Topic Introduction

Identification and Computational Analysis of Gene Regulatory Elements

Leila Taher,^{1,2} Leelavati Narlikar,³ and Ivan Ovcharenko^{1,4}

¹Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; ²Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, 18051 Rostock, Germany; ³Chemical Engineering and Process Development Division, National Chemical Laboratory, CSIR, Pune 411008, India

Over the last two decades, advances in experimental and computational technologies have greatly facilitated genomic research. Next-generation sequencing technologies have made de novo sequencing of large genomes affordable, and powerful computational approaches have enabled accurate annotations of genomic DNA sequences. Charting functional regions in genomes must account for not only the coding sequences, but also noncoding RNAs, repetitive elements, chromatin states, epigenetic modifications, and gene regulatory elements. A mix of comparative genomics, high-throughput biological experiments, and machine learning approaches has played a major role in this truly global effort. Here we describe some of these approaches and provide an account of our current understanding of the complex landscape of the human genome. We also present overviews of different publicly available, large-scale experimental data sets and computational tools, which we hope will prove beneficial for researchers working with large and complex genomes.

THE REGULATORY LANDSCAPE OF THE HUMAN GENOME

The Architecture of the Genome

As is the case for most eukaryotic genomes, the human genome has a small ratio of protein-coding to nonprotein coding (or, simply, noncoding) DNA. Indeed, the proportion of the human genome that encodes proteins accounts for only ~1.5% of its DNA (International Human Genome Sequencing Consortium 2004). The remaining portion of the genome includes introns, noncoding RNA genes, pseudogenes, regulatory elements, and repetitive DNA, and has largely unknown function (Lander et al. 2001). The identification of functional noncoding elements has proved to be nontrivial. Cross-species genomic comparisons have revealed a large number of noncoding sequences under evolutionary constraint, and conservative estimates suggest that ~3.5% of the noncoding genome is indeed under purifying natural selection, and thus, probably is functional (Waterston et al. 2002; Boffelli et al. 2004; Lindblad-Toh et al. 2011). Detecting and interpreting these elements is particularly relevant to human medicine, since many of the common diseases show complex inheritance patterns and are likely associated with changes in gene expression, rather than in the genes themselves (Altshuler et al. 2008).

⁴Correspondence: ovcharen@nih.gov

Protein-Coding Genes

There are about 25,000 protein-coding genes in the human genome (International Human Genome Sequencing Consortium 2004). Most protein-coding genes are organized into multiple exons, which are spliced to produce mRNAs that are subsequently translated into protein molecules. A typical human coding gene contains eight exons and (including introns) is 54-kb long (Hsu et al. 2006; Fujita et al. 2011). However, the size of protein-coding genes varies considerably, from a few hundred base pairs in the case of genes encoding histones to several millions of base pairs, in the case of the dystrophin gene (*DMD*), which is altered in muscular dystrophy patients (Blake et al. 2002). Genes are nonrandomly distributed in the genome, both within and between chromosomes, and gene density is strongly correlated with the GC content (Bernardi 2000). Moreover, the human genome can be separated into two gene spaces, each space containing approximately half of all genes: a GC-rich, gene-rich space (12% of the genome), and a GC-poor, gene-poor space (88% of the genome). The two gene spaces are characterized by distinct structural and functional features that are associated with different levels of transcription (Bernardi 2005).

Transcriptional control of gene expression is mediated through the interaction between *trans*-acting transcription factor proteins and *cis*-regulatory sequence elements. Thus, most genes possess a promoter element, immediately upstream of their transcription start site, which recruits the preinitiation complex (PIC), including the RNA polymerase II, and is, in most cases, sufficient to drive basal expression. Additionally, single or multiple distal *cis*-regulatory elements are responsible for precise spatiotemporal expression control (Levine and Tjian 2003) (see Fig. 1).

Cis-Regulatory Elements

Promoter Elements

Promoter sequences comprise both core and proximal elements. In eukaryotes, the core promoter represents the minimal set of elements required to initiate transcription. The average core promoter sequence encompasses ~100 bp and contains the transcription start site. This core region interacts directly with the complete set of general transcription factors and RNA polymerase known collectively as the preinitiation complex (PIC) (Roeder 1996). RNA polymerase II catalyzes RNA synthesis, but—on its own—is unable to recognize the transcription start site or to “melt” or open the

FIGURE 1. Schematic representation of a typical gene locus. The promoter comprises a core promoter and a proximal promoter region, and contains binding sites for the general transcription machinery as well as specific transcription factors. Distal enhancers and silencers also provide binding sites for transcription factors, and activate or prevent the initiation of the transcription of the regulated genes. Distal enhancers and silencers would interact with the promoter through a mechanism that involves looping out the intervening DNA. Insulators act as boundaries, insuring that only the appropriate gene is transcribed.

DNA, and therefore cannot initiate transcription. These functions are accomplished by general transcription factors, typically including TFIIA, TFIIB, TFIID, TFIIIE, TFIIF, and TFIIF (Lee and Young 2000). The assembly of the PIC usually starts with the binding of TFIIA, TFIIB, and TFIID to sequence elements in the core promoter including the TATA box, BRE (TFIIB-recognition element), Inr (initiator element), and DPE (downstream promoter element). Human core promoters are surprisingly diverse, and although most promoters contain one or more of these elements, none seems to be essential for promoter function (Smale and Kadonaga 2003). Other factors are then recruited to the complex and assembled with RNA polymerase II (see Fig. 1). Thus, TFIIF binds RNA polymerase II and can assist in recruiting it to the promoter region. TFIIIE also interacts with RNA polymerase and can stimulate the kinase and helicase activities of TFIIF, which in turn phosphorylates the carboxy-terminal domain of RNA polymerase II and melts the DNA in the promoter region. TFIIIE and TFIIF are partially dispensable, but have an impact on the efficiency of transcription initiation (Hahn 2004). In addition to the core promoter, human genes often contain a proximal promoter element that spans up to 1000 bp, and comprises binding sites for additional specific transcription factors. Because the expression of these transcription factors is cell-specific rather than ubiquitous, proximal promoters, in conjunction with distal regulatory elements, determine the expression program of the target gene.

Distal Elements

A single genome directs the process of cell differentiation during development that leads to hundreds of distinct cell types present in the adult human body. This morphological and developmental complexity is not a function of the number of genes in the genome, but rather of the number of possible spatial and temporal patterns in which genes are expressed within the organism. The decoupling of transcriptional regulation from the gene promoter region and distributing the regulatory controls among multiple distal regulatory elements provide an exquisite level of precision and robustness for gene expression control.

Known distal regulatory elements have been classified on the basis of their behavior in synthetic assays. Enhancers and silencers mediate positive and negative regulation of transcription, respectively. They are position- and orientation-independent. Multiple enhancers and silencers often act additively, synergistically, or competitively (Buttgereit 1993; Lin et al. 2007; Perry et al. 2011). Similar to proximal promoters, enhancers and silencers often consist of clusters of binding sites for transcription factors; the particular nature, number, and spatial arrangement of these binding sites, together with the availability of the cognate transcription factors, establish the unique pattern of expression of a gene. Although not well understood, distal transcriptional regulation involves long-range interactions between the transcription factors bound to enhancers and silencers, and promoters, with concomitant looping of the intervening DNA (see Fig. 1). Such interactions may be established either by simple diffusion within the nucleus or by an active “tracking” mechanism, in which the enhancers and silencers migrate along the chromatin fiber until they encounter a promoter. Specific transcription factors bound to enhancers, silencers, and promoters would then stimulate the assembly of a functional PIC on the promoter (Bulger and Groudine 2011). Because the assays are well established, most distal regulatory elements studied to date are enhancers.

A different class of distal regulatory elements, known as insulators (boundary elements), establish discrete transcriptional domains. Insulators have been shown to possess either one or both of the two activities, blocking enhancers (enhancer-blocking insulators) and/or protecting against heterochromatin spreading (barrier insulators) (Gaszner and Felsenfeld 2006). The most extensively characterized vertebrate insulator—5' HS4 (DNase I hypersensitivity site 4) chicken β -globin—was first identified at the 5' end of the chicken β -globin locus, and has both barrier and enhancer-blocking activities (Chung et al. 1993). Most vertebrate enhancer-blocking insulators, including 5' HS4 chicken β -globin, contain binding sites for the transcription factor CTCF, a zinc finger protein with multiple roles, including transcription activation and repression (Bell et al. 1999; Yusufzai and Felsenfeld 2004). CTCF has also been implicated in the function of barrier insulators, but it has been suggested

that additional proteins may be required for specificity (Cuddapah et al. 2009). In the case of HS4, binding sites for CTCF are neither necessary nor sufficient for protection against adjacent heterochromatin or repressive domain-mediated effects (Recillas-Targa et al. 2002).

Regulatory elements may be either discrete or clustered to create locus control regions (LCRs) that mediate complex transcriptional programs involving several genes within a genomic locus. LCRs were first discovered in the human β -globin locus, which comprises five β -type globin genes arranged in the order of their developmental stage-specific expression in erythroid cells (Stamatoyannopoulos 1991). The switches of human globin gene expression from embryonic (ϵ) in the yolk sac, to fetal (γ) during intrauterine life, and to adult (β) after birth, are controlled by regulatory sequences located both proximal and distal to the genes. The most prominent distal regulatory element is a LCR that has been shown to be required for high-level globin gene expression at all developmental stages (Grosveld et al. 1987). In addition to integrating the instructions from multiple regulatory elements, the human β -globin LCR modulates the overall chromatin architecture (Liang et al. 2008). LCRs have now been described for a broad spectrum of human and mammalian gene systems, suggesting that they play an important role in the control of gene expression (Li et al. 2002).

Distal regulatory elements have been identified within introns or within several hundred base pairs upstream or downstream from the transcription start site of their target genes, but may also reside as distant as several millions of base pairs away, within the introns of neighboring genes with unrelated functions, or even on different chromosomes (Visel et al. 2009a; Williams et al. 2010). Together with their heterogeneous sequence properties, this flexibility makes the comprehensive identification of distal regulatory elements in the human genome an extremely challenging task.

Other Noncoding DNA

Noncoding RNA Genes

Even though only $\sim 1.5\%$ of the human genome encodes proteins, a substantially larger fraction appears to be transcribed into noncoding RNA (Bertone et al. 2004; Johnson et al. 2005; Birney et al. 2007; Kapranov et al. 2007). Noncoding RNA genes (ncRNA), transcribed from the genome by one of the three nuclear DNA-dependent RNA polymerases, produce abundant and important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as the highly heterogeneous small nucleolar RNA (snoRNA), involved in RNA processing. These genes also encode microRNA (miRNA), small interfering RNA (siRNA), and piwi-interacting RNA (piRNA), responsible for gene silencing. RNA gene products are involved in processes as diverse as splicing, messenger RNA (mRNA) turnover, gene silencing, and translation, and often elicit their biological responses by base pairing with their targets transcripts (Storz et al. 2005). Some ncRNAs may have important roles in transcriptional regulation. For example, a recent study uncovered $\sim 12,000$ enhancers in the mouse genome that are transcribed into RNA (eRNAs) in levels correlated with mRNA synthesis from nearby genes (Kim et al. 2010). Thus, eRNAs would serve not only as a robust indicator of enhancer activity, but would alter transcriptional programs by activating promoters, facilitating the adoption of an open chromatin structure, or even binding to other enhancers (Ling et al. 2004; Ong and Corces 2011; Wang et al. 2011). A second class of ncRNAs showing transcriptional activity was identified in the late 1990s (Brannan et al. 1990; Pfeifer et al. 1996). The means by which such long (>200 nucleotides) noncoding RNAs (lncRNAs) regulate transcription encompass diverse mechanisms, such as cleavage into shorter RNAs, specific binding to chromatin, and mediation of epigenetic control (Carninci 2008; Mercer et al. 2009). The quantity and nature of lncRNAs encoded within the human genome remains unclear; an increasing number of transcriptomic and bioinformatic studies suggest the existence of thousands of these transcripts (Guttman et al. 2009; Marques and Ponting 2009; Hung et al. 2011).

