Genome Medicine

REVIEW

Identification of *cis*-regulatory sequence variations in individual genome sequences

Rebecca Worsley-Hunt[1,2], Virginie Bernard[1] and Wyeth W Wasserman[1]*

Abstract

Functional contributions of *cis*-regulatory sequence variations to human genetic disease are numerous. For instance, disrupting variations in a HNF4A transcription factor binding site upstream of the Factor IX gene contributes causally to hemophilia B Leyden. Although clinical genome sequence analysis currently focuses on the identification of protein-altering variation, the impact of *cis*-regulatory mutations can be similarly strong. New technologies are now enabling genome sequencing beyond exomes, revealing variation across the non-coding 98% of the genome responsible for developmental and physiological patterns of gene activity. The capacity to identify causal regulatory mutations is improving, but predicting functional changes in regulatory DNA sequences remains a great challenge. Here we explore the existing methods and software for prediction of functional variation situated in the *cis*-regulatory sequences governing gene transcription and RNA processing.

Introduction

Cis-regulatory sequences control when, where and with what intensity genes are expressed. Key to this control is the recruitment of transcription factors (TFs) that bind to regulatory sequences, such as promoters, enhancers, repressors and insulators. These target sequences are spread across DNA. Probably reflecting the secondary structure properties of looped DNA within a nucleus, there are confirmed cases of *cis*-regulatory elements up to about 10^6 bp distant from the transcription initiating promoter of a gene [1]. Mutations in TF binding sites (TFBSs) can disrupt the essential protein-DNA interactions required for the appropriate patterning or magnitude of gene expression. Similarly, mutations can disrupt other sequence-specific regulatory controls, such as elements regulating RNA splicing or stability. Although much emphasis in the age of exome sequencing has been placed on variation within protein-encoding sequences, it is apparent that regulatory sequence disruptions will become a key focus as full genome sequences become widely accessible for medical genetics research.

The emerging collection of *cis*-regulatory variations that cause human disease or altered phenotype is growing [2-4]. Reports identify *cis*-acting, expression-altering mutations observed within introns, far upstream of genes, at splicing sites or within microRNA target sites. For example, *cis*-regulatory mutations have roles in hemophilia, Gilbert's syndrome, Bernard-Soulier syndrome, irritable bowel syndrome, beta-thalassemia, cholesterol homeostasis and altered limb formation [5-11]. The number of *cis*-regulatory variants reported in the literature has continued to expand over the past 2 years [12-18]. In addition, compilations of *cis*-regulatory variants have been reported [4,19,20]. Although many studies associate *cis*-regulatory variations with phenotype, it is rare for researchers to conclusively demonstrate causality. The strongest causal evidence is obtained with transgenic approaches, in cell culture or animal models, to identify phenotypes triggered by such variations [21,22]. The importance of regulatory changes is nevertheless apparent.

Ultimately genetics researchers seeking regulatory mutations are best served by high-quality annotations of the human genome, with clearly designated functional elements. Most routinely expressed protein coding exons are known, making initial identification of protein-altering genetic changes simple. In contrast, despite ongoing ambitious efforts to annotate non-coding genome features, the inventory of *cis*-regulatory elements is far from complete. Large-scale chromatin immunoprecipitation (ChIP) experiments provide the vast majority of data, eclipsing the compiled information of the past 25 years derived from targeted studies of specific regulatory elements. Many of the new ChIP-derived data, however, highlight segments of DNA (about 200 to 1,000 bp) containing a functional element rather than a

*Correspondence: wyeth@cmmt.ubc.ca
[1]Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada
Full list of author information is available at the end of the article

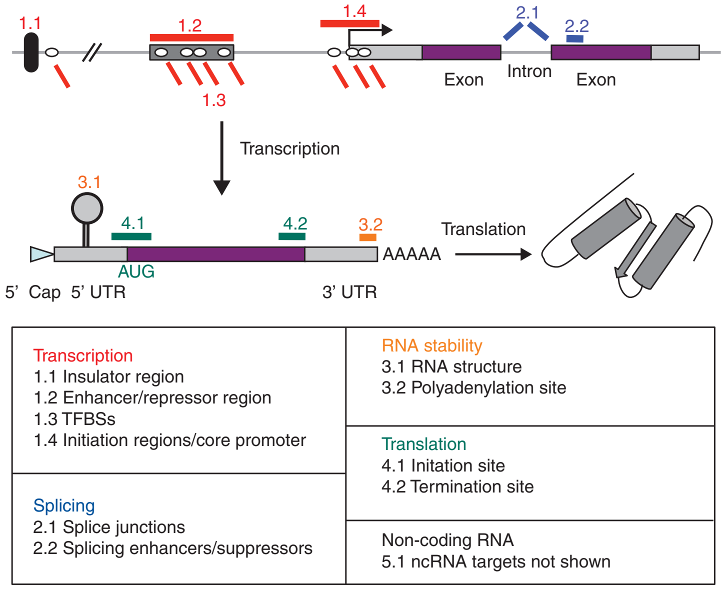


Figure 1. Classification of the regulatory sequences in a gene that can be modified by genetic variation. The four colored groups reflect mutations that have a deleterious impact on gene transcription (class 1 in red), splicing (class 2 in blue), RNA stability (class 3 in orange) and translation (class 4 in green). A fifth class, non-coding RNA interaction sites, is not shown as sites can occur throughout DNA and RNA sequences.

specific element (average <15 bp). Similarly, DNase I hypersensitivity analysis specifies regions likely to contain regulatory elements [23]. Thus, the experimentally defined regions must be coupled to additional methods to assess the potential for a specific DNA variation to affect gene activity. Some of the key data resources reporting regulatory regions and delineating specific elements are introduced below.

Our perspective is biased to elements with sequence-specific properties, including TFBSs, microRNA, splice-regulating target sequences, and immediate core-promoter sequences critical to the initiation of transcription (Figure 1). Although diverse types of *cis*-regulatory variations will become accessible for future studies, at present the bioinformatics resources for the study of variation within TFBSs are the most accessible and therefore our primary focus here.

We begin by outlining an example workflow. Then we step through elements of the workflow in greater detail, including a brief overview of the discovery of sequence variations from high-throughput sequence data. Finally, we review TFBS identification approaches and strategies for the prioritization of *cis*-regulatory variations for further analysis. We conclude with a brief mention of two emerging experimental techniques that may be used in the future to associate *cis*-regulatory variants and their gene targets. Our aim is to assist medical genetics researchers to identify potential regulatory variants within non-coding regions of the human genome.

A workflow for identifying disease-causing *cis*-regulatory variants

An example workflow for the identification of *cis*-regulatory TFBSs linked to a disease is outlined in the following section, and is illustrated in Figure 2. Given a set of high-throughput sequencing data from an individual, the short sequences are individually aligned to a reference human genome sequence. Sets of overlapping reads are analyzed to determine the genotype at each position for which sufficient aligned sequences are available. Common polymorphisms and rare variants are distinguished relative to the reference genome. For familial studies, variations segregating with the phenotype can be determined, and researchers emerge with a set of disease-causing candidates. For each candidate for causality, it is desirable to assess the potential for the sequence mutation to disrupt a biological function. Many researchers will be content to focus on the subset of candidates predicted to create a severe alteration in a protein sequence. Software for the prediction of damaging changes, including modules from SIFT [24] and PolyPhen-2 [25], identify variants that substitute amino acids expected to cause changes in protein structure, alter a critical position in a protein domain, or change an amino acid of high evolutionary conservation. Many researchers will stop at this step.

