantly impaired. A follow-up study confirmed that the wild-type MCS functioned as a tissue-specific enhancer in vivo [24]. Transgenic embryos containing the wild-type MCS directed RET-like expression to the neural crest-derived enteric ganglia and other sites. Unfortunately, the impact of the HSCR MCS on in vivo reporter activity was not described. The molecular mechanism by which the HSCR-associated variants causes a reduction in RET expression remains to be determined. Presumably one or all of the cis-acting variants alter the recruitment of transcriptional activators to these sites. Identifying the trans regulators of RET expression should offer novel insights into the regulatory network underlying enteric nervous system development and HSCR disease pathogenesis.

In another study, a combination of genetic and genomic approaches was taken to investigate the molecular etiology of α-thalassemia in a group of individuals from Melanesia who, despite the absence of known mutations, showed a significant reduction in α-globin transcription [25]. The human α-globin cluster resides on chromosome 16 and contains an embryonic gene, two minor α-like genes, two α-globin genes and two pseudogenes. Severe anemia results when α-globin expression falls below 50% of its normal level. Genetic studies in Melanesian individuals indicated that the α-thalassemia phenotype mapped to a 213 kb genomic interval spanning the α-globin cluster [25]. To identify the causative SNP under the peak segregating with α-thalassemia in affected individuals and not found in 131 nonthalassemic, Melanesian individuals. Chromatin immunoprecipitation (ChIP) assays were used to show that the regulatory mutation created a new GATA-1-binding site that recruited an erythroid-specific transactivation complex, which interfered with the normal transcription of the α-globin genes, thus causing α-thalassemia [25].

**GENE EXPRESSION BY REMOTE CONTROL**

Long-range enhancers have also been identified as targets of mutation in human disease [22]. A particularly compelling example is the association of preaxial polydactyly (PPD), extra digits on the anterior (thumb) side of the hands and/or feet, with mutations in the Sonic hedgehog (Shh) limb bud enhancer (Figure 2) [26]. Shh is a secreted morphogen,
expressed in the posterior portion of the developing limb bud, a region coined by classical developmental biologists as the zone of polarizing activity (zpa) [27]. The polarizing properties of the zpa were realized by experimental manipulations of this tissue in chick embryos. Transplanting the posterior limb mesoderm to the anterior side resulted in chicks with supernumerary digits [26]. Shh mediates the polarizing properties of the zpa and is required to promote the growth and identity of digits along the anteroposterior axis of the limbs [27]. Identification of the Shh limb bud enhancer (zrs, zone of polarizing regulatory sequence) was aided by the serendipitous discovery of a mouse line carrying a transgene insertion on mouse chromosome 5, in close proximity to the zrs [28]. Mice harboring the transgene presented with PPD, due to ectopic Shh expression in the anterior limb [28]. Subsequent genomic analysis identified a highly conserved 800 bp DNA sequence that was both necessary and sufficient to direct Shh expression to the zpa [29–31]. The zrs proved to be a good candidate for mutation screening in familial cases of PPD mapping to 7q36, the corresponding location of human SHH. Distinct point mutations in the zrs were identified in four unrelated familial cases of PPD [29]. Moreover, point mutations were also identified in the zrs of various lines of polydactylous mice and domestic cats, including the famed Hemingway’s cat [29,30,32]. In each case, PPD appeared to manifest from the ectopic expression of Shh in the anterior limb bud (Figure 2) [32].

The molecular mechanisms, by which the zrs functions to activate and repress Shh transcription in the posterior and anterior limb bud, respectively, are unclear. The finding of 12 independent mutations scattered throughout the 800 bp zrs may reflect the need for several transcription factors acting in combination to repress Shh expression. This model is consistent with genetic data in the mouse, where mutations in multiple transcription factors including, Twist1, Alx4 and Gli3, lead to ectopic Shh expression and a PPD phenotype [33–35]. Alternatively, the various point mutations in the zrs may disrupt the chromosomal dynamics needed to maintain Shh transcription in a repressed state in the anterior limb bud [36]. It is particularly intriguing that no point mutations have been described in the zrs that cause a loss of Shh transcription. Since the targeted inactivation of the zrs in mouse embryos causes a loss of digits [31], it is conceivable that activation of Shh transcription in the posterior limb may not wholly depend on any one binding site.

The zrs is not the only long-range Shh enhancer associated with a human developmental disorder. A mutation that disrupts a binding site for the Six3 homeodomain protein in Shh brain enhancer-2 (SBE2) was identified in a patient with holoprosencephaly (HPE) [37]. HPE is a structural brain defect resulting from haploinsufficiency in Shh, Six3 or several other genes [38]. The mutation in SBE2, a previously characterized Shh forebrain enhancer mapping 470 kb upstream of Shh, was instrumental in determining that Six3 acts as a direct regulator of Shh transcription [37,39,40].

Mutations in critical binding sites of long-range enhancers have also been described in other congenital anomalies affecting craniofacial development. Misexpression of the transcriptional regulator, SOX9, due to disruption of an Msx1-binding site in a mandibular enhancer mapping ~1.5 Mb distal to SOX9 is a cause of Pierre Robin sequence, a severe form of cleft palate [41]. Additionally, a common variant in an IRF6 enhancer that disrupts an AP-2α-binding site was recently shown to increase the risk of nonsyndromic cleft lip [42].

**CONCLUDING REMARKS**

What stands out about several of the aforementioned studies is the tremendous advantage of combining genetic and genomic approaches to narrow down disease-causing cis-regulatory mutations. In several instances, the fine mapping of disease-causing variants has led to the discovery of novel cis-acting regulatory elements. While in other cases, the prior knowledge of the regulatory landscape surrounding the gene of interest aided in the selection of enhancers for mutation screening.

Most of these studies relied on multi-species sequence alignment to predict DNA with regulatory potential. Evolutionary constraint is an extremely powerful tool for identifying functional regulatory elements in the human genome [4–6]. However, since not all human regulatory elements are conserved across phyla, additional methods are needed to identify newly evolved human enhancers. Genome-wide approaches of identifying functional regulatory elements have recently been described that nicely complement comparative sequence-based methods. For instance, the genome-wide mapping of histone modification patterns has identified...
55,000 potential tissue-specific human enhancers [8]. High-resolution mapping of DNase I hypersensitive sites has also been successfully applied to identify functional regulatory elements [9]. Interestingly, algorithms based on the structural profile of DNA can predict cis-regulatory sequences in the human genome [10]. A database has been constructed with changes in the structural profile of all known human SNPs [10].

A large number of genome-wide association studies have identified common variants in noncoding regions of the genome that associate with increased risk of a variety of human diseases including, diabetes, autism, Crohn’s, colorectal cancer and asthma [3]. Databases that compile information on the regulatory potential of the human genome should greatly assist in identifying the functional rSNPs that associate with these and other human diseases.

### Key Points
- Variation in gene expression is highly heritable and a significant determinant of human disease susceptibility.
- The fine mapping of disease-causing variants has led to the discovery of novel cis-acting regulatory elements that, in some instances, are located as far away as 1.5 Mb from the target gene.
- The integration of genetic and genomic approaches should greatly assist in identifying the functional noncoding regions of the genome that confer increased risk for a variety of human diseases.

### Acknowledgement
This manuscript is dedicated to the memory of Richard S. Spielman, a brilliant colleague and friend.

### FUNDING
The National Institutes of Health (NS039421) and March of Dimes (#1-FY08-421).

### References