Repeats

Repetitive DNA makes up $>45\%$ of the human genome, although this fraction is expected to increase as sequencing methods improve. The various types of repetitive DNA can be broadly separated into

tandem and interspersed repetitive DNA, based on their distribution within the genome. Tandem repeats are composed of short sequences that are repeated multiple times, immediately adjacent to each other. These repeats are further classified as microsatellites, minisatellites, or satellites according to their length. Microsatellites are randomly distributed in the genome. Because they are present in highly variable copy numbers among different individuals of the same species (including humans), microsatellite DNAs are widely used for genetic mapping and population genetics studies (Dib et al. 1996). Minisatellites are also very variable in size, but generally longer than microsatellites. Minisatellite sequences are mainly confined to subtelomeric and telomeric regions of the chromosome, where they are believed to play a role in DNA replication (Fajkus et al. 2005). Satellites are usually organized as large clusters at telomeres, and, particularly, at centromeres, where they may play a structural role (Schueler and Sullivan 2006).

Tandem repeats are thought to have originated by the expansion of a template sequence, either by “replication slippage” or by recombination (Charlesworth et al. 1994). In contrast, most interspersed repeats have arisen by transposition. Transposable elements in the genome can be separated into two main classes, DNA transposons and retrotransposons. DNA transposons constitute ~3% of the human genome, can excise themselves from the genome, move as DNA, and insert themselves into new genomic sites. Retrotransposons comprise ~40% of the human genome, and duplicate via RNA intermediates that are reverse transcribed and inserted at new genomic locations. The best documented example of interspersed repeats in the human genome is the class of Alu retrotransposons, which cover ~10% of the genome (Batzner and Deininger 2002). Other repeats include segmental and whole-genome duplications, and processed pseudogenes. Although repetitive DNA is often regarded as genomic “junk”, the biological significance of repeats was recognized early. Indeed, Barbara McClintock, who discovered transposons in the 1940s, observed that they affected the expression of neighboring genes and suggested that they might play a regulatory role (McClintock 1956). Several studies in the last years have associated different human diseases with mutations in noncoding repetitive DNA, supporting this initial hypothesis (Boby et al. 2005; Sha et al. 2009; Hagerman et al. 2010; Renton et al. 2011). The association between repetitive DNA and human disease is likely to extend far beyond these few known cases.

The Role of Mutations in Noncoding Regulatory Regions

Mutations in *cis*-regulatory elements do not disrupt the amino acid sequence of genes, do not create alternative transcripts, do not introduce premature stop codons, and do not affect the three-dimensional structure of proteins (with the exception of splice enhancers and similar cases). Instead, noncoding regulatory polymorphisms modulate the dynamics of gene expression, including modifications to regulatory pathways, changing the level of gene expression, impacting the speed of transcription, and/or producing other changes that result in remodeling of the chromatin state and architecture. The role of regulatory mutations as the key driving force behind the evolution of species and origin of many phenotypic differences was advocated in a pioneering work of King and Wilson in 1975 (King and Wilson 1975). Since then, many regulatory mutations leading to phenotypic human variants have been reported. For example, intronic mutations in the gene responsible for oculocutaneous albinism type II (*OCA2*) have been reported to explain most of the variation in human eye-color (Duffy et al. 2007). Additionally, a regulatory mutation located 21 kb upstream of *OCA2* is perfectly associated with blue and brown eye colors in a large population from Denmark (Eiberg et al. 2008). These and other similar, parallel regulatory mutations have been part of a “fine tuning genome toolkit” during vertebrate evolution. One of the adaptive traits in some human populations is the ability to digest milk in adulthood. This trait has become independently fixed in African and European populations since the domestication of cattle in the early Neolithic Era through a genetic process called convergent evolution (Tishkoff et al. 2007). Lactose intolerance varies drastically among different human populations (characteristic to ~10% of Americans of northern European descent, 10% of Africa’s Tutsi tribe, 50% of French, and 99% of Chinese, for example), and relates to whether the gene *LCT* that encodes the protein lactase-phlorizin hydrolase (LPH) continues

to be expressed in adulthood or not. Expression of *LCT* in adulthood depends on a regulatory mutation located ~20 kb upstream of *LCT* (Wang et al. 1995; Enattah et al. 2002).

In addition to being associated with phenotypic diversity between and within species, regulatory mutations can also lead either to a disease or to increased susceptibility to a disease. For example, a recent study revealed a strong association between Hirschsprung disease risk and the presence of a regulatory mutation located in the first intron of the *RET* proto-oncogene (Emison et al. 2005). *RET* was previously linked with Hirschsprung disease (Puliti et al. 1993) and the development of multiple types of cancer (Donahue and Hines 2009; Pacini et al. 2010). This particular mutation, however, has a 20-fold greater contribution to risk than do rare alleles, and the authors of the study showed that the acquired mutation is capable of significantly decreasing the level of *RET* expression. Another classical example of a disease causal regulatory mutation is a single nucleotide polymorphism (SNP) in an enhancer of Sonic hedgehog homolog (*SHH*). The enhancer mutation is associated with preaxial polydactyly, a frequently observed congenital limb malformation (Masuya et al. 1995). This *SHH* enhancer is located 1 Mb away from the gene, within an intron of another gene (*LMBR1*) (Lettice et al. 2003). Its genomic location and distance from the affected gene reflect the complexity of the gene regulatory landscape in the human genome.

With the development of new sequencing technologies and SNP arrays capable of simultaneously genotyping hundreds of thousands of SNPs, genome-wide association studies (GWAS) became a reality and a practical tool for identifying genotypes specific to a population with a particular disease (Mathew 2008). In 2008, a group of manual curators at the National Human Genome Research Institute (NHGRI) started assembling published GWAS results in a systematic manner (Hindorff et al. 2009). At the time of the original publication in 2009, the NHGRI GWAS catalog covered 80 traits and diseases. In November 2014, the catalog included more than 2000 curated publications and more than 14,000 SNPs (www.genome.gov/gwastudies). Only 4% of the trait- and disease-associated SNPs were identified in the coding regions and those potentially disrupt or modify the protein sequence encoded by the host gene. It is likely that some of the remaining noncoding SNPs do not represent causal variants, but are rather associated with unknown coding mutations in the same haplotype block with the identified noncoding SNPs. On the other hand, more than 7000 (50% of total) noncoding SNPs are located at least 10 kb from the nearest transcription start site, and thus, are likely to correspond to mutations in distant regulatory elements that disrupt a *cis*-regulatory mechanism. For example, an unusually long intergenic interval on human chromosome 8q24 hosts several variants associated with prostate, breast, and colorectal cancers (Amundadottir et al. 2006; Easton et al. 2007; Tomlinson et al. 2007). In a follow up study of this region, multiple enhancers were mapped to this intergenic interval, one of which contains the prostate cancer-associated variant rs6983267 (Wasserman et al. 2010). This study also reported that the cancer-associated variant significantly increases the enhancer activity during early prostate development and throughout prostate maturation, highlighting rs6983267 as a key regulatory mutation contributing to prostate cancer and disease susceptibility.

To directly assess the impact that regulatory mutations have on the level of gene expression, expression levels were measured in 210 unrelated individuals and the results compared to their SNP and copy-number variant (CNV) genome-wide profiles (Stranger et al. 2007). Noncoding SNPs accounted for >80% of the observed variation in expression and at least one strong SNP–gene expression association was recorded for 888 nonredundant genes. The majority of causal SNPs were in *cis* with the affected gene. These SNPs were located mainly outside of the proximal promoter region, arguing yet again for the importance of distant regulatory mutations in phenotypic diversity and individual disease susceptibility in the human population.

THE EVOLUTION OF THE HUMAN GENOME

Functional Sequences Are Generally Conserved

Genetic mutation is a crucial process in evolution that provides raw material for natural selection. Mutations that impact the function of a gene or a regulatory element are likely to have a negative effect

on fitness and will thus be under negative selection. In contrast, mutations at nonfunctional sites are free to be fixed by drift. Together with the local mutation rate, the balance of selection and drift determines the rate of evolution. Consequently, nonfunctional sequences accumulate neutral substitutions and become less similar between different species as the evolutionary distance between them increases, whereas sequences whose function is maintained diverge at a much slower rate (Kimura 1968; King and Jukes 1969). That is, sequences conserved over large evolutionary distances are more likely to be functional than those conserved over lesser distances (Boffelli et al. 2004). Thus, comparisons of orthologous sequences across multiple species have provided a wealth of information enabling gene finding (e.g., van Baren et al. 2007), prediction of gene function (Huynen et al. 2004), identification of protein domains (e.g., Geer et al. 2002; Marchler-Bauer and Bryant 2004; Marchler-Bauer et al. 2007; Marchler-Bauer et al. 2011), prediction of functional amino acid residues (Lichtarge and Sowa 2002), and detection of natural selection at the level of genes (O'Neill et al. 2007). Regulatory elements are subject to the same process of molecular evolution as is the rest of the genome. However, regulatory mutations are more likely to alter only gene expression levels, thereby affecting the functioning of the organism in a subtle manner, as compared to the more obvious effects resulting from mutations within genes. Only mutations that disrupt sequences essential for regulatory function, such as transcription factor binding sites, could be deleterious (Wittkopp and Kalay 2011). As a result, regulatory sequences are hypothesized to evolve, in general, faster than genes (Borneman et al. 2007).

DNA sequence comparison, known as comparative genomics, have identified hundreds of thousands of noncoding sequences conserved in the human genome (Thomas et al. 2003; Ureta-Vidal et al. 2003; Boffelli et al. 2004; Cooper et al. 2005; Dermitzakis et al. 2005; Margulies et al. 2007). Estimates for the total extent of noncoding constraint are roughly twice that for coding constraint (e.g., Waterston et al. 2002). Conserved noncoding sequences (CNS) are nonuniformly distributed. These sequences tend to be found in gene-poor regions, in locations that are consistent with the long-distance or distance-independent interactions between distal regulatory elements and genes (Boffelli et al. 2004; Ovcharenko et al. 2005b). Furthermore, CNS in the human genome are found to be overall enriched in the neighborhood of transcription factors and developmental (“trans-dev”) genes (Bejerano et al. 2004; Woolfe et al. 2005). However, subsets of CNS that came under selection at distinct evolutionary times are associated with distinct functions. For example, CNS exclusive to placental mammals are actually enriched in regions near genes involved in posttranslational protein modification, including those in intracellular signaling pathways (Lowe et al. 2011). With the increasing number and availability of high-quality genomes, the methods used to define noncoding sequence conservation have become increasingly sophisticated (Cooper et al. 2005; Siepel et al. 2005; Prabhakar et al. 2006; Lindblad-Toh et al. 2011).

Comparative Genomics Uncovers Regulatory Elements

It has long been understood that sequence conservation is a reliable strategy to infer regulatory functions in noncoding DNA. One of the first analyses that showed the reliability of purely comparative genomics approaches to identify regulatory elements of human genes involved the search for highly conserved noncoding sequences between the human and the mouse within a 1 Mb locus on human chromosome 5. (“Highly conserved” sequences display at least 70% sequence identity >100 bp.) This locus comprises three cytokine genes (*IL4*, *IL13*, and *IL5*), which typically show coordinated expression (Kelly and Locksley 2000), as well as ~90 CNS, most of which are conserved across other vertebrates. The activity of the longest CNS, CNS-1, was assayed in transgenic and knockout mouse assays. The assay results showed that this element, separated by >120 kb of sequence from the cytokine genes, modulates the expression of all three genes (Loots et al. 2000). Soon afterward, methods such as Genome Evolutionary Rate Profiling (GERP) and phylogenetic shadowing (Boffelli et al. 2003; Cooper et al. 2005) were developed to compare the genomes of more closely related species, such as primates. These methods search, not for regions that are similar, but rather for regions that are significantly depleted in mutations. Such analyses have been proven successful for discovering lineage-specific

regulatory elements (Prabhakar et al. 2006; Wang et al. 2007). These results provide evidence for the role of CNS as transcriptional regulators. The increasing availability of sequenced genomes has provided the scientific community with an invaluable resource for comparative genomics to understand the molecular and genetic basis of many complex traits and diseases.