For those interested in potential *cis*-regulatory changes, a panel of computational analyses can be performed on the candidate variations. At the core of the processes is a

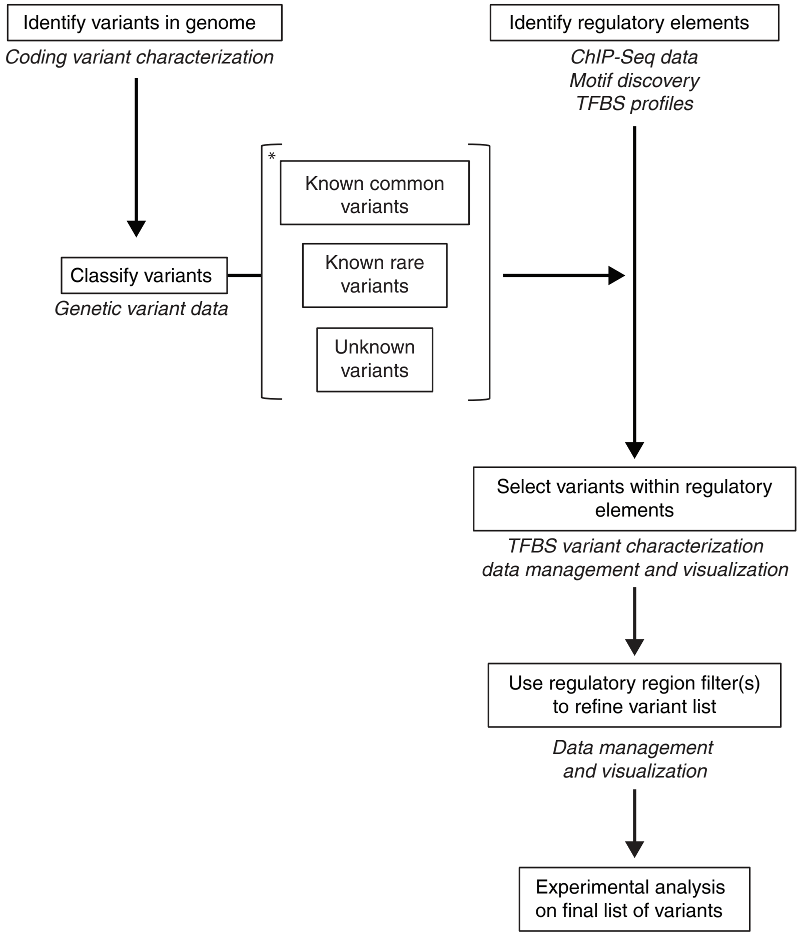

Identify variants in genome
Coding variant characterization
Identify regulatory elements
ChIP-Seq data
Motif discovery
TFBS profiles
*
Known common variants
Known rare variants
Unknown variants
Classify variants
Genetic variant data
Select variants within regulatory elements
TFBS variant characterization
data management and visualization
Use regulatory region filter(s) to refine variant list
Data management and visualization
Experimental analysis on final list of variants

Figure 2. Overview of a workflow for *cis*-regulatory variant detection. The boxes represent steps in the workflow, and the italicized descriptions under the boxes correspond to analysis resources listed in Table 1. For identification of regulatory elements and regions, the order in the workflow may be changed without loss of information. Common variants, flagged by an asterisk, may be eliminated from the analysis or alternatively flagged for later tracking.

method for TFBS prediction based on position weight matrices, known alternatively as position-specific scoring matrices (PSSMs, called ‘possums’). Each matrix is a quantitative description of the frequency of each nucleotide at each position of a set of known TFBSs for a specific TF. Methods related to the generation and application of the matrices are described below. Such matrices can be useful for predicting the biochemical capacity of a TF to interact with a specific DNA sequence, but the models have no capacity to assess whether a specific DNA segment in the genome will be accessible to the TF. Thus, TFBS predictions are almost always combined with one or more ‘filters’ to specify regions of the genome expected to function as *cis*-regulatory regions. Such filters may include data about epigenetic modifications or DNA accessibility, the observed binding of TFs, or sequence conservation (phylogenetic footprinting). The process will be explored in greater detail in the following sections.

For researchers who do not have bioinformatics tools in the laboratory, the comparison of coordinate positions between datasets (between variants and TFBSs, or variants and regulatory region filters) can be done using the Galaxy tools (a set of web-based, fundamental bioinformatic tools for extracting and manipulating text-based data) [26]. Tutorials and help documentation are accessible through the Galaxy wiki.

Overview of variant identification from high-throughput sequence data

The first step in the identification of disease-causing regulatory variants in individual genomes requires both sequencing technologies and software for processing the data to distinguish technical errors from true variations.

Table 1. Data and analysis tools (open-source)

Genetic variant data		
	dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
	1000 Genomes	http://www.1000genomes.org/
	HapMap Project	http://hapmap.ncbi.nlm.nih.gov/
Coding variant characterization		
	SIFT	http://sift.jcvi.org/
	Polyphen-2	http://genetics.bwh.harvard.edu/pph2/
ChIP-Seq data		
	Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
	ENCODE project	http://genome.ucsc.edu/ENCODE/
	PAZAR	http://www.pazar.info/
Motif discovery		
	Meme-ChIP	http://meme.nbcr.net/
TFBS profiles		
	JASPAR	http://jaspar.genereg.net/
	PAZAR	http://www.pazar.info/
TFBS databases		
	PAZAR	http://www.pazar.info/
	ORegAnno	http://www.oreganno.org/
TFBS variant characterization		
	Variant effect predictor	http://uswest.ensembl.org/tools.html
	is-rSNP	http://www.genomics.csse.unimelb.edu.au/is-rSNP/
	rSNP-MAPPER	http://genome.ufl.edu/mapper/
RNA-Seq splice analysis		
	TopHat	http://tophat.cbcb.umd.edu/
	MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSplice
Splice enhancer discrimination		
	SFmap	http://sfmap.technion.ac.il/
	ESE Finder	http://rulai.cshl.edu/tools/ESE
Data management and visualization		
	Galaxy (tool kit)	http://galaxy.psu.edu/
	UCSC Genome Browser	http://genome.ucsc.edu/
	Ensembl BioMart	http://www.ensembl.org/biomart/martview/

Both of these methods are in a period of rapid development [27,28] and it is unlikely that they will become stable for several years. We will therefore begin this section by outlining general concepts for working with high-throughput sequencing data, before highlighting the most promising recent developments. In Table 1 we provide specific examples of well-maintained open-resource databases and software, including resources for conceptual classes of software mentioned below.

The process of identifying a variation or mutation anywhere in a DNA sequence begins with mapping the sequenced DNA of interest to a reference human genome sequence. Sequencing generates DNA segments termed reads. Given a combination of technological sequence errors and genetic variation, coupled to the extensive sequence repetition in the human genome, many reads cannot be uniquely mapped to a single reference co-ordinate region. False variations may arise, in part, from the incorrect mapping of reads. One can identify and set aside reads that map to multiple locations in the genome with nearly equal alignment quality, denoting such cases as potential sources of variant-calling errors.

Once a set of uniquely mapped reads is determined, the next step in an analysis pipeline is the identification of genetic variations from the reference genome based on the shared characteristics of overlapping reads. Greater read depth (the number of times a nucleotide is sequenced) is advantageous for more reliable statistical confidence in the determination of a genotype at a position. But depth is not the only influence to consider, as systematic mismapping of reads can result in false calls even with many reads overlapping. The complex nature of the data causes genotype 'calling' software to be one of the most rapidly changing components of the variant identification process, with continuously improving methods emerging in a constant stream of publications [29].

Given a set of variations, it is common to classify observed variants as common or rare. Common variations, such as single nucleotide polymorphisms (SNPs), occur with at least a minimum frequency in a population (commonly a minimum allele frequency of 1% is applied). Lists of common variants can be obtained from HapMap [30,31], dbSNP [32], and 1000 Genomes [33] databases, but recently dbSNP has been consolidating data from most of the major resources and may be sufficient for most users. Please note that contrary to the formal meaning of SNP, the dbSNP resource also includes rare variations.

After filtering common variants from a dataset, the remaining variants are categorized as rare (and consequently considered by many researchers as more likely to be causal for extremely rare phenotypes) [34]. Although it has been frequent practice in genome sequencing studies of familial disorders to exclude all previously observed variants (both rare variants and polymorphisms), it is our perception that the rapidly growing collection of genome sequence data and variants makes it increasingly likely that such screens will exclude relevant causal mutations. For each individual genome sequenced, the researcher emerges with a categorized set of variants - common polymorphisms, previously reported rare variants, and novel variants. A recent technical report presents a framework for variation discovery and genotyping that reflects the above concepts [35].