Deeply Conserved Sequences Act as Enhancers In Vivo

Further studies confirmed the validity of using comparative genomics for predicting regulatory elements on a genome-wide scale. In particular, Pennacchio et al. showed that almost half (45%) of 167 noncoding elements conserved from human to pufferfish or ultraconserved (Bejerano et al. 2004) in human, mouse, and rat function as tissue-specific enhancers at embryonic day 11.5 in transgenic mice. Interestingly, the majority of the enhancers assayed were active in the central nervous system (Pennacchio et al. 2006). Assayed sequences are available through the VISTA Enhancer Browser (<http://enhancer.lbl.gov>), a growing database that currently contains information on 2164 in vivo tested elements and represents a valuable resource for the scientific community (Visel et al. 2007).

Conserved Gene Deserts Harbor Regulatory Elements

Approximately 25% of the human genome comprises long regions with no protein-coding genes and no evident function; these regions are known as gene deserts (Lander et al. 2001; Venter et al. 2001). While some of these gene deserts can be deleted from the mouse genome without any detectable phenotypic difference and only minor alterations in gene expression (Russell et al. 1982; Nobrega et al. 2004), some of these regions have been shown to contain distal regulatory sequences that control the transcription of neighboring genes (Bishop et al. 2000; Lettice et al. 2003; Nobrega et al. 2003; Kimura-Yoshida et al. 2004; Uchikawa et al. 2004). However, most gene deserts conserved across multiple vertebrates appear to constitute an integral unit with their flanking genes, and show characteristics suggesting regulatory function (Ovcharenko et al. 2005b). On these grounds, a recent study was performed to explore a possible link between the genetic susceptibility to coronary artery disease (CAD) and response to inflammatory signaling. The study focused on a 200-kb gene desert, comprising several SNPs associated with CAD and type 2 diabetes, that was also enriched in different enhancer signatures.

Sequence Conservation Does Not Imply Functional Constraint

Comparative genomics has proven effective in identifying both known and novel regulatory elements in the human genome. However, the degree of functional constraint is not directly correlated with sequence conservation, and sequence conservation is only an estimator of function. Thus, not all conserved regions are involved in gene regulation. Targeted deletion of extremely conserved sequences does not necessarily produce the expected phenotype (Ahituv et al. 2007), indicating that the regulatory activity of the deleted sequences might be showed in genes other than the neighboring genes; or that there might be functional redundancy with other enhancers; or the assay used might not be suitable to detect the corresponding phenotypes. Moreover, conserved regions may have functionally diverged, and regulatory sequences that are conserved across species often display different activities (Borneman et al. 2007). Conversely, constrained functions may be encoded in sequences showing little sequence conservation. For instance, only a small fraction of in vivo binding sites for key transcription factors appears to be shared by any two mammalian species, suggesting frequent gain and loss of individual regulatory elements (Kunarso et al. 2010; Schmidt et al. 2010). In the same line of reasoning, a study relying on ChIP-Seq analysis with the enhancer-associated protein p300 to identify regulatory elements involved in heart development showed that these elements are only weakly conserved across vertebrates (Blow et al. 2010). Therefore, current evidence suggests that regulatory turnover might represent the norm, rather than the exception.

Transcriptional regulation would be maintained by the conservation of the overall transcriptional architecture, rather than of the sequence. Flexibility in the relative order and spacing of binding sites for relevant transcription factors would permit substantial sequence evolution while maintaining the overall function of a regulatory element (Arnosti and Kulkarni 2005). Because transcription factor binding sites are usually short and degenerate, standard alignment algorithms often fail to align them correctly. To address this issue, alignment-free approaches relax the definition of conservation. These approaches consider a binding site to be conserved across a set of orthologous sequences if the site occurs anywhere within the sequences, irrespective of relative distance and orientation, and thereby extend the power of traditional comparative genomics methods (Palin et al. 2006; Gordan et al. 2010; Taher et al. 2011).

COMPARATIVE SEQUENCE ANALYSIS: METHODS AND TOOLS

Sequence Alignment

Comparative sequence analysis is based on sequence alignments that identify matching nucleotides (or amino acids), mismatches, and unalignable segments (known as gaps) in homologous sequences either from the same or from different genomes. One of the original alignment methods, which is still one of the most popular, is BLAST—Basic Local Alignment Search Tool (Altschul et al. 1990). BLAST is a pairwise alignment tool that allows the comparison of given nucleotide or protein sequences as well as searches against a database of known sequences. A wide range of genome databases is supported by the National Center for Biotechnology Information (NCBI) including GenBank, the database of complete genome sequences of species (Benson et al. 2012). Modern BLAST is a fully customizable toolkit consisting of a wide variety of different alignment tools; the accuracy, speed, and multiple other alignment parameters of the BLAST toolkit can be adjusted by the user depending on a particular task.

Because conservation patterns are often subtle and/or lineage specific, pairwise alignments could fail to capture all sequences under evolutionary constraint (Margulies et al. 2006). Instead, a larger proportion of such sequences can be identified by the use of multiple sequence alignment. In particular, whole-genome multiple alignments are currently considered of utmost importance to elucidate the functional landscape and evolutionary history of the human genome.

Whole-genome multiple alignment strategies can be broadly divided into two classes. Local aligners work by “stacking” pairwise alignments and are therefore very specific, whereas global aligners predefine collinear segments and show better sensitivity. An example of the local strategy is Threaded Blockset Aligner (TBA)/MultiZ; MultiZ is the component within TBA responsible for the dynamic programming alignment step (Blanchette et al. 2004). TBA/MultiZ uses BLASTZ (Schwartz et al. 2003), a BLAST-like method, to find gapped local alignments which are then arranged together in a partially ordered block-set representing a “global” set of local alignments. These alignments are available as tracks in the UCSC Genome browser (Kent et al. 2002). Examples of the global strategy are MAVID (Bray and Pachter 2004), MLAGAN (Brudno et al. 2003), and Pecan (Paten et al. 2008). MAVID is an alignment approach that integrates maximum-likelihood inference of ancestral sequences, automatic guide-tree construction, protein-based anchoring of ab-initio gene predictions, and constraints derived from a global homology map of the sequences. More specifically, MAVID recursively defines a global alignment using maximally nonrepetitive ungapped subsequences before applying standard pairwise alignment methods to join these segments. MLAGAN assumes that a set of orthologous regions, or “anchors,” can be identified correctly, and that there are no genomic rearrangements. Nearby and consistent “anchors” are joined together to construct an orthology map. Dynamic programming is then applied to compute a global alignment around the “anchors.” MLAGAN alignments can be visualized using VISTA (Mayor et al. 2000). Finally, Pecan works by first partitioning the input genomes into a set of collinear segments, a process which essentially handles rearrangements, including duplications, and then building alignments in these collinear segments. Pecan alignments are publicly available within the Ensembl Genome Browser (Flicek et al. 2012). A

recent assessment (Chen and Tompa 2010) of the ENCODE pilot regions found that the agreement among the alignments computed with these tools is not always satisfactory, and that, overall, Pecan is the most accurate and sensitive alignment program.

Genome-Wide Identification of CNS

A common, straightforward approach to scoring conservation and defining discrete evolutionarily conserved elements in the genome employs pairwise nucleotide alignments between the locus of interest and its orthologous counterpart from another species. Based on the alignment, each nucleotide in the locus is assigned a conservation score, which corresponds to the proportion of identical nucleotides in a window centered on that particular nucleotide. A classical, empirically set threshold to call a CNS is 70% identity across 100 bp. When species within the mammalian lineage are compared, more CNS than coding exons are commonly identified in a given locus. Thus, the number of human/mouse CNS exceeds the number of coding exons by about fivefold. Graphical conservation analysis tools are often not only helpful but also necessary to effectively process and analyze the distribution of CNS in long and/or multigenic loci. Multiple tools are dedicated to this task, including VISTA (Mayor et al. 2000) and zPicture (Ovcharenko et al. 2004a). These tools plot the level of nucleotide conservation along the reference sequence and superimpose the annotation of genic elements (coding exons, UTRs, and introns) over the conservation profile. For example, the conservation analysis of the *IL13/IL4* locus in human and mouse performed using zPicture identifies about 20 CNS in the locus, including the aforementioned CNS-1, which regulates both *IL13* and *IL4* as well as the more distant *IL5* (Loots et al. 2000). CNS-1 is located 3 kb downstream from *IL13* (Fig. 2A), is 1.6 kb long, and features 79% identity (Fig. 2C). Although the mechanisms of the *IL5* regulation have not been studied in detail, it is known that CNS-1 targets *IL5* specifically and does not affect the expression of the *RAD50* gene located in between CNS-1 and *IL5* (Loots et al. 2000) (Fig. 2D). A similar analysis, based on VISTA, was performed to analyze the conservation landscape of the *RET* proto-oncogene, leading to the identification of a well-conserved intronic CNS, mutation in which underlies Hirschsprung disease risk (Emison et al. 2005).

Using the same strategy, multiple sequence alignments often provide essential information for the identification of putative regulatory elements. The tool Mulan (Ovcharenko et al. 2005a) integrates TBA/MultiZ with alignment visualization based on zPicture (Ovcharenko et al. 2004a). We applied Mulan to show the use of multispecies alignments in the conservation analysis of 20 kb segments around two genes—iroquois homeobox 4 (*IRX4*; Fig. 3A) and hypermethylated in cancer 1 (*HIC1*; Fig. 3B). Often, the poor quality of genome sequences might result in missing alignment segments: A clear case is the human *IRX4* alignment with chicken and chimp, which misses coding exons and chunks of noncoding DNA. Similar effects can be traced in the *HIC1* alignment as well, and should be ignored. However, the alignable parts of the sequences provide a good representation of the divergence rate at these two loci. In case of *IRX4*, there is extensive conservation in all mammals, and the identification of the most conserved segments is simple only in the alignment with frog or fugu fish that feature five and one CNS, respectively. The human/fish CNS located ~3 kb downstream from *IRX4* is the only CNS conserved across all vertebrates, and is a known brain enhancer (Tena et al. 2011), likely playing a key role in regulating the expression of this gene. The conservation profile of *HIC1* is different: The alignment of the human sequence with the distant fish and frog genomes does not reveal any CNS. In the case of this gene, the most refined evolutionary analysis of *HIC1* comes from the human/mouse alignment, exposing seven CNS upstream of the gene transcription start site. Performing multiple sequence alignments with species that span as large a part of the phylogenetic tree as is feasible for a particular locus is often a necessity in a conservation analysis of a locus. This method has been proven to work as a rapid screen for the identification of putative proximal and distant regulatory elements (Nobrega et al. 2003; de la Calle-Mustienes et al. 2005; Savic et al. 2011).

From these analyses, it is evident that the divergence rate varies considerably across the sequence of a genome, and that other definitions of conservation might be more appropriate, depending on the context. Usually, housekeeping gene loci diverge slowly under strong purifying selection, while the loci

FIGURE 2. Graphical display of conservation analysis. (A) The conservation profile of the *IL13/IL4* locus was generated using zPicture (zPicture ID: 1209116710196). The level of sequence identity in a 100-bp sliding window is plotted on a 50%–100% scale. The direction of transcription of *IL13* and *IL4* is indicated by blue arrows. CNS are color coded based on their genic characteristics (intergenic, intronic, coding, or UTR). (B) *IL13/IL4* locus is displayed in the PhyloP and GERP conservation tracks of the UCSC Genome browser (Kent et al. 2002) (C) The alignment of human and mouse CNS-1 orthologous sequences reveals the extent of identity between the two sequences. (D) A schematic representation of the CNS-1 regulatory profile shows the relative locations of the elements.

of tissue-specific genes accumulate mutations at a faster rate (Zhang and Li 2004; Vinogradov and Anatskaya 2007). For example, there are 335 human/mouse CNS in the locus of the orthodenticle homeobox 2 (*OTX2*) gene, implicated in medulloblastoma, *OTX2* encodes a transcription factor that plays a role in brain and sensory organ development (Balikova et al. 2011; Bunt et al. 2011) [Mulan ID m12130214635891]. In contrast, there are no human/mouse CNS in the locus of transmembrane and immunoglobulin-domain containing 2 (*TMIGD2*) gene.

Several statistical methods exist that attempt to score conservation by evaluating changes in the mutation rate relative to neutrally evolving sequence. Most of these methods consider either positive or negative selection, and assume that selection acts uniformly across the branches of a phylogeny. A standard likelihood ratio test (LRT) to detect changes in the mutation rate involves comparing the likelihood of an observed alignment for two models with different evolutionary rate parameters, where one model represents neutral evolution. The LRT statistic is the log of the ratio of these two likelihoods. Conserved regions can be ranked using the LRT statistic (e.g., Pollard et al. 2006). If a

FIGURE 3. Mulan multisequence alignment. The Mulan display shows the conservation analysis of multiple species around the *IRX4* locus [m1213053390883] (A) and around the *HIC1-SMG6* locus [m12140221589700] (B).

locus is under purifying selection, deleterious mutations have a higher chance of being rejected than nondeleterious mutations. Consequently, the entire phylogenetic tree of the alignment will show a lower number of mutations than would be expected under a neutral model.

Based on this strategy, GERP (Cooper et al. 2005) identifies constrained elements in a multiple alignment by quantifying “rejected substitutions.” Relying on a multiple alignment model and a published phylogenetic model (Structural EM Phylogenetic Reconstruction [SEMPHY]), GERP first estimates “observed” rates of evolution across the tree for each individual nongap nucleotide in the reference genome, using the maximum-likelihood-based SEMPHY program (Friedman et al. 2002). GERP also computes “expected” rates of evolution based on the neutral model; in this step, GERP uses the Hasegawa, Kishino, and Yano (HKY) model of neutral evolution. In this model, gaps in genomes are removed from the computation of both the observed and the expected rates of evolution, making the method robust to missing data; sites lacking sufficient expected substitutions are excluded from the analysis. Candidate constrained elements are discovered by identifying stretches of alignment positions that show ratios of observed to expected rates below a certain threshold. The GERP score represents the degree of conservation of an element. Positive scores indicate fewer substitutions than expected, suggesting that the element is under evolutionary constraint. Negative scores indicate higher substitutions than expected, and may be weak evidence of accelerated rates of evolution. The more recent GERP++ program (Davydov et al. 2010) also uses “rejected substitutions” as a metric, but a significantly faster and more statistically robust maximum likelihood estimation procedure to compute expected rates of evolution. GERP, GERP++, and precomputed elements for human and mouse genomes (hg18, hg19, and mm9) using the alignments available at UCSC can be downloaded from the website, <http://mendel.stanford.edu/SidowLab/downloads/gerp/>. GERP scores also constitute a comparative genomic track in the UCSC Genome Browser (Karolchik et al. 2011). Another program, phastCons (Siepel et al. 2005), combines phylogenetic models, as described for GERP, with a Hidden Markov Model (HMM). PhastCons is part of the freely available PHAST (PHYlogenetic Analysis with Space/Time models) software, which includes several phylo-HMM-based programs and related tools for phylogenetic analysis and functional element identification (<http://compugen.bscb.cornell.edu/phast/>). This type of statistical model—phylo-HMM—considers not only the process by which substitutions occur at each site in the genome, but also how this process changes between neighboring sites. The phylo-HMM implemented in phastCons consists of two states: a state of conserved regions and a state of nonconserved regions. Each state is associated with a phylogenetic model. By default, the phylogenetic models are identical except for a scaling parameter, which is applied to the branch lengths, and represents the average rate of substitution in conserved regions relative to nonconserved regions. Thus, as input, phastCons uses a multiple alignment, a phylogenetic model for conserved regions, and, optionally, a phylogenetic model for nonconserved regions. In addition to the scaling parameter, two parameters that define all state-transition probabilities are

learned from the data. PhastCons deals with missing data in a manner similar to that used by GERP. Because phastCons is based on an HMM, each position in the alignment can be assigned a log-odds score, indicating whether the alignment is more likely under the conserved state of the phylo-HMM than under the nonconserved state. These scores can be displayed as a track in the UCSC Genome Browser (Karolchik et al. 2011). Moreover, phastCons produces predictions of discrete conserved elements, as displayed in the “most conserved” tracks in the UCSC browser. Another tool, phyloP (Pollard et al. 2010), allows the detection of sites evolving both faster and more slowly (conservation) than expected under a neutral model, while allowing changes in evolutionary rates in a clade-specific manner. For this purpose, the phyloP package implements four alternative methods for testing for nonneutral evolution, an LRT, a score test, a phastCons-like test, and a GERP-like test. The score test is similar to the LRT, but only the null model is fitted to the data. Independently of the statistical test, given a multiple alignment between different genomes, phyloP first estimates the mean number of substitutions in each genome to estimate the neutral evolution rate. Then, the probability of observed substitutions is analyzed under the hypothesis of neutral evolution. Sites predicted to be conserved are assigned positive scores, whereas sites predicted to be fast-evolving are assigned negative scores. The absolute values of the phyloP score is the $-\log(p\text{-value})$ under a null hypothesis of neutral evolution. Because phyloP ignores neighboring sites in the calculation of the phyloP score of a given site, phyloP is useful for evaluating signatures of selection at particular nucleotides (e.g. third codon positions, first positions of miRNA target sites, etc.). PhyloP has been applied to the human and mouse genomes to reveal patterns of positive/negative selection in functional elements identified by the ENCODE project (Birney et al. 2007; Myers et al. 2011). Tracks displaying phyloP scores are available in the UCSC Genome Browser (Karolchik et al. 2011). PhyloP is also part of PHAST (<http://compugen.bscb.cornell.edu/phast/>). PhyloP scores are plotted along the genome and displayed as part of the “Conservation” track in the UCSC Genome Browser (Karolchik et al. 2011). For the purpose of comparison, we return to the IL13/IL4 locus example; note that CNS-1 is clearly visible in both the GERP and PhyloP tracks of the UCSC Genome Browser (Karolchik et al. 2011) (Fig. 2B). The methods we have just considered model conservation as a decrease in the rate of mutation, independently of the pattern of mutation. Instead, the SiPhy software (Garber et al. 2009) identifies evolutionary selection by uncovering substitution patterns characteristic of sequence undergoing natural selection. For example, transcription factor-binding sites are known to tolerate degeneracy without affecting binding affinity. Therefore, such sites may be misidentified by methods focusing on changes in the substitution rate as unconstrained. Given a multiple alignment and a phylogenetic model, SiPhy uses a probabilistic framework to describe molecular evolution, and models the evolution of sequences along the branches of the phylogenetic tree as a continuous time Markov process. The model has basically two parameters: the fraction of sequence conserved and the typical length of a conserved region. SiPhy is implemented as a Java software package freely available at <http://www.broadinstitute.org/science/software/>. In general, all these programs produce similar results, but differ substantially over specific clades (Pollard et al. 2010). Other widely used programs for detecting conservation include BinCons (Margulies et al. 2003) and SCONE (Asthana et al. 2007). Most programs do not distinguish between contig gaps (incomplete genomes) and inserted or deleted regions; thus, inferences based on alignments comprising a large number of gaps should be considered with caution. Finally, each of these methods has been devised to identify particular conservation patterns, and may not generalize well.

Visualization

When studying the conservation of multiple genes or even one gene in many species, the identification of all necessary individual species’ DNA sequences, mapping repetitive elements, and preparing gene annotation files could be a tedious and time-consuming process. Genome conservation browsers have been created to facilitate evolutionary DNA sequence analysis across multiple loci and species. Among the many conservation browsers currently available, the UCSC Genome Browser (Karolchik et al. 2011) and Ensembl Genome Browser (Flicek et al. 2012) are probably the most popular. These two browsers were developed to provide the deepest possible level of genomic annotation for any locus of a

genome, and feature a constantly growing collection of experimental and computational DNA annotation data, including sequencing, phenotypic, disease, gene prediction, expression, regulation, and variation information. The drawback of the all-inclusive data presentation is in a common lack of detailed information. For example, the conservation data might be limited to a single overview track providing little or no indication of where the pairwise-comparison CNS are. Specialized conservation browsers, such as the VISTA Browser (Dubchak 2007) or the Evolutionary Conserved Regions (ECR) Browser (Ovcharenko et al. 2004b), are more applicable when a high-level resolution of interspecies alignments is necessary. The ECR Browser features an interface and output display similar to zPicture and Mulan and is integrated together with these and other related tools into the NCBI Dcode.org Comparative Genomics Developments portal (Loots and Ovcharenko 2005). We used ECR Browser to investigate the conservation profile of the *IRF6* gene associated with cleft lip (Kondo et al. 2002). The comparison of the human intergenic sequence upstream of *IRF6* with seven other vertebrate species revealed an interesting evolutionary history of the *IRF6* locus. The human interval is exceptionally well conserved in cow and chimp, with CNS covering the intergenic space almost completely, whereas there is no noncoding conservation with distant chicken, frog, and Fugu fish genomes (Fig. 4A). It is difficult to determine which of the 24 human/mouse intergenic CNS might be the best candidates for the key regulatory elements of *IRF6*, but the conservation with opossum pinpoints to two most likely candidates located ~15 kb upstream of *IRF6* (Fig. 4A). To narrow down the search of the top candidate, we used the “core ECR” approach (Ovcharenko et al. 2004c) implemented in ECR Browser. Core ECRs are longer (350 bp and longer) and feature a higher level of evolutionary conservation (at least 77% identity) than classical 100 bp/70% identity ECRs. This additional filtering identified a sole human/mouse core CNS in the locus, which coincides with one of the two human/opossum CNS (Fig. 4B). The core ECR human/mouse conservation profile of the *IRF6* locus was next subjected to the identification of SNPs in ECRs (another function of ECR Browser), and two SNPs located in the human/mouse core CNS—rs642961 and rs11582607—were identified. It is likely that either one of these SNPs (or both) might disrupt the regulatory activity of the CNS and impact the level of *IRF6* expression—a variation putatively associated with the disease involvement of the gene. Indeed, the rs642961 SNP has been

FIGURE 4. ECR browser analysis of the *IRF6*-*C1orf107* 30 kb human locus conservation. The comparative analysis was performed with chimp, cow, mouse, opossum, fugu, chicken, and frog (A) and in opossum and mouse (B). The standard 100 bp/70% identity threshold has been used to identify 24 human intergenic CNS (A), and a strict—core ECRs (350 bp/77% identity)—threshold was used to identify the sole well-conserved human-mouse CNS in the locus (B). Gray vertical lines in the middle of the plot indicate the position of the disease SNP, rs642961.

previously documented to disrupt a TFAP2A-binding site within the CNS and is strongly associated with cleft lip (Rahimov et al. 2008). Although the analysis presented was based on our familiarity with this SNP and the corresponding disease, it illustrates a step-by-step approach for disease SNP prioritization using ECR Browser that could be applied to any locus and any species of interest.

GENOME ANNOTATION WITH EXPERIMENTAL HIGH-THROUGHPUT TECHNOLOGIES

High-Throughput ChIP Experiments

Chromatin immunoprecipitation (ChIP) directly profiles protein–DNA binding in vivo (Solomon et al. 1988). In a ChIP experiment, all DNA–protein interactions are cross-linked, the cells lysed, and resulting chromatin sonicated to form short fragments of ~500 bp long. Subsequent analysis with PCR using primers specific to DNA regions of interest can be used to assess protein–DNA binding at those regions. Although ChIP does not directly identify the functionality of the protein-bound regions, ChIP assays are inherently scalable. Immunoprecipitated regions can be hybridized to a microarray (ChIP-chip) to identify bound regions on a large scale. An alternative approach, makes it possible to follow ChIP by next-generation massively parallel sequencing (ChIP-Seq). ChIP-Seq provides better coverage and resolution than ChIP-chip and is becoming increasingly more affordable.