For each individual sequenced, the number of variants in each of the three categories will be large. The number of variants to consider can be reduced by focusing on variations that segregate with a phenotype in a family study [36]. In addition, it has been common practice to focus on mutations predicted to severely alter an encoded protein, such as nonsense mutations. However, there is now a growing interest in the impact of variants located in the non-coding portion of a genome. The following sections highlight methods that researchers can use to focus on variants located within *cis*-regulatory elements, in particular TFBSs.

Detection of *cis*-regulatory elements

The identification of variants situated in TFBSs is dependent on the identification of *bona fide* TFBSs, which is a challenge. Common approaches include three primary components: (i) databases of known TFBSs; (ii) computational prediction of TFBSs using software models; and (iii) prediction of regions likely to contain TFBSs - a process we term 'filtering'. The first two components will be introduced below, followed by the introduction of a method highly related to the second component, for predicting the subset of putative regulatory variants likely to alter the binding energy of a TF-DNA interaction. The third component - filtering - is described in the subsequent section.

Databases of experimentally defined TFBSs

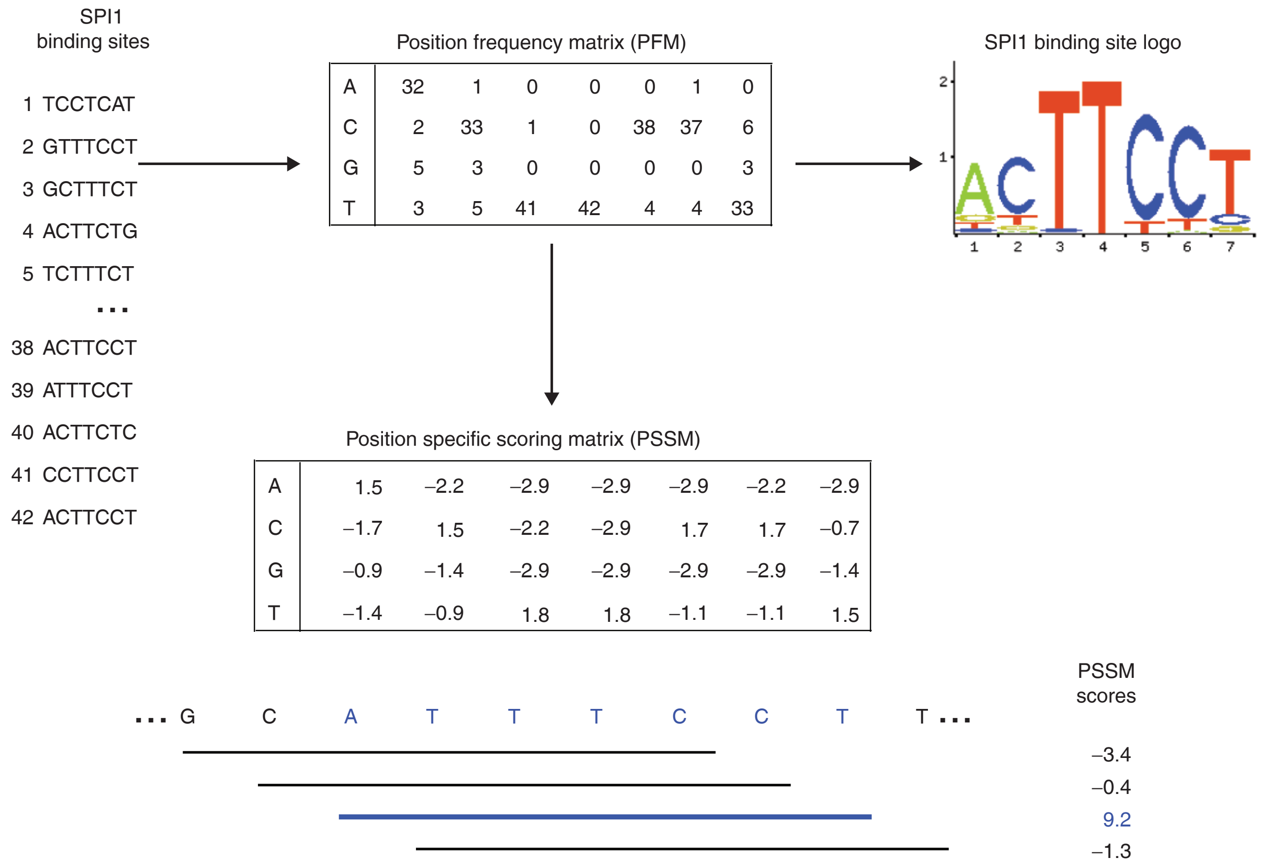
At some point in the future, there will be a reliable database reporting every TFBS in the human genome, the gene promoter(s) that each acts on, the biological condition(s) under which each TFBS is active and the TF(s) that interact with each. There have been attempts over previous decades to develop databases housing such information; nevertheless, none of these are ideal, and all are constrained by a lack of high-resolution experimental data. One of the most widely known databases of this kind is Transfac, which operates under a commercial access model [37]. Open-access proponents have implemented alternative databases. The ORegAnno database aims to collect a broad range of TFBS data, using a convenient data format that allows rapid submission [38]. The PAZAR database has a more complex data model, which allows rich annotation of the evidence underlying each TFBS, as well as a full description of the TFBS (conditions, interacting TFs, cell types, and so on) [39]. There are many databases restricted to lower resolution data, such as ChIP with sequencing (ChIP-Seq), that specify regions of DNA bound by a TF (such data are also contained by ORegAnno, PAZAR, and the ENCODE data center at the University of California, Santa Cruz - UCSC [40]).

The data from the open-access collections can be downloaded for high-throughput comparisons of TFBS positions with variant positions. The logistics of such comparisons are described in the subsection below on 'Assessing the impact of sequence variations on TF-DNA interactions'.

Computational prediction of TFBSs

Although a small set of TFBSs has been experimentally validated as functional and recorded in reference databases, this probably represents an insignificant portion of the entire set of TFBSs in the human genome. We must therefore, for now, rely heavily on computational methods to predict TFBSs. A description of the most common method follows.

A TFBS motif model summarizes what is known, at the sequence level, about the properties of a set of TFBSs for



Position frequency matrix (PFM)

A	32	1	0	0	0	1	0
C	2	33	1	0	38	37	6
G	5	3	0	0	0	0	3
T	3	5	41	42	4	4	33

Position specific scoring matrix (PSSM)

A	1.5	–2.2	–2.9	–2.9	–2.9	–2.2	–2.9
C	–1.7	1.5	–2.2	–2.9	1.7	1.7	–0.7
G	–0.9	–1.4	–2.9	–2.9	–2.9	–2.9	–1.4
T	–1.4	–0.9	1.8	1.8	–1.1	–1.1	1.5

Figure 3. Transcription factor binding site (TFBS) motif model. Sequences known to be bound by a specific transcription factor (for example, SPI1) are aligned. A position frequency matrix (PFM) is generated by counting the number of times each type of nucleotide occurs at each position of the alignment. The PFM is then converted to a log-scale position-specific scoring matrix (PSSM). The score of any DNA sequence window, having the same length as the matrix, is calculated by summing the corresponding nucleotide values from the PSSM. The PFM may also be represented as a binding site logo, depicting the nucleotide properties of the TFBSs.

a given TF (Figure 3). Once constructed, such a model can be used to scan a DNA sequence of interest in 'windows' of the same length as the regulatory element that the motif model represents. A computer program incrementally moves along a DNA sequence in 1 bp steps, returning a score at each step to indicate the strength with which each window of the DNA sequence matches the motif. The top scoring windows are the most similar to the consensus sequence of the TFBS bound by a TF.