Both ChIP-chip and ChIP-Seq have been used successfully to identify regions bound by hundreds of transcription factors in several organisms and cell-types. An early notable study was performed in yeast, in which Harbison and coworkers profiled more than 100 transcription factors in multiple environmental conditions (Harbison et al. 2004). An advantage of high-throughput ChIP is that it is not restricted to transcription factors that bind DNA directly. The approach has been applied to activator proteins such as P300 and CBP that mediate transcription by interacting with numerous DNA-binding transcription factors (Kasper et al. 2006). Because P300 and CBP localize at enhancer regions, albeit not bound directly to the DNA, profiling one of them obviates to some extent profiling the transcription factors separately. Both proteins have been profiled on a genome-wide scale to reveal enhancers in mammals (Heintzman et al. 2009; Visel et al. 2009b; Blow et al. 2010; Ramos et al. 2010).

High-throughput ChIP has also been applied to probe the local chromatin structure along the genome, which is known to be highly variable. Specifically, the histones around which the DNA is wrapped to form nucleosomes can be chemically modified and exchanged for variants. Together, nucleosome positioning and histone variants and modifications determine the primary structure of the chromatin, which exists in a structural continuum between the closed heterochromatic and open euchromatic conformations. Chromatin conformation is dynamically controlled by a number of processes and factors—DNA methylation, histone modifications (e.g., methylation, acetylation, and ubiquitination of histone tails), and small RNAs and DNA-binding proteins (e.g., the polycomb group), which, consequently, play an important role in transcriptional regulation. The need to identify mechanisms that could alter chromatin conformation, and thus transcription, in a transient but potentially stable and inheritable manner has given rise to the field called epigenetics. Epigenetics is commonly referred as the “second code,” in that epigenetic modifications define how different genetic programs can be executed from the same genome, in different tissues and developmental stages. Aberrant epigenetic marks have been directly implicated in common human diseases, including diabetes, cardiopulmonary diseases, neuropsychiatric disorders, autoimmune diseases, and cancer as well as in ageing (Roth et al. 2009; Grolleau-Julius et al. 2010; Portela and Esteller 2010). Importantly, the epigenome represents the interface between genetics and the environment. Moreover, the fact that epigenetic modifications are potentially reversible and can be induced by environmental and nutrition changes has significant implications for the prevention, diagnosis, and treatment of major human diseases.

High-resolution “nucleosome maps” arising from ChIP experiments profiling core nucleosomes in yeast (Shivaswamy et al. 2008), worm (Valouev et al. 2008), and later in human (Schones et al. 2008; Li et al. 2011) have revealed significant nucleosome depletion at regulatory regions. Although not all nucleosome-free regions are necessarily regulatory in nature or transcriptionally active, there is some

correlation between the nucleosome-free state and regulatory or transcriptional activity, which has been exploited to identify regulatory regions (Narlikar et al. 2007; Rosenbloom et al. 2010). Histone variant H2A.Z has also received much attention and been shown to be associated with regulation. Several histone methylation and acetylation marks have been profiled in human T cells. Based on these whole-genome marks, distinct chromatin signatures have been shown to be associated with a variety of regulatory elements (Barski et al. 2007; Heintzman et al. 2007, 2009).

Assays to Identify Open Chromatin

As discussed in the previous section, most DNA regions bound by transcription factors are devoid of nucleosomes. As a result, assays detecting open chromatin regions can be used to identify regulatory elements. These assays are especially beneficial when the identity of the DNA-binding protein is unknown or its antibody is not available. Two kinds of sequencing assays target open chromatin regions: DNase-Seq and Formaldehyde-Assisted Identification of Regulatory Elements (FAIRE)-Seq. DNase-Seq assays rely on the use of the DNaseI endonuclease, which nonspecifically digests regions of open DNA, called DNaseI hypersensitive sites. Two different techniques have been proposed for mapping DNaseI hypersensitive sites on a genome-wide scale (Crawford et al. 2004; Sabo et al. 2004). In both methods, the nuclei of the cells are digested briefly with DNaseI. Crawford et al. (2004) attach a biotinylated tag to the cleaved DNA ends, which is used to extract the cleaved ends. The DNA ends are then identified using microarrays or by high-throughput sequencing. Sabo et al. (2004) exploited the fact that two cleavage events are more likely to occur near each other in accessible regions, and therefore extracted the small-length fragments from the digested chromatin for identification using microarrays. The quality of the resulting maps can be estimated by comparing them with the “gold standard” of Southern blotting.

A FAIRE-Seq assay (Giresi et al. 2007) is relatively simpler: the steps are similar to ChIP-Seq, but without the use of antibody. Instead, the noncross-linked DNA (presumably free of histones) is segregated from the sonicated DNA mixture and then sequenced. Both DNase-Seq and FAIRE-Seq report a large number of common regions. However, sites identified by using only one assay are enriched for binding sites of certain specific DNA-binding proteins (Furey 2012), indicating that these assays may be biased toward open chromatin marked by certain kinds of protein–DNA complexes. Furthermore, there are differences in the classes of genomic regions identified by the two assays: Promoter regions are better detected by DNase-Seq, whereas distal regulatory regions are more likely to be identified by FAIRE-Seq (Song et al. 2011).

Current Large-Scale Genome and Epigenome Annotation Projects

Rapid progress of experimental technologies in molecular biology has given rise to several genome-wide initiatives to identify functional elements. These projects have spurred unprecedented innovation in the assessment, application, improvement, and development of experimental and bioinformatics methods. The projects led by the Encyclopedia of DNA Elements (ENCODE) and the Roadmap Epigenomics Mapping consortia have generated exceptional publicly available resources for the scientific community to advance basic biology and disease-oriented research.

ENCODE Project

The ENCODE Consortium is an international collaboration of research groups with diverse backgrounds and expertise funded by the National Human Genome Research Institute (NHGRI) that aims to identify all functional elements in the human genome. This public research project was launched in 2003. During its initial 4-yr pilot phase, ENCODE tested and compared existing techniques, technologies, and strategies to exhaustively map and validate functional elements on a set of 44 regions comprising 1% (30 Mb) of the genome. Among the findings in this phase were the confirmation that the human genome is pervasively transcribed, even though only a small fraction of transcripts are subsequently translated into proteins; the identification of multiple overlapping transcripts and

rare, poorly conserved splice variants for many known genes; and the observation that chromatin accessibility and histone modifications and transcription activity are well correlated. Perhaps most surprising was the revelation that only 50% of regulatory sequences appear to be conserved. The project expanded to a whole-genome scope in 2007 (Huynen et al. 2004; Birney et al. 2007; Myers et al. 2011). The ENCODE Consortium conducts multiple assays on various cell types, under a set of experimental conditions. For example, transcriptome analyses are performed on poly(A)+ or poly(A)– fraction of RNA extracts from different subcellular compartments (Rosenbloom et al. 2010). To date, the ENCODE project has generated data for more than 30 different assays and 211 cell types (Table 1; Rosenbloom et al. 2012). Primary data from ENCODE are available at the NCBI GEO (<http://www.ncbi.nlm.nih.gov/projects/geo/info/ENCODE.html>) and the EBI ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) public array data repositories. The ENCODE web portal at UCSC (<http://encodeproject.org> or <http://genome.ucsc.edu/ENCODE>), the ENCODE resource page at Ensembl (http://www.ensembl.org/Homo_sapiens/encode.html), and the ENCODEdb portal at NHGRI (<http://research.nhgri.nih.gov/ENCODEdbGRI>) provide convenient links for access (Elnitski et al. 2007; Rosenbloom et al. 2010). Unreleased data can be viewed on the UCSC Preview Browser (<http://genome-preview.ucsc.edu/>). All ENCODE data are freely available for download and analysis; however, publication of global analyses is restricted for 9 mo following data release. ENCODE is currently (in 2014) entering the seventh year of its production phase.

Roadmap Epigenomics Program

Whereas the ENCODE project focuses on determining functional elements in the genome, the National Institutes of Health (NIH) Roadmap Epigenomics Program (<http://www.roadmapepigenomics.org>) investigates the epigenetic patterns associated with these elements, and thus, complements

TABLE 1. ENCODE data types, as of January 2012

Data type	Description
5C	Chromatin conformation capture
BIP	Bi-directional promoters, identified informatically
CAGE	5' cap analysis of gene expression
ChIA-pet	Chromatin interaction analysis with paired-end tags
ChIP-seq	DNA fragments from ChIP purifications, measured by sequencing
Genotype	CNV and SNP as determined by the Illumina Human 1M-Duo Infinium HD BeadChip assay and circular binary segmentation (CBS)
DNase-DGF	Digital DNase genomic footprinting
DNA PET	DNA fragments measured by paired-end di-tag sequencing
DNase-seq	Sequencing of DNase-digested DNA
Exon-array	RNA expression measured by Affymetrix exon microarrays
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
Gencode	Curated gene annotation
Integ Cluster	Element cluster by integrative analysis
Mapability	Level of sequence uniqueness within the genome
Methyl27	DNA methylation, measured by Illumina bead arrays
Methyl-RRBS	DNA methylation, measured by reduced representation bisulfite sequencing
Methyl-seq	DNA methylation, measured by sequencing
NRE	Negative regulatory elements
Nucleosome	Nucleosome positioning data
ORChID	Predicted hydroxyl radical cleavage intensity on naked DNA
Proteogenomics	Mass Spectrometry Proteogenomic Mapping
RIP-chip	RNA-protein interactions, measured by microarrays
RIP-chip Gene ST	RIP-chip on Affymetrix GeneChip (R) Human Gene 1.0 ST Arrays
RIP-chip Tiling Array	RIP-chip on Affymetrix GeneChip ENCODE 2.0R Tiling Arrays
RIP-seq	RNA-protein interactions, measured by sequencing
RNA PET	RNA expression measured by paired-end di-tag sequencing
RNA-chip	Tiling arrays measuring expression in various cell compartments
RNA-seq	RNA expression measured by sequencing
SwitchGear	SwitchGear Genomics 3' UTR reporter assay

ENCODE. Initiated in 2008, the goals of the Roadmap Epigenomics Program are to establish multiple reference epigenomes; develop, standardize, and disseminate protocols and reagents for epigenetic research; identify public resources for high-quality stem and differentiated cells, and tissues; and provide publicly accessible data and analytical tools to enable the generation of new hypotheses on epigenetic roles in human health and disease.

The Roadmap Epigenomics Mapping Consortium uses next-generation sequencing technologies to map DNA methylation (MethylC-seq, BS-seq, MeDIP-seq, MRE-seq), histone modifications (ChIP-seq), chromatin accessibility (DNase-seq), and small RNA transcripts (RNA-seq, microarrays) in stem cells and primary *ex vivo* tissues to represent the normal counterparts of tissues and organ systems frequently involved in human disease. High-value cell types, such as human embryonic stem cells (hESCs), are subjected to deep exploration of >30 histone modifications and comprehensive analysis of DNA methylation (using single-nucleotide resolution MethylC-seq). To capture a wide scope of the epigenomic differences among cell types, the consortium also intends to describe DNA methylation as well as six major histone modification patterns (H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3, and H3K36me3) in >100 human cell types and tissues (Bernstein et al. 2010). The first main contribution of the Roadmap Epigenomics Mapping Consortium was the publication of DNA methylation profiles at single-nucleotide resolution across the genomes of two human cell lines, H1 hESC and IMR90 fibroblasts. This approach showed a correlation between DNA strand-specific methylation and transcription, and highlighted differences in the prevalence of non-CG methylation among different cell types, suggesting that non-CG methylation may play a role in maintaining pluripotency (Lister et al. 2009). Histone modifications profiles across multiple cell types are beginning to reveal different classes of regulatory elements, their cell-type specificities, and their functional interactions (Ernst et al. 2011).

Data generated by the Roadmap Epigenomics Mapping Consortium are released immediately, with a 9-mo embargo for publication of genome-wide analyses, and are archived in the Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/epigenomics>). Epigenomic maps of major tissues and cell lines can be accessed through the Human Epigenome Atlas (<http://www.genboree.org/epigenomeatlas/>).