As defined above, the most commonly used motif models for predicting the location of TFBSs are termed PSSMs. A single TF will recognize similar DNA sequences, but will tolerate variation from the consensus TF binding site pattern [41]. A PSSM is generated from a DNA sequence alignment of experimentally confirmed TFBSs for a TF (Figure 3). Such alignments are commonly generated using pattern discovery software such as MEME [42]. Once aligned, a matrix is created that reports the frequency of each nucleotide (A, C, G, and T) at each position of the alignment - the resulting matrix is called a position frequency matrix (PFM). The last step in obtaining a PSSM is to convert the PFM using a logarithmic function that weights the frequency of each nucleotide at each position by the frequency of that nucleotide in the genomic background (in many software implementations the default background frequency is set to 0.25 for each nucleotide) [43]. The widths of most published TFBS PSSMs fall in the range 8 to 14 bp. The scores produced are analogous to binding energies [44] and can thus be considered a prediction of the strength of association of a TF protein with a specific DNA sequence. Software for scanning DNA sequences using the matrix models is widely available through computer programming modules (TFBS [45]), downloadable software (RSAT [46]) or online websites (ORCAtk [39]). Active discussions are ongoing in the bioinformatics field about how the models can be improved in light of the increasing amount of TFBS data arising from ChIP-Seq studies [47].

PSSMs are available from open-resource databases such as JASPAR [48], PAZAR [49] or Uniprobe [50]. Until recently, the number of TFs with PSSMs has increased slowly, but high-throughput laboratory approaches for the profiling of TF-bound sequences have resulted in a striking increase in the number and quality of PSSMs [48,50]. Such high-throughput experimental data typically arise from either *in vivo* ChIP, such as ChIP-Seq [51], or from *in vitro* protein binding studies, such as protein binding microarrays [52]. In ChIP-Seq experiments, protein-DNA complexes are isolated using an antibody specific to the TF of interest, and the recovered DNA sequence is determined. For protein binding microarrays, double-stranded DNA of known sequence is affixed to the microarray surface and the adherence of a fluorescently labeled protein preparation of a TF (or frequently just the DNA-binding domain from a TF) is measured; the bound sequences are subsequently analyzed to determine the DNA sequence patterns targeted by the protein. Driven by these new technologies, the number of TFs with PSSMs in the open access JASPAR database has increased fivefold in the past year, rising from 100 to 500 PSSMs (about 25% of the 1,500 vertebrate TFs are now represented).

The prediction of functional regulatory elements by PSSMs, although having good sensitivity (most true positives are found), suffers from poor specificity (many false positives are predicted) [53]. With regard to specificity, a simple biochemical explanation of the problem is that the TFBS-predicting PSSMs determine sequences that a TF can bind *in vitro*, but *in vivo* the DNA may not be accessible to the TF. For instance, a predicted TFBS may be buried in compact chromatin. Thus, a prediction of a TFBS in isolation has limited relevance to the probability that a segment in the human genome will function as a *cis*-regulatory element. Approaches to reduce the specificity problems by filtering are discussed in the section below on ‘Refining *cis*-regulatory predictions with filters’.

The same concepts underlying the use of PSSMs to predict TFBSs apply to most motif discrimination methods for sequence-specific regulatory elements, ranging from splice enhancers to translation start sites [54-56]. We focus here principally on TFBSs, which are DNA-related; more information about RNA-related *cis*-regulatory elements can be found in a recent review [57].

Assessing the impact of sequence variations on TF-DNA interactions

Given a PSSM for a specific TF, and both the reference DNA sequence and the DNA sequence containing a variant, one can predict whether the variant alters the DNA sequence in a manner that strengthens or weakens the biochemical interaction of the TF with the DNA. The reference sequence and the variant sequence are both scanned and scored by the PSSM model. If the difference between observed scores is large, and at least one of the sequence isoforms is a known TFBS or is assigned a score that exceeds a user defined threshold for TFBS presence, the variation is predicted to have a functional impact. Such thresholds depend on the software used. The impact of the variant is calculated as the reported difference between the two scores. The higher-scoring allele is predicted to be bound by the TF with greater affinity. The calculation of PSSM score differences has two directions as variants have the potential to either knock out or create a TFBS. The action of the variant is captured by the sign (+/-) of the score difference in the above calculation.

Software for prediction of TFBS alterations by sequence variations includes the variant effect predictor, is-rSNP, and rSNP-MAPPER tools [58-60]. Although useful for identifying mutations overlapping known TFBSs, in the absence of additional information, such comparisons have limited value for predicted TFBSs (including all cases of *de novo* generation of TFBSs) [61,62], as the high rate of false TFBS predictions by PSSMs remains unaddressed. In the following section we outline additional data that may be incorporated to improve TFBS prediction specificity.

Similar allele comparison programs have been developed to predict altered microRNA target sites [63] and splicing elements [64].

Refining *cis*-regulatory predictions with filters

As stated above, predictions of TFBSs are unreliable because of a high false positive prediction rate (poor specificity). Predictions of *cis*-regulatory elements can be overlaid with genome annotations or experimental data to focus attention on the regions that are more likely to be functional [18,65]. An increase in specificity can be obtained by filtering predicted regulatory elements against complementary data, such as: (i) gene structure (topology filters); (ii) regions of sequence conservation (phylogenetic footprinting); (iii) TF-bound regions defined experimentally (such as ChIP-Seq for TFs); or (iv) structurally accessible (or inaccessible) regions (such as ChIP-Seq for epigenetic marks or DNase I hypersensitivity analyses). All the filters can be used individually or in combination, where it is functionally relevant; their main purpose is to add supporting evidence that a predicted regulatory element is functional. Although biologically relevant filters can dramatically increase the specificity of *cis*-regulatory element predictions, there may be a loss in sensitivity with the use of multiple filters, so it is recommended that a researcher assess results based on one filter before incorporating additional filters.

Topology filter

The activity of many *cis*-regulatory elements is spatially dependent (see Figure 1 for locations of *cis*-regulatory elements). For instance, splice-regulating elements are positioned adjacent to splice sites (reviewed in [57]) and the target sequences for non-coding RNA, such as microRNAs, may be preferentially situated within 3' untranslated regions [66]. Specific types of TFBSs within the core and proximal promoters, such as the TATA box and the downstream proximal element, are topologically constrained to occupy a specific location relative to the transcription start site (TSS) [67]. Genome annotations and laboratory data can specify TSS locations, allowing researchers to focus on variants situated with functionally relevant spatial localization. Existing annotations from high-throughput profiling of 5' capped RNA [68] and cDNA sequencing in genome annotation databases can delineate such regions. Increasingly, however, the annotation of exons is defined by RNA-Seq experiments applied to patient samples [69].

Each of these genomic data types can be retrieved as genomic positions from either a genome browser (for example, using the Galaxy tools [26] or Ensembl BioMart [70]) or from laboratory data, and should be chosen for their relevance to the type of regulatory element of interest. The positions of topological annotations can be compared with the positions of the predicted regulatory elements, using data analysis tools such as those that the Galaxy system provides. Where topological features are proximal to or overlap with corresponding variant-altered regulatory element predictions, the variants may have greater reliability than predictions lacking such support.

Conservation filters (phylogenetic footprinting)

Sequence conservation in the human genome can focus attention on regions with functional roles, a process termed phylogenetic footprinting. Using conservation scores based on multiple species alignments, such as the Phylogenetic P-values (PhyloP) [71] (obtainable using the Galaxy system or directly from the UCSC genome annotation database), researchers can restrict attention to regions more likely to have sequence-specific function. Although there is evidence of functional regulatory sequences being conserved over moderate periods of evolution [72], there is also ample evidence of plasticity in regulatory sequences [73]. Conservation-based filters can enrich for functional sequences, but, as for all filters, functionally relevant sequences without sequence conservation may be lost [65]. If the position of a predicted variant-altered regulatory element overlaps a conserved region, then the *cis*-regulatory potential of the variant is considered to have functional support.

TF binding filters

Increasing access to high-throughput profiles of ChIP data is key to improved regulatory sequence studies. In the ChIP method, a specific antibody targeting a protein of interest is used to recover DNA sequences bound by the protein [51]. The nucleotide sequence of the recovered DNA is increasingly being identified by high-throughput sequencing, resulting in the procedure known as ChIP-Seq. Regions containing a site bound by a targeted protein are identified in ChIP-Seq experiments as displaying a higher abundance of sequence reads recovered relative to a control set of data at a specific position in the genome. The method delivers two important advances for *cis*-regulatory element detection over past methodologies. First, it can be applied to detect TF or transcription co-activator bound regions across the entire genome of any species that has been sequenced [74]. Second, the results provide improved resolution of the boundaries for functional regulatory regions, providing the researcher with a refined search space for determining the active *cis*-regulatory element(s) in the region.

We focus here on two classes of ChIP-Seq experiments - those that profile interactions between a sequence-specific binding TF with DNA and those proteins that associate in a sequence-independent manner with regulatory regions (discussed in the next section).

With ChIP-Seq it is common to map the protein-DNA interactions to regions as small as about 300 bp. Although useful, the study of genetic variants requires more precise mappings of individual TFBSs. Thus, ChIP-Seq-defined regions are used as filters to refine the computational predictions generated with PSSMs. If no PSSM is available for the TF, the ChIP-Seq regions can be used in a motif discovery program (such as MEME [42]) to generate a PSSM that can be used in turn to predict TFBS positions. As with the previous filters, for each predicted TFBS-altering variant, those overlapping a ChIP-Seq-delineated region can be considered of sufficient reliability to motivate further laboratory studies.

Regulatory region accessibility filters

In addition to data pertaining to the sequence-specific binding of TFs, ChIP-Seq data can be obtained that delineate regions of a genome that are likely to contain elements involved in gene regulation. Such approaches may be based on antibodies that recognize specific epigenetic marks associated with *cis*-regulatory activity (for example, histone modifications), or antibodies that recognize proteins, such as co-activators, associated with regulatory sequences that interact with DNA-bound TFs. DNase I accessibility analysis likewise reveals regions with potential regulatory roles.

Studies focused on individual histone modifications have shown that certain marks, such as H3K4me3, associate with active promoter regions, whereas others, such as H3K27me3, are pronounced at silent promoters [75]. Combined with *cis*-regulatory predictions, combinations of epigenetic marks can be used to more precisely delineate regulatory regions with potential active roles. Focusing *cis*-regulatory element predictions proximal to or within epigenetic regions associated with active regulation improves the specificity of *cis*-regulatory element prediction [76,77].

The transcriptional co-activator p300, a component of many regulatory protein associations, has been targeted in ChIP-Seq studies to define transcriptional enhancers - genomic regions containing multiple TFBSs that collectively enhance transcription [78]. Visel *et al.* [78] took 86 regions associated with p300, tested them for regulatory activity *in vivo*, and found that 88% of the predicted regions were active regulatory regions. They found that using p300-predicted enhancer regions reduced the rate of false-positive predictions made by alternative methods by four-fold. Given the difficulty in obtaining high quality antibodies to proteins, and as the number of co-activators (about 10^1) is small relative to the number of sequence-specific TFs (about 10^3), it is likely that ChIP-Seq data for the complete set of co-activators will become a preferred means of delineating likely regulatory regions active in each cell type.

In both classes of ChIP-Seq experiments, the defined regulatory regions from the ChIP-Seq studies can be used to select the predicted regulatory elements and variants most likely to affect *cis*-regulatory function.

Examples of applied regulatory sequence variation prediction

Use of such filters as outlined above for the prediction of regulatory variants is starting to emerge in the literature. As interest rises with respect to the non-coding regulatory portions of the genome, we can expect to see more examples similar to the two we briefly outline below.

A key paper has recently emerged, highlighting the potential power behind combining SNP identification and different lines of regulatory evidence. Ernst *et al.* [79] combined non-coding disease-associated SNPs derived from multiple genome-wide association studies with epigenetic evidence for potential regulatory enhancer regions, and TFBS predictions. Non-coding SNPs were found to significantly overlap with enhancers predicted by epigenetic analyses, and the SNP-containing enhancers tended to be detected in cell types relevant to the disease. For instance, SNPs associated with systemic lupus erythematosus coincided with enhancer regions detected in lymphoblastoid cells. The authors further investigated these regulatory variant predictions by examining the potential of these disease-associated SNPs to interrupt or strengthen predicted TFBSs in the overlapping enhancer regions, such as an ETS1 binding motif in the aforementioned lupus example.

Oishi *et al.* [80] assessed multiple layers of regulatory evidence, such as topology, conservation, TFBS prediction, and TF binding, for regulatory variation prediction. They focused on the *KLF5* gene locus because of the potential role of its upstream regulatory programs in hypertension. Genotyping array experiments on 20 hypertensive individuals identified an associated SNP located upstream of a *KLF5* transcription initiation site. The SNP was observed to be situated at a position conserved between humans, mice, and rats. Bioinformatics analysis of TFBS motifs predicted an overlapping binding site for MEF2a, which was subsequently confirmed by ChIP analysis (in human aortic smooth muscle cells). Experimental analysis demonstrated that the SNP altered the MEF2a binding affinity.

These two examples highlight the effectiveness of using regulatory annotation data and predictions to focus attention on variants that might have an impact on gene regulation.

Case studies of filter-based support of regulatory variant predictions

We provide two examples of disease-associated variants and some data filters that, if these variants were unknown, could be overlaid to support a *cis*-regulatory variant prediction (Figures 4 and 5). These case reviews were in part chosen because they lend themselves to viewing in the UCSC Genome Browser. The first variant (Figure 4) is causal for hemophilia B Leyden and the second (Figure 5) is associated with protection against a disease; the latter is a case of a mutation creating a new regulatory element. We selected genome annotations from four tracks available on the UCSC Genome Browser: (i) distance from a TSS (topology) - RefSeq gene track; (ii) two databases for literature- derived TFBSs - PAZAR and ORegAnno tracks; (iii) mammalian conservation - PhyloP track; and (iv) known sequence variants - dbSNP track. The UCSC Genome Browser provides instructions for researchers wanting to add their own data tracks to a display, or wishing to extract data from tracks to apply to their own analysis.

A classic example of an inactivating mutation in a regulatory sequence is the disruption of a binding site for the TF HNF4A present in the promoter of the Factor IX endopeptidase gene (*F9*) involved in blood coagulation (Figure 4). The mutation is causal for hemophilia B Leyden [5]. A single nucleotide change of T to A at the most strongly conserved position of the HNF4A binding site results in reduced transcription of *F9*. The position of the variation relative to the regulatory element is displayed in the HNF4A PSSM sequence logo (Figure 4b). Figure 4c depicts a view of the region surrounding the