IHEC Project

Despite currently being the largest epigenomics project, the Roadmap Epigenomics Program will only describe a limited number of epigenomes. The goal of the International Human Epigenome Consortium (IHEC), an initiative begun in 2010, is to generate at least 1,000 reference epigenomes for multiple human cell types and developmental stages, laying the foundation to study the epigenetic mechanisms of human diseases (<http://ihec-epigenomes.org/>; Abbott 2010). The IHEC also intends to help developing good practices and standards for epigenomic data generation, analysis, and integration. Although plans for the consortium are not finalized, it would expand the number of cell types characterized by the Roadmap Epigenomics Program, possibly including nonhuman cells (Satterlee et al. 2010).

COMPUTATIONAL METHODS FOR PREDICTING REGULATORY ELEMENTS

The methods and tools described in the section on comparative sequence analysis are crucial for visualizing genomic regions and selecting them for further analysis. Selection is typically done through manual inspection or with a predetermined cutoff on the extent of sequence conservation. Although these methods have been used successfully to winnow possible functional regions from noise, they do have several inherent limitations, such as the number of genomic regions that can be scanned manually, subjectivity in the calls, inability to identify the functionality across cell-types, and the danger of missing diverged elements. On the other hand, high-throughput experimental methods that profile the binding of specific proteins on a genome-wide scale (described in the preceding section on

genome annotation) get around these limitations. These approaches, however, are restricted by the cost of the experiments, the number of proteins that can be profiled, and the number of different cell types or conditions that can be used during the experiments. As a result, computational approaches that use information from different sources such as genomic sequence, epigenetic information, and curated databases are rapidly gaining ground. Here we describe a few representative approaches with the acknowledgment that there are many more methods with different features used by the scientific community. The purpose of this section is to familiarize the reader with a few core concepts popularly used in computational regulatory element identification.

Using Sequence Conservation to Identify Functional Elements

Comparative genomics is one of the most prominent approaches taken to identify functional regions in a genome. Regions conserved across species, i.e., regions under negative selection, are likely to be important for the functioning of the organism. In the case of gene identification, several tools using multiple genomes have shown more proficiency in terms of increased sensitivity and specificity when compared with single genome methods (Miller et al. 2004). In the case of regulatory elementary elements as well, conservation often signifies functionality. Indeed, several predicted elements based on sequence conservation only have been shown to act as enhancers (Emorine et al. 1983; Loots et al. 2000; Kellis et al. 2003; Nobrega et al. 2003; Pennacchio et al. 2006; Navratilova et al. 2009). Kellis et al. (2003) identified putative motifs in *S. cerevisiae* using genomes of multiple yeast species. Their approach relied on both the frequency of occurrence of the motif as well as its conservation. The authors later extended the method to mammals (Xie et al. 2005) and fly (Stark et al. 2007).

Sequence conservation, when used in conjunction with other large-scale data, can prove especially useful. For example, de novo motif discovery within regions reported from high-throughput ChIP experiments or within coregulated promoters can benefit from using sequence conservation in the model. Several tools exist for this purpose, some based on alignments (Sinha et al. 2004a; Siddharthan et al. 2005) and others based on word-frequencies (Kantorovitz et al. 2007; Gordan et al. 2010). The latter approaches are now gaining more attention, especially for the application of regulatory element detection. In contrast to genes, the very structure of regulatory regions makes them susceptible to rapid turnover (Dermitzakis and Clark 2002). A regulatory element is typically bound by several proteins: the individual protein–DNA binding sites, which are ~5–15 bp long, can move around or change orientation during evolution, while maintaining the regulatory property of the region. As a result, regions that appear diverged in the traditional sense, where alignments do not score well, in reality may perform the same regulatory role (Fisher et al. 2006). As a result, searching for conserved *cis*-regulatory *modules* (CRMs), where the goal is to look for conservation of the *structure* of the element instead of individual nucleotides is more effective in locating regulatory regions. The next few sections describe methods that identify CRMs. In all situations, sequence conservation can, in principle, be incorporated in the models, but must be done with caution. Using species from different levels of divergence can yield different results: conserved regions found in very closely related organisms may not signify function, whereas those regions in detected in diverged genomes may not contain the same regulatory apparatus.

Incorporating Known Transcription Factor Binding Motifs

Databases such as TRANSFAC (Matys et al. 2006) and JASPAR (Sandelin et al. 2004) maintain a catalog of binding specificities of eukaryotic transcription factors derived from experimental data. In addition, communities working on specific model organisms such as yeast, worm, fly, etc. maintain more detailed regulatory databases of the respective organism (Tweedie et al. 2009; Cherry et al. 2012; Yook et al. 2012). Several computational tools use this information directly to identify clusters of potential transcription factor-binding sites. Position weight matrices (PWMs) (Staden 1984) are commonly used for modeling binding specificities of a transcription factor. For every position within the binding site of a transcription factor, its PWM stores the binding preference of the factor toward each nucleotide in terms of a probability. It implicitly assumes that nucleotide prefer-

ences show independence across positions. While this assumption has been shown to be invalid for many transcription factors (Bulyk et al. 2002; King and Roth 2003), the simplicity of the PWM and lower number of parameters make it a popular choice for representing binding specificities. Furthermore, it is fairly straightforward to scan a new genomic sequence with a PWM to look for potential matches. Visually, the binding site is represented as a sequence logo (Crooks et al. 2004) where the height of a nucleotide at each position indicates the binding preference (Fig. 5).

When the only information available is a database of known binding specificities, predictions of regulatory elements can be made in two steps. First, the genomic region is scanned for possible matches to PWMs to identify putative binding sites of transcription factors. The region is then scored using some statistic that evaluates the significance of the density of putative binding sites. This second step is crucial to account for spurious matches to PWMs that are not bound *in vivo*, but occur just by chance. When several matches are found together, the possibility of them being spurious is reduced. Furthermore, regulatory elements are typically targeted by multiple transcription factors, which act together for gene activation or repression. The significance of the matches is typically assessed by one of two methods. One is to count the number of hits in a given window of some size and set an explicit minimum number or compare it with random expectation (Wagner 1999; Berman et al. 2002; Rebeiz et al. 2002; Johansson et al. 2003; Blanchette et al. 2006). This is a more intuitive approach with few parameters; however, it requires thresholds for determining a binding site match and window length, which may be organism dependent. The second approach involves representing the entire regulatory element with a probabilistic model. Hidden Markov models (HMMs; Wu and Xie 2010), typically used for this purpose, can be applied to make more mathematically sound predictions of regulatory elements and can also incorporate conservation information in their predictions. Several methods use a variant of HMMs (Frith et al. 2001, 2002; Sinha et al. 2004b). Interestingly, approaches involving only one PWM for scanning the genome have shown that homotypic clusters of putative binding sites are indicative of functionality and have great potential in identifying regulatory elements (Lifanov et al. 2003; Gotea et al. 2010).

One disadvantage of probabilistic models is that they are less intuitive, more complex and may not scale well with a large number of motifs. As a result, when investigating a certain cell-type or process, knowing the identity of the transcription factors involved can result in more accurate predictions (Frith et al. 2001, 2003; Berman et al. 2002). Ensuring that only the clusters of binding sites belonging to transcription factors of interest are retained in the analysis will lead to detection of regulatory elements active in the process under study. In a contrasting situation, not knowing the transcription factors, but knowing the identity of genes involved in a common process can yield more accurate predictions of elements regulating those genes (Aerts et al. 2003; Gotea and Ovcharenko 2008; Van Loo et al. 2008). The loci of such genes are searched for matches to PWMs that are overrepresented

FIGURE 5. Using position weight matrices (PWMs) to identify transcription factor binding sites. The genomic region of interest is scanned with known binding preferences of transcription factors to find potential binding sites. The spatial distribution of these predicted binding sites is used to identify regulatory elements.

with respect to loci of other genes. The idea is that co-expressed genes are likely to be regulated by a common set of transcription factors and hence will be enriched for the binding sites of those transcription factors. As a result, as well as the regulatory elements, the set of transcription factors involved in the regulation can be simultaneously identified.

Having a set of regulatory elements known to be active in a specific cell type can be particularly useful. One can then predict binding sites within these elements and build a model based on the binding sites that help distinguish between regulatory elements and nonregulatory regions. Several methods attempt to learn such a classifier; one of the earliest was developed by Wasserman and Fickett for predicting muscle (Wasserman and Fickett 1998) and later liver (Krivan and Wasserman 2001) regulatory elements. In this approach, the authors used PWMs of transcription factors involved in the respective tissue and used logistic regression based on matches to those PWMs in the regulatory elements. Most of these tools require a set of binding specificities as input, while some learn binding specificities *de novo* as we shall see in the next section.

It is important to note that most of the methods described in this section have been applied only to promoter regions, and care must be taken when using them to identify distant regulatory elements such as enhancers and silencers. Some assumptions made for promoters such as knowledge of the target gene, high background GC-content, and proximal core-promoter signals are not necessarily applicable to other regulatory elements.

Using Sequence Features and De Novo Binding Sites

Our current knowledge of transcription factor-binding specificities is far from complete. Not only are binding sites of several transcription factors uncharacterized, but also many transcription factors themselves have not been identified. The databases of binding sites noted previously are useful only for dealing with regulatory processes for which the regulators and their binding sites are known. Here we discuss a few representative methods that get around this issue, some of which are also available as web servers (Kazemian et al. 2011; Taher et al. 2011b). All these methods learn model parameters from known regulatory regions and can be used to scan genomic regions to identify novel regulatory elements.

One of the early methods that identified clusters of *de novo* PWMs was CisModule. From larger regions known to be active in a certain cell-type or tissue, CisModule simultaneously learns binding sites and cluster boundaries. This ability has been extended to use HMMs in more complex models (Gupta and Liu 2005; Shi et al. 2008). These methods are restricted by the number of motifs that can be included in the model and work well with a smaller number of motifs. In practice, a large number of PWMs implies a larger number of parameters to learn, which is time-consuming with no guarantee of reaching the optimal parameters. Methods like CLARE (Taher et al. 2012) include both known PWMs from databases and *de novo* PWMs in their feature-based approach. A linear classifier is learned that selects a small set of PWM features, which are crucial for discriminating between regulatory elements of a certain tissue and other noncoding regions. This process ensures that, although the starting set of PWMs is large, the final model contains only a few relevant PWMs.

In a novel approach, Sinha et al. propose models without PWMs, but based instead on frequencies of words. Their non-PWM-based methods work when a set of known regulatory elements are available (Kantorovitz et al. 2009) and also when elements are to be searched for in loci of a set of co-regulated genes (Ivan et al. 2008). While their methods show good discriminating power across organisms, caused by the inherent non-PWM-based approach, they are not able to decipher the structure of the regulatory element in terms of transcription factor binding sites.

Incorporating Epigenetic Information

Recent studies have shown that different parts of the genome have different local chromatin structures, and these structures are often specific to the cell type. Indeed, as we have discussed, it is well established that chromatin remodeling plays an important role in gene regulation. It has been shown

that regulatory elements are marked with distinct chromatin signatures based on the function of the region. These chromatin signatures include the several possible histone modifications, DNA modifications, and presence of histone variants in the nucleosome complex.

Histone H3 lysine 4 (H3K4) methylation is one of the histone modifications that has been most extensively studied on a genome-wide scale (Hon et al. 2009). H3K4me1 (mono-methylation) is typically enriched at enhancers whereas H3K4me3 (tri-methylation) is associated with promoters. Hon et al. (2009) show the presence of many more distinct combinations of modifications at putative enhancers, silencers, insulators, and promoters. However, these observations are mostly based on correlations, and there has been only one attempt thus far to build a predictive model from these modifications. Ren et al. have developed ChromaSig (Hon et al. 2008), a probabilistic model that uses genome-wide epigenetic information to identify regulatory elements.