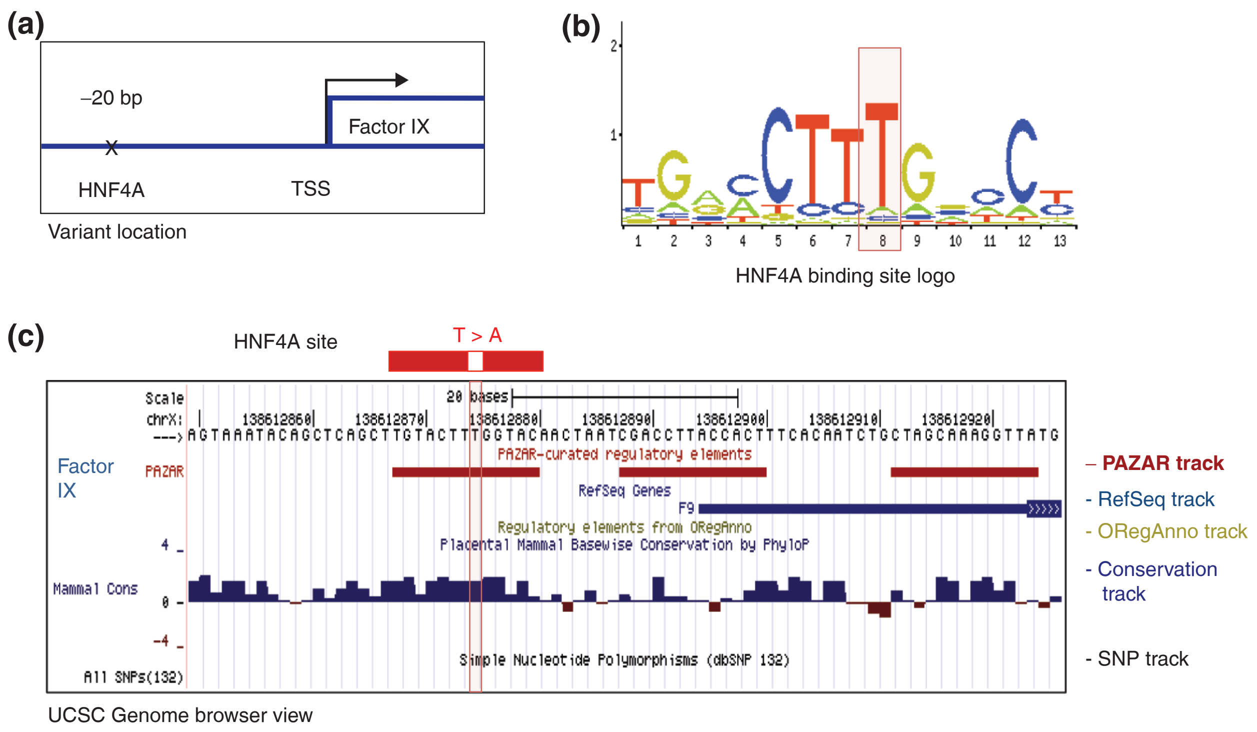

(a)
–20 bp
Factor IX
X
HNF4A
TSS
Variant location
(b)
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13
HNF4A binding site logo
(c)
HNF4A site
T > A
Scale
20 bases
chrX: 138612860 138612870 138612880 138612890 138612900 138612910 138612920
---> AGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTTCACAATCTGCTAGCAAAGGTTATG
PAZAR-curated regulatory elements
Factor IX
PAZAR
RefSeq Genes
F9
Regulatory elements from ORegAnno
Placental Mammal Basewise Conservation by PhyloP
4
Mammal Cons
0
-4
Simple Nucleotide Polymorphisms (dbSNP 132)
All SNPs(132)
– PAZAR track
- RefSeq track
- ORegAnno track
- Conservation track
- SNP track
UCSC Genome browser view

Figure 4. Visualization of a causal *cis*-regulatory mutation for hemophilia B Leyden that alters an HNF4A binding site in the promoter of the Factor IX gene. In this example, at least three sets of data support the hypothesis that the variant modifies a functional transcription factor binding site (TFBS): a TFBS database, proximity to a transcription start site (TSS; topology filter), and conservation data. **(a)** Variant location in base pairs between the regulatory element and the TSS. **(b)** Logo of nucleotide binding preferences for HNF4a. The mutated position in the HNF4A binding site is highlighted. The height of the letters 'T' and 'A' at that position indicate the frequency of a thymine or an adenine nucleotide in HNF4a TFBSs. The thymine at position 8 is strongly maintained among known TFBSs. **(c)** UCSC Genome Browser view highlighting four of the data types that may be used to support the prediction of a *cis*-regulatory variant. Red and gold tracks, PAZAR and ORegAnno databases of TFBSs; dark blue track, mammalian conservation; black track, data from dbSNP.

HNF4A binding site in the UCSC Genome Browser. The PAZAR track shows the location of binding sites in the region, one of which is for HNF4A, and the RefSeq track illustrates that the binding sites are proximal to the gene's TSS. The PhyloP mammalian conservation track reports sequence conservation at the HNF4a binding site, contributing support to this site as a functional regulatory element. For this region no SNP has been reported, as represented by the empty track for dbSNP at the bottom of Figure 4c. As dbSNP was originally intended as a database for sequence polymorphisms and has only recently started to acquire rare variants, the empty track signifies that this variant may not be common.

Activating mutations may create binding sites. A variation in the promoter of the *Alox15* gene creates a binding site for the TF SPI1 (Figure 5), which results in elevated expression of the gene. The SPI1 TFBS-creating variation has been linked to protection against atherosclerosis [81]. Comparing Figure 4 with Figure 5 we see that a TFBS common to the population (Figure 4, HNF4A) has various data annotations supporting its functional importance, whereas a TFBS that has been newly generated (Figure 5, SPI1) lacks most such annotations. The RefSeq track shows the variant to be within the promoter region of the *Alox15* gene. However, consistent with a variant creating a new regulatory element, no overlapping TFBS has been annotated in either of the PAZAR or ORegAnno databases, nor is there evidence of significant sequence conservation at the location of this element. The position of the variation is reported by the dbSNP track, even though it is not a polymorphism, which suggests this variant is not a technical error. However, in this instance it may be present owing to the dataset the variant was derived from being uploaded to dbSNP. In a case such as this, where a variant is predicted to have created a new regulatory element, the only useful data filters are likely to be topological, unless a researcher has access to ChIP-Seq data generated from the individual(s) carrying the variant.

With the conclusion of a workflow, such as that overviewed here, a researcher will possess a refined list of putative *cis*-regulatory variants. Determining the causality

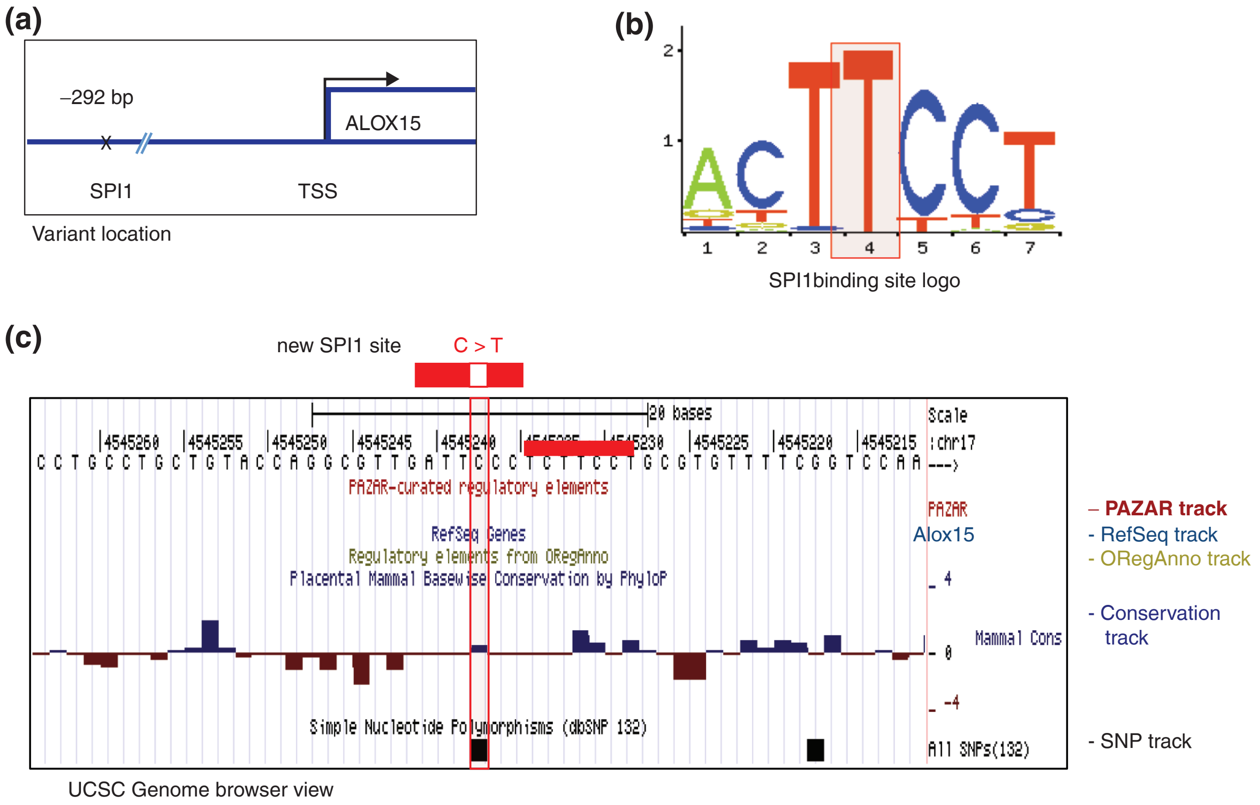


Figure 5. Visualization of a *cis*-regulatory mutation creating a functional SPI1 binding site in the promoter of *ALOX15*. In this example, the variant occurs near the transcription start site (TSS; topology filter), there is slight evidence that the site of the variant is conserved against mutations, and the variant is found in dbSNP; thus, it is unlikely to be a technical error. **(a)** Variant location. **(b)** Logo of nucleotide binding preferences for SPI1. The altered position of the *ALOX15* sequence relative to a SPI1 binding site is highlighted. The mutation converts the nucleotide at that position from a cysteine to the most strongly conserved nucleotide in all SPI1 binding sites - a thymine. **(c)** UCSC Genome Browser view as in Figure 4.

of a variant for a disease and the gene it regulates currently lies in the hands of experimental researchers.