As described in the section on genome annotation, a plethora of data is available on a genome-wide scale for dozens of chromatin modifications. The authors strongly believe in the potential of more powerful computational approaches that take into account sequence conservation, sequence features, and epigenetic information while making predictions.

Validating Predicted Elements

Validation of putative elements can be done in two ways: computationally and experimentally. In a computational approach, some of the available information is typically omitted while building the model; the information initially withheld is used subsequently to assess whether the predictions made by the model are supported. An effective and popular approach is the k -fold or leave-one-out cross-validation. In a k -fold cross-validation, the training is conducted on a smaller training set by leaving out $1/k$ th of the set. These “left-out” data are then used for testing the trained model. This whole process is repeated k times, each time leaving out a different $1/k$ th of the set. In other words, k different models are trained in such a manner that each of the available data samples is used for training $k-1$ models, but each is used for validating only once. This strategy ensures that the models are always tested on “unseen” data. This method works when a set of known elements is available for training. However, some of the approaches described earlier make predictions *ab initio*, with no characterized known data for training. Such predictions can also be corroborated by using data that are not incorporated while training the model. For example, methods that identify elements functional in a certain cell-type, based only on genomic sequence, can look for enrichment of the predictions in loci of genes expressed in that cell type. Similarly, sequence conservation and epigenetic modifications can be used as additional sources of information to validate predictions.

A reliable validation of regulatory elements is the use of a reporter gene assay. In such an assay, a plasmid carrying the putative regulatory element and a reporter gene is introduced into cells of the same cell-type for which the prediction was made. The predicted role of the element—enhancing, silencing, or enhancer-blocking—dictates the structure of the plasmid. To test for an enhancer role, the putative element is placed upstream of a minimal promoter and the reporter gene. The minimal promoter is not sufficient to drive expression of the reporter gene without the presence of an enhancer. In the case of a silencer, the features of the plasmid are the same with one exception: here, the minimal promoter is replaced by a constitutive promoter (a promoter that is always active unless silenced by other factors). In the former case, the tested element is positive if the reporter gene is highly expressed; whereas in the case of the constitutive promoter, the tested element is positive if the gene is repressed. To test for an enhancer-blocker role, the putative element is inserted between a known enhancer and a corresponding minimal promoter. In such an assay, if the gene is not expressed, the putative element can be said to have an enhancer-blocker role. See Carey and Smale for further details on constructing these assays (Carey and Smale 1999).

In all cases, expression of the reporter gene indicates whether the putative element is indeed functional. However, this approach suffers from three major limitations. First, it is not easily scalable. Only a few elements can be tested at a time and the number is even smaller when the experiments must be done in an animal. Second, animal testing is restricted to a few model organisms and may not

be possible in the organism of interest, e.g., in humans. Finally, these assays test elements outside of their native contexts. As a result the construct may not be able to emulate the chromatin structure and other genomic regions vital for the function of the regulatory element.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and by an Early Career Fellowship to L.N. from the Wellcome Trust-DBT India Alliance.

REFERENCES

- Abbott A. 2010. Project set to map marks on genome. *Nature* **463**: 596–597.
- Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. 2003. Computational detection of *cis*-regulatory modules. *Bioinformatics* **19**: ii5–ii14.
- Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**: e234.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* **322**: 881–888.
- Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsson KR, Cazier JB, Sainz J, et al. 2006. A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**: 652–658.
- Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.
- Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* **3**: e254.
- Balikova I, de Ravel T, Ayuso C, Thienpont B, Casteels I, Villaverde C, Devriendt K, Fryns JP, Vermeesch JR. 2011. High frequency of submicroscopic chromosomal deletions in patients with idiopathic congenital eye malformations. *Am J Ophthalmol* **151**: 1087–1094 e1045.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.
- Batzner MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.
- Bell AC, West AG, Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**: 387–396.
- Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. *Nucleic Acids Res* **40**: D48–D53.
- Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci* **99**: 757–762.
- Bernardi G. 2000. The compositional evolution of vertebrate genomes. *Gene* **259**: 31–43.
- Bernardi G. 2005. Isochores. *eLS*.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045–1048.
- Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242–2246.
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.
- Bishop CE, Whitworth DJ, Qin Y, Agoulunik AI, Agoulunik IU, Harrison WR, Behringer RR, Overbeek PA. 2000. A transgenic insertion upstream of *sox9* is associated with dominant XX sex reversal in the mouse. *Nat Genet* **26**: 490–494.
- Blake DJ, Weir A, Newey SE, Davies KE. 2002. Function and genetics of dystrophin and dystrophin-related proteins in muscle. *Physiol Rev* **82**: 291–329.
- Blanchette M, Bataille AR, Chen X, Poitras C, Laganier J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**: 656–668.
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–715.
- Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**: 806–810.
- Boby T, Patch AM, Aves SJ. 2005. TRbase: A database relating tandem repeats to disease genes for the human genome. *Bioinformatics* **21**: 811–816.
- Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**: 1391–1394.
- Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**: 456–465.
- Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317**: 815–819.
- Brannan CI, Dees EC, Ingram RS, Tilghman SM. 1990. The product of the H19 gene may function as an RNA. *Mol Cell Biol* **10**: 28–36.
- Bray N, Pachter L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res* **14**: 693–699.
- Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**: 721–731.
- Bulger M, Groudine M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**: 327–339.
- Bulyk ML, Johnson PL, Church GM. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* **30**: 1255–1261.
- Bunt J, Hasselt NE, Zwijsen DA, Hamdi M, Koster J, Versteeg R, Kool M. 2011. OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells. *Int J Cancer* **131**: E21–E32.
- Buttgereit D. 1993. Redundant enhancer elements guide beta 1 tubulin gene expression in apodemes during *Drosophila* embryogenesis. *J Cell Sci* **105**: 721–727.
- Carey M, Smale ST. 1999. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Carninci P. 2008. Non-coding RNA transcription: Turning on neighbours. *Nat Cell Biol* **10**: 1023–1024.

- Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215–220.
- Chen X, Tompa M. 2010. Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotechnol* 28: 567–572.
- Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces genome database: The genomics resource of budding yeast. *Nucleic Acids Res* 40: D700–D705.
- Chung JH, Whiteley M, Felsenfeld G. 1993. A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* 74: 505–514.
- Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglu S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15: 901–913.
- Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, Bouffard G, Young A, Masiello C, Green ED, Wolfsberg TG, et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci* 101: 992–997.
- Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Res* 14: 1188–1190.
- Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19: 24–32.
- Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglu S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6: e1001025.
- de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 15: 1061–1072.
- Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: Conservation and turnover. *Mol Biol Evol* 19: 1114–1121.
- Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences—An unexpected feature of mammalian genomes. *Nat Rev Genet* 6: 151–157.
- Derrien T, Johnson R, Bussotti G, Tanzer A, Diebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* 22: 1775–1789.
- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152–154.
- Donahue TR, Hines OJ. 2009. CXCR2 and RET single nucleotide polymorphisms in pancreatic cancer. *World J Surg* 33: 710–715.
- Dubchak I. 2007. Comparative analysis and visualization of genomic sequences using VISTA browser and associated computational tools. *Methods Mol Biol* 395: 3–16.
- Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA. 2007. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* 80: 241–252.
- Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struwing JP, Morrison J, Field H, Luben R, et al. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447: 1087–1093.
- Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L. 2008. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* 123: 177–187.
- Elnitski LL, Shah P, Moreland RT, Umayam L, Wolfsberg TG, Baxevas AD. 2007. The ENCODEdb portal: Simplified access to ENCODE Consortium data. *Genome Res* 17: 954–959.
- Emission ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. 2005. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature* 434: 857–863.
- Emorine L, Kuehl M, Weir L, Leder P, Max EE. 1983. A conserved sequence in the immunoglobulin J kappa-C kappa intron: Possible enhancer element. *Nature* 304: 447–449.
- Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30: 233–237.
- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43–49.
- Fajkus J, Sykorova E, Leitch AR. 2005. Telomeres in evolution and evolution of telomeres. *Chromosome Res* 13: 469–479.
- Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 312: 276–279.
- Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. *Nucleic Acids Res* 40: D84–D90.
- Friedman N, Ninio M, Pe'er I, Pupko T. 2002. A structural EM algorithm for phylogenetic inference. *J Comput Biol* 9: 331–353.
- Frith MC, Hansen U, Weng Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 17: 878–889.
- Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* 31: 3666–3668.
- Frith MC, Spouge JL, Hansen U, Weng Z. 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res* 30: 3214–3224.
- Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2011. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* 39: D876–D882.
- Furey TS. 2012. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* 13: 840–852.
- Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25: i54–i62.
- Gaszner M, Felsenfeld G. 2006. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7: 703–713.
- Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: Protein homology by domain architecture. *Genome Res* 12: 1619–1623.
- Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17: 877–885.
- Gordan R, Narlikar L, Hartemink AJ. 2010. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res* 38: e90.
- Gotea V, Ovcharenko I. 2008. DiRE: Identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* 36: W133–W139.
- Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* 20: 565–577.
- Grolleau-Julius A, Ray D, Yung RL. 2010. The role of epigenetics in aging and autoimmunity. *Clin Rev Allergy Immunol* 39: 42–50.
- Grosveld F, van Assendelft GB, Greaves DR, Kollias G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* 51: 975–985.
- Gupta M, Liu JS. 2005. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci* 102: 7079–7084.
- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458: 223–227.
- Hagerman R, Hoem G, Hagerman P. 2010. Fragile X and autism: Intertwined at the molecular level leading to targeted treatments. *Mol Autism* 1: 12.
- Hahn S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* 11: 394–403.
- Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431: 99–104.

- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459: 108–112.
- Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39: 311–318.
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 106: 9362–9367.
- Hon G, Ren B, Wang W. 2008. ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol* 4: e1000201.
- Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* 18: R195–R201.
- Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* 22: 1036–1046.
- Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P, et al. 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* 43: 621–629.
- Huynen MA, Snel B, van Noort V. 2004. Comparative genomics for reliable protein-function prediction from genomic data. *Trends Genet* 20: 340–344.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945.
- Ivan A, Halfon MS, Sinha S. 2008. Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. *Genome Biol* 9: R22.
- Johansson O, Alkema W, Wasserman WW, Lagergren J. 2003. Identification of functional clusters of transcription factor binding motifs in genome sequences: The MSCAN algorithm. *Bioinformatics* 19: 169–176.
- Johnson JM, Edwards S, Shoemaker D, Schadt EE. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 21: 93–102.
- Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson GE, Gottgens B, Halfon MS, Sinha S. 2009. Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. *Dev Cell* 17: 568–579.
- Kantorovitz MR, Robinson GE, Sinha S. 2007. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* 23: i249–i255.
- Kapranov P, Cheng J, Dike S, Nix DA, Duttgupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316: 1484–1488.
- Karolchik D, Hinrichs AS, Kent WJ. 2011. The UCSC Genome Browser. *Curr Protoc Hum Genet* Chapter 18: Unit18 16. doi: 10.1002/0471142905.hg1806s71.
- Kasper LH, Fukuyama T, Biesen MA, Boussouar F, Tong C, de Pauw A, Murray PJ, van Deursen JM, Brindle PK. 2006. Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development. *Mol Cell Biol* 26: 789–809.
- Kazemian M, Zhu Q, Halfon MS, Sinha S. 2011. Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res* 39: 9463–9472.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.
- Kelly BL, Locksley RM. 2000. Coordinate regulation of the IL-4, IL-13, and IL-5 cytokine cluster in Th2 clones revealed by allelic expression patterns. *J Immunol* 165: 2982–2986.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* 12: 996–1006.
- Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465: 182–187.
- Kimura-Yoshida C, Kitajima K, Oda-Ishii I, Tian E, Suzuki M, Yamamoto M, Suzuki T, Kobayashi M, Aizawa S, Matsuo I. 2004. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development* 131: 57–71.
- Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217: 624–626.
- King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science* 164: 788–798.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107–116.
- King OD, Roth FP. 2003. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res* 31: e116.
- Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, de Lima RL, Daack-Hirsch S, Sander A, et al. 2002. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32: 285–289.
- Krivan W, Wasserman WW. 2001. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* 11: 1559–1566.
- Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42: 631–634.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* 34: 77–137.
- Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12: 1725–1735.
- Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* 424: 147–151.
- Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. 2002. Locus control regions. *Blood* 100: 3077–3086.
- Li Z, Schug J, Tuteja G, White P, Kaestner KH. 2011. The nucleosome map of the mammalian liver. *Nat Struct Mol Biol* 18: 742–746.
- Liang S, Moghimi B, Yang TP, Strouboulis J, Bungert J. 2008. Locus control region mediated regulation of adult beta-globin gene expression. *J Cell Biochem* 105: 9–16.
- Lichtarge O, Sowa ME. 2002. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol* 12: 21–27.
- Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic regulatory clusters in *Drosophila*. *Genome Res* 13: 579–588.
- Lin Q, Chen Q, Lin L, Smith S, Zhou J. 2007. Promoter targeting sequence mediates enhancer interference in the *Drosophila* embryo. *Proc Natl Acad Sci* 104: 3237–3242.
- Lin T, Ray P, Sandve GK, Uguroglu S, Xing EP. 2008. BayCis: A Bayesian hierarchical HMM for cis-regulatory module decoding in metazoan genomes. *Research in Computational Molecular Biology Lecture Notes in Computer Science* Vol. 4955, pp. 66–81. Springer, Berlin.
- Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478: 476–482.
- Ling J, Ainol L, Zhang L, Yu X, Pi W, Tuan D. 2004. HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *J Biol Chem* 279: 51704–51713.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.
- Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288: 136–140.
- Loots GG, Ovcharenko I. 2005. Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res* 33: W56–W64.
- Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. *Science* 333: 1019–1024.

- Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwazd M, Hao L, He S, Hurwitz DI, Jackson JD, et al. 2007. CDD: A conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35: D237–D240.
- Marchler-Bauer A, Bryant SH. 2004. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res* 32: W327–W331.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al. 2011. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39: D225–D229.
- Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* 13: 2507–2518.
- Margulies EH, Chen CW, Green ED. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet* 22: 187–193.
- Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* 17: 760–774.
- Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness. *Genome Biol* 10: R124.
- Masuya H, Sagai T, Wakana S, Moriwaki K, Shiroishi T. 1995. A duplicated zone of polarizing activity in polydactylous mouse mutants. *Genes Dev* 9: 1645–1653.
- Mathew CG. 2008. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet* 9: 9–14.
- Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCOMP: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34: D108–D110.
- Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16: 1046–1047.
- McClintock B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* 21: 197–216.
- Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: Insights into functions. *Nat Rev Genet* 10: 155–159.
- Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu Rev Genomics Hum Genet* 5: 15–56.
- Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, et al. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9: e1001046.
- Narlikar L, Gordan R, Hartemink AJ. 2007. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol* 3: e215.
- Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B, Becker TS. 2009. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol* 327: 526–540.
- Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302: 413.
- Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* 431: 988–993.
- O'Neill MJ, Lawton BR, Mateos M, Carone DM, Ferreri GC, Hrbek T, Meredith RW, Reznick DN, O'Neill RJ. 2007. Ancient and continuing Darwinian selection on insulin-like growth factor II in placental fishes. *Proc Natl Acad Sci* 104: 12404–12409.
- Ong CT, Corces VG. 2011. Enhancer function: New insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* 12: 283–293.
- Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, Hardison RC, Stubbs L, Miller W. 2005a. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res* 15: 184–194.
- Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L. 2004a. zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res* 14: 472–477.
- Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005b. Evolution and functional classification of vertebrate gene deserts. *Genome Res* 15: 137–145.
- Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. 2004b. ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* 32: W280–W286.
- Ovcharenko I, Stubbs L, Loots GG. 2004c. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics* 84: 890–895.
- Pacini F, Castagna MG, Cipri C, Schlumberger M. 2010. Medullary thyroid carcinoma. *Clin Oncol (R Coll Radiol)* 22: 475–485.
- Palin K, Taipale J, Ukkonen E. 2006. Locating potential enhancer elements by comparative genomics using the EEL software. *Nat Protoc* 1: 368–374.
- Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. 2008. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 18: 1814–1828.
- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444: 499–502.
- Perry MW, Boettiger AN, Levine M. 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proc Natl Acad Sci* 108: 13570–13575.
- Pfeifer K, Leighton PA, Tilghman SM. 1996. The structural H19 gene is required for transgene imprinting. *Proc Natl Acad Sci* 93: 13876–13883.
- Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res* 20: 110–121.
- Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443: 167–172.
- Portela A, Esteller M. 2010. Epigenetic modifications and human disease. *Nat Biotechnol* 28: 1057–1068.
- Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16: 855–863.
- Puliti A, Covone AE, Biccocchi MP, Bolino A, Lerone M, Martucciello G, Jasonni V, Romeo G. 1993. Deleted and normal chromosome 10 homologs from a patient with Hirschsprung disease isolated in two cell hybrids through enrichment by immunomagnetic selection. *Cytogenet Cell Genet* 63: 102–106.
- Queck XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2014. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* doi: 10.1093/nar/gku988.
- Queck XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2014. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*. doi: 10.1093/nar/gku988.
- Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, et al. 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40: 1341–1347.
- Ramos YF, Hestand MS, Verlaan M, Krabbendam E, Ariyurek Y, van Galen M, van Dam H, van Ommen GJ, den Dunnen JT, Zantema A, et al. 2010. Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Res* 38: 5396–5408.
- Rebeiz M, Reeves NL, Posakony JW. 2002. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc Natl Acad Sci* 99: 9888–9893.
- Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, Gaszner M, Felsenfeld G. 2002. Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci* 99: 6883–6888.
- Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L, et al. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72: 257–268.
- Roeder RG. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* 21: 327–335.
- Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, et al. 2012. ENCODE whole-genome data in the UCSC Genome Browser: Update 2012. *Nucleic Acids Res* 40: D912–D917.

- Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res* 38: D620–D625.
- Roth TL, Lubin FD, Sodhi M, Kleinman JE. 2009. Epigenetic mechanisms in schizophrenia. *Biochim Biophys Acta* 1790: 869–877.
- Russell LB, Montgomery CS, Raymer GD. 1982. Analysis of the albino-locus region of the mouse: IV. Characterization of 34 deficiencies. *Genetics* 100: 427–453.
- Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, Shafer A, Kawamoto J, Hall R, Mack J, Dorschner MO, et al. 2004. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci* 101: 16837–16842.
- Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32: D91–D94.
- Satterlee JS, Schubeler D, Ng HH. 2010. Tackling the epigenome: Challenges and opportunities for collaboration. *Nat Biotechnol* 28: 1039–1044.
- Savic D, Ye H, Aneas I, Park SY, Bell GI, Nobrega MA. 2011. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res* 21: 1417–1425.
- Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328: 1036–1040.
- Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132: 887–898.
- Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet* 7: 301–313.
- Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* 13: 103–107.
- Sha BY, Yang TL, Zhao LJ, Chen XD, Guo Y, Chen Y, Pan F, Zhang ZX, Dong SS, Xu XH, et al. 2009. Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. *J Hum Genet* 54: 199–202.
- Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* 6: e65.
- Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1: e67.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034–1050.
- Sinha S, Blanchette M, Tompa M. 2004a. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 5: 170.
- Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. 2004b. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. *BMC Bioinformatics* 5: 129.
- Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* 72: 449–479.
- Solomon MJ, Larsen PL, Varshavsky A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53: 937–947.
- Song L, Zhang Z, Grassefder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* 21: 1757–1767.
- Staden R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* 12: 505–519.
- Stamatoyannopoulos G. 1991. Human hemoglobin switching. *Science* 252: 383.
- Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450: 219–232.
- Storz G, Altuvia S, Wassarman KM. 2005. An abundance of RNA regulators. *Annu Rev Biochem* 74: 199–217.
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848–853.
- Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* 21: 1139–1149.
- Taher L, Narlikar L, Ovcharenko I. 2012. CLARE: Cracking the language of regulatory elements. *Bioinformatics* 28: 581–583.
- Tena JJ, Alonso ME, de la Calle-Mustienes E, Splinter E, de Laat W, Manzanares M, Gomez-Skarmeta JL. 2011. An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat Commun* 2: 310.
- Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788–793.
- Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell M, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39: 31–40.
- Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, et al. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 39: 984–988.
- Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. 2009. FlyBase: Enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res* 37: D555–D559.
- Uchikawa M, Takemoto T, Kamachi Y, Kondoh H. 2004. Efficient identification of regulatory sequences in the chicken genome by a powerful combination of embryo electroporation and genome comparison. *Mech Dev* 121: 1145–1158.
- Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* 4: 251–262.
- Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* 18: 1051–1063.
- van Baren MJ, Koebe BC, Brent MR. 2007. Using N-SCAN or TWINSKAN to predict gene structures in genomic DNA sequences. *Curr Protoc Bioinformatics* Chapter 4: Unit 4.8. doi: 10.1002/0471250953.l00408s20.
- Van Loop P, Aerts S, Thienpont B, De Moor B, Moreau Y, Marynen P. 2008. ModuleMiner - improved computational detection of cis-regulatory modules: Are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol* 9: R66.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* 291: 1304–1351.
- Vinogradov AE, Anatskaya OV. 2007. Organismal complexity, cell differentiation and gene expression: Human over mouse. *Nucleic Acids Res* 35: 6350–6356.
- Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA. 2009a. Functional autonomy of distant-acting human enhancers. *Genomics* 93: 509–513.
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009b. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457: 854–858.
- Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA enhancer browser—A database of tissue-specific human enhancers. *Nucleic Acids Res* 35: D88–D92.
- Wagner A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 15: 776–784.
- Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, et al. 2011. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474: 390–394.

- Wang QF, Prabhakar S, Chanan S, Cheng JF, Rubin EM, Boffelli D. 2007. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol* 8: R1.
- Wang Y, Harvey CB, Pratt WS, Sams VR, Sarnier M, Rossi M, Auricchio S, Swallow DM. 1995. The lactase persistence/non-persistence polymorphism is controlled by a *cis*-acting element. *Hum Mol Genet* 4: 657–662.
- Wasserman NF, Aneas I, Nobrega MA. 2010. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a MYC enhancer. *Genome Res* 20: 1191–1197.
- Wasserman WW, Fickett JW. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278: 167–181.
- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.
- Williams A, Spilianakis CG, Flavell RA. 2010. Interchromosomal association and gene regulation in trans. *Trends Genet* 26: 188–197.
- Wittkopp PJ, Kalay G. 2011. *Cis*-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* 13: 59–69.
- Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3: e7.
- Wu J, Xie J. 2010. Hidden Markov model and its applications in motif findings. *Methods Mol Biol* 620: 405–416.
- Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y. 2014. NONCODEv4: Exploring the world of long non-coding RNA genes. *Nucleic Acids Res* 42(Database issue): D98–D103.
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434: 338–345.
- Yang L, Froberg JE, Lee JT. 2014. Long noncoding RNAs: Fresh perspectives into RNA world. *Trends Biochem Sci* 39: 35–43.
- Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, de la Cruz N, Duong A, Fang R, et al. 2012. WormBase 2012: More genomes, more data, new website. *Nucleic Acids Res* 40: D735–D741.
- Yusufzai TM, Felsenfeld G. 2004. The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc Natl Acad Sci* 101: 8620–8624.
- Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21: 236–239.

Cold Spring Harbor Protocols

Identification and Computational Analysis of Gene Regulatory Elements

Leila Taher, Leelavati Narlikar and Ivan Ovcharenko

Cold Spring Harb Protoc; doi: 10.1101/pdb.top083642

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

- [Alignment of Sequences](#) (46 articles)
- [Alignment of Sequences, general](#) (24 articles)
- [Bioinformatics/Genomics, general](#) (192 articles)
- [Computational Biology](#) (100 articles)
- [Gene Prediction](#) (12 articles)
- [Genetic Variation](#) (86 articles)
- [Genome Analysis](#) (191 articles)
- [Genome Wide Association Studies \(GWAS\)](#) (11 articles)
- [High-Throughput Analysis, general](#) (155 articles)
- [Phylogenetic Prediction](#) (19 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>