The emerging challenge: associating variant-altered TFBSs with target genes

Unaddressed in the workflow reviewed up to this point is the challenge of defining the associations between potential *cis*-regulatory variants and target genes. Possibly as a result of DNA looping, regulatory sequences can act specifically on distant genes, skipping intervening genes in some cases [82,83]. Within the nucleus, DNA sequences that are not proximal in sequence may be brought into proximity by the three-dimensional looping of chromatin. Emerging methods that detect such DNA proximity (such as the Hi-C method described in [84,85]) may provide data suitable for integration into future bioinformatics methods for predicting the impact of a variant. Alternatively, methods are emerging that specify the edges of accessible DNA regions, features termed insulators. Such methods, largely based on ChIP-based experiments using antibodies to the insulator binding protein CTCF, could be used to determine which promoter regions are accessible to a TF bound between insulator sequences [86,87]. As these techniques are still maturing, bioinformatics approaches continue to associate regulatory elements and putative target genes based on distance measurements (the closest gene is the target), or through predictions arising from linkage disequilibrium or differential gene expression studies. However, we anticipate a time in the future when three-dimensional maps of nuclei can be generated experimentally.

Conclusions and future directions

At present the tools for the study of genome-wide regulatory sequence variations are limited, leading researchers to focus on variations predicted to alter genomic regions with well developed annotation - protein-coding sequences. This is due in part to the nature of the regulatory target and in part the availability of data. The *cis*-regulatory elements are short in length, widespread throughout the genome and are not confined to specific genomic landmarks - they can be both proximal and distal to their gene targets. Computational predictions of regulatory elements, in turn, are faced with extracting short and variable signal from a large genomic space, in which there is a mixture of functional elements and apparent randomly occurring non-functional sequences. Regulatory predictions are further complicated by the fact that the cellular environment and stage of development affects the functional activity of regulatory elements - an element that is active in one cellular context may not be active in another, an aspect important to the study of disease. However, from the current era of high-throughput technology, we can anticipate an increased understanding of the biological dynamics of *cis*-regulatory elements to feed into and improve computational algorithms predicting the locations of *cis*-regulatory elements. With improved predictions we will increase our ability to predict *cis*-regulatory-associated variants and their functional impact on the regulatory elements they coincide with.

The ability to look with increased resolution at the non-coding space of the genome has recently encouraged an increasing number of laboratories to investigate the impact of *cis*-regulatory-associated variants on disease, which as a result has motivated the development of bioinformatics tools for linking variants with regulatory elements. Bioinformaticians are still in the early stages of developing methods to integrate high-throughput regulatory data, such as ChIP-Seq and RNA-Seq, with regulatory element prediction, variant calling, and databases of known regulatory elements and variants. At the current time, researchers are best served by following a workflow such as that described here. However, a critical mass of interest in regulatory variants is being reached, and automated workflows will become publicly available in the near future.

The increased affordability of whole-genome sequencing has dramatically expanded the potential for studying *cis*-regulatory-related diseases in a familial context. The added power of having related genomes to study segregation of familial sequence variants with a phenotype dramatically improves the ability to predict disease-associated *cis*-regulatory variants. We anticipate that the next few years will see a rapid expansion of such family-associated studies.

Use of the aforementioned filters and tools in a workflow, as outlined here, coupled to the improved detection of causal variants provided by genome-wide data for multiple related individuals, provides medical genetic researchers with the means to prioritize the potential regulatory impact of a given a set of variants. In the future, integrated tools will consolidate the analysis process, bringing diverse analysis methods and data sources into a self-contained workbench for regulatory variation analysis.

Abbreviations

ChIP, chromatin immunoprecipitation; PFM, position frequency matrix; PSSM position-specific scoring matrix; SNP, single nucleotide polymorphism; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site; UCSC, University of California, Santa Cruz.

Acknowledgements

The work was supported by the National Institutes of Health (USA) grant 1R01GM084875, the Canadian Institutes for Health Research and the National Science and Engineering Research Council (NSERC). RWH was supported by fellowships from the Canadian Institutes of Health Research and the Michael Smith Foundation for Health Research.

Author details

[1]Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada. [2]Bioinformatics Graduate Program, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada.

Published: 10 October 2011

References

1. Gordon CT, Tan TY, Benko S, Fitzpatrick D, Lyonnet S, Farlie PG: Long-range regulation at the SOX9 locus in development and disease. *J Med Genet* 2009, **46**:649-656.
2. Wray GA: The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 2007, **8**:206-216.
3. Wang GS, Cooper TA: Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 2007, **8**:749-761.
4. Epstein DJ: Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic* 2009, **8**:310-316.
5. Reijnen MJ, Sladek FM, Bertina RM, Reitsma PH: Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci U S A* 1992, **89**:6300-6303.
6. Bosma PJ, Chowdhury JR, Bakker C, Gantla S, de Boer A, Oostra BA, Lindhout D, Tytgat GN, Jansen PL, Oude Elferink RP: The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N Engl J Med* 1995, **333**:1171-1175.
7. Ludlow LB, Schick BP, Budarf ML, Driscoll DA, Zackai EH, Cohen A, Konkle BA: Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. *J Biol Chem* 1996, **271**:22076-22080.
8. Zhu H, Tucker HM, Grear KE, Simpson JF, Manning AK, Cupples LA, Estus S: A common polymorphism decreases low-density lipoprotein receptor exon 12 splicing efficiency and associates with increased cholesterol. *Hum Mol Genet* 2007, **16**:1765-1772.
9. Kapeller J, Houghton LA, Mönnikes H, Walstab J, Möller D, Bönisch H, Burwinkel B, Autschbach F, Funke B, Lasitschka F, Gassler N, Fischer C, Whorwell PJ, Atkinson W, Fell C, Büchner KJ, Schmidtmann M, van der Voort I, Wisser AS, Berg T, Rappold G, Niesler B: First evidence for an association of a functional variant in the microRNA-510 target site of the serotonin receptor-type 3E gene with diarrhea predominant irritable bowel syndrome. *Hum Mol Genet* 2008, **17**:2967-2977.
10. de Vooght KM, van Wijk R, van Solinge WW: Management of gene promoter mutations in molecular diagnostics. *Clin Chem* 2009, **55**:698-708.
11. Vandermeer JE, Ahituv N: cis-regulatory mutations are a genetic cause of human limb malformations. *Dev Dyn* 2011, **240**:920-930.
12. Laurila K, Lahdesmaki H: Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. *In Silico Biol* 2009, **9**:209-224.
13. Wasserman NF, Aneas I, Nobrega MA: An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* 2010, **20**:1191-1197.
14. Murphy SM, Polke J, Manji H, Blake J, Reiniger L, Sweeney M, Houlden H, Brandner S, Reilly MM: A novel mutation in the nerve-specific 5'UTR of the GJB1 gene causes X-linked Charcot-Marie-Tooth disease. *J Peripher Nerv Syst* 2011, **16**:65-70.

15. Gallione CJ, Solatycki A, Awad IA, Weber JL, Marchuk DA: **A founder mutation in the Ashkenazi Jewish population affecting messenger RNA splicing of the CCM2 gene causes cerebral cavernous malformations.** *Genet Med* 2011, **13**:662-666.
16. Garone C, Pippucci T, Cordelli DM, Zuntini R, Castegnaro G, Marconi C, Graziano C, Marchiani V, Verrotti A, Seri M, Franzoni E: **FA2H-related disorders: a novel c.270+3A>T splice-site mutation leads to a complex neurodegenerative phenotype.** *Dev Med Child Neurol* 2011, **53**:958-961.
17. Luo X, Yang W, Ye DQ, Cui H, Zhang Y, Hirankarn N, Qian X, Tang Y, Lau YL, de Vries N, Tak PP, Tsao BP, Shen N: **A functional variant in microRNA-146a promoter modulates its expression and confers disease risk for systemic lupus erythematosus.** *PLoS Genet* 2011, **7**:e1002128.
18. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA: **9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response.** *Nature* 2011, **470**:264-268.
19. Rockman MV, Wray GA: **Abundant raw material for cis-regulatory evolution in humans.** *Mol Biol Evol* 2002, **19**:1991-2004.
20. Buckland PR, Hoogendoorn B, Coleman SL, Guy CA, Smith SK, O'Donovan MC: **Strong bias in the location of functional promoter polymorphisms.** *Hum Mutat* 2005, **26**:214-223.
21. van Wijk R, van Solinge WW, Nerlov C, Beutler E, Gelbart T, Rijksen G, Nielsen FC: **Disruption of a novel regulatory element in the erythroid-specific promoter of the human PKLR gene causes severe pyruvate kinase deficiency.** *Blood* 2003, **101**:1596-1602.
22. Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM, Noonan JP: **Human-specific gain of function in a developmental enhancer.** *Science* 2008, **321**:1346-1350.
23. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS: **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).** *Genome Res* 2006, **16**:123-131.
24. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4**:1073-1081.
25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248-249.
26. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.
27. Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA: **Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies.** *Mutat Res* 2008, **659**:147-157.
28. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11**:473-483.
29. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12**:443-451.
30. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, *et al.*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.
31. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, *et al.*: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467**:52-58.
32. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308-311.
33. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061-1073.
34. Ng SB, Nickerson DA, Bamshad MJ, Shendure J: **Massively parallel sequencing and rare disease.** *Hum Mol Genet* 2010, **19**:R119-R124.
35. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**:491-498.
36. Hedges DJ, Hedges D, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S: **Exome sequencing of a multigenerational human pedigree.** *PLoS One* 2009, **4**:e8232.
37. Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.** *Brief Bioinform* 2008, **9**:326-332.
38. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ: **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Res* 2008, **36**:D107-D113.
39. Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW: **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.** *Nucleic Acids Res* 2009, **37**:D54-D60.
40. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
41. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML: **Diversity and complexity in DNA recognition by transcription factors.** *Science* 2009, **324**:1720-1723.
42. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets.** *Bioinformatics* 2011, **27**:1696-1697.
43. King OD, Roth FP: **A non-parametric model for transcription factor binding sites.** *Nucleic Acids Res* 2003, **31**:e116.
44. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors.** *Science* 2007, **315**:233-237.
45. Lenhard B, Wasserman WW: **TFBS: Computational framework for transcription factor binding site analysis.** *Bioinformatics* 2002, **18**:1135-1136.
46. Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohee S, van Helden J: **RSAT: regulatory sequence analysis tools.** *Nucleic Acids Res* 2008, **36**:W119-W127.
47. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity.** *Nat Biotechnol* 2011, **29**:480-483.
48. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2010, **38**:D105-D110.
49. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: **PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation.** *Genome Biol* 2007, **8**:R207.
50. Robasky K, Bulyk ML: **UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions.** *Nucleic Acids Res* 2011, **39**:D124-D128.
51. Orlando V: **Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation.** *Trends Biochem Sci* 2000, **25**:99-104.
52. Berger MF, Bulyk ML: **Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors.** *Nat Protoc* 2009, **4**:393-411.
53. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**:276-287.
54. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 2002, **3**:285-298.
55. Chasin LA: **Searching for splicing motifs.** *Adv Exp Med Biol* 2007, **623**:85-106.
56. Sparks ME, Brendel V: **MetWAMer: eukaryotic translation initiation site**

prediction. *BMC Bioinformatics* 2008, **9**:381.

57. Wang Z, Burge CB: Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 2008, **14**:802-813.
58. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, *et al.*: Ensembl 2011. *Nucleic Acids Res* 2011, **39**:D800-D806.
59. Macintyre G, Bailey J, Haviv I, Kowalczyk A: is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* 2010, **26**:i524-530.
60. Mapper 2 - Multi-genome analysis of positions and patterns of elements of regulation [http://genome.ufl.edu/mapper]
61. Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJ: A survey of genomic properties for the detection of regulatory polymorphisms. *PLoS Comput Biol* 2007, **3**:e106.
62. Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J: In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* 2008, **4**:e5.
63. Barenboim M, Zoltick BJ, Guo Y, Weinberger DR: MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum Mutat* 2010, **31**:1223-1232.
64. Churbanov A, Vorechovsky I, Hicks C: A method of predicting changes in human gene splicing induced by genetic variants in context of cis-acting elements. *BMC Bioinformatics* 2010, **11**:22.
65. King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, Chiaromonte F, Miller W, Hardison RC: Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* 2007, **17**:775-786.
66. Sevignani C, Calin GA, Siracusa LD, Croce CM: Mammalian microRNAs: a small world for fine-tuning gene expression. *Mamm Genome* 2006, **17**:189-202.
67. Bernard V, Brunaud V, Lecharny A: TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics* 2010, **11**:166.
68. Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J: A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* 2010, **7**:521-527.
69. Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert JJ, Goldstein DB: Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol* 2010, **11**:R57.
70. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: BioMart Central Portal--unified access to biological data. *Nucleic Acids Res* 2009, **37**:W23-W27.
71. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010, **20**:110-121.
72. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003, **2**:13.
73. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT: Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 2010, **328**:1036-1040.
74. Park PJ: ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009, **10**:669-680.
75. Zhou VW, Goren A, Bernstein BE: Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 2011, **12**:7-18.
76. Whitington T, Perkins AC, Bailey TL: High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res* 2009, **37**:14-25.
77. Ramsey SA, Knijnenburg TA, Kennedy KA, Zak DE, Gilchrist M, Gold ES, Johnson CD, Lampano AE, Litvak V, Navarro G, Stolyar T, Aderem A, Shmulevich I: Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* 2010, **26**:2071-2075.
78. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009, **457**:854-858.
79. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011, **473**:43-49.
80. Oishi Y, Manabe I, Imai Y, Hara K, Horikoshi M, Fujiu K, Tanaka T, Aizawa T, Kadowaki T, Nagai R: Regulatory polymorphism in transcription factor KLF5 at the MEF2 element alters the response to angiotensin II and is associated with human hypertension. *FASEB J* 2010, **24**:1780-1788.
81. Wittwer J, Bayer M, Mosandl A, Muntwyler J, Hersberger M: The c.-292C>T promoter polymorphism increases reticulocyte-type 15-lipoxygenase-1 activity and could be atheroprotective. *Clin Chem Lab Med* 2007, **45**:487-492.
82. Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, Lenhard B: Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res* 2010, **38**:1071-1085.
83. Dean A: In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics* 2011, **10**:3-10.
84. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009, **326**:289-293.
85. Vassetzky Y, Gavrilov A, Eivazova E, Priozhkova I, Lipinski M, Razin S: Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification. *Methods Mol Biol* 2009, **567**:171-188.
86. Gaszner M, Felsenfeld G: Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 2006, **7**:703-713.
87. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K: Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 2009, **19**:24-32.

doi:10.1186/gm281
Cite this article as: Worsley-Hunt R, *et al.*: Identification of *cis*-regulatory sequence variations in individual genome sequences. *Genome Medicine* 2011, **3**:65